

**Characterisation of polymorphic inversions and complex regions
in the human genome from long-read data**

Ricardo Emanuel Moreira Pinhal

ADVERTIMENT. L'accés als continguts d'aquesta tesi queda condicionat a l'acceptació de les condicions d'ús establertes per la següent llicència Creative Commons:  <https://creativecommons.org/licenses/?lang=ca>

ADVERTENCIA. El acceso a los contenidos de esta tesis queda condicionado a la aceptación de las condiciones de uso establecidas por la siguiente licencia Creative Commons:  <https://creativecommons.org/licenses/?lang=es>

WARNING. The access to the contents of this doctoral thesis it is limited to the acceptance of the use conditions set by the following Creative Commons license:  <https://creativecommons.org/licenses/?lang=en>



UNIVERSITAT AUTÒNOMA DE BARCELONA

DOCTORAL THESIS

Characterisation of polymorphic inversions and complex regions in the human genome from long-read data

Author: Ricardo Emanuel Moreira Pinhal

Co-Supervisors: Mario Cáceres Aguilar and Marta Puig Font

Departament de Genètica i de Microbiologia
Facultat de Biociències

February 2026

Characterisation of polymorphic inversions and complex regions in the human genome from long-read data

Memòria presentada per Ricardo
Emanuel Moreira Pinhal per optar al
grau de Doctor en Genètica per la
Universitat Autònoma de Barcelona

Autor:

Ricardo Emanuel Moreira Pinhal

Codirectors:

Mario Cáceres Aguilar

Marta Puig Font

Cerdanyola del Vallés, Febrer 2026

Contents

Acknowledgements	v
Abstract	ix
List of Figures	xi
List of Tables	xiii
Glossary	xix
1 Introduction	1
1.1 Genetic variation	1
1.1.1 Variation discovery	4
1.2 Inversions	7
1.2.1 An historical perspective	7
1.2.2 A mechanistic view	10
1.3 Functional consequences	15
1.3.1 Indirect effects	16
1.3.2 Direct effects	23
1.4 Inversion detection	32
1.4.1 Inversion validation	34
1.4.2 New methodologies	37
1.4.3 Reference genome	46
1.5 Human variation resources	48

1.5.1	Large scale initiatives to characterise human variation	49
1.5.2	Inversion data and catalogues	53
1.6	Mosaicism	57
1.7	Complex genomic rearrangements (CGRs)	59
1.8	Objectives	62
2	Materials and methods	65
2.1	Datasets	65
2.1.1	Inversions	65
2.1.2	Samples	67
2.2	Methods	70
2.2.1	Merging of variants	70
2.2.2	Genotyping performance evaluation	75
2.2.3	AF calculation	76
2.2.4	Population differentiation	76
2.2.5	Genotype quality control	77
2.2.6	Hardy–Weinberg equilibrium (HWE)	78
2.2.7	Linkage disequilibrium (LD) analysis	78
2.2.8	Manual sequence analysis	79
2.2.9	Complex regions analysis	79
3	Results	91
3.1	Genotyping of inversions using long-read data	91
3.2	Application of GeONTIpe to other datasets	140
3.3	Generation of a complete inversion catalogue	144
3.3.1	Merge of variants	144
3.3.2	Selection of variants for further analysis	147
3.3.3	Breakpoint definition	149
3.3.4	Inversion region genotyping	152

3.3.5	Inversion characteristics analysis	159
3.3.6	Inversion frequency in human populations	162
3.3.7	Size effects	167
3.3.8	Genotype validation	169
3.3.9	Hardy–Weinberg equilibrium (HWE) analysis	174
3.3.10	Linkage disequilibrium (LD) with nearby variants	177
3.4	Complex genomic regions	182
3.4.1	HsInv0822 Region	182
3.4.2	HsInv0786 Region	184
3.4.3	Pipeline for the analysis of complex structural haplotypes	188
3.4.4	HsInv0401 Region	192
3.4.5	HsInv0390 Region	196
3.4.6	HsInv0231 Region	200
4	Discussion	205
4.1	GeONTIpe advantages and limitations	207
4.2	CHARONTE advantages and limitations	213
4.3	Inversion catalogue	217
4.3.1	Inversion characteristics	219
4.3.2	Genotype accuracy	220
4.3.3	LD analysis	223
4.3.4	Inversion frequency	226
4.3.5	F_{ST}	227
4.4	Detailed analysis of candidate regions	229
4.4.1	HsInv0012-0659	229
4.4.2	HsInv0231	231
4.4.3	HsInv0348	232
4.4.4	HsInv0390	235
4.4.5	HsInv0394-816-822	236

4.4.6	HsInv0401	238
4.4.7	HsInv0786	239
4.4.8	HsInv1117	243
4.5	Future perspectives	245
5	Conclusions	248
	Bibliography	250
	Appendix A	305

Acknowledgement

Esta es la última página que escribo como parte de esta tesis, no porque sea la más extensa ni la más compleja, sino porque considero que merece una atención especial. Agradecer puede parecer algo superfluo o incluso insignificante; sin embargo, sin el apoyo de las personas que me han acompañado a lo largo de estos años, este trabajo no habría sido posible.

En primer lugar, me gustaría agradecer a mis directores, Marta y Mario, que me han acompañado durante todo este viaje. Por suerte —o por desgracia para ellos— les ha tocado leer cientos de páginas escritas por alguien que, desde luego, no será recordado por su estilo literario. Soy plenamente consciente de que nada de esto habría sido posible sin la oportunidad que se me ofreció dentro del grupo. A pesar de mi peculiar currículum académico, espero haber demostrado que aceptarme no fue, después de todo, una mala decisión. Durante estos años, además de experimentar lo que significa formar parte de un grupo de investigación ya consolidado, también he tenido la oportunidad de ser partícipe del nacimiento de otro. Aunque la financiación todavía no acompañe, creo firmemente que se sabrá reconocer su potencial y se le dará la oportunidad de florecer.

Aunque mis directores han tenido un papel fundamental en este camino, los verdaderos protagonistas han sido las personas con las que he compartido quejas, dramas y chascarrillos. Algunas de estas personas, como Elena o Aina, ya no forman parte de mi día a día, pero tras haber pasado antes por lo mismo supieron servirme de guía y doy gracias por eso. Otras, en cambio, siguen ahí, como Elías y María, que se convirtieron en mis patitos de goma personales durante interminables jornadas de café en el parking del IBB. No puedo imaginar el camino recorrido sin vosotros y solo espero que también me recordéis con el mismo cariño. Elías, prácticamente ya estás al otro lado, así que solo puedo desearte mucha suerte. María, tú estás a las puertas del final del camino; a ti sí que puedo desearte ánimo en esta recta final y decirte que voy a estar ahí para lo que necesites. Odei y Adrià a vosotros aún os queda largo camino por recorrer, pero nunca está de más desearos mucho ánimo.

Para mis compañeros de UAB, Jordi e Iris, independientemente de lo que nos depara el futuro quiero agradecer esas sobremesas inacabables de conversaciones superfluas de las que tanto hemos disfrutado.

También quiero agradecer a mis Tocinetes, con quienes he compartido ya más de una década de mi vida. Aunque ya no disfrutemos de horas infinitas de cartas como antaño o aunque algunos hayan tomado caminos lejanos para construir su vida, cada reencuentro con vosotros me devuelve con cariño a momentos en que todo era mucho más sencillo. Solo por eso no me queda más que decir que gracias.

Por último, quizá el agradecimiento más importante está dirigido a mi familia. Estos momentos están siendo los más difíciles debido a las circunstancias que estamos atravesando, sin embargo sé que todo va acabar saliendo bien y que nos ayudará a apreciar esos pequeños momentos que vivimos a diario. Para todos ellos voy a dirigir unas palabras que espero que les alegre y reconforte.

Felipe, espero que para el momento en que este texto esté concluido, seas capaz de leerlo y comprender lo que en él se expresa. Quiero decirte que gracias por estar ahí, me has demostrado que es la valentía, independientemente de las dificultades que se han cruzado en tu vida las has sobrepasado, convirtiéndote en un ejemplo a seguir para mí y para tus hijas. Esta valentía la he visto reflejada en vosotras, Daniela, Diana y Sara, que a pesar de pasar por lo imposible habéis sido capaces de manteneros de pie. Espero haber demostrado yo también esa valentía cuando ha hecho falta y espero haber estado a la altura de cuando más me habéis necesitado.

Este agradecimiento no estaría completo sin mencionar a mi compañera de viajes, mi hermana, la cual sé que se une a cualquier aventura que se me pueda pasar por la cabeza. Quizá la frase que mejor puede resumir mi agradecimiento es: "Japón ahí vamos".

Creo que para el siguiente agradecimiento es necesario un cambio de idioma, así que espero no liarla mucho. Papa e mama. Sei que foi um percurso difícil até chegarmos a este momento. Emigrar para Andorra, deixando para trás família e amigos, à procura de um futuro mais promissor, tanto para vocês como para mim não deve

ter sido nada fácil. Espero ter demonstrado que esse esforço não foi em vão e espero continuar a prová-lo todos os dias, para que possam dizer, de cabeça erguida, que se sentem orgullosos de mim. Assim, resta-me apenas dizer: obrigado.

Hay muchas más personas a las que agradecer, pero las ganas de escribir se van agotando y la memoria empieza a fallar, así que voy a limitarme a un último agradecimiento. Gracias a MI. A lo largo de los años ha habido innumerables ocasiones para rendirse y acomodarse, y no lo has hecho. Te prometiste que llegarías a obtener el doctorado, y aquí estas, en la recta final. Solo tú sabes realmente las horas y el esfuerzo que has dedicado, los retos que has afrontado y superado, por mucho que costara. Aunque el camino aún no ha terminado, sí puedo decir que esta etapa ha llegado a su fin. Los próximos desafíos no serán menores, así que solo queda decir: ánimo a mi yo del futuro.

Abstract

Structural variants (SVs) constitute one of the most significant sources of genetic diversity, although their study has historically lagged behind due to technical limitations. Inversions are particularly challenging due to their balanced nature and the frequent presence of inverted repeats (IRs) at their breakpoints (BPs). Recent technological advances have made it possible to overcome many of these limitations, expanding considerably the catalogue of inversions and other SVs in humans. Nevertheless, most studies have only analysed a small number of individuals, which limits the comprehensive evaluation of the functional effects of these variants. To address this gap, we merged inversion candidates from multiple studies employing diverse new genomic technologies, thus generating the most exhaustive inversion catalogue to date. In addition, to enable accurate genotyping of these inversions, we developed GeONTIpe, a tool that uses Oxford Nanopore Technologies (ONT) long-read data to detect inversion orientation by identifying reads spanning inversion breakpoints with probe sequences located inside and outside the inverted regions. Using this approach, we analysed 1,235 variants, including some flanked by IRs up to 190 kb in length, in 1,409 samples, and identified 680 polymorphic inversions showing both orientations, 191 of which exhibited a global minor allele frequency (MAF) larger than 1%. Moreover, 740 regions contained additional SVs in at least one sample, with 362 of them exhibiting variants in more than 1% of samples, highlighting the complexity present in repetitive regions. Several quality controls were performed to assess the reliability of the inversion genotypes, including Mendelian consistency, comparisons with experimental genotypes, linkage disequilibrium (LD) analysis with surrounding SNPs, and Hardy–Weinberg equilibrium, identifying candidate variants that require further validation. One possible explanation for some of the problematic genotypes is the presence of haplotypes with different structural configurations, consistent with the multiple SVs found in some regions. However, GeONTIpe has a limited ability to correctly characterise those regions. In fact, in depth manual analyses revealed numerous haplotypes in several regions, including 19 haplotypes within the functionally relevant HsInv0786 locus. To scale up

this analysis, we developed CHARONTE, which is based on the mapping of unique k-mers to ONT long reads, and exploits their position, orientation, and spacing to reconstruct individual structural haplotypes of multiple samples in complex regions of the genome, allowing population-level analyses of haplotype frequency and distribution. The application of this new methodology to characterise three complex regions in a set of 213-455 samples identified 18 structural haplotypes for the HsInv0401 region, 8 for HsInv0390, and 5 for HsInv0231, some of which were not previously described. These structural haplotypes harbour deletions, insertions, inversions, and inverted duplications, including a recurrence event for the HsInv0231 inversion. Also, these haplotypes were largely in agreement with previous studies, suggesting that CHARONTE is a valuable resource for the detailed characterisation of complex genomic regions. Therefore, long-read sequencing analysis offers a great potential for resolving and genotyping multiple genomic regions in large numbers of individuals, facilitating a deeper understanding of the diversity of SVs in the human genome and their functional consequences.

List of Figures

1.1	Overview of the different forms of genetic variation	2
1.2	Classification of inversions, by location (A) or by mechanism of generation (B, C and D)	14
1.3	Paracentric and pericentric inversions and their effect on recombination	18
1.4	Genomic structure of two human inversion supergenes	22
1.5	Inversions associated with diseases and functional effects	26
1.6	Genotyping inverted alleles using different techniques	43
1.7	Haplotype structures affecting <i>NOTCH2NL</i> and <i>NBPF</i> genes	62
2.1	Sample population assignment	69
2.2	Schematic overview of CHARONTE	80
3.1	Distribution of variants according to genotyping outcome	142
3.2	Sharing variants by study	148
3.3	Origin of variants with less than 1% allele frequency	149
3.4	Inversion region genotypes by sample	153
3.5	Genotype success by dataset	154
3.6	Genotyping success by number of reads and N50	155
3.7	Genotyping success according to distances between probes	157
3.8	Distribution of inversions and breakpoints sizes	160
3.9	Distribution of inversions across chromosomes	161
3.10	Correlation between inversion distribution and chromosome size across the genome	162

3.11 Inversion allele frequency spectrum	163
3.12 Inversion heterozygote fraction by ancestry	164
3.13 Global and population-specific MAF for polymorphic inversions	166
3.14 Distribution of MAF according to the breakpoint size/inversion length ratio	169
3.15 Allelic imbalance detection	176
3.16 Linked variants per inversion	178
3.17 Linked variants per ancestry	179
3.18 HsInv0822 genomic region	183
3.19 HsInv0786 genomic region	185
3.20 Workflow overview of CHARONTE	189
3.21 HsInv0401 genomic region	193
3.22 Synteny plot comparing haplotypes from HsInv0401	195
3.23 HsInv0390 genomic region	196
3.24 Synteny plot comparing haplotypes from HsInv0390	198
3.25 Analysis of a read from region HsInv0390 in individual HG00096	199
3.26 HsInv0231 genomic region	201
3.27 Synteny plot comparing haplotypes from HsInv0231	203
3.28 Inversion recurrence of HsInv0231	204

List of Tables

2.1	Information of analysed studies	67
2.2	Sample sources	68
3.1	Overview of inversion candidate types	140
3.2	Review of candidate merging by study	146
3.3	Complex SDs regions	151
3.4	N50 and number of reads by dataset	155
3.5	Genotyping success by inversion	156
3.6	Features of additional SVs in inversion regions	158
3.7	Summary of inversions with Global Fixation index (F_{ST}) differences .	167
3.8	Verification of Mendelian inconsistencies	170
3.9	Verification with experimental inconsistencies	173
3.10	Assessment of deviation from HWE	175
3.11	Haplotype distribution of HsInv0822	184
3.12	Haplotypes found for the HsInv0786 region	187
3.13	Summary of the haplotypes in region HsInv0401 identified by CHARONTE194	

Glossary

1000GP 1000 Genomes Project. 6, 39, 44, 49–55, 63, 66–69, 74, 76–78, 83, 140, 146–148, 171, 214, 244

AF allele frequency. ii, 21, 49, 51, 71, 72, 76, 78, 142, 144–146, 157, 162, 165, 167, 168, 180, 220, 226, 227, 235, 237

AFLPs amplified fragment length polymorphisms. 8

BAC bacterial artificial chromosome. 32, 46

BISRS break-induced serial replication slippage. 12, 13

BPs breakpoints. ix, 2, 7, 11–15, 20, 22, 23, 27, 33–36, 38, 40, 43, 45, 55, 59, 66, 70–75, 77, 140, 141, 144–146, 149–152, 155, 156, 160, 168, 180, 184, 185, 192, 202, 207, 209, 213, 219, 224, 225, 235, 236, 242, 243

CCS Circular Consensus Sequencing. 41

CGH comparative genomic hybridisation. 5

CGRs complex genomic rearrangements. ii, 2, 48, 59–61

CLR continuous long reads. 41, 51

CNVs copy number variants. 2, 4, 5, 9, 20, 29, 30, 50, 58, 60, 192–194, 239

CSVs complex structural variants. 59, 60

CTCF CCCTC-binding factor. 29

ddPCR digital droplet PCR. 36, 55, 182, 219, 235, 236

DGV Database of Genomic Variants. 52, 53

DNA deoxyribonucleic acid. 2, 4, 5, 7, 9–15, 33, 35, 38–44, 46, 47, 49

DSB double-strand break. 10, 12, 13

eQTL expression Quantitative Trait Loci. 27, 56

F_{ST} Fixation index. iii, xiii, 21, 76, 77, 165–167, 227, 228, 234, 236, 238

FISH fluorescence in situ hybridisation. 5, 9, 32, 35, 55, 58

FoSTeS fork stalling and template switching. 12–14

GIAB Genome in a Bottle Consortium. 51, 52

gnomAD The Genome Aggregation Database. 53, 60, 63, 66, 67, 73, 146–148, 151

GWAS genome-wide association study. 3, 25–27, 49, 56, 184, 239, 243

HGSVC The Human Genome Structural Variation Consortium. 50, 51, 54, 67, 68, 152, 153, 208, 210, 216

HMM hidden Markov model. 37

HPC high-performance computing. 210, 216

HPRC The Human Pangenome Reference Consortium. 48, 68, 152

hPSCs human pluripotent stem cells. 59

HR homologous recombination. 12

HWE Hardy–Weinberg equilibrium. ii, iii, xiii, 78, 174–176, 178, 222

InvFEST the INVersion Functional & Evolutionary STudies. 22, 54, 55, 63, 65–67, 91, 144–147, 152, 181, 205, 246

iPCR inverse PCR. 35, 36, 55, 172, 222, 233, 243, 244

IRs inverted repeats. ix, 11, 15, 24, 25, 29–32, 34, 36, 55, 56, 62, 91, 147, 159, 168, 169, 180, 182, 207, 211–213, 219, 220, 224, 227

KIR the killer cell immunoglobulin-like receptor. 48, 61

LCL lymphoblastoid cell lines. 171, 221

LCR low copy repeats. 10

LD linkage disequilibrium. ii, iii, ix, 11, 19, 20, 36, 49, 55, 78, 177–180, 222–226, 230, 231, 233, 236, 239, 240, 244, 245

LINE long interspersed nuclear element. 10

LSFS frequency spectrum of linked sites. 22

LTR long terminal repeat. 10

MAF minor allele frequency. ix, xii, 76, 141, 144, 148, 157, 158, 161–164, 166, 168, 169, 182–184, 226, 227

MEIs mobile element insertions. 59

MHC major histocompatibility complex. 60, 61

MLPA multiplex ligation-dependent probe amplification. 36, 55, 172

MMBIR microhomology-mediated break-induced replication. 12, 13

MMEJ microhomology mediated end-joining. 11, 12

MPS massively parallel sequencing. 6

mSNVs mosaic single nucleotide variants. 57

mSVs mosaic structural variants. 58, 59

NAHR non-allelic homologous recombination. 10, 11, 14, 15, 23–25, 30, 31, 55, 56, 59, 62, 159, 219, 220, 224, 226, 248

NCBI National Center for Biotechnology Information. 53

NGS next generation sequencing. 6, 32–34, 58

NH non homologous. 10, 11, 14, 55, 56, 159, 226

NHEJ non-homologous end-joining. 11, 12

NUMTs nuclear mitochondrial insertions. 50

OLAP online analytical processing. 54

ONT Oxford Nanopore Technologies. ix, x, 6, 40–46, 51, 59, 63, 67, 68, 80, 81, 85, 91, 152, 159, 171, 173, 188, 190, 206, 209–212, 214, 215, 217–219, 222, 224, 227, 232, 233, 243–245, 248

PacBio Pacific Biosciences. 6, 40–46, 50, 51, 59, 66, 67, 206

PBWT positional Burrows–Wheeler Transform. 37

PCR polymerase chain reaction. 5, 34–36, 55, 66, 67, 77, 171, 172, 222, 228

PEM paired-end sequencing and mapping. 33–35, 66

PEV position effect variegation. 28

RFLPs restriction fragment length polymorphisms. 8

SCS Single-cell sequencing. 58

SDs segmental duplications. xiii, 10, 11, 14, 15, 20, 22, 26, 31, 34–37, 40, 46, 47, 51, 59–62, 72, 73, 84, 141, 150, 151, 158, 160, 180, 182, 184–188, 192, 193, 196, 197, 200–203, 213, 219, 229–245

SINE short interspersed nuclear element. 10

SLE systemic lupus erythematosus. 20

SMC simple matching coefficient. 85, 88

SMRT single-molecular real-time. 41

SNPs single nucleotide polymorphisms. ix, 1–3, 5, 8, 11, 21, 25, 32, 36, 37, 47, 49, 50, 52, 53, 55, 56, 58, 66, 69, 76–78, 83, 165, 172, 177, 178, 180, 190, 200, 201, 205, 214, 223, 225–228, 230, 231, 233, 234, 240, 244, 246

SNVs single nucleotide variants. 1, 3, 5, 6

SRS serial replication slippage. 12, 13

Strand-seq Single-cell template strand sequencing. 34, 39, 40, 43, 45, 47, 50, 51, 59, 66, 206

STRs short tandem repeats. 1, 2

SVs structural variants. ix, x, xiii, 1–6, 9, 10, 15, 17, 29, 32–34, 37–40, 45–47, 50–54, 58–60, 62, 63, 65–67, 70, 74, 79, 140, 157, 158, 175, 176, 184–188, 191, 195, 196, 198, 199, 203–206, 208–212, 215, 221, 222, 226, 229, 231, 232, 234–236, 239, 240, 246, 247, 249

T2T telomere-to-telomere. 6, 44, 47, 48, 51, 60, 62, 69, 76, 81, 82, 89, 145, 183, 186, 190, 192, 193, 195, 201, 237, 241, 242

TADs topologically associated domains. 28, 29

WGS whole genome sequencing. 49, 50, 52, 67, 68

1. Introduction

The rediscovery of Mendel’s laws of inheritance in the early twentieth century marked a turning point in the emergence of genetics as a scientific discipline. Since then, genetics has sought to understand how information is transmitted from parents to offspring and how this information shapes the traits we exhibit in the world. Research has spanned from the study of entire chromosomes to the identification of tiny genetic variants, gradually revealing what defines and distinguishes us as individuals. These efforts have unveiled a fraction of the immense diversity encoded in the genome, both at the sequence level and in its structural organisation. However, far from being a complete journey, each discovery uncovers new questions and challenges, highlighting that the exploration of the human genome is a continuously unfolding adventure.

1.1 Genetic variation

Currently, the world is home to a diverse array of human populations. Despite the large phenotypic differences that can be observed between populations or unrelated individuals, we share 99% of the genome. The differences observed between genomes are called genetic variation (Nevo, 1978; Talseth-Palmer & Scott, 2011; Alzu’bi et al., 2019). These genetic variants, which differentiate individuals from each other, come in various shapes and sizes. Some affect only a few base pairs, such as single nucleotide variants (SNVs) in which single nucleotide polymorphisms (SNPs) are included, indels and short tandem repeats (STRs), whereas others, such as structural variants (SVs), can affect millions of bases (Figure 1.1).

Within small variants, SNPs are characterised by the substitution of a single base for another. In contrast, indels involve small deletions or insertions that affect at least 50 bases, altering the amount of sequence. Meanwhile, STRs are 1-6 base sequences that are repeated in different numbers of copies on different chromosomes, becoming one of the most variable genomic elements (Weischenfeldt et al., 2013; Verbiest et al., 2023). On the opposite side, SVs are defined as variants in which a rearrangement affects more than 50 base pairs. SVs can be divided into two main groups; the first includes copy number variants (CNVs), which are variants that modify the number of copies of a specific genomic fragment. Within CNVs, deletions and insertions are included as unbalanced variants, as they alter the sequence dosage. The second group are the balanced variants, in which translocations, both interchromosomal and intrachromosomal, and inversions are included. Sometimes, a locus can present multiple of these variants, generating complex genomic rearrangements (CGRs) (Zhang et al., 2009; Weischenfeldt et al., 2013; Carvalho & Lupski, 2016; Mahmoud et al., 2019; Ho et al., 2020) (Figure 1.1). These complex rearrangements usually involve multiple breakpoints (BPs), which can be defined as novel junctions generated by a structural mutation. At the same time, these rearrangements cannot be explained by a single-end joining or an deoxyribonucleic acid (DNA) exchange event (Zhang et al., 2009; Quinlan & Hall, 2012; Schuy et al., 2022).

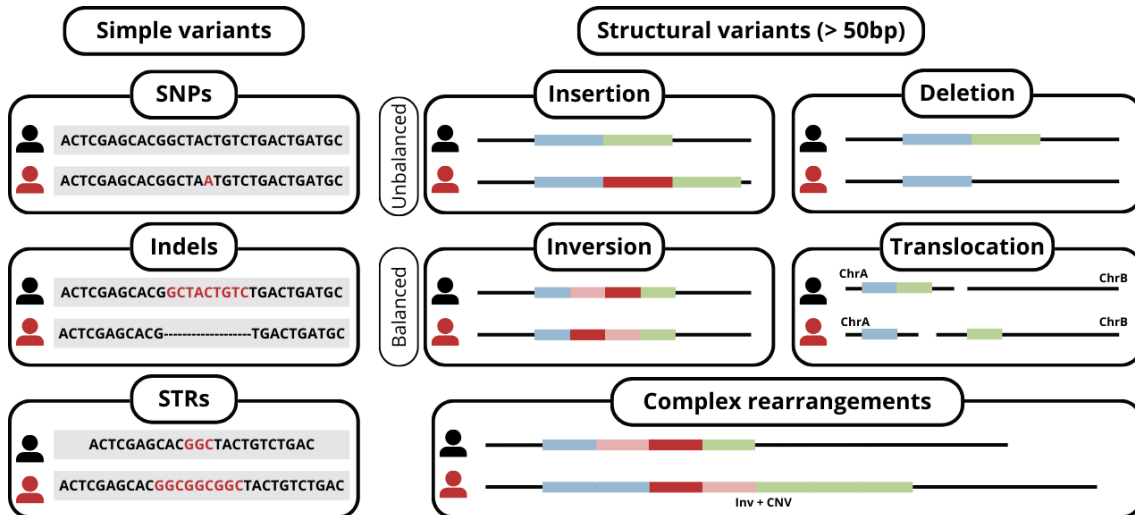


Figure 1.1: **Overview of the different forms of genetic variation.** Small or simple variants are those that affect less than 50 bp. When the affected region has at least 50 bp, the changes are considered SVs.

Among the various forms of genetic variation, SNPs are the most abundant, accounting for roughly 90% of all detected variants (Alzu'bi et al., 2019). When comparing two individuals, SNPs represents between about 0.1% of the genome, corresponding to approximately 4 to 5 million variable bases. However, when structural variants (SVs) are also considered, the total genomic divergence increases to about 0.5–1%, since SVs can span much larger genomic regions, affecting between 20 and 130 million bases, depending on the study (Weischenfeldt et al., 2013; Cook et al., 2020). This variation is notably significant in the telomeric regions due to the abundance of SVs in these genomic areas (Mahmoud et al., 2019). From the perspective of phenotypic effects, SVs have been shown to be approximately three times more likely to be associated with a genome-wide association study (GWAS) signal compared to SNVs (Uffelmann et al., 2021). This is particularly relevant given that GWAS are widely used to identify associations between genotypes and phenotypes by comparing differences in variant allele frequencies between individuals of the same population who exhibit different phenotypes (Uffelmann et al., 2021). Typically, these studies involve comparisons between affected individuals (cases) and healthy individuals (controls). When considering only the largest SVs, those exceeding 20 kilobases, the probability of influencing gene expression increases dramatically, reaching up to 50 times higher than SNVs (Chaisson et al., 2019).

Human diversity was not possible to study until the development and improvement of sequencing technologies, which allowed us to characterise multiple individuals. As more samples were analysed, it became increasingly clear that human diversity was greater than initially expected. However, while these advances had a significant impact on the characterisation of SNPs, SVs lagged behind due to the technical challenges involved in their detection and analysis (Mahmoud et al., 2019; Hollox et al., 2022). The underdetection of SVs and rare variants could partially explain the problem of missing heritability, which can be defined as the gap between the heritability estimated from family studies and that linked to traits and diseases through GWAS (Manolio et al., 2009; De Coster et al., 2021). An example of this could be the fact that SVs play a crucial role in some human pathology, such as autism (Pinto et al., 2010) and schizophrenia (Marshall et al., 2017), which are

related to CNVs. Currently, we have a growing understanding of certain genetic variations and their effects. However, much of this knowledge has been accumulated over many years of research, and we are far from fully characterising SVs and their functional consequences.

1.1.1 Variation discovery

The nineteenth century marked the beginning of genetics with the discovery of Mendel's principles of inheritance. They were the key to understanding that certain traits in organisms were heritable. The understanding of inheritance expanded with the realisation that genes govern specific phenotypes. In the next years, these discoveries were followed by the description of the genome structure, a double-stranded, helical, complementary, and antiparallel model of DNA. The next milestone in our understanding was reached with the discovery of the genetic code, which revealed that a sequence of three adjacent nucleotides, known as a codon, encodes a specific amino acid (Watson & Crick, 1953; Durmaz et al., 2015). This understanding led to the emergence of numerous methods for visualising and analysing chromosome structure and function.

The first genetic analyses were performed in the middle of the twentieth century, when cytogenetics allowed the counting of chromosomes and the identification of significant chromosomal anomalies, such as the trisomy of chromosome 21, the cause of Down syndrome (Lejeune et al., 1959; Durmaz et al., 2015). Despite the success of classical cytogenetic techniques in detecting large-scale chromosomal alterations, new methods were later developed to identify additional types of genomic variants. Among these classical approaches, G-banding remains a fundamental technique, which through specific chemical treatments and staining procedures reveals characteristic light and dark banding patterns along chromosomes (Yunis, 1976). The analysis of these bands allows for the detection of multiple types of SVs, such as deletions, translocations, or inversions. Although numerous large-scale SVs (> 3 Mb) have been identified, these generally affect only a small number of individuals and are commonly associated with diseases, which does not reflect typical human

variability (Durmaz et al., 2015; Balajee & Hande, 2018). The next step in variant discovery came with the introduction of fluorescence in situ hybridisation (FISH), which was initially based on radioactive labelling and was later replaced by fluorescent probes to detect chromosomal rearrangements (Pardue & Gall, 1969; John et al., 1969; Shakoori, 2017). Subsequently, the development of multiple probes and their combinations greatly increased the versatility and resolution of the technique. As a result of FISH, submicroscopic SVs were detected for the first time, allowing the discovery of multiple congenital anomalies. The evaluation of chromosomal rearrangements in the entire genome using FISH was extremely costly, leading to the development of comparative genomic hybridisation (CGH) based on arrays, which allows the detection of CNVs at high resolution (Kallioniemi et al., 1992; Durmaz et al., 2015).

At that time, the idea of studying single-base variations was not feasible. To achieve this goal, polymerase chain reaction (PCR), which allowed the generation of thousands of copies of a DNA sequence, was necessary (Saiki et al., 1985). PCR was a great success, especially when first-generation sequencing methods were applied to determine the specific nucleotide composition. These sequencing technologies, first using Maxam-Gilbert chemical sequencing and then Sanger sequencing, allowed the analysis of SNVs and were crucial in advancing our sequencing capabilities. These efforts were materialised at the beginning of the 2000s with the achievement of the first near-complete human genome sequence by the Human Genome Project in parallel with Celera Genomics (Lander et al., 2001; Venter et al., 2001; International Human Genome Sequencing Consortium, 2004). The completion of these large-scale projects not only provided a first picture of the genes on the genome but also led to improvements in sequencing technology. The first genome was only the starting point; later, multiple international projects were developed to study in detail the variation responsible for our phenotypic traits. The initial HapMap Project (Gibbs et al., 2003; The International HapMap Consortium, 2005; Altshuler et al., 2010) provided one of the first large-scale resources that describes patterns of DNA sequence variation between human populations, with a particular emphasis on SNPs. However, these early efforts relied on technologies that were costly and limited in

their ability to capture the full spectrum of genetic variation.

The advent of massively parallel sequencing (MPS) technologies, later known as next generation sequencing (NGS), marked a major turning point (Margulies et al., 2005; Heather & Chain, 2016). Using these advances, 1000 Genomes Project (1000GP) (McVean et al., 2012; Auton et al., 2015; Byrska-Bishop et al., 2022) generated a comprehensive catalogue of human sequence variation between diverse populations and a wide range of allele frequencies. Subsequently, numerous NGS-based sequencing projects were launched to further characterise human genetic diversity, focussing either on specific populations (Wong et al., 2013; Genome of the Netherlands Consortium, 2014; UK10K Consortium et al., 2015; Besenbacher et al., 2015; Nagasaki et al., 2015; Ameer et al., 2017; Choudhury et al., 2017; Chheda et al., 2017; Kim et al., 2018; 100,000 Genomes Project Pilot Investigators et al., 2021; Kurki et al., 2023) or including individuals from multiple global populations (Gurdasani et al., 2015; Mallick et al., 2016).

Such research efforts have been highly successful in identifying and describing genetic variation, becoming the basis in which to search for associations between genetic variation and phenotypic effects. However, the main focus has been largely on SNVs, while SVs have remained in the background. This imbalance has been reduced with the development of new sequencing techniques that allow for a more accurate analysis of SVs. These advances are currently being made, particularly through the implementation of these third-generation sequencing technologies, such as Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) (Dijk et al., 2018; Amarasinghe et al., 2020). At the same time, larger-scale population studies are being carried out using not only NGS but also this third-generation sequencing technologies (Beyter et al., 2021; Ebert et al., 2021; Quan et al., 2021; Gustafson et al., 2024; Schloissnig et al., 2025). Part of these efforts and uses of third-generation sequencing technologies have been reflected in the generation of the first telomere-to-telomere (T2T) genome in 2022 (Nurk et al., 2022). Recently, the first version of the human pangenome has been created to capture diversity within large sample collections using graph models. The samples used to develop this

pangenome were sequenced using various methodologies, including third-generation sequencing technologies (Liao et al., 2023). However, to accurately represent diversity, the improvement of the detection and characterisation of genetic variants is essential, especially those historically overlooked, such as inversions, which still constitute a challenge to study.

1.2 Inversions

Inversions are a common class of structural variants characterised by a change in orientation of a segment of DNA. Inversions occur when two breaks on the same chromosome are generated and the internal sequence rotates 180 degrees and is reinserted at the same location, but in reverse orientation (Kirkpatrick, 2010). Those remain as one of the least studied variants due to their balanced nature and the frequent presence of repetitive regions becoming the BPs (Puig, Casillas, et al., 2015). Although inversions were identified early, only a few have been thoroughly researched in humans and their clinical relevance is still largely unknown. However, inversions have been linked to various phenotypic traits and adaptations in multiple species (Puig, Casillas, et al., 2015; Vicente-Salvador et al., 2017; Wellenreuther & Bernatchez, 2018; Puig et al., 2020). Before analysing the various impacts and outcomes of inversions, it is essential to consider the history of this structural variant.

1.2.1 An historical perspective

The history began in 1913, when the first genetic map of *Drosophila melanogaster* was generated based on the frequency of cross-over events between normal and mutant alleles of six genes located on the X chromosome (Sturtevant, 1913). These crossovers were instrumental in the establishment of the linear arrangement of genes, allowing Alfred H. Sturtevant to realise that certain chromosomes in *Drosophila* carried “crossover modifiers”, factors that reduced the frequency of recombination at specific loci in some individuals (Sturtevant, 1917; Fuller et al., 2019). Later, when Sturtevant compared chromosome X of *Drosophila melanogaster* and *Drosophila simulans*, he discovered that five genes regulating similar traits were arranged in

the same order, indicating conservation among related species (Sturtevant, 1921b). Despite the general similarity, Sturtevant identified three identical loci on the third chromosome that were arranged in reverse order, constituting the first published evidence on an inversion (Sturtevant, 1921a; Wellenreuther & Bernatchez, 2018; Huang & Rieseberg, 2020). In the following years, Sturtevant obtained enough mutants with chromosomes containing the inversion in *Drosophila melanogaster* to establish that the inversions were indeed dominant crossover suppressors (Sturtevant & Mather, 1938; Wellenreuther et al., 2019). This statement was later validated by the observation of giant polytene chromosomes in the salivary glands of *Drosophila* larvae. These chromosomes allowed the direct visualisation of inversion loops in heterozygotes, formed to maintain homologous pairing during somatic chromosome alignment (Kapun & Flatt, 2019). Although cytogenetic methods were initially limited, the clear banding patterns of polytene chromosomes provided a powerful tool to detect chromosomal rearrangements (Dobzhansky & Sturtevant, 1938; Brehm & Krimbas, 1991; Kirkpatrick, 2010; Huang & Rieseberg, 2020).

Inversions became one of the first variants studied (Kapun & Flatt, 2019) and were widely used as genetic markers to infer phylogenetic relationships based on fixed inversions among closely related species. In addition, differences in chromosomal banding map patterns were observed between individuals of the same species, finding thousands of polymorphic inversions (Dobzhansky & Sturtevant, 1938; Wellenreuther & Bernatchez, 2018; Fuller et al., 2019).

The 1970s decade and the improvement of molecular biology techniques caused a progressive reduction in the popularity of inversion research (Kirkpatrick, 2010; Wellenreuther & Bernatchez, 2018). These new techniques included the establishment of different genetic markers such as microsatellites, restriction fragment length polymorphisms (RFLPs) (Botstein et al., 1980), amplified fragment length polymorphisms (AFLPs) (Vos et al., 1995), and later SNPs (Wellenreuther et al., 2019). With the low-cost and high-throughput methods for genotyping numerous SNPs with array-based techniques, SNPs became the most studied type of variants. In addition, these techniques allowed the identification of large unbalanced variants,

CNVs, including deletions and insertions. However, the study of inversions complicated by their balanced nature with no gain or loss of DNA lag behind, compared to the other variants that were easier to detect SVs (Alkan et al., 2011).

Despite the use of sequencing technologies, the study of inversions in humans remained limited to basic cytogenetic techniques, such as G-band karyotyping, which cannot detect submicroscopic variations. Here, several pericentric inversions, those that encompass the centromere, affecting chromosomes 1, 2, 9, 10, 12, 16, X, and Y were identified (Hsu et al., 1987; Thomas et al., 2008; Feuk, 2010). Among these, the inversion of the centromere of chromosome 9 is the most relevant, showing a frequency ranging from 0 to 4% in human populations. However, these large inversions have not been associated with negative phenotypic effects or fertility problems, and in many cases they simply appear to be consequences of the repetitive architecture of pericentromeric regions (Yamada, 1992; Mozdarani et al., 2007; Dana & Stoian, 2012; Kosyakova et al., 2013; Nonaka et al., 2019). Furthermore, cytogenetic analysis has been implemented as routine in hospitals, where it is used to find *de novo* rearrangements that appear in patients or families and are often associated with clinical symptoms (Vermeulen et al., 2005; Partida-Pérez et al., 2012; Oliveira et al., 2013; Sotoudeh et al., 2017) or cancer (Beau et al., 1983; Gorello et al., 2013; Kaneko et al., 2014; Lv et al., 2020). Other techniques were used to try to characterise inversions, usually focussing on target-based approaches such as FISH, however, they allowed only for the identification of a limited number of inversions (Antonacci et al., 2009). The lack of high-throughput discovery and genotyping methods for inversions was a limitation in obtaining a representative catalogue of inversions and defining their functional impact (Feuk, 2010). New computational genomics in conjunction with improved sequencing techniques, such as long-read sequencing, is leading to the rediscovery of inversions, improving their detection, and their involvement in evolutionary and functional processes (Wellenreuther & Bernatchez, 2018; Fuller et al., 2019; Chaisson et al., 2019; Ebert et al., 2021; Porubsky et al., 2022). With all this in mind, the following sections aim to explore in greater depth the intricacies of inversions, from their mechanistic origins to the phenotypic associations that have been identified to date.

1.2.2 A mechanistic view

Traditionally, inversions were classified by examining whether they involve the centromere (pericentric) or not (paracentric) (Figure 1.2 - A). Paracentric inversions are more common and some estimates suggest that each person could have approximately 100 of those inversions in their genome (Chaisson et al., 2019; Ebert et al., 2021; Porubsky et al., 2022). In addition, inversions can be classified according to their mechanism of formation. Inversions can occur during various cellular processes: DNA recombination, repair, and replication, and can be generated by different mechanisms that can involve homologous or repetitive sequences or not: non-allelic homologous recombination (NAHR) and non homologous (NH) mechanisms (Escaramís et al., 2015).

NAHR inversions

NAHR inversions occur between a pair of repeats in opposite relative orientation. These repeats can be transposable elements (TE) or low copy repeats (LCR), which can also be called segmental duplications (SDs). SDs consist of DNA blocks with more than 90% identity between both copies and more than 1 kilobase in size. They can contain genes, pseudogenes, gene clusters, or other paralogous sequences (Eichler, 2001; Burssed et al., 2022). Repetitive sequences play a key role in the promotion of NAHR mechanisms. Although SDs are well-known contributors, mobile elements such as long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs), and long terminal repeats (LTRs) can also drive this process by providing homologous sequences that facilitate recombination (Startek et al., 2015). Their similarity, along with factors such as size and orientation, influences the architecture of the genome and creates regions susceptible to structural rearrangements. NAHR variations are generated by the misalignment of these repetitive elements after double-strand break (DSB) during mitosis or meiosis, causing an unequal crossing event that generates a rearrangement in the daughter cells. Depending on the relative orientation and location of the repetitive sequences, different SVs can emerge from the process. When the repeats are in inverted orientation and

in the same chromosome, it generates an inversion (Figure 1.2 - C), whereas when the repeats are in the same orientation, a deletion arises (Figure 1.2 - B). When repeats are located on different chromosomes, they lead to a translocation (Burssed et al., 2022). Generally, the genomic elements that generate NAHR inversions can be called inverted repeats (IRs), which encompass SDs and other repetitive elements in inversion BPs.

The presence of IRs constitutes one of the main problems in inversion detection. IRs act as obstacles to achieving the inverted region, because they are usually longer than the reads or library inserts (Puig, Casillas, et al., 2015; Porubsky et al., 2022). However, the development of third-generation sequencing methods is helping to overcome this limitation, due to the obtention of larger reads. The other main feature of this type of inversion is recurrence, which occurs when an inversion arises repeatedly in independent events, both within and between species. This phenomenon can also be called "inversion toggling" (Fickelscher et al., 2007; Zody et al., 2008; Antonacci et al., 2009; Salm et al., 2012; Giner-Delgado et al., 2019; Porubsky, Sanders, Höps, et al., 2020; Puig et al., 2020; Porubsky et al., 2022). Recurrence can occur as long as repetitions are maintained (Cáceres et al., 2007). In addition, the length and distance between these IRs have been shown to influence the occurrence of inversions. The proximity and large size of these IRs increase the likelihood of inversion formation and subsequent recurrence events (Giner-Delgado et al., 2019; Puig et al., 2020). One consequence of recurrence is the reduction in linkage disequilibrium (LD) inversions with other variants in the region, as well as the presence of shared SNPs between the two orientations, which complicates genotyping by association with other variants easier-to-genotype and reduces the precision of prediction methods based on imputation. As a direct consequence of these features, the NAHR inversions have been understudied and their functional impact is in most cases unknown (Aguado et al., 2014; Giner-Delgado et al., 2019; Porubsky et al., 2022).

NH inversions

The other main pathway of inversion formation involves the repair of DNA breaks through mechanisms such as non-homologous end-joining (NHEJ) and microhomol-

ogy mediated end-joining (MMEJ), as well as replication-related errors, which include processes such as fork stalling and template switching (FoSTeS), serial replication slippage (SRS), microhomology-mediated break-induced replication (MM-BIR) and break-induced serial replication slippage (BISRS) (Escaramís et al., 2015). These mechanisms are collectively known as non-homologous pathways and can produce inversions as a result of unique events (Lee et al., 2007).

Repair-based inversions

DSB are essential for processes such as meiotic recombination and immune system development, but can also be a negative consequence resulting from endogenous (e.g. reactive oxygen species) or exogenous (e.g. radiation) factors. If these breaks are not correctly repaired, they can threaten genome stability and produce mutations, and for that reason multiple repair mechanisms are needed (McVey & Lee, 2008; Jiang, 2022). Between all these repair-based and replication-based mechanisms, homologous recombination (HR) and NHEJ are the main pathways for repair DSB in mammals. In the case of NHEJ, when both DNA strands are broken, the ends are directly ligated and rejoined with little or no sequence homology (<4 bp). Although NHEJ can generate gain or loss of nucleotides, or even translocations at BPs, it generally operates with high fidelity and is not considered a major source of genomic instability (Lieber et al., 2010; Escaramís et al., 2015; Burssed et al., 2022).

On the other hand, MMEJ is responsible for DSB repair in few cases and has traditionally been considered a backup mechanism when NHEJ and HR are not functional (Escaramís et al., 2015; Seol et al., 2018). However, MMEJ has been shown to be a prominent repair pathway in multiple models, even when the main pathways are active. The main feature of MMEJ is that it uses sequence microhomologies (2-20 bp) during the rejoining of broken ends, resulting in the production of deletions at the BPs junctions (McVey & Lee, 2008; Carrara et al., 2025). In MMEJ, the repair process involves resection of the 5' end at BPs, generating 3' single-stranded overhangs. These overhangs can align via the mentioned microhomologies, leading to deletion of the intervening sequence (Burssed et al., 2022). Due to this process, MMEJ is

considered more prone to error and produces a higher genome instability. In general, inversions arising from DSB repair are distributed throughout the genome without a clear pattern and are accompanied by short insertions and deletions at the BPs.

Replication-based inversions

Replication-based processes generate complex rearrangements that can involve multiple BPs and variants such as insertions, deletions, or inversions (Lee et al., 2007). During replication, genomic rearrangements can arise when the replication fork stalls, leading to a mechanism known as FoSTeS, in which DNA polymerase switches templates and the DNA synthesis is re-started (Zhang et al., 2009; Burssed et al., 2022). This template switching is driven by microhomologies in nearby single-stranded DNA, which can occur repeatedly and produce rearrangements that range from kilobases to megabases (Escaramís et al., 2015; Carvalho & Lupski, 2016). If switching involves sequences in opposite orientation, inversions can be generated (Burssed et al., 2022), often with flanking deletions (Vicente-Salvador et al., 2017) (Figure 1.2–D).

SRS can produce small complex rearrangements as a consequence of multiple rounds of forward and backward replication slippage, producing a provisional dissociation of the primer and template strands followed by a misaligned reassociation that leads to loss or duplication of the region in the presence of short direct repeats (Chen et al., 2005; Escaramís et al., 2015). Other replication-based mechanisms include BISRS and MMBIR. Unlike FoSTeS, these mechanisms are initiated by a replication fork collapse, resulting in a single-ended DSB. In this context, the break arises from the failure of the replication machinery, leaving one broken DNA end rather than a canonical two-ended DSB. The exposed 3' single-stranded overhang can then invade a different genomic region to reinitiate DNA synthesis (Hastings et al., 2009).

In BISRS, this invasion is mediated by long stretches of homology, whereas in MMBIR it is based on short microhomologies. In contrast, FoSTeS does not require the formation of a DSB and instead involves template switching by a stalled but unbroken replication fork.

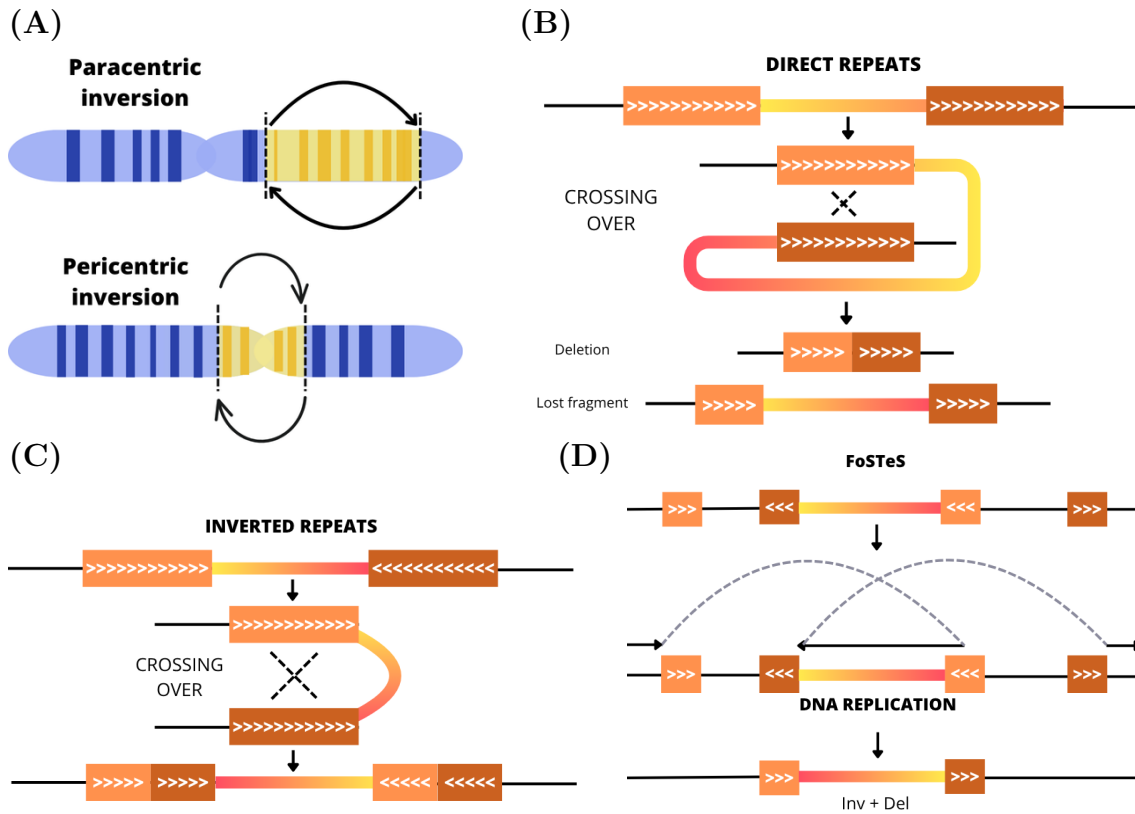


Figure 1.2: **Classification of inversions by location (A) or by mechanism of generation (B, C and D).** (A) The involvement of the centromere in an inversion determines its classification: pericentric when the centromere is affected and paracentric when it is not. (B and C) NAHR events occur between direct SDs (B), generating deletions, or between inverted copies (C), generating inversions. (D) FoSTeS can produce inversions when the polymerase stalls and switches templates in microhomology regions. This mechanism often leads to the formation of additional structural variants surrounding the inversion BPs. Arrows indicate the path and direction of the DNA synthesis and dashed lines show the template switching events between DNA strands.

The underlying characteristics of an inversion can provide information on their mechanism of origin. NAHR inversions are typically characterised by homologous sequences at their BPs, tend to occur recurrently, and are less likely to be associated with other structural variants. In contrast, NH inversions generally lack breakpoint homology, arise from unique events, and are more frequently accompanied by additional structural variants at BPs (Vicente-Salvador et al., 2017; Giner-Delgado et al., 2019). Due to their molecular characteristics, NH inversions have historically been easier to detect, particularly with short-read sequencing technologies, which are

often blind to inversions mediated by IRs (Lledó & Cáceres, 2013). Over the years, numerous studies have been aimed at unravelling the mechanisms underlying the formation of SVs, showing that NAHR inversions tend to appear more frequently (Kidd et al., 2010; Pang et al., 2013; Chaisson et al., 2015). However, the true number of NAHR variants is likely much higher, given the intrinsic properties of these inversions, which make them difficult to detect using traditional sequencing approaches with short reads that cannot span large repeats at inversion BPs. An indication supporting this statement is the high abundance of SDs in the human genome ranging from 5% (Bailey et al., 2002) to 6.6% (Nurk et al., 2022), and the widespread presence of repetitive elements, which comprise between half and two thirds of the genome (Lander et al., 2001; Koning et al., 2011; Nurk et al., 2022), which are substrates for NAHR events.

1.3 Functional consequences

Having reviewed the history of inversion discovery and classified inversions into groups based on their characteristics, the next step is to address their potential functional effects. An important consideration is that, in some cases, the genetic content between the two orientations remains identical, raising the question of the impact that a simple change in linear DNA order can have.

Large inversions were discovered more than a century ago and some have been systematically detected using cytogenetic methods such as karyotyping. Although many have not been clearly linked to functional effects in most cases, mainly due to the technical difficulty of accurately genotyping them. However, several *de novo* inversions, as well as inversions segregating within families, have been demonstrated to cause specific diseases (Lakich et al., 1993; Bondeson et al., 1995; Osborne et al., 2001; Pagnamenta et al., 2023, 2024; Bozkurt-Yozgatli et al., 2025). Over the years, chromosomal inversions have also been found to exert a variety of direct and indirect effects at the molecular level, with functional and evolutionary implications. In the following section, we discuss these effects in more detail and provide illustrative examples from the literature.

1.3.1 Indirect effects

Inhibition of recombination

As was mentioned before, since the beginning of inversion discovery as "crossing-over modifiers", it was clear that inversions were able to reduce the recombination between opposite orientations, or which is the same, in heterozygote individuals. This feature is key to understanding the impact of inversions in an evolutionary and functional perspective. In diploid organisms, homologous chromosomes pair for recombination during the first meiotic division. When there is a heterozygous inversion, disrupted homology within the inverted region leads to the formation of a chromosomal loop, allowing homologous chromosomes to align in opposite orientations along this segment (Hoffmann & Rieseberg, 2008; Kirkpatrick, 2010; Wellenreuther & Bernatchez, 2018). As mentioned above, this phenomenon was observed for the first time in *Drosophila* polytene chromosomes (Dobzhansky & Sturtevant, 1938).

Depending on the location of the inversion (whether it involves a centromere or not), the resulting gametes differ. If a crossover occurs within the chromosomal loop, the resulting unbalanced products will vary accordingly. In the case of paracentric inversions, four products will be generated, a dicentric bridge, an acentric fragment, and two chromosomes with all loci (one in standard order and the other in reverse order)(Figure 1.3). In the next steps of meiosis, the acentric chromosome will be lost due to the inability to be drawn to the cell pole. Meanwhile, the dicentric bridge will break in a random position, leaving behind two products with deletions as a result of the bridges being drawn to opposite poles. If such recombinant chromatids are transmitted to the offspring, the result will be a segmental aneuploidy in the zygote (Figure 1.3). Instead, if the crossover is produced within a pericentric inversion, four products are also generated: two chromosomes with all loci (one in standard order and the other in reverse order), and two duplication and/or deletion products, in which some loci can be duplicated or deleted, depending where the crossover occurs (Wellenreuther & Bernatchez, 2018) (Figure 1.3). Therefore, inversions are likely to affect fertility due to the production of unbalanced chromosomes, which

cannot provide viable progeny. To avoid this, recombination within inversions must be suppressed. This can occur either through a true inhibition of crossover formation within the inverted region or through selection against recombinant gametes that are rendered unbalanced by crossing over (Puig, Casillas, et al., 2015; Yahaya et al., 2021). It is important to note that, unlike other SVs, inversions have an inherent disadvantage due to their negative effects on fertility. Therefore, the fact that some inversions are maintained—and even increase in frequency—within populations likely reflects the presence of selective advantages that counterbalance this reduction in fertility (Giner-Delgado et al., 2019). This negative effect on fertility has an exception, *Drosophila* species and other Diptera, in which males do not recombine during meiosis and in females these aneuploid products are pushed to the polar bodies rather than the egg nucleus (Lande & Wilkinson, 1999; John et al., 2016).

The ultimate consequence of recombination suppression by an inversion is the accumulation of different mutations within each rearranged segment, which can no longer undergo recombination. This leads to a separate evolution and divergence between the chromosomes that present each orientation, creating new allelic combinations. These two chromosomal orientations can sometimes maintain together advantageous combinations of alleles and promote the accumulation of adaptive mutations within the inverted region, which can increase the frequency of the inverted allele in the population (Puig, Casillas, et al., 2015). The presence of large inversions, which inhibit recombination in large blocks, can lead to "supergenes", which can promote complex phenotypes, environmental adaptation, and speciation (Schwander et al., 2014). An example of this is the evolution of mammalian sex chromosomes. In this case, a series of inversions occurred to prevent recombination between the proto-sex chromosomes, allowing the emergence of distinct sex-determining alleles (Bachtrog, 2013; Schwander et al., 2014). Now, as a direct consequence of the suppression of recombination, we will talk in more detail about "supergenes" and the role of inversions in complex phenotypes and local adaptation.

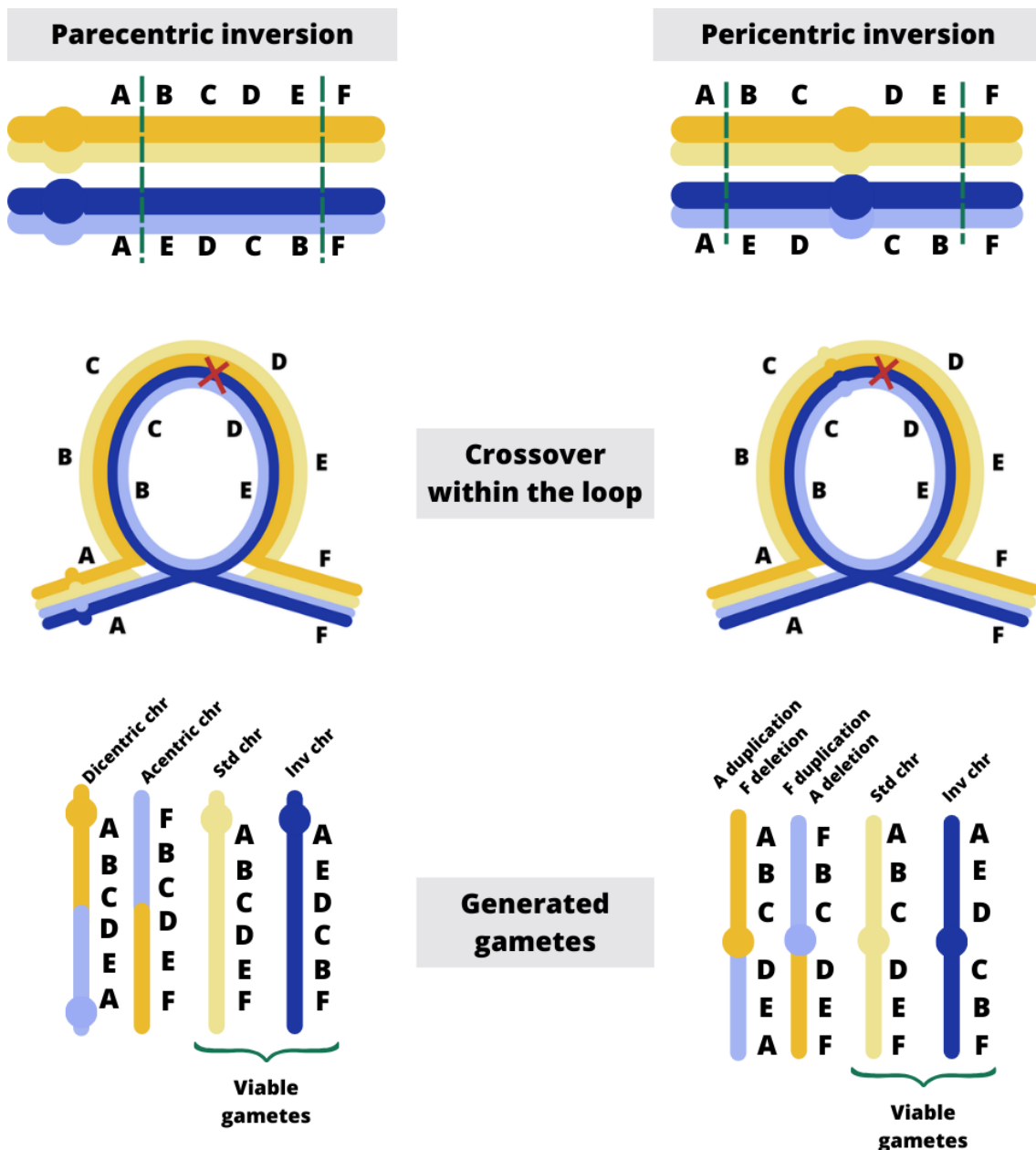


Figure 1.3: **Paracentric (left) and pericentric (right) inversions and their effect on recombination.** In individuals heterozygote for an inversion the chromosomes form an inversion loop to allow proper pairing between homologous regions with opposite orientations during recombination. Crossovers within this loop (red cross) can generate multiple types of gametes, depending on the position of the inversion relative to the centromere and the location of the crossover event. In the figure, the two chromosome orientations are shown in yellow and blue. Letters represent the sequence order, and dots indicate the centromeres of the corresponding chromosomes.

Complex phenotypes and local adaptation

Inversions in some species can span several megabases of length affecting multiple genes. By limiting recombination, inversions can promote the emergence of

"supergenes", which are defined as a group of two or more tight clusters of loci under balanced selection, that control complex phenotypes, such as developmental or behavioural characters, and are inherited as a unit due to their location in a non-recombining region (Schwander et al., 2014; Thompson & Jiggins, 2014). The presence of these supergenes has significant evidence in multiple species, attributed to mating systems, social behaviour, and climate adaptation (Wellenreuther & Bernatchez, 2018). Over evolutionary timescales, supergenes can contribute to reproductive isolation and even speciation. The ability of inversions to preserve large haplotype blocks underlies their role in supergene formation and, in some cases, in the creation of supergene complexes (Hoffmann & Rieseberg, 2008; Wellenreuther & Bernatchez, 2018). The concept of supergenes was first proposed in 1930 to explain Batesian mimicry in butterflies, where gradual adaptation through multiple small changes, rather than a single mutation of large effect, was required. It was suggested that several genes, with LD within a non-recombining region, jointly controlled this trait. However, it was not until 1949 that Darlington and Mather formally introduced the term supergene (Thompson & Jiggins, 2014; Gutiérrez-Valencia et al., 2021).

Among species, there are multiple examples of supergenes, starting from Batesian mimicry in butterflies already mentioned. Batesian mimicry is a protection mechanism that consists in the exhibition of specific traits such as wing patterns, body colour, or hindwing tails, to mimic toxic species. In the case of males of *Papilio polytes*, males never present this mimicry while females can exhibit the mimetic or non-mimetic pattern (Schwander et al., 2014; Gutiérrez-Valencia et al., 2021). In recent studies, two differentiated haplotypes, called H and h, were associated with the mimetic and non-mimetic forms. The haplotypes contain a region of 130 kb located on chromosome 25 that includes the double sex gene (*dsx*), a key gene in insect sexual dimorphism. These haplotypes are distinguished by the presence of an inversion. Another example of supergenes induced by inversions includes the region of locus P in *Heliconius numata*, which also has effects on Batesian mimicry (Gutiérrez-Valencia et al., 2021; Komata et al., 2022). Inversion-derived supergenes are not restricted to insects. In birds, for example, the white-throated sparrow (*Zonotrichia*

albicollis) exhibits two plumage morphs associated with distinct reproductive behaviours. The morphs, known as white-striped and tan-striped, differ in mating strategies: white males tend to be more promiscuous and invest in securing multiple mates, whereas tan males are typically monogamous and provide greater paternal care (Tuttle et al., 2016). These behavioural and morphological differences are controlled by a supergene located on chromosome 2, which presents two chromosomal configurations (ZAL2m and ZAL2), which differ by the presence of two large inversions. Tan-striped individuals are homozygous (ZAL2/ZAL2), while white-striped ones are predominantly heterozygous (ZAL2/ZAL2m). These two configurations are maintained polymorphic due to a near-obligate mating between tan and white morphs (disassortative mating) (Tuttle et al., 2016; Maney et al., 2020).

In the case of humans, there are two main candidates to act as supergenes. The first is the 8p23.1 region that has a size of 4.2 Mb and is associated with an inversion mediated by two large blocks of SDs called REPP (proximal) and REPD (distal). The region contains multiple genes that vary in copy number, such as defensins or FAM90A (Bosch et al., 2009; Campoy et al., 2022). The region has shown low LD with nearby variants in most populations, excluding Europeans, suggesting the presence of double crossovers in the central regions of the inversion. The 8p23.1 region has been linked to several phenotypes and differences in gene expression. Reference orientation (O1) has been associated with an increase in the risk of systemic lupus erythematosus (SLE). Inverted orientation (O2) has been associated with increased risky behaviour, a reduction in hypertension and obesity, protection of Barrett’s oesophagus inflammation, an increased risk of autoimmune thyroid disease and lower levels of liver enzymes (González et al., 2020; Campoy et al., 2022) (Figure 1.4-A). The second case is the 17q21.31 region, which spans 900 kb and is also flanked by complex blocks of SDs. Unlike the 8p23.1 region, this region presents a high LD, originating two clearly divergent haplotypes, the reference orientation (H1) and the inverted one (H2) which do not present recombination between them (Stefansson et al., 2005; Campoy et al., 2022). Furthermore, in addition to the small multiple variants found between H1 and H2, the inversion BPs of the 17q21.31 region present three different CNVs that affect genes. Each of the orientations of this re-

gion also shows association with multiple phenotypic traits: the H2 haplotype shows protective effects against certain neurodegenerative disorders, changes in cell blood composition, reduction of some parameters of lung function, increased risk of cancer and autoimmune diseases, and affects numerous physiological and behavioural traits. In addition, it has been associated with increased fertility in women, showing a positive selection in the Icelandic population, which is also reflected by the increased frequency of this inversion in Europeans (Stefansson et al., 2005). In the study of this region, multiple SNPs affect genes such as *CRHR1*, *KANSL1*, *MAPT* and *SPPL2C*, which could explain some of the observed phenotypic traits. The 17q21.31 region is one of the clearest examples of a human supergene (Campoy et al., 2022) (Figure 1.4-B).

A well-studied example is the 3RP inversion in *Drosophila melanogaster*, which exhibits consistent latitudinal clines across continents that parallel climatic gradients. This inversion has been associated with variation in body size, as well as traits related to thermal resistance (Anderson et al., 2005). Inversion clines can also be found in other species such as *Anopheles*, seaweed fly (*Coelopa frigida*) and maize (*Zea mays*) (Fang et al., 2012; Ayala et al., 2017; Wellenreuther et al., 2017). Sometimes inversions can also act as an adaptive response to environmental changes, such as the case of inversion frequency changes caused by increased temperature due to climate change, reported in *Drosophila subobscura* (Balanyá et al., 2006; Rodríguez-Trelles & Tarrío, 2024). These adaptation mechanisms can, over time, lead to reproductive barriers that can ultimately result in speciation. Some examples of this are the speciation of *Noccaea caerulea* with the union of pre-existing metal homeostasis genes (Mandáková et al., 2015; Huang & Rieseberg, 2020), the flowering differences generated in *Boechera stricta* (Lee et al., 2017) and differences in flowering time and growth-related traits in yellow monkeyflower (*Mimulus guttatus*) (Lowry & Willis, 2010).

In humans, some inversions also show signs of selection, including the 17q21.31 mentioned above. Sometimes, this selection could result in significant AF differences between populations based on Fixation index (F_{ST}) statistics, which are observed in

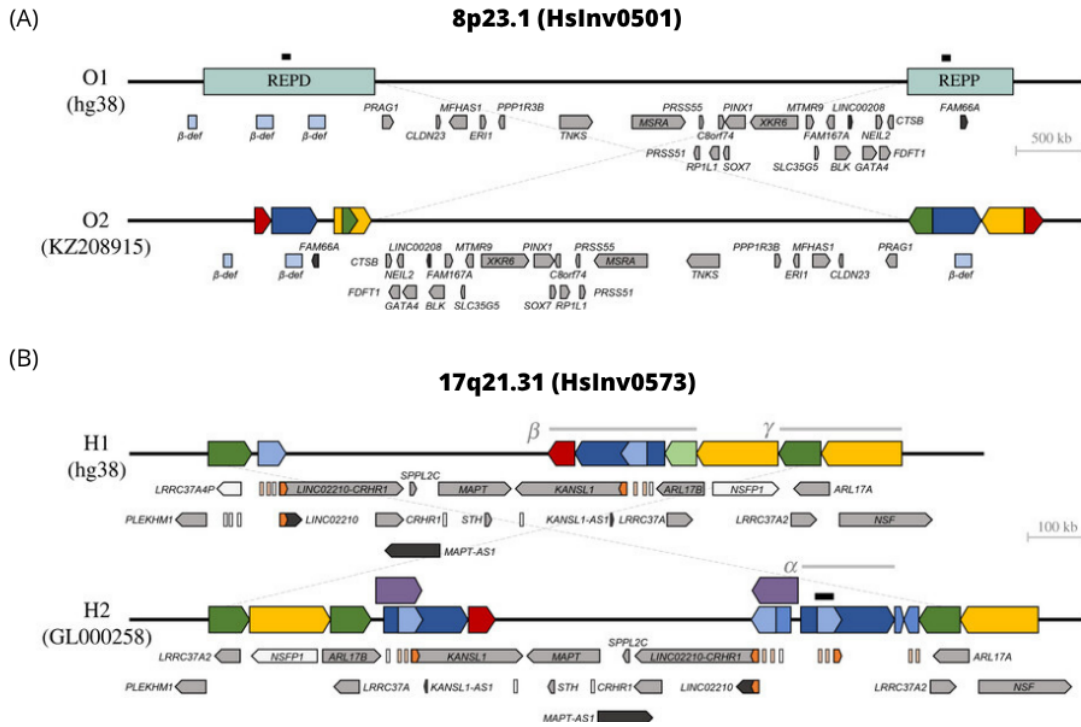


Figure 1.4: **Genomic structure of two human inversion supergenes.** (A) The 8p23.1 inversion region. The proximal (REPP) and distal (REPD) repeat blocks in orientation O1 are represented as light green rectangles. Protein-coding genes are shown as grey arrows, non-coding genes as black arrows, and β -defensin clusters as light blue boxes. (B) The 17q21.31 region containing the inversion that gives rise to the H1 and H2 haplotypes. At the BPs there are variable, complex segmental duplications (SDs) blocks of 200–800 kb indicated in different colours. Partial blocks are shown in lighter shades. The α , β and γ blocks correspond to copy number variants. Protein-coding genes (grey), non-coding genes (black), and pseudogenes (white) are represented as arrows below each orientation. The duplication of the first *KANSL1* exons and two pseudogenes is shown as small orange boxes. Figure adapted from (Campoy et al., 2022)

HsInv0040, HsInv0389 and HsInv0059 (the notation "HsInv" used throughout this thesis comes from the INVersion Functional & Evolutionary STudies (InvFEST) (<https://invfestdb.researchmar.net>), in European, African and East Asian populations, respectively (Vicente-Salvador et al., 2017; Giner-Delgado et al., 2019). This population difference can also be observed in the 16p11.2 inversion, which is more frequent in Northern Europe (González et al., 2014a; Kosuthova & Solc, 2023). In addition, the use of frequency spectrum of linked sites (LSFS) allowed the identification of positive selection signs in HsInv0006 and HsInv0031, while also showing balancing selection in HsInv0201 (Giner-Delgado et al., 2019). All of these examples

give us some clues about the importance of inversions in evolution.

1.3.2 Direct effects

Inversions can not only cause phenotypic effects by the accumulation of changes within the non-recombining inverted sequence, but can also have multiple direct effects, especially when one or both of their BPs disrupt coding or regulatory sequences, contributing to gene dysfunction, or modify gene expression by changing the position of genes or regulatory elements. For that reason, accurately identifying inversions and their genomic context is essential to infer their potential functional and phenotypic effects. Inversions can also cause disease directly, including *de novo* events in individual patients, as well as recurrent rearrangements observed in unrelated cases that can affect gene structure or regulation. Furthermore, inversions can facilitate the formation of novel rearrangements such as copy number changes and translocations (Sharp et al., 2006). Although there are some large inversions in humans, many lack associations with phenotypic effects (Puig, Casillas, et al., 2015). In the next paragraphs, some of the direct effects of inversions will be commented on in more detail, and some examples will be provided.

Inversions causing diseases

One of the most known cases of disease caused by inversion is haemophilia A, which is a coagulopathy in which 45% of severe cases are a consequence of genetic deficiency in clotting factor VIII (encoded by the *F8* gene) located on chromosome X. This mutation is caused by an inversion generated by NAHR mechanisms. In this case, three repeat sequences of 9.5 kb (called int22h) located close to or within the factor VIII gene are responsible for the crossing-over (Antonarakis et al., 1995; Gouw et al., 2012; Puig, Casillas, et al., 2015; Berntorp et al., 2021). In the reference genome (GRCh38), one copy is located inside intron 22 of that gene (int22h-1), and the other two are located 565 kb away (closer to the telomere), being int22h-2 in direct orientation and int22h-3 in reverse orientation with respect to int22h-1 (Lakich et al., 1993; Bagnall et al., 2005; Turner et al., 2006). Repetitive sequences can lead to mispairing and generation of an inversion that disrupts the *F8* gene, splitting it

into two parts, one containing the promoter with exons 1 to 22 and the other containing exons 23 to 26. The presence of 40-kb IRs in opposite orientations flanking int22h-2 and int22h-3 generates a polymorphic inversion known as HsInv0608. Each orientation results in a distinct genomic structure, which can both lead to a disruption of the factor VIII gene. In the reference genome, int22h-3 is inverted relative to int22h-1, whereas in the alternative orientation created by HsInv0608, it is int22h-2 that is inverted relative to int22h-1. Then the inversion causing haemophilia A can be generated by the interchange of int22h-1 and int22h-3 (inversion type 1) or by int22h-1 and int22h-2 (inversion type 2), depending on the status of HsInv0608 (Bagnall et al., 2005; Turner et al., 2006).

In the same region, an additional inversion was also found that could be the explanation of between 2-5% of severe cases of haemophilia A. This inversion is produced by recombination of another repeat with a copy in intron 1 of the *F8* gene (called int1h-1) and another located 125 kb upstream (called int1h-2), resulting in the generation of hybrid transcripts between the *F8* gene and the *VBP1* gene (Bagnall et al., 2002; Gouw et al., 2012; Oldenburg et al., 2014; Puig, Casillas, et al., 2015).

Another disease caused by an inversion in chromosome X is Hunter syndrome, which is also known as mucopolysaccharidosis type II. This is a rare disease caused by the deficiency of the lysosomal enzyme iduronate 2-sulfatase (encoded by the *IDS* gene), which is involved in the catabolism of glycosaminoglycans. The disruption of the *IDS* gene generates the accumulation of glycosaminoglycans in body tissues. In Hunter syndrome, the inversion is generated by a NAHR inversion between the *IDS* gene and a nearby pseudogene called *IDS2*, generating a disruption in intron 7 of the *IDS* gene. This inversion is found in 13% of the patients (Bondeson et al., 1995; Puig, Casillas, et al., 2015). In addition, inversions have also been linked to the disruption of coding genes and the creation of fusion genes related to prognosis in cancer in several unrelated patients (Soda et al., 2007; Gruber et al., 2012; Puig, Casillas, et al., 2015). Despite clear examples of inversions that directly cause disease, many conditions or traits are complex and are influenced by multiple loci, as well as environmental factors. In such cases, the impact of individual variants is more

moderate, and inversions are studied in terms of their association with the trait rather than as the sole causal agents.

Inversions associated with traits and diseases

Within the pool of inversions that have been identified in humans, associations with phenotypes have been determined using GWAS information, since in some cases inversion alleles can be predicted from nucleotide variation data. Some examples were already mentioned, including the 17q21.31 (Stefansson et al., 2005) and 8p23.1 (Salm et al., 2012) inversions. Other cases include inversion HsInv0031 that has been associated with GWAS signals for the risk of Alzheimer’s disease, as well as lower levels of *FAM92B* expression in the cerebellum (Giner-Delgado et al., 2019) or 16p11.2 (HsInv0786) that displays associations with diseases such as asthma, obesity, paediatric autoimmune diseases, and the presence of autoantibodies with type 1 diabetes (González et al., 2014a; Puig et al., 2020).

As we mentioned, some inversions disrupt protein-coding genes. That is the case of HsInv0379. This inversion of 415 kb is found on chromosome 19 and it has been reported that the second breakpoint of the inversion, located in an intron, disrupts the protein-coding gene *ZNF257*, which encodes a transcription factor. This gene presents two main transcripts which differ by the presence of an extra non-coding second exon. In this case, inversion removes the first two exons and the promoter, which is confirmed by the lack of expression of the gene in individuals carrying both inverted alleles and the decrease in expression in heterozygotes. In addition, a fusion transcript was detected at breakpoint 1, which could alter gene expression and potentially have functional consequences. This inversion is found almost exclusively in individuals of East Asian ancestry, with a mean frequency of 5% (Puig, Castellano, et al., 2015). In addition, one of the SNPs associated with this inversion was found to be associated with type 2 diabetes in East Asian individuals (Spracklen et al., 2020).

In some cases, inversions can affect two genes placed at each breakpoint. This can happen in NAHR inversions with genes within IRs that recombine and exchange se-

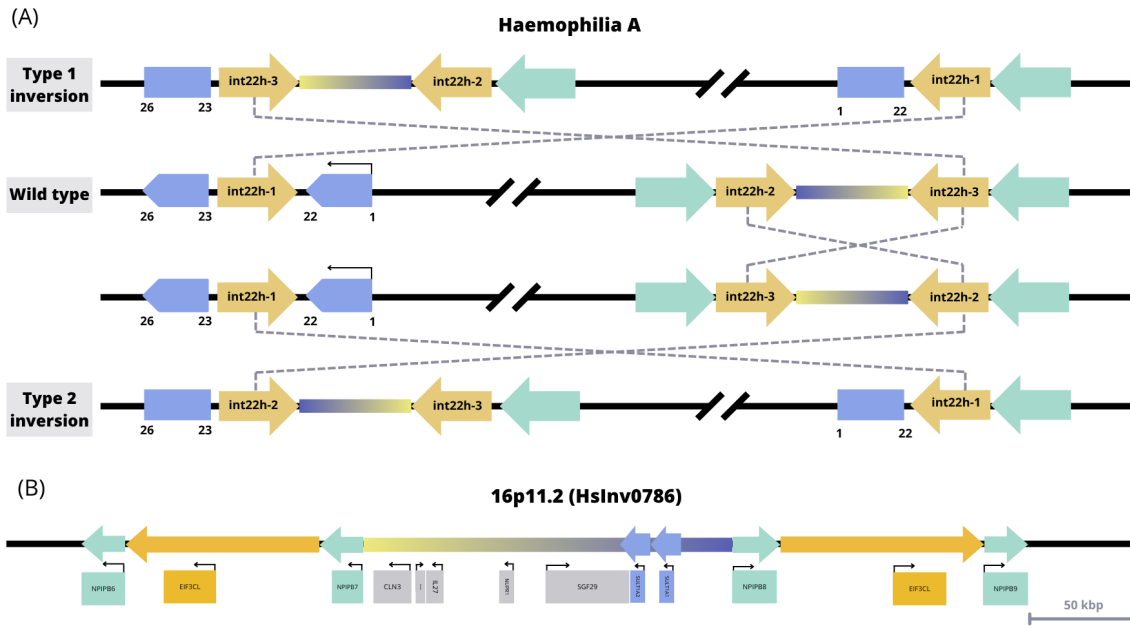


Figure 1.5: **Inversions associated with diseases and functional effects.** (A) Rearrangements within the inversion region related to Haemophilia A. Within the Factor VIII gene (shown in blue) lies a repeated sequence, int22h-1 (in yellow). This sequence can undergo recombination with either of two homologous repeats, int22h-2 or int22h-3, located approximately 550 kb away. Such recombination separates exons 1–22 and the promoter region from exons 23–26 of the Factor VIII gene. (B) Reference structure of the 16p11.2 region in GRCh38. The 16p11.2 region contains multiple sequences with high similarity, represented by arrows of the same color. Some of these repetitive sequences include genes; genes from the same family are shown in the same color, whereas unique genes are shown in grey. The high sequence identity across this region suggests a potential for multiple rearrangements, and the presence of genes within these repeats may help explain associations reported in GWAS hits involving 16p11.2.

quences between them when inversion occurs. The effect produced by this exchange will depend on the final transcripts created; if the genes have enough differences, a hybrid transcript will be created and the expression or function can be affected. An example of this is HsInv0030, which is an inversion of 16 kb located on chromosome 16 between two inverted SDs containing the two chymotrypsinogen precursor genes *CTRB1* and *CTRB2*. When inversion occurs, an exchange of promoter, exon 1, and part of intron 1 is produced between *CTRB1* and *CTRB2* (Pang et al., 2013; Puig, Casillas, et al., 2015; Giner-Delgado et al., 2019). This inversion has been associated with changes in the expression ratio of the *CTRB1-CTRB2* isoforms, affecting the protective capacity to degradation of trypsinogen and, ultimately, being associated

with an increased risk of pancreatitis in inverted carriers (Rosendahl et al., 2018; Seltsam et al., 2020). In addition, HsInv0030 shows several GWAS signals involving type 1 and 2 diabetes, pancreatic cancer, insulin secretion, and cholesterol and triglyceride levels (Giner-Delgado et al., 2019). In the study of this region, a 585 bp deletion was also found that affects 6% of the inverted alleles, eliminating exon 6 of *CTRB2* (Puig, Casillas, et al., 2015), but no association has been found for this variant (Lerga Jaso, 2020).

There are other cases in which inversions alter the gene structure, for example when the BPs of the inversion are found within the same gene. An example of this is HsInv0102, which inverts a non-coding exon of an *RHOH* isoform eliminating it from the transcript. However, this effect is masked by the presence of a lead expression Quantitative Trait Loci (eQTL) that acts in the same direction. For that reason, more studies on the expression difference between isoforms are needed to unravel the potential functional effect of some inversions (Puig, Casillas, et al., 2015; Giner-Delgado et al., 2019). Other examples found are HsInv0201 in which an associated deletion removes an exon of *SPINK14*, HsInv0340 in which the long non-coding RNA *LINC00395* is disrupted, HsInv1051 which breaks *CCDC144B* gene and creates a new fusion transcript with new sequences at 3', and HsInv0124 in which an alternative *IFITM2* isoform is disrupted (Giner-Delgado et al., 2019). Despite the multiple effects found in inversions, their associations with a phenotypic trait are largely unknown due to the limited number of inversions and individuals that can be studied due to their challenging genotyping.

Position effect

In addition to the effects of inversions on genes or regulatory regions, inversions can influence expression without a direct mutation on the gene sequence, produced by a change in the position of a gene relative to its original chromosomal environment. That effect is called the "position effect" and was coined in the early twentieth century. This kind of effect should not cause alteration in the transcription unit and the minimal promoter of the rearranged gene (Kleinjan & van Heyningen, 1998;

Chen & Zhang, 2016). Based on this definition, inversions by their own nature can be drivers of this effect due to the change in position and orientation of genes.

There are multiple situations that can explain the change in expression caused by the placement of a gene with respect to their environment. The first case is the separation of a gene from its regulatory elements, which removes the effect of these regulators on the gene. Depending on the regulatory element, the effect on expression can change. On the one hand, if the inverted allele causes a displacement from an enhancer, a reduction in gene transcription can be produced. However, if the affected element is a silencer, the inversion can cause inappropriate gene activation in certain tissues or cell types. The second case could occur when a gene is placed under the control of regulatory elements of another gene, which can cause an increase in its expression. In addition, a change in expression could be caused by competition for the interaction between both genes and regulatory elements (Kleinjan & van Heyningen, 1998). An example of the effects caused by this kind of position effect would be the one shown in a translocation in Burkitt lymphoma, which places the c-MYC gene under the control of an immunoglobulin enhancer, activating c-MYC (Kleinjan & van Heyningen, 1998; Boerma et al., 2009). Finally, inversions can produce a change in expression by position effect variegation (PEV), which is the alteration of gene expression caused by changes in local chromatin structure, either by locating genes from euchromatic regions to heterochromatic regions or vice versa (Kleinjan & van Heyningen, 1998). A known example can be found in *Drosophila* in which the white gene (w) located at the distal tip of the X chromosome is located next to the pericentric heterochromatin by an inversion called white-mottled-4. At the molecular level, promoter methylation is produced, generating a reduction in gene expression. Mutants of this variation have red and white tissue patches in the eye, while the wild type has a red eye phenotype (Vogel et al., 2009).

Position effects can also have effects on the tridimensional structure of the genome, affecting topologically associated domains (TADs), which can be defined as chromatin segments with a high intra-domain frequency of chromatin contact (Dixon et al., 2012; Rajderkar et al., 2023; Pande et al., 2025). TADs define neighbour-

hoods of chromatin often flanked by CCCTC-binding factor (CTCF) sites in which regulatory sequences can interact with genes within TADs while at the same time interaction with genes outside of those TADs is prevented. SVs have been proposed to disrupt the boundaries of TADs, causing expression changes in some human diseases (Lupiáñez et al., 2015; Flavahan et al., 2016; Franke et al., 2016; Hnisz et al., 2016; Kraft et al., 2019). An illustrative example of the impact of rearrangements on TADs is provided by the study by Kraft et al. (2019), who induced a series of inversions at the *Epha4* locus in mice (Lupiáñez et al., 2015). This locus has previously been reported to be susceptible to genomic rearrangements causing limb and craniofacial malformations. In their experiment, inversions disrupted the normal TAD configuration, leading to ectopic gene activation and resulting in congenital malformations such as severe polydactyly (Kraft et al., 2019). In conclusion, inversions that are not modifying gene sequences are not necessarily functionally neutral; when an inversion encompasses a non-coding segment it can modify the regulatory machinery through the position effect. However, until now position effects have not been demonstrated for any polymorphic inversion in humans.

Inversions predispose to other rearrangements

Another consequence of inversions can be the increase in the predisposition to generate further rearrangements that can lead to diseases. Sometimes, inversions present large and highly identical IRs that can act as drivers for the generation of these secondary rearrangements. Often these rearrangements are CNVs (such as microdeletions and microduplications) with phenotypical effects. In these situations, inversions can act by changing the orientation of a repetitive element and increasing the length of repeats in direct orientation (that can generate duplications and deletions through errors in recombination) (Sharp et al., 2006; Feuk, 2010; Puig, Casillas, et al., 2015; Porubsky et al., 2022). This effect can be observed in parents of individuals affected by genomic syndromes that tend to have higher frequencies of the inverted allele compared to the general population (Puig, Casillas, et al., 2015). In 2022, Porubsky and collaborators found a significant co-localisation between morbid CNVs, associated with paediatric developmental delay and neuropsychiatric disor-

ders, and inversions. In addition, this association shows an enrichment of 5-fold for recurrent inversions, suggesting a clear relationship between inversion recurrence and morbid CNVs (Porubsky et al., 2022).

Among known polymorphic inversions, several cases have been associated with secondary rearrangements in humans. One of the classic examples is the already mentioned 17q21.31 inversion, which presents two main haplotypes H1 and H2 that do not recombine between them. Inside these main groups, there are several sub-haplotypes that present copy number variable regions. One of those is H2D ($H_2\alpha_2\gamma_2$) that has two duplications of 150 kb in direct orientation, which has been associated with a predisposition to deletion of the entire region that encompasses multiple genes, including *MAPT*, *KANSL1*, *SPPL2C*, *STH*, and *CRHR1*. This deletion can affect 500 to 650 kb and is generated through NAHR mechanisms involving IRs in direct orientation (Steinberg et al., 2012; Campoy et al., 2022). The result of these deletions is the Koolen-de Vries syndrome (KdVS) (also known as 17q21.31 microdeletion syndrome), which is characterised by developmental delay, intellectual disability, hypotonia, epilepsy, congenital malformations, and recognisable facial features (Koolen et al., 2006; Boettger et al., 2012; Puig, Casillas, et al., 2015; Koolen et al., 2016). Almost all 17q21.31 microdeletion syndrome cases reported belong to patients of European ancestry (Mefford et al., 2009; Cooper et al., 2011), which is explained by the fact that the H2D haplotype has a higher frequency in the European population ($\sim 20\%$) compared to Asian and African populations in which it is a rare haplotype (Boettger et al., 2012). In addition, all parents of affected individuals present the inverted allele, which is a higher frequency than 20% observed in the European normal population (Koolen et al., 2008).

Another relevant example of secondary rearrangements associated with diseases is the deletion of the chromosome region 7q11.23, which includes 28 genes in a 1.5-1.8 Mb region. The consequence of this deletion is Williams-Beuren syndrome, which is a rare disease characterised by mental retardation, vascular stenosis, visual problems, over-friendliness, short stature, and specific facial characteristics (Osborne & Mervis, 2007; Merla et al., 2010). The 7q11.23 region presents a known polymorphic

inversion flanked by IRs generated by NAHR mechanisms. As before, the parents of affected individuals are more frequently carriers of the inverted allele (25%) than the general population (5%) (Bayés et al., 2003; Hobart et al., 2010; Lühmann et al., 2023). In the cases mentioned above, it is assumed that the origin of the disease is the haploinsufficiency of the gene contained in the removed region.

The final examples are Angelman syndrome and Sotos syndrome. The former involves a 4-Mb deletion of maternal origin and an inversion located in the region 15q11-q13. In this case, the inversion shows a frequency of 67% in parents compared to 9% in the general population (Gimelli et al., 2003; Puig, Casillas, et al., 2015). The latter, Sotos syndrome, involves a deletion derived from the paternal chromosome in the 5q35 region, where all fathers were heterozygous for an inversion at this locus (Visser et al., 2005). In contrast to the other cases, the 3q29 microdeletion syndrome, which is caused by a deletion of ~ 1.6 Mb in chromosome 3, presents a polymorphic inversion that seems to play a protective role against microdeletion (Ballif et al., 2008; Porubsky et al., 2022). Despite all the cases explained, not all secondary rearrangements are deletions, there are also cases in which inversions promote translocations. This is the case of a region on chromosome Y that presents an inversion of 4 Mb that predisposes to the a translocation between the PRKX and PRKY genes located on chromosomes X and Y, respectively. Sex reversal and infertility issues are related to translocation (Jobling et al., 1998).

All the effects we have identified throughout the previous sections highlight the importance of continuing to study inversions. Despite the many cases described above, human inversions remain, to a large extent, a mystery. The knowledge we have today has only been possible due to significant efforts devoted to the detection of these variants and the genotyping of individuals. However, much of the existing research has focused on inversions that are not flanked by SDs, as these are generally easier to detect and characterise. In contrast, SD-mediated inversions remain challenging to resolve due to the high sequence similarity and complex genomic architecture of these regions, and their functional consequences are still poorly understood. Ultimately, to uncover the functional impact of inversions, it is necessary first to detect

and genotype them. In the following section, we examine how the techniques developed for this purpose have evolved.

1.4 Inversion detection

Now we go back to the beginnings of this text, remembering and amplifying what was commented on SVs detection, but pointing directly to the case of inversions with their particularities. As mentioned above, inversions were detected initially using cytogenetic techniques such as G-band karyotyping, fluorescence in situ hybridisation (FISH), Southern blot hybridisation or pulsed field gel electrophoresis. However, these techniques only allow for the detection of large SVs (3 Mb) and have limited power to perform large-scale studies (Aguado et al., 2014; Puig, Casillas, et al., 2015; Durmaz et al., 2015; Porubsky & Eichler, 2024). In the early 2000s, bacterial artificial chromosome (BAC)-based, oligonucleotide, and SNPs-based microarrays were widely used to detect SVs, allowing the identification of variants not visible at the karyotype level and observing that SVs represent a higher amount of genetic variation in phenotypically normal individuals than previously thought. Due to these techniques, hundreds of variants were described for the first time, but inversions have lagged behind due to their balanced nature and association with IRs (Feuk, 2010; Escaramís et al., 2015; Porubsky & Eichler, 2024). Advancements in inversion detection were achieved with the use of sequencing methods, which facilitated comparison among independently assembled genomes, enabling identification of inverted segments between assemblies. This approach was also used to compare sequences between humans and primates, to find variants between species, and to reveal the evolution of inversions along lineages (Feuk et al., 2005; Levy et al., 2007; Ventura et al., 2007; Puig, Casillas, et al., 2015). However, this methodology was considerably expensive, not only because it required sequencing the entire genome, but also due to complications introduced by SD-mediated inversions, which hinder accurate reconstruction through genome assembly. Improvements in both SVs detection algorithms and NGS technologies were required to allow a more accurate identification of inversions between samples and populations, thus paving the way

for new studies to uncover the functional and evolutionary functions of inversions.

However, the key innovation came with paired-end sequencing and mapping (PEM), which enabled a substantial improvement in the detection of all forms of SVs, becoming the major advance in the identification of inversions (Kidd et al., 2008; Feuk, 2010; Puig, Casillas, et al., 2015). This methodology consists of the following: For a given sample, the two ends of a DNA segment with a known size are sequenced. The resulting reads are mapped against a reference genome. The expected result is that the two resulting sequences map in the reference genome with a distance approximately of the fragment size and in positive and negative orientation, respectively. However, if an inversion is present, the end sequences of fragments that span the BPs of the inversion will exhibit a positive/positive or negative/negative strand with respect to the reference genome and also a difference in the expected distance between them. One consideration to take into account is the fact that when the two end sequences are located inside the inversion, this pattern is lost. Then, by searching for the read clusters that show this grouping of patterns, it is possible to identify inversion BPs (Feuk, 2010; Escaramís et al., 2015; Puig, Casillas, et al., 2015). Moreover, when discordant paired reads show differences in the distance between them, it suggests the presence of deletions or insertions, while if the paired reads map in different chromosomes, it could be indicative of translocations. PEM was applied for the first time in fosmid clones using traditional Sanger sequencing, identifying 56 inversion BPs in a library of a single human genome (Tuzun et al., 2005; Feuk, 2010). Later, the same methodology was applied to eight samples, detecting 217 inversions (Kidd et al., 2008; Feuk, 2010). PEM was later adapted to use NGS platforms in the sequencing of paired-end fragments (Korbel et al., 2007; Ahn et al., 2009). *De novo* genome assembly, split reads or read depth became other techniques that were used in combination with PEM (Sudmant et al., 2015; Hehir-Kwa et al., 2016; Abel et al., 2020; Collins et al., 2020). Despite the remarkable success of PEM in detecting SVs, several limitations remain. Complex genomes, such as the human genome, which contain extensive repetitive regions can hinder accurate assembly. This challenge becomes particularly critical when paired-end reads are mapped onto a reference genome that itself may contain misassembled regions. In

these circumstances, repetitive sequences often lead to spurious SVs (Aguado et al., 2014; Puig, Casillas, et al., 2015; Vicente-Salvador et al., 2017). In addition, inversions mediated by IRs frequently generate false positive inversion calls, mainly due to difficulties in assigning unique read placements near the BPs. Estimates indicate that more than 80% of inversions flanked by SDs, and between 5 and 50% of those flanked by repetitive elements remain undetected by PEM (Lledó & Cáceres, 2013). In general, typical short reads produced by NGS, would be missing more than 70% of human SVs (Logsdon et al., 2020). Moreover, the size of the fosmids used limits the detection of variants within large IRs, and regions containing multiple SVs produce complex patterns that are extremely difficult to resolve with PEM. Consequently, these technical constraints substantially limit the spectrum of variants that can be reliably detected using this approach. The other main problem with the use of this technology is that it usually requires to perform targeted validations using independent methods. Although PEM can detect regions with discordant read mappings, it does not precisely define the exact location of BPs, nor the sequence changes produced by the variant. Consequently, targeted experimental validations, most commonly using PCR-based techniques are necessary for this validation and also for genotyping at the population level.

In recent years, new techniques have been developed that have revolutionised SVs detection and genotyping, enabling the identification of a greater number of variants with greater reliability. These methods include optical mapping, Single-cell template strand sequencing (Strand-seq), and long-read sequencing technologies. Although we will discuss all of them later, we first introduce some experimental approaches used for the validation and large-scale genotyping of inversions.

1.4.1 Inversion validation

As mentioned, PEM and other detection techniques were able to identify candidate regions that potentially contained SVs, and in particular inversions. However, some of these proved to be false positives, either because of the complexity of the human genome or because of limitations in the reference genome used for comparison.

The first method used to validate PEM inversion predictions, and also one of the methods used at the beginning of our story, is FISH. FISH is a cytogenetic technique consisting of the use of fluorescent probes designed to hybridize to complementarily specific genomic regions. When a candidate inversion is found, two or three probes are labelled with different colours, which are located outside and inside the inversion. Through the analysis of the distance and order of the probes, the orientation of the inversion region can be determined. However, FISH is restricted to large regions, and its application to the simultaneous analysis of many samples and inversions has become a major limitation (Kallioniemi et al., 1992; Antonacci et al., 2009; Bosch et al., 2009; Hobart et al., 2010; Sheen et al., 2011; Nomura et al., 2018). Southern-blot hybridisation (Small et al., 1997) and pulsed-field gel electrophoresis (Osborne et al., 2001) have also been used for the validation of inversions. However, both methods face limitations in scalability, making them unsuitable for population-scale studies or large cohorts, which are necessary to perform association analyses.

The other main approach used for inversion validation and genotyping is PCR-based. This technique relies on primers designed at both sides of a breakpoint to produce orientation-specific amplicons that allow us to determine whether an individual carries the normal or inverted orientation (or both) in their chromosomes (Feuk et al., 2005; Korbelt et al., 2007; Pang et al., 2013). However, traditional PCR is capable of generating amplicons of a few kb in maximum size, preventing the analysis of inversion BPs with large repeats (Shagin et al., 1999; Aguado et al., 2014). To solve this problem, variations were introduced in the PCR protocols, especially to improve the genotyping of inversions with SDs at inversion BPs.

For instance, inverse PCR (iPCR), which includes steps prior to amplification of DNA. In iPCR, the first step consists in the digestion of DNA using restriction enzymes and the self-ligation of the generated fragments to create circular molecules. The junction site is then amplified with primers located on both sides, avoiding amplification throughout the breakpoint. This method has been used for the genotyping of the inversion that affects factor VIII that causes haemophilia A and also for the validation of multiple polymorphic inversions (Rossetti et al., 2005; Aguado

et al., 2014; Puig, Casillas, et al., 2015; Vicente-Salvador et al., 2017). Other variations used in inversion genotyping include long-range PCR and haplotype-fusion PCR (Turner et al., 2006, 2008). Another variation of PCR is digital droplet PCR (ddPCR), which has shown excellent performance in the genotyping of inversions with long repeats at their BPs. It detects the linkage between amplicons located on both sides of the IRs that contain the inversion BPs, allowing the identification of rearranged orientations over long genomic distances (Puig et al., 2020). These methods allow for the genotyping of inversions in a relatively large number of samples. In some cases multiple inversions can be genotyped at the same time using techniques such as multiplex ligation-dependent probe amplification (MLPA) or inverse MLPA (iMLPA) combining iPCR and MLPA detection, which allow the use of multiple primers to genotype multiple inversions in a single experiment (Schouten et al., 2002; Giner-Delgado et al., 2019). Despite their strong performance, all these experimental approaches have technique-specific requirements: iPCR depends on the presence of restriction sites in particular positions, ddPCR requires sufficiently large inversions or specific restriction sites, and MLPA/ iMLPA requires simple and well-annotated BPs. Consequently, not all inversions can be analysed using any single method, and some, particularly those with extremely long IRs (>100 kb) or those that cannot be resolved by digestion of restriction enzymes, cannot be accurately validated or genotyped using these experimental methods (Puig et al., 2020). Moreover, the number of samples that can be studied using these experimental approaches is limited, so multiple bioinformatic methodologies have been designed to complement or substitute experimental methods.

Most bioinformatic approaches are based on the presence of other variants in strong LD with inversions. An example is the use of tagSNPs, whose detection requires inversion genotypes for multiple samples. In this case SNPs with alleles in high LD with inversion alleles can be used as a proxy to infer the inversion status in additional samples (Steinberg et al., 2012). The main limitation of this methodology arises from inversions mediated by SDs. The recurrent nature of these inversions — appearing multiple times on different sequence backgrounds — often results in low LD with the surrounding variants. Consequently, many SNPs are shared between

chromosomes that carry the inversion and those without it, so there are no tagSNPs (Vicente-Salvador et al., 2017; Giner-Delgado et al., 2019). Other methodologies are based on imputation, which consists in inferring the inversion state from the information of surrounding SNPs. This is the case for tools such as IMPUTE2 and IMPUTE5, which compare the SNP haplotypes of an unknown sample with those in a known reference panel to identify the most similar haplotypes, thus attributing the corresponding inversion status. The difference between these two tools is that one uses the hidden Markov model (HMM) (IMPUTE2) and the other uses the positional Burrows–Wheeler Transform (PBWT) model (Howie et al., 2009; Rubinacci et al., 2020). Moreover, in the case of inversion identification and genotyping, specific software to call variants has been designed based on the detection of signatures in nucleotide variation patterns, such as PFIDO, invClust or scoreInvHap (Bansal et al., 2007; Sindi & Raphael, 2010; Cáceres & González, 2015; Ruiz-Arenas et al., 2019; Balagué-Dobón et al., 2022). The main limitation of these programmes is the reliability of the genotypes obtained, especially in inversions mediated by SDs that are probably not captured by any combination of linked genetic variants.

The combination of these experimental and bioinformatic methods has made possible the study of specific inversions and the determination of their worldwide distribution, their frequencies in different populations, and, in some cases, their association with functional consequences. It is important to note that without reliable techniques to identify inversions and obtain accurate genotypes, which are the first steps, the study of the effects of inversions would not be feasible. However, the new approaches discussed below open a new chapter in this field and promise to finally discover how many common inversions exist, their size distribution, and the degree to which inversions are linked to human diseases (Puig, Casillas, et al., 2015).

1.4.2 New methodologies

In this section, we will discuss new technologies that emerge and show promising results in the identification and genotyping of inversions. In addition, new software will also be discussed to include callers for SVs, new assembly methodologies, and

new reference approaches. We will start by explaining experimental methodologies.

Optical maps

Although this section being about new technologies, optical maps have been present since 1993, when they were used in the assembly of microorganism genomes (Schwartz et al., 1993). However, low throughput, high error rate, and imprecise measurements of DNA length hampered the use of the methodology in the study of more complex species. However, due to the improvements introduced by Bionano Genomics, the technology has more recently become a relevant tool in genomic research (Yuan et al., 2020). The experimental design of optical maps is based on the obtention of long DNA molecules (~ 1 Mb), which are digested using an endonuclease to create single-strand nicks. Those nicks are repaired using fluorescently-labelled nucleotides that generate fluorescent marks along the DNA molecules. The DNA molecules are then linearised in a nanochannel array present in a chip, and the images of the extended marked DNA molecules are taken using high-resolution fluorescent microscopy. The resulting images are converted into a format accepted by software capable of performing *de novo* assembly and phasing haplotypes to generate the optical maps. In the case of SVs detection, the final step consists of recognising the location of the restriction site labels on each DNA molecule and comparing it with the pattern observed on the reference sequence. The detection of changes in the pattern of the label indicates the presence of SVs that can be characterised based on the type of change in the pattern (Figure 1.6-B). For example, if you find certain labels located farther away than in the in-silico map generated as a reference, it would be indicative of insertion, or deletion if they are closer than expected. The same applies to duplicated or inverted patterns (indicative of duplications or inversions, respectively). There are multiple tools that can be used throughout the optical mapping workflow, including alignment, variant discovery, *de novo* assembly, and scaffolding (Li et al., 2017; Yuan et al., 2020). The main advantage of the technique is the ability to study large rearrangements, especially those within repetitive regions. One of the limitations of optical maps is that they only represent the position of the selected nicking enzymes without generating nucleotide information. Therefore, the BPs of

the variants cannot be defined with precision because they provide only breakpoint estimates based on the closest nicks. Moreover, because of the use of nicks, optical maps have a relevant limitation in the identification of small variants that do not change the label pattern. For that reason, it can be complemented with the use of short-read sequencing technologies to improve the detection of small SVs. The use of multiple endonucleases can provide different patterns that can resolve some of the limitations (Ho et al., 2020; Yuan et al., 2020). Optical maps have also been used to characterise inversions in recent years and are becoming an important tool for the identification of clinically relevant variants (Iqbal et al., 2023). Optical maps have been used in the analysis of the inversion that affects factor VIII, helping in the identification of novel variants in the region (Liu et al., 2025). In healthy individuals, optical mapping has been applied to 154 genomes from 1000GP to identify multiple SVs, including insertions (~ 6500), deletions (~ 9100), and inversions (380) (Levy-Sakin et al., 2019).

Strand-seq

The next methodology is Strand-seq, a sequencing technique that allows the sequencing of a single strand of DNA. This is achieved by incorporating 5-bromo-2'-deoxyuridine (BrdU), a thymidine analogue, into the newly synthesised (non-template) strand during DNA replication in a cell line of the individual analysed. Subsequent UV-induced photolysis at BrdU sites selectively removes the newly synthesised strand (positive BrdU), leaving only the original template strand. The remaining template strand (BrdU-negative) is then used for library construction, and sequenced preserving their directionality. After single-cell DNA sequencing, the reads are aligned with the reference genome. Since all reads come from a single strand, they should all map in the same strand of the reference genome, unless there are inversions, in which case the reads will switch their mapping to the complementary strand of the reference (Falconer et al., 2012; Sanders et al., 2016, 2017; Ho et al., 2020). Sequencing of several individual treated daughter cells is needed to analyse all the strands of the two chromosomes of the tested individual. The main point of this technology is the preservation of genomic directionality, which makes

that especially interesting in the study of inversions. Inversions can be observed as genomic regions in which the reads belonging to the inverted region align with the complementary DNA strand contrary to the surrounding sequences (Figure 1.6-C). Moreover, other SVs such as deletions and insertions can be detected by analysing changes in the depth of sequencing. The limitation of this methodology is that due to the multiple enzymatic clean-up steps, the average coverage obtained is low (0.01-0.05x), which is not useful for the detection of small SVs (Sanders et al., 2016; Ho et al., 2020). Strand-seq has been successfully applied to the identification of inversions in humans, allowing the detection of variants in a wide size range (1 kb–4.5 Mb), including regions located between large blocks of SDs (Sanders et al., 2016), as well as in apes (Porubsky, Sanders, Höps, et al., 2020). In addition to experimental advances, such as the simultaneous construction of multiple libraries and their automation (Sanders et al., 2017), significant software improvements have also been achieved, increasing both the number and the performance of tools dedicated to inversion detection (Hanlon et al., 2021; Weber et al., 2023; Bai et al., 2024).

Long-read sequencing

In this section, we discuss the two main technologies for single-molecule sequencing, PacBio and ONT long reads, which enable the generation of continuous reads spanning up to hundreds of kilobases directly from native DNA (Ho et al., 2020; Logsdon et al., 2020). Long-reads due to their size are able to span low complexity and repetitive sequences such as SDs and mobile elements, which sometimes are found at the BPs of SVs, including inversions. This feature allows for an improvement in the detection of SVs compared to short-read approaches (Pollard et al., 2018; De Coster et al., 2021) (Figure 1.6-D). In this context, while short-read technologies allow the identification of fewer than 10,000 SVs per human genome, long-read sequencing allows the detection of more than 20,000 variants (Ho et al., 2020). Long-read sequencing technologies are grouped into third-generation sequencing methods (Amarasinghe et al., 2020; Olson et al., 2023).

PacBio long reads

The first of the two technologies we will discuss emerged in 2011 (Amarasinghe et al., 2020). PacBio introduced the single-molecular real-time (SMRT) platform, which initially offered two sequencing modes: continuous long reads (CLR) (now deprecated) and Circular Consensus Sequencing (CCS), producing HiFi reads. CLR long reads are obtained by constructing SMRTbell template libraries containing DNA inserts greater than 30 kb, allowing the polymerase to traverse the template only once or a few times. This process generates reads that typically range from 5 to 60 kb, with maximum lengths approaching 100 kb. However, the accuracy of this mode is relatively low, averaging between 85 and 92%. In contrast, the use of smaller libraries of SMRTbell templates (10-30 kb) in CCS mode allows the polymerase to perform multiple passes through the template, generating a long continuous polymerase read (up to 150 kb long) that includes both forward and reverse passes of the DNA template. These multiple passes are then combined to produce a consensus, resulting in high-fidelity (HiFi) reads with a greatly increased accuracy (99%) and lengths corresponding to the size of the insert (10–30 kb) (Espinosa et al., 2024) (Figure 1.6-D). The increased accuracy of PacBio long reads has enabled improvements in variant discovery, reduced assembly time, and better access to complex genomic regions (Ho et al., 2020; Logsdon et al., 2020).

ONT long reads

Oxford Nanopore Technologies (ONT) long reads, first introduced in 2014, are generated by passing single-stranded DNA molecules through a protein nanopore. The DNA sequence is then inferred from characteristic changes in electrical current as individual bases move through the pore. Similarly to PacBio long reads, ONT sequencing is also a single molecule sequencing technique (Amarasinghe et al., 2020; Ho et al., 2020). The main distinction between the two long-read sequencing technologies lies in the fact that in ONT no DNA synthesis takes place and in the final read length. ONT can generate considerably longer reads, sometimes reaching hundreds of kilobases, generating reads able to span large complex repetitive genomic

regions (Figure 1.6-D). However, such ultra-long reads represent only a small fraction of the data. The read length in ONT sequencing is primarily determined by the extraction and preparation of high molecular weight DNA. Based on this, ONT reads are typically classified into two categories: standard long reads (10–100 kb) and ultra-long reads (>100 kb) (Logsdon et al., 2020).

A key limitation of ONT sequencing is its reduced accuracy per base, which generally ranges from 87% to 98% (De Coster et al., 2021; Logsdon et al., 2020; De Coster et al., 2021). Moreover ONT long reads show many specific small deletions produced from basecalling errors (Ho et al., 2020). Due to the relatively low accuracy per base of ONT long reads, studies that require high base-level precision often rely on complementary error correction or polishing tools. These approaches may be non-hybrid, using only long-read data, or hybrid, combining long-reads with short-read data to improve accuracy (Amarasinghe et al., 2020). The precision of ONT sequencing depends both on the chemistry used and on the algorithms applied during basecalling. Guppy, the most widely used basecaller in recent years, provides three modes: fast, high accuracy (HAC), and super accuracy (SAC), which differ in computational demand and resulting accuracy (Karakurt et al., 2023). More recently, Dorado has emerged as a promising successor, with the aim of becoming the new standard for ONT basecalling (“Nanoporetech/Dorado”, 2022, May 17/2025). Improvements in sequencing chemistry have also contributed substantially to an enhanced base-level accuracy in more recent sequencing data. The latest chemistries and basecallers are approaching accuracy levels comparable to those achieved with PacBio long reads. However, a trade-off remains: generating a higher proportion of ultra-long reads often comes at the expense of reduced overall throughput (Logsdon et al., 2020; Ferguson et al., 2022; Espinosa et al., 2024).

In recent years, several approaches have been developed to reduce the cost of ONT sequencing. One of the most promising strategies is the enrichment of specific loci, which increases the sequencing depth in the targeted regions. Among the proposed methods, CRISPR-based enrichment has been explored as a powerful tool for selective targeting (Geng et al., 2022). However, an alternative that is gaining con-

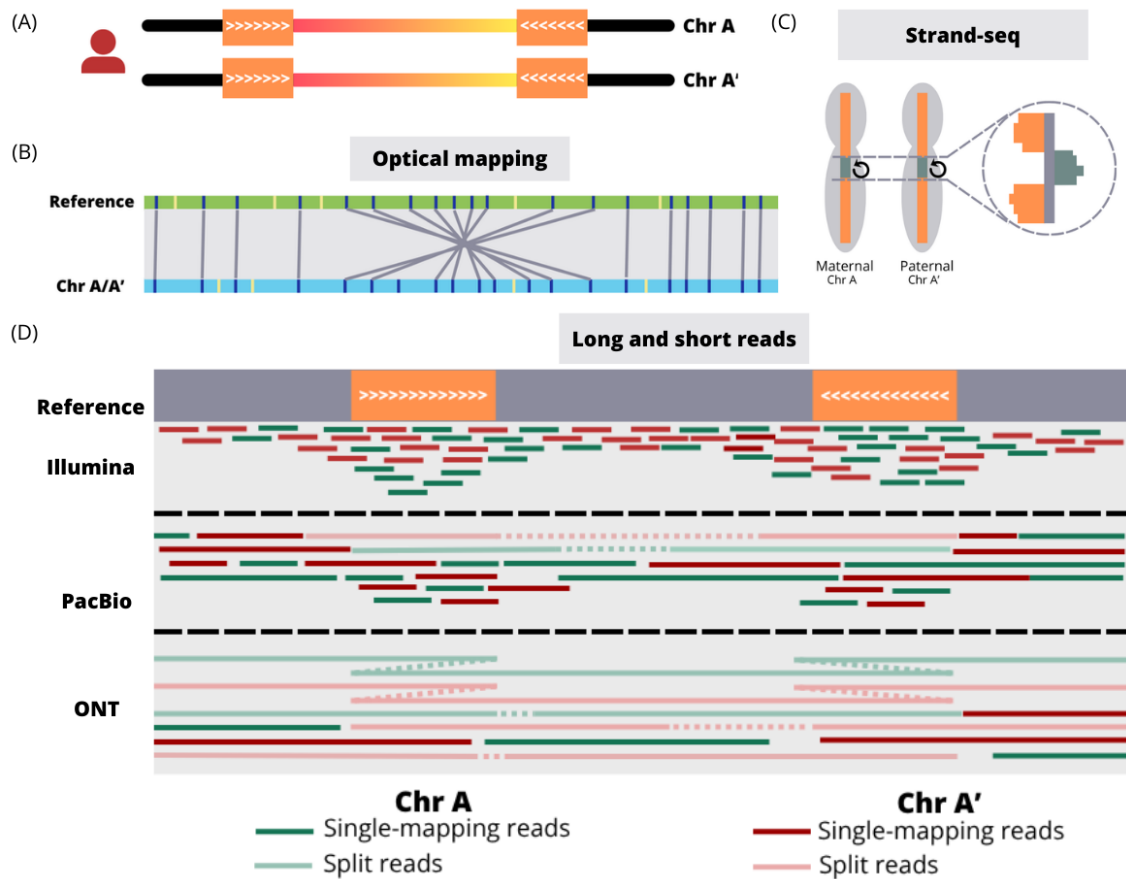


Figure 1.6: **Genotyping inverted alleles using different techniques.** (A) Example of a sample carrying two inverted alleles (ChrA and ChrA') relative to the reference genome. (B) Optical mapping detects inversions by anchoring fluorescent labels (blue as aligned and yellow as unaligned) to specific nucleotide motifs and comparing their order along the molecule with the reference genome. (C) Strand-seq sequences template-specific DNA strands, allowing the detection of inversions through characteristic changes in the read mapping strand in the reference genome. Draw represents the view of an homozygous inversion. (D) Comparison of mapping patterns among sequencing technologies. Short reads, which do not span large repetitive blocks (in orange at the top), often produce double-depth signals in these regions due to multiple mapping hits. PacBio long reads, owing to their greater length (10-30 kb), can span some inversion BPs (light colours), while others map continuously (dark colours) and therefore are uninformative for inversion genotyping. ONT long reads, with even greater length (reaching more than 100 kb), span a higher proportion of inversion BPs (light colours) and show fewer single-mapping reads (dark colors). ChrA is shown as green lines and ChrA' as red lines. Coloured guide lines join different mapping segments of the same read.

siderable attention is adaptive sampling. Adaptive sampling enables the selection of specific DNA sequences in real time, without the need for prior enrichment steps during library preparation. This is achieved through algorithms that align the first

nucleotides of a read (typically 400 bp) to a reference genome, allowing the system to determine whether the read corresponds to predefined target genomic coordinates or not (Chevrier et al., 2025). If the signal does not match the selected target, the DNA molecule is ejected from the nanopore, freeing it to start sequencing a new fragment. This approach is particularly valuable in clinical diagnostics, as it not only reduces sequencing costs but also shortens turnaround times to obtain results (Yang et al., 2025).

In general, long reads generated by the PacBio and ONT technologies have been applied in numerous studies, demonstrating a significant impact in several areas of genetics, including variant discovery (Sudmant et al., 2015; Audano et al., 2019; Beyter et al., 2021), genome assembly (Chaisson et al., 2015, 2019), disease association (Mantere et al., 2019), and human genetic diversity (Hsieh et al., 2019). To support these applications, the development of specialised tools has been essential.

In the field of variant discovery, multiple structural variant callers have been designed, many of which show strong performance. Among the most widely used are Sniffles2, SVIM, PBSV, and cuteSV (Helal et al., 2024; Moustakli et al., 2025). In addition, several studies have used third-generation sequencing methods to detect structural variants such as deletions, insertions, and inversions, demonstrating a substantial increase in detection compared to short-read sequencing. Initially, the cost of large-scale studies using these technologies was prohibitive; however, advances in parallelisation and the continuous reduction in the cost per sample have made it possible to scale up considerably (Tafazoli et al., 2025). Recently, ONT long read sequencing has been applied for the first time to study structural variation in a dataset of more than 1000 samples (Schloissnig et al., 2025), and at the same time, initial efforts are being made to complete all 1000GP samples using ONT long read sequencing (Gustafson et al., 2024).

Genome assembly is another area where long-read sequencing has provided major advantages compared to short-read approaches. The ability of long reads to span complex genomic regions enables the resolution of repeats and substantially increases contiguity, which was crucial for the assembly of the first T2T human genome (Nurk

et al., 2022). Tools such as Canu (Koren et al., 2017), Flye (Kolmogorov et al., 2019), Hifiasm (Cheng et al., 2021), and Verkko (Rautiainen et al., 2023) have been widely applied in genome assembly with long-read data, sometimes relying on a single technology and in other cases combining multiple platforms to leverage the strengths of each of them.

Multipatform

All of the technologies mentioned above have had a significant impact in different genomic fields, particularly in the detection and genotyping of SVs. However, several studies have explored the combined use of these approaches, using their complementary strengths to further improve the results. Combining different techniques and algorithms increases both the sensitivity and specificity of SVs detection (Chaisson et al., 2019; Ebert et al., 2021; Porubsky et al., 2022; Logsdon et al., 2025; Porubsky et al., 2025). Several strategies exist to integrate different technologies to improve detection. A common approach is to use long-read sequencing at low coverage (10x) to identify variants and resolve complex regions, while relying on short reads at high coverage (30x) to refine BPs and improve sequence accuracy (Li et al., 2017; Ho et al., 2020). Furthermore, hybrid assemblies that combine PacBio, ONT, and short-read sequencing have shown strong performance in genome assembly (Di Genova et al., 2021; Rautiainen et al., 2023). Multiple studies in the literature have demonstrated the benefits of combining complementary techniques. For example, in 2019, Chaisson and colleagues analysed three family trios using PacBio, ONT, Strand-seq, optical maps and short reads, detecting more than 27,000 SVs. In this study, Strand-seq performed particularly well for inversion calling (Chaisson et al., 2019). In 2022, Porubsky et al. applied three complementary approaches—Strand-seq, assemblies generated from PacBio Hi-Fi reads and optical maps to 41 unrelated samples, successfully identifying 729 inversion sites, including 292 balanced inversions (Porubsky et al., 2022). The most recent study, published in 2025, analysed 65 samples using PacBio and ONT long reads, Strand-seq, optical maps, and Hi-C sequencing. This approach enabled the detection of 118,500 SVs, including 298 non-redundant inversions, representing the largest effort to date that uses multiple

technologies for the discovery and genotyping of structural variants (Logsdon et al., 2025).

1.4.3 Reference genome

Some methodologies for the detection and validation of inversions and others SVs rely on comparing sequencing reads from a given sample to a reference genome. For this reason, in this section, we briefly review the evolution of the human reference genome. As mentioned above, in 2001 the Human Genome Project Consortium released the first draft of the Reference Human Genome (Lander et al., 2001). In parallel, Celera Genomics also released the Alternate Human Genome assembly (Venter et al., 2001; Singh et al., 2022). But it was not until 2004 that the finished euchromatic genome was finally released (International Human Genome Sequencing Consortium, 2004). This first release has been revised and modified over time, reaching the actual version (GRCh38) which was released in 2013. The main idea of this reference is to represent a linear coordinate system that allows mapping reads against it, serving as a standard for the identification of variations (Singh et al., 2022). The GRC reference genome was constructed using BAC clones, which were ordered and orientated along the human genome using genetic linkage maps, radiation hybrid maps and fingerprint maps. Consequently, this genome underrepresents repetitive sequences, and the assembly of BAC clones from multiple individuals resulted in a mosaic of haplotypes. As a result, the reference contains multiple gaps, especially centromeric and telomeric regions, with some regions remaining unfinished or incorrectly assembled (Nurk et al., 2022; Singh et al., 2022). The current version of the human reference genome (GRCh38) contains approximately 151 Mb of unresolved gaps, distributed throughout the genome and including pericentromeric and subtelomeric regions, recent SDs, and arrays of ribosomal DNA (rDNA).

As mentioned above, long-read sequencing technologies such as PacBio and ONT demonstrated the ability to resolve complex SVs and previously inaccessible genomic gaps. Consequently, a combination of these technologies was used to assemble challenging centromeric regions, including the Y chromosome (Jain, Olsen, et al., 2018)

and even the complete X chromosome (Miga et al., 2020). These achievements set the stage for the sequencing and assembly of the first T2T genome in 2022, which fully resolves centromeric regions, previous missing euchromatic sequences, large SDs and the short arms of five acrocentric chromosomes. These regions account for hundreds of megabases of previously uncharacterised sequences, which can harbour biologically relevant features, including repeat-mediated SVs hotspots and loci associated with chromosomal instability and disease (Nurk et al., 2022). The cell line chosen for this effort was CHM13hTERT (known as CHM13), a essentially homozygous line derived from a complete hydatidiform mole (CHM) with a 46,XX karyotype. CHM13 had already been used to close the gaps in the human reference genome (Schneider et al., 2017), to benchmark genome assemblers, variant callers (Li et al., 2018), and to study human SDs (Vollger et al., 2019; Miga et al., 2020). The resulting genome, known as T2T-CHM13, represents a major advance in genomic analysis, which provides a high quality map of the human genome allowing the discovery of a large number of variants that were previously invisible. For example, in the case of SNPs, T2T-CHM13 allowed the identification of 3.7 million additional variants in regions that are non-syntenic to GRCh38 (Aganezov et al., 2022; Liao et al., 2023). This assembly has also facilitated the discovery of SVs such as inversions: using Strand-seq data from 41 unrelated samples, the researchers identified 63 putative novel inversion sites, several of which lie within pericentromeric regions and are associated with repetitive DNA which are usually not well represented in GRCh38 (Porubsky et al., 2023). Further progress was achieved with the first complete assembly of the Y chromosome, comprising 62 Mb, about 30 Mb more than in GRCh38. Since CHM13 lacks a Y chromosome, the assembly was generated from the HG002 sample, ultimately producing T2T-Y. This sequence includes comprehensive annotations of genes, repeats, and overall structural organisation (Rhie et al., 2023). In addition to the discovery of new variation, the new reference genome allows a more comprehensive annotation of gene families inside SDs and a better understanding of centromere biology, epigenetic regulation, and repeat expansions related to disease. The T2T reference closes the last blind spots of the human genome, establishing a new foundation for genetic and clinical studies (Nurk et al.,

2022).

Despite the substantial improvement provided by T2T-CHM13+Y, limitations remain in the detection and characterisation of variants using a linear reference. One major limitation is that the reference genome does not capture the full extent of variation present in human populations. To address this gap, the first human pangenome was released in 2023 (Liao et al., 2023). Pangenomes are designed to represent human genetic variability using graph-based approaches, typically Bruijn graphs, which were originally developed in 2001 for the assembly of genomes from short reads and have since been adapted for other applications (Alipanahi et al., 2021). The initial pangenome release performed by the The Human Pangenome Reference Consortium (HPRC) includes 47 fully phased diploid genomes (94 haplotypes), with the ultimate goal of generating 350 diploid assemblies (700 haplotypes) (Liao et al., 2023). In addition to the Human Pangenome Consortium, other initiatives are creating population-specific pangenome references, such as the Arab Pangenome Reference (UPR) (Nassir et al., 2025) and the Chinese Pangenome Consortium (CPC) (Gao et al., 2023). Analyses using the human pangenome have shown an improved representation of human variation, improved read mapping, more accurate variant calling, and better genotyping between samples (Liao et al., 2023). Pangenomes are also particularly useful for studying CGRs, allowing the characterisation of different structural haplotypes, as exemplified by the analysis of the amylase locus (Bolognini et al., 2024), AGAP duplicons (Fornezza et al., 2024), the killer cell immunoglobulin-like receptor (KIR) gene complex (Hung et al., 2024), and the family of TBC1D3 genes (Guitart et al., 2024). However, additional haplotypes are required to fully capture the diversity in these CGRs and to accurately genotype the variants they contain (Ebert et al., 2021).

1.5 Human variation resources

Throughout the years, numerous techniques have been developed to detect and genotype genomic variants, with the overarching goal of representing human genetic variability. All of these methodologies have been used in large cohorts of populations,

including healthy and patients, creating resources for the identification of functional and evolutionary consequences for genetic variants.

1.5.1 Large scale initiatives to characterise human variation

In 2003, the International HapMap Project was launched to characterise common variations in the human genome DNA sequence and to make this information publicly available (Gibbs et al., 2003). This resource has since become invaluable for a wide range of studies, including candidate-gene analysis, linkage studies GWAS, and investigations of genetic factors influencing disease susceptibility or drug response. HapMap Phase I, which analysed 269 samples, aimed to genotype common SNPs (with allele frequencies greater than 5%) at an average spacing of one SNPs every 5 kilobases. In this phase, more than one million SNPs were successfully identified and genotyped (The International HapMap Consortium, 2005). Phase II, released in 2007, significantly expanded this effort, characterising more than 3.1 million SNPs in 270 individuals of four different ancestry. This corresponds to approximately one SNPs per kilobase, covering 25 to 35% of all common SNPs (allele frequency (AF) $>5\%$). Beyond cataloguing genetic variation, the HapMap Project has also provided valuable information on recombination rates and the LD structure of the human genome (Frazer et al., 2007). Finally, the last phase of the HapMap Project was published in 2010, featuring a dataset of 1,844 reference individuals from 11 global populations to study both SNPs and copy number polymorphisms (CNP). In this phase, the project focused on genotyping 1.6 million common SNPs using microarrays (Altshuler et al., 2010).

In 2010, the 1000GP was also launched, creating another valuable resource for the study of human variation. Its primary objective was to characterise all forms of human genetic variation, with the aim of capturing 95% variants with an AF of 1% or greater in accessible genomic regions, using the most advanced sequencing technologies available (Illumina short reads) at the time. In the pilot phase, whole genome sequencing (WGS) was carried out for 179 samples from five main population groups. Most samples were sequenced at low-coverage ($3.63\times$ per sample) and their

variant information was completed by imputation based on the haplotypes detected in a set of samples sequenced at high coverage. This phase identified more than 14 million SNPs, 1.3 million indels, and 20,000 large structural variants (SVs) (Durbin et al., 2010). Phase II, launched in 2012, expanded the analysis to 1,092 samples from 14 populations. This effort uncovered 38 million SNPs, 1.4 million indels, and 14,000 deletions (McVean et al., 2012). Phase III, released in 2015, carried out WGS in 2,504 individuals from 26 populations, identifying 88 million variants, most of which were SNPs, along with 3.6 million indels and more than 60,000 SVs (Auton et al., 2015). Specifically, SVs included 42,279 deletions, 6,025 duplications, 2,929 CNVs, 786 inversions, 168 nuclear mitochondrial insertions (NUMTs), and 16,631 insertions of mobile elements. Despite the challenges of detecting variants in repetitive or complex genomic regions, due to read depth limitations and sequencing technology constraints, the 1000GP reference panel has been widely used for the imputation of other variants (Sudmant et al., 2015). Finally, in 2022, 1000GP released an improvement in this resource, increasing the coverage from $7\times$ in Phase III to $30\times$ and adding 698 new samples to the original 2,504. As in previous phases, they identified SNPs (111 million), indels (14 million), and SVs (173,366), including 920 inversions and 3,568 complex rearrangements. The main advantage of 1000GP in all phases was the absence of sharing restrictions, providing full access to individual genotypes and enabling a more accurate analysis of variant effects (Byrska-Bishop et al., 2022).

The Human Genome Structural Variation Consortium (HGSVC), has emerged to continue the analysis of SVs in various human populations. Its primary objective is to characterise the full spectrum of SVs between different populations. The central approach of the consortium relies on assembly-based strategies to both discover and phase variants. In its first phase, the consortium analysed three trios with fathers previously described in the 1000GP Phase III, while children were used in the development of new human reference genomes (Chaisson et al., 2015). To maximise detection of SVs, the study used multi-scale mapping and multiple sequencing technologies to capture variants of different types and sizes. A combination of short reads of Illumina, PacBio, optical mapping, Hi-C, 10X Chromium, Strand-seq, and

ONT was used, resulting in increased sensitivity and combined coverage exceeding 223x. Then a comparison between the variants detected with short read and long read technologies was performed. Using short reads, 20,395 non-redundant SVs were identified in the three children, while long-read sequencing detected 48,635 non-redundant SVs, but calls only included deletions and insertion. When inversions were considered, a high-confidence callset of 308 inversions was identified among the nine individuals, corresponding to an average of 156 inversions per genome (Chaisson et al., 2019). A few years later Porubsky et al. (2021) designed a workflow to generate diploid *de novo* assemblies by combining Strand-seq data with PacBio CLR or HiFi long reads. Shortly thereafter, Ebert and colleagues, in the context of HGSVC2 assembled 70 haplotypes (64 from unrelated individuals and 6 from children) representing diverse populations. This approach enabled the detection of an average of 24,596 SVs per genome, creating a nonredundant data set of $\sim 108,000$ SVs. In the case of inversions, 117 events per sample were identified, resulting in a non-redundant call set of 316 inversions (Ebert et al., 2021). Finally, in 2025, the next stage was achieved with HGSVC3, which combined PacBio and ONT long reads, Strand-seq, optical mapping, and Hi-C to assemble 65 diploid human genomes from the 1000GP. The goal was to generate gapless chromosome assemblies, including centromeres and complex SDs. In this phase, a comparison of SVs was performed using T2T-CHM13 and GRCh38 as reference genomes. On average, 12,903 insertions and 12,505 deletions per genome were detected when using T2T-CHM13, compared to 17,458 insertions and 10,868 deletions per genome when using GRCh38. The excess of variants in GRCh38, particularly at high AF, was attributed to reference errors (Logsdon et al., 2025).

Some initiatives focused on the idea of deep characterisation and benchmarking of human genomes. This was the case of Genome in a Bottle Consortium (GIAB) (<https://www.nist.gov/programs-projects/genome-bottle>), which is hosted by the National Institute of Standards and Technology (NIST). Its work has focused on the deep sequencing and analysis of the NA12878 genome (HG001) from the HapMap project, as well as two trios of Ashkenazi Jewish and Han Chinese ancestry. Beyond sequencing, GIAB develops and integrates pipelines to generate high-confidence vari-

ant calls, including SVs (Zook et al., 2014, 2020). More recently, GIAB has expanded into cancer genomics by generating the first widely agreed open access tumour-normal resource, specifically designed to serve as a reference for the detection of somatic structural variants (McDaniel et al., 2024).

The success of projects such as 1000GP spurred the development of new initiatives designed to include larger cohorts and to address more specific research goals. One such example is the UK Biobank, a large prospective study launched in 2006 to build a comprehensive phenotypic resource and genotype dataset. It comprises 500,000 participants and provides data on 96 million variants (Young et al., 2016; Bycroft et al., 2018). In the United Kingdom, the 100,000 Genomes Project was established to investigate genetic variation in patients with rare diseases, with the goal of sequencing 100,000 samples from affected individuals (Samuel & Farsides, 2017). Another large-scale initiative is TOPMed (Trans-Omics for Precision Medicine), which focuses on improving the understanding of genetic risk factors for heart, lung, blood, and sleep disorders. By integrating WGS with other omics data in 80 participating studies, TOPMed has reported more than 410 million variants from 53,831 samples, enabling the detection of variants with frequencies as low as 0.01%. However, its scope remains limited to SNPs and small indels, excluding larger SVs (Taliun et al., 2021). Finally, the "All of us" research programme in the United States aims to sequence or genotype more than one million individuals. By 2023, the project had generated 245,000 whole genome sequences and more than 312,000 genotyping arrays, with nearly 50% of the participants representing ethnic minority groups (Ginsburg et al., 2023).

The heterogeneous samples, experimental protocols, and bioinformatic pipelines used by these and other initiatives generates a vast and often fragmented pool of variants. To address this, several programmes have been launched to collect, organise, and integrate predicted variants from multiple studies, thus creating more comprehensive and standardised catalogues of human structural variation. In this field, there are examples such as Database of Genomic Variants (DGV) (<http://dgv.tcag.ca>), which was one of the first resources designed to provide a comprehensive cata-

logue of structural variation in healthy individuals. It aggregates data from 75 studies, comprising more than 8 million entries, including 4,083 inversions. However, this resource has not been updated since 2019 (MacDonald et al., 2014). Another example is dbVar (<https://www.ncbi.nlm.nih.gov/dbvar>), maintained by National Center for Biotechnology Information (NCBI), which is a database of human genomic structural variation that includes pathogenic and benign variants (Lappalainen et al., 2013). Currently, dbVar contains more than 6 million variants from 185 studies (NCBI, 2025). Unlike DGV, dbVar remains active and continues to accept new submissions, with the expectation that its content will expand further in the coming years. DECIPHER (DatabasE of genom*i*C variation and Phenotype in Humans using Ensembl Resources) (<https://www.deciphergenomics.org>) is a resource focused on the identification and clinical interpretation of variants in patients, encompassing both benign and pathogenic cases. To date, it contains data from more than 50,000 uploaded cases (Foreman et al., 2023). Another example would be The Genome Aggregation Database (gnomAD) (<https://gnomad.broadinstitute.org>) that integrates sequencing data from a wide range of projects. The latest version (v4.1.0) includes data from 730,947 exomes and 76,215 whole genomes from unrelated individuals of diverse ancestry. Although gnomAD is best known for its catalogues of SNPs, it also includes SVs. Specifically, gnomAD v2 reported 445,857 SVs identified in 10,738 unrelated individuals (Collins et al., 2020; Gudmundsson et al., 2022).

1.5.2 Inversion data and catalogues

A closer look at structural variation detection over time reveals a progressive increase in the number of variants identified in the past decade. This trend has been driven in part by the growing number of samples studied and the higher coverage achieved with short-read sequencing, as demonstrated by the 1000GP (Byrska-Bishop et al., 2022), gnomAD (Collins et al., 2020), and population-specific studies such as the Dutch cohort (Hehir-Kwa et al., 2016). However, short-read technologies have inherent limitations, particularly for variants located within repetitive or structurally complex genomic regions. The next major step was the adoption of new sequencing strategies, single- or multiplatform approaches, that significantly improved discovery rates,

although initially in relatively small cohorts (Chaisson et al., 2019; Ebert et al., 2021; Porubsky et al., 2022). With decreasing sequencing costs, larger datasets are now being generated using these advanced technologies (Schloissnig et al., 2025). A clear result of these methodological advances is the steady increase not only in the total number of SVs detected, but also in the average number identified per genome. For example, phase II of HGSVC reported approximately 108,000 SVs in 32 samples (Ebert et al., 2021), compared to the 60,000 variants detected in 2,504 genomes during phase III of 1000GP (Auton et al., 2015). This trend is also evident in the case of inversions: 786 were reported in 1000GP phase III (Sudmant et al., 2015), 308 were identified in three trios in 2019 (Chaisson et al., 2019), and 729 were detected in 41 unrelated samples in 2022 (Porubsky et al., 2022). However, despite their power, multi-platform approaches remain relatively costly, which continues to limit the number of samples analysed in such studies. Despite the large number of inversions identified, they are spread over different studies and careful merging and quality control are required before creating a comprehensive catalogue that can serve as a resource for researchers.

The the INVersion Functional & Evolutionary Studies (InvFEST) project

As discussed previously, several studies and repositories report SVs; however, inversions are often under-represented or poorly characterised. To address this gap, InvFEST (<https://invfestdb.researchmar.net>) was developed to provide the most comprehensive catalogue of non-redundant polymorphic inversions in human populations. The database compiles inversion predictions from multiple studies carried out in healthy individuals. Since these predictions are generated using various methods and reported in different formats, online analytical processing (OLAP) is applied to harmonise the data and build a coherent catalogue. In addition, each prediction is assigned a status label based on its reliability, allowing users to identify high-confidence inversion calls. Over time, the content of InvFEST has grown from 1,092 candidates reported in 34 studies (85 of which had experimental validation) to 1,819 candidates in 49 studies. Within this pool, 252 inversions have been classified as validated, 412 as unreliable, 274 as false, and one as ambiguous. The classification

criteria are as follows: an inversion is considered validated when at least one breakpoint has experimental support; unreliable when no experimental validation exists and the candidate fails quality standards (for example, if more than 90% of its BPs overlap with a SD); false when experimental evidence disproves the candidate or is deemed incorrect; and ambiguous when results from two or more validation assays are inconsistent (Martínez-Fundichely et al., 2014).

In the early years of InvFEST, only a small set of candidates were genotyped using FISH (Antonacci et al., 2009), PCR (Pang et al., 2013), or iPCR (Aguado et al., 2014; Giner-Delgado et al., 2019). However, the development of inverse MLPA iMLPA (Giner-Delgado et al., 2019) and the application of digital droplet PCR (ddPCR) (Puig et al., 2020) assays allowed the genotyping of inversions that had previously remained inaccessible, particularly those with large IRs at the BPs. Using iMLPA in combination with PCR, iPCR, and MLPA, it was possible to genotype 45 paracentric inversions in 551 samples from seven populations included in large-scale projects such as HapMap and 1000GP. These inversions ranged in size from 83 bp to 415 kb and included 24 variants generated by the NH mechanisms and 21 by the NAHR mechanisms with IRs shorter than 30 kb (Giner-Delgado et al., 2019). At this stage of the project, this set represented the majority of variants experimentally verified using PCR-based techniques, accounting for about half of the predicted inversions segregating at a frequency greater than 5% in human populations. Similarly, the use of ddPCR allowed the validation and genotyping of 19 additional inversions in 95 samples of various ancestry drawn from the panel mentioned above of 511 individuals. All inversions in this dataset were generated by NAHR mechanisms, with repeat sizes of up to 134 kb and overall lengths of 3.1 kb to 741.7 kb (Puig et al., 2020).

These studies revealed important differences between NH and NAHR inversions. NH inversions, typically generated by a single mutational event, usually present perfect tag SNPs. In contrast, NAHR inversions often present recurrence, with inversions being generated independently multiple times on different haplotypes, resulting in a low LD with neighbouring variants and a high number of SNPs shared between

both orientations. In addition, recurrence was shown to depend on the proximity and length of IRs, with repeats located closer to each other and longer being more prone to recombination (Giner-Delgado et al., 2019). Inversion recurrence has been confirmed in subsequent studies (Porubsky et al., 2022) and has also been observed in other primate species (Giner-Delgado et al., 2019) and throughout mammalian evolution (Cáceres et al., 2007). In some studies, recurrence rates were estimated, yielding inversion rates ranging from 3.4×10^{-6} to 2.7×10^{-4} per site per generation, which is $\sim 10^3$ times higher than that of single-base mutations (Giner-Delgado et al., 2019; Porubsky et al., 2022). Due to this recurrence, the imputation of NAHR inversions from SNPs is generally unreliable, representing a key limitation for functional analyses. However, there are some exceptions, such as HsInv1057 and the well-characterised 17q21.31 inversion, where a single origin mediated by IRs allows imputation using tag SNPs (Puig et al., 2020; Campoy et al., 2022). Inversions have been proposed to have a maximum tolerated length with respect to fertility. Below this threshold, inversions are likely neutral, consistent with the observed inverse relationship between inversion frequency and size (Giner-Delgado et al., 2019; Puig et al., 2020). In contrast, larger inversions may persist only if they confer advantageous functional effects under positive selection, compensating for their negative impact on fertility (Giner-Delgado et al., 2019). A paradigmatic example is the 17q21.31 inversion, which exhibits strong functional consequences and evidence of positive selection in European populations, likely compensating for its deleterious effects on fertility (Puig et al., 2020). These studies also discovered multiple functional effects on both NH and NAHR inversions. In addition to 17q21.31, examples include HsInv0786, HsInv0124, and HsInv1051, some of which act as eQTL for gene expression, disrupt exonic sequences, or show enrichment in GWAS signals. Several additional candidates remain promising for further functional investigation (Giner-Delgado et al., 2019; Puig et al., 2020).

1.6 Mosaicism

Up to this point, we have discussed genetic variation primarily in the context of population polymorphisms that were originally generated in germline cells. However, variants can also occur in any cell within a single individual, and then only a subset of cells will carry the change. This phenomenon is known as mosaicism, defined as the presence of two or more distinct genomic haplotypes in an individual derived from a single zygote. Another far more rare mechanism that generates genetically distinct cell populations is chimerism, which arises from the fusion of lineages originating from different zygotes (King et al., 2015). Mosaicism emerges through *de novo* mutations that can occur during any cell division, at any stage of development, and in any tissue, which means that, in principle, all humans are mosaics. When postzygotic mutations affect only somatic cells, the condition is termed somatic mosaicism (Thorpe et al., 2020). In contrast, when mutations occur in germline precursor cells, they give rise to germline mosaicism, which can be transmitted to the offspring and thus become non-mosaic in the next generation (Truty et al., 2023). Somatic mosaicism is generally more widespread than germline mosaicism, and its frequency varies between tissues and developmental stages, with higher mutation rates during early embryogenesis and neurogenesis compared to adulthood (Spinner & Conlin, 2014). Most mosaic variants are benign; however, they can also be pathogenic, particularly when they reach high clonality (affecting a substantial proportion of cells) or involve genes associated with Mendelian disorders, thus influencing disease onset, progression, or inheritance (King et al., 2015; Thorpe et al., 2020; Truty et al., 2023). The functional impact depends not only on the clonality but also on the timing of the mutation and its genomic context. Variants in coding regions, especially nonsense or missense substitutions, are more likely to be deleterious than synonymous changes or mutations in non-coding regions, which are often less constrained (Sharp et al., 1995; Rech et al., 2014).

Somatic mutations arise from a combination of endogenous processes and environmental exposures, accumulating with age. Age-related mosaicism has been extensively documented, including both mosaic single nucleotide variants (mSNVs) and

mosaic structural variants (mSVs) (Machiela, 2019; Truty et al., 2023). However, due to mSVs mediated by repetitive elements, these motifs exhibit higher generation rates are expected to be more frequent. Germline mosaicism also increases with parental age (Jónsson et al., 2018). Somatic mosaicism contributes to late-onset disorders such as Alzheimer’s and Parkinson’s disease (Lodato & Walsh, 2019; Thorpe et al., 2020), and has been observed in genes underlying X-linked and autosomal disorders, such as *PCDH19* and *VHL*, with phenotypes ranging from cancer predisposition and neurological conditions to developmental syndromes. Mosaic SVs, including copy number changes, have been shown to be particularly relevant for developmental disorders. However, their observed frequency and size distribution depend on some factors, such as the sensitivity of the detection methods and the fitness consequences of the variants. For example, large CNVs are often deleterious, making them easier to detect compared to variants that have more subtle effects (Thorpe et al., 2020). Mosaicism has been especially well studied in hereditary cancer and paediatric overgrowth syndromes (Truty et al., 2023).

Detecting mosaic variants remains challenging because they are often present at low frequencies and restricted to certain cell populations. The early methods relied on molecular cytogenetics. Conventional karyotyping, the reference standard for chromosomal mosaicism, can identify large structural variants but lacks resolution for smaller alterations (Wan & Ma, 2012). FISH improved sensitivity, allowing the detection of smaller variants and analysis in interphase cells, which has been shown to be valuable in cancer cytogenetics (Thorpe et al., 2020). SNPs arrays further advanced detection by quantifying allele frequencies and copy number changes, although they typically miss variants below the 5% frequency and exclude repetitive or low-complexity regions (King et al., 2017). NGS expanded the discovery of mosaic variants by allowing the comparison of multiple tissues of the same individual and using trio-based designs to distinguish inherited mosaic events. Nevertheless, bulk NGS has limitations: false positives from molecular or computational errors, poor sensitivity in low-frequency variants, and inability to resolve mosaics specific to cell types (Biesecker & Spinner, 2013; Grimes et al., 2024). Single-cell sequencing (SCS) offers a powerful alternative by isolating, amplifying, and sequencing individual cells.

Although whole-genome amplification introduces artefacts, these can be mitigated by validation with targeted bulk sequencing or by analysing multiple single cells and retaining only recurrent variants. Long-read single-molecule platforms such as PacBio and ONT further improve detection, eliminating the need for amplification and facilitating variant phasing and analysis in repetitive regions. However, their higher error rates remain a challenge (Thorpe et al., 2020). Moreover, Strand-seq has become a valuable tool for studying inversions, due to its single-cell sequencing design. Dedicated software has also been developed to support its analysis (Porubsky, Sanders, Taudt, et al., 2020; Hanlon et al., 2023; Weber et al., 2023). Using this technology, Grimes and colleagues (2024) investigated the impact of large-scale mosaicism in normal human pluripotent stem cells (hPSCs), reporting that 16 out of 19 donors carried mSVs and identifying 51 independent variants (Grimes et al., 2024). Despite these technological advances, mosaicism and its impact on human health remain underexplored, particularly in complex diseases where low-frequency variants may play important roles.

1.7 Complex genomic rearrangements (CGRs)

Throughout the text, we have focused primarily on isolated SVs, which correspond to cases in which only a structural variant is found at a specific locus. However, at the beginning, we introduced the concept of CGRs, defined as regions harbouring multiple SVs in cis (Schuy et al., 2022). More recently, the term complex structural variants (CSVs) has been used to describe events composed of simple SVs spanning more than one repair junction (Logsdon et al., 2025). Both concepts capture a similar idea: loci where multiple variants coexist on the same chromosome. Typically, the BPs of these structures are located within repetitive sequences such as SDs or mobile element insertions (MEIs), and are generated by NAHR mechanisms (Collins et al., 2020; Logsdon et al., 2025). As noted earlier, the presence of these repetitive sequences makes these loci prone to additional rearrangements, resulting in highly diverse regions. For the remainder of the text, we define a “complex region” as a locus where the combination of simple SVs generates CSVs or CGRs in different

samples, producing a population-level pattern of more than two haplotypes.

Complex regions encompass large genomic segments, often reaching megabase scales, and typically exhibit intricate patterns of CNVs and balanced variants such as inversions (Ebert et al., 2021). These regions, with high diversity and the number of haplotypes in the population, remain a challenge for most technologies. Even today, the use of multiplatform approaches has limited power in resolving them. As estimated by Ebert and colleagues, approximately 35 regions greater than 50 kbp and containing five or more distinct SVs haplotypes remain unresolved per phased assembly (Ebert et al., 2021).

At the haplotype level, large-scale studies such as gnomAD have reported that 1.6% of SVs may actually represent CSVs or CGRs. With improvements in detection methods, this proportion has increased from 2.5% (Collins et al., 2020) to 3.3% (Abel et al., 2020), suggesting that the true number of complex rearrangements and their combinations between samples may be higher than previously estimated. Third-generation sequencing technologies now enable the detection of such rearrangements and the resolution of regions enriched in SDs, centromeres, and telomeres (Schuy et al., 2022). The most recent studies using multiplatform approaches identified 5,122 CSVs, which were merged into 1,247 unique variants in 130 phased haplotypes, corresponding to an average of 72 CSVs per sample. These included duplications, inversions, and complex structures such as INVDUP-INV-DEL, some consistent with previous findings. The mean size of these variants was 91.5 kbp when mapped to the reference genome T2T-CHM13 (Audano et al., 2024; Logsdon et al., 2025). The use of human pangenomes, which integrate multiple sequencing technologies and construct de Bruijn graphs to represent variability, has already been applied to the characterisation of complex regions such as the major histocompatibility complex (MHC) class II locus and the amylase locus (Bolognini et al., 2024). However, properly capturing this variability requires dedicated tools, such as PGR-TK, which provide a set of tools that were already used in the visualisation of complex variants (Chin et al., 2023).

Classical examples of complex regions in the human genome include the 17q21.31

locus mentioned above, the amylase gene cluster, or the MHC and KIR genomic regions. Owing to their high haplotype variability, some studies have focused on characterising single regions in detail. For example, the amylase locus is among the most studied, with between 28 and 30 distinct haplotypes identified to date (Bolognini et al., 2024; Yilmaz et al., 2024). The MHC is one of the most complex regions of the genome, with estimates of haplotype diversity ranging from several dozen to thousands of haplotypes; however, no consensus has yet been reached, as different studies use different strategies to infer variation (Norman et al., 2017; Kulski et al., 2022; Scigliuolo et al., 2023; Mosbrugger et al., 2025). Other studies have investigated regions where human haplotypes differ markedly from those of other mammals, particularly in loci enriched for SDs, such as 15q11–q13 (Paparella et al., 2023). Some regions show extreme haplotypic complexity or very large sizes, whereas others display more limited variation. For example, in the 3q29 region, 18 structural haplotypes have been identified, involving an inversion polymorphism, making this region especially interesting due to its association with a microduplication syndrome (Ebert et al., 2021). Similarly, the locus containing the *NOTCH2NL* and *NBPF* genes shows three structural haplotypes: the reference haplotype (13.7% of alleles); a variant in which *NBPF8* is inverted and *NOTCH2NLR* and *NBPF26* are deleted (35.39%) (Figure 1.7); and a haplotype in which *NBPF8* is replaced by *NBPF9* (50.8%) (Logsdon et al., 2025). In addition, ten additional structural conformations associated with *NOTCH2NL* were also identified (Real et al., 2025).

Despite progress in analysing several of these loci, many complex regions remain uncharacterised. It is important to note that state-of-the-art methodologies are costly as they require combining multiple sequencing technologies. In addition, when using graph-based references such as the pangenome, the number of samples that can be analysed remains limited, making it difficult to draw definitive conclusions about the functional and evolutionary functions of these regions, which often encompass a substantial number of genes due to their size.

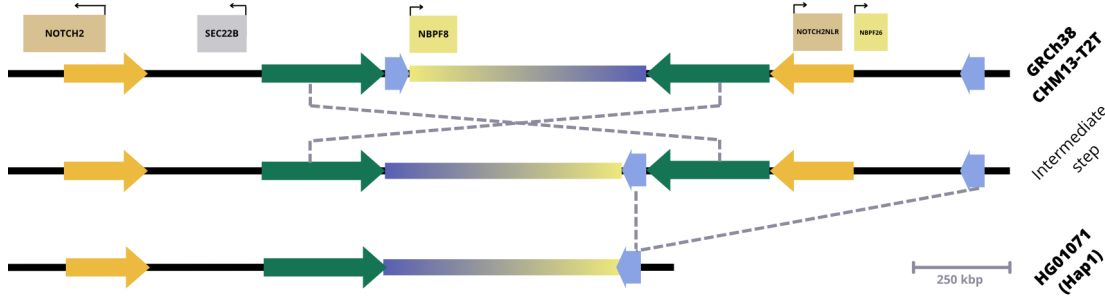


Figure 1.7: **Haplotype structures affecting *NOTCH2NL* and *NBPF* genes.** The region affecting *NOTCH2NL* and *NBPF* genes spans chr1:119,900,000–121,490,000 (1.59 Mb) and contains multiple annotated genes according to GRCh38. Within this locus, several SDs (same-colour arrows) and unique sequence blocks are present, with some genes embedded in these repetitive segments. Reference genome structures (GRCh38 and T2T-CHM13) and haplotype 1 of HG01071 (Logsdon et al., 2025) are shown, highlighting the *NBPF8* inversion and a deletion of *NOTCH2NLR* and *NBPF26*. An intermediate step, potentially present in other samples, is also depicted.

1.8 Objectives

The main objective of this thesis is to characterise and genotype human polymorphic inversions, with a particular emphasis on those generated through NAHR mechanisms, and their associated complex regions. By assembling a comprehensive dataset of inversions that have so far remained insufficiently explored, this work seeks to provide the basis for an in-depth evaluation of their functional consequences and evolutionary significance in the human genome. The specific objectives are outlined below:

- 1) **Develop novel high-throughput genotyping assays for repeat-mediated inversions that cannot be accurately genotyped with other techniques.** New techniques have emerged for the detection of SVs, all of which present advantages and limitations. However, only a limited number of methods are available for a reliable characterisation of SVs at the population level. This problem is particularly relevant in the case of inversions mediated by IRs, which are challenging to genotype due to their intrinsic characteristics. For this reason, the creation and evaluation of a tool for inversion genotyping

is essential, especially one that uses ONT long-read sequencing, which makes possible to span large repeats.

- 2) **Validate and genotype inversion candidates to generate the most complete catalogue of human polymorphic inversions.** Over the years, numerous candidate inversions have been reported, both in large-scale studies such as the 1000GP and gnomAD, and more recently through the combination of novel genomic technologies. These candidates need to be systematically compiled, reviewed, and organised to obtain a curated set of independent variants. Subsequently, a thorough characterisation is required to validate these variants. This information will enable the update of InvFEST, with the aim of establishing the most comprehensive catalogue of human inversions to date, providing a key reference resource for the field.
- 3) **Genotype human inversions in a large set of samples.** To date, most studies that employ new technologies for the detection and genotyping of genomic inversions have relied on relatively small sample sets, which limits the robustness of the conclusions drawn. Thus, taking advantage of the increasing availability of ONT sequencing data, we aim to compile samples from various sources to construct the largest ONT-based dataset for inversion genotyping. This resource will enable a more accurate characterisation of the population frequency and distribution of the whole set of studied inversions, constituting a crucial first step towards more comprehensive functional and evolutionary analyses in the future.
- 4) **Develop a new method for structural haplotype resolution and accurate characterisation of highly variable genomic regions.** Some repeat-rich genomic regions harbour multiple SVs and exhibit unexpectedly high variability among individuals, which complicates their characterisation. Although this diversity can be manually assessed in a limited number of samples, such an approach is not scalable to larger cohorts. To overcome these challenges, it is necessary to develop new tools capable of resolving haplotypes in complex genomic regions without relying exclusively on genome assembly and which

can take advantage of available long-read data.

2. Materials and methods

2.1 Datasets

2.1.1 Inversions

One of the main goals of this thesis is to create a comprehensive catalogue of inversions. To achieve this, we compiled a list of inversion calls and other SVs involving inverted segments, such as inverted duplications and complex rearrangements, from multiple studies. Even though our object of study are inversions, inverted duplications were included because structural variant callers can sometimes confuse inversions and inverted duplications and make the wrong call. These include the large-scale studies mentioned in the introduction, as well as recent investigations that leverage emerging technologies, which are improving our ability to detect and genotype SVs. For each study, we provide a brief description and, finally, we summarise all identified variants in a table that can serve as a reference guide. Collectively, these 13 studies encompass 20,920 variants that include inversions, inverted duplications, and complex rearrangements that involve inversions. However, some of the variants could overlap between the different studies, so to avoid redundancies, multiple transformations will be performed to correctly merge the variants. This catalogue will be refined to capture as many human polymorphic inversions as possible, which will then be genotyped. The resulting information and derived statistics will serve as the basis for creating an exhaustive and publicly accessible catalogue that can act as a resource for other researchers in the field. To simplify the explanation, studies are organised into three groups: the INVersion Functional & Evolutionary

STudies (InvFEST) and other studies in the group; large-scale projects; and studies focused on the use of specific technologies. Here, we present additional information about each group of studies and include a summary table for each study (Table 2.1).

- **InvFEST project:** This is a comprehensive catalogue of non-redundant polymorphic inversions in human populations, integrating predictions from 49 studies conducted between 1997 and 2018 (Martínez-Fundichely et al., 2014). These predictions were derived from various technologies, including short-read sequencing and paired-end sequencing and mapping (PEM). The different detected variants were harmonised using automatic tools to ensure consistency. The catalogue currently includes 1,819 variants, 252 of which are experimentally validated using multiple polymerase chain reaction (PCR)-based assays (Giner-Delgado et al., 2019; Puig et al., 2020). As part of the InvFEST project, we included the study of (Campoy, 2024), which reported a total of 197 SVs, including 134 inversions, 61 inverted duplications and two associated deletions. These variants were genotyped both experimentally, incorporating the validation assays previously described in InvFEST, or *in silico* using BreakSeq (Lledó & Cáceres, 2013; Puig, Castellano, et al., 2015; Giner-Delgado et al., 2019). The genotyping effort covered between 95 samples for some variants and up to 3,202 samples for others, all from the 1000 Genomes Project (1000GP). In addition to variant genotyping, breakpoints (BPs) of these SVs were also manually curated.
- **Large-scale studies:** This group includes two major population-scale projects: 1000GP and The Genome Aggregation Database (gnomAD). Both used short-read sequencing technologies that aimed primarily to identify single nucleotide polymorphisms (SNPs), although SVs were also detected. For this analysis, we used the high-coverage dataset from 1000GP (Byrsk-Bishop et al., 2022), and gnomAD SVs v4.1.0 (Collins et al., 2020).
- **New technologies studies:** This group comprises the majority of studies (eight in total), all employing emerging technologies such as Bionano optical mapping, Single-cell template strand sequencing (Strand-seq), Pacific Biosciences

(PacBio), or Oxford Nanopore Technologies (ONT) long-read sequencing. These studies are heterogeneous, ranging from single- to multi-platform approaches and involving datasets from only a few samples to as many as 1,019. Several of them are part of The Human Genome Structural Variation Consortium (HGSVC), reflecting the ongoing efforts to develop and refine tools for the detection and genotyping of SVs.

Table 2.1: Information of analysed studies. A comprehensive list of all studies contributing to the identification of the variants analysed in this work, which will form part of the new variant catalogue. The listed studies include those reporting inversions, inverted duplications, and other complex genomic rearrangements. For each study, details should be provided on the total number of variants detected, the number of samples analysed, and the sequencing or analytical technologies employed for variant discovery.

Study	Variants	Samples	Technologies
1000GP	2986	3202	WGS Illumina
gnomAD	13064	63046	WGS Illumina
InvFEST	1819	–	Multiple technologies
Audano et al. (2019)	222	15	PacBio long reads
Chaisson et al. (2019)	443	9	Multi-platform
Levy-Sakin et al. (2019)	380	156	Optical Mapping
Ebert et al. (2021)	316	34	Multi-platform
Porubsky et al. (2022)	399	44	Multi-platform
Porubsky et al. (2023)	63	44	Multi-platform
Campoy (2024)	197	95-3202	Short-reads; PCR-based
Logsdon et al. (2025)	298	65	Multi-platform
Schloissnig et al. (2025)	733	1019	ONT long reads

In this study, inversions are referred to using the notation "HsInv" as defined in the InvFEST database (<https://invfestdb.researchmar.net>). Any inversion not previously catalogued in InvFEST is assigned the temporary notation "HsCan" for the purposes of this thesis. Due to the size of the dataset, this paper does not include a comprehensive table containing the coordinates of all inversions. However, the coordinates defined for each inversion are available from the author upon request.

2.1.2 Samples

To perform our analyses, it was necessary to collect available samples. When this project began, the number of publicly available datasets generated using ONT long

reads was very limited because it was still a relatively new technology with applications restricted to specific research areas. However, this has changed substantially over the years, allowing us to collect a significant number of samples. In this section, we describe the sources of these samples and summarise their main characteristics.

In total, we found 1,409 unique samples, which come from different sources listed in Table 2.2. Beyond the information provided in the table, we outline here the origin and relevant context of each dataset.

Table 2.2: **Sample sources.** Summary table describing the origin of each sample, specifying the original publication when available, and otherwise the generating consortium which the sample was obtained. Sample counts are reported per study; samples included in multiple studies are counted separately in each study.

Source	N ^o Samples
HGSVC	64
The Human Pangenome Reference Consortium (HPRC)	294
Gustafson et al. (2024)	100
Schloissnig et al. (2025)	1019
Sequenced by us	30

Some of our samples come from emerging consortia dedicated to studying human genomic variability, such as HGSVC and HPRC. For HPRC, our dataset includes not only samples from the first Pangenome release, but also those now in development for the forthcoming second release. Regarding Gustafson et al. (2024), the samples represent the first ONT-based reanalysis of individuals from the 1000GP using long and ultra-long reads. For Schloissnig et al. (2025), their study also includes samples from the 1000GP, and this project involved direct collaboration, which will be commented on in section 3.2. Additionally, during the course of this thesis, our group sequenced several samples. These were used to evaluate the performance of ONT adaptive sampling compared to traditional WGS. Part of this work is presented in the first article in the section 3.1.

It is important to note the following: although the dataset contains 1,409 samples, we only performed read mapping for 476 of them. Samples appearing exclusively in Schloissnig et al. (2025) were excluded from our mapping because that study already

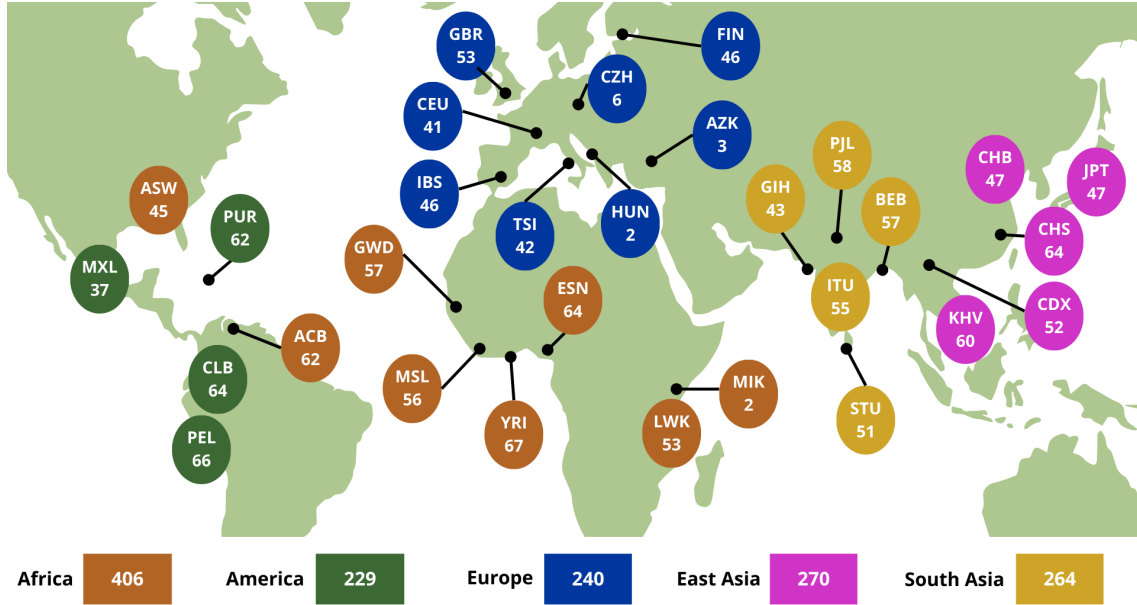


Figure 2.1: **Sample population assignment.** Itemisation of self-identified geographical ancestries for 1409 samples representing 30 populations from 5 continental regions. The three-letter codes used are equivalent to those used in the 1000GP phase III.

provided its own alignments. For samples shared across datasets, if we generated our own alignments, these were merged to increase depth. However, in cases where our group and Schloissnig et al. (2025) produced mappings for the same sample, we kept only our own, as they were generated using our own specific mapping and filtering criteria.

Of all samples, 677 correspond to male individuals and 732 to female individuals, T2T-CHM13 is included in the latter group, as it lacks a Y chromosome. At the population level, the dataset includes individuals representing five ancestries, as shown in Figure 2.1

Within these samples, we identified a total of 114 samples belonging to trios, corresponding to 38 families. Additionally, when considering duos, defined as an offspring and one of their parents, we found 164 samples, representing 82 duos. These related samples were used primarily to detect potential genotyping errors.

Among the 1,409 samples, 69 did not belong to 1000GP, which limited our ability to perform certain SNPs analyses exactly as originally designed. Finally, we also as-

sessed how many samples had experimental validation available for specific variants, which allowed comparison with the calculated genotypes; this subset comprised 231 samples.

Together, these observations summarise some of the key characteristics of our sample dataset. Additional characteristics, whether related to the samples themselves or to specific datasets, will be introduced throughout the Results section as required by each analysis.

2.2 Methods

2.2.1 Merging of variants

To create the most comprehensive catalogue of genomic inversions, we merged the variants reported in the different studies described above. The first step consisted of homogenising the information available for each variant. This harmonisation was achieved by integrating data from all studies into a common set of fields shared between datasets. These fields included the chromosome, four coordinates that define the two BPs of the variant, the type of variant, the frequency and the samples in which the variant was identified. Each breakpoint is defined by two coordinates that correspond to the start and end positions of inverted repeats, and, if they are not present, to the external limits of any elements found at the BPs (like other associated SVs as deletions, microhomologies, etc.) or to the two genomic positions flanking a clean break.

For several studies, additional processing steps were required to obtain these fields. Most studies reported only a single set of coordinates per variant, so a confidence interval of ± 5 kb was introduced around each defined breakpoint. Given the uncertainty associated with the breakpoint resolution of different techniques, this value was selected as a conservative estimation.

In some cases, it was necessary to convert the reported coordinates from a previous version into the GRCh38 reference genome. The coordinate conversion was

performed using LiftOver (Hinrichs et al., 2006), applying the appropriate chain file for each conversion. Both external coordinates (corresponding to the first coordinate of breakpoint 1 and the second coordinate of breakpoint 2) and internal coordinates (corresponding to the second coordinate of breakpoint 1 and the first coordinate of breakpoint 2) were converted. During the LiftOver process, multiple hits were allowed, and the minimum remapping fraction was set to 0.01. Once the candidate hits were obtained, they were evaluated to determine which conversions were acceptable. This evaluation was based on the successful conversion of six specific positions: four single-base coordinates (BP1-1, BP1-2, BP2-1, BP2-2) and the two defined BPs (BP1 and BP2). First, we verified that the converted coordinates matched exactly the point conversion values. Second, we assessed whether the size of the converted region differed by less than 25% from the original region. If this threshold was exceeded, only conversions with an absolute size difference of less than 10 kb were accepted. Variants that did not meet any of these criteria were excluded from further analysis.

For some studies, allele frequencies (AFs) were not provided directly and therefore had to be calculated. When genotype information was reported as presence or absence, the frequency was calculated as the percentage of individuals carrying the variant. When complete genotypes were available, AFs were calculated without accounting for relatedness, since in many cases the sample sizes underlying the reported frequencies were extremely small.

Following these steps, a unified table was generated that contains all relevant information for each variant of each study. Then we performed an initial merge of the variants within each study to avoid redundancies. Two variants were considered identical if their two breakpoint intervals overlapped with the corresponding BPs of another variant. When variants were merged, their breakpoint intervals were updated by taking the minimum and maximum coordinates, thereby progressively expanding the breakpoint ranges.

To validate this merging process, we verified which variants should be merged using bedtools pairtopair (Quinlan & Hall, 2010), with the option "both" (requiring both

ends of feature A to overlap feature B) and a "slop" value of 0 to avoid extending the overlap region. After completion of this within-study merge, the same procedure was repeated to merge variants from different studies.

Variant filtering

After merging, a filtering step was applied to the resulting set of variants. The first exclusion criterion was AF: only variants with a frequency of at least 1% were retained. The second criterion was the size of the variant. Variants larger than 10 Mb were excluded, as inversions of this size are highly unlikely to present a high frequency and are likely artefacts generated by variant callers (MacDonald et al., 2014; Cuenca-Guardiola et al., 2023).

Breakpoint definition

To define inversion BPs, we first assessed whether the two breakpoint intervals overlapped with segmental duplications (SDs). If a variant overlapped with a single isolated pair of SDs, the inversion BPs was extended to include the entire repeat. If the breakpoint interval overlapped multiple SDs, all were combined to form a larger breakpoint. So, for all variants sharing the same overlapping SDs, the same BPs were defined; however, they did not merge. To know how many regions share this same large BPs, they were grouped in regions that were termed complex, since the presence of multiple pairs of SDs provides a favourable genomic context for the generation of multiple rearrangements. When these complex regions were defined, we evaluated how many additional variants overlapped the same regions, as some variant calls could represent alternative haplotypes generated through interactions among multiple SDs.

All variants overlapping with one complex region were assigned the same BPs, with breakpoint intervals overlapping multiple SDs merged into a single extended breakpoint region. As a result, variants whose breakpoint intervals overlapped the same set of SDs were assigned identical extended BPs. These variants were not merged; rather, they were retained as independent calls sharing the same breakpoint defi-

dition. Then these variants were compared with each other to identify groups of variants that share the same extended BPs. Such regions were termed complex regions, since the presence of multiple pairs of SDs constitutes a genomic context prone to the generation of multiple potentially related rearrangements.

Once complex regions were defined, additional variants for which BPs had not yet been assigned were evaluated for overlap with these regions. When a variant overlapped a given complex region, it was assigned the same extended BPs as that complex region. This approach allowed for the consistent definition of breakpoint coordinates between variants that could represent alternative haplotypes generated by interactions among multiple SDs.

Next, we examine variants classified as complex in gnomAD, as the corresponding VCF files provide explicit coordinates for variants located near the inversion boundaries, such as deletions or duplications. In these cases, the gnomAD coordinates were used as BPs. When such variants were merged with others, we verified that the single-base breakpoint defined in the original study was consistent with the gnomAD coordinates. In cases where a variant was not consistent, the breakpoint was extended to encompass the variant coordinates.

For the remaining variants (including both simple inversions and inversions associated with other variants), all merged individual calls were examined and the minimum and maximum coordinates of each set were used to define their BPs, which were then extended by 100 bp on each side.

Finally, an additional filtering step was applied to the final list of merged variants defined BPs to exclude from the analysis those variants with an internal size smaller than 300 bp, those with two BPs larger than 200 kb, or those corresponding to inverted duplications.

Inversion genotyping pipeline

The following steps are integrated into the pipeline designed for inversion genotyping, termed GeONTIpe (<https://github.com/RMoreiraP/GeONTIpe>). For clarity,

a brief overview of the pipeline and its main components is provided here. A more detailed description of the developed tool can be found in Section 3.1, within the Materials and Methods section.

GeONTIpe is a tool developed for the genotyping of genomic inversions. The initial stages of the pipeline include the download and subsequent analysis of sequencing reads, followed by a filtering step. Specifically, reads shorter than 5 kb and with a quality score below 7 are removed. The next step corresponds to the read mapping, which includes optimisation of the parameters to improve the detection of inversions and other SVs. Additionally, only alignments with a mapping quality (MAPQ) greater than 20 are retained.

When FASTQ files are available for the samples, they are prioritised, due to the control of mapping criteria. However, precomputed mapping files can be used directly, as was the case for samples obtained from the 1000GP FTP repository associated with the study by (Schloissnig et al., 2025). For overlapping samples, the mappings generated by GeONTIpe were prioritised and mapping files were not merged, as mapping criteria differ between studies.

Once the samples are mapped, the inversion genotyping is performed. Based on the coordinates of the BPs defined for each variant, four probes are designed: two internal probes and two external probes (one upstream of breakpoint 1 and one downstream of breakpoint 2). A minimum of three out of the four probes is required for a candidate inversion to be included in the analysis.

These probes are the elements used to detect the orientation of the variant of interest. GeONTIpe identifies the reads that span the defined BPs and places the designed probes on each read using BLAST (Boratyn et al., 2013). Subsequently, the probe position and orientation within each read are analysed to assign a genotype at the read level. By aggregating the information across all supporting reads, the genotype of the sample is defined. Throughout the text, Std (standard) will be used to denote the orientation present in the GRCh38 reference genome; alternative orientations will be specified explicitly when applicable. The criteria used for genotype determination

are described in the mentioned section. In addition, the distances between probes are used to detect the presence of additional unbalanced variants by comparing the expected and observed inter-probe distances. This approach allows for not only the determination of the genotype for a given sample and inversion but also the identification of additional structural variants.

Probe design and genotyping

First, based on the parameters provided to GeONTIpe, a 20 kb region surrounding the defined BPs for each inversion was selected. Probes of 300 bp were designed with an overlap of 275 bp, maximising the number of probes generated. In this step, probes were not filtered according to repetitive element content (0% unique sequence requirement). Designed probes were required to be unique within a 500 kb window upstream and downstream of each inversion region. Among the probes that met these criteria, those closest to the defined BPs were selected for inversion genotyping.

These initial probes were tested to assess their ability to correctly identify both alleles in a test sample set that included at least one inverted allele for each candidate inversion region. Genotypes classified as erroneous or ambiguous were manually reviewed using dot plot analyses. Variants deemed structurally invalid were excluded from further analysis, while variants with a coherent structural interpretation were retained. For retained variants, a second set of probes was designed to improve genotyping by increasing the minimum unique sequence requirement to 25%.

The final genotyping with GeONTIpe was performed using the following internal parameters for megablast: a minimum probe coverage of 90% and a minimum sequence identity of 85%.

2.2.2 Genotyping performance evaluation

After genotyping, an evaluation of the resulting genotypes was performed. This assessment was based on the proportion of valid genotypes obtained (excluding undetermined calls and low-confidence) relative to the total number of variants or sam-

ples for which genotyping was attempted. In addition, genotyping performance was compared between studies by grouping samples according to their study of origin. For samples shared by more than one study, each sample was counted independently for each study to which it belonged.

To assess the effect of sequencing quality metrics on genotyping performance, the N50 and the total number of reads were extracted from FASTQ data using NanoPlot v.1.40.2 (De Coster & Rademakers, 2023). The reported values corresponding to each sample were obtained after the GeONTIpe read filtering step. To evaluate the effect of breakpoint size on performance, the minimum distance between probes was used, under the assumption that the smallest inter-probe distance would be the main factor that determines the number of informative reads spanning that region.

2.2.3 AF calculation

AFs were calculated excluding offspring from trios and duos. The T2T-CHM13 sample was treated as carrying a single allele and partial genotypes in which only one allele was determined were also counted as single alleles. The same counting criteria were applied when computing AF within each ancestry. Because certain loci in GRCh38 represent minor inverted configurations relative to population frequencies, and given the absence of reliable ancestral state information for all inversions, frequencies were consistently reported as MAF to avoid reference-dependent polarisation biases. When reporting AF within each ancestry, the frequencies were referenced to the allele defined as minor in the global dataset.

2.2.4 Population differentiation

Fixation index (F_{ST}) to evaluate differences in AF between populations was calculated both globally and in pairwise comparisons between ancestries. The F_{ST} values obtained for each inversion were compared with an empirical null distribution derived from SNPs from the 1000GP high-coverage dataset. To construct this null distribution, SNPs located in easy-to-sequence regions, as defined by the Genome in a Bottle (GIAB) Consortium (Krusche et al., 2019), were selected. The AFs for

both inversions and SNPs were calculated excluding related individuals, and variants were grouped into frequency bins ranging from 0 to 0.5 with increments of 0.01.

For each inversion, SNPs belonging to the same frequency bin and located on the same chromosome type (autosome or chromosome X) were randomly sampled. A total of 10,000 SNPs were selected for autosomes and 1,000 SNPs for chromosome X, reflecting the distinct population genetic properties of chromosome X relative to autosomes (Holsinger & Weir, 2009). The F_{ST} values were then computed globally and pairwise between ancestries for inversions and SNPs using *vcftools* (v0.1.16) with the *weir-fst-pop* option (Danecek et al., 2011). Negative F_{ST} values were set to zero, as F_{ST} is constrained to the $[0,1]$ interval. Inversion F_{ST} values were considered significant when they fell within the top or bottom 5% of the SNP-based null distribution, corresponding to an empirical significance level of 90%. In addition, high F_{ST} values were also used as an indicator of possible selective pressures acting on specific variants (Holsinger & Weir, 2009).

2.2.5 Genotype quality control

To assess the reliability of the genotypes, several validation analyses were performed. First, the genotypes of the offspring of the trios were compared with those of their parents, considering only the cases in which the genotypes were available for all three family members. Each comparison was classified according to whether the genotype of the offspring was compatible with the parental genotypes under Mendelian inheritance.

Second, we compared our genotypes with a high-confidence genotyping dataset reported in (Campoy, 2024). In that study, genotypes for inversions with simple BPs were generated using BreakSeq (Lam et al., 2010) in 3,202 samples of high-coverage 1000GP, and repeat-mediated inversions were genotyped using several experimental PCR-based approaches. Many of these experimentally validated genotypes have previously been published (Giner-Delgado et al., 2019; Puig et al., 2020). Genotypes were considered comparable only when both GeONTIpe-derived and reference genotypes were valid. In addition, for 33 samples and 155 polymorphic inversions, geno-

types were compared with the results of the Pangenome project (see Moreira-Pinhal et al. 2025 for details of the methods).

2.2.6 Hardy–Weinberg equilibrium (HWE)

To evaluate which variants deviated from HWE only genotypes from unrelated samples were included to avoid over-representation of alleles from related individuals. HWE was assessed separately for autosomal and chromosome X inversions; for chromosome X, the analysis was adapted to account for the distributions of sex-specific genotypes. All tests were performed independently for each ancestry. The mid-p-values were computed using the HardyWeinberg R package (Graffelman, 2015) (R version 4.3.1). Variants with a mid-p-value below 3×10^{-5} (Bonferroni correction for multiple tests, using an initial significance level of 0.05) were considered to deviate from HWE.

2.2.7 Linkage disequilibrium (LD) analysis

Pairwise LD was calculated between the inversion alleles and other biallelic SNPs located within 0.5 Mb upstream and downstream of the external coordinates of each inversion region. SNPs were obtained from high-coverage 1000GP data. LD was computed both globally and separately for each ancestry using PLINK v1.9.0 (Chang et al., 2015) following the framework and code described in Giner-Delgado et al. (2019). LD was measured as the squared correlation coefficient (r^2), which in PLINK is normalised by the AF of each variant, representing the proportion of variance in an SNP explained by the inversion. Only samples belonging to high-coverage 1000GP were used in the analysis due to the absence of high-quality SNP information for samples from other sources.

Two LD analyses were performed: one that included all genotypes and a second that excluded homozygous genotypes supported by only 5 to 7 reads. LD was considered high when a value of at least 0.95 was observed with at least one SNP. This analysis allowed for the identification of inversion regions whose genotypes would require further refinement before subsequent functional analyses.

2.2.8 Manual sequence analysis

Throughout the different stages undertaken to achieve the objectives of this thesis, manual sequence analysis was frequently required. This was applied to the inspection of both specific reads or assembled sequences, allowing us to accurately determine which variants could be directly observed in the sequence. To perform these analyses, several visualisation and annotation tools were used.

IGV 2.17.4 (Robinson et al., 2023) was used for the direct visualisation of mapping files, allowing visual detection of certain SVs. BLAST (Boratyn et al., 2013), used via the command-line interface or the web application, was used to align reference genome sequences with the query sequence of interest. This facilitated the identification of sequence blocks that differ in size, orientation, or copy number between the reference and the read query. BLAST results were usually inspected using supplementary tools such as Artemis v18.2.0 (Carver et al., 2012), which provides graphical representations of BLAST hits, allowing quick identification of the position and orientation of sequence blocks. However, when the complexity of the genomic region increased, due to the expansion of repetitive content or intricate genomic structures, the QIAGEN CLC Main Workbench 25.0 (QIAGEN Digital Insights, 2025) was used to annotate BLAST-derived sequence blocks and take advantage of other useful features in manual sequence analysis.

In the subsequent sections, instances of manual sequence analysis will be noted whenever this step was essential to understand the structural changes observed with the tools developed during this work.

2.2.9 Complex regions analysis

To enable a more systematic analysis of structurally variable genomic regions across a large number of samples, we designed CHARONTE (**C**omplex **H**aplotype **A**nalysis and **R**econstruction using **O**xford **N**anopore **T**echnology **E**evidence) a pipeline that, as described in detail below, uses a set of unique k-mers that are directly mapped onto long reads. By analysing the position, orientation, and spacing of these k-

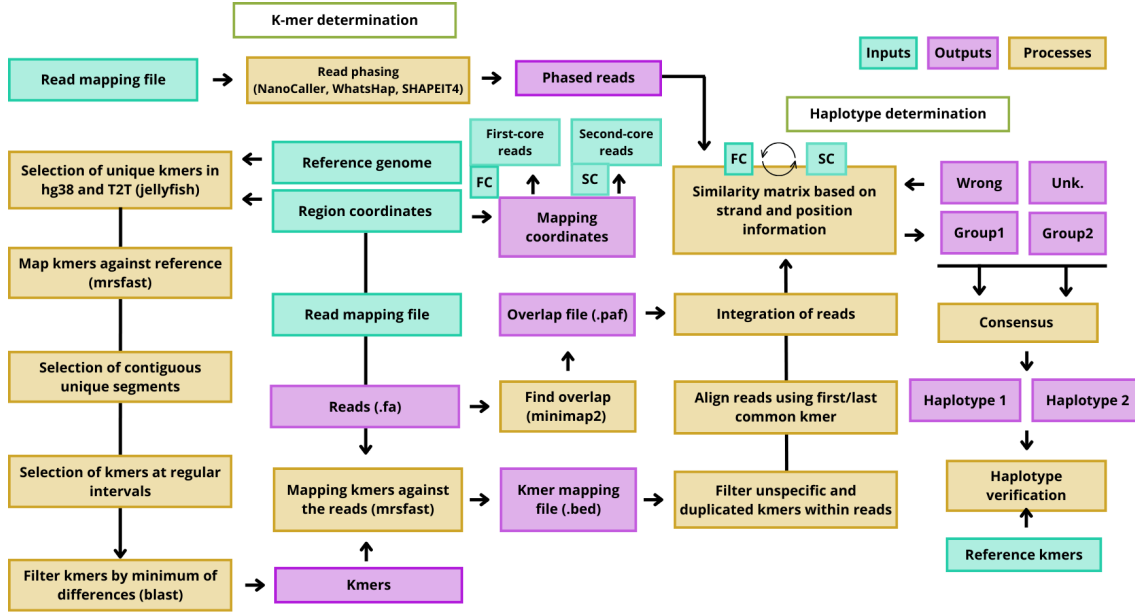


Figure 2.2: **Schematic overview of CHARONTE.** The figure illustrates the two main code blocks: k-mer determination and haplotype determination. Major steps include the selection and mapping of unique k-mers, phasing of ONT reads, identification of first-core and second-core reads, construction of a similarity matrix using positional and strand information, clustering of reads into haplotype groups, and generation of consensus haplotypes followed by verification using reference k-mers. Colours indicate the type of elements: turquoise represents inputs, yellow denotes processing steps, and lilac represents outputs.

mers within each read, the pipeline identifies reads that follow the same structural path and incrementally merges them, ultimately extending across the entire region of interest. This process allows us to reconstruct the structural haplotypes of the sample under study. Once the haplotypes of all individuals are inferred, the final stage of the pipeline integrates this information to explore the diversity of haplotypes present in the cohort. As the pipeline is still under development, a formal reference is not yet available.

The pipeline is organised into three code blocks that will be named "K-mer selection", "Haplotype determination", and "Synteny". An entire code is implemented in Bash and R v4.3.3 (R Core Team, 2021). To facilitate its interpretation, a schematic overview has been generated (Figure 2.2), and each section will be introduced and explained in three parts: description of the input parameters, workflow logic, and summary of the output produced.

K-mer selection

This first code block is responsible for generating all the information required and subsequently used by the second block of the pipeline.

Inputs:

The inputs necessary to execute this initial stage include a file containing the coordinates of the region of interest, the reference genome, and a mapping file with the alignments of the ONT reads to that reference. As the pipeline relies on two reference genomes for k-mer identification, two coordinate files must be provided, one for GRCh38 and one for T2T-CHM13, along with both reference genome sequences (FASTA format).

Workflow:

In this code section, the coordinates provided as input are first used to extract the target genomic region from the mapping file and to retrieve the corresponding sequence from each reference genome. These tasks are performed using SAMtools v1.15 (Li et al., 2009; Danecek et al., 2021). From the resulting mapping file, additional information is obtained, including read mapping flags, read sequences, and the position of each read relative to the reference. The mapping information is converted to BED format using bedtools v2.30 (Quinlan & Hall, 2010). In addition, overlaps between reads are identified using the overlap function of minimap2 v2.24 (Li, 2018), with optimised parameters for ONT long reads.

From the extracted GRCh38 reference sequence, all possible k-mers of fixed size are generated. In our implementation, the length of the k-mer can be defined by the user and can range from 20 to 50 bp (the optimal k-mer size was empirically determined by testing multiple k-mer lengths within a specific genomic region across a small set of individuals, and selecting the parameter that yielded the highest concordance with prior knowledge of the structural configuration of that region). The overlap between k-mers is set to one base, so that k-mers with a single base of difference with respect to the previous one are generated. K-mers are generated using Jellyfish v2.3.0

(Marçais & Kingsford, 2011), producing a maximum number of k-mers calculated as:

$$N_{\text{k-mers}} = \frac{\text{region size} - \text{k-mer size}}{\text{k-mer size} - \text{overlap between k-mers}} + 1$$

Jellyfish is also used to select unique k-mers. Finally, k-mers are filtered to keep only those found once in the analysed region in the second reference genome (T2T-CHM13).

After finalising the unique list of k-mers, consecutive k-mers are mapped back to the reference genome using mrsFAST v3.4.2 (Hach et al., 2014) with a zero-error tolerance (accepting only perfect matches) to identify uninterrupted sequence blocks. From each unique sequence block, a subset of k-mers is selected. At least one k-mer per block is generated, and the maximum number is calculated as:

$$N_{\text{selected k-mers}} = \frac{\text{block size}}{\text{window size}}$$

where the window size is a user-defined parameter that varies from a few hundred to several thousand bases. If a block is smaller than the window size, one k-mer is selected from each end of the block.

To minimise potential non-specificities, users can define a minimum number of nucleotide differences that a k-mer must have relative to other sequences in the region. This threshold defines a minimum identity for each k-mer, calculated as:

$$\text{Minimum Identity (\%)} = \left(1 - \frac{\text{minimum differences}}{\text{kmer size}} \right) \times 100$$

Using this identity threshold, BLAST v2.16.0 (Boratyn et al., 2013) is used to identify k-mers with multiple hits exceeding the threshold and then being discarded to prevent non-specific mapping.

Finally, all remaining k-mers are mapped to the reads using mrsFAST v3.4.2, spec-

ifying the desired error rate. In our workflow, only exact matches are accepted. The output is a BED-format table that indicates, for each read, which k-mers were mapped, their positions, and their strand orientation.

To facilitate downstream classification of reads, a final step was incorporated in a later stage of the development of CHARONTE, allowing the phasing of reads in the target region. For this purpose, NanoCaller (Ahsan et al., 2021) is used to perform the call of SNPs, identifying heterozygous variants in the individual analysed. Then phasing is performed using WhatsHap (Martin et al., 2016), to determine which reads belong to each haplotype and the number of phase blocks. Within each block, the reads are assigned to one of the two haplotypes or classified as unphased when no reliable assignment is possible. When multiple phasing blocks are detected, SHAPEIT4 (Delaneau et al., 2019) is used to merge them into a single coherent phasing structure, using known SNPs from 1000GP.

Outputs:

The outputs generated in this first section of the pipeline include: (i) the mapping information of k-mers against the reads and of the reads against the reference; (ii) the phasing information of the reads; and (iii) the overlap characteristics among reads. Subsequently, these components are used as inputs for the second section of the developed software.

Haplotype determination

This second block constitutes the core of the pipeline. Its main purpose is to determine the structural configuration of the two chromosomes of the individual sample. This section generates multiple information records that can be examined later, including an estimate of the structural differences identified between the unknown sample and the reference genome.

Inputs:

In addition to the files generated in the previous section, several basic inputs are required to enable parallelisation, such as the region ID, coordinates, and sample ID. Finally, four coordinates must be supplied that delimit the region where variation is hypothesised to occur; these coordinates often correspond to SDs.

Workflow:

The initial steps of this code block involve the progressive introduction of input data used to define the elements of interest that will later enable the classification of reads.

First, several rounds of correction are applied to the k-mers detected in the reads. When a given k-mer is mapped more than once within the same read, it is immediately discarded; similarly, k-mers located in the external ends of reads and outside of the analysed region are also removed. Then, to properly align each read to the reference, it is necessary to translate the k-mer mapping coordinates observed within the read into reference-based coordinates. Therefore, a procedure is implemented to determine an anchor point between each read and the reference. This method identifies the first k-mer shared by the read and the reference. Once this anchor k-mer is found, its position within the read is reassigned according to the reference coordinates. The positions of all other k-mers in the same read are then recalculated by preserving their relative distances to the anchor.

After this step, k-mers that do not follow the positional trends of the remaining k-mers within the read are removed. When multiple inconsistent k-mers appear within the same read, the algorithm attempts to detect whether a structural variant may explain the anomalous pattern. To this end, it searches for other reads exhibiting similar inconsistencies and identifies the corresponding regions of the reference in which they occur. This procedure allows for the identification of duplications and inverted duplications within the reads. When such events are detected, the k-mers belonging to the conflicting region are removed, and the region is annotated as a

variant.

Once each read contains a pool of unique k-mers that follow homogeneous and consistent positional and strand patterns, reads with too few k-mers are filtered out because they are not informative for classification. These correspond to cases in which less than 0.5% of the total number of k-mers in the region, or less than two k-mers in total, are detected within the read.

The next stage involves read integration. Here, longer reads with more k-mers are used as reference containers to incorporate shorter reads that may contribute additional k-mers initially not found in the longer read due to the high error rate of ONT reads and the mapping of k-mers requiring a perfect match. This step uses the sequence overlap information between different reads obtained in the previous pipeline stage. Using this information, a sequential integration process is applied when: (i) the reads overlap by more than 90% of the sequence length, (ii) present less than 5% sequence divergence, (iii) belong to the same phasing group, and (iv) the shared k-mers in the overlapping region are identical. After each integration event, the shorter read is discarded to avoid repeated incorporation, and the short-read specific k-mers are incorporated into the overlapping longer read. Additional consistency reviews are then performed to ensure that incoherent k-mers are not introduced.

To determine whether two reads are considered identical or distinct based on the k-mers they contain, we use a simple matching coefficient (SMC), which evaluates the similarity between independent attributes in a binary fashion (0 or 1). Two attributes are considered in this evaluation: position and strand. The comparison of strands is direct, identical strands score 1, and mismatched strands score 0. For each k-mer position, we computed the ratio between the k-mer position in the read and its position in the reference (after realignment). If this ratio is higher than 95%, the k-mer is scored as 1 (equal). If not, the absolute position of the k-mer is analysed and if the difference between position within the read and in the reference is smaller than 5000 bp, it is also scored as 1. Otherwise, it is scored as 0.

By repeating this procedure for all the k -mers in the read, a global similarity score between two reads is obtained, defined by the following formula:

$$\text{similarity score} = \frac{\sum_{k=1}^p \text{pos_sim}_k + \sum_{k=1}^s \text{strand_sim}_k}{p + s}$$

where p is the number of comparable positional observations, s is the number of comparable strand observations, pos_sim_k is the binary indicator of positional similarity for the k -th k -mer, and strand_sim_k is the binary indicator of strand similarity for the k -th k -mer. This similarity score serves as the basis for all read–read comparisons performed throughout the pipeline, and it will be referenced in the following sections.

The following step uses the mapping information of reads aligned to the reference to determine which reads cross at least one predefined point of interest. These points correspond to the four coordinates within the analysed region introduced as input. Reads meeting this criterion are labelled as first-core reads, whereas the remaining ones are classified as second-core. This distinction establishes the prioritisation of the reads used during the classification process.

To classify the reads, four groups are defined, and each read is assigned to one group as it is processed. Two of these groups correspond to the two haplotypes (Group 1 and Group 2), while the remaining two groups contain reads that could not be assigned to either haplotype, either because they were not comparable (Unknown) or because their similarity score was too low (Wrong). A read is assigned only to an haplotype group if the phasing information and the similarity score match. However, unphased reads may be assigned to the haplotype group with which they show the highest similarity.

In the first classification stage, the reads from the first-core are analysed. The read with the highest number of k -mers is selected as the initial target for comparison. Sequentially, each read is compared to this target based on shared k -mers. If its similarity score exceeds the threshold, the read is assigned to that group. The first

read that shows low similarity or conflict with the phasing label becomes the first read and target for the second haplotype group. In subsequent comparisons, each read is used as a query against both groups and assigned to the one that provides the best match. Reads that do not fit either group are moved to the unassigned categories.

This stage ends with an internal validation of each haplotype group. All reads within a group are compared pairwise to detect inconsistencies. The read with the lowest similarity score in the resulting matrix is removed and checked if its removal substantially improves the overall mean pairwise similarity of the group. This process is repeated until no further improvement is observed and all remaining reads in the group are considered internally consistent.

Once the groups have been validated, a k-mer consensus is generated for each haplotype group. This expands the k-mer set that represents each group, providing an updated target for comparison in the second round. In this second round, the remaining unassigned reads from the first-core are compared against the new consensus profiles of Groups 1 and 2. After this stage, a new consensus is generated and all reads from the second core are compared against it. This iterative classification of unassigned reads against the newly created consensus for each group is repeated twice for first-core and twice for second-core reads to maximise the number of reads assigned to each haplotype.

As a final classification step, the proportion of reads in each group is evaluated to determine whether a reconstructed haplotype is valid. If more than 30% of all reads remain unassigned, the classification process is considered unsuccessful. Additionally, a haplotype is considered invalid if its corresponding group contains fewer than 25% of the total reads.

Finally, a k-mer consensus is generated for each reconstructed haplotype and evaluated in several ways. First, the k-mer composition is compared with that of the reference to identify divergent k-mers and characterise the nature of these differences. Then, the internal consistency of each haplotype is assessed by examining

the number of k-mers used in its reconstruction, its total size, and whether differences in the size difference can be explained by the observed variation relative to the reference. Haplotypes showing inconsistent k-mer distributions are classified as erroneous, whereas haplotypes that exhibit substantial k-mer loss or size discrepancies are categorised as partial.

Outputs:

The outputs generated in this step consist of the two reconstructed haplotypes, each represented as a set of k-mers with their corresponding positions and assigned strands. In addition, several images are produced to facilitate the verification of the consistency of the process, along with a log file containing the statistics generated throughout the procedure and the final evaluation of each reconstructed haplotype.

Synteny

Unlike the previous module of the pipeline, which analyses the haplotype structure of individual samples, the final stage performs a comparative analysis across all samples. The aim is to identify haplotypes from different individuals that are identical, so they can be recognised as the same haplotype, and those that are different. This enables a final representation of all distinct haplotypes detected in the population, together with the estimation of their frequencies and differences between them and with respect to the two reference genomes used for the k-mer design.

Inputs:

As input, this third module uses the haplotypes generated in the previous block and the sample IDs provided by the user.

Workflow:

In this third block, all haplotypes that were not previously classified as erroneous or partial are included. The correctly inferred haplotypes are compared pairwise using the SMC metric described above. Haplotypes with a score of 1 are assigned to the

same haplotype group, and the k-mer information from the contributing samples is used to generate a consensus, either increasing the k-mer coverage in the final haplotype or improving the positional accuracy of each k-mer. The frequency of each haplotype is determined.

To characterise the differences among all detected haplotypes, they are first divided into two groups based on strand changes: those with strand changes indicating the presence of an inverted segment (inverted group) and those without (standard group). Within each group, the haplotype with the highest frequency is selected as the central reference and then the central haplotypes of both groups are compared with each other. The pipeline then identifies the most complex haplotype within each group. Complexity is defined either by being the most dissimilar relative to the other haplotypes in the same group or by being the longest haplotype. These outer, complex haplotypes are compared last and only against the adjacent haplotype in terms of similarity, reflecting their peripheral position. In contrast, the remaining haplotypes are compared sequentially against their nearest neighbours in similarity.

Once the haplotype pairs to be compared are defined, each pair of k-mers is evaluated to pinpoint the location of sequence variation. Variants are classified as deletions or insertions when changes in k-mer spacing are observed, or as inversions when strand orientation switches. All detected variants are annotated and stored, providing a qualitative profile of the differences between each pair of haplotypes. After annotation, variant coordinates are formatted for use in a synteny visualisation. The resulting synteny plot is generated using customised modifications of the Synteny-Plotter package (Quigley et al., 2023).

Finally, individual genotypes are inferred for each sample on the basis of their haplotype combinations. A summary table is also produced, reporting the haplotype frequencies and the individuals in which each haplotype is detected. Each haplotype is additionally compared with both reference genomes (GRCh38 and T2T-CHM13) to provide a simple record of haplotype-specific differences relative to each reference.

Outputs:

The outputs generated by this module include individual genotype calls, summary tables, pairwise haplotype comparisons, and comparisons between each haplotype and the reference genomes.

3. Results

3.1 Genotyping of inversions using long-read data

The first chapter of this thesis Results comprises a manuscript currently in press that presents GeONTIpe, a novel computational package for genotyping of inversions mediated by inverted repeats (IRs) based on the use of Oxford Nanopore Technologies (ONT) long-read data. In addition to the description of the methodological framework of the tool, the publication also presents its application to an initial dataset comprising 612 inversion candidate regions. These inversions were compiled from multiple studies that employed recent sequencing technologies, as well as from the INVersion Functional & Evolutionary STudies (InvFEST), a database that, as discussed throughout this thesis, aggregates inversion calls reported in numerous independent studies. This inversion dataset was genotyped in a cohort of 54 individuals for whom ONT long-read sequencing data was available.

The publication reports the main findings derived from the analysis of the inversion dataset, including an evaluation of genotyping performance between variants and samples. Furthermore, it explores the use of adaptive sampling as a strategy to reduce sequencing costs in ONT experiments and assesses the impact of this approach on inversion genotyping.

Finally, the article presents several genotype concordance analyses to assess accuracy and reliability, comparing GeONTIpe genotypes with those reported in the original studies from which inversions were obtained. It also includes a preliminary evaluation of the Human Pangenome as a potential source of inversion genotypes.

Resolving missing human polymorphic inversions and other complex variants from ultra-long read data

Ricardo Moreira-Pinhal^{1,2}, Konstantinos Karakostis^{1,2}, Illya Yakymenko^{1,2}, Oscar Conchillo^{1,3}, Maria Díaz-Ros^{1,2}, Andrés Santos³, Miquel Àngel Senar⁴, Jaime Martínez-Urtaza³, Marta Puig^{1,3}*, Mario Cáceres^{1,2,5}*

¹ Institut de Biotecnologia i de Biomedicina, Universitat Autònoma de Barcelona, Bellaterra, Barcelona, Spain

² Research Programme on Biomedical Informatics (GRIB), Hospital del Mar Research Institute, Barcelona, Spain

³ Departament de Genètica i de Microbiologia, Universitat Autònoma de Barcelona, Bellaterra, Barcelona, Spain

⁴ Àrea d'Arquitectura i de Tecnologia de Computadors, Universitat Autònoma de Barcelona, Bellaterra, Barcelona, Spain

⁵ ICREA, Barcelona, Spain

* Corresponding authors

Running title: Resolving human inversions with long reads

ABSTRACT

Inversions are a unique type of balanced structural variants (SVs) with important consequences in multiple organisms. However, despite considerable effort, these and other complex SVs remain poorly characterized due to the presence of large repeats. New techniques are finally allowing us to identify the full spectrum of human inversions, but the number of individuals analyzed is still quite limited. Here, we take advantage of Oxford Nanopore Technologies (ONT) long reads to characterize an exhaustive catalogue of 612 candidate inversions between 197 bp and 4.4 Mb of length and flanked by <190-kb long inverted repeats (IRs). For that, we have developed a bioinformatic package to identify inversion alleles reliably from long-read data. Next, using a combination of different DNA extraction, library preparation, and ONT sequencing protocols, we show that ultra-long reads (50-100 kb) and adaptive sampling are an efficient method to detect most human inversions. Lastly, by analyzing ONT data from 54 diverse individuals, 87-99% of the inversions can be genotyped in each sample, depending mainly on read and IR length and genome coverage. Both orientations are observed for 155 of the analyzed regions (frequency 0.01-0.49), which multiplies by three the number of polymorphic IR-mediated inversions studied in detail so far. Moreover, we found more than 300 additional independent SVs in the studied regions and resolved several complex rearrangements. Therefore, our work provides an accurate benchmark of those inversions that typically escape most analyses, and it demonstrates the potential of nanopore sequencing to characterize missing human genomic variation.

INTRODUCTION

Over the past few decades, there has been a global effort to identify all human genomic variation and its association with phenotypic traits and disease susceptibility in different populations (The 1000 Genomes Project Consortium 2015; Collins et al. 2020; Ebert et al. 2021; Karczewski et al. 2020; Choudhury et al. 2020; Choi et al. 2023). However, structural variants (SVs), which involve changes of more than 50 bp, are often generated through non-allelic homologous recombination (NAHR) between large repeats and tend to occur in complex regions harboring many segmental duplications (SDs), making their detection very difficult by most techniques (Alkan et al. 2011; Logsdon et al. 2020; Ebert et al. 2021). Thus, a big fraction of SVs are still under-studied and new high-throughput methods to determine reliably their genotypes in multiple individuals are needed.

This is especially problematic for inversions, which are generally undetected with most sequencing techniques due to their balanced nature, changing only the sequence order (usually with no gain or loss of DNA), and the frequent presence of highly identical inverted repeats (IRs) at their breakpoints (Puig et al. 2015; Ebert et al. 2021). In addition, compared to the rest of SVs, inversions are unique as they inhibit recombination between the two orientations, leading to potential negative consequences on fertility (Puig et al. 2015). Accordingly, those that reach a certain frequency in natural populations could be more likely to have positive functional effects that compensate for these costs. In fact, a growing number of inversions have been involved in adaptation and phenotypic variability in diverse organisms,

from plants to birds (Wellenreuther and Bernatchez 2018), including humans (Campoy et al. 2022). Moreover, around one third of human inversions are associated with gene expression or epigenetic changes, and they tend to have larger effects than other variants (Giner-Delgado et al. 2019; Puig et al. 2020; Lerga-Jaso et al. in prep.).

Large-scale projects using a combination of novel genomic techniques, such as long-read sequencing, Strand-seq or Bionano optical maps (Levy-Sakin et al. 2019; Chaisson et al. 2019; Audano et al. 2019; Ebert et al. 2021; Porubsky et al. 2022) are finally making it possible to identify the full spectrum of human inversions. However, each technique has its own limitations and error sources. In addition, in most cases, just a reduced number of individuals has been studied, which hinders the analysis of the effects of the detected variants and their association with phenotypic traits. To overcome these problems, different methods have been developed to genotype specific inversions, either individually (e.g. regular PCR, inverse PCR (iPCR) or droplet digital PCR (ddPCR)) or in sets of several of them together (e.g. multiplex ligation-dependent probe amplification (MLPA) or inverse MLPA (iMLPA)) (Giner-Delgado et al. 2019; Puig et al. 2020). This has allowed the detailed analysis of up to 134 inversions in 95-551 individuals from diverse human populations, including 55 inversions with IRs at their breakpoints (Lerga-Jaso et al. in prep.). Nevertheless, the variants that can be studied with each method are extremely dependent on inversion and IR characteristics, such as their length or the presence of restriction enzyme target sites, which means that a single

technique cannot detect all inversions and a significant fraction of them are still not characterized. In addition, many inversions generated by NAHR have appeared recurrently in the human lineage and they are not linked to nearby SNPs or particular haplotypes (Giner-Delgado et al. 2019; Puig et al. 2020; Porubsky et al. 2022). In fact, we have seen that highly recurrent inversions are very difficult or even impossible to impute, compared to unique inversions that typically have perfect tag SNPs (Giner-Delgado et al. 2019). As a result, their potential effects have not been detected in common genome-wide association studies (GWAS). Therefore, it is especially important to develop methods that allow us to directly determine the genotype of an individual by investigating the actual DNA sequence order, over other indirect methods, like imputation, that can have a high error rate.

Long-read sequencing methods, such as PacBio and Oxford Nanopore Technologies (ONT), represent a unique opportunity for the complete characterization of complex genomic variants (Logsdon et al. 2020; De Coster et al. 2021; Mastroiosa et al. 2023; Grochowski et al. 2024). Specifically, ONT sequencing is based on identifying the nucleotide bases as individual DNA molecules pass through multiple nanopores, which can generate extremely long reads of up to several Mb in a relatively short time (Logsdon et al. 2020). Despite its higher error rate, ONT ultra-long reads (>50-100 kb) have been instrumental in the generation of the new Telomere-to-Telomere (T2T) and draft human pangenome references, helping to resolve centromeres and other hard-to-characterize regions (Nurk et al. 2022; Liao et al. 2023). Furthermore, adaptive sampling promotes the sequencing

of targeted molecules in real-time (Payne et al. 2021), enriching regions of interest in a simple and cost-effective way. This strategy has already been applied to identify rare genomic rearrangements and genotype difficult-to-study variants, such as tandem repeats (Miller et al. 2021; Stevanovski et al. 2022). Also, some studies are starting to generate long-read resources of multiple individuals of diverse populations and diseases (De Coster et al. 2021; Beyter et al. 2021; Wu et al. 2021; Gustafson et al. 2024; Schloissnig et al. 2025), although in most cases read length is moderate. Therefore, it is necessary to undertake further analyses to establish optimized protocols that can be widely adopted and define the current limits of these technologies. In addition, it is important to develop easy to use bioinformatic tools that can extract accurate information of different types of genomic variants from these new data.

In this work, we take advantage of long-read information to generate an accurate catalogue of human polymorphic inversions mediated by a wide range of IR lengths and to obtain an initial estimate of their frequency and distribution. Moreover, we investigate the applicability of ONT ultra long reads for efficient genotyping of inversions, as a first step to characterize their functional impact. Finally, we compare our results with other studies and the draft human pangenome reference to check how well these variants are represented.

RESULTS

Dataset of IR-mediated inversions

In order to build the most exhaustive dataset of IR-mediated human inversions to date, candidate inversion regions were selected from four different sources: (1) all 58 independently validated inversions with different types of breakpoint IRs (ranging from 140 bp to 134 kb), including the 55 genotyped in multiple individuals using different PCR-based techniques (Giner-Delgado et al. 2019; Puig et al. 2020; Lerga-Jaso et al. in prep.); (2) the merging of 1528 candidate inversions from five recent studies (Levy-Sakin et al. 2019; Chaisson et al. 2019; Audano et al. 2019; Ebert et al. 2021; Porubsky et al. 2022) that identify SVs using long-read sequencing, Strand-seq or optical mapping in a total of 195 individuals; (3) previous paired-end mapping (PEM) predictions that could not be analyzed with available methods or were identified as possible human genome assembly errors caused by highly-identical IRs (Korbel et al. 2007; Martínez-Fundichely et al. 2014; Vicente-Salvador et al. 2017); and (4) pairs of inverted SDs annotated in the human reference genome (GRCh38/hg38) that have the potential to generate inversions (see Methods). Candidate inversions from the different sources were merged into independent variants when their breakpoints were located within or close to the same pair of IRs. Also, we selected only inversion candidates with at least one IR copy smaller than 200 kb and not located in very complex regions full of repeats and SDs (which excludes most of the centromeric and telomeric sequences) or gaps in hg38, for which specific probe sequences could be designed around the breakpoints (see below).

Of the merged list of candidate inversions by recent studies, 622 regions were filtered out since they did not meet the above criteria (mostly because no IRs were identified in the region or sequence was extremely complex) and 11 more due to unexpected complex behavior during inversion genotyping. This resulted in 150 new possible independent inversions with IRs that could be tested (including 65 with breakpoints within or close to the same IR pair that were predicted by at least three independent studies), not counting 58 matching those already known (Lerga-Jaso et al. in prep.). In addition, there were 85 PEM inversion predictions overlapping IRs (Martínez-Fundichely et al. 2014) that were not present in the previous set. Finally, we selected an additional group of 319 pairs of inverted SDs 1-50 kb long and distance of <250 kb that do not correspond to any of the other inversions. In total, 612 regions with potential inversions ranging from 197 bp to 4.4 Mb and flanked by IRs between 140 bp and 190 kb were included in the analysis (Supplemental Table S1). However, the inversion selection could exclude some cases in which the IRs are not found in the current hg38 human reference genome assembly.

Development of a new inversion genotyping package from long reads

Thanks to the sequence contiguity, ONT ultra-long reads offer the opportunity to analyze a wide range of inversions and resolve other challenging variants, without being affected by sequencing error rates or other methods limitations (Giner-Delgado et al. 2019; Puig et al. 2020). Thus, we have developed a novel inversion genotyping software package called GeONTipe. The genotyping

strategy is based on the detection of reads spanning inversion breakpoints and determining their sequence order and orientation by searching for specific probe sequences at either side of each IR (A, B, C and D) (Fig. 1A). That way, we can identify reads with A-B/C-D combinations that support the reference orientation (O1) or A-C/B-D combinations that support the inverted orientation (O2) (Fig. 1B).

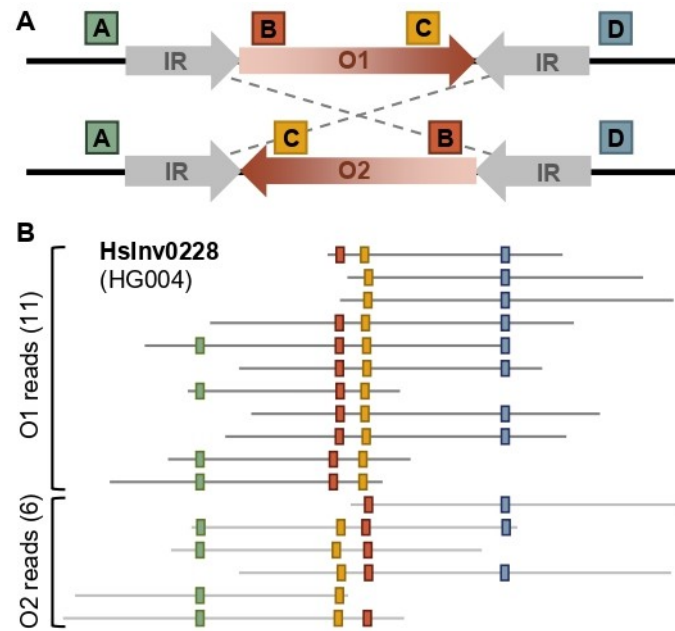


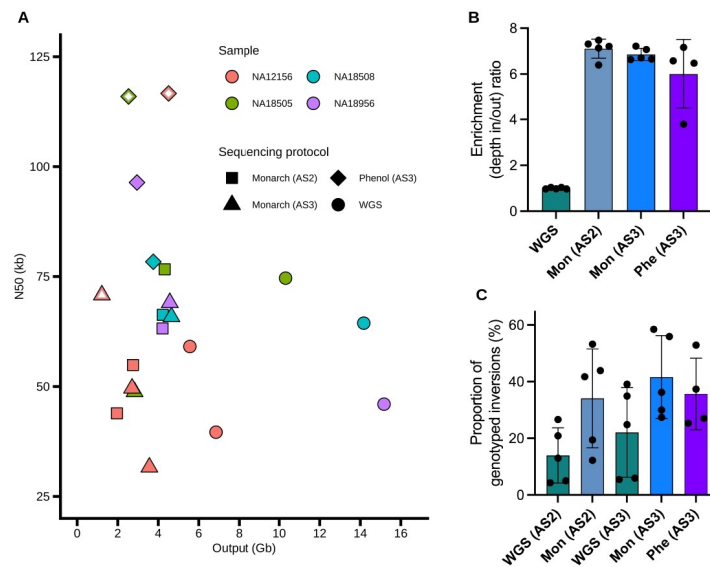
Figure 1. Inversion genotyping from ONT long reads. **A.** Schematic representation of candidate inversions analyzed in this work, with the inverted region shown as a red gradient flanked by two inverted repeats (IR) in grey. Probe sequences located at both sides of each IR (labelled as A, B, C and D) are indicated as colored rectangles and have different order in O1 and O2 orientations. **B.** Example of HsInv0228 inversion genotyping results in an heterozygote sample, showing ONT individual reads as dark grey (O1) or light grey (O2) lines and the probe sequences indicated with colored rectangles. Probe sequence order together with read number and proportion of reads supporting each orientation determines the inversion genotype.

The analysis has several steps: quality filtering of the reads, mapping against hg38, generation of inversion genotypes, and identification of additional SVs (see Methods; Supplemental Fig. S1). One of the key steps is the design of the probe sequences of a few hundred bp to identify unique regions flanking the breakpoint IRs, which were manually revised to avoid specificity problems and possible errors caused by the presence of unexpected haplotypes. This allowed us to interrogate accurately 612 potential inversions, including 40 inversions analyzed only using one breakpoint due mainly to the presence of large IRs or the lack of specific probe sequences. Once these sequences were mapped to the reads, inversion genotype was determined according to the number of reads supporting each orientation as homozygote, heterozygote or low confident (see Methods). Moreover, to recover additional genotypes, when possible inversion homozygotes were supported just by 2-5 reads, known SNPs were checked to see if the two alleles were present. Finally, we took advantage of the distance between the probe sequences to identify additional insertions (increased distance) or deletions (reduced distance) in the analyzed regions. Therefore, the developed method makes it possible to resolve a large majority of human inversions mediated by a wide range of IRs for the first time, while it characterizes also additional SVs.

Benchmark of ONT sequencing strategies for inversion genotyping

In parallel, to assess the most efficient methodology for genotyping the whole set of human inversions, we tested several ONT ultra-long read sequencing protocols using a MinION device. The analysis was multi-dimensional and includes

two main groups of experiments on four human female samples of diverse origins for which most of the PCR-validated inversions had been previously analyzed (Lerga-Jaso et al. in prep.): NA12156 (European), NA18505 and NA18508 (African), and NA18956 (East Asian) (Fig. 2A). First, we used the Monarch® HMW DNA extraction method (NEB) to compare the genotyping efficiency of whole-genome sequencing (WGS) and adaptive sampling (AS) of two sets of inversions: an initial set, including 132 candidate inversions plus 40-kb flanking regions (AS2), and an extended one, targeting additional predicted inversions and inverted SDs plus 70-kb flanking regions (AS3), which covers almost all the inversions interrogated in the genotyping pipeline (Supplemental Table S1). Second, we evaluated the effect on read length and total output of an alternative phenol-based genomic DNA purification protocol using AS3.



or phenol) are indicated by different shapes, whereas empty symbols indicate the use of fresh cells as starting material. **B.** Enrichment ratio between the read depth inside and outside the targeted regions (depth in/out) of each sample for diverse types of sequencing (black dots), showing a 6-7-fold average enrichment in AS2 or AS3 compared to WGS (colored bars). **C.** Average proportion of genotypes obtained for the analyzed inversion regions normalized to the number of AS2 or AS3 targets (% , colored bars). Values for each sample (black dots) show considerable variation due mainly to sequencing output differences.

In the different sequencing experiments with the Monarch-extracted DNA, we obtained an N50 of 32-79 kb, with consistent values when the same library was sequenced and no significant differences between WGS and AS (Fig. 2A; Supplemental Table S2). However, the phenol-based DNA extraction method resulted in significantly larger N50, ranging from 79 to 118 kb (Fig. 2A). Also, long-term storage of the cells appears to reduce genomic DNA integrity and read length, with DNA isolated from fresh cells by phenol resulting in higher N50 values than DNA isolated from cells stored at -80 °C for a year (mean of 117.7 vs 87.9 kb, respectively), whereas the N50 of the DNA extracted with the Monarch kit from NA12156 cells kept for 8 years at -80 °C showed clearly the lowest average N50 (50.6 kb) (Fig. 2A). In terms of output, both Monarch AS2 and AS3 experiments yielded an average of 3.3 and 3.1 Gb compared to 9.5 Gb for WGS, which corresponds to a ratio of 2.4, 2.0 and 5.5 Mb/nanopore, respectively (Fig. 2A; Supplemental Table S2). Thus,

our results suggest that read length depends mainly on sample preparation, while the amount of sequence data is reduced by ~3-fold in AS compared to WGS.

Regarding the efficiency and specificity of AS sequencing, we observed a similar average increase on the coverage of the targeted inversion regions for the Monarch AS2 (7.1-fold) and AS3 (6.9-fold) results, which is quite consistent across samples (Fig. 2B). Dividing the genome into 10-kb windows showed a clearly different distribution of the average fold enrichment between targeted and non-targeted windows (Supplemental Fig. S2; Supplemental Fig. S3A) and a significant correlation of the mean enrichment of the different target regions common in AS2 and AS3 ($r = 0.35$, $P < 0.001$; Supplemental Fig. S3B). Still, in both AS datasets a small proportion of non-targeted windows display increased coverage values comparable to that of targeted windows (Supplemental Fig. S2; Supplemental Fig. S3A), which could be due to a significant excess of repetitive elements and/or SDs in these regions and homology between AS targets and repeated sequences (Supplemental Fig. S3C).

When considering inversion genotyping, Monarch AS2 and AS3 sequencing experiments provide respectively an average of 1.9 and 1.7 times more informative reads for inversion genotyping than WGS (Supplemental Table S2), which results from a trade-off between the higher enrichment and the lower output of AS. Similarly, depending on each run results, up to 65% of the targeted inversions could be successfully genotyped in a single MinION flow cell, and this value is 1.9-2.5-

times higher in AS2 (mean = 36.9%) or AS3 (mean = 46.0%) than in WGS (mean = 15.0% and 24.3% for AS2 and AS3 targets, respectively) (Fig. 2C; Supplemental Table S2). Even though phenol-extracted DNA generated significantly longer reads, with N50 >100 kb from fresh cells, this did not result in better inversion genotyping (mean = 39.4%), probably because the longer molecules yielded less total sequence reads (Fig. 1A,C). When the data of the different sequencing experiments of the same sample are merged, 87-91% of the inversions could be genotyped in each individual.

Large-scale inversion genotyping in humans

To obtain a good picture of human polymorphic inversions mediated by IRs, we applied the inversion genotyping pipeline to a diverse set of 54 individuals with publicly available ONT data, including those sequenced in this study (Supplemental Table S3). The analysis showed a high genotyping success (>87%) of the candidate inversion regions in all the samples (Fig. 3A). As expected, the sample genotyping rate is correlated with the sequencing depth and especially N50, with 99% of the inversions being genotyped in the samples with N50 >100 kb (Fig. 3B). Similarly, the distance between probe sequences at both sides of the breakpoints reduces inversion genotyping success due to a lower number of reads crossing the IRs (Fig. 3C). In fact, inversions with probe sequences separated by >100 kb could be genotyped on average in only 47.0% of the samples.

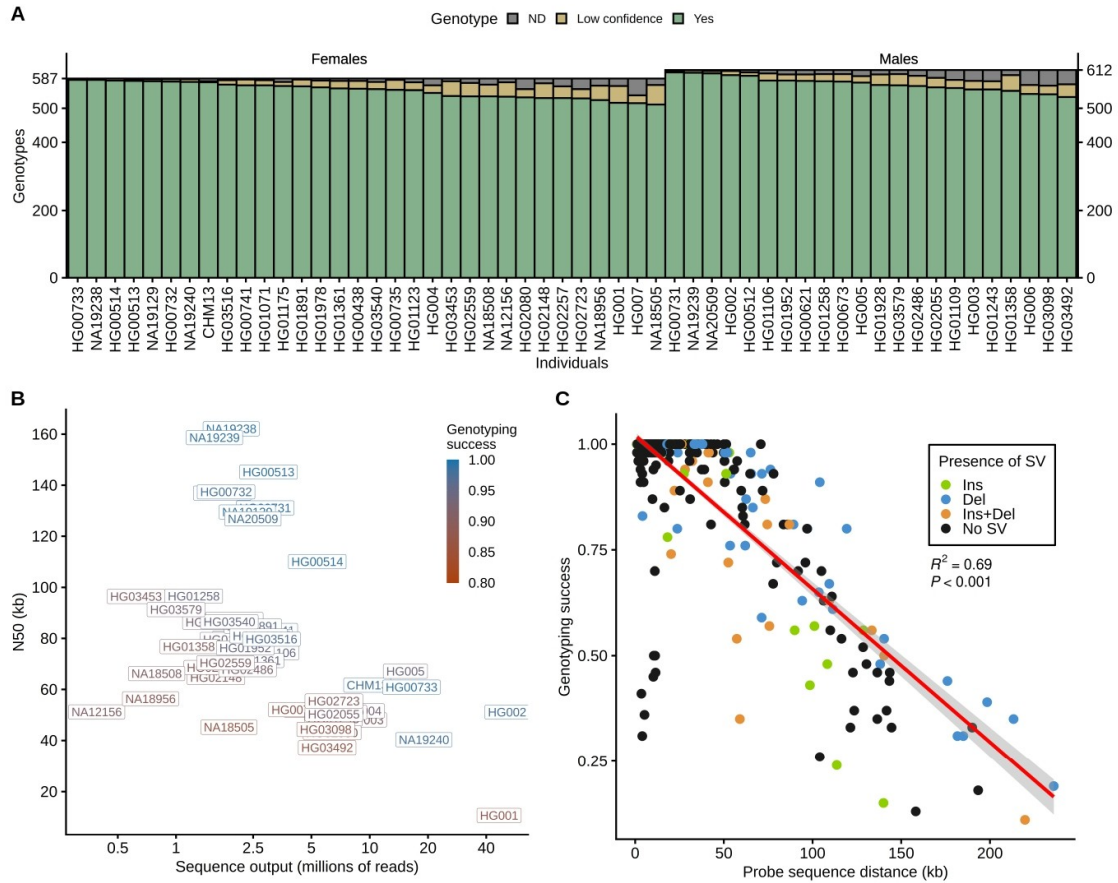


Figure 3. Genotyping success of candidate inversion regions from ONT data in 54 individuals. **A.** Summary of inversion genotyping results by individual (612 for males and 587 for females, excluding 25 inversions located in Chr. Y), with resolved genotypes in green, low confident genotypes due to a reduced number of reads in yellow, and those with no supporting reads (not determined, ND) in grey. **B.** N50 values and total number of reads for each sample, showing the genotyping success rate in a red to blue scale. **C.** Genotyping success of candidate inversion regions as a function of the average distance between the probe sequences flanking the breakpoints. Additional SVs detected in the analyzed regions, some of which can affect genotyping success, are indicated in different colors.

We detected the two orientations in at least one of the 54 individuals (not counting the hg38 reference genome to avoid possible assembly errors) in 155 of the 612 inversion regions tested (25.3%). The minor allele frequency (MAF) of the detected inversions ranged from 0.01 to 0.49, with 78.1% of the polymorphic inversions being widespread across populations (Supplemental Table S1; Supplemental Table S3). However, the regions identified as polymorphic inversions varied depending on their origin, with two alleles detected in these individuals in all but three of the 58 previously known inversions (94.8%), 83 of the additional candidate inversions merged from five different studies (55.3%), 11 of other PEM predictions that had not been validated before (12.9%), and, as expected, only 6 of the inverted SD pairs (1.9%) (Fig. 4A). The previously known inversions not found here correspond to two very low-frequency inversions (HsInv0486 and HsInv0626), and HsInv1869, which could be an hg38 assembly error since only inverted alleles are found in all analyzed samples (Supplemental Table S3). Regarding the data from recent SV studies, a similar proportion of real inversions was found using a multiplatform approach or long reads (75.2-81.0%), with a slightly lower success for those relying only in Bionano optical maps (66.2%), although it is also the study analyzing a larger number of samples (Fig. 4B). The remaining 67 inversion candidates not detected as polymorphic were predominantly detected only in one or two studies (47), which could represent false predictions or low frequency variants. Those found in three or more studies correspond mainly to possible hg38 reference genome assembly errors, including 13 that had been already confirmed experimentally (Vicente-Salvador et al. 2017; Antonacci et al. 2010). Also, there were

a few inversion candidates in complex regions with different breakpoint SD pairs that resulted in different possible inversions in which only one was validated, or that showed low genotyping success caused by the length of the repeat blocks. Also, thanks to the ONT data, two validated reference genome errors (Vicente-Salvador et al. 2017) turned out to be low frequency inversions (HsInv0057, HsInv1119). We also investigated the reasons why the Porubsky et al. (2022) study that focuses on inversions, did not detect 37 polymorphic inversions identified here. We found that 57% of these inversions had low frequency or were small, 24% overlapped with variants that had been classified as inverted duplications (and therefore excluded from our list of candidates) or with slightly different breakpoints, and for the remaining 19% there was no variant called without a clear explanation. This highlights the ability of our approach to detect inversions that had been previously missed.

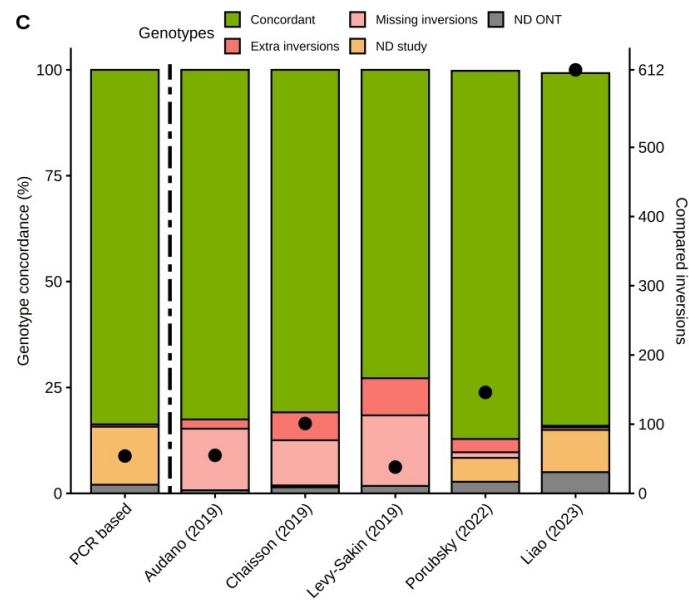
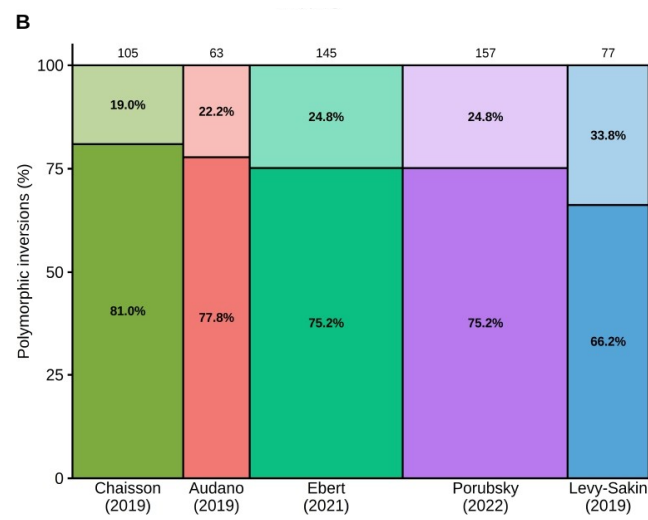
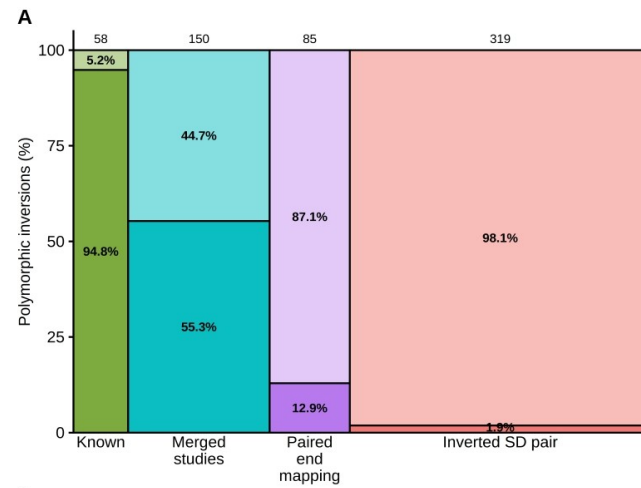


Figure 4. Comparative analysis of inversions identified in this and previous studies. **A.** Mosaic plot representing the proportion of the 612 analyzed regions derived from different sources (bar width, with the number of regions indicated in the top part) and those that were detected in ONT data as polymorphic inversions (dark color) or not (light color). **B.** Mosaic plot representing the candidate regions from each of the merged SV studies included in our analysis (bar width with the total number indicated in the top part) and the percentage detected as polymorphic inversions (dark color) or not (light color). **C.** Concordance of ONT genotypes with PCR-based experimental data (separated by dashed line) and five other studies analyzing the same samples and inversion regions (left Y axis). Genotypes are classified as concordant, having extra or missing inverted alleles, or not determined (ND) in ONT data or the other study used for comparison. Black dots represent the number of inversions compared in each dataset (right Y axis).

Accuracy of ONT inversion genotypes

We performed different comparisons to demonstrate the accuracy of the methodology. First, we checked the genotypes of the polymorphic inversions in five parent-child trios included in our dataset (Supplemental Table S3). Only 2 of the 652 trio comparisons showed discordant results (99.7% concordance): HsInv0808 in the HG002 Ashkenazi trio child that showed a single O2 read and 31 O1 reads, which was insufficient to call it as heterozygote, and HsInv0278 in the HG00731 Puerto-Rican trio father, with 35 O2 reads, that in previous studies already showed discrepant calls as O2 homozygote (Chaisson et al. 2019) and as heterozygote

(Porubsky et al. 2022). In addition, by checking the PCR-based genotypes of 54 inversions in 10 individuals included in the present analysis (Giner-Delgado et al. 2019; Puig et al. 2020; Lerga-Jaso et al. in prep.), we found that 452 of the 455 genotypes (99.3%) were concordant (Fig. 4C), and all three discrepancies correspond apparently to errors in the experimental data. These involved a new inverted duplication identified from the long read data instead of the expected inversion (HsInv0397 in NA12156), a complex duplicated region (HsInv0348 in NA18508) or an inversion with a high genotyping error rate (Giner-Delgado et al. 2019) in which the 1000 Genome Project (1KGP) high-coverage data supports also the ONT results (Byrska-Bishop et al. 2022) (HsInv0045 in NA20509).

Next, we compared our results with those of four previous studies, taking into account the full genotype, if available, or just the presence or absence of the inverted allele (Levy-Sakin et al. 2019; Chaisson et al. 2019; Audano et al. 2019; Porubsky et al. 2022) (Fig. 4C). A lower genotype concordance was observed with older studies, that rely either in one detection technique, such as Bionano optical maps (74.1% from 38 candidate inversions in 6 individuals) (Levy-Sakin et al. 2019) and PacBio long reads (83.1% from 55 candidate inversions in 5 individuals) (Audano et al. 2019), or a multi-platform approach integrating the previous two techniques and Strand-seq (82.4% from 101 candidate inversions in 9 individuals) (Chaisson et al. 2019). Conversely, a recent large multi-platform analysis using the same techniques shows 95.1% genotype concordance from 146 inversion regions compared in 12 individuals (Porubsky et al. 2022).

Finally, we checked how well the analyzed inversion regions are represented in the available diploid genomes of 14 males and 19 females from the draft human pangenome reference (Liao et al. 2023). We could compare 84.2% (16614) of all the genotypes, which showed almost perfect concordance with our results (98.8%, Fig. 4C). However, most of the remaining genotypes could not be compared mainly due to the absence of assembled sequences crossing the breakpoints in some inversions and individuals in the draft human pangenome reference (10.7%, 2120). Thus, even though this new resource provides a good representation of polymorphic inversions, it still misses information in a significant fraction of regions or individuals.

Identification of other SVs and characterization of complex regions

After mapping the probe sequences on each read, the comparison of the actual distances between them with those expected in the hg38 reference genome or its inverted version allowed us to identify additional SVs in the tested regions. In total, we detected 577 deletions and insertions between 250 bp and 469 kb in either O1 or O2 chromosomes (Supplemental Table S4), which correspond to ~322 independent SVs: 195 deletions, 89 insertions, and 38 possible copy number variants (CNVs) (see Methods for details). Of those, there were 243 SVs between the probe sequences flanking one breakpoint, 65 between the two internal probe sequences within the inverted segment, which could be only analyzed for part of the inversions shorter than 50-100 kb, and 14 deletions that remove one or more of

the internal probe sequences. We observed that regions with polymorphic inversions contain more additional SVs (137 SVs in 76 regions (49.0%), with an average of 1.8 SVs per region) than those without inversions (185 SVs in 141 regions (30.9%), with an average of 1.3 SVs per region) (Fisher's exact test $P = 6 \times 10^{-5}$). This could be due in part to a higher SD content in the regions with real inversions (58.6% versus 40.7%) and it is consistent with increased structural diversity associated with inversions (Porubsky et al. 2022). In addition, although our data only analyzed presence or absence of each SV in the reads of an individual, their frequency ranges from those found in only 1 out of 54 analyzed individuals to those found in all of them, corresponding probably to an error or low-frequency haplotype in the reference genome (Supplemental Table S4), with 65.5% of them (211 SVs) detected in several individuals, which indicates that they are relatively common. The association of these additional SVs with inversion alleles could not be analyzed in more detail because for these SVs we do not have complete genotypes, but presence/absence information, and it is also possible that these data are incomplete for some individuals with a low number of reads showing a discordant distance. Also, in IR-mediated inversions it is not clear which is the ancestral or derived orientation and many of these additional SVs have been detected in both O1 and O2 chromosomes, so we cannot be sure of the inversion allele in which they were generated.

For some regions that accumulate several SVs, we combined the pipeline output and manual read analysis to identify multiple structural haplotypes. For

example, in HsInv0401 there is an 8.7-kb polymorphic inversion and a 12.7-kb CNV at one or both breakpoints that can have 2-5 copies per chromosome and result in nine configurations (Fig. 5A). We also characterized other types of complex regions, like that of HsInv0822-0816-0384, where three nested polymorphic inversions can generate up to eight different haplotypes (Porubsky et al. 2022) of which we detected six. Similarly, in the HsInv0012-0659 region, a polymorphic inversion, a CNV with 0-2 copies, two insertions and a deletion generate at least five different haplotypes (Fig. 5B).

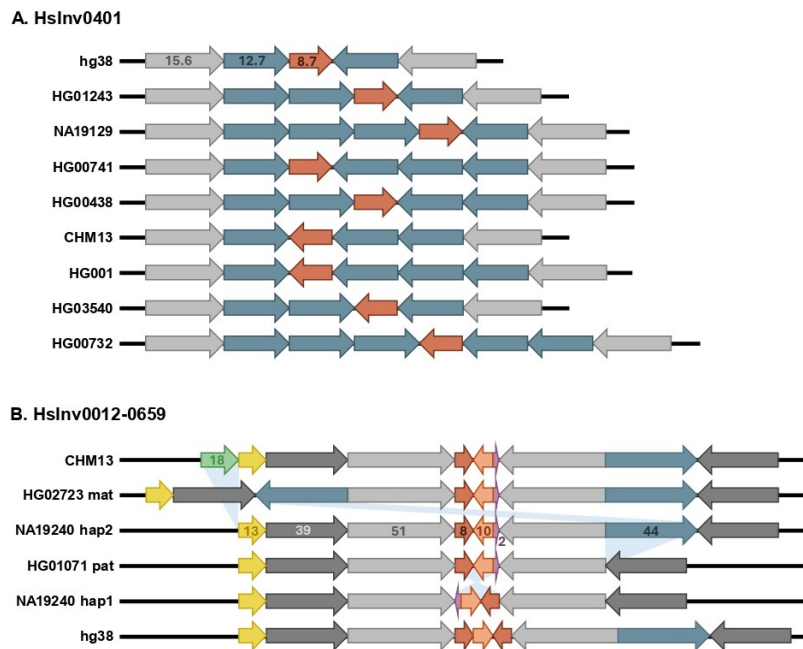


Figure 5. Examples of structural diversity at complex inversion regions resolved with ONT long reads. A. HsInv0401 includes a 12.7-kb CNV with up to five copies per chromosome (blue arrows) located within the inverted repeats at both sides of the inverted sequence (red arrow). B. HsInv0012-0659 includes a polymorphic inversion (red, orange and purple arrows), a 44-kb duplication and deletion (blue arrows) and an 18-kb insertion (green arrow). In this case, the hg38 reference genome sequence is not supported by any other assembly we analyzed so far and might represent an

assembly error. In both images, grey arrows correspond to non-variable regions of inverted SDs and numbers indicate the size of the different segments in kb.

DISCUSSION

Despite ongoing efforts to identify the whole genetic variation in humans (Sudmant et al. 2015; Byrska-Bishop et al. 2022; Levy-Sakin et al. 2019; Chaisson et al. 2019; Audano et al. 2019; Collins et al. 2020; Ebert et al. 2021; Porubsky et al. 2022; Gustafson et al. 2024; Schloissnig et al. 2025), a detailed characterization of the different types of SVs found is needed to determine precisely how many variants there really are and their mutational effects. Here, by doing a comprehensive analysis of the latest human SV information, we have created an accurate catalogue covering a large fraction of human inversions mediated by IRs, which typically escape most analysis. In addition, we have developed an entire bioinformatic and experimental approach to genotype complex inversions not matched by other available techniques. This includes a new easy-to-use bioinformatic package that generates highly reliable inversion genotypes from different types of long read data (Schloissnig et al. 2025). Furthermore, by testing different DNA extraction, library preparation and sequencing protocols, we have shown that the combination of ultra long ONT reads and AS constitutes an efficient method to analyze most human inversions mediated by IRs in a single experiment, which reduces enormously the amount of time and effort required. Finally, by taking advantage of ONT long read data from multiple individuals, we have been able to: (1) confirm the existence of 155 polymorphic inversions, including several

not previously detected; (2) identify more than 300 additional SVs in the analyzed regions; and (3) resolve several complex regions with different rearrangements generating structurally diverse haplotypes that deserve further study.

In fact, ONT long reads represent a more direct way to identify inversions and characterize complex regions than other available techniques, since it just depends on the mapping of long stretches of DNA, which works better than that of shorter sequences, especially for repeats. For example, around half of the IR-mediated inversions from recent studies were not detected in our analysis (Fig. 4A). Part of the discrepancies can be because not the same individuals are interrogated and others correspond to potential assembly errors in the human reference genome (Vicente-Salvador et al. 2017). However, our results suggest that there is still a significant proportion of false predictions, which is illustrated by the 6-60 extra inversion calls in the different studies with respect to the ONT data in the few individuals compared (Fig. 4C). This is especially evident in the case of older PEM inversion predictions (Martínez-Fundichely et al. 2014), with 87.1% of those not already validated or present in the recent studies not detected in our analysis (Fig. 4C), despite half of the analyzed individuals in both cases are the same, which suggests that they were probably errors caused by individual variation within SDs. Moreover, most previous studies capture only part of the actual polymorphic inversions. Therefore, our high-quality inversion benchmark constitutes a very useful resource for SV detection projects and future attempts to represent all

human structural diversity (Wagner et al. 2022), such as the new draft human pangenome reference (Liao et al. 2023).

Another important contribution is the generation of precise genotypes for most inversions mediated by IRs, including many challenging variants that could not be analyzed before. Specifically, we show that ONT inversion results allow us to correct three existing errors from PCR-based techniques (Giner-Delgado et al. 2019; Lerga-Jaso et al. in prep.). Long read mapping has the advantage that avoids some problems of other methods, like the need for restriction sites in specific positions, and it is more robust to different types of variation that cause genotyping errors, such as gene conversion (Giner-Delgado et al. 2019). Also, since the analysis depends on sequences located outside the IRs, any type of recurrence event that occurs between them can be detected. Accordingly, the design of the assays is very flexible and can cover a wide range of inversions, the only limitation being having reads longer than the IRs and finding uniquely-mapping probe sequences at each side. Furthermore, the actual read sequence provides extra information that allows us to identify SNPs or resolve other SVs, as in the low frequency inverted duplication of HsInv0397 and complex regions analyzed here. Still, the high variability of these repeated regions can make that some specific haplotypes are not detected properly and give inconsistent results, as happened with several of the initial candidate inversions. Regarding previous studies, apart from the draft human pangenome reference, we found the best genotype concordance with the most recent

multiplatform analysis (Porubsky et al. 2022), although ONT sequencing achieves this performance with a lower cost and effort than the different techniques used.

The optimization of the sequencing protocol showed that ONT AS strategy obtains a significant and quite-specific enrichment of sequence reads from the regions of interest, which is somehow compensated by a lower total sequencing output of each run (probably as a result of reduced activity and lifetime of pores during AS) (Payne et al. 2021; Kovaka et al. 2021), resulting finally in just a ~2-fold increase in the inversions genotyped compared to sequencing the whole genome. The combination of AS and the appropriate DNA extraction and library preparation methods to generate ultra long reads with N50 up to 70-100 kb allows us to identify consistently 40-60% of inversions in each individual MinION flow cell, which increases to ~90% by pooling the data from 3-4 flow cells together. This is equivalent to the results obtained from other available high-coverage genomes (Fig. 3A) and it means that the whole set of inversions could likely be analyzed in a single PromethION flow cell.

As seen in Fig. 3C, one of the main limitations is the size of inversion breakpoint IRs, with genotyping success decreasing quickly for IRs over 50 kb. Similarly, genotyping success of the samples increases as the reads get longer (higher N50), although good genotyping can also be obtained with lower N50 and more coverage, since there are still enough long reads to genotype most inversions (Fig. 3B). In addition, according to our results, using fresh cells is important to

extract the longest DNA molecules and increase N50 (Fig. 1A). However, there is a well-known trade-off between read length and sequencing output with ONT, so that higher N50 values do not necessarily mean more genotyped inversions, as happens for the samples extracted using the phenol protocol (Fig. 3C). Thus, ONT inversion genotyping depends on a balance between several factors with opposing effects, such as read length, number of sequences and sequence selection affecting pore survival.

Accurate genotypes of SVs in multiple individuals are crucial to determine their functional effects and their association with phenotypic traits and disease susceptibility. By extending the current work to the increasing amount of available ONT genome data (Gustafson et al. 2024; Schloissnig et al. 2025), it would be possible to investigate a diverse set of inversions in a large number of individuals. These genotypes will help to improve inversion imputation in functional and phenotypic data, especially for recurrent inversions, of which there is very limited information (Giner-Delgado et al. 2019; Puig et al. 2020; Lerga-Jaso et al. in prep.; Porubsky et al. 2022). Also, the additional genotypes will provide a better estimate of inversion frequency and distribution across human populations. It is important to mention that a small fraction of inversions with extremely long IRs (>150-200 kb) cannot be detected with long reads and require more specialized techniques like Strand-seq (Sanders et al. 2016; Porubsky et al. 2022). In addition, as illustrated by the Fig. 5 examples and other recent studies (Ebert et al. 2021; Porubsky et al. 2022; Schuy et al. 2022; Bolognini et al. 2024; Logsdon et al. 2025) there is an increasing

number of complex genomic regions with multiple structural haplotypes, including different inversions, that need to be investigated in more detail. Nevertheless, ONT sequencing improves extraordinarily current inversion analysis and opens the door to the full characterization of one elusive type of human genetic variant for the first time. Due to the potential of the technology, it is expected that continued advances and method standardization would likely improve sequence accuracy, output and targeting. This will likely result in reduced costs and increase its widespread use to analyze inversions and other complex variants in basic and translational research (Dixon et al. 2023; Xu et al. 2023; Miller et al. 2021; Stevanovski et al. 2022) contributing to a more complete picture of human genetic variation.

METHODS

Human samples

Lymphoblastoid cell lines (LCLs) of four unrelated females included in the 1KGP were obtained from the Coriell Cell Repository and they were cultured in T75 flasks at 37°C, 5% CO₂ in DMEM medium, supplemented with antibiotics, 2 mM L-glutamine (Gibco/Invitrogen) and 10% fetal serum (HyClone), until saturation. Cells were harvested and pellets were stored at -80°C or used directly for DNA isolation. Also, we used available ONT sequencing data from 50 additional individuals of diverse origins generated mainly by the T2T Consortium (1) (Nurk et al. 2022), the Genome in a Bottle Consortium (7) (Zook et al. 2016), the Human Structural Variation Consortium (11) (Porubsky et al. 2022) and the Human Pangenome Reference Consortium (31) (Liao et al. 2023)) (Supplemental Table S1). All

procedures that involved the use of human samples were approved by the Animal and Human Experimentation Ethics Committee (CEEAH) of the Universitat Autònoma de Barcelona.

Generation of the inversion dataset

To generate the dataset of candidate inversions, first we merged all 58 previously characterized inversions with IRs (Giner-Delgado et al. 2019; Puig et al. 2020; Lerga-Jaso et al. in prep.) and 1528 candidate inversions from five different studies: Levy-Sakin et al. (2019) (314), Chaisson et al. (2019) (320), Audano et al. (2019) (220), Ebert et al. (2021) (315), and Porubsky et al. (2022) (359). IRs at the predicted breakpoints, formed by SDs or other type of repetitive elements, were identified from the UCSC Genome Browser hg38 human assembly SD track (<https://genome.ucsc.edu>) or by BLAST (>90% identity) of the extended inverted region against itself. Those variants with both breakpoints located respectively within the same IR pair were considered to represent the same inversion. When several SD pairs overlapped the breakpoints, then different possible inversions were generated and tested. Second, we took all the inversion predictions generated by PEM of longer fragments (Korbel et al. 2007; Kidd et al. 2008) in the InvFEST database (Martínez-Fundichely et al. 2014) and removed those matching regions already identified in the previous step. The remaining predictions with inverted SDs or IRs at the breakpoints, assessed as before, were added to the dataset. In both cases, regions without IRs, containing complex repeat structures or large gene families, breakpoints within SD blocks >200 kb, or inverted regions <50 bp, were excluded. Third, a more limited set of

other IRs that could result in inversions were obtained by selecting inverted SDs of 1-50 kb of length and separated by less than 250 kb from the UCSC hg38 SD track. According to SD definition, these sequences have a minimum of 90% identity and cannot be formed entirely by other types of repeat sequences, such as transposable elements (like LINEs or SINEs) or simple repeats. Inverted SDs that had a reciprocal overlap larger than 80% were merged with BEDTools v2.30.0 (Quinlan and Hall 2010) and the coordinates of the merger were used. Only those regions not already included were added to the final inversion list. Finally, additional extra regions previously identified as assembly errors in the reference genome that had IRs at both breakpoints (Vicente-Salvador et al. 2017) were also added to the analysis.

DNA isolation and library preparation

High molecular weight genomic DNA was obtained by two different methods following the protocol specifications: (1) the Monarch HMW DNA Extraction Kit for Cells & Blood (New England Biolabs) starting from ~6 million cells; and (2) a phenol-based isolation protocol optimized to obtain ultra-long reads for ONT sequencing (Gong et al. 2019) starting from 20-30 million cells. DNA was resuspended in 200 to 760 µl of EEB buffer (10 mM Tris-HCl, 1 mM EDTA pH 9 and 0.5% Triton X-100). DNA concentration was measured shearing 10 µl by vortexing and using the Qubit Fluorometric Quantification and the AccuGreen Broad Range dsDNA Quantitation Kit reagents (Biotium). Sequencing genomic libraries were prepared with the Ultra-Long DNA Sequencing Kit V14 (ONT) following the manufacturer's instructions. DNA libraries were dissolved in 300 µl of Elution Buffer (10 mM Tris-HCl, pH 8.0) and kept

at 4 °C. During these steps, DNA samples were always handled with care using wide-bore tips, without being vortexed, shaken, or centrifuged at high speed to avoid DNA shearing.

ONT sequencing

Sequencing was performed using a MinION Mk1B or Mk1C device and R10.4.1 flow cells (ONT). Flow cells ran approximately for 3-5 days with flushes every 24 hours and each time 37 µl of library were loaded into the flow cell. The MinKNOW software (version 22.12.5 to 23.04.3) was used to run the sequencing assays, and high-accuracy (HAC) basecalling was performed with Guppy (version 6.4.6 to 6.5.7) (<https://nanoporetech.com/>). Real-time adaptive sampling was done using MinKNOW in a Windows 10 2009 computer with an AMD Phenom II X4 955 processor, 16 GB of RAM and a PNY Nvidia GTX 1080 GPU. Two different versions of a FASTA file with the target sequences were used: AS2, an initial set including only 132 candidate inversions regions; and AS3, the final set that contains 587 of the 612 inversions interrogated (Supplemental Table S1). Targeted regions in AS2 and AS3 files spanned, respectively, 40 or 70 kb of sequence at both sides of the IRs at inversion breakpoints, excluding the actual IRs to avoid generation of non-informative reads that do not cross the breakpoint. Average sequencing enrichment in adaptive sampling was obtained by the ratio of the coverage of targeted and non-targeted regions in each run. Supplemental Table S6 summarizes the information of all the sequencing runs generated in this work, which were

merged together in the final genotyping of each of these samples to maximize the number of inversions resolved.

Inversion genotyping

As explained above, inversions were identified by mapping a set of four specific probe sequences flanking the inversion breakpoints on ONT reads (Fig. 1A). Probe sequences were designed by a combination of manual analysis and a custom script that generates 500-bp sequences with a 250-bp overlap along the inversion region (extending up to 100 kb at each side of the external limits of the breakpoints). Manually and automatically generated sequences were used as query on BLASTN v.2.12.0 (Altschul et al. 1990) against the inversion region $\pm$ 500 kb in the hg38 and T2T reference genomes with parameters `-perc_identity 80` and `-qcov_hsp_perc 90`, and those closest to the breakpoint and showing a single hit were selected. Selected sequences were tested for inversion genotyping (see below) and they were manually adjusted when needed to avoid specificity problems caused by unexpected variation.

The inversion genotyping pipeline consisted of multiple steps represented in Supplemental Fig. S1 that use different computer programs and custom scripts. These programs and scripts, including the probe sequence design, have been joined together in a new package named GeONTipe (v1.0), which was developed with Snakemake v.7.32.4 (Köster and Rahmann 2012). Briefly, ONT reads with less than 5 kb of length and quality of 7 were filtered out using NanoFilt v.2.8.0 (De

Coster and Rademakers 2023) and total sequence output, number of reads and N50 of each sequencing run were obtained with NanoPlot v.1.40.2 (De Coster and Rademakers 2023). Next, filtered reads were mapped against the hg38 reference genome with minimap2 v.2.23-r1111 (Li 2021) using the following parameters: `ax map-ont -z 400,100 -r 100,1000 --secondary=no`. Unmapped and mapped reads with `MAPQ < 20` were removed from the alignment file using SAMtools v.1.14 (Danecek et al. 2021), whereas reads mapping to each inversion region (external inversion breakpoint coordinates ± 100 kb) were selected. Specific inversion probe sequences were then mapped to the reads of the corresponding region using BLASTN v.2.12.0 with parameters `-task megablast -outfmt "6 sstart send slen sstrand" -qcov_hsp_perc 90 -sorthits 4`, and the inversion allele of the individual reads was determined from the relative order, orientation and distance of the probe sequences at both sides of the breakpoint. To ensure that genotypes for each inversion and sample are as reliable as possible, they were defined by the number of informative reads supporting each orientation as: (1) homozygote, if one orientation is supported by ≥ 5 reads for autosomes and Chr. X in females (corresponding to a binomial p-value of being heterozygote of less than 0.031), which is reduced to ≥ 2 reads for haploid Chr. X and Y in males, and there are $< 5\%$ reads of the other orientation (which allows for the presence of 1-2 inconsistent minority reads in inversions with high coverage); (2) heterozygote, when there are ≥ 1 O1 and O2 reads and they represent at least 15% of the total reads, to avoid incorrect genotype calls based on a few spurious reads; and (3) low confident, when there are insufficient reads or a small proportion of reads (5-15%) of one

orientation, which could correspond to different types of errors or other rearrangements that are harder to characterize. Moreover, for low confident cases with <5 reads supporting the same inversion orientation, we analyzed known SNPs to test if both chromosomes are already represented in these few reads and recover additional genotypes. Biallelic SNPs within the inversion region were selected from dbSNP (Sherry et al. 2001) and those with <15% frequency or identified as homozygotes in the target samples in the 1KGP data were filtered out (Byrsk-Bishop et al. 2022). SNPs present in informative reads with alleles concordant to those expected were used to calculate a distance between reads (value between 0 and 1) and create a dendrogram with the Hclust function of the R 4.1.2 Stats package (R Core Team 2017). Read clusters showing a distance higher than 0.5 were considered as different haplotypes belonging to an homozygote for the supported orientation.

Additional SVs of more than 200 bp in the inversion regions were detected by a >5% discrepancy between the expected and the observed distance among the normal combinations of the four probe sequences (AB, CD, AC, BD, BC, and AD) in >15% of reads with the same orientation (Supplemental Table S4). The distance difference was then used to classify the SV as either an insertion or deletion and to determine its size. The number of independent additional SVs was determined by removing the variants present in all O1 or O2 chromosomes (which could have been generated in the same mutational event than the inversion or represent possible errors in the reference genome or in the generated inverted reference) and

duplicated SVs with the same size detected separately in O1 and O2 chromosomes in the same location (including insertions and deletions exchanging positions between O1 and O2 chromosomes that can be moved by the inversion). Multiple insertions and deletions located in the same region with sizes varying by a similar amount were considered CNVs and counted as a single SV. Cases with no internal B and C sequences and a size difference between A and D sequences compatible with a deletion in >15 % of the reads were classified as AD deletions (deletion of the entire inversion region) and one of the alleles of the individual was labelled as Del. Finally, when probe sequences showed incompatible orientation conformations, the reads were classified as errors, suggesting the presence of potential new haplotypes that require further analysis.

Genotype comparison

Comparison of ONT inversion genotypes with those already available was done based on the inversions from the merged dataset detected in each of the previous studies depending on whether the information available included actual genotypes (Porubsky et al. 2022; Chaisson et al. 2019; Lerga-Jaso et al. in prep.) or just the presence or absence of the inversions (Levy-Sakin et al. 2019; Audano et al. 2019). Cases when several inversions in a study are associated with the same one in the ONT data were not taken into account. Analysis of the draft human pangenome reference samples with available paths involved the conversion of the pggp draft (Liao et al. 2023) into sequences. Hg38 coordinates were used as input to select paths and nodes with Odgi v0.8.3 extract function (Guarracino et al. 2022) and odgi

paths with parameter -f was applied to convert the odgi into FASTA format. The obtained output was used to run the inversion genotyping pipeline with default parameters, considering that there is only one sequence per haplotype.

DATA ACCESS

All sequencing data generated in this study have been submitted to the NCBI Sequence Read Archive (SRA) (<https://www.ncbi.nlm.nih.gov/sra>) under accession number PRJNA1271287. The code of the GeONTIpe inversion genotyping package is available in GitHub (<https://github.com/caceres-lab/GeONTIpe>) and as Supplemental code (distributed under the MIT License). Inversion genotypes are available in the Supplemental Material and they have been deposited in dbVar (accession number nstd248).

COMPETING INTEREST STATEMENT

The authors declare no competing interests.

ACKNOWLEDGEMENTS

We thank Jim Thomas, Olga Francino, Michael Vieyra and the ONT Technical Support team and multiple ONT users for advice and technical support about DNA extraction and nanopore sequencing, Carles Acosta for help with the

implementation of the genotyping pipeline in a supercomputer environment, David Porubsky and Jan Korbel for suggestions about the development of the Snakemake package, Odei Blanco for individual inversions quality control information, and Adrià Mompart for thorough GeONTipe testing. The data processing and analysis tools have been developed, implemented and operated in collaboration with the Port d'Informació Científica (PIC) data center. PIC is maintained through a collaboration agreement between the Institut de Física d'Altes Energies (IFAE) and the Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT). This work was supported by research grants PID2019-107836RB-I00 and PID2022-137615OB-I00 to MC and PID2023-146193OB-I00 to MAS funded by the Agencia Estatal de Investigación of the Ministerio de Ciencia, Innovación y Universidades (MICIU/AEI/10.13039/501100011033, Spain) and the European Regional Development Fund (ERDF, EU), 2021 SGR 00526 support to research groups from the Departament de Recerca i Universitats (Generalitat de Catalunya, Spain) to MC, the predoctoral program AGAUR-FI Joan Oró grants 2021FI-B 00240 to RMP and 2023 FI-1 00746 to MDR funded by the Departament de Recerca i Universitats (Generalitat de Catalunya, Spain), as well as the European Social Plus Fund, a Maria Zambrano Grant from the MICIU (Spain) funded by the European Union-NextGenerationEU to KK, and Predoctoral Contract PRE-2020-092440 to IY (MICIU/AEI/10.13039/501100011033, Spain). MP is a Serra Húnter Fellow.

AUTHOR CONTRIBUTIONS

MC conceived the project, devised the study and oversaw all the steps. RMP and IY developed the inversion genotyping pipeline with help from AS, JMU, MP, and MC. RMP, MP, MDR, and MC selected the inversion regions. KK, OC, MAS, and MP performed the ONT sequencing. RMP, KK, and MP analyzed the previously available and newly generated ONT sequencing data. RMP, KK, MP, and MC wrote the manuscript, and all the authors contributed comments to its final version.

REFERENCES

- Alkan C, Coe BP, Eichler EE. 2011. Genome structural variation discovery and genotyping. *Nat Rev Genet* **12**: 363–376.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215**: 403–410.
- Antonacci F, Kidd JM, Marques-Bonet T, Teague B, Ventura M, Girirajan S, Alkan C, Campbell CD, Vives L, Malig M, et al. 2010. A large and complex structural polymorphism at 16p12.1 underlies microdeletion disease risk. *Nat Genet* **42**: 745–750.
- Audano PA, Sulovari A, Graves-Lindsay TA, Cantsilieris S, Sorensen M, Welch AE, Dougherty ML, Nelson BJ, Shah A, Dutcher SK, et al. 2019. Characterizing the major structural variant alleles of the human genome. *Cell* **176**: 663-675.e19.
- Beyter D, Ingimundardottir H, Oddsson A, Eggertsson HP, Bjornsson E, Jonsson H, Atlason BA, Kristmundsdottir S, Mehringer S, Hardarson MT, et al. 2021. Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *Nat Genet* **53**: 779–786.
- Bolognini D, Halgren A, Lou RN, Raveane A, Rocha JL, Guarracino A, Soranzo N, Chin C-S, Garrison E, Sudmant PH. 2024. Recurrent evolution and selection shape structural diversity at the amylase locus. *Nature* **634**: 617–625.
- Byrska-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, Corvelo A, Clarke WE, Musunuri R, Nagulapalli K, et al. 2022. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**: 3426-3440.e19.

- Campoy E, Puig M, Yakymenko I, Lerga-Jaso J, Cáceres M. 2022. Genomic architecture and functional effects of potential human inversion supergenes. *Philos Trans R Soc B* **377**: 20210209.
- Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, Gardner EJ, Rodriguez OL, Guo L, Collins RL, et al. 2019. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat Commun* **10**: 1784.
- Choi J, Kim S, Kim J, Son HY, Yoo SK, Kim CU, Park YJ, Moon S, Cha B, Jeon MC, et al. 2023. A whole-genome reference panel of 14,393 individuals for East Asian populations accelerates discovery of rare functional variants. *Sci Adv* **9**: eadg6319.
- Choudhury A, Aron S, Botigué LR, Sengupta D, Botha G, Bensellak T, Wells G, Kumuthini J, Shriner D, Fakim YJ, et al. 2020. High-depth African genomes inform human migration and health. *Nature* **586**: 741–748.
- Collins RL, Brand H, Karczewski KJ, Zhao X, Alföldi J, Francioli LC, Khera A V., Lowther C, Gauthier LD, Wang H, et al. 2020. A structural variation reference for medical and population genetics. *Nature* **581**: 444–451.
- Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM. 2021. Twelve years of SAMtools and BCFtools. *Gigascience* **10**: 1–4.
- De Coster W, Rademakers R. 2023. NanoPack2: population-scale evaluation of long-read sequencing data. *Bioinformatics* **39**: btad311.
- De Coster W, Weissensteiner MH, Sedlazeck FJ. 2021. Towards population-scale

- long-read sequencing. *Nat Rev Genet* **22**: 572–587.
- Dixon K, Shen Y, O'Neill K, Mungall KL, Chan S, Bilobram S, Zhang W, Bezeau M, Sharma A, Fok A, et al. 2023. Defining the heterogeneity of unbalanced structural variation underlying breast cancer susceptibility by nanopore genome sequencing. *Eur J Hum Genet* **31**: 602–606.
- Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, Sulovari A, Ebler J, Zhou W, Mari RS, et al. 2021. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**: eabf7117.
- Giner-Delgado C, Villatoro S, Lerga-Jaso J, Gayà-Vidal M, Oliva M, Castellano D, Pantano L, Bitarello B, Izquierdo D, Noguera I, et al. 2019. Evolutionary and functional impact of common polymorphic inversions in the human genome. *Nat Commun* **10**: 4222.
- Gong L, Wong C-H, Idol J, Ngan CY, Wei C-L. 2019. Ultra-long read sequencing for whole genomic DNA analysis. *J Vis Exp* **145**: e58954.
- Grochowski CM, Bengtsson JD, Du H, Gandhi M, Lun MY, Mehaffey MG, Park K, Höps W, Benito E, Hasenfeld P, et al. 2024. Inverted triplications formed by iterative template switches generate structural variant diversity at genomic disorder loci. *Cell genomics* **4**: 100590.
- Guarracino A, Heumos S, Nahnsen S, Prins P, Garrison E. 2022. ODGI: understanding pangenome graphs. *Bioinformatics* **38**: 3319–3326.
- Gustafson JA, Gibson SB, Damaraju N, Zalusky MPG, Hoekzema K, Twesigomwe D, Yang L, Snead AA, Richmond PA, De Coster W, et al. 2024. High-coverage

- nanopore sequencing of samples from the 1000 Genomes Project to build a comprehensive catalog of human genetic variation. *Genome Res* **34**: 2061–2073.
- Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, et al. 2020. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**: 434–443.
- Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F, et al. 2008. Mapping and sequencing of structural variation from eight human genome. *Nature* **453**: 56–64.
- Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, et al. 2007. Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**: 420–426.
- Köster J, Rahmann S. 2012. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **28**: 2520–2522.
- Kovaka S, Fan Y, Ni B, Timp W, Schatz MC. 2021. Targeted nanopore sequencing by real-time mapping of raw electrical signal with UNCALLED. *Nat Biotechnol* **39**: 431–441.
- Lerga-Jaso J, Campoy E, Puig M, Yakimenko I, Gómez-Graciani R, Moreira-Pinhal R, Soos T, Vilella A, Ramírez C, Giner-Delgado C, et al. Towards a complete characterization of human polymorphic inversions and their functional effects. In preparation.
- Levy-Sakin M, Pastor S, Mostovoy Y, Li L, Leung AKY, McCaffrey J, Young E, Lam ET, Hastie AR, Wong KHY, et al. 2019. Genome maps across 26 human

- populations reveal population-specific patterns of structural variation. *Nat Commun* **10**: 1025.
- Li H. 2021. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37**: 4572–4574.
- Liao WW, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, Lu S, Lucas JK, Monlong J, Abel HJ, et al. 2023. A draft human pangenome reference. *Nature* **617**: 312–324.
- Logsdon GA, Ebert P, Audano PA, Loftus M, Porubsky D, Ebler J, Yilmaz F, Hallast P, Prodanov T, Yoo D, et al. 2025. Complex genetic variation in nearly complete human genomes. *Nature* **644**: 430–441.
- Logsdon GA, Vollger MR, Eichler EE. 2020. Long-read human genome sequencing and its applications. *Nat Rev Genet* **21**: 597–614.
- Martínez-Fundichely A, Casillas S, Egea R, Ràmia M, Barbadilla A, Pantano L, Puig M, Cáceres M. 2014. InvFEST, a database integrating information of polymorphic inversions in the human genome. *Nucleic Acids Res* **42**: D1027–32.
- Mastrorosa FK, Miller DE, Eichler EE. 2023. Applications of long-read sequencing to Mendelian genetics. *Genome Med* **15**: 42.
- Miller DE, Sulovari A, Wang T, Loucks H, Hoekzema K, Munson KM, Lewis AP, Fuerte EPA, Paschal CR, Walsh T, et al. 2021. Targeted long-read sequencing identifies missing disease-causing variation. *Am J Hum Genet* **108**: 1436–1449.
- Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze A V., Mikheenko A, Vollger MR, Altomose N, Uralsky L, Gershman A, et al. 2022. The complete sequence of a human genome. *Science* **376**: 44–53.

- Payne A, Holmes N, Clarke T, Munro R, Debebe BJ, Loose M. 2021. Readfish enables targeted nanopore sequencing of gigabase-sized genomes. *Nat Biotechnol* **29**: 442–450.
- Porubsky D, Höps W, Ashraf H, Hsieh PH, Rodriguez-Martin B, Yilmaz F, Ebler J, Hallast P, Maria Maggiolini FA, Harvey WT, et al. 2022. Recurrent inversion polymorphisms in humans associate with genetic instability and genomic disorders. *Cell* **185**: 1986–2005.e26.
- Puig M, Casillas S, Villatoro S, Cáceres M. 2015. Human inversions and their functional consequences. *Brief Funct Genomics* **14**: 369–379.
- Puig M, Lerga-Jaso J, Giner-Delgado C, Pacheco S, Izquierdo D, Delprat A, Gayà-Vidal M, Regan JF, Karlin-Neumann G, Cáceres M. 2020. Determining the impact of uncharacterized inversions in the human genome by droplet digital PCR. *Genome Res* **30**: 724–735.
- Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.
- R Core Team. 2017. R: a language and environment for statistical computing. *R Found Stat Comput*.
- Sanders AD, Hills M, Porubský D, Guryev V, Falconer E, Lansdorp PM. 2016. Characterizing polymorphic inversions in human genomes by single cell sequencing. *Genome Res* **26**: 1575–1587.
- Schloissnig S, Pani S, Ebler J, Hain C, Tsapalou V, Söylev A, Hüther P, Ashraf H, Prodanov T, Asparuhova M, et al. 2025. Structural variation in 1,019 diverse humans based on long-read sequencing. *Nature* **644**: 442–452.

- Schuy J, Grochowski CM, Carvalho CMB, Lindstrand A. 2022. Complex genomic rearrangements: an underestimated cause of rare diseases. *Trends Genet* **38**: 1134–1146.
- Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. DbSNP: The NCBI database of genetic variation. *Nucleic Acids Res* **29**: 308–311.
- Stevanovski I, Chintalaphani SR, Gamaarachchi H, Ferguson JM, Pineda SS, Scriba CK, Tchan M, Fung V, Ng K, Cortese A, et al. 2022. Comprehensive genetic diagnosis of tandem repeat expansion disorders with programmable targeted nanopore sequencing. *Sci Adv* **8**: 17.
- Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Hsi-Yang Fritz M, et al. 2015. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**: 75–81.
- The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature* **526**: 68–74.
- Vicente-Salvador D, Puig M, Gayà-Vidal M, Pacheco S, Giner-Delgado C, Noguera I, Izquierdo D, Martínez-Fundichely A, Ruiz-Herrera A, Estivill X, et al. 2017. Detailed analysis of inversions predicted between two human genomes: errors, real polymorphisms, and their origin and population distribution. *Hum Mol Genet* **26**: 567–581.
- Wagner J, Olson ND, Harris L, McDaniel J, Cheng H, Fungtammasan A, Hwang Y-C, Gupta R, Wenger AM, Rowell WJ, et al. 2022. Curated variation benchmarks for challenging medically relevant autosomal genes. *Nat Biotechnol* **40**: 672–680.

- Wellenreuther M, Bernatchez L. 2018. Eco-evolutionary genomics of chromosomal inversions. *Trends Ecol Evol* **33**: 427–440.
- Wu Z, Jiang Z, Li T, Xie C, Zhao L, Yang J, Ouyang S, Liu Y, Li T, Xie Z. 2021. Structural variants in the Chinese population and their impact on phenotypes, diseases and population adaptation. *Nat Commun* **12**: 1–12.
- Xu L, Wang X, Lu X, Liang F, Liu Z, Zhang H, Li X, Tian S, Wang L, Wang Z. 2023. Long-read sequencing identifies novel structural variations in colorectal cancer. *PLoS Genet* **19**: e1010514.
- Zook JM, Catoe D, McDaniel J, Vang L, Spies N, Sidow A, Weng Z, Liu Y, Mason CE, Alexander N, et al. 2016. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data* **3**: 1–26.

3.2 Application of GeONTIpe to other datasets

In the first section, the development of GeONTIpe and its initial evaluation on a dataset comprising 612 inversions and 54 individuals were described. As part of my research stay at the EMBL, we decided to apply this tool to genotype inversions to a larger dataset that was being generated at the time. This dataset included 1,019 samples from the 1000 Genomes Project (1000GP) and a set of inversions and inverted duplications previously detected using Sniffles2 and Delly. The initial validation and classification of these variants had been performed using dot plots, which confirmed the presence of each variant in a specific read and individual. Based on this preliminary validation, I was provided with a BED file containing two breakpoint coordinates for each variant, along with the variant type and the sample in which it had been confirmed by dot plot inspection. The file contained a total of 558 variants, including simple inversions, inversions mediated by repetitive elements, and several complex rearrangements that could involve sequence deletions and other more intricate SVs (Table 3.1).

Table 3.1: **Overview of inversion candidate types.** Distribution of inversion candidate classes provided across the dataset. Counts indicate the number of inversion candidates assigned to each structural class.

Variant class	Count
Simple inversion	221
Inversion and deletion	170
Short repeat-mediated inversion	50
Inverted duplication	38
Complex inversion	20
SD-mediated inversion	40
Paired deletion inversion	12
Two serial inversions	5
Cut and paste inversion	2

From these 558 variants, we removed those classified as inverted duplications, since the GeONTIpe package is not designed to analyse this type of SVs. After filtering, 520 inversions of different categories remained for downstream analysis. The first step was to define the putative BPs for these inversions, since GeONTIpe requires

four coordinates to operate correctly. Given the uncertainty regarding the correct breakpoint coordinates, we extended each provided position by 50 bp on both sides, generating two potential 100 bp breakpoint regions for each variant.

Once these coordinates were established, we proceeded to design the probes necessary for inversion genotyping. Several rounds of probe design were performed, modifying both the percentage of required unique sequence and the probe length, as many inversions were extremely small. Through this iterative optimisation, we successfully designed probes for 407 of the 520 variants, corresponding to a success rate of 78.26%.

With the probe set finalised, we first genotyped the variants in the sample in which each inversion was initially confirmed by a dot plot. A probe was considered valid when the corresponding sample showed evidence of the inverted allele on at least one read. Although some comparisons between the dot plot inspection and GeONTIpe results were not concordant, we proceeded to genotype all samples to determine whether the inversion identified by the dot plot might be present in other additional individuals.

In this inversion dataset, GeONTIpe has a mean genotyping performance of 95% per inversion, indicating that most inversions could be successfully genotyped in the majority individuals. In addition, only four variants exhibited performance below 50%. In particular, three of these variants were annotated as being mediated by segmental duplications (SDs). In such cases, an imprecise definition of BPs can negatively affect the genotyping process for these variants.

Using this approach, 393 inversions showed evidence of both alleles across the 1,019 available samples, supporting their validity. Of these, 130 variants exhibited a minor allele frequency (MAF) of at least 1%. Additionally, for 36 variants, the major allele corresponded to the orientation opposite to the GRCh38 reference. The remaining 14 inversions were not validated for several reasons. In 5 cases, all individuals carried the inverted allele, potentially indicating an assembly error in the reference genome. In 4 cases, although the inverted allele was detected by GeONTIpe, the genotype

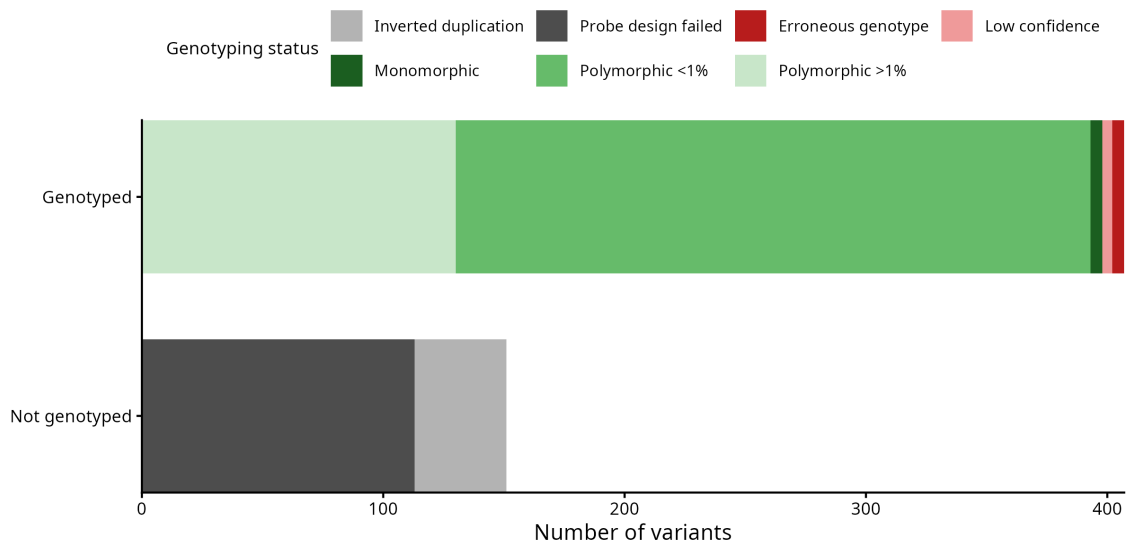


Figure 3.1: **Distribution of variants according to genotyping outcome.** Stacked bar chart showing the distribution of variants by genotyping outcome. Genotyped variants are subdivided into polymorphic ($>1\%$ or $<1\%$), monomorphic, low-confidence, and erroneous calls. Non-genotyped probes include inverted duplications and failed probe designs. Colors reflect these categories (grey = non-genotyped, green = correct calls, red = problematic calls).

was classified as low confidence due to the low proportion of reads in which this allele was observed. In addition, no additional samples showed evidence of the inversion. These four cases exhibited genotyping performances above 97% and were distributed as follows: in two of them, two samples were classified as low confidence genotypes, while the remaining samples were homozygous for the standard orientation. In the other two cases, only a single low confidence genotype was observed, and all remaining samples were classified as standard. This pattern suggests that even if these cases were conservatively interpreted as heterozygous genotypes, they would likely correspond to singletons or inversions with extremely low allele frequency (AF) (Figure 3.1).

Finally, in the remaining five cases, even though all of them presented a performance of at least 94%, the only genotypes different from the standard orientation consistently indicated errors. These corresponded to instances in which a substantial proportion of reads within a sample exhibited probe patterns that were inconsistent with the changes expected from the presence of an inversion. For these variants, a

manual inspection of the selected reads was performed.

In some instances, the issue originated from an incorrect dot plot classification, where an inversion was mistaken for an inverted duplication. In other cases, the region exhibited greater complexity than initially expected, and the reads suggesting the inverted allele were unreliable, ultimately invalidating the genotypes. Therefore, in the end, a total of 407 inversions could be analysed in the available samples, finding 393 polymorphic inversions.

All work described in this section of the Results is included in the article entitled “Structural variation in 1,019 diverse humans based on long-read sequencing”, which is provided as Appendix A of this thesis.

3.3 Generation of a complete inversion catalogue

Once the development of GeONTipe as a tool for inversion genotyping was completed, it was applied to a couple of inversion datasets. As a next goal, we aim to generate the most comprehensive catalogue of inversions to date. The objective was to identify and genotype all types of inversions with a MAF of at least 1%. This catalogue was generated by integrating the numerous variants reported in 13 different studies (Table 2.1), which include the dataset of 612 inversion candidates used to create and benchmark GeONTipe.

In addition to this primary objective, we also wanted to establish the basis for a future release of InvFEST, which will incorporate both the information from the original studies and the results of the various analyses performed in this section.

3.3.1 Merge of variants

The first step to build this catalogue was the creation of a harmonised table of all the studies containing, for each variant, the following fields: chromosome, four breakpoint-related coordinates, variant type, reported AF, samples carrying the variant, and a unified identifier following a scheme similar to that used in InvFEST.

One of the most challenging aspects of constructing this table was obtaining all four breakpoint coordinates, since most studies report only two per variant. Because the exact precision of the different technologies used to generate these variant calls was unknown, we defined a ± 5 kb interval around each reported coordinate, as previously done for InvFEST. Two exceptions were the InvFEST dataset and Campoy (2024), both of which already provided all four breakpoint positions and could therefore be used directly.

The next step consisted of merging all inversion predictions, both within and between studies, that seem to correspond to the same inversion using their four-coordinate breakpoint intervals. Specifically, when two independently reported variants showed overlap intervals in both breakpoints (BPs), they were considered to represent the

same inversion and were merged accordingly.

This strategy was first applied independently to each study in order to identify and collapse redundant variants reported within the same dataset. Table 3.2 summarizes the merging procedure applied to each study, reporting the initial number of variants, the number of merged variants, and the final set of independent (non-redundant) inversion candidates. In most studies, the procedure consisted solely of collapsing overlapping variants based on their BPs within the same study. However, additional processing steps were required for a subset of datasets, which includes whether AF calculation or LiftOver.

For Chaisson et al. (2019), Levy-Sakin et al. (2019), Porubsky et al. (2023), Logsdon et al. (2025) and Schloissnig et al. (2025), AFs were not directly provided and therefore calculated from available genotype or presence/absence information prior to variant merging. Furthermore, studies based on the T2T-CHM13 assembly (Porubsky et al. (2023) and Schloissnig et al. (2025)) required the conversion of the coordinates to GRCh38 using LiftOver. In the case of Porubsky et al. (2023), only 25 of the 63 reported variants (39.7%) were successfully converted and no overlaps were detected between the converted variants. Although Schloissnig et al. (2025) 731 variants were correctly converted, including 558 variants analysed in the previous Results chapter of this thesis.

Among all datasets, InvFEST required the most extensive additional processing, since variants were originally reported in the NCBI36 assembly. First, breakpoint overlap was evaluated within the dataset, applying an additional reciprocal overlap threshold of 75% using the external coordinates of the BPs (90% for variants larger than 5 Mb), which reduced the number of variants from 1,819 to 1,594. The coordinates of the variants affected by the assembly gaps were then corrected and the variants located entirely within the gap regions were removed, retaining 1,592 variants. Then LiftOver was applied to convert the coordinates from NCBI36 to GRCh38, resulting in 1,500 successfully converted variants. Finally, breakpoint overlap was re-evaluated in GRCh38, yielding a final set of 1,493 independent, non-redundant variants.

Additionally, the complete set of 612 refined candidate inversions described in Section 3.1 of this thesis was incorporated into the final merged catalogue as another independent study.

After this initial round of redundancy filtering, the total number of variants was reduced from 21,530 to 20,321. Table 3.2 summarises the evolution of the number of variants and the processing steps applied to each study.

Table 3.2: Review of candidate merging by study. For each study, the number of variants at the initial stage is reported. An “X” indicates whether allele frequency (AF) calculation of the reported variants was required and if LiftOver was used to convert variant coordinates between reference genomes, the number of resulting variants is indicated. Finally, the number of merged variants is reported, together with the total number of non-redundant variants identified for each study.

Study	Initial	AF count	LiftOver	Merged variants	Final
1000GP	2986	-	-	219	2767
gnomAD	13064	-	-	398	12666
InvFEST	1819	-	1500	326	1493
Audano et al. (2019)	222	-	-	15	207
Chaisson et al. (2019)	443	X	-	107	336
Levy-Sakin et al. (2019)	380	X	-	48	332
Ebert et al. (2021)	316	-	-	10	306
Porubsky et al. (2022)	399	-	-	8	391
Porubsky et al. (2023)	63	X	25	0	25
Campoy (2024)	195	-	-	0	195
Schloissnig et al. (2025)	733	X	731	38	693
Logsdon et al. (2025)	298	X	-	0	298
Moreira-Pinhal et al. (2025)	612	-	-	0	612
Total	21530	-	-	1169	20321

Once this first merging step was completed, the same strategy was applied to combine the predictions between studies, merging candidate inversions only when both BPs were identical in the respective study predictions. As a result, the total number of variants was reduced from 20,321 to 15,605 (76.8%). Of these, 13,099 variants (83.94%) were reported in a single study. Among these, gnomAD contributed the largest number of study-specific variants, with a total of 10,782 variants not detected in any other dataset. This outcome was expected, since gnomAD includes the largest sample size and reports the highest number of variants overall. In contrast, Campoy (2024) did not report any variants unique to that study. This result

was also anticipated, as InvFEST served as the primary source for the variants. For the same reason, 1000GP, a previous version of which (1000GP Phase3) also contributed to InvFEST and was analysed in detail in Campoy (2024), exhibited a very high proportion of shared variants.

On average, individual studies shared 64% of their variants with at least one other study, although this proportion was highly heterogeneous between datasets. Campoy (2024) showed complete overlap, with 100% of its variants also reported in other studies, while gnomAD shared only 14.8% of its variants. For studies based on newer sequencing technologies, between 60% and 82% of the variants were shared among studies (Figure 3.2). This pattern may be explained by the frequent use of overlapping samples between studies and by the fact that Moreira-Pinhal et al. (2025) already compiled some of the variants of the other studies. In addition, this may also be due to the detection of common variants in multiple studies, especially for the 247 variants shared by more than four studies.

3.3.2 Selection of variants for further analysis

After examining how these variants were shared between studies, a set of inversion candidates was defined to be retained without further filtering. This subset included the variants analysed by Moreira-Pinhal et al. (2025), Campoy (2024), and Schloissnig et al. (2025) were included.

In the case of Moreira-Pinhal et al. (2025), although a substantial fraction of these variants did not show evidence of polymorphic inversions in the initial analysis, the availability of a well-established probe set motivated their continued evaluation in a larger number of samples. Similarly, all variants reported by Schloissnig et al. (2025) were also retained without filtering steps, as some probe sets used in the previous Results section were directly applicable to this second analysis. Furthermore, some variants reported in Campoy (2024) were also excluded from genotyping, with the exception of the 56 variants mediated by inverted repeats (IRs) also included in Moreira-Pinhal et al. (2025) and HsInv0573 (17q21.31). This was supported by the availability of reliable genotypes for this dataset, obtained experimentally or by

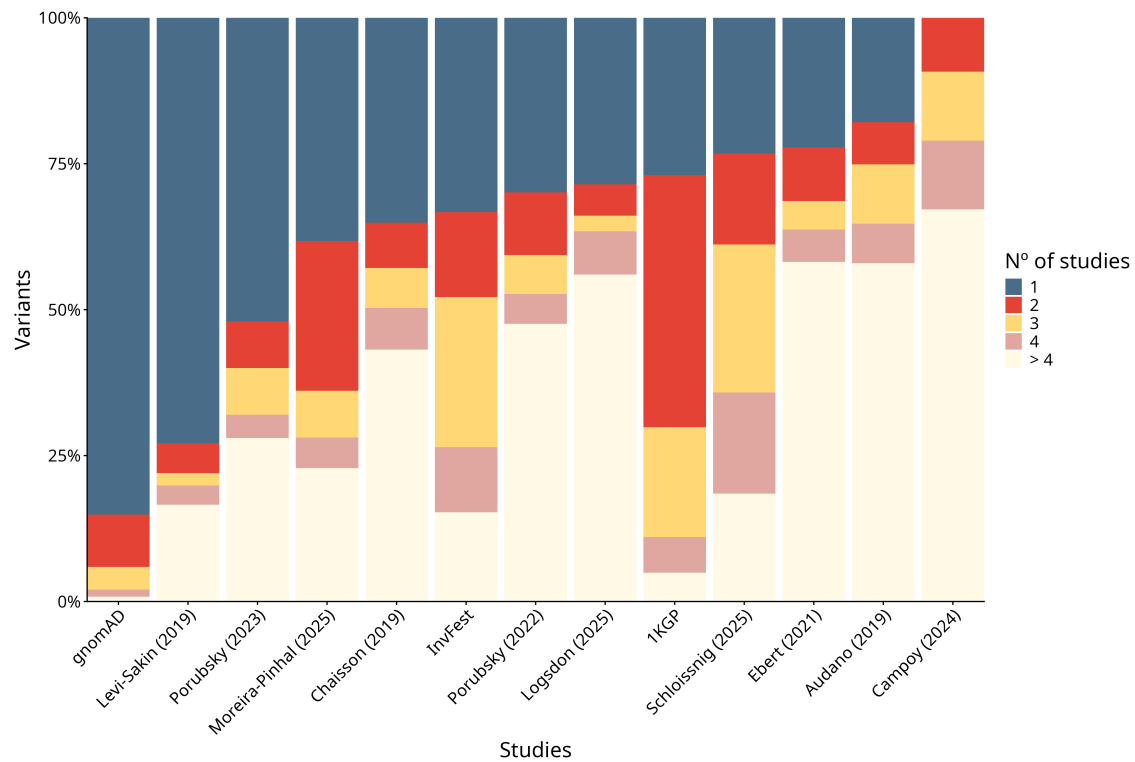


Figure 3.2: **Shared variants by study.** Bar plot showing the percentage of variants shared by each study, classified according to the number of studies in which they were detected, from study-specific variants to variants shared by one, two, three and up to more than four different additional studies.

BreakSeq genotyping in 3,202 individuals from 1000GP.

After excluding these three studies from the filtering procedure, a total of 14,273 variants remained. For this pool of 14,273 variants, the first applied exclusion criterion was MAF of more than 1% in at least one of the studies reported. According to this criterion, 1,631 variants were selected. A second exclusion parameter was the size of the variant. Variants larger than 10 Mb were discarded, as polymorphic inversions of this magnitude were considered unlikely to be genuine and more likely to represent detection artefacts. After applying this second exclusion criterion, a total of 982 putative inversions remained and were selected.

Before proceeding to the next stage of the analysis, we examined the behaviour of the 12,642 variants with a frequency below 1%. The study that contributed the largest number of low-frequency variants was gnomAD, with 11,414 variants (90.28%) identified in total. Of these, 10,098 variants were exclusively reported in

this dataset. This result was expected given the substantially larger sample size of this study compared to the others (Figure 3.3). The remaining variants (9.71%) came from other studies, in which variant frequencies were often not reported, either due to limited confidence in certain genotypes or due to missing data.

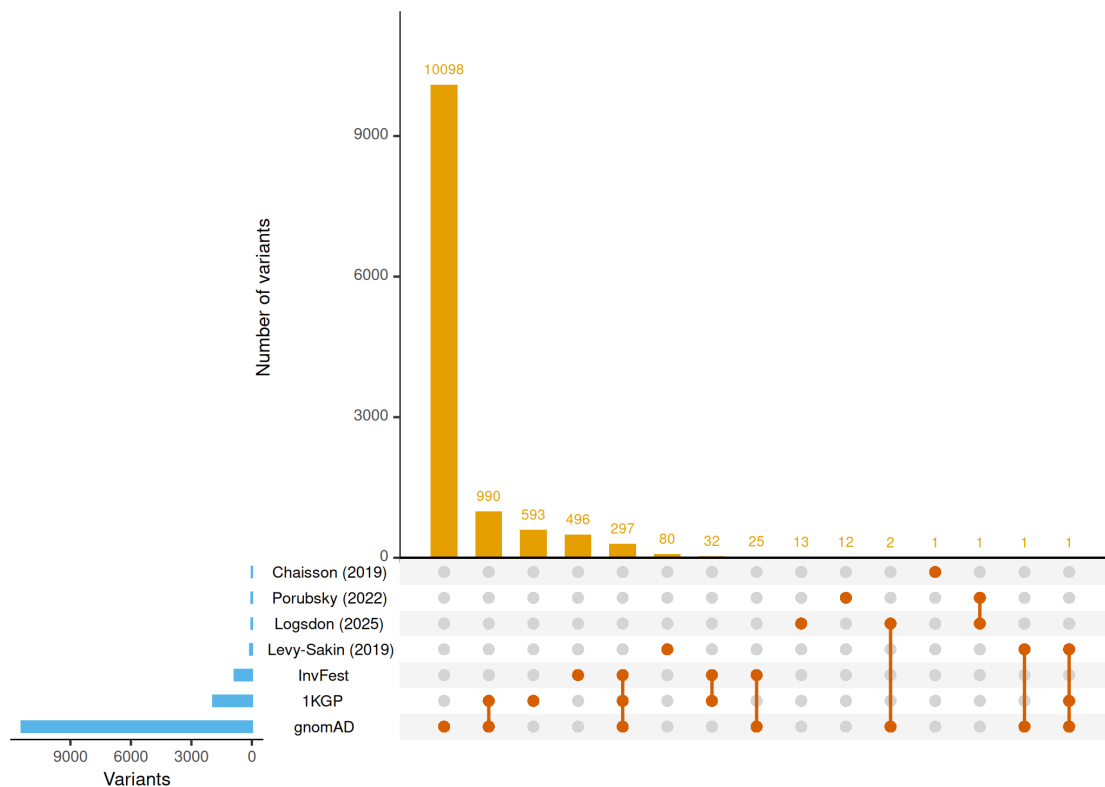


Figure 3.3: **Origin of variants with less than 1% allele frequency.** UpSet plot showing variants with an allele frequency below 1% across the different studies. Orange dots indicate the presence of variants in individual studies, and connecting lines denote variants shared among studies. The upper bars represent the number of variants in each combination, while the side bars indicate the total number of variants per study.

3.3.3 Breakpoint definition

For the correct functioning of GeONTIpe, precise inversion BPs is required. For variants reported by Moreira-Pinhal et al. (2025), this step was not necessary, since BPs had already been manually curated and accurately defined in that study. In contrast, variants reported by Schloissnig et al. (2025) lack this precise breakpoint resolution and therefore required further refinement.

Consequently, variants from Schloissnig et al. (2025) were included in the breakpoint definition step together with the 982 candidate inversions retained after the filtering procedure. This resulted in a final dataset comprising 1,564 variants for which breakpoint refinement was performed.

Once this final dataset was established, breakpoint coordinates were further refined. As an initial step, we explored the presence of segmental duplications (SDs) that overlaps with BPs derived directly from the merging of variants. When a predicted breakpoint overlapped with a pair of SDs in direct or inverted orientation, we verified whether the original coordinates of each contributing variant coincided with that pair. If this was not the case, the breakpoint was extended to also capture the original coordinate. Within this subset of variants, cases in which a single pair of SDs could be assigned unambiguously were defined directly as the BPs of that variant. In contrast, when multiple sets of SDs overlapped a given variant, these loci were classified as complex regions. This classification reflects the fact that multiple SDs, in direct or inverted orientation, can produce distinct structural configurations compatible with the detected variants. Such regions are particularly challenging for automated probe design.

Using this criterion, 38 variants with simple SDs defining their BPs were identified: 7 involving SDs in direct orientation and 31 in inverted orientation. Although SDs in direct orientation cannot generate inversions, they were retained as BPs, as they can lead to deletions or participate in rearrangements in which one duplication is inverted, effectively resulting in an inverted configuration.

In addition to these 38 variants with simple SDs, 25 regions containing multiple overlapping SDs were identified. For these regions, a breakpoint region was generated by merging the different SDs. Then, how many variants overlapped these large breakpoint regions was assessed. In total, 163 variants overlapped these regions, which are listed in Table 3.3. Several of these loci correspond to well-known inversion hotspots characterised by high structural complexity, such as the amylase locus. Given the substantial difficulty associated with the design of the probes in such regions and the low confidence expected for the resulting genotypes, most

of these 163 variants were excluded from the subsequent steps. Through this first approximation, BPs were defined for a total of 201 variants.

Table 3.3: **Complex SDs regions.** Summary of inversion breakpoint regions containing multiple segmental duplications (SDs). For each region (CR), the external coordinates are reported together with additional contextual information, including associated genes and whether the region was included in subsequent analyses. Finally, the table indicates the number of the 1,564 variants analysed that overlap each region and were therefore incorporated into the subsequent analyses.

Name	Coordinates	Inversion candidates	Included	Genes related	Additional information
CR1	chr1:12809207-13393767	7	Yes	Multiples genes	Characterized as complex region
CR2	chr1:16474442-16954362	10	No	-	Not reliable GRCh38 region
CR3	chr1:103589768-103774845	5	No	<i>AMY</i>	Characterized as complex region
CR4	chr1:120192772-120705592	7	No	<i>NBPF</i>	Characterized as complex region
CR5	chr1:143184587-143275986	5	No	-	Characterized as complex region
CR6	chr1:144044798-144915702	8	No	<i>NBPF15</i>	Related to extremely large SDs
CR7	chr1:146066805-149366538	9	No	<i>NBPF</i> and <i>NOTCH</i>	Characterized as complex region
CR8	chr10:46332689-47780368	11	No	<i>AGAP</i>	Characterized as complex region
CR9	chr11:4217829-4341822	5	Yes	<i>SSUL72L</i>	Characterized as complex region
CR10	chr15:23123714-28399999	11	No	<i>GOLGA</i>	Characterized as complex region
CR11	chr15:75253045-75299270	2	No	<i>GOLGA</i>	Characterized as complex region
CR12	chr16:31977208-33214595	12	No	-	Characterized as complex region
CR13	chr16:55776428-55820004	3	No	<i>CES1</i>	Palindromic SDs
CR14	chr17:21304469-22100583	8	No	<i>KCNJ</i>	Not reliable GRCh38 region
CR15	chr17:36222498-38250159	13	No	<i>TBC1D3</i>	Characterized as complex region
CR16	chr19:7018679-7052022	2	No	<i>MBD3L</i>	Palindromic SDs
CR17	chr2:109736854-110634615	3	No	<i>RGPD</i>	Related to extremely large SDs
CR18	chr4:49200000-49592471	3	No	-	Centromeric region
CR19	chr5:69635627-71364489	11	No	<i>SERF</i> and <i>SMN</i>	Not reliable GRCh38 region
CR20	chr7:72858572-75515341	6	No	Multiple genes	Related to extremely large SDs
CR21	chr8:7147340-8239446	7	No	<i>FAM90A</i>	Characterized as complex region
CR22	chr8:85640339-85829481	4	No	-	Not reliable GRCh38 region
CR23	chrY:6234811-9919592	6	No	<i>TSPY</i>	Related to extremely large SDs
CR24	chrY:21924887-22661521	3	No	<i>RBMY</i>	Related to extremely large SDs
CR25	chrY:23061402-23359290	3	No	<i>DAZ</i>	Related to extremely large SDs

For other variants that do not overlap SDs, information was incorporated from variants classified as complex in gnomAD, which typically involve deletions and duplications flanking the inversion. Using these annotations, BPs were defined to encompass the associated deletions and duplications. When a variant was supported by multiple studies, we verified that the coordinates reported in those studies were consistent with the gnomAD annotations. Using this approach, BPs were defined for 307 additional variants.

For the remaining variants (1056), an analogous approach to that described in Section 3.2 was incorporated. Specifically, a ± 100 bp window was defined around each set of coordinates in the original coordinates. When multiple variants were included within the same region, the minimum and maximum coordinates were used to define

the boundaries of the variants, and ± 100 bp was applied from these positions.

After the BPs were determined, a second round of variant filtering was performed. In this step, variants that were consistently annotated as inverted duplications in the different studies were excluded from further analysis. Although these variants could not be genotyped, they had not been filtered previously because refining their breakpoint coordinates could provide useful information to update InvFEST and enrich the variant catalogue. As observed in Table 3.3 most variants located within complex genomic regions were removed from the analysis. Moreover, only those variants with an internal region of at least 300 base pairs and for which at least one of the two BPs was less than 200 kb were retained. After applying these criteria, 1,082 variants remained of the initial set of 1,564.

Variants passing all filters were submitted to GeONTIpe to design specific 300 bp probes containing at least 25% unique sequence. Based on these criteria, GeONTIpe could generate probes for a total of 623 variants (57.56%). These were combined with 612 variants previously reported by Moreira-Pinhal et al. (2025), resulting in a final set of 1,235 variants that were genotyped across the exhaustive set of 1,409 samples with available ONT data (Section 2.1.2).

3.3.4 Inversion region genotyping

Once the genotypes of the regions analysed were obtained, the first step was to evaluate the performance of GeONTIpe for each sample (Figure 3.4). We observed that 503 samples achieved at least 90% genotyped inversion regions. On average, 86.37% of inversion regions could be successfully genotyped per sample, and only 14 samples showed a genotyping success below 50%.

The performance of the samples in the different datasets was then assessed. To do this, samples were classified into the following subgroups: HGSVC phase 3, HPRC, Gustafson et al. (2024), Schloissnig et al. (2025) and Multiple Sources (including samples not assigned to the remaining datasets and those sequenced by us). For samples present in several studies, performance values were assigned to all the cor-

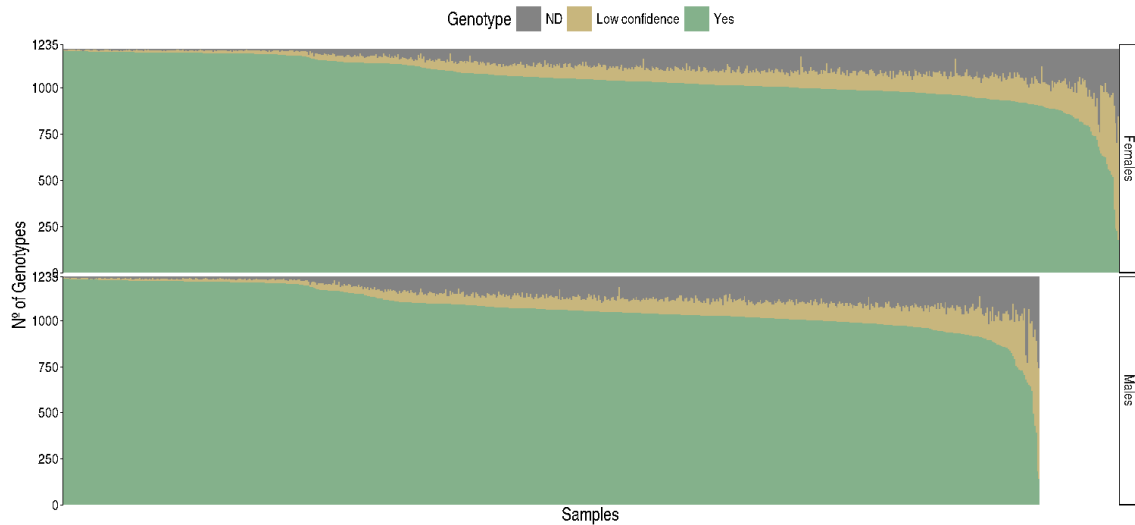


Figure 3.4: **Inversion region genotypes by sample.** Bar plot showing the number of genotypes obtained per sample. The maximum number of genotypes was 1,235 for male samples and 1,209 for female samples, after excluding inversions on the Y chromosome. Valid genotypes are shown in green, low confidence genotypes supported by a low number of reads and therefore excluded from downstream analyses are shown in mustard, and genotypes without any read are shown in grey.

responding datasets. The mean genotyping success and its distribution across the datasets are shown in Figure 3.5. Among the datasets, HGSVC phase 3 was the best-performing, achieving a mean of 98.75% of successfully genotyped inversions. Two datasets, Multiple sources and Schloissnig et al. (2025), exhibited mean genotyping success below 85%, and all samples with performance less than 50% were from the latter, as expected by the low sequencing coverage.

Next, we tried to identify sample-level variables that influence inversion genotyping success. Following the approach used for the dataset described in section 3.1, we examined N50 and read count as the main potential predictors. To do this, the percentage of genotyped regions against N50 and the number of reads for each sample was plotted (Figure 3.6-A). This analysis showed that samples with performance below 50% were also those with the lowest number of reads, while the N50 values were comparable to those of other samples, indicating that the read count was a limiting factor.

To isolate more clearly the effect of N50 we focused the analysis on samples with $\geq 95\%$ inversions genotyped and assessed whether the success was still driven primarily

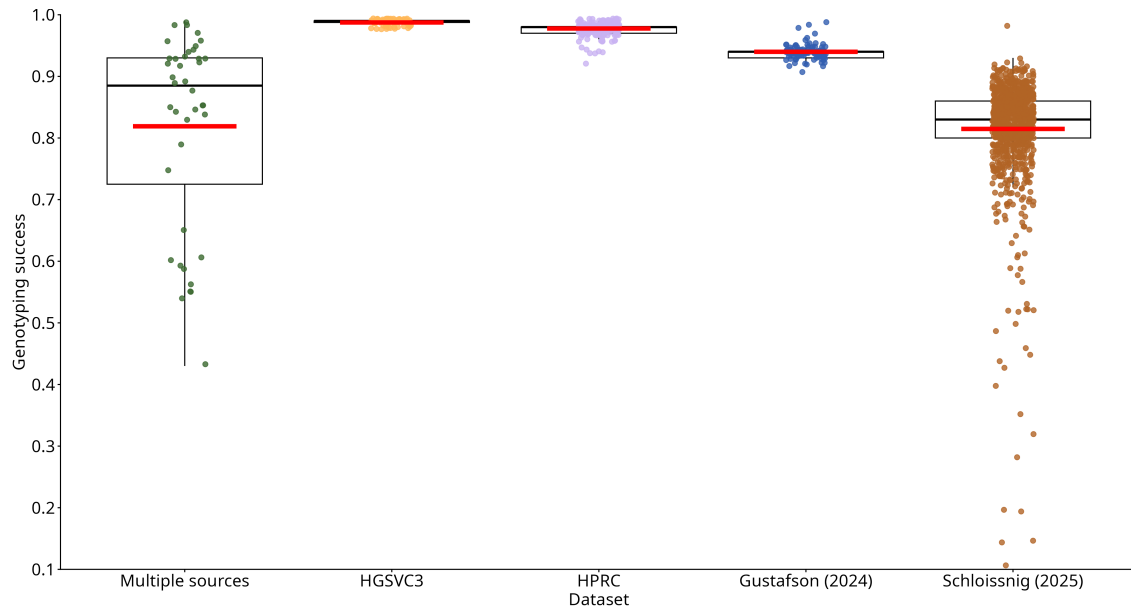


Figure 3.5: **Genotype success by dataset.** Boxplots describe the genotyping success of inversion regions across different datasets, including the median and interquartile range. Coloured points represent each individual sample, while the red line denotes the mean genotyping success within each dataset.

by the number of reads or instead became more dependent on N50. Most of the samples with performance $\geq 95\%$ had N50 values between 50 and 80 kb. In contrast, samples that achieved 99% performance had a considerably higher number of reads or, if they were lower, exhibited N50 values above 100 kb (Figure 3.6-B).

Then, the differences in N50 and the number of reads between the datasets were examined to assess whether they could explain the observed performance differences. A summary is provided in Table 3.4. In general, datasets with smaller N50 values tended to show reduced performance, despite having a similar or even higher number of reads, as can be observed by the dataset of Schloissnig et al. (2025). One paradigmatic case corresponds to the Multiple Source dataset, which had an N50 larger than expected given its genotyping success distribution. This may be explained by the relatively small number of samples and heterogeneity in this dataset, combined with the larger variation of observed performance that ranges from 43% to 99%. In addition, samples that given their number of reads should present a lower performance, but the use of adaptive sampling to enrich target regions during sequencing partially compensated for this effect.

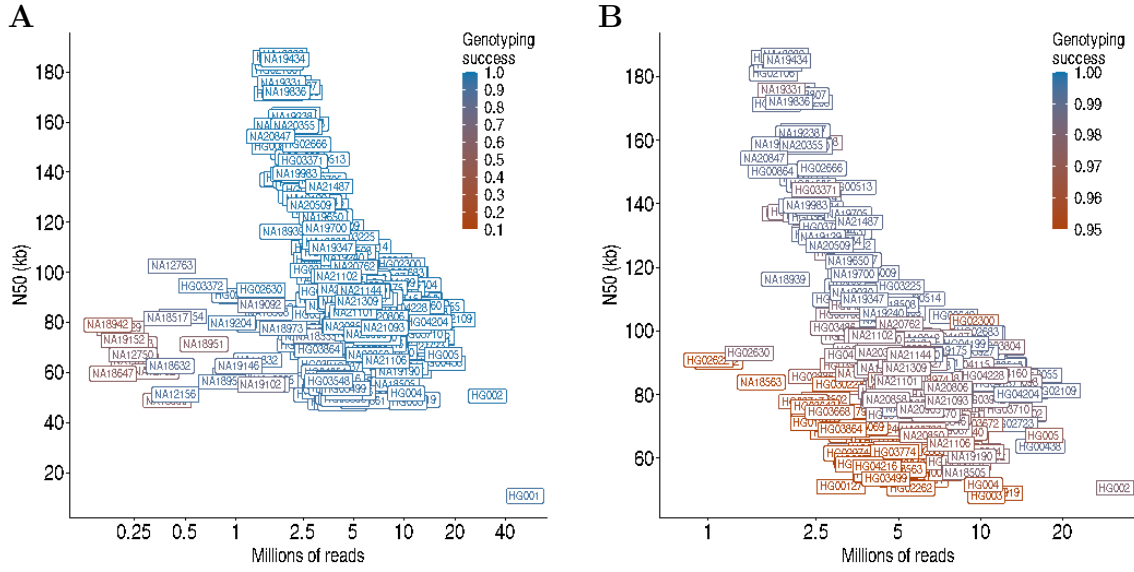


Figure 3.6: **Genotyping success by number of reads and N50.** Genotyping success indicated on a colour gradient from high (blue) to low (red), represented as a function of N50 (Y axis) and the number of reads per sample (X axis). (A) Overall performance across all samples, showing the effect of the number of reads on GeONTIpe genotyping ability. (B) Subset of samples with performance $\geq 95\%$, highlighting the effect of N50 on genotyping success.

Table 3.4: **N50 and number of reads by dataset.** Summary of average N50 and number of reads across datasets, with corresponding genotyping success.

Dataset	N50	N ^o of reads	Genotyping success (%)
Multiple sources	67720	3593154	81.9
HGSVC3	138249	2305098	98.75
HPRC	85251	5542246	97.78
Gustafson (2024)	62014	3468593	93.99
Schloissnig (2025)	21454	3098610	81.47

After examining the performance at the sample-level, the performance at the variant level was evaluated. Of all inversions, 1078 (87.29%) achieved at least 50% genotyping success and 909 (73.60%) more than 90%. Next, we assessed the effect of increasing BPs sizes on inversion genotyping. To do that, we transformed the breakpoint size into the probe distance and grouped the inversions by this distance, as shown in Table 3.5. The data showed that the distance of the probe had a substantial impact on genotyping: performance decreased dramatically once the breakpoint exceeded 10 kb, with a particularly severe decline above 50 kb. This relationship could be observed in Figure 3.7, where the probe distance was plotted against the

genotyping success. As noted, larger probe distances were associated with lower performance, and the decrease in genotyping success fits quite well a decreasing exponential curve ($R^2 = 0.88$, $P < 0.001$).

In Table 3.5 the length of the inversion had an effect on the genotyping performance was also evaluated. By contrast, taking into account the inversion size showed a more moderate reduction in genotyping performance as the inversion length increases.

Table 3.5: **Genotyping success by inversion.** Genotyping success per inversion, considering the distance between probes as a proxy for breakpoint size (upper table) and the inversion size, defined as the internal distance between the two BPs (lower table).

Probe distance (bp)		
Probe distance (bp)	Variants	Genotyping success (%)
<1000	467	97.69
1000-10000	482	94.84
10000-50000	187	63.66
>50000	99	20.68
Internal size (bp)		
Internal size (bp)	Variants	Genotyping success (%)
<1000	108	95.5
1000-10000	337	94.04
10000-50000	359	88.19
>50000	431	73.31

Moreover, we quantified how many polymorphic inversions were identified in this large dataset. Of the 1,235 inversions tested, 680 showed both the standard and inverted alleles, corresponding to 55.06% of the variants. Of these inversions, three showed a heterozygote genotype only in the offspring of duos and trios. Of these cases, two (Mos-15-30033302-30515994 and HsCan14840) showed extremely low genotyping success, likely due to the large distance between probes or the presence of additional haplotypes that could interfere with genotyping. The third inversion (HsCan7452) appeared to be a singleton. For all three variants, we compared the genotypes with related samples and no discordances were observed.

In addition to these 680 inversions that presented both alleles, eight showed only the inverted allele in all genotyped individuals. Six of them presented a genotyping success of at least 97%, suggesting possible errors in the reference genome. When

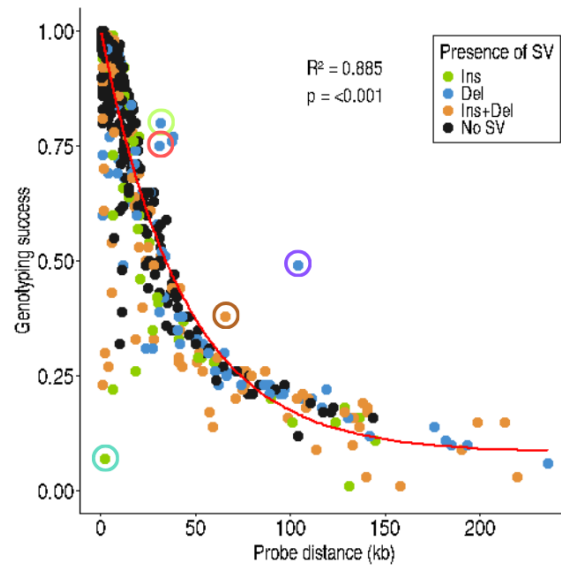


Figure 3.7: **Genotyping success according to distances between probes.** Genotyping success as a function of probe distance. Each point represents an inversion region, and colours indicate the presence of deletions (blue), insertions (green), or a combination of both (orange) in at least 1% of the genotyped samples. A red trend line is shown, for which the coefficient of determination (R^2) and the associated p-value are reported. Moreover six regions, marked in colour circles where analysed to verify the presence of these additional SVs.

the GRCh38 reference genome was checked, five of the eight presented an assembly fix that could correspond to the detected inverted allele. On the other hand, we found that 191 of the 680 variants (28.08%) had a minor allele frequency (MAF) of at least 1%. This low proportion of variants with a MAF greater than 1%, despite the fact that most variants have been selected using this threshold, can be explained by the contribution of studies based on small sample sizes, which likely led to an overestimation of the true AFs of these variants. Moreover, in 44 of these 191 inversions, the minoritarian allele corresponds to the orientation of the reference genome (GRCh38), which increases to 53 variants considering all inversions showing both alleles.

Beyond inversions, GeONTipe can also detect additional SVs by analysing the distances between probes (Section 3.1). In our dataset, 740 regions displayed at least one additional variant in at least one sample. Of these, 362 were found in more than 1% of the samples. As GeONTipe is not optimised for the detection of deletions and insertions, an extended analysis was not conducted. However, among these

362 variants, 156 corresponded to deletions, 78 to insertions, and the remaining corresponded to mixed events between samples involving increases and decreases in distances. Of the 740 inversions that present additional SVs, 232 were already present in Moreira-Pinhal (2025). The higher detection of deletions compared to insertions was expected, as probe distance reductions are generally easier to detect due to the genotyping strategy used.

In Figure 3.7 were also highlighted cases with additional SVs detected in at least 1% of the samples (Figure 3.7). Some points deviated markedly from the general trend commented on before, most of which corresponded to regions with additional SVs. Five of these regions (circled in Figure 3.7) presented additional SVs of substantial size and affected a large proportion of samples, and a summary is provided in Table 3.6.

Table 3.6: Features of additional SVs in inversion regions. Analysis of five regions for which additional SVs were detected and found to affect GeONTipe performance. For each variant, its position in Figure 3.7 is indicated by the corresponding circle colour. The table reports the detected variants, classified as deletions (Del) or insertions (Ins), their genomic location, the orientation in which the SVs are found, and whether the associated inversion is polymorphic, with the MAF and the corresponding orientation provided in parenthesis. Finally, the size of the detected variants and the percentage of samples in which they were observed are reported.

Variant ID	Circle colour	Variant type	Location	MAF (%)	Orientation	Variant size	Variant presence (%)
HsCan5223	Turquoise	Ins	BP2	3.7 (Std)	Std Inv	12000-242000	99.32
HsInv0231	Brown	Del	BP1	1.9 (Std)	Std	70000	47.11
		Ins	BP2		Inv	8000	1.9
Mos-12-9293902-9679442	Purple	Del	BP2	-	Std	96000	87.98
HsCan9082	Red	Del	BP2	-	Std	24000	85.52
HsInv1115	Green	Del	BP1	11.5 (Inv)	Std	3000	7.21
					Inv	37000	79.23

Among these cases, the HsCan5223 inversion stands out due to its very low genotyping success rate (7%). However, the small subset of samples carrying an extremely large additional insertion could not fully explain this poor performance. Probe analysis showed that inversion was genotyped using only the second breakpoint, since probe A could not be designed because the upstream region is enriched with multiple SDs. These SDs clusters likely reflect local structural instability, compromising probe specificity and genotyping accuracy.

The support for this interpretation comes from the nearby, partially overlapped variant HsCan5220, which showed a higher genotyping success (68%), but also showed multiple insertions and a minor allele corresponding to the reference orientation. This suggests that the reference genome may represent a low-frequency haplotype in this region, negatively impacting probe design.

In general, these observations indicate that incorrect assumptions about the predominant structural configuration can affect probe placement. In structurally complex regions, improved breakpoint resolution and haplotype-aware design are likely required to enhance genotyping performance and reduce artefacts.

3.3.5 Inversion characteristics analysis

Based on the analyses described above, we were able to describe the general features of the genotyping process, including factors that affected genotyping performance and several global characteristics of the dataset. However, these sections did not explore in detail the characteristics of the polymorphic inversions, which represent the most comprehensive inversion catalogue in humans so far.

When we compared the characteristics of this new dataset with previous studies by our group, such as Giner-Delgado et al. (2019), Puig et al. (2020), and Campoy (2024), we observed a substantial increase in the number of polymorphic inversions detected. The most extensive previous study, Campoy (2024), identified 134 polymorphic inversions, of which 57 are shared with the present dataset. Thus, our study multiplies the number of known polymorphic inversions five times. This improvement is largely attributable to the use of ONT long-reads, as well as to the integration of inversion calls from multiple recent studies that applied novel technologies for variant detection. Ideally, we would also have compared the inversions generated by NAHR versus NH mechanisms, since one of the main advantages of the use of long reads is the analysis of inversions mediated by IRs that cannot be easily genotyped with any other technique. However, due to the lack of reliable breakpoint annotation and origin mechanism classification, it is not possible to perform this comparison with sufficient precision.

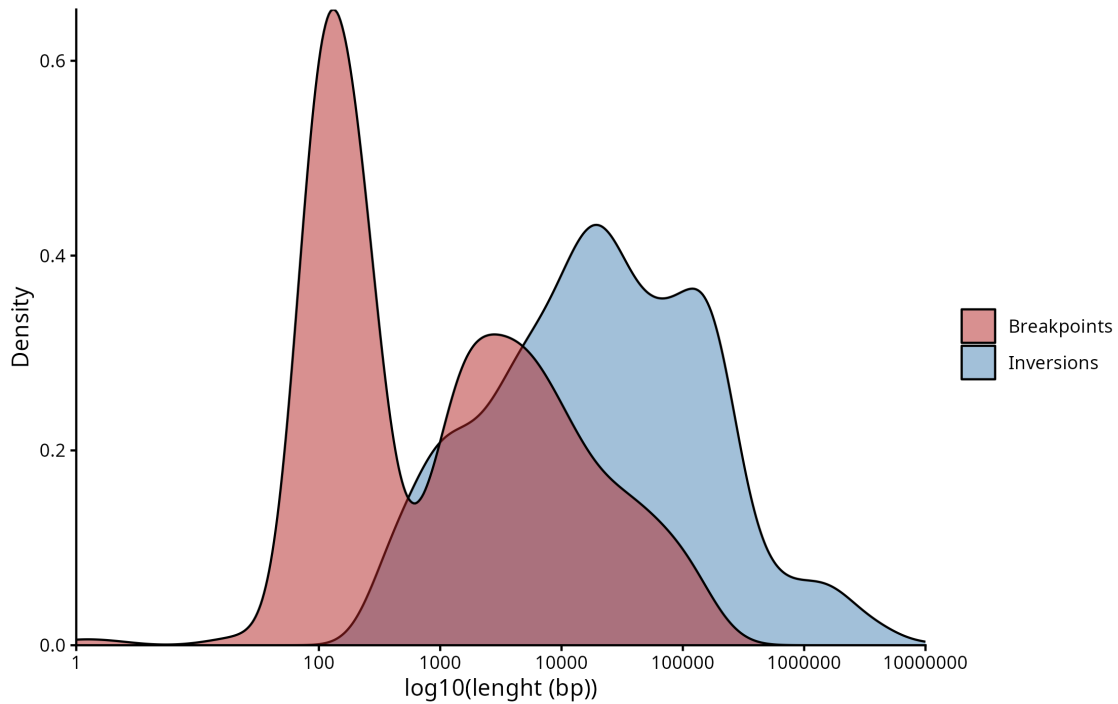


Figure 3.8: **Distribution of inversions and breakpoint sizes.** Density plot representing inversion (blue) and breakpoint sizes (red) on a $\log_{10}$ scale. Breakpoint sizes correspond to the smaller of the two breakpoints per inversion.

Then, the size distribution of the detected inversions was examined. In general, polymorphic inversions ranged in size from 301 bp (HsCan1073) to 4.4 Mb (HsInv1928), excluding breakpoint regions, with a mean inversion length of approximately 140 kb. The sizes of BPs also varied widely, from a few bases to repeat blocks of 200 kb. Of the 677 polymorphic inversions, 286 (42.24%) have less than 10 kb, 239 inversions (35.30%) ranged between 10 kb and 100 kb, 124 (18.31%) spanned 100 kb to 1 Mb, and 28 (4.13%) exceeded 1 Mb in length (Figure 3.8).

Next, we assessed the genomic distribution of the inversions along the chromosomes (Figure 3.9). Consistent with previous studies, chromosome X showed a particularly high number of inversions (56), making it the most enriched chromosome in our dataset. This enrichment is likely explained by the high density of SDs present on that particular chromosome. Other chromosomes with a large number of inversions included chromosome 1 (50), chromosome 2 (50), and chromosome 7 (53). At the opposite extreme, chromosome Y harboured only six inversions. This low number may reflect its poor characterisation rather than its true biological scarcity, as

previous studies have reported extensive large-scale structural variability on chromosome Y (Hallast et al., 2023), making it one of the most challenging chromosomes to analyse.

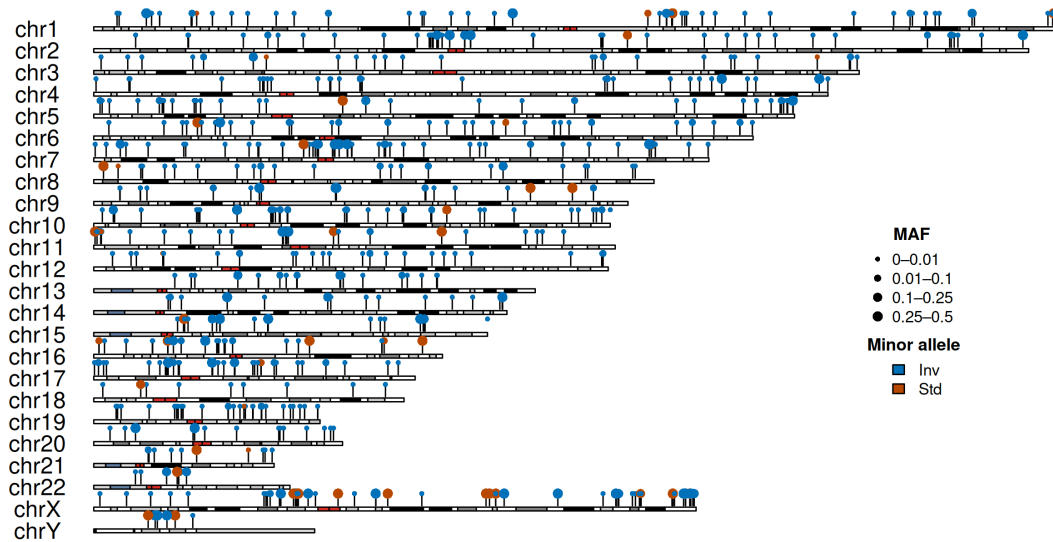


Figure 3.9: **Distribution of inversions across chromosomes.** Representation of the distribution of inversions across the genome, with different colours representing the minor allele and the circle size its frequency (MAF).

When normalising inversion counts by chromosome length, chromosome 19 emerged as the most enriched, with approximately one inversion per 1.5 Mb, followed by chromosome 17 (one per 2.44 Mb) and chromosome X (one per 2.78 Mb). To visualise these differences, the relationship between chromosome size and inversion count was plotted. The size of the chromosome accounted for approximately 44% of the variation in the inversion distribution ($R^2 = 0.443$, $p = 1.81 \times 10^{-16}$) (Figure 3.11). Among the chromosomes, 19, 17, X, and, in addition, chromosome 7 showed the largest deviations from expected counts based on the size of the chromosome.

However, these estimates should be interpreted with caution, because not all inversions in the dataset are not completely independent of each other. When enrichment was instead assessed based on the maximum proportion of bases potentially involved in inversions, chromosome 16 appeared to be the most enriched, with nearly 17.5% of its sequence located within inverted regions. This signal was largely driven by two overlapped inversions of 4.4 Mb which counted independently, inflating the estimated inverted fraction. Overall, treating each inversion as an independent entity,

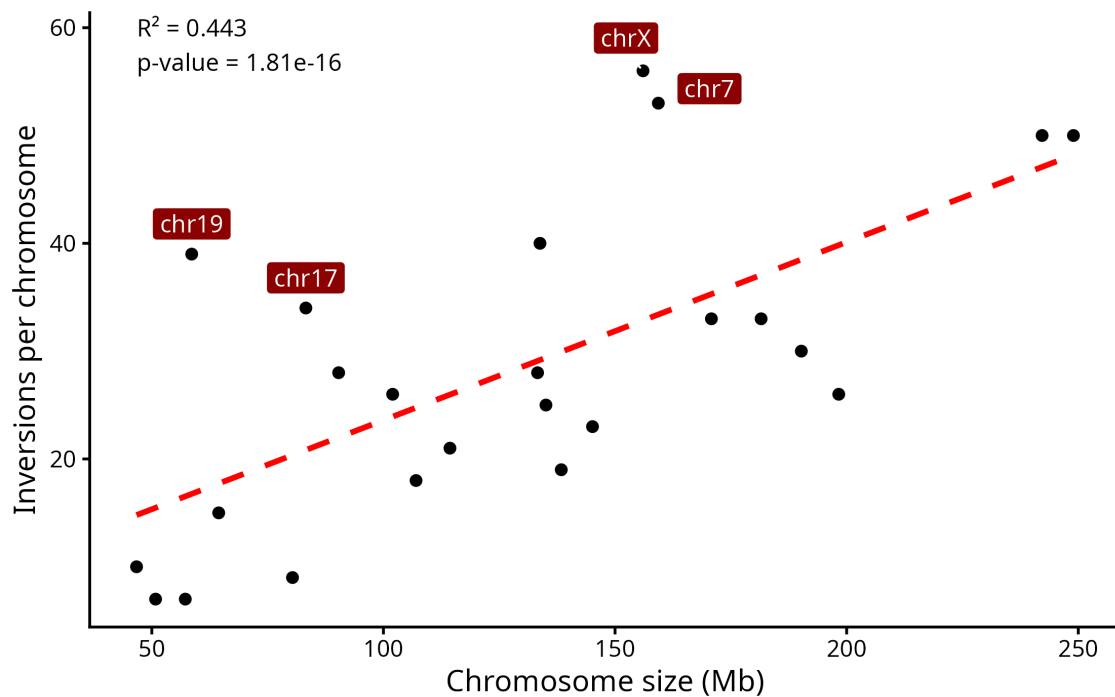


Figure 3.10: **Correlation between inversion distribution and chromosome size across the genome.** Relationship between chromosome size and inversion distribution, highlighting chromosomes with the largest deviations from the expected counts (red labels).

due to the lack of information about how many of them are truly independent, limits our ability to accurately quantify the chromosome-level inversion burden, although these metrics still provide a useful approximation.

3.3.6 Inversion frequency in human populations

As the next step, we focused on AF of the inversions, since it provides information on the distribution of inversions in human populations and can shed light on the action of natural selection within populations. As described previously, AF were calculated using only unrelated individuals, excluding offspring from trios and duos. The global frequency distribution of inversions is shown in Figures 3.9 and 3.13), while Figure 3.13 also presents frequencies stratified by ancestry. In our study, we defined MAF as the frequency of the rarest allele considering all the samples analysed from all populations together. In our case, that value ranged from 0.04% to 49.91%. When variants were grouped into 5% allele-frequency bins, the distribution was

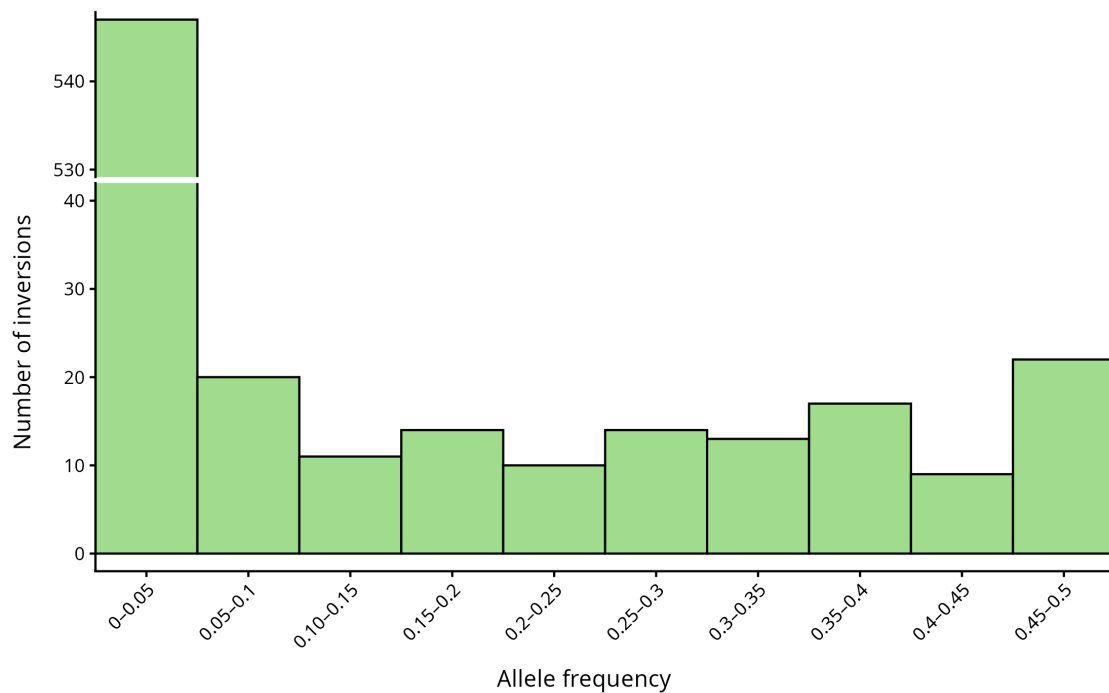


Figure 3.11: **Inversion allele frequency spectrum.** Green bars represent the global distribution of inversion MAF, grouped into 5% allele-frequency bins.

strongly skewed towards low frequencies. In total, 547 inversions (80.79%) showed a MAF below 5%, while a smaller enrichment was observed in the MAF bin 45-50% (3.24%).

Individuals of African ancestry exhibited the highest number of polymorphic inversions (406), while European populations showed the lowest number (273). This difference may be explained by two factors: the larger number of African individuals in the dataset and the greater genetic diversity typically observed in African populations. To disentangle these effects, we calculated the total number of inversion heterozygotes in the population, normalised by the number of successfully genotyped variants. As shown in Figure 3.12, African populations showed a higher proportion of inversion heterozygotes, consistent with increased genetic diversity, in agreement with previous reports (Schloissnig et al., 2025).

We then examined inversions that were exclusive to each ancestry. In Africans, 125 inversions were not detected in any other ancestry, although 106 of these had a MAF below 1% in that ancestry. An inversion that exceeded that 1% of frequency was

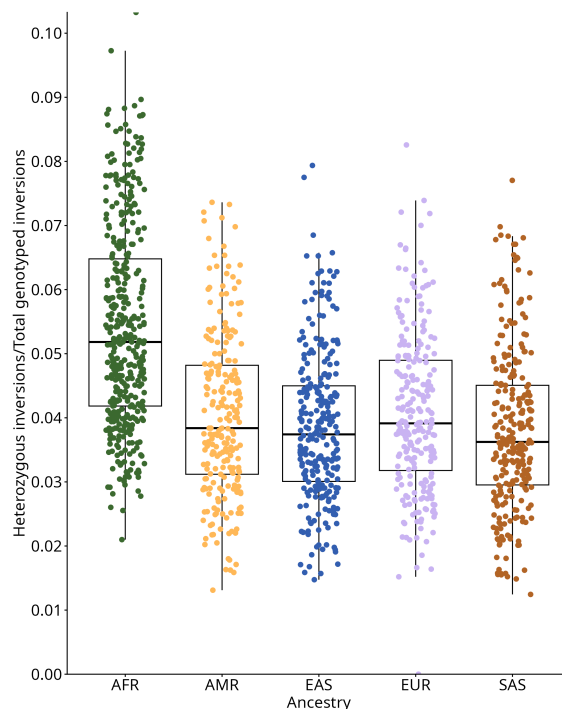


Figure 3.12: **Inversion heterozygote fraction by ancestry.** Boxplots represent the number of heterozygote genotypes divided by the total inversions genotypes in all samples, grouped by ancestry.

HsInv0365, which showed a global frequency of 8.11% and reached 11.54% within the African populations, which is unusually high for a variant absent elsewhere. However, this inversion had an overall genotyping success of only 1%, suggesting that its apparent uniqueness may be due to the low genotypes rather than true population specificity. Europeans presented 31 unique inversions, none exceeding a frequency of 1%, with the most frequent (HsInv1144) reaching 0.11% globally and 0.86% within Europeans. The Americans showed 44 unique inversions, of which only two surpass the 1% frequency in that ancestry. East Asian populations had 57 unique inversions, all below the 2% frequency in that ancestry, while South Asian populations exhibited 82 unique inversions, seven surpassing the 1% of MAF in that ancestry.

Population differentiation

In addition to identifying inversions belonging to one specific ancestry, our objective was to estimate the variation in inversion MAF between ancestry. To this end, we

used the F_{ST} , which ranges from 0 to 1, where lower values indicate similarity in AF between populations, while higher values reflect greater population differentiation (Holsinger & Weir, 2009). If natural selection favours an orientation in one population but not in others, the AF at that locus will diverge more rapidly between populations and the F_{ST} at that locus will tend to be higher than at neutral loci, where AF differences arise solely from genetic drift. In contrast, low F_{ST} values can indicate balanced selection, maintaining similar AF between populations.

To get a first estimate of the action of selection acting on inversions in different human populations, we calculated the F_{ST} for all variants, both globally and within each pair of ancestries. For each variant, its value F_{ST} was compared to the distribution derived from 10,000 random SNPs of similar global frequency throughout the genome (1,000 for X-chromosome inversions). Using this approach, we identified 23 inversions with a global F_{ST} in the top 5% of the genome-wide distribution. However, when considering pairwise comparisons between ancestries, 64 of 677 inversions (9.45%) had at least one comparison within this top 5%, and 18 of these were in the top 1%. These variants are summarised in Table 3.7, which lists all inversions with a significant global F_{ST} , and in Figure 3.13, which highlights all inversions with at least one ancestry showing a significant F_{ST} value.

Among the variants showing global frequency differences within ancestry, seven showed strong differentiation between African populations and others with F_{ST} values of 0.15-0.78 (HsInv0055, HsInv0114, HsInv0297, HsInv0340, HsInv0348, HsInv0389, Mos-17-36222071-36490245). For East Asian populations, no variant was significant exclusively in this ancestry. However, when excluding South Asian populations, HsInv0385 showed differentiation relative to the others. Similarly, for South Asian populations, no variant exhibited a significant F_{ST} relative to the remaining populations. In Europeans, one inversion met this criterion (HsInv0124), while for American populations, no inversion showed such a pattern. The pronounced effect observed in African populations likely reflects the effect of the out-of-Africa bottleneck from which the other human populations derive. Nevertheless, some clear cases suggest genuine population-specific differences in certain inversions.

Figure 3.13: **Global and population-specific MAF for polymorphic inversions.** Summary of the global and ancestry-specific minor allele frequency (MAF). Frequencies were calculated using unrelated individuals. Bars represent the global MAF, while coloured triangles (Std minor allele) or circles (Inv minor allele) indicate the MAF within each ancestry: Africans (green), Americans (yellow), East Asians (blue), Europeans (purple), and South Asians (brown). Black asterisks denote significant genetic differentiation between some of the ancestries based on F_{ST} values with $P \leq 0.05$.

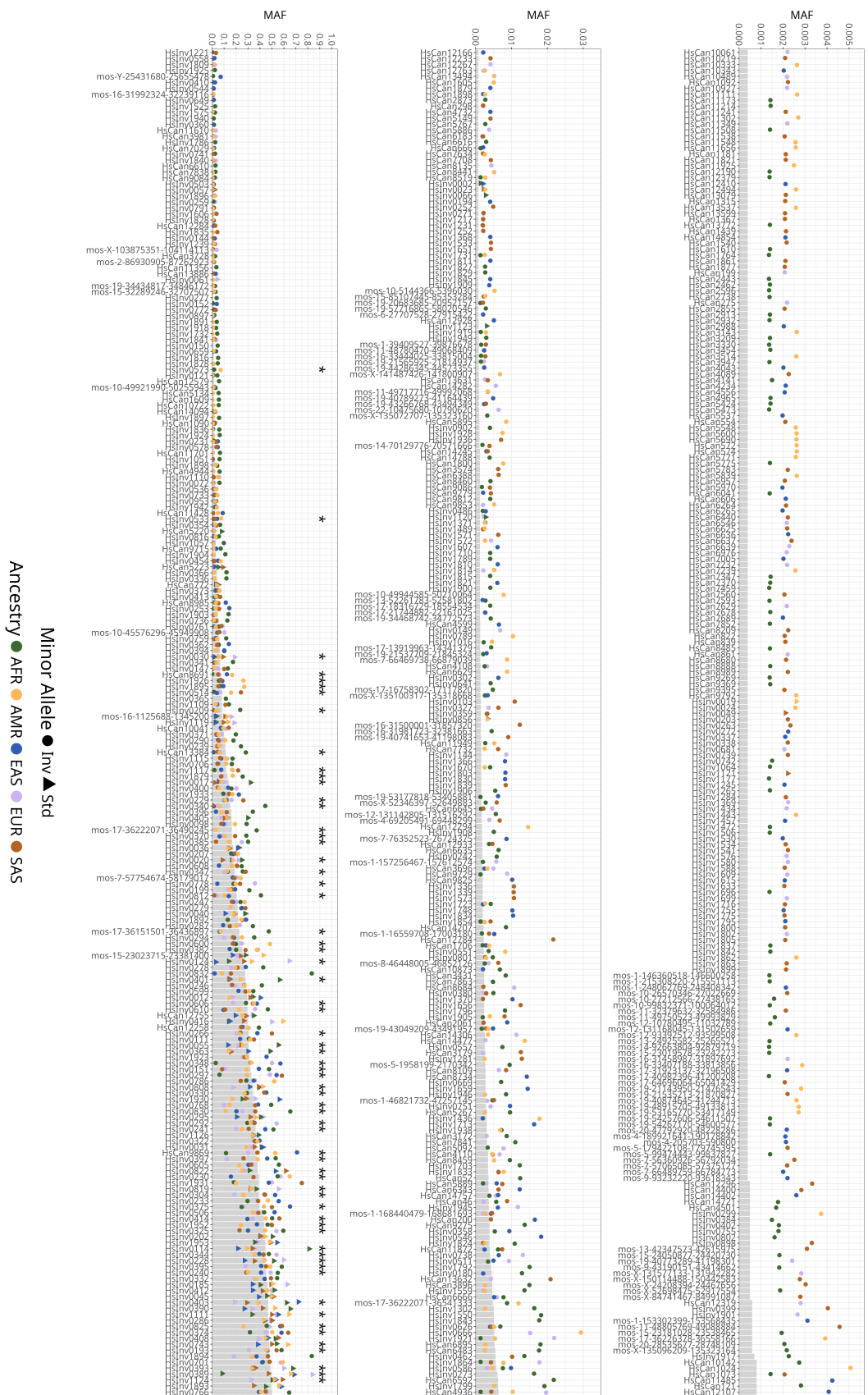


Table 3.7: **Summary of inversions with Global Fixation index (F_{ST}) differences.** Variants with significant F_{ST} global values. The table shows the P -values for the global F_{ST} and also for the pairs of ancestries. Non-significant values are marked as ns.

Variant	Global	AFR AMR	AFR EAS	AFR EUR	AFR SAS	AMR EAS	AMR EUR	AMR SAS	EAS EUR	EAS SAS	EUR SAS
HsInv0340	0.003	0.013	0.012	0.013	0.007	ns	ns	ns	ns	ns	ns
HsInv0297	0.004	0.001	0.006	0.004	-	ns	ns	ns	ns	ns	ns
HsInv0114	0.007	0.005	0.005	0.011	0.009	ns	ns	ns	ns	ns	ns
HsInv0124	0.008	ns	ns	0.005	ns	0.006	0.003	ns	0.001	0.018	0.001
HsInv0389	0.009	0.012	0.01	0.004	0.027	ns	0.025	ns	ns	0.083	-
HsInv0348	0.01	0.011	0.003	0.003	0.014	ns	0.028	ns	ns	ns	ns
mos-17-36222071-36490245	0.01	0.036	0.035	0.031	0.048	-	ns	-	ns	-	ns
HsCan9869	0.011	0.007	0.005	ns	0.043	ns	0.007	ns	0.018	0.016	ns
mos-17-36151501-36436897	0.011	ns	0.031	0.017	ns	ns	-	-	ns	ns	-
HsInv0812	0.015	ns	ns	0.021	0.016	ns	0.02	0.007	ns	0.06	ns
HsInv0230	0.021	ns	0.021	ns	ns	ns	0.024	ns	0.013	ns	0.047
HsInv0017	0.028	0.04	ns	0.084	0.025	ns	ns	ns	ns	ns	ns
HsInv0055	0.03	0.046	0.018	0.031	0.047	ns	ns	ns	ns	ns	ns
HsInv0403	0.03	ns	ns	0.027	ns	0.028	0.026	ns	0.011	0.014	ns
HsInv0151	0.032	ns	ns	0.026	0.014	ns	0.003	0.001	0.047	0.009	ns
HsInv0209	0.033	ns	ns	ns	0.037	ns	ns	ns	ns	ns	ns
HsCan13384	0.034	ns	0.043	ns	0.045	ns	ns	ns	ns	ns	ns
HsInv0401	0.034	ns	0.048	0.034	0.039	ns	ns	ns	ns	ns	ns
HsInv0375	0.035	ns	0.029	ns	ns	ns	0.026	ns	0.021	ns	0.014
HsInv0385	0.037	ns	0.03	ns	ns	0.038	0.03	ns	0.024	ns	0.013
HsInv0768	0.039	0.044	ns	0.02	ns	ns	ns	0.014	ns	ns	0.001
HsInv0363	0.048	ns	ns	0.033	0.037	ns	ns	ns	ns	ns	ns
HsInv0830	0.049	0.012	ns	ns	ns	ns	ns	0.031	ns	ns	ns

3.3.7 Size effects

Given the large number of genotyped samples and variants, AF were highly variable, a pattern also evident in Figure 3.9. As discussed above, inversions, particularly large ones, can have mutational consequences by disrupting genes or regulatory elements, potentially leading to deleterious effects. More importantly, long inversions are expected to have a higher fertility cost as a result of the generation of unbalanced gametes by recombination in inversion heterozygotes. Therefore, inversions that persist at the population level, especially at high frequencies, are likely to confer some selective advantage. This expectation is particularly relevant for large inversions, whereas smaller inversions may be less likely to affect functional elements or have negative effects associated with recombination and therefore may evade strong selective pressures (Campoy et al., 2022). In fact, a negative correlation between inversion frequency and genetic size has previously been observed in a much smaller dataset (Puig et al., 2020).

Based on this hypothesis, we tested whether inversion characteristics were associated with the AF. Two analyses were performed: one using all polymorphic inversions identified in this study, and another restricted to the subset analysed by Moreira-Pinhal et al. (2025), for which breakpoints were manually curated and the inversion and breakpoint sizes are therefore more accurately defined.

First, we assessed whether the size of the inversion was associated to the MAF using a linear regression. When considering the complete dataset, larger inversions tended to have lower frequencies, although this trend was not statistically significant (Global MAF vs. inversion size: adjusted $R^2 = 0.0039$, $\beta = -2.068 \times 10^{-8}$, $p = 0.0576$). Repeating the analysis with the Moreira-Pinhal et al. (2025) subset yielded a similar negative trend that was statistically significant, although the size of the inversion still explained only a small fraction of the variability of MAF (Global MAF vs. inversion size: adjusted $R^2 = 0.0291$, $\beta = -5.308 \times 10^{-8}$, $p = 9.997 \times 10^{-4}$).

Next, we evaluated the ratio between the mean size of BPs and the inversion size in the inversions, which seems to be mediated by IRs, focussing on the 612 inversions by Moreira et al.(2025) in which the BPs were assigned manually.

This metric was previously shown to be correlated with AF in Giner-Delgado et al. (2019), Puig et al. (2020), and Campoy (2024). Using this ratio, we observed a significant positive association (Global MAF vs. mean breakpoint size/inversion size: adjusted $R^2 = 0.0303$, $\beta = 0.0016$, $p = 0.000792$), although the explained variance was lower than that reported by Campoy (2024).

To explore whether MAF varied across the range of breakpoint-to-inversion size ratios, we grouped the inversions into ratio intervals and compared MAF among these categories. This grouping, as previously reported in Giner-Delgado et al. (2019), Puig et al. (2020), and Campoy (2024), revealed a general trend in which inversions with low ratios were associated with low AF, while higher ratios tended to correspond to increased MAF. Consistent with this pattern, a linear model fitted using the ratio as a continuous predictor showed a positive association between the ratio and MAF (adjusted $R^2 = 0.2913$, $\beta = 0.144$, $p < 2 \times 10^{-16}$). These

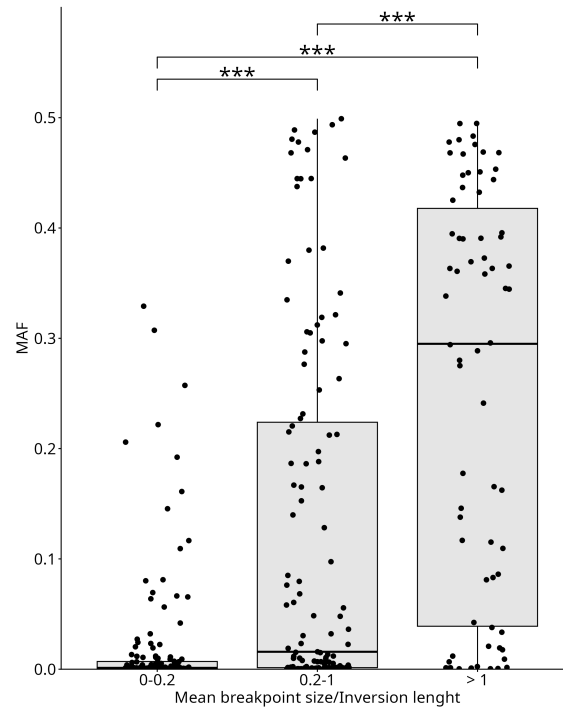


Figure 3.14: **Distribution of MAF according to the breakpoint size/inversion length ratio.** Boxplot display the distribution of MAF grouped into three categories of breakpoint size/inversion length ratio defined based on segmented regression: 0–0.2, 0.2–1, and >1. Each point represents an inversion. Pairwise comparisons between groups were performed using Tukey’s HSD following one-way ANOVA, and asterisks (***) indicate highly significant differences ($p < 0.001$).

patterns are illustrated in Figure 3.14. In general, inversions with a higher ratio between the size of IRs and the length of the inversion tended to be more frequent, consistent with the idea that this ratio influences the recurrence of the inversion, as previously reported by Giner-Delgado et al. (2019) and Puig et al. (2020).

3.3.8 Genotype validation

Once the genotyping results were completed, we wanted to verify their reliability by performing several comparative analyses. First, we evaluated the genotypes of samples belonging to trios by directly comparing the child’s genotypes with those of both parents, with the objective of identifying Mendelian-inconsistent genotypes. The initial analysis included 38 families for which genotype data was available for all three members. In total, 28,022 genotypes were assessed, of which 28,002 (99.92%) were consistent with Mendelian inheritance, while 20 showed incompatibilities. To

better elucidate the underlying causes of these inconsistencies, the results are summarised in Table 3.8, where the potential sources of the observed discrepancies were systematically examined.

Table 3.8: Verification of Mendelian inconsistencies. Table reports all trios in which the child presents a genotype that is non-concordant with parental genotypes. Each row corresponds to a inversion and indicates the family for which Mendelian inheritance fails, together with the genotype of each family member. Values in parentheses represent the number of reads supporting each orientation, with those for heterozygous genotypes always in Std/Inv order.

Inversion	Family	Genotype child	Genotype father	Genotype mother
HsCan4110	PR22	Het (4/1)	Std (7)	Std (7)
HsCan6645	1463	Het (8/5)	Std (19)	Std (11)
HsCan6895	GB97	Het (7/3)	Std (5)	Std (8)
HsCan12258	VN68	Het (48/9)	Std (6)	Std (4)
HsInv0036	1463	Het (34/24)	Inv (8)	Inv (9)
HsInv0111	IT220	Het (50/36)	Std (62)	Std (73)
HsInv0207	3150	Het (22/6)	Std (6)	Std (14)
HsInv0207	GB97	Het (2/1)	Std (2)	Std (10)
HsInv0278	PR05	Std (57)	Inv (35)	Std (28)
HsInv0297	PR05	Std (17)	Std (16)	Inv (10)
HsInv0348	PEL051	Het (45/32)	Std (5)	Std (10)
HsInv0366	Y077	Het (21/19)	Std (8)	Std (6)
HsInv0701	2424	Std (14)	Inv (22)	Het (23/16)
HsInv0808	3140	Std (28/1)	Std (7)	Inv (11)
HsInv0812	2424	Std (11)	Inv (5)	Std (10)
HsInv1115	PEL019	Inv (53)	Het (9/3)	Std (10)
HsInv1933	IT220	Het (21/8)	Std (6)	Std (8)
HsInv1933	IT3	Het (28/17)	Std (5)	Std (9)
HsInv1933	PEL003	Het (32/7)	Std (11)	Std (8)
HsInv1933	VN46	Het (23/12)	Std (7)	Std (10)

We first observed that inconsistencies were detected in a total of 15 of the 38 families analysed, with only 5 families showing inconsistencies for 2 inversions. When examining which inversions were more prone to these discrepancies, HsInv1933 was the most affected followed by HsInv0207, with inconsistencies observed in four and two families, respectively. Among the discordant genotypes, two have already been reported in the original GeONTIpe publication (Moreira-Pinhal et al. (2025)), corresponding to families 3140 and PR05.

The origin of these inconsistencies may be multifactorial. Potential sources include

probe-related errors that affect a specific family member, possibly due to low sequencing quality of the supporting reads, sample contamination, or genotyping errors. For example, in 12 cases, one of the trio individuals has a homozygote genotype supported by 7 or fewer reads, which could result in the inability to capture the alternative allele required for Mendelian consistency. However, there are cases in which this is unlikely, as all members of the trio present a sufficiently high number of reads to allow for confident genotype assignment. For such cases, the most plausible explanation is the presence of artefacts introduced during the genotyping process.

An alternative explanation for some of these discrepancies could be a true orientation change that occurs as a somatic mutation for any member of the family, arising directly in the individual or during one of the lymphoblastoid cell lines (LCL) divisions. However, identifying such events is not feasible without detailed information on cell lineage passages and comparative genotyping across those passages. Moreover, these somatic changes probably affect a reduced number of individuals. Therefore, the comparison of trio-based genotypes resulted in the identification of a subset of candidate inversions for future targeted validation, with the aim of elucidating the underlying causes of the observed inconsistencies.

Once this first verification was completed, we proceeded to compare the genotypes of 57 inversions for which previous genotypes were available from Campoy (2024). These data included both experimentally validated genotypes obtained using different PCR-based techniques, or genotypes generated using BreakSeq (Lam et al., 2010) in short reads of all samples from 1000GP.

For comparison of the genotypes obtained from BreakSeq, only one inversion was shared between both datasets (HsInv0098). The analysis included 1,311 genotypes obtained with both methods (GeONTipe and BreakSeq) and all of them were fully concordant. Next, we compared our results with the experimental genotypes available for 57 inversions in a set of 231 samples, which also included HsInv0098. Of a total of 13,167 theoretically shared genotypes, 7,180 (54.53%) could be directly compared. The remaining genotypes could not be evaluated because genotype was missing in both datasets (1,458, 24.35%), in the data based on ONT (1,285, 21.46%)

, or in the experimental data (3,244, 54.18%). Among the 7,180 comparable genotypes, 7,108 (99.00%) were consistent between both studies.

Of the remaining 72 genotypes that showed discrepancies, three differences had already been reported in the original GeONTIpe publication, which apparently were mainly due to possible errors in the PCR genotypes (Moreira-Pinhal et al. (2025)). We then wanted to investigate the remaining 69 discordant genotypes to determine whether they were more likely to be attributable to experimental genotyping errors or to inaccuracies in the genotypes generated in this study. We observed that 46 of the 69 discordant genotypes (66.67%) were concentrated in only four inversions: HsInv0241, HsInv0278, HsInv0344 and HsInv0348. These four inversions had already been suspected to contain unreliable experimental genotypes when they were performed using iPCR and MLPA (Giner-Delgado et al., 2019). If discrepancies associated with these four inversions were excluded, the number of discordant genotypes was reduced to 23, resulting in a general concordance of 99.67%.

Despite the good concordance observed, we decided to evaluate the number of reads that supported each discordant genotype. This metric is particularly relevant for homozygote calls, as the requirement of a minimum threshold of five reads with the same orientation may not be sufficient to exclude the possibility that all observed reads originate from just one chromosome.

We found that, excluding the inversions already mentioned, 10 genotypes were supported by five or fewer reads (Table 3.9). Of these, four were heterozygous, indicating that at least one read supported the alternative orientation. Among the remaining six homozygous genotypes supported by five or fewer reads, two had been genotyped using the SNPs methodology implemented in GeONTIpe. However, the number of SNPs that supports the reads belonging to different haplotypes was quite low (2 and 1 respectively). So, those genotypes are likely to be errors due to the low coverage of the samples.

For heterozygous genotypes with low read support, the most plausible source of error appears to be the presence of artefact reads, particularly in cases where only a single

read supports the alternative orientation. This scenario applies to five of the eight heterozygous genotypes. Specifically, we checked for some recurrent artefact reads classified as “mirror reads”, as they contain the read sequence in both forward and reverse orientations, resembling the structure expected for inverted duplication. For these cases, we generated dot plots to assess whether the read supporting the minor orientation exhibited the structural features characteristic of these mirror reads, or whether it instead appeared to represent a genuine signal. However, none of the cases showed a structure indicative of the artefact reads described above (Table 3.9).

Table 3.9: **Verification with experimental inconsistencies.** Table reports the 23 discrepancies observed with experimental data. Each row correspond to an inversion indicating the experimental genotype as well as the genotype provided by GeONTIpe, indicating the number of reads supporting it. with those for heterozygous genotypes always in Std/Inv order.

Inversion	Sample	Experimental	GeONTIpe
HsInv0031	NA10859	Het	Std (7)
HsInv0031	NA19027	Het	Inv (2)
HsInv0031	NA19310	Het	Std (9)
HsInv0045	NA19119	Het	Inv (5)
HsInv0193	NA20540	Het	Std (7)
HsInv0209	NA19468	Het	Inv (69)
HsInv0228	NA19118	Std	Het (4/1)
HsInv0325	NA19118	Het	Inv (6)
HsInv0340	NA19137	Std	Het (2/1)
HsInv0340	NA19307	Std	Het (3/3)
HsInv0340	NA20801	Std	Het (5/1)
HsInv0341	NA19189	Het	Inv (4)
HsInv0370	NA19240	Std/Del	Std (20)
HsInv0374	GM18501	Inv	Het (34/7)
HsInv0374	NA20585	Het	Std (6)
HsInv0374	NA20785	Het	Inv (5)
HsInv0382	NA12750	Het	Std (6)
HsInv0414	NA19118	Std	Het (4/1)
HsInv0786	NA19204	Het	Inv (7)
HsInv1057	NA19092	Het	Std (5)
HsInv1057	NA19118	Std	Het (2/1)
HsInv1124	NA21099	Het	Inv (5)
HsInv1124	NA21111	Std	Het (4/5)

In contrast, three of the discordant genotypes were supported by a large number of reads. In these cases, the probability that the discrepancy originates from the ONT data is substantially reduced, unless confounding factors such as sample contam-

ination or alternative haplotypes are present. An example of such contamination was detected in sample HG02554, in which several chromosome X genotypes were defined as heterozygous despite the individual being a male. SNP-based validation confirmed contamination in this sample and was therefore excluded from subsequent analyses. For these ten well supported genotypes, discrepancies are more likely to be attributable to errors in the experimental data. However, additional validation is required to conclusively determine the source of the error, whether in GeONTipe-generated genotypes or in experimental methods. This need for further verification is particularly relevant for inversions supported by an intermediate number of reads (between 6 and 10), where genotype confidence is inherently less robust.

3.3.9 Hardy–Weinberg equilibrium (HWE) analysis

Finally, as a last step in quality control, we sought to assess whether our genotypes exhibited systematic bias or whether certain variants consistently failed to be genotyped in a manner that might have gone unnoticed. To this end, an analysis of Hardy–Weinberg equilibrium (HWE) was performed to evaluate the internal consistency of the genotype frequencies within each ancestry. Deviations from HWE were assessed solely as a quality control measure to detect possible genotyping errors, rather than to investigate the underlying biological or evolutionary processes. This is because the population structure and other demographic factors were not extensively controlled in this analysis. For this analysis, genotype counts for each inversion were performed in unrelated individuals of different ancestries. However, inversions located on the Y chromosome were excluded, as the haploid nature of this chromosome violates the assumptions of the HWE model. For inversions on the X chromosome, the expected sex-specific allele distribution was taken into account to properly perform the allele count. Variants were considered to deviate from HWE when the mid p-value was below 3×10^{-5} , equivalent to a Bonferroni correction for the number of tests performed (1566 tests).

This analysis revealed that each ancestry had at least one inversion that deviated from HWE. At the global level, 10 of the 680 polymorphic variants did not reach

HWE in at least one ancestry. Of these, 1 variant (HsInv0111) deviated from HWE in all ancestries in which these variants were polymorphic, suggesting that this case may reflect a systematic problem in genotyping rather than population-specific effects. For the remaining 9 variants, African populations exhibited the highest number of variants that did not achieve HWE, with 6 inversions. However, the south Asian ancestry did not show any variant that deviated from HWE. Moreover, between those 10 variants, 3 of them have less heterozygote with respect to the expected, while the other presented an excess of them (Table 3.10).

Table 3.10: **Assessment of deviation from HWE.** Each row represents an inversion that deviated from HWE in one or more ancestries. For each inversion, the total number of alleles is reported. The table also indicates the ancestries showing deviations, the direction of the discrepancy, and the possible cause of the observed deviation.

Inversion	Genotyped alleles	MAF (%)	Pop desviation	Desviation	Cause desviation
HsCan12258	266	39.1	AFR	Lack Het	Presence of SV
HsInv0111	670	30.54	ALL	Excess Het	Bad breakpoint location
HsInv0151	290/192	32.13	AFR/EAS	Excess Het	Fail probes
HsInv0199	186	31.13	AFR	Excess Het	Fail probes
HsInv0322	120	36.13	EUR	Excess Het	Not determined
HsInv0348	464	42.88	EAS	Lack Het	Not determined
HsInv0768	168	49.4	AFR	Excess Het	Not determined
HsInv0812	196	40.82	AFR	Lack Het	Presence of SV
HsInv1933	362	22.37	EAS	Excess Het	Alternative haplotype
Mos-7-57754674-58179017	578/280	19.22	AFR/AMR	Excess Het	Not determined

As already mentioned, this information was used solely as an indicator to guide future, more focused analyses aimed at investigating the genotyping results of these variants in greater depth and determining the underlying causes of the observed deviations. One possible explanation for these deviations is the presence of previously unknown alleles in certain populations, which could generate haplotypes that GeONTipe may misclassify during genotyping. These new haplotypes could have other SVs that affect some of the studied regions. If such SVs were preferentially related to one orientation, this could lead to systematic over-genotyping of one of the orientations, resulting not only in deviations HWE, but also in a general deficit of heterozygotes for the affected inversions.

To assess whether an additional structural variant could bias allele genotyping, variants deviating from HWE due to the absence of heterozygotes in a given ancestry

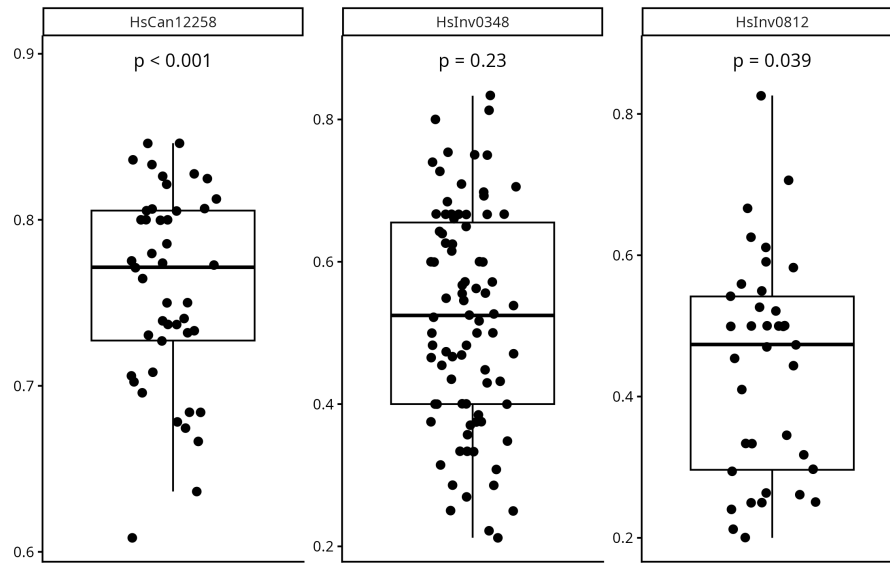


Figure 3.15: **Allelic imbalance detection.** Boxplots showing the distribution of the standard-to-total read proportion in heterozygous individuals. Individual dots represent each heterozygous sample in the ancestry analysed. The p-value indicates the significance of deviation from the 50% ratio expected under HWE.

were examined. Heterozygous samples from the affected ancestry were selected and the ratio of standard (Std) reads to total reads was calculated for each variant. In an unbiased genotyping, this ratio is expected to be centred around 50%. The observed distributions were therefore compared to this expectation. Two of the variants analysed (HsCan12258 and HsInv0812) showed a significant deviation from the expected allelic ratio (Figure 3.15). In both cases, additional SVs associated with the inverted allele were detected. For HsCan12258, all inverted alleles carried a 69 kb insertion, while for HsInv0812 an associated deletion was observed, both of which may impact genotyping accuracy and, consequently, HWE calculations.

For inversions showing an excess of heterozygotes, a subset of heterozygous samples was manually inspected to identify potential causes of genotyping errors. In some cases, errors were linked to probe failure, with reads showing multiple occurrences of the same probe, suggesting probe nonspecificity or the presence of an additional sequence containing the same motif. In other cases, genotyping difficulties could be explained by poorly positioned breakpoints or the complexity of the region (Table 3.10). However, for several samples, no clear cause of the genotyping error was detected, leaving open the possibility of other factors, including a potential advan-

tage of heterozygotes, although this could not be confirmed with the available data.

3.3.10 Linkage disequilibrium (LD) with nearby variants

As the last analysis step, we checked whether any of the inversions showed a strong LD with nearby SNPs. This information can provide both an idea of the unique or recurrent origin of inversions and indirect evidence of genotyping precision, particularly in cases where the values of LD approach 1. However, this measure is not definitive, as low linkage values are compatible with a recent inversion origin or recurrent events that reduce LD with surrounding variants due to the exchange of SNPs between orientations over time. To try to address this, LD was calculated both globally and separately for each ancestry, using the full set of genotypes, as well as an alternative set of more reliable genotypes in which the threshold for defining homozygosity was increased from five to seven reads supporting the same orientation (which corresponds to a binomial probability of being a heterozygote of $P = 0.0078$). The application of this stricter criterion resulted in the conversion of 65,336 previously valid genotypes (4.38%) into undetermined genotypes.

We first examined the LD results using the complete set of genotypes. At the global level, at least one SNPs with a LD greater than 0.95 was identified in 145 of the 677 polymorphic inversions, with a total of 2,104 linked SNPs. Among inversions showing high LD, 37 were associated with a single variant, while the remainder showed associations with at least two SNPs, with a maximum of 280 observed for HsInv1923. In particular, when HsInv1923 was genotyped, the inverted allele was consistently associated with a 6 kb insertion located at breakpoint 1, which already suggested that this inversion occurred only once. In addition, 41 variants presented a LD between 0.8 and 0.95 in the global population (Figure 3.16).

An indirect approach to evaluate whether genotyping errors constitute the main source of low LD values was to remove genotypes that are more likely to be false, namely those supported by a small number of reads. For this reason, we excluded homozygotes supported by 5 to 7 reads in autosomal inversions and chromosome X inversions in females. If a substantial fraction of errors originated from such geno-

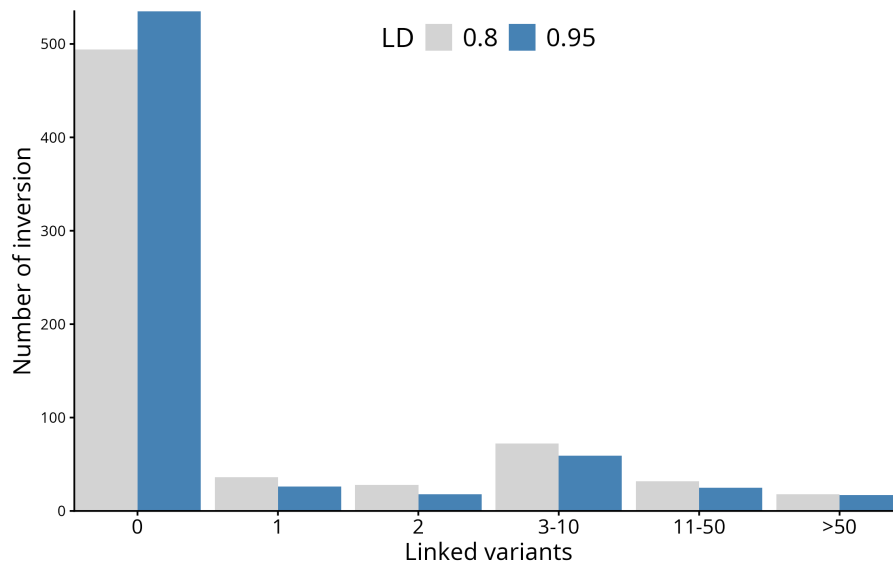


Figure 3.16: **Linked variants per inversion.** Number of inversion per LD bin under two scenarios: variants with $LD \geq 0.8$ (gray) and $LD \geq 0.95$ (blue). Each X-axis bin shows the number of nearby variants linked to an inversion. The “0” bin indicates inversions with no variants in LD.

types, we would expect a marked increase in the number of variants that show high LD after filtering. Instead, we observed that only two additional variants showed LD greater than 0.95 compared to the original data set (HsCan13631 and HsCan8109). This result suggests that the chosen minimum-read threshold is reasonable. Nevertheless, it is still possible that there are additional sources of genotyping errors that we are currently not taking into account.

We then assessed whether the 10 inversions suspected of genotyping errors due to significant differences with HWE in at least one population showed associated SNPs. Among them, only HsInv0151 and HsInv0348 presented SNPs with high LD, suggesting that the inversion genotypes could be correct. This case was particularly striking, as it exhibited 167 variants with values of LD greater than 0.95. Similarly, we analysed the four inversions that showed the largest discrepancies between the experimental data and the GeONTIpe generated genotypes (HsInv0241, HsInv0278, HsInv0344 and HsInv0348). Of these, only HsInv0348 displayed LD values greater than 0.95. However, the variant with the highest LD reached 0.969, suggesting that a small number of genotypes may still be incorrectly assigned.

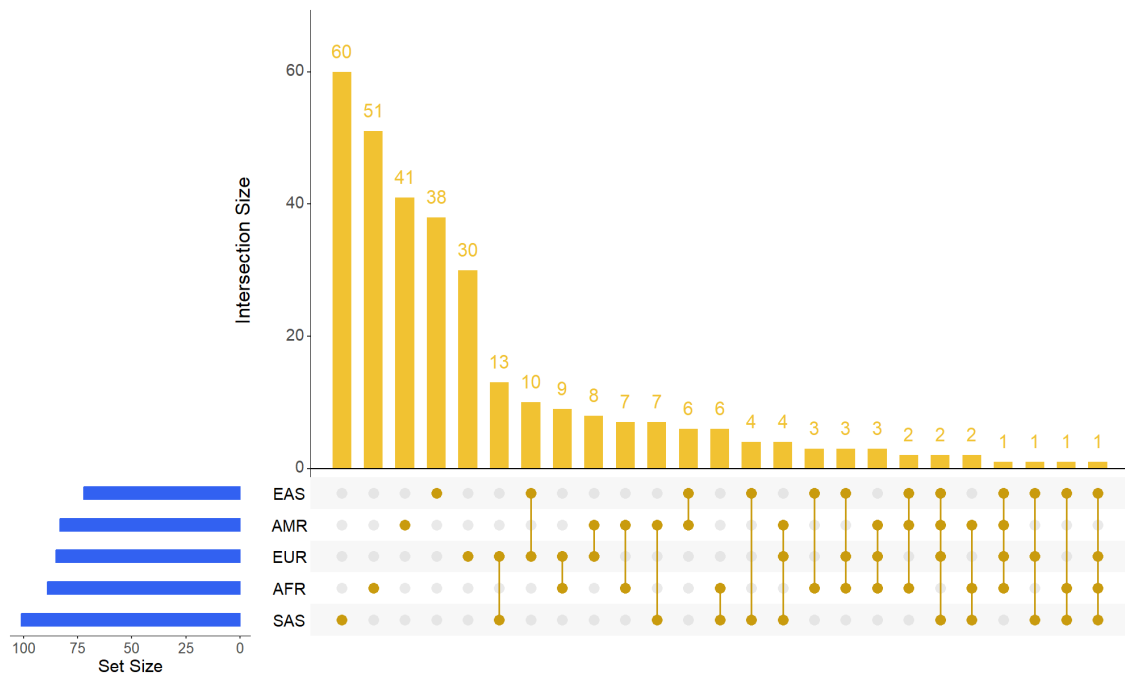


Figure 3.17: **Linked variants per ancestry.** Upset plot showing unique variants absent in global LD across five ancestries (AFR, EUR, AMR, SAS, EAS). Vertical bars represent the number of shared variants between populations, horizontal bars indicate the total number of variants per population.

Subsequently, we investigated whether inversions lacking high LD at the global level might still show strong LD within specific ancestries. In African populations, we identified 89 additional inversions with LD greater than 0.95 that were not detected at the global level, while in the remaining ancestries this number ranged between 83 and 101, with South Asian populations showing the highest number of these cases. In general, 313 inversions showed LD in at least one ancestry but not globally, being 220 population specific (Figure 3.17). This observation suggests that in these ancestries there are no potential genotyping errors for the above inversions, or alternatively, that no recurrence events have occurred, at least among the genotyped samples, although bottleneck effects cannot be discarded for the out-of-Africa populations.

When we checked whether any of these variants were shared between multiple ancestries, we observed that 73 variants showed a high LD in two ancestries, 16 in three and 4 in four (Figure 3.17). For those present in three or four ancestries, we found that 8 of the 20 variants exhibited a global LD greater than 0.8 globally, which makes them strong candidates for additional genotype validation to ensure

the absence of genotyping errors.

Several factors may explain the lack of very high LD values. As already mentioned, one possibility is the recurrent nature of many inversions, particularly those mediated by SDs, which may arise repeatedly over time and recombine, thus reshuffling SNPs and obscuring associations between the inversion and linked SNPs. Addressing this issue would require accurate haplotype resolution and phasing, followed by pairwise comparisons between haplotypes, while ensuring that reduced sample sizes do not generate spurious SNP associations.

In the case of inversions without IRs in BPs, the most likely underlying explanations of the absence of SNPs in high LD are different. The first factor that could contribute is the inversion age: very young inversions may not have had sufficient time for SNPs to become fixed with the inversion, showing an incomplete linkage. The second factor could be genotyping errors that originate in GeONTIpe or in the SNPs used to verify the link. The latter is expected to be more pronounced in variants with low AF, which could be easy to miss, especially if the sequencing coverage is low.

Regarding possible inversion genotyping errors introduced by GeONTIpe, it is also important to note that their impact depends on the success rate of the specific inversion: a single error in a large set of samples has a minor effect, whereas the same error in a variant genotyped in only a few samples has a greater impact. This effect is also more pronounced for variants with low AF, since a single genotyping error that affects the minor allele has a proportionally greater influence in LD calculations. Therefore, some inversion genotypes can be explored in more detail to determine if a small number of errors can explain the observed reduction in LD, especially for variants with a linkage close to 1.

By combining these 680 polymorphic inversions found here with the highly reliable polymorphic inversions reported by Campoy (2024) that are not shared with this study, the resulting dataset comprises 774 polymorphic inversions genotyped in a large number of samples. This constitutes the largest global catalogue of polymorphic inversions to date and likely represents a very large fraction of all common

inversions in the human genome, providing a solid foundation for the improvement and update of the InvFEST database.

3.4 Complex genomic regions

The analysis of 612 regions across 54 samples conducted in the first result section of this thesis revealed that the structural diversity present in several of the target regions was substantially greater than originally anticipated. Although some examples of this diversity were already described in Moreira-Pinhal (2025), additional cases emerged during the design and validation of the probes that were later used to genotype inversions in these regions. In the following sections, we present selected examples that required manual inspection through the analysis of individual long reads or local assemblies and the development of a new pipeline to facilitate the automatic characterisation of complex regions.

3.4.1 HsInv0822 Region

The first case study focuses on an inversion located on the X chromosome (HsInv0822), affecting approximately 40 kb. This inversion had previously been analysed and genotyped in our group in Puig et al. (2020), where ddPCR was used to genotype inversions mediated by IRs. In the genotypes obtained in the first result section, the analysis of this region in 54 individuals yielded a minor allele frequency (MAF) of 29%, the Std allele being the less common. The increase in the number of samples raised that value to 38.18%. Upon checking the coordinates of our study regions, we observed that two nearby regions could potentially form a complex haplotype that includes this polymorphic inversion

Specifically, HsInv0822 is located within a genomic interval containing two additional pairs of nested inverted SDs, each potentially capable of mediating inversion events, corresponding to HsInv0394 (approximately 165 kb) and HsInv0816 (approximately 355 kb). The arrangement of these SDs could generate three nested polymorphic inversions (HsInv0822 by recombination within the most internal pair, HsInv0394 within the second pair, and HsInv0816 within the most external SDs). The SDs associated with HsInv0394 are approximately 50 kb away from HsInv0822, while those mediating HsInv0816 are roughly 90 kb and 197 kb from HsInv0822 on

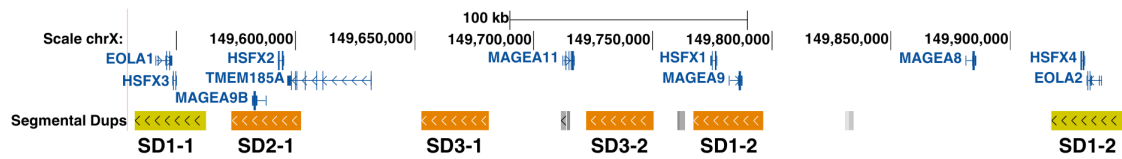


Figure 3.18: **HsInv0822 genomic region.** Adaptation from the UCSC Genome Browser view of the HsInv0822 region showing the SD structure track and MANE gene annotations. For each SD, the two copies share the same SD identifier and are distinguished by a second digit, corresponding to the first and second copy (e.g., SD1-1 and SD1-2; SD2-1 and SD2-2).

each side, respectively. Together, these three inversions form a structurally complex genomic region spanning 415 kb (chrX:149,532,424 - 149,947,803). This interval contains several genes, including members of the *XAGEA*, *HSFX*, and *EOLA* families (Figure 3.18).

Using the inversion-genotyping pipeline developed in this thesis, we observed that, similar to HsInv0822, both HsInv0394 and HsInv0816 are also polymorphic in humans, though at substantially lower frequencies (MAF of 6.5% and 3.2%, respectively). Taking into account these three nested polymorphic inversions, the region can theoretically accommodate up to eight distinct haplotypes. To analyse how many of those haplotypes exist, we filtered our data set, restricting the analysis to males with genotype information for all three inversions ($n=262$), homozygous females for all three variants ($n=134$) and heterozygous females for only one the three inversions ($n=134$), identifying six of the eight possible haplotypes in a total of 798 alleles. This filtering step was necessary because phasing information was not available to distinguish between paternal and maternal haplotypes. All commented conformations are referenced with respect to GRCh38, with a Std/Std/Std genotype. Two haplotypes, Inv/Inv/Inv and Inv/Inv/Std (order HsInv0816-394-822), were not observed in any individual. Based on this dataset, we inferred haplotype frequencies. The most common haplotype was Std/Std/Inv, with 474 occurrences (59.39%), which is also shown by T2T-CHM13, followed by the haplotype that carried the other HsInv0822 allele, which represented 37% of all observed haplotypes (Table 3.11).

Table 3.11: **Haplotype distribution of HsInv0822.** All possible HsInv0822 haplotype combinations, listed in the order HsInv0816–394–822. For each combination, the number of alleles observed in males, homozygous females, and females heterozygous for a single inversion is presented. Total allele counts and estimated frequencies are included.

Combinations	Males	Hom_females	Het_females	Total	Frequency
Std/Std/Std	101	78	120	299	0,37
Std/Std/Inv	154	190	130	474	0,59
Std/Inv/Std	2	0	4	6	0,01
Std/Inv/Inv	3	0	10	13	0,02
Inv/Std/Std	1	0	0	1	0
Inv/Inv/Std	0	0	0	0	0
Inv/Inv/Inv	0	0	0	0	0
Inv/Std/Inv	1	0	4	5	0,01
Allele count	262	268	268	798	

3.4.2 HsInv0786 Region

The second region that drew our attention corresponds to inversion HsInv0786 (chr16:28,288,647 - 28,874,074), located within the 16p11.2 region. This inversion is mediated by 134-kb SDs and was also previously genotyped by our group in Puig et al. (2020) (Figure 3.19). Furthermore, analyses of these genotypes had revealed multiple GWAS hits associated with various traits, including asthma, diabetes, and paediatric autoimmune diseases (González et al., 2014a; Campoy, 2024). This information made HsInv0786 particularly interesting due to its potential functional impact, as well as the presence of several members of the *NPIB* gene family located in the region (Dishuck et al., 2025). Initially, it was thought that the region harboured only the inversion; however, the presence of multiple SDs in both direct and inverted orientation suggested the possibility of additional structural rearrangements.

The initial analysis of 54 samples suggested that the region was more complex than expected. Although the inversion itself exhibits a MAF of 38%, we identified multiple additional SVs at both BPs. Within this first dataset, we detected eight distinct variants, observed in both the standard and inverted orientations. After expanding the dataset, we found that 278 of the 303 individuals with available genotypes

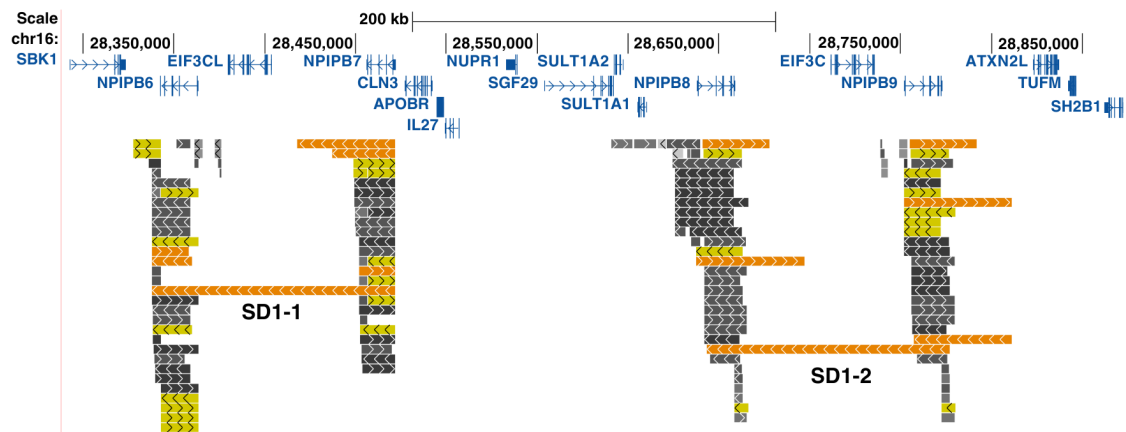


Figure 3.19: **HsInv0786 genomic region.** Adaptation from the UCSC Genome Browser view of the HsInv0786 region showing the SD structure track and MANE gene annotations. The main SDs two copies are labelled.

(78.54%) carried at least one of these additional variants. In total, these individuals represented 284 occurrences, indicating that some samples contained more than one variant. The detected variants correspond to the following:

- **114-kb deletion:** Observed in 79 individuals. It can be found within both inversion BPs, but it is more frequently located at the second breakpoint (73 individuals), with a similar distribution observed between both orientations.
- **11-kb deletion:** Found in 188 individuals in one or both BPs and in both inversion alleles. The most common configuration was the deletion in the second breakpoint of the Std allele (132 individuals).
- **12-kb insertion:** Detected in 14 samples within the internal region of the inversion and found in both orientations.

The combination of these SVs generates a highly diverse set of potential haplotypes. Because phasing information was not available, we then performed a manual inspection of individual long reads and assemblies that captured specific combinations of variants. This detailed analysis allowed us to identify additional SVs: a 114 kb duplication found at both BPs and a deletion of 12 kb in the internal section of the inversion, all detected in 1 haplotype in the dataset of 1409 samples. These new

variants further increase the number of possible haplotype combinations. Moreover, the insertion of 114 kb substantially extends the length of the SDs that generates the inversion, which can negatively affect its genotyping precision, as it increases the distance between the probes used for inversion detection. This not only reduces the number of samples that can be reliably genotyped but also explains why this insertion was not detected in the initial analysis.

Beyond these newly identified variants, the manual analysis of samples carrying additional variants enabled the detection of 19 different haplotypes, confirmed either directly through long-read evidence or through assemblies available for the same individuals. These 19 haplotypes are summarised in Table 3.12, indicating the samples in which they were found and the data source used for their identification. Although this analysis allowed the characterisation of several combinations of variants in the region, the large number of SVs detected suggests that additional haplotypes may exist, and the total number of haplotypes present in human populations remains unknown.

Manual analysis of these sequences allowed us to identify the specific regions in which the variants occur and to assess whether SDs could be responsible for their generation. In the case of the 114 kb deletions and insertions, these elements correspond to SDs in direct orientation, located at both breakpoint 1 and breakpoint 2 of the inversion, with sizes of approximately 20 kb and 36 kb, respectively. For the 11 kb deletions, no repetitive elements could be identified that might explain their origin. Regarding insertions and deletions within the inverted region, both were found in the same segment, which contains a 12 kb repetitive sequence that encompasses the *SULT1A* gene. The presence of both duplications and deletions in these locations suggests that they may harbour a variable number of such sequences.

These two regions (HsInv0822 and HsInv0786) discussed above represent only two examples in which we examined in greater detail all the available information. However, there are several other regions where the analysis of just a few individuals already indicated the presence of additional and often complex haplotypes. An illustrative case is HsInv0295 (chr7:144,135,693 – 144,423,179), where the T2T-CHM13

Table 3.12: **Haplotypes found for the HsInv0786 region.** Each row corresponds to one haplotype, indicating the sample where it was found and the specific combination of SVs detected, including inversion status and the presence of deletions or insertions at breakpoint 1, breakpoint 2, or within the internal region of the inversion. The ‘Evidence’ column indicates the source used to detect each haplotype, coded as (A) for assembly or (R) for read.

Hap	Sample	Orientation	BP1	Internal	BP2	Evidence
H1	GRCh38	Std	-	-	-	A
H2	HG002	Std	Del.114 kb	-	-	A + R
H3	NA19317	Std	Ins.114 kb	-	-	R
H4	HG00358	Std	-	Ins.12 kb	-	A
H5	HG00512	Std	-	-	Del.11 kb	A
H6	NA18989	Std	-	Del.12 kb	Del.11 kb	A
H7	HG002	Std	-	Ins.12 kb	Del.11 kb	A + R
H8	NA19705	Std	Del.11 kb	Ins.12 kb	Del.11 kb	R
H9	HG00673	Std	-	-	Del.114 kb	A + R
H10	HG02953	Std	-	Ins.12 kb	Del.114 kb	A
H11	CHM13	Inv	-	-	-	A
H12	HG02148	Inv	Del.11 kb	-	-	A + R
H13	NA20509	Inv	Del.114 kb	-	-	R
H14	HG01071	Inv	Ins.114 kb	-	-	A + R
H15	HG00513	Inv	Del.11 kb	Ins.12 kb	-	A + R
H16	HG01071	Inv	-	-	Del.114 kb	A + R
H17	HG02148	Inv	Del.11 kb	-	Del.114 kb	A
H18	HG02554	Inv	-	Ins.12 kb	Del.114 kb	R
H19	HG02059	Inv	Del.11 kb	Ins.12 kb	Del.114 kb	A + R

reference genome already showed a 100-kb deletion corresponding to a segment that we initially assumed to be inverted. Despite this, we were able to detect the inverted allele in other individuals, with a frequency of 63.93%. In addition to these two variants (inversion and deletion), manual analysis of long reads revealed a copy-number increase that involves both the SDs assumed to mediate the inversion, and the inverted region itself. In total, seven distinct haplotypes were confirmed in this region. Another notable example is HsInv0375 (chr17:18,985,128 – 19,266,295), where manual analysis identified six haplotypes, encompassing different combinations of deletions, duplications, and inversions. These variants were found to affect multiple sequences, including duplicated and unique regions in GRCh38.

All the cases mentioned correspond to large genomic regions spanning more than 100

kb. When reviewing the coordinates of several candidate regions analysed throughout the different chapters presented here, we found many more *loci* that share characteristics similar to those of these manually analysed regions, such as the presence of multiple pairs of SDs, several members of the same gene families, and large overall size. Although GeONTipe has proven to be very useful for genotyping polymorphic inversions, in repeat-rich regions with additional SVs, inversion genotypes do not completely describe the nature of alleles present in the analysed individuals. For this reason, we decided to develop a bioinformatic tool capable of identifying different structural haplotypes without relying on manual analysis, which usually requires data from multiple sources and is feasible only for a limited number of samples due to its time-consuming nature. Establishing specific structural haplotypes for each analysed sample will be a key step toward disentangling the functional consequences associated with each combination of variants.

3.4.3 Pipeline for the analysis of complex structural haplotypes

The development of our pipeline was motivated by the need to uncover genome structures directly from ONT long reads, which we had already mapped and which provide the advantage of spanning hundreds of kilobases. Initially, assembly-based strategies were considered. However, tests performed with assemblers such as Flye (Kolmogorov et al., 2019) and Canu (Koren et al., 2017) using only ONT long reads failed to generate complete local assemblies of the regions of interest, probably due to the extensive presence of repeated sequences. Furthermore, the lack of full control over the assembly process introduced uncertainty. For these reasons, we opted for a strategy that leverages the ONT reads directly, without relying on prior assembly. An additional advantage of designing our own tool is that, although high-quality assemblies such as the Human Pangenome are now available, the number of samples remains limited, which constrains the power to discover structural haplotypes.

Thus, to discover and analyse structural haplotypes we have developed a novel pipeline called CHARONTE (Complex Haplotype Analysis and Reconstruction

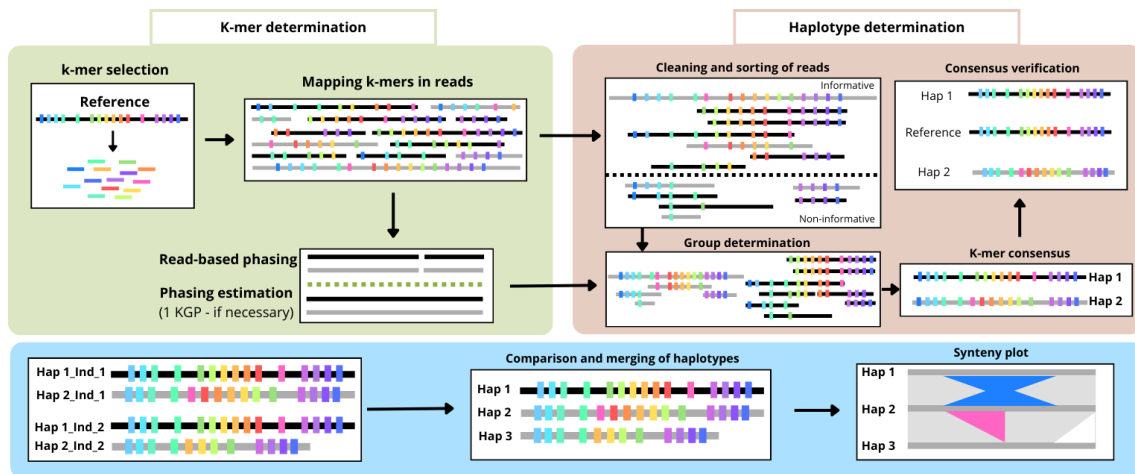


Figure 3.20: **Workflow overview of CHARONTE.** Schematic representation of the main processes carried out by CHARONTE to determine structural haplotypes in complex genomic regions. The diagram summarizes the selection of k-mers (shown as coloured segments) and their mapping to ONT long reads, followed by the removal of non-specific hits and the priority-based sorting of reads. This information, together with phasing information, are then used for read classification and subsequent consensus generation, which is compared against the reference genome. After multiple individuals are processed, the resulting haplotypes are compared and merged to obtain a set of unique haplotypes, which are finally used to generate the synteny plot and characterize the set of structural haplotypes detected in the region

using Oxford Nanopore Technology Evidence). The core principle of this pipeline is the identification of unique k-mers within the target sequence, and through the analysis of their position, orientation, and spacing, locate and connect sequence blocks across different reads sharing the same properties to progressively extend the analysed segment until the complete structure of each chromosome is resolved. Once individual structures are determined, shared k-mers allow population-level analyses, determining the frequencies of each distinct haplotype (Figure 3.20).

The pipeline is organised into three modules, the technical specifications of which are detailed in the Materials and Methods section of this thesis. These modules will be described in greater detail in the following sections, highlighting how they contribute to a systematic and scalable approach to analysing complex genomic regions.

K-mer selection

The first part of the pipeline generates all the input files that will be used as a base for the identification of the structural configurations present in the region of interest. A key element of this section of the pipeline is the construction of a set of unique k-mers designed using two reference genomes (GRCh38 and T2T-CHM13). Ensuring that these k-mers are strictly unique and highly specific within the region is essential, as duplicated or ambiguous k-mers could negatively impact downstream analyses, particularly given the high error rate of ONT long reads and the abundance of repetitive elements present in the complex regions that the pipeline is specifically designed to resolve. To guarantee this level of specificity, several user-defined parameters are introduced at the beginning of the process, as detailed in Section 2.2.9. The selection of specific parameters was empirically determined by systematically testing multiple combinations in 15 samples within a defined genomic region. The final parameter set was chosen based on the highest concordance between CHARONTE-derived genotypes and the previously established structural configuration of the region, as determined by manual analysis or using GeONTIpe.

In parallel with the design and subsequent mapping of k-mers onto the reads, the pipeline also extracts additional information from the individual's mapping file. This includes read mapping positions, alignment flags, and overlap relationships between reads. These data are used to classify the reads and determine which segments of the reference genome are spanned by each read. This information is crucial for identifying the most relevant reads to infer the underlying structural haplotype. Finally, the pipeline generates phasing information derived from the analysis of heterozygote SNPs in the same individual, which is provided as an additional input to guide the reconstruction of the haplotype structure. A phasing result was considered correct when a single block was obtained and this block covered at least 80% of the length of the original sequence.

Haplotype determination

This block represents the central component of the pipeline as it constructs the structural haplotypes present in each sample. After correcting, filtering, and integrating the k-mers detected in the long reads, the pipeline classifies the reads into groups in which the two haplotypes can be found. This classification relies on similarity patterns derived from k-mer positions and strand orientation, prioritising the more informative reads (containing a higher number of k-mers and overlapping the regions where SVs could be expected to occur) and subsequently incorporating the remaining reads through iterative refinement (see Section 2.2.9 for details).

Throughout this process, the pipeline identifies structural inconsistencies, resolves duplicated or inverted regions, and removes conflicting information, ensuring that each haplotype is supported by a coherent set of k-mers. The final classification step evaluates whether each haplotype meets the required read-support thresholds, allowing the pipeline to distinguish complete, partial, or erroneous haplotype reconstructions.

The resulting output consists of the two reconstructed haplotypes per sample (or one haplotype for regions on the X chromosome in males), represented as consensus sets of k-mers with their positions and strands. Additional diagnostic images and a detailed log file are also generated, allowing visual inspection of the classification process and the assessment of structural differences between the sample and the reference genome.

Synteny

The final stage of the pipeline performs a comparative analysis on all analysed samples to identify all the different haplotypes detected. The generated haplotypes are grouped according to similarity, and the differences between haplotypes are characterised, classifying the variants in multiple types, including inversions, insertions, and deletions. The frequency of each haplotype is also calculated in the analysed group of samples.

After these comparisons, different structural haplotypes are named, and a list of individual genotypes is generated, together with a summary table reporting haplotype frequencies. Additionally, each haplotype is compared with the reference genomes GRCh38 and T2T-CHM13, providing a record of haplotype-specific differences. Finally, a synteny plot is produced to visualise the structural relationships and variation among the haplotypes.

Over time, the design of the pipeline has undergone systematic changes and improvements. To illustrate these developments, we will describe three regions that were used as test cases for the pipeline. In the initial version, no phasing information was used, and the read classification relied solely on structural differences. As a result, homozygous samples for a given structure presented a single haplotype, in which reads from both chromosomes were combined. In the initial stage of development, to simplify the design and benchmark, the selected regions were located on the X chromosome, where males have only a single possible haplotype.

3.4.4 HsInv0401 Region

The first test case focused on the HsInv0401 region located on the X chromosome (chrX:52,467,739 – 52,551,803), spanning approximately 85 kb (Figure 3.21). This region had previously been analysed using GeONTipe showing the presence of an 8.3-kb inversion and additional insertions of different sizes at both inversion BPs. In the manual analysis performed for the publication associated with GeONTipe, nine distinct haplotypes were identified. These haplotypes included the inversion located between 30-kb inverted SDs, as well as two CNVs embedded within these duplications, each affecting fragments of approximately 12 kb that include two copies of the *XAGE1* family gene. In the previous analysis, this 12-kb sequence was found to reach up to three copies per side, with a maximum of five total copies within a single haplotype.

To test the pipeline, we first selected a set of 15 individuals for whom the expected haplotypes were already known from manual analysis, enabling the identification of clear reads defining each haplotype. The k-mer parameters used were: k-mer length

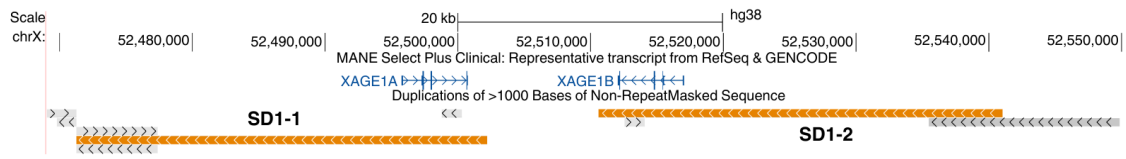


Figure 3.21: **HsInv0401 genomic region.** Adaptation from the UCSC Genome Browser view of the HsInv0401 region showing the SD structure track and MANE gene annotations. The main SDs two copies are labelled.

of 27 bp, separation between k-mers of 500 bp, and a minimum of three differences with any other sequence within the region. In this dataset, the pipeline recovered eight distinct haplotypes, correctly capturing both inversion and CNVs, achieving complete concordance with manually curated haplotypes.

After completing this initial test, we applied the pipeline to a larger dataset comprising 213 individuals from our mapped sample cohort. In this analysis, GRCh38 was also included as a haploid haplotype. The complete dataset consisted of 80 males, 131 females, and two haploid samples (GRCh38 and T2T-CHM13), which yielded a total of up to 344 possible haplotypes. The first objective was to evaluate the performance of the pipeline. In total, 299 haplotypes were resolved successfully (86.91%). Among the 45 (13.08%) unresolved cases, 42 (93.33%) were classified as partial, 39 of which originated from female samples. The remaining three haplotypes were classified as erroneous (HG02051 and the two haplotypes of HG01281) showing a substantial number of reads that could not be assigned to any constructed haplotype.

In this evaluation, the pipeline showed clear gender-specific performance differences. Male samples, together with haploid references (T2T-CHM13 and GRCh38), achieved the highest performance, reaching 95.12% success, while the performance in females was 84.35%, which was expected given that males are effectively haploid for the X chromosome and only a single structural configuration needs to be reconstructed. Among the 299 resolved haplotypes, 18 distinct structural configurations were identified. To examine the structural variation identified, we converted the observed insertions and deletions into differences in the copy number of the 12 kb segment at each breakpoint, as well as the orientation of the internal region. This

Table 3.13: **Summary of the haplotypes in region HsInv0401 identified by CHARONTE.** Each row represents a distinct haplotype and reports the copy number of the 12 kb segment within each segmental duplication, together with the orientation of the intervening segment. Additionally, the table includes the number of alleles detected for each haplotype and their corresponding frequencies in the dataset.

Variants	CNV SD1	CNV SD2	Orientation	N ^o haplotypes	Frequency (%)
H1	1	1	Std	7	2.34
H2	1	2	Inv	63	21.07
H3	1	3	Inv	103	34.45
H4	2	1	Std	25	8.36
H5	3	1	Std	28	9.36
H6	1	1	Inv	27	9.03
H7	1	2	Std	12	4.01
H8	1	3	Std	14	4.68
H9	2	2	Std	2	0.67
H10	1	4	Inv	5	1.67
H11	4	1	Std	1	0.33
H12	2	1	Inv	4	1.34
H13	3	2	Inv	1	0.33
H14	1	5	Inv	1	0.33
H15	2	2	Inv	1	0.33
H16	2	3	Inv	3	1
H17	2	3	Std	1	0.33
H18	3	1	Inv	1	0.33

information is summarised in Table 3.13, and the resulting comparison between haplotypes is illustrated in the synteny plot shown in Figure 3.22.

Among the detected haplotypes, the maximum total number of copies of the 12-kb sequence was six, with a minimum of two. When assessed per side of the inversion, the copy number located in the first segmental duplication ranged from one to four copies, whereas the one within the second duplication ranged from one to five.

Regarding inversion status, the inverted allele was the most frequent, reaching 69.90%, a proportion similar to that observed using GeONTIpe (71.99%). In contrast, for CNVs, haplotypes carrying four copies of the sequence were the most common (represented in six haplotypes), accounting for 49.83% of alleles (149 alleles). Most of the alleles (82.27%) belonged to one of the five main haplotypes. The most frequent, H3, was detected in 103 alleles and carries the inverted orientation, one copy of the CNV in SD1 and three copies in the CNV in SD2. Six haplotypes

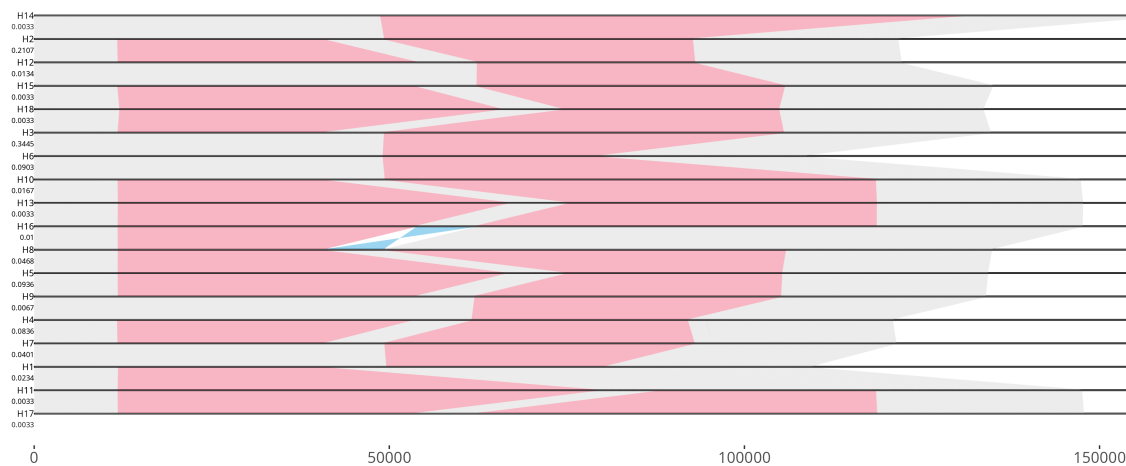


Figure 3.22: **Synteny plot comparing haplotypes from HsInv0401.** Consecutive k-mer blocks are displayed in pairwise (1-vs-1) haplotype comparisons. Grey segments indicate no detectable differences in distance or orientation between haplotypes. Pink segments highlight insertions and deletions, while blue segments mark inversion events. All haplotypes above the inversion event carry the inverted allele with respect to the reference genome GRCh38. The left panel lists the name of each haplotype along with its corresponding frequency.

were represented by a single allele, each accounting for 0.33% of the frequency. The two reference genomes showed different structures: GRCh38 corresponds to H1, a rare allele found only in a frequency of 2.34%, while T2T-CHM13 carries H2, which has a frequency of 21%. An alternative explanation, beyond the selection of a rare allele in GRCh38, is that this configuration may result from assembly artefacts, such as the collapse of the 12-kb repeat sequence into a single copy or an incorrect assignment of the inversion orientation.

To validate the results, we compared the inversion status obtained with CHARONTE with that reported by GeONTIpe. Among the 192 samples for which both methods produced results, the concordance reached 99.47% (191 out of 192). The single discordant case corresponded to sample HG02559, for which CHARONTE detected only the standard allele. Manual inspection showed that the second haplotype lacked sufficient read support to be validated by the pipeline. Next, we compared the additional SVs reported by GeONTIpe based on the presence of reads with insertions compared to the reference genome (these are not complete genotypes), with those detected by CHARONTE. Among the 192 comparable samples, 167 (86.98%)

showed complete concordance, while 25 presented discrepancies. To confirm the presence of these variants, we inspected the raw GeONTIpe read data, finding that 17 of these samples indeed contained additional SVs in the reads; however, these were not present in enough proportion to be reported by GeONTIpe. Taking this into account, the concordance increased to 95.83%. The remaining eight discrepant samples corresponded to cases in which GeONTIpe reported SVs not identified by CHARONTE. In seven of these, CHARONTE produced a homozygous structural call, suggesting that the alternative structure was not detected due to insufficient supporting reads for the second allele. In the remaining case (HG03548), the source of the discrepancy is not straightforward. CHARONTE reported an H1/H8 genotype, and while H8 could be validated based on shared variant signals, H1 differed due to the presence of an additional 12 kb copy in SD1 according to GeONTIpe. In total 9 individuals out of 192 were incorrectly genotyped by CHARONTE, which results in a 95.3% rate of correct genotypes.

3.4.5 HsInv0390 Region

After the successful outcome obtained for the first region, we decided to test a second region with a polymorphic inversion where additional SVs had also been identified by GeONTIpe. The selected candidate was HsInv0390, a region of 250 kb in chromosome X (chrX:103,880,293-103,133,167) characterised by several sets of SDs, the largest of which are approximately 70 kb in length. Another reason for selecting this region was the presence of multiple copies of the *TMSB15* and *H2BW* genes (Figure 3.23).

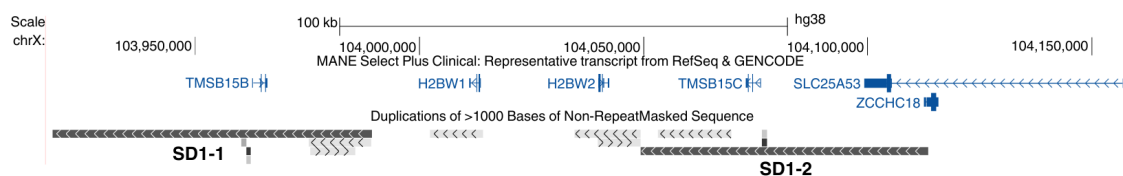


Figure 3.23: **HsInv0390 genomic region.** Adaptation from the UCSC Genome Browser view of the HsInv0390 region showing the SD structure track and MANE gene annotations. The two copies of the main SDs are labelled.

The same strategy used for the previous region was followed here. A small subset of

samples was used to evaluate and refine the pipeline. Once this initial set had been validated both through the pipeline and by manual read inspection, we proceeded with large-scale genotyping using the same 213 samples analysed for the previous region. Given the strong performance of k-mers observed in the earlier analysis, we chose to apply the same parameters in this case.

Using this approach, CHARONTE was able to resolve a total of 240 haplotypes (69.77%), which is a lower performance compared to HsInv0401. Among the 104 unresolved haplotypes, 60 corresponded to partial haplotypes; of these, 2 were found in male samples, while the remainder were associated with female samples. The remaining 44 unresolved cases were characterised as erroneous haplotypes. As observed in the previously tested region, male samples tend to perform better because only a single haplotype must be generated, achieving a success rate of 75.60%, while female samples showed a success rate of 67.94%.

Among the 240 resolved haplotypes, we identified eight distinct haplotypes. These included the 60-kb inversion previously reported by GeONTIpe, which is found with a frequency of 47.04%, as well as a 10-kb deletion located within the second 70-kb segmental duplication, which is present in all haplotypes except H1, found only in GRCh38 and other two samples. This deletion does not appear to be associated with any specific repeat. In addition to this deletion, the region that showed the highest diversity is located near the start of the region, where we detected an insertion and a deletion of approximately 5.5 kb involving sequences within a LINE1, more specifically an L1PA2 element. This short deletion of LINE1 was present in 14.79% of the haplotypes, while the version with the insertion was found in a single haplotype (H6). A final variant of interest was a 45-kb inverted duplication that lies within the single-copy segment located between the SDs in the reference genome, while the second copy is located near the beginning of the region, upstream of the first segmental duplication. However, this variant was observed only on a single chromosome of individual HG00096. The variation found within this region is shown in the synteny plot generated by CHARONTE (Figure 3.25).

As in the previous region, GeONTIpe data were used to compare both the inversion

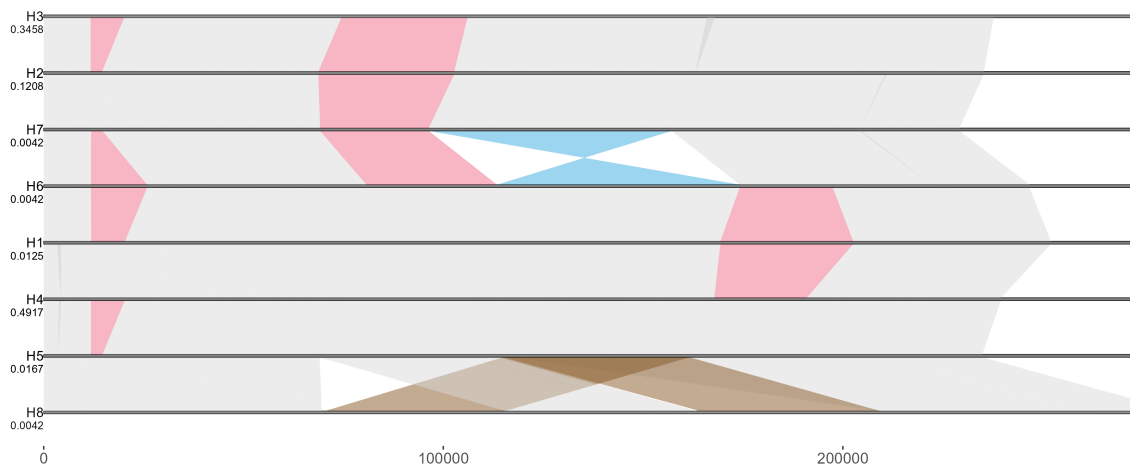


Figure 3.24: **Synteny plot comparing haplotypes from HsInv0390.** Consecutive k-mer blocks are displayed in pairwise (1-vs-1) haplotype comparisons. Grey segments indicate no detectable differences in distance or orientation between haplotypes. Pink segments highlight insertions and deletions, while blue segments mark inversion events. The inverted duplication is shown in brown. The left panel lists the name of each haplotype along with its corresponding frequency.

status and the presence of additional SVs. However, in this case, there are additional aspects to consider. The LINE-associated variants could not be assessed with GeONTIpe because the target sequences generated by this technique did not capture this region, so we could not check the accuracy of this part of the structural haplotypes. However, to confirm that these were real insertions and deletions and not artefacts generated by the calculations within the analysis pipeline, these variants were validated using an alternative approach: by calculating the distance between the two diagnostic k-mers flanking the variable LINE element in reads containing both of them. The three alleles (like the reference genome, deletion and insertion) were validated in this analysis.

Regarding the inverted duplication identified in the individual HG00096, a dot plot (Figure 3.25) was generated using a read that covered the entire region and comparing it with the reference genome. This analysis was performed because the inverted duplication caused an heterozygous genotype in a male individual that had been classified by GeONTIpe as erroneous.

Regarding the comparison of variants identified by CHARONTE and GeONTIpe,

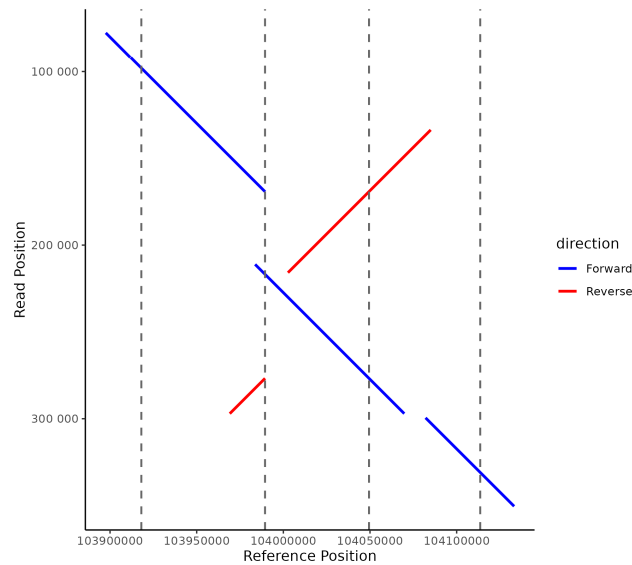


Figure 3.25: **Analysis of a read from region HsInv0390 in individual HG00096.** Dot plot of the read aligned to the reference genome, showing the alignment segments (hits) and illustrating both the continuity and interruptions in the correspondence between the two sequences. In the plot, grey dashed vertical lines mark the coordinates of segmental duplications in the region. Line colours indicate the strand orientation of each alignment relative to the reference, blue for forward and red for reverse. The inverted duplication is visible as one forward and one reverse alignment within the internal section of the region, while alignment discontinuities reflect insertions or deletions present in the read.

the following observations were made. Among the 152 individuals for whom the inversion orientation could be compared, 148 (97.36%) showed concordant genotypes. The four discordant samples corresponded to female individuals who were classified as homozygous for the same haplotype (H4). Upon inspection of the CHARONTE output files for these cases, it was observed that although some reads were assigned to a separated group, their proportion was not sufficiently high to support the generation of a second haplotype for the region.

Regarding the remaining SVs, 148 out of 152 individuals also showed compatible calls, particularly for the 10 kb deletion located in the second segmental duplication. Of the four discrepancies, three were attributable to the absence of the variant in the GeONTipe output. However, when examining the raw data for these individuals, at least one read carrying the variant was present in each case; GeONTipe did not report it because of its internal threshold of requiring more than a single supporting

read to avoid false positives. The remaining discordant call appeared to be the result of CHARONTE incorrectly attempting to construct two different haplotypes in a male sample on the X chromosome (this version of CHARONTE does not use the sex of the individual as input).

The analysis of these two regions (HsInv0401 and HsInv0390) provided the first insight into both the strengths and limitations of the newly developed software. In these initial tests, no phasing information was incorporated, which occasionally led to haplotype mixing or erroneous homozygous genotypes. Although the proportion of genotypes that can be determined and the rate of genotypes with independent concordant data supporting them is high, it was revealed that some samples carrying two distinct haplotypes are the most challenging and generate most of the errors.

For this reason, when extending the analysis to autosomal regions (described in the following section), we incorporated phasing information obtained using NanoCaller, WhatsHap, and SHAPEIT4, as detailed in Materials and Methods. This integration was expected to improve the read classification, as the combination of heterozygous SNPs with k-mer-based assignments could be critical for the accurate characterisation of complex structural haplotypes. Another improvement derived from the phasing step was the automatic generation of two haplotypes even when both showed identical genomic structures.

3.4.6 HsInv0231 Region

The genomic region selected to evaluate the updated version of CHARONTE was HsInv0231. This region of ~ 180 kb is located on chromosome 1 (chr1:25,248,883 – 25,435,328). It contains two pairs of SDs, one pair in inverted orientation (60 kb) and another in direct orientation (9 kb) flanking the first copy of the inverted pair (Figure 3.26). This architecture initially suggests the potential presence of an inversion of the sequence between the inverted SD pair, as well as a deletion between the direct copies that could remove the intervening sequence. In this context, such a deletion can directly affect the *RHD* gene, which is located entirely within the first copy of the long SDs in GRCh38, along with its analogue *RHCE* inside the sec-

ond copy. Additionally, the region susceptible to inversion includes the *TMEM50A* gene. Previous analyses of this locus using GeONTIpe revealed a deletion of 70 kb occurring at a relatively high frequency generated by the recombination of the direct short SDs, whereas the inversion itself appeared to be considerably less common.

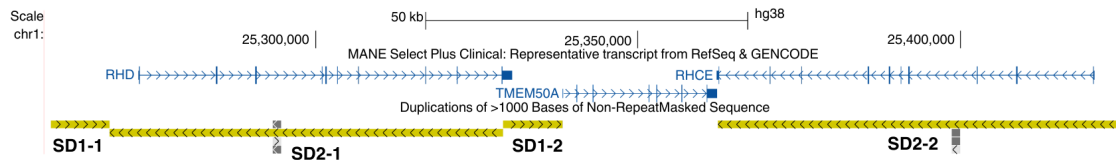


Figure 3.26: **HsInv0231 genomic region.** Adaptation from the UCSC Genome Browser view of the HsInv0231 region showing the SD structure track and MANE gene annotations. The main SDs two copies are labelled.

The strategy used to assess this region followed the same general approach as in previous analyses. A small subset of samples was first selected to optimise the execution parameters. In this case, the testing phase also served to refine and validate the pipeline phasing step, which had not yet been incorporated in earlier analyses. Once preliminary tests yielded reliable results, the analysis was extended to a larger dataset. For these tests of the updated version of the pipeline, an even greater number of samples was included compared to the previous analyses: a total of 456 samples (205 males, 249 females, T2T-CHM13 and GRCh38), corresponding to 910 possible haplotypes.

The first step was to determine the proportion of the region that could be successfully phased, and we considered samples with more than 80% of the region confidently phased as passing the threshold. Based on this criterion, 337 out of 455 samples were successfully phased in this region, while 12 samples completely failed this step, showing no heterozygous SNPs. Despite phasing failures in some samples, all samples were included in the structural haplotype analysis to evaluate whether a lower phasing coverage could negatively affect genomic structure reconstruction. Of the 910 haplotypes considered (including GRCh38), CHARONTE successfully determined 809 haplotypes, which corresponds to a success rate of 88.90%.

Upon examination of the performance of the samples according to phase quality. Among the 337 samples with more than 80% of the region successfully phased,

CHARONTE success rate was 93.76%, compared to 74.89% in the group of samples that were unphased or had phasing coverage below this threshold.

Among the 809 resolved haplotypes, five distinct haplotypes were identified. These included the previously reported 70-kb deletion, originally detected by GeONTipe, which was present in 16.81% of all chromosomes. In addition to this variant, 11 chromosomes showed evidence of a 32-kb inversion. In addition, a second inversion of approximately 115 kb was also detected on three chromosomes of three different samples. The BPs of this longer inversion are located inside the inverted SD pair, so this inversion includes the 32-kb sequence of the smaller inversion and also additional k-mers at both sides that were also inverted (Figure 3.28). This represents the first instance in which CHARONTE has captured a recurrent inversion event. This recurrence was supported by three distinct alleles, all originating from individuals of South East Asian (SAS) and East Asian (EAS) ancestry. Finally, an inverted duplication was identified that affects 21 kb within the internal region between the SDs in three haplotypes, also belonging to individuals from the same populations mentioned above.

Subsequently, these structural haplotypes were subjected to validation through comparison with previous information. First, we compared the inversion status of the detected haplotypes with GeONTipe results. However, it was not possible to discriminate between the two recurrent inversion events identified by CHARONTE. Among the 13 individuals with inverted alleles for whom the comparison was feasible, 12 showed concordant genotypes. For the remaining sample, inspection of the raw data revealed that CHARONTE could not assign a substantial number of reads to any haplotype and the standard allele was lost. Examination of several of these reads showed support for the inverted configuration, suggesting that CHARONTE mixed reads from different haplotypes into one.

We then proceeded to validate the deletion genotypes. For the 70-kb deletion, all 113 comparable samples showed complete concordance between both methods. Additional deletions are shown in the comparison between the H2 and H4 haplotypes in the synteny plot (Figure 3.27). However, these deletions are an artefact derived

from the comparison represented in the figure involving H2 (70-kb deletion, Std orientation) with H4 (non-deleted, Inv orientation). In fact, both haplotypes derive from H1 (non-deleted, Std orientation), and no other SVs were detected in this region.

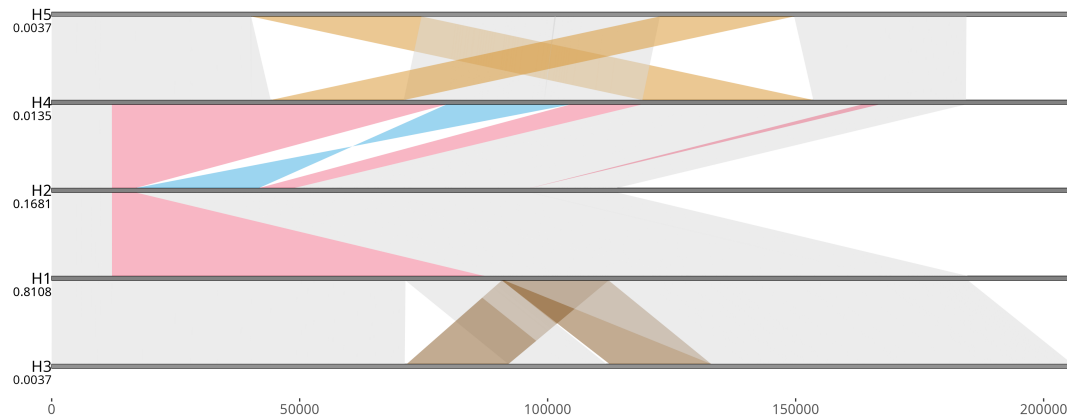


Figure 3.27: **Synteny plot comparing haplotypes from HsInv0231.** Consecutive k-mer blocks are displayed in pairwise (1-vs-1) haplotype comparisons. Grey segments indicate no detectable differences in distance or orientation between haplotypes. Pink segments highlight insertions and deletions, blue segments mark the inversion event, and orange segments indicate the k-mer blocks that are shifted between the two inverted haplotypes. In brown are shown the inverted duplication found, being two segments, one corresponding to the original copy and other showing the location of the second one. The left panel lists the name of each haplotype along with its corresponding frequency.

For the inverted duplication, as in the case of HsInv0390, we searched for reads that spanned the entire region and generated a dot plot to compare it with the reference genome to confirm the observed structure. Due to the detection of a recurrent inversion event involving a longer sequence, we further investigated the locations of the associated breakpoints. According to the k-mer analysis, the 32-kb inversion showed breakpoints located close to the most internal part of the 60-kb SDs. In contrast, the 115-kb inversion displayed breakpoints located in the central region of these SDs. To better visualise these variants, a synteny representation was generated using H1 as a reference and showing each inversion event on either side (Figure 3.28).

In general, we applied our analysis methodology to three complex genomic regions in datasets of 213 to 455 individuals with long-read data, where we identified 5 to 18

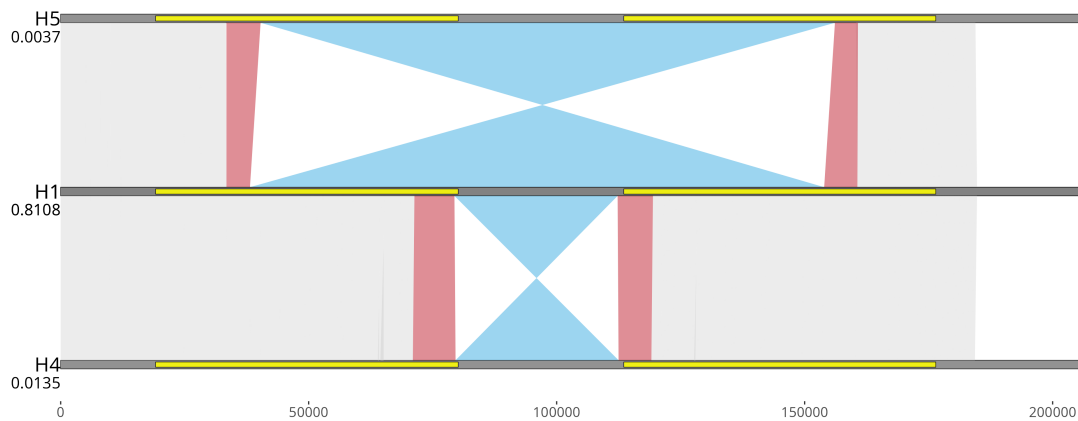


Figure 3.28: **Inversion recurrence of HsInv0231.** Synteny plot comparing haplotype H1 with H4 and H5, illustrating the two detected inversions. Grey regions indicate segments with no differences between consecutive k-mers, whereas inverted segments are shown in blue. Yellow bars mark the 62-kb segmental duplications arranged in an inverted orientation in the three haplotypes. Red blocks highlight the k-mer intervals in which the inversion breakpoints are located. The haplotype names and their corresponding frequencies are shown on the left side of the figure.

different structural haplotypes in different regions and estimated their frequencies, revealing extensive and previously uncharacterized haplotypic diversity. Across regions, our method determined 70–89% of the haplotypes analysed, and comparison with previous information on individual SVs showed an accuracy of 95–99% for these genotypes. Overall, this approach enables systematic characterisation of complex structural haplotypes and provides a robust foundation for studying their functional and evolutionary relevance in human populations.

4. Discussion

More than a century has passed since the first evidence of variation within the genome (Sturtevant, 1917), opening the door to the study of genetic diversity and its underlying determinants. Early discoveries focused on large structural variants (SVs). However, the intrinsic complexity of studying many of these variants led to a gradual shift in attention towards simpler forms of variation, particularly single nucleotide polymorphisms (SNPs), whose relevance increased substantially as sequencing costs decreased. As a consequence, the study of SVs, and inversions in particular, has remained comparatively underexplored.

The main focus of this thesis has been on inversions and associated complex regions, which have frequently been excluded from large-scale human genomic studies due to their intrinsic properties. As a result, their functional consequences remain largely unknown but for a limited number of notable exceptions that include large and well-characterised inversions, such as those in 8p23.1 and 17q21.31 (Campoy et al., 2022), as well as inversions with direct clinical consequences, exemplified by those associated with haemophilia A (Turner et al., 2006). To shed light on the number, population frequency, and functional effects of human inversions, the the INVersion Functional & Evolutionary STudies (InvFEST) project was established, with the goal of collecting and validating inversions reported in multiple studies over several years. This effort enabled the construction of a comprehensive catalogue of polymorphic inversions in humans to facilitate the investigation of their functional and evolutionary impact (Martínez-Fundichely et al., 2014).

In recent years, the InvFEST catalogue has been expanded through additional vali-

dation and genotyping efforts, particularly by Giner-Delgado et al. (2019) and Puig et al. (2020). These studies included the characterisation of 64 common inversions and introduced novel genotyping strategies that enabled more detailed analyses of their functional effects. More recently, the catalogue of inversions analysed was extended to 197 variants in the study by Campoy (2024), which provided an opportunity not only to increase both number of variants and genotyped samples, but also to implement methodological improvements to improve the detection of functional effects associated with inversions. However, despite these efforts, the fraction of variants that could be studied was still quite limited, and there was considerable room to improve the inversion catalogue.

At the same time, the continued development of long-read sequencing technologies from Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) and other technological advances like optical mapping or Single-cell template strand sequencing (Strand-seq) has enabled not only the identification of an increasing number of SVs (Ho et al., 2020; Audano et al., 2019; Ebert et al., 2021; Porubsky et al., 2022; Logsdon et al., 2025), but also a deeper characterisation of their properties. Although many recent studies have focused mainly on deletions and insertions, largely due to their abundance and easier detection, several studies have identified a growing number of inversion candidates that remain uncharacterized and may exert highly relevant functional effects (Levy-Sakin et al., 2019; Chaisson et al., 2019; Logsdon et al., 2025).

For this reason, the present thesis has focused both on the development of tools that leverage these emerging technologies, and on their application to the characterisation of inversions and other complex regions, with the aim of further expanding the catalogue of known genomic variants in humans and making this information accessible to the wider research community. Initially, we concentrated on the development of tools that exploit the advantages of ONT long-read sequencing for inversion genotyping. The result was GeONTipe, a pipeline that can detect both alleles of an inversion from long reads ONT. After developing and applying this approach on an initial set of 612 inversion candidates genotyped in 54 samples, we sought to

maximise both variant and sample coverage by integrating data from more than 13 independent studies. This integration encompassed over 21,000 reported variants, including both inversions and inverted duplications, resulting in a curated dataset of 1,235 inversions. Using this dataset, 1409 samples were genotyped and a detailed evaluation of the performance of our tool (GeONTIpe) was performed, examining how different factors influence our ability to resolve inversions. Subsequently, we focused on assessing the reliability of genotypes, as accurate genotyping constitutes a fundamental prerequisite for the downstream detection of functional effects.

The analysis performed revealed a higher variability around inversion breakpoints (BPs) than initially anticipated, consistent with the increasing number of complex genomic regions identified by recent studies (Ebert et al., 2021; Logsdon et al., 2025; Yilmaz et al., 2024). Recognising the limitations of our first approach, we developed a second methodology specifically designed to investigate regions exhibiting high structural variation. We hypothesised that such regions, due to their size and extensive rearrangements, are key contributors to genomic diversity. This second tool was applied to three distinct genomic regions, each with unique characteristics. Iterative improvements were incorporated after the analysis of each region, ultimately converging toward a framework that we believe will be essential to generate a final version capable of resolving variation within a wide range of complex regions. Therefore, collectively, the methods and analyses developed throughout this thesis contribute to filling a significant gap in our understanding of human genomic variation, and provide a robust foundation for future functional and evolutionary characterisation of these variants.

4.1 GeONTIpe advantages and limitations

One of the main objectives of this thesis was the development of a tool capable of accurately genotyping all types of inversion, and especially those mediated by IRs, most of which were missing from previous studies. This objective has been achieved with GeONTIpe, which has demonstrated its ability to fulfil this goal through the work performed in the thesis. The underlying concept of the tool is both simple

and powerful: the use of four probes, two external and two internal, that allow identification of the inversion state through analysis of the position, orientation, and distance of these probes. As discussed in more detail in the companion research article (Moreira-Pinhal et al. 2025), this methodology is highly useful and reliable in a large proportion of cases, with good concordance with previous genotypes and data from other studies, especially the Pangenome draft. However, throughout the analyses, regions emerged that proved difficult to resolve using this approach, highlighting some of the limitations currently faced by the current methodology.

First, it is important to emphasise the potential role of complex regions, defined here as regions in which there are multiple haplotypes with different structural conformations across individuals. Such regions do not appear to be merely rare exceptions: studies from the The Human Genome Structural Variation Consortium (HGSVC) have revealed several examples of quite complex regions (Ebert et al., 2021; Logsdon et al., 2025). In our case, regions containing a limited number of variants can be resolved with the help of the information generated by GeONTIpe, as shown by the HsInv0401 haplotypes described in Moreira-Pinhal et al. (2025). However, larger and more complex regions, such as the amylase locus or the Major Histocompatibility Complex locus (Yilmaz et al., 2024; Mosbrugger et al., 2025) remain too variable to be effectively handled by GeONTIpe. The use of only four probes, and in particular their size (300 bp), provides a limited number of independent SVs that can be tested and captured, identifying only a fraction of the true complexity of these regions. This limitation can affect all downstream analyses, making GeONTIpe not the best option for the complete resolution of these regions. We have partially attempted to address this issue through the development of CHARONTE, which promises to resolve some of these problems, although it introduces other limitations that will be discussed later.

A second relevant limitation that we aim to address is the current inability to resolve inverted duplications, or, more critically, that these variants can be frequently misclassified as heterozygous inversions. At present, GeONTIpe employs a read-filtering strategy in which reads containing multiple hits from the same probe are discarded.

By detecting an anomalously high number of such reads, we have identified some inverted duplications. However, this is limited to cases in which the duplicated sequence is small and close to the original location, and therefore there are many reads that can capture both copies of the duplicated probe sequence. For larger variants, where only one copy of the inverted duplication is detected, the interpretation depends on which side of the variant is analysed. This potentially leads to the classification of some reads as standard and others as inverted, resulting in the classification of that variant as an heterozygous inversion. We are currently exploring solutions for detecting these cases, either through prior analysis of mapping hits to identify duplicated sequences or by using CIGAR strings to detect sequence duplications. Similar approaches have been used in other studies to detect BPs in other SVs, such as tandem repeats (Wu et al., 2013). Incorporating a dedicated methodology to detect inverted duplications into GeONTIpe would not only increase the number of variants that can be analysed, but also improve the genotyping accuracy for variants already under study. Cases such as those of specific samples of HsInv0390 and HsInv0397 exemplify this limitation, since only deeper manual analyses allowed us to identify the presence of inverted duplications, which were previously misclassified as inversions.

Such improvements could also help address another issue that we have detected over time: the presence of what we have termed “mirror reads.” These reads appear to be sequencing artefacts and have been observed both in publicly available datasets and in samples sequenced by us. For example, Schloissnig et al. (2025) reported multiple variants that we suspect are a result of such artefacts. Furthermore, recent studies have begun to specifically target the detection of these artefacts in ONT reads (Heinz et al., 2025). Although the overall magnitude and impact of mirror reads are currently unknown, their direct mistake with true inversions may represent a source of genotyping inconsistency, especially if other features such as structural mosaicism want to be studied. Consequently, their detection and removal should be considered a high-priority task to improve GeONTIpe.

Finally, a further limitation of the software is related to computational constraints.

Although GeONTIpe was designed to run on high-performance computing (HPC) systems, particularly for mapping and genotyping steps, where computational demands are high and parallelisation significantly improves performance, there are still specific stages of the pipeline that act as bottlenecks. One such step is the use of BLAST for the detection of probes within reads. Several tools have been developed to perform similar tasks with substantially higher speed, such as MMseqs2 (Kallenberg et al., 2025). Improvements in this area would increase scalability, allowing the analysis of more samples and more variants in less time. This should be considered an important goal, as reductions in ONT sequencing costs and technological improvements, whether through increased output or advances such as adaptive sampling, are likely to substantially increase both the amount of data generated and the number of detectable inversions. An early example of this trend is the sequencing of 1,019 samples using ONT long-reads (Schloissnig et al., 2025), which highlights the potential of this technology for the study of inversions.

Having outlined some of the current limitations of GeONTIpe and the strategies we intend to pursue to address them, we now turn to the advantages this software offers over existing methods and the future perspectives we believe can be achieved through its continued development.

To understand the main advantage of GeONTIpe, it is necessary first to consider the strategy that is most commonly used for the characterisation not only of inversions, but also for SVs in general. Most studies, particularly those associated with large consortiums such as HGSC or HPC, are based on the use of multiple technologies to generate high-quality assemblies capable of resolving regions of varying complexity (Ebert et al., 2021; Porubsky et al., 2022; Logsdon et al., 2025). This approach offers several advantages over the use of a single technology, as reflected in the resolution of extremely complex regions, such as the MHC, and the wide characterisation of multiple types of variants (Logsdon et al., 2025). However, it also presents several relevant limitations.

The first limitation is cost. The production of high-quality assemblies requires not only substantial computational resources, but also the integration of multiple tech-

nologies, which significantly increases overall costs, thereby limiting the number of samples that can be studied. This constraint is exemplified by the maximum of 65 samples analysed using this approach in Logsdon et al. (2025). With such limited sample sizes, the scope of the subsequent analyses and the robustness of the conclusions drawn from the resulting genotypes are necessarily restricted. The use of a single technology mitigates this issue by reducing costs, which in our case allowed the analysis of more than 1,400 samples. This strategy is also reflected in several recent initiatives that aim to generate exhaustive SV catalogues in large cohorts using a single technological approach (Gustafson et al., 2024; Schloissnig et al., 2025).

On the other hand, the competition to establish a gold standard for long-read sequencing remains ongoing, as each technology offers distinct advantages and limitations, and research groups select platforms according to their specific objectives. In our case, for the study of SVs, we consider ONT to be the most suitable technology, having a great advantage of its ability to generate reads that can reach hundreds of kilobases in length, which makes it particularly well suited to solve large genomic rearrangements. In addition, apart from the cost associated with assembly generation, GeONTIpe enables a more focused analysis of specific genomic regions, a task with which global assembly-based approaches cannot compete without the addition of complementary region-specific analyses.

Moreover, genome assemblies inherently generate consensus representations, which results in the loss of certain types of information, such as the presence of mosaicism for specific variants. In contrast, by analysing individual reads, our approach preserves this information, allowing it to be leveraged to obtain a first estimate of mosaicism rates for inversions. In particular, given the recurrent nature of inversions mediated by inverted repeats (IRs), we hypothesise that these variants may exhibit elevated levels of mosaicism, making inversions ideal candidates for such analyses and positioning GeONTIpe as a suitable tool for the detection and characterisation of this phenomenon. In fact, we are currently working on incorporating mosaicism detection capabilities into GeONTIpe to enable the reliable study of inversion mosaic reads in samples sequenced using ONT long-read data.

To conclude, it is relevant to say that the field of inversion genotyping remains relatively niche, with only a limited number of tools that demonstrate robust and consistent performance. Beyond GeONTIpe, one of the few methods with a comparable objective is npINV, which also leverages ONT long reads for inversion detection and genotyping. npINV is primarily alignment-driven: it interrogates read mappings to identify subreads aligned to the same chromosome but in opposite orientations, and subsequently examines mapping overlaps to infer the presence of IRs potentially mediating the inversion (Shao et al., 2018).

Although conceptually sound, this alignment-centric strategy has several limitations compared to GeONTIpe. First, it is highly dependent on mapping accuracy, making it susceptible to spurious hits arising from alignment artefacts. This issue is particularly pronounced for smaller inversions, where reads can span the entire region without producing a clear split or discordant mapping signal, thus reducing sensitivity (Shao et al., 2018). Furthermore, regions harbouring additional SVs can generate complex sub-mapping patterns that obscure the inversion signature itself, rendering the approach insufficient in complex genomic contexts (Shao et al., 2018). The presence of “mirror reads” further compounds this problem, as these reads can introduce additional noise that interferes with accurate inversion identification and genotyping.

Some of these limitations are reflected in the genotyping consistency reported of 94% for npINV, which is lower than the values obtained with GeONTIpe in the comparative analyses performed in this work. As is also the case for GeONTIpe, the read length constitutes a critical limiting factor for both the detection of inversion and the characterisation of IRs. In the npINV study, the detection of IRs was limited in practice to repeats of up to 7 kb in length, since the tool only reported IRs within this size range (Shao et al., 2018). This is substantially shorter than several of the IRs associated with the inversions genotyped in this thesis. This constraint further highlights the importance of optimising strategies for handling long and complex repeat-mediated rearrangements. If we consider earlier methodologies for inversion genotyping, such as BreakSeq (Lam et al., 2010), the limitations become even more

evident. BreakSeq relies on predefined breakpoint libraries and is therefore restricted to inversions with clean and well-characterised BPs. Its performance is heavily dependent on the precise delineation of those. Moreover, this framework is inherently unsuitable for genotyping inversions mediated by IRs, where the resolution of the breakpoint is often ambiguous or structurally complex.

Together, these considerations underscore a clear methodological gap in the field. GeONTIpe addresses several of the limitations associated with alignment-dependent or breakpoint-library-based approaches, particularly in the context of repeat-mediated inversions and complex genomic backgrounds. In doing so, it establishes itself as a relevant and timely contribution to the landscape of inversion genotyping.

4.2 CHARONTE advantages and limitations

In the previous section, we discussed the limitations of GeONTIpe, several of which motivated the development of a new pipeline designed to overcome its main weaknesses. This pipeline, referred to as CHARONTE, was designed specifically to resolve complex genomic regions, in which the use of just four probes lacks sufficient discriminatory power to resolve a wide range of structural configurations that may arise in different chromosomes when multiple SDs are present.

To address this type of region, we opted to use k-mers of a few tens of bases, with a sufficiently large number of sequences distributed along the entire region to allow a more detailed resolution. Although in principle this approach is well suited for the characterisation of such complex regions, it has additional limitations, most of which are related to the scalability and reliability of the pipeline. In this section, we discuss the limitations identified during the different validation tests performed, propose potential solutions that should be explored in future work, and outline the prospective applications and problems that this tool may help to address.

The first major limitation identified concerns the classification of the reads of each individual in two haplotypes, which constitutes the foundation of our strategy for characterising these regions. Part of this issue has been addressed through the

incorporation of tools specifically developed for read phasing. The current strategy consists of using NanoCaller (Ahsan et al., 2021) to generate a VCF containing SNPs detected in the reads, followed by WhatsHap (Martin et al., 2016) to identify phasing blocks. When more than one phasing block is generated, we attempt to apply SHAPEIT4 (Delaneau et al., 2019), which uses SNPs derived from high-coverage 1000GP samples (Byrska-Bishop et al., 2022) to identify variants that can bridge the phasing blocks initially obtained.

This approach is inspired by previous studies aimed at resolving complex genomic regions, such as Mosbrugger et al. (2025). Although this strategy performs well in a substantial proportion of cases, we have identified scenarios in which reads that should be mutually incompatible are assigned to the same phasing block. The clearest examples correspond to reads harbouring large deletions that are phased together with reads lacking those deletions. Our current hypothesis is that a large fraction of the shared sequence between deleted and non-deleted reads exerts a disproportionate influence during phasing, leading to these misclassifications. Increasing the weight assigned to the absence of shared variants may therefore be necessary to improve read separation. However, we are also testing other software to improve read phasing.

To date, only a single test has been conducted aimed at improving phasing of SNPs, using LRphasing (Holmes et al., 2023). However, the results do not suggest that this constitutes an optimal solution. Consequently, a more promising avenue may involve improving variant calling itself by testing alternative callers that have demonstrated strong performance on ONT long reads, such as Clair3 (Su et al., 2022).

Beyond strategies directly related to read classification, other additional improvements could be implemented. These include analysing read depth along the sequence or incorporating mapping-derived information to pre-characterise reads, thereby providing additional features that could be weighted during classification. Several of such strategies are currently being considered. All of these measures are intended to improve the overall performance of the pipeline by being able to resolve a larger number of haplotypes, but also to reduce the number of erroneous or partial haplo-

types that have been identified throughout the tests conducted.

Finally, additional improvements to the framework should address the determination of optimal parameters for structural analysis. Currently, parameter selection has been performed manually, based on the inspection of a small subset of samples. Although this strategy is pragmatic, it inherently relies on prior knowledge of the expected structural configuration to assess whether a given parameter combination yields accurate results. Consequently, this approach is not fully generalisable to regions or datasets where ground truth is unavailable. Automating parameter optimisation represents a non-trivial challenge, given the multidimensional parameter space and the context-dependent behaviour of SVs. Nevertheless, several avenues could be explored. For instance, an adaptive strategy could evaluate the distribution and abundance of mapped k-mers across candidate regions, using quantitative metrics to identify parameter sets that maximise signal-to-noise ratio. In addition, incorporating genomic context into the optimisation procedure may prove beneficial.

A critical aspect of this optimisation process is the intrinsic trade-off associated with k-mer length. Short k-mers increase sensitivity but reduce specificity as they are more likely to map to multiple genomic loci. In contrast, longer k-mers provide higher specificity but are more vulnerable to sequencing errors, particularly in ONT reads, where base-calling inaccuracies may prevent exact matches. Given that the current implementation requires exact k-mer matches to accept mappings, excessively long k-mers may not be detected even when biologically present, thereby reducing sensitivity. Thus, parameter selection must balance these competing constraints to achieve robust inversion genotyping. Overall, while the current manual strategy has proven effective in controlled settings, the development of a systematic data-driven parameter selection approach would improve scalability, reproducibility, and applicability to previously uncharacterised genomic regions.

A second limitation concerns the lack of robust approaches to compare the results obtained, which we consider critical at this stage of pipeline development to improve confidence in the generated outputs. So far, our primary strategy has been to compare CHARONTE-derived haplotypes with those obtained using GeONTipe.

However, given that the two approaches differ conceptually, this comparison may not be fully accurate. Alternative strategies include analysing the genome assemblies generated for many of the samples used in our tests. These include samples from the first phase of the Pangenome developed by the HPC (Liao et al., 2023), as well as those from the second phase, which is expected to be released in the near future. In addition, other genome assemblies, such as those generated as part of the HGSVC project (Ebert et al., 2021; Porubsky et al., 2022; Logsdon et al., 2025) could also be considered as a validation method. Within this framework, the k-mers used for read characterisation could also be applied directly to characterise the corresponding assemblies, comparing the structure observed in both strategies.

A complementary strategy would involve the use of tools specifically designed for direct Pangenome analysis. A candidate could be PGR-TK (Chin et al., 2023), which enables multiscale decomposition of nodes and haplotype paths within human pangenome graphs. This framework restructures the paths into hierarchical blocks that capture shared structural patterns between individuals. Such a representation could facilitate comparison with the structural organisation inferred by CHARONTE. However, a current limitation lies in establishing an automated comparison framework, as both approaches rely on fundamentally different representations.

A third limitation of the complex region analysis is related to the intrinsic characteristics of the reads themselves. Samples with low N50 values or reduced read quality can negatively affect the continuity of the reconstructed haplotypes, which throughout this thesis we have referred to as partial haplotypes. The use of k-mers, and in particular the requirement for a 100% identity to accept a mapping on the reads, makes the sequence quality a critical factor. Limited overlap, either due to short read length or sequence errors, impairs the ability to compare reads and assign them to consistent groups. For this reason, we chose to exclude from the analysis the samples from Schloissnig et al. (2025), whose N50 values and read quality are relatively low. At this early stage of development and benchmarking, we considered it unnecessary to address the additional complexity associated with such data.

However, the ongoing efforts of multiple groups to generate high-quality ONT data (Gustafson et al., 2024; Logsdon et al., 2025), allow us to anticipate that it will be possible to analyse a substantially larger number of samples in the coming years. In parallel, continued technological improvements are expected to further increase sequencing output, read quality, and N50 values, a trend that has already been observed in recent years (Sereika et al., 2022; Zhang et al., 2024). Consequently, these advances will enable the generation of more robust conclusions regarding the diversity and functional or evolutionary consequences of complex genomic regions.

Despite these main limitations identified during the development and application of CHARONTE, we believe that this tool fills an important gap in current knowledge of human genome variation, particularly with respect to the characterisation of regions of intermediate complexity. Specifically, by this, we refer to genomic regions that span hundreds of kilobases and in which a limited number (few tens) of distinct haplotypes are expected to be present. Although such regions often contain multiple genes, they have not received the same level of attention as other extensively studied loci, such as the amylase region (Bolognini et al., 2024) or the KIR gene cluster (Mosbrugger et al., 2025). Despite that our efforts to date have focused on regions of this scale, we do not exclude the possibility that CHARONTE could be extended to the analysis of substantially larger regions.

A further major advantage of CHARONTE is its reliance on a single sequencing technology, an advantage it shares with GeONTipe. This significantly reduces the characterisation costs and facilitates the analysis of a much larger number of samples, thereby enabling future studies to uncover functional effects or reconstruct the evolutionary history of many of these regions.

4.3 Inversion catalogue

Leaving aside the software developed, one of the primary objectives of this thesis was to generate the most comprehensive inversion catalogue to date. The catalogue presented here comprises a total of 1,235 inversions, although only 680 of these

exhibit evidence of the presence of both alleles. Compared with previous datasets generated by our group, this represents a substantial increase in the number of successfully genotyped inversions. The largest catalogue previously reported by the group was presented by Campoy (2024), which included 134 inversions. Thus, in this study, we genotyped 5.05 times more variants than previously reported. In addition, since 77 inversions included in that study were intentionally excluded from the analyses performed here, the effective increase is even more pronounced. If we compare with even earlier studies of the group, such as Giner-Delgado et al. (2019) and Puig et al. (2020), the observed increase in variant genotyping reaches between 15 times. This highlights the considerable potential of ONT sequencing for the detection and genotyping of inversions.

In contrast with datasets generated by other research groups, our catalogue is also among the most extensive in terms of the number of variants analysed. For example, Schloissnig et al. (2025) reported 733 variants. However, a substantial fraction of these correspond to inverted duplications, and only 393 polymorphic inversions were identified with GeONTIpe. Furthermore, the number of inversions analysed in this work is consistently higher than in other studies that contributed data to our catalogue. Importantly, beyond the size of the catalogue itself, a key strength of this study lies in the large number of samples used to genotype each variant. In addition, we deliberately prioritised detailed genotype resolution and extensive validation steps to minimise potential errors. Although time constraints prevented us from implementing corrective analyses for certain regions that require additional scrutiny, the present work has provided a clear estimate of the number of variants likely to benefit from targeted re-evaluation.

Finally, due to current methodological limitations, we chose to exclude inverted duplications from the analysis. Nevertheless, we are actively exploring strategies to incorporate those variants from the catalogue, as well as to recover other inversions that could not be analysed due to the lack of suitable probes or the filtering steps.

4.3.1 Inversion characteristics

The complete catalogue of 1,235 inversions exhibits a mean inversion size of 140 kb, with variants broadly distributed across size classes and sizes ranging from 301 bp to 4.4 Mb, as shown in Table 3.5. By comparison, studies such as Audano et al. (2019) and Schloissnig et al. (2025) tend to report inversions within more restricted size ranges, largely reflecting the characteristics of the sequencing data used, although other studies such as Levy-Sakin et al. (2019) have reported much larger variants. In particular, the size distribution observed in our catalogue resembles more closely that reported in studies that combine multiple sequencing technologies, such as Chaisson et al. (2019), Ebert et al. (2021), and Porubsky et al. (2022), consistent with the integration of multiple datasets used to construct our catalogue.

It is important to note that particular emphasis was placed on inversions mediated by IRs. This focus is reflected by the systematic search for all segmental duplications (SDs) in inverted orientation across the genome (variants reported whose name begins with "Mos"), which were used as candidate regions for the discovery of new inversions. Moreover, the relevance of the detailed analysis of variants generated by non-allelic homologous recombination (NAHR) mechanisms can be observed in the extensive manual probe design applied to the initial dataset of 612 inversions used to develop and benchmark GeONTipe. Despite their intrinsic complexity, the original objective of the study was to genotype specifically this class of inversions. However, the success achieved motivated the inclusion of inversions with diverse mechanistic and structural characteristics, which ultimately led to the present dataset.

Another noteworthy aspect of this work was the ability to analyse inversions with BPs greater than those previously studied by our group. Before, digital droplet PCR (ddPCR) had been the method of choice for genotyping individual inversions with large inverted repeats (IRs), as demonstrated in Puig et al. (2020). Nevertheless, ONT long reads have proven to be highly effective in genotyping inversions with these characteristics. Despite the utility of this technology, considering the typical read length of the available data, we imposed an upper limit of 200 kb on the size of

IRs, thus excluding variants that are likely to be polymorphic at the population level. In fact, we have seen that the longer the IRs are, the more difficult the inversion is for the genotype. Despite this conservative choice, ongoing improvements in sequencing technology make us optimistic that such variants will become tractable in the near future.

4.3.2 Genotype accuracy

To be able to conduct the analysis of this thesis, we want to ensure that current genotypes are highly accurate. Therefore, we performed several comparative analyses to evaluate their reliability.

First, we examine the consistency of Mendelian transmission of our genotypes by comparing the genotypes of offspring from trios with those of their parents. Although this analysis could in principle have been extended to offspring from duos, we decided to exclude them at this stage, as such comparisons would be inherently less informative. In this way, we identified 20 Mendelian incompatibilities. In some cases, we suspect that the use of a permissive threshold for the minimum number of supporting reads may underlie these inconsistencies. However, this explanation is less convincing in other cases, particularly where all family members exhibit a high number of supporting reads (Table 3.8), and several alternative hypotheses can be proposed to explain these observations.

- **Sample contamination:** One possible explanation for the observed inconsistencies is contamination affecting one or more family members. However, contamination would be expected to affect not only one or two variants, as observed here, but multiple variants, particularly those with high AF, where contamination effects should be easier to detect.
- **Recurrence:** Inversions mediated by NAHR mechanisms can undergo recurrent recombination events, generating the alternative orientation and thus leading to apparent Mendelian inconsistencies (Giner-Delgado et al., 2019; Puig et al., 2020; Porubsky et al., 2022). This mechanism, combined with the

use of immortalised lymphoblastoid cell lines (LCL) that have undergone multiple passages, could explain the emergence of an inversion in a subset of cells within a cell line, followed by clonal expansion that increases its frequency in a given passage. Such events have already been reported for other SVs (Tosca et al., 2015). Recurrence could also manifest as somatic mosaicism, which has been described in several studies (Grimes et al., 2024) and is also suggested by our data, where reads supporting the opposite orientation were detected in samples classified as homozygous. Although this phenomenon was not explored further in this thesis, it has previously been proposed as a potential avenue for improving the developed software.

- **Genotyping errors:** As already shown, inversion structural rearrangements that deviate substantially from expectations or exhibit increased complexity may generate artefacts during genotyping. In such cases, one might expect at least one progenitor to carry a similar structure, resulting in comparable artefacts across samples. However, factors such as the number of reads long enough to fully span the structure may play a relevant role in these comparisons. Future work should include a systematic review of variants showing Mendelian inconsistencies, for example using dot plot visualisation for each observed orientation, and validating them using an independent methodology.
- **Sequencing artefacts:** Another potential source of error arises from sequencing artefacts. Throughout the analyses, we identified reads that appear to be artefacts, which we termed "mirror reads", also described in the literature as "fold-back reads" (Heinz et al., 2025). In our data, one of the two copies of the duplicated sequence seems to show a reduced sequence identity, suggesting that it corresponds to the secondary read, while the other represents the original read. As the proportion of such reads is currently unknown, it is possible that some of the observed genotyping errors arise from incorrect genotype classification driven by these artefacts. However, without a more detailed characterisation of mirror read properties, their precise impact on genotyping accuracy remains uncertain.

Importantly, however, the overall frequency of the observed Mendelian discrepancies is extremely low, 0.071%, corresponding to 20 inconsistencies out of 28,002 comparisons. This is 7-70 times lower than the values reported by Porubsky et al. (2022), who observed an overall consistency of 95% or 99.5% when restricting the analysis to high-confidence genotypes. However, this difference may partially reflect methodological differences between studies or variations in the way inconsistencies were quantified.

The explanations discussed above may also explain some of the discrepancies observed when comparing our genotypes with previous experimental results. The general consistency with the experimental data was high, reaching 99%. However, in this case, errors may also arise from the PCR-based approaches used to genotype the inversions. In fact, when excluding variants for which experimental genotypes are suspected to be erroneous, the consistency increased to 99.66%. The case of HsInv0348 illustrates this issue well. Due to the presence of several duplications, the reliability of the iPCR assays has long been questioned within the group and was ultimately reassessed using ONT data. Although the iPCR results for this variant appear to include more errors than the genotypes produced in this thesis, some uncertainty remains. Specifically, analysis of LD revealed values close to, but not equal to, 1, suggesting that a small subset of ONT genotypes may still be incorrect. To address this issue, four alternative probe sets were designed for this variant. However, none of them produced substantial improvements: all probe sets showed similar levels of discrepancy with the experimental data, and the first set exhibited not only the strongest linkage but also the best overall performance. Therefore, the original probe set was retained.

These genotyping inconsistencies could also have an effect on the analyses HWE, where 10 inversions were found to deviate from the expected proportions of genotype. We investigated whether a large amount of additional SVs associated with one of the orientations could explain these deviations, but this did not appear to be a satisfactory explanation in most cases. For some inversions, a possible explanation was provided in Table 3.10, but for others it remains unknown. Consequently,

some of the alternative hypotheses previously discussed could be the most plausible sources of error.

4.3.3 LD analysis

In general, LD showed that 145 of the 680 polymorphic inversions exhibited a strong linkage ($LD \geq 0.95$) with nearby variants at the global level. The application of stricter genotype-quality filters resulted in only marginal increases in high-LD signals, suggesting that read-support thresholds do not seem to represent the main source of the observed low LD values. In particular, several inversions lacking strong global LD displayed high LD within specific ancestries, indicating a possible influence of population-specific haplotype structures. For inversions in which strong linkage was not detected, multiple factors can explain the observed patterns, including recurrent inversion events or the age of the inversion, both of which can substantially affect LD with surrounding variants.

Recombination processes may contribute to the observed discrepancies. Although inversions typically suppress single crossovers in heterozygous individuals, residual recombination can still occur through gene conversion within the inverted segment or via crossovers in the flanking regions, reshuffling haplotype backgrounds. These mechanisms may generate apparent transfers of variants between orientations or locally disrupt otherwise strong linkage patterns, particularly in cases with LD values close to 1. To evaluate these hypotheses, several complementary analyses would be required. Ideally, an approach similar to that previously used to assess recurrence should be applied (Giner-Delgado et al., 2019; Puig et al., 2020; Porubsky et al., 2022). By constructing a phylogenetic tree based on the alleles detected in all the SNPs in the region and associating them with inversion orientation, a unique origin could be inferred if the haplotypes with the same orientation cluster together, even in the absence of fixed variants. In contrast, if different haplotypes of one orientation cluster with those of the opposite orientation, this could suggest recurrence events or gene conversion, depending on the observed distribution of variants (Giner-Delgado et al., 2019). However, so far this analysis is relatively manual and time consuming,

and was out of the scope of this thesis.

Therefore, it would be important to distinguish well between the different types of variants under study. An initial approach would be to determine the underlying mechanisms responsible for generating each variant, which will provide information into which inversions could be expected to show LD values close to 1.

To address these questions in a comprehensive way, a more accurate characterisation of the inversion BPs would be required. Although this has traditionally been performed by manual analysis, the increasing number of variants currently under study renders this approach impractical. Therefore, the development and implementation of automated alternatives would be highly desirable. One possible strategy involves leveraging the read-mapping information to distinguish between two scenarios. The reads aligned on both sides of the inversion that do not overlap indicate the presence of clean BPs, while the reads that present overlapping hits suggest the presence of repetitive sequences in the inversion BPs, as expected for inversions mediated by IRs. However, attempts to apply this strategy using ONT data have yielded ambiguous results, likely due to the presence of multiple spurious hits caused by the “mirror reads” previously discussed. Despite these limitations, improving the resolution of IRs and the real BPs constitutes a critical step towards a deeper characterisation of the variants and could ultimately contribute to the detection and correction of potential genotyping errors.

For inversions mediated by NAHR, we would not expect to observe strong LD patterns, particularly for highly recurrent variants, as previously reported (Giner-Delgado et al., 2019; Porubsky et al., 2022; Campoy, 2024). In cases where linkage is detected, as observed for some inversions in our analysis, this could suggest either a low recurrence of the inversion or that no individuals carrying alternative recurrent events were genotyped in our dataset.

By contrast, for most inversions mediated by non-homologous mechanisms, LD values close to 1 would be expected, as previously reported (Giner-Delgado et al., 2019; Campoy, 2024). Nevertheless, several factors could still explain the absence of link-

age. As already mentioned, one possible explanation is related to the age of the inversion. At the initial emergence of an inversion, if its formation mechanism does not generate novel variants at the BPs, no variants fixed to the new orientation would be expected. However, in successive generations, the loss of ancestral haplotypes, combined with the accumulation of new mutations and the absence of recombination between the two orientations, could progressively lead to the appearance of variants fixed in the inverted orientation. This scenario would be particularly plausible in the initial stages after inversion generation and for low-frequency inversions, where genetic drift may have a strong impact and frequencies have not yet increased substantially. For inversions at intermediate frequencies, the absence of fixed variants could indicate that the frequency increase was driven by positive selection, allowing the inversion to rise rapidly before linked variants had time to accumulate.

Additional explanations for the lack of strong LD between inversions and surrounding markers include genotyping errors, either in inversion genotypes produced by GeONTIpe or in the genotyping of nearby variants. In fact, we have already used LD as a metric for genotype validation, especially for those variants that displayed LD values close to 1 but did not show perfect linkage with nearby SNPs. The impact of inversion genotyping errors on LD is expected to be multifactorial, as both the number of successfully genotyped samples and the inversion frequency influence linkage estimates. For example, inversions genotyped in a small number of samples are more susceptible to a decrease in linkage due to errors than those supported by many genotypes. A similar effect is expected for low-frequency inversions, where a single genotyping error in one allele can disproportionately affect LD estimates.

Thus, taken together, the analysis suggest that our genotypes are generally robust and reliable, despite the presence of a limited number of variants that warrant further investigation to ensure the reliability of the genotypes or determine if it is necessary to adjust some GeONTIpe parameters. Such follow-up should include verifying that the BPs are accurately defined, given their critical importance to GeONTIpe, as well as searching for additional structural variation that may be behind genotyping failures. Future efforts to better characterise these possible errors will likely focus

on variants with values of LD very close to 1, as these suggest that the majority of genotypes are consistent with nearby SNPs, and that only a small number of samples may be driving the observed discrepancies.

4.3.4 Inversion frequency

Due to the large number of genotypes obtained, we were able to investigate the population frequency and distribution of the 680 polymorphic inversions analysed in this study. One of the most relevant properties examined was AF, which was calculated both globally and separately for each ancestry represented in our dataset (AMR, AFR, EAS, EUR and SAS). Because information on which inversion orientation corresponds to the ancestral state is not available in most cases, we opted to calculate MAF instead of the derived allele frequency (DAF).

In general, we expect the estimated frequencies to be robust. Nevertheless, for variants with very low genotyping performance, for example, cases with success rates as low as 1%, AF reliability of the estimates may be affected. Another potential source of bias arises from the presence of additional SVs, particularly when these were frequent and large. Given the underlying principles of the GeONTIpe strategy, increases or decreases in the distance between probes can enhance or impair genotyping accuracy, especially if the additional variant is consistently associated with one of the inversion orientations.

In contrast to previous studies by our group (Campoy, 2024), we did not formally distinguish between inversions generated by the non homologous (NH) or NAHR mechanisms. Although breakpoint information allows us to infer the likely mechanistic origin of many variants, the lack of a systematic classification of the origin mechanism imposes limitations on some downstream analyses. In particular, recurrence analyses of NAHR inversions, have not yet been conducted here. Although such analyses could be performed once the genotypes have been further refined, they would likely face the constraints inherent to the sequencing methodology. These approaches typically rely on the association of the inversion orientation with the surrounding variation, identifying SNPs profiles characteristic of one orientation in

individuals carrying the opposite (Giner-Delgado et al., 2019; Porubsky et al., 2022). However, the use of ONT data for a reliable call of SNPs can be challenging, suggesting that recurrence analyses may need to be restricted to samples for which high-quality SNPs calls are already available.

Returning to the MAF patterns, we observed a wide range of inversion frequencies, from extremely rare variants to those approaching 50%. Our validation and genotyping efforts focused primarily on variants with a frequency of at least 1%. Despite this threshold, a substantial proportion of inversions (487 of 680) fell below this value. This is probably due to the relatively small sample size of most of the included studies, which results in the selection of rare variants that are present in those variants. In contrast, due to the large sample size analysed here, we were able to detect variants with global frequencies as low as 0.04%. This information will be instrumental in guiding future decisions regarding which variants should be prioritised for functional impact analyses.

Using these more comprehensive data, we revisit the observed relationship between inversion size and AF. First, we did not find a consistent negative correlation between inversion size and MAF, which could be due in part to the higher proportion of rare variants analysed. However, in agreement with previous results, we detected a positive correlation between the breakpoint-to-inversion size ratio and MAF, with higher ratios associated with increased frequencies (Giner-Delgado et al., 2019; Puig et al., 2020; Campoy, 2024). This pattern is consistent with the observation that inversions mediated by IRs, particularly when these repeats are closely spaced, tend to exhibit higher recurrence rates, ultimately driving MAF towards intermediate values.

4.3.5 F_{ST}

The complementary analysis of F_{ST} had the objective of identifying which of the variants analysed exhibited significant differentiation between ancestries. Variants showing elevated F_{ST} values can be considered candidates for being affected by natural selection, potentially favouring specific alleles in particular populations (Holsinger &

Weir, 2009). Rather than limiting the analysis to the raw values of F_{ST} , we compared them against a null distribution derived from SNPs, allowing us to assess whether the variants of interest were over-represented among the most extreme values.

Using all polymorphic variants identified in our dataset, we detected 23 inversions displaying significant global F_{ST} values. However, when pairwise comparisons between ancestries were performed, 64 inversions were found to fall within the top 5% of the F_{ST} distribution relative to SNPs. Several of the reported variants had already been associated with population differentiation in previous studies. Notable examples include HsInv0389, for which the Inv allele reaches a high frequency in African populations and is the major allele, in contrast to the remaining ancestries (Giner-Delgado et al., 2019); HsInv0114, where Std orientation has been reported as the major allele worldwide, while Africans only present a 18.9% frequency for this allele; HsInv0124, characterised by a lower proportion of inverted alleles in European populations; HsInv0266, with South Asian populations showing the highest frequency of inverted alleles; and HsInv0340, for which African populations have the highest frequency of inverted allele (Campoy, 2024).

Despite the confirmation of previously reported variants, our analysis also identified a novel set of inversions exhibiting significantly high F_{ST} values. Among these, some correspond to variants for which earlier studies in our group did not observe population differentiation, including HsInv0030, HsInv0055, HsInv0209, HsInv0228, HsInv0241, HsInv0344, HsInv0347, HsInv0348, HsInv0370, HsInv0382, HsInv0397, HsInv0403, HsInv0414, HsInv0822, HsInv0830, HsInv1111, HsInv1117, and HsInv1124. These discrepancies may be partly explained by genotyping some of these variants in only 95 samples as presented in Campoy (2024) and the inclusion of additional ancestries. In addition, in the case of HsInv0348, the different results could be due to the high error rate of the PCR data, which apparently overestimated the inversion frequency in African populations.

Among the significant hits identified in this thesis, several variants had already been proposed as candidates for functional effects in previous studies, including HsInv0030, HsInv0055, HsInv0347, HsInv0348, and HsInv1117 (Giner-Delgado et

al., 2019; Campoy, 2024). Additionally, although direct functional effects for some other variants were not detected, they overlapped with genes or regulatory elements that could potentially have functional consequences (Campoy, 2024). Overall, these results suggest that a larger number of inversions may be influenced by selection than previously detected. Consequently, this analysis should be interpreted as a first step in identifying candidates for future studies aimed at detecting additional signals of selection.

4.4 Detailed analysis of candidate regions

Finally, as the last part of the thesis, we will summarise a subset of variants for which we were able to perform a more in-depth structural analysis and integrate all the available information. This set includes variants that were characterised manually and those analysed using GeONTIpe and CHARONTE. These variants represent strong candidates for further analysis in more individuals, particularly in the context of their potential functional effects as complex genomic regions.

4.4.1 HsInv0012-0659

This region of chromosome 1 (chr1:248,359,787–248,688,406) is characterised by the presence of two pairs of inverted SDs of 38 kb and 59 kb, respectively, and has been reported to contain inversion candidates in several of the studies used to populate our catalogue (Chaisson et al., 2019; Levy-Sakin et al., 2019; Ebert et al., 2021; Byrska-Bishop et al., 2022), particularly a 10 kb inversion mediated by the 59 kb SDs (HsInv0012). However, the structural organisation of the region suggested the presence of a second larger inversion of 172 kb that encompassed the smaller one (HsInv0659). The analyses performed with GeONTIpe suggested that the frequency of the inverted orientation of the smaller inversion was 70.58%, while that of the larger inversion was 1.33% (Moreira-Pinhal et al., 2025). Nevertheless, it is important to note that this second inversion could be genotyped in only about 30% of individuals, mainly because probes A and B were placed more than 100 kb apart.

As part of this analysis, in addition to the inversions, several SVs were observed,

including multiple variants located within the internal region between the SDs, indicating that the reference genome (GRChr38) represents a minor haplotype. Moreover, there was a 40 kb deletion that removed the space between the second inverted SDs of both inversions. Finally, an inverted duplication was also found affecting the same segment, which relocated that sequence upstream of the first copy of the 59 kb segmental duplication in sample HG02723. For this region, 6 different haplotypes were found. in the set of 54 samples.

Although several of the previous studies reported variants in this region, only Ebert et al. (2021) appears to indicate the presence of the larger inversion (HsInv0659) based on the provided coordinates, which emphasises the importance of our detailed analysis. Regarding the distribution in the population, although 12 of the 14 individuals who exhibit inverted orientation of HsInv0659 belong to African populations, no significant population-specific differences were observed for either variant. Moreover, LD analysis revealed that, for the smaller inversion (HsInv0012), several SNPs in East and South Asian populations (EAS and SAS) exhibited LD values higher than 0.85, with top values of 0.927 and 0.892, respectively. For the larger inversion, LD values of 1 were observed in East Asian and American populations. However, this appears to result from a single individual carrying the variant in both ancestries.

From a perspective of gene content, this region contains multiple members of the olfactory receptor (OR) gene family, which is known to display a large variation in copy number throughout the genome, including a cluster in 1q44 (Young et al., 2008). Among members of the OR family, there was already evidence of *OR2T10* deletion, consistent with the 40 kb deletion described above. According to Young et al. (2008), this deletion was estimated to occur with a frequency of 9%, while GeONTIpe suggests a lower frequency of approximately 1.8%. Despite these discrepancies, this region does not appear to have a high haplotype diversity and could be a suitable candidate for further analysis using CHARONTE. In addition, the possible effects on the OR genes make it also a good candidate for functional analysis.

4.4.2 HsInv0231

This genomic region has already been highlighted in Section 3.4.6, which describes the implementation of CHARONTE to resolve the different structural haplotypes present. However, in the current section we will focus on the existing knowledge about this region, previously reported findings, and their consistency with the results obtained in this study. As mentioned above, the HsInv0231 region was previously detected in the studies of Ebert et al. (2021) and Byrska-Bishop et al. (2022), and was analysed using both software tools developed in this thesis, revealing some discrepancies that have already been addressed. Genotyping revealed the presence of a low frequency inversion (1.92%), as well as a 70 kb deletion with a frequency of 16.81%. This deletion appears to be caused by the presence of SDs in direct orientation, whereas the inversion is likely generated by SDs arranged in inverted orientation.

Focussing only on the inversion, no significant differences in frequency were observed between populations. In addition, SNPs with LD equal to 1 was detected in the American and European populations, although none of the identified variants was shared between these populations. The absence of globally fixed variants is not unexpected, given not only the presence of other SVs in the region, but also the detection of a recurrence event of the inversion detected by CHARONTE in Asian populations, which could explain the observed lack of LD on a global scale.

Interestingly, this region has been extensively investigated due to the presence of genes involved in the Rh blood group antigen system. In fact, several studies have reported the existence of SVs in this region, including polymorphisms affecting *RHD* and *RHCE* genes, such as deletions of *RHD* (Okuda & Kajii, 2002; Srivastava et al., 2018). However, the deletions described in those studies do not appear to correspond to the 70 kb deletion observed here. In addition, multiple *RHD/RHCE* hybrid genes have also been reported (Wheeler et al., 2019).

This same region was analysed in the first release of the Human Pangenome, where a total of ten different haplotypes were identified. Two of these haplotypes present

in Liao et al. (2023) had complete deletion of *RHD*, named RHCE and RHCE-RHD(2)-RHCE, respectively, with a combined frequency of 18%, a value consistent with that observed in our study. Furthermore, an inverted haplotype (RHD-RHCE(10);inv;RHCE-RHD(10)) was detected at a frequency of 1%. Nevertheless, several hybrid configurations involving both genes have also been described that cannot currently be detected by either GeONTipe or CHARONTE. This limitation is partly due to the fact that these configurations arise within regions characterised as SDs, where the density of k-mers is reduced and limits our analysis capabilities.

However, due to the relatively small number of samples interrogated in that study, other variants such as the 115 kb inversion or the inverted duplication detected by CHARONTE were not identified. Therefore, despite the differences observed between the studies, the software developed in this work provides a robust overview of the structural configurations reported in the Human Pangenome, suggesting that CHARONTE can be highly useful for performing an initial exploratory analysis of structural diversity within a genomic region.

4.4.3 HsInv0348

The HsInv0348 region on chromosome 14 (chr14:105536882–105859988) has already been briefly discussed in previous sections in relation to the high error rate observed when comparing the data generated by ONT with experimental genotypes. However, given the particular interest of this locus, it needs a more detailed discussion. This region was originally described by our group in Campoy (2024), and has been reported to harbour several SVs in multiple studies that were used to build our inversion catalogue (Audano et al., 2019; Chaisson et al., 2019; Ebert et al., 2021; Byrska-Bishop et al., 2022).

First, when this region was examined in more detail for analysis using the ONT reads, a potential source of error was identified. A highly similar sequence with approximately 95% of identity is located around 115 kb upstream of the HsInv0348 inversion. However, this sequence appears to be consistently orientated in the opposite direction to that of the reference genome for the HsInv0348 region. To confirm

this observation, a specific set of probes (called HsInv1891) was designed and used to genotype the 54 samples included in the study by Moreira-Pinhal et al. (2025), which showed no evidence of an orientation change at this upstream locus. However, in the dataset of 1,409 samples, HsInv1891 was also identified as polymorphic, albeit at a much lower frequency (1.19

When comparing the HsInv0348 genotypes inferred from the ONT data with the corresponding experimental results, several discrepancies were detected. As discussed previously, we tested multiple sets of alternative probes, particularly extending them beyond the set of SDs included within the region. However, none of these alternative designs resulted in a significant improvement compared to the experimental genotypes. In addition, inconsistencies were also observed among genotypes generated from ONT data by the different probe sets, suggesting the presence of other underlying confounding factors.

To further investigate this issue, we compared LD between variants detected by each probe set, under the assumption that a variant that shows a perfect or near-perfect linkage (LD close to 1) should imply more reliable genotypes. For three of the probe sets tested, we identified SNPs with a global LD greater than 0.95, seven of which were shared between the three designs. This suggests that despite the observed discrepancies, the ONT genotypes are largely consistent. Moreover, it indicates that the discrepancies with the experimental genotypes likely result mainly from the complexity of the iPCR assay due to the presence of the upstream duplicated region. Given the absence of significant differences in performance among the probe sets, the original design was retained, as it exhibited the best overall performance.

Currently, our working hypothesis is that the small discrepancies between probe sets are caused by reads originating from the upstream region that would be erroneously incorporated into the analysis of HsInv0348, thus confounding the genotyping process. An alternative explanation is the presence of additional haplotypes that carry other rearrangements. Specifically, the existence of two direct repeats could facilitate the deletion of the HsInv0348 inversion region, which, due to the high sequence similarity to the upstream segment, might result in reads derived from such haplo-

types being incorrectly interpreted as supporting HsInv0348. Although this deletion, together with other SVs affecting the region, has been reported in previous studies, this hypothesis cannot be conclusively validated at present (Jana et al., 2025).

Using these reliable majoritarian genotypes, by performing a F_{ST} analysis, we were able to detect significant differences in the frequency of inversion of HsInv0348 between populations. The African ancestry showed a markedly lower frequency of the inverted allele (approximately 3%), while in the remaining populations the frequency was between 30% and 57%. Importantly, no substantial differences in GeONTipe performance were observed between populations, suggesting that these frequency differences may reflect genuine biological variation rather than technical artefacts.

In previous analyses, when the potential functional effects of inversion were examined using experimental genotypes, it was found to be associated with chronic spontaneous urticaria, as well as with expression changes in the lncRNA gene *RP11-731F5.2* and the pseudogenes *IGHGP* and *ELK2BP* (Campoy, 2024). Furthermore, it was shown that the region overlapped an area enriched with H3K27ac and H3K4me1 marks and was located near two enhancers that interact with the protein-coding gene *TMEM121* (Campoy, 2024). From a perspective of gene content, the region contains several genes from the immunoglobulin heavy chain constant region (IGH) family, which have very important functions in adaptive immunity. Among them, *IGHA1* is the gene closest to inversion, although it does not overlap with it directly. In addition to these genes, the region contains a large number of SDs in direct and inverted orientations with identities that can reach up to 96%.

Four *IGHG* genes (*IGHG4*, *IGHG2*, *IGHG1* and *IGHG3*), which are close to our inversion, have previously been studied, harbouring multiple variants that generate different allotypes in populations (Bashirova et al., 2021). This further supports the complexity of haplotypes in this region, at least at the SNPs level. Recently, in addition to the inversion itself, several deletions and insertions have been identified that affect multiple members of the *IGHG* gene family (Jana et al., 2025). In particular, a 120 kb insertion duplicates several genes in the region, potentially giving rise to a third copy of the high-identity duplicated region associated with HsInv0348.

In addition, a fusion between *IGHA2* and *IGHA1* was also identified accompanied by the loss of *IGHE*, *IGHG4*, and *IGHG2*, probably causing the deletion of the HsInv0348 sequence (Jana et al., 2025). Together, these SVs further support our hypothesis that alternative haplotypes can interfere with the genotyping process in this region, although the number of individuals affected by these additional SVs is relatively small. Thus, these findings highlight this region as a compelling candidate for a more comprehensive structural and functional characterisation.

4.4.4 HsInv0390

This genomic region was also used for the optimisation of CHARONTE, and its structural configuration has already been described in Section 3.4.5. The region was previously characterised in the study of Beck et al. (2015), where the authors proposed the existence of a mechanism capable of exchanging variants between the SDs present in the region, as well as the occurrence of possible rearrangements caused by polymerase stalling during repeat replication. These rearrangements were reported to affect *PLP1* gene, which is located approximately 140 kb upstream of the region analysed here (Lerga Jaso, 2020; Puig et al., 2020). The inversion was previously genotyped by the group in 95 individuals using ddPCR (Puig et al., 2020) and was therefore incorporated into our inversion catalogue as a previously known variant. However, its existence is further supported by several additional independent studies (Chaisson et al., 2019; Ebert et al., 2021).

In our analysis, the inverted allele was detected with a AF of 53.16%, a value slightly higher than that reported in the previous study (Puig et al., 2020) and in the 54 samples analysed in Moreira-Pinhal et al.(2025), which was close to 48%. However, there were marked differences in genotyping performance between datasets: while Moreira-Pinhal et al. (2025) achieved a success rate of 98%, genotyping of 1,409 samples showed a success rate of only 30%. This reduction is most probably due to the inclusion of samples with read lengths insufficient to span the 70 kb BPs of the inversion. However, we expected that the frequency value obtained here is closest to reality due to the much higher number of genotyped samples. Analysis of

F_{ST} revealed that there was no population differentiation and no evidence of LD was detected in any ancestry, which is consistent with previous observations (Lerga Jaso, 2020).

In addition to the inversion itself, several other SVs were identified, most notably a deletion of approximately 10 kb located at the second breakpoint. This variant was not previously detected, as its size is unlikely to produce a measurable effect on ddPCR results when genotyping HsInv0390 with this technique (Puig et al., 2020). Furthermore, no evidence was found to suggest functional consequences of this deletion, either through effects on *TMSB15C* or *SLC25A53*, which are the genes closest to the variant. The analyses also did not identify evidence that indicates a recurrence of this deletion. Moreover, the shift of this variant from breakpoint 2 to breakpoint 1 was also not observed in any of the samples, suggesting that the true BPs of this inversion may be located more internally within the SDs, at least for the samples carrying the deletion.

4.4.5 HsInv0394-816-822

The next genomic region that could be considered a suitable candidate for analysis using CHARONTE is composed of the variants HsInv0394, HsInv0816, and HsInv0822. This region is of particular interest due to the presence of the melanoma antigen gene family cluster (MAGE) located in Xq28. As described in the corresponding results section (Section 3.4.1), this region is characterised by the presence of three inversions, the most frequent of which is the one located between the innermost SDs. This inversion was previously genotyped using ddPCR (Puig et al., 2020) and later confirmed that the minor allele corresponded to the orientation present in GRCh38 (Porubsky et al., 2023). Besides the previous knowledge of HsInv0822, the other two inversions were also detected in multiple independent studies that were used to create our inversion catalogue.

This region was initially analysed using GeONTipe, allowing us to estimate the frequency of each inversion independently. When assessing F_{ST} among ancestries, no significant differences were detected globally. However, pairwise comparisons be-

tween ancestries revealed that HsInv0822 showed significant differentiation between American and European populations compared to Asian populations (both EAS and SAS), which was not previously reported in other studies (Puig et al., 2020). Inspection of the AF across populations revealed that this differentiation was driven by a higher frequency of the inverted allele in European, American and African populations (68-74%) relative to Asian populations (38-46%). However, no differences are observed in either of the other two inversions. None of the variants showed evidence of linkage in any of the ancestries analysed either, indicating that they are probably highly recurrent, as already suggested for HsInv0822 (Puig et al., 2020; Porubsky et al., 2022).

As discussed previously, given the structural complexity of the region, we performed an additional analysis to identify the possible structural configurations that could arise from the combination of the three inversions. Of the eight theoretically possible configurations, six were observed in our dataset. This combination of inversions was previously analysed in Pangenome data by Porubsky et al. (2023) and we evaluated whether the samples reported as carrying each structural haplotype were consistent with our observations. In that study, five haplotypes were described, four corresponding to different inversion configurations and an additional inverted duplication located between the SDs termed R5 and R6. In our framework, this inverted duplication would affect the second breakpoint of HsInv0394 and the second breakpoint of HsInv0816.

The authors reported that the most frequent configuration was H1, corresponding to the entire region being in direct orientation, which also matches the orientation of the T2T-CHM13 reference genome. This is fully consistent with our results, where we estimated an approximate frequency of this configuration is 59%, closely matching the 69.3% suggested by their haplotype counts. For the remaining haplotypes, we examined the specific samples reported in both studies, bearing in mind that in our analysis the inversions were treated as independent variants, requiring a conversion of the reported orientations. After accounting for this, configuration of all common samples was found to be compatible. However, the previous study proposed the

existence of three independent inversion events mediated by distinct SDs, rather than a model based on nested inversions, which seems more parsimonious.

Overall, despite the necessarily limited nature of our analysis, primarily due to the inability to phase haplotypes, we have shown that the general structural landscape observed is highly consistent with the findings reported in a previous study, supporting the reliability of the haplotypes identified in our analysis. Importantly, the inclusion of a larger number of samples allowed us to identify two additional haplotypes beyond those originally reported. This observation underscores the importance of analysing structurally complex regions in larger datasets, as increased sample sizes can substantially refine haplotype resolution and strengthen the conclusions that can be drawn from such regions.

4.4.6 HsInv0401

The HsInv0401 was the first region used to optimise CHARONTE and its main characteristics were previously described in the corresponding Section 3.4.4. This variant of chromosome X was identified in several studies used to build the inversion catalogue (Audano et al., 2019; Chaisson et al., 2019; Ebert et al., 2021). In the initial genotyping analysis using 54 samples, the inverted allele showed a frequency of 73%, which was highly consistent with the frequency observed in the larger dataset of 1,409 samples (72%). When stratifying by ancestry, the inverted allele was less frequent in African populations, with a frequency of 43%, with significant differences observed in the F_{ST} analysis.

A distinctive feature of this variant was the presence of insertions and deletions of approximately 12 kb, or multiples thereof. This pattern suggested the existence of a copy number variant involving a pre-existing sequence in the reference genome (GRCh38). Sequence inspection revealed that this element is located within the SDs that mediate the observed inversion. Due to the fact that the haplotype diversity had already been manually characterised as part of Moreira-Pinhal et al. (2025), this region was selected for the development and validation of CHARONTE.

Using this tool, we were able to identify a total of 18 distinct haplotypes, resulting from the combination of the inversion and the change in the number of copies of the 12 kb sequence at each breakpoint. This extensive haplotype diversity likely explains the absence of detectable LD with nearby variants. Furthermore, we observed that this variable sequence encompasses the *XAGE1* gene, making this region a compelling example of the potential effects of changes in gene dosage associated with SVs.

XAGE1 belongs to the GAGE gene family and has been implicated in the avoidance of apoptosis. It is known to be expressed in germ cells and is frequently observed in malignant cancers, including breast cancer. Four main transcripts of this gene have been described, with *XAGE-1b* being the predominant isoform (Jumaa & Ramadhan, 2021). Despite previous studies, to date, variation in the copy number of this gene has not been reported. Given its potential effect in cancer, this makes the region a candidate of high-interest for further investigation, particularly to assess the functional relevance of the different identified haplotypes.

4.4.7 HsInv0786

Another region that was manually analysed to further unravel the underlying structural complexity was HsInv0786. This region is involved in different GWAS hits and the inversion was previously associated with an asthma-obesity phenotype and expression changes in several genes (González et al., 2014a; Puig et al., 2020; Campoy, 2024). As mentioned in the previously presented Section 3.4.2, this region contains some SDs, with one inverted pair of approximately 134 kb that is involved in the generation of the inversion. However, it is important to note that these SDs belong to one of five clusters of SDs. These five clusters are called BP1 to BP5 and together span approximately 4.5 Mb within the chromosomal band 16p11.2.

The 16p11.2 locus is characterised by the presence of extensive structural rearrangements, specifically between two of these clusters SDs (BP4 and BP5). These SDs clusters have undergone rapid expansion in humans, facilitating the emergence of CNVs in this region. In particular, the BP4–BP5 interval also harbours the recurrent

16p11.2 microdeletion and microduplication syndromes, which have been associated with developmental delay and other mental disorders (Auwerx et al., 2024; Taylor et al., 1993).

The HsInv0786 region analysed here is found between the BP1 and BP2 clusters, spanning approximately 500 kb. This interval contains several genes, among which the *NPIP6-9* genes are particularly notable, since paralogous copies are also present in other groups of SDs within the same 4.5 Mb region. However, the function of these genes remains unknown and further characterisation is needed (Dishuck et al., 2025). Additional candidate genes in this region include *EIF3CL* and *SULT1A*, the latter of which shows reduced expression associated with the inversion (González et al., 2014a).

The region was initially analysed using GeONTIpe, which detected the known inversion with a frequency for the minority inverted allele of 33.48%, which is similar to the 25% reported by Puig et al. (2020). No significant differences in frequency were detected among the ancestries analysed. However, when LD was evaluated with nearby variants, a single SNP showing a strong link (0.938) was identified in European populations. Relaxing the LD threshold to 0.8 revealed that four ancestries harboured at least one variant exceeding this value, with eight variants shared by three of the ancestries. In this context, both potential genotyping biases, arising from haplotypes that may be more easily genotyped due to the presence of deletions, or the recurrence of the inversion, as previously suggested (Puig et al., 2020), could be responsible for the lack of associated SNPs. However, without a more comprehensive analysis, it is not possible to clearly discriminate between these two scenarios.

Beyond the inversion itself, GeONTIpe reported several additional SVs, including deletions of 114 kb and 11 kb, as well as a 114 kb insertion. All of these variants were subsequently validated by manual inspection. This manual analysis also revealed a previously unreported 12 kb insertion that was not captured by the HsInv0786 probes. The two deletions and the 114 kb insertion are located within the 134 kb SDs, with the 114 kb deletion being mediated by direct repeats of 20 kb nested

within both copies of the 134 kb duplication. Interestingly, the 114 kb deletion and the insertion are expected to directly affect copies of the *EIF3CL* gene, whereas the 12 kb insertion impacts the *SULT1A* gene. Despite the limited number of samples analysed, we were able to identify 19 distinct structural configurations.

When comparing these structures with the 13 haplotypes reported by Dishuck et al. (2025), the relative proportions of inversion orientations were generally concordant between both studies. However, in that work, the T2T-CHM13 genome is considered the reference, and it is assigned the "direct" orientation for the HsInv0786 region, which for us corresponds to the inverted allele. Consequently, the D7 haplotype reported in their study (T2T-CHM13) would correspond to our H11 haplotype (Table 3.12). Similarly, our haplotypes with the 114 kb deletion (H2, H9, H13, and H16) could correspond to haplotypes I14, I11, and D10, which suggests that the standard allele (according to GRCh38) carrying the deletion at BP1 may not have been detected. Additionally, the 114 kb insertion at BP1 does not appear to be reported in that study either. Nevertheless, an increase in *SULT1A* copy number was described, which is consistent with our observations. Further comparisons are currently limited due to the lack of a more exhaustive sequence-level analysis of the haplotypes reported in that work. However, the sample reported to contain haplotype I9 was classified in our analysis as harbouring the standard orientation with an 11 kb deletion in BP2, which is compatible with their findings.

The Dishuck et al. (2025) study also reported three additional inversions that affect several *NPIP* genes. Inspection of our inversion catalogue revealed multiple candidates whose genomic coordinates could overlap with these regions. For example, in the 16p11.3 region, HsCan4599, HsInv0363, HsInv0551, HsInv1936, and Mos-16-14815630-15240398 are compatible with the reported coordinates. Among these, HsInv0363, an inversion mediated by approximately 72 kb SDs, emerges as the most plausible candidate, as the minor allele (Std in our framework, "direct" in theirs) shows a frequency of approximately 28% in both cases. In the sample used as an example for haplotype I3, GeONTIpe detected a pair of reads supporting the Inv orientation, which is consistent with their observations. Other candidate inversions

in this region were also found to harbour inverted alleles, although at much lower frequencies, potentially reflecting the presence of additional rearrangements among the multiple SDs present in this locus.

In the second reported region, 16p13.11–12.3, two variants (HsCan4609 and HsInv0365) appear to be plausible candidates, with HsInv0365 being the most likely. HsInv0365 corresponds to an inversion mediated by SDs of approximately 148 kb. For this inversion, the inverted allele showed a frequency of approximately 5%, consistent with that observed in Dishuck et al. (2025) study. Given the large size of the associated BPs, the genotyping performance in this region was extremely low. Thus, a direct comparison of the genotype of the only sample with the inversion in common was not possible, as only two supporting reads were reported and the genotype was classified as low confidence.

Finally, the region reported as 16p12.2 does not appear to clearly correspond to any of the inversions analysed in our study. The closest candidate is HsInv0152, for which we inferred BPs of approximately 102 kb in GRCh38. However, this is substantially smaller than the 262 kb SDs visible in the T2T-CHM13 assembly. For this variant, we detected the inverted allele in approximately 1.8% of the chromosomes, which could be compatible with the reported frequency. Nevertheless, when examining the sample analysed in the previous study, the inversion status contradicts our results. Given that our genotype is supported by 58 reads with the standard orientation, it is unlikely that one of the two alleles would have been missed in our analysis. Examination of probe location using Pangenome assemblies for this sample (HG03516) showed that probes A, B, and C exhibited configurations consistent with those reported by GeONTIpe, while probe D showed a double hit, making it unreliable. This suggests that the observed discrepancy is most likely due to the fact that the two variants are not equivalent.

In summary, this work provides a preliminary overview of this complex genomic region, which, although not analysed as comprehensively as in Dishuck et al. (2025), shows substantial concordance between the GeONTIpe-based observations and previously reported results. In addition, GeONTIpe enabled the detection of smaller-

scale variants that can be overlooked when focussing on large regions or when variants do not directly affect *NPIP* genes, which are the main focus of the referenced study. The manual analysis of HsInv0786 allowed the disentanglement of several additional haplotypes, some of which show clear internal consistency. Future work will aim to incorporate this region into CHARONTE, enabling a more detailed structural analysis and substantially increasing the number of individuals that can be analysed.

4.4.8 HsInv1117

The final region examined corresponds to the HsInv1117 inversion, located on chromosome 10 (chr10:4,925,933–5,031,863). This variant was originally identified through a comparative analysis between the reference genome available at the time (HG18) and the HuRef genome derived from J. Craig Venter (Venter et al., 2001), and then validated and genotyped in our group through sequence analysis (Vicente-Salvador et al., 2017). This region is flanked by SDs of approximately 57 kb. However, manual inspection of these sequences identified the inversion BPs within a much smaller region of just 520 bp, which allowed genotyping of multiple individuals using iPCR (Campoy, 2024).

Regarding the possible functional consequences of the HsInv1117 inversion, these SDs overlap with the *AKR1C1* and *AKR1C2* genes, as well as several pseudogenes. In addition, the experimental genotyping of this inversion revealed several potential functional effects (Lerga Jaso, 2020; Campoy, 2024). Specifically, 15 phenotypic associations with variants in the GWAS Catalog were detected, together with a new fusion transcript identified in inverted orientation. Furthermore, several changes in gene expression that affected more than ten tissues were associated with this inversion (Lerga Jaso, 2020; Campoy, 2024).

In this case, for ONT genotyping, manual analysis of the SDs mediating the inversion allowed the placement of the external probes (A and D) within the respective repeat elements. Nevertheless, due to uncertainty as to whether the most external portions of these SDs might also contribute, albeit to a lesser extent, to inversion

formation, a second probe set was designed that encompassed the entire sequence of the SDs (HsInv1117L). To assess the relative performance of both probe sets, we compared their genotyping results with previously obtained experimental data and found no discrepancies for either configuration. We then directly compared the genotypes between both probe sets and observed 753 differences, although they all corresponded to a transition from valid genotypes to non-determined (ND) or low-confidence genotypes. The only exception involved a change from a standard homozygote (HsInv1117) to an heterozygote (HsInv1117L).

Both probe sets exhibited high linkage with nearby SNPs, with the internal set showing an r^2 of 0.99 compared to 0.97 for the external set, despite both being linked to the same nearby variant. The slight deviation from perfect LD suggests that only a very small number of samples differed in genotype, with the internal probes performing slightly better, probably due in part to the larger number of supporting reads. Inspection of raw data revealed that for both probe sets, the sample showing a discrepancy genotype between inversion and SNP was HG03856, which belongs to the South Asian population (SAS). Its homozygous standard genotype reported had support from only three reads, but the GeONTIpe methodology identified three SNPs showing differences between the reads, which supported the presence of both alleles. However, when we look at the 1000GP data, the linked variant exhibited a heterozygous genotype.

Based on these observations, the origin of the discrepancy remains unclear, whether it arises from a genotyping error derived from ONT or from a modification of the linked variant due to a localised mutational event, gene conversion or a single, highly specific recurrence. This linked SNP was located within the second segmental duplication, suggesting that a recurrent rearrangement event may explain the observed loss of linkage. A similar inconsistency in LD had previously been observed in iPCR data, where the sample NA19474 from the African (AFR) population showed a homozygous standard genotype for the inversion, while the same linked variant was heterozygous. To rule out an experimental artefact, the iPCR was recently repeated, yielding the same genotype. Unfortunately, no ONT data were available

for this sample, leaving the underlying cause of these discrepancies unresolved.

Given these results, we opted to retain the internal probe genotypes, as they exhibits stronger LD with nearby variants and overall better performance. However, the possibility of a slightly different inversion mediated by more external regions of the SDs cannot be excluded, which would be a recurrence event. For this reason, this region emerges as a strong candidate for further analysis using CHARONTE, as the relatively low overall sequence identity of these repetitive elements suggests that the strategic incorporation of multiple k-mers could be critical to elucidating the underlying structural dynamics of this locus.

In this work, we propose several solutions to try to properly characterise inversions and other complex variants. Nevertheless, a certain number of regions still had to be excluded from our analyses, and further technical and methodological improvements will be necessary to enable a reliable analysis of a broader set of regions.

4.5 Future perspectives

To conclude the discussion of the results obtained in this thesis, it is important to emphasise that this work has been made possible by the development of ONT sequencing, with the increasing availability of reads and their growing length, and by the continuous expansion of the number of samples generated using this technology. These advances have enabled not only the discovery and genotyping of a larger number of inversions, many of which could not be studied before, but also a more reliable characterisation of complex regions. Such improvements are reflected across the studies used to curate and expand the inversion catalogue, which now encompasses a substantially larger number of events and allows their genotyping with progressively higher confidence.

Nevertheless, important challenges remain and must be addressed in order to fully exploit the methodological advances that are currently emerging. Although further refinements are still required for the characterisation of some of the regions analysed in this thesis, we are closer than ever to achieving a comprehensive view of a class of

SVs that has been underexplored for a long time. As the end of one path becomes visible, new crossroads emerge, particularly those related to a level of structural diversity far greater than initially anticipated.

One of the future perspectives arising from the work presented in this thesis is the further refinement of the genotypes reported here through the multiple complementary analyses performed, with the aim of generating a highly reliable genotype set that will serve as a foundation for subsequent studies. The goal is to update InvFEST by incorporating all information obtained from the merged studies together with the newly derived genotypes, making these data openly available for the research community. This effort will transform InvFEST into the most comprehensive inversion dataset to date, which can subsequently be complemented with functional annotations as new evidence emerges.

As discussed above, accurate characterisation of SVs constitutes a fundamental prerequisite for downstream genomic and functional studies, and novel methodological approaches are required to fully address this challenge. Many of the inversions identified in this study, as well as previously described ones, can be used for the imputation of inversions in large-scale cohorts, such as the UK Biobank or disease-specific datasets for complex traits. This would enable a systematic investigation of the contribution of inversions to phenotypic variation. In particular, it would facilitate genome-wide association study (PheWAS) analyses for inversions that have no linkage with SNPs, a scenario that is especially relevant for recurrent inversions. For many inversions, imputation is currently based on a quite limited number of genotypes, and the great genotype increase generated in this work could improve considerably the imputation of recurrent inversions. However, for inversions that cannot be reliably imputed, the methodology developed in this work provides a framework for direct genotyping, provided that long-read sequencing data are available. Given the continuous reduction in sequencing costs, the availability of such data is expected to increase substantially in the coming years.

Regarding complex regions, future perspectives are conceptually similar, but involve an additional layer of complexity. Dissecting the functional impact of these

regions will likely require strategies to associate specific structural haplotypes with phenotypic outcomes, including the development of haplotype-based imputation approaches. This will be necessary to disentangle the effects of unique combinations of SVs within these regions. Complex regions are also of particular interest from an evolutionary point of view, especially in cases where haplotype diversity differs markedly from that observed in other species (Guitart et al., 2024; Paparella et al., 2023). Among these candidates, some may provide insight into genetic features that contribute to human-specific traits.

From a clinical perspective, this work opens the possibility of investigating unresolved cases in which exome sequencing has failed to identify a causal variant. Since exome-based approaches remain standard practice in clinical settings, SVs are often underdiagnosed. Incorporating the analysis of structural variation, particularly inversions and complex haplotypes, represents a novel avenue for the diagnosis of rare genetic diseases and will help clarify the contribution of SVs to monogenic disorders.

Finally, although our current expertise in this domain is not yet as mature as that of simpler variants, this work opens the door to the discovery of combinations of SVs whose functional or evolutionary effects have remained largely unnoticed due to longstanding methodological limitations in their accurate detection and characterisation. We believe that we are still only at the threshold of what next-generation sequencing technologies can offer and that many additional associations will continue to emerge as new haplotypes are identified and systematically characterised in the future.

5. Conclusions

The main conclusions derived from this work are as follows:

1. Long-read sequencing data from ONT have proven to be highly effective in characterising and genotyping genomic variants that were previously inaccessible to standard approaches. This capability is particularly relevant for the detection and analysis of inversions generated by NAHR mechanisms.
2. We have generated the most comprehensive catalogue of polymorphic human inversions to date, integrating 13 independent studies and 21,530 variants. Of 15,605 non-redundant candidates, 1,235 were finally analysed in a large set of samples.
3. GeONTIpe was specifically designed for accurate inversion genotyping and has demonstrated strong performance compared to previously obtained experimental data. Nevertheless, several candidate inversions were identified that require additional revision to further validate and refine the inferred genotypes.
4. A total of 680 polymorphic inversions were identified, exhibiting a wide spectrum of frequencies. Among these, 64 variants showed significant population-specific frequency differences, some of which corroborate previously reported findings, while also introducing new candidates for the investigation of potential selection signatures.
5. The dataset comprising 1,409 samples is the largest compilation of samples sequenced with ONT and our inversion genotypes for 680 polymorphic inversions in these samples represents by far the largest dataset on human inversions

to date. This data constitutes an extremely valuable resource for uncovering novel functional implications of inversions in human disease and phenotypic traits.

6. The application of GeONTipe enables the detection of additional SVs within inversion regions, revealing a degree of structural complexity driven by repetitive sequences that is higher-than-expected. This has led to the identification of multiple loci that display distinct structural conformations that are defined as “complex regions”.
7. Based on GeONTipe results, several previously unreported structural conformations were identified in different complex regions, including different SVs that were accurately characterized through detailed manual inspection of the sequence data.
8. CHARONTE is a newly developed tool for the characterisation of complex genomic regions, and was applied to three loci, allowing the detection of multiple structural haplotypes in a large number of individuals and with high accuracy. This tool has demonstrated its ability to genotype complex structural haplotypes in repeat-rich complex regions harbouring multiple SVs that had not been previously explored in depth.

Bibliography

- 100,000 Genomes Project Pilot Investigators, Smedley, D., Smith, K. R., Martin, A., Thomas, E. A., McDonagh, E. M., Cipriani, V., Ellingford, J. M., Arno, G., Tucci, A., Vandrovcova, J., Chan, G., Williams, H. J., Ratnaike, T., Wei, W., Stirrups, K., Ibanez, K., Moutsianas, L., Wielscher, M., ... Caulfield, M. (2021). 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care - Preliminary Report. *The New England Journal of Medicine*, *385*(20), pp. 1868–1880. <https://doi.org/10.1056/NEJMoa2035790>
- Abel, H. J., Larson, D. E., Regier, A. A., Chiang, C., Das, I., Kanchi, K. L., Layer, R. M., Neale, B. M., Salerno, W. J., Reeves, C., Buyske, S., Matise, T. C., Muzny, D. M., Zody, M. C., Lander, E. S., Dutcher, S. K., Stitzel, N. O., & Hall, I. M. (2020). Mapping and characterization of structural variation in 17,795 human genomes. *Nature*, *583*(7814), pp. 83–89. <https://doi.org/10.1038/s41586-020-2371-0>
- Abyzov, A., Li, S., Kim, D. R., Mohiyuddin, M., Stütz, A. M., Parrish, N. F., Mu, X. J., Clark, W., Chen, K., Hurles, M., Korbel, J. O., Lam, H. Y. K., Lee, C., & Gerstein, M. B. (2015). Analysis of deletion breakpoints from 1,092 humans reveals details of mutation mechanisms. *Nature Communications*, *6*(1), pp. 7256. <https://doi.org/10.1038/ncomms8256>
- Aganezov, S., Yan, S. M., Soto, D. C., Kirsche, M., Zarate, S., Avdeyev, P., Taylor, D. J., Shafin, K., Shumate, A., Xiao, C., Wagner, J., McDaniel, J., Olson, N. D., Sauria, M. E. G., Vollger, M. R., Rhie, A., Meredith, M., Martin, S., Lee, J., ... Schatz, M. C. (2022). A complete reference genome improves

- analysis of human genetic variation. *Science (New York, N.Y.)*, 376(6588), pp. eabl3533. <https://doi.org/10.1126/science.abl3533>
- Aguado, C., Gayà-Vidal, M., Villatoro, S., Oliva, M., Izquierdo, D., Giner-Delgado, C., Montalvo, V., García-González, J., Martínez-Fundichely, A., Capilla, L., Ruiz-Herrera, A., Estivill, X., Puig, M., & Cáceres, M. (2014). Validation and genotyping of multiple human polymorphic inversions mediated by inverted repeats reveals a high degree of recurrence. *PLoS genetics*, 10(3), pp. e1004208. <https://doi.org/10.1371/journal.pgen.1004208>
- Ahn, S.-M., Kim, T.-H., Lee, S., Kim, D., Ghang, H., Kim, D.-S., Kim, B.-C., Kim, S.-Y., Kim, W.-Y., Kim, C., Park, D., Lee, Y. S., Kim, S., Reja, R., Jho, S., Kim, C. G., Cha, J.-Y., Kim, K.-H., Lee, B., ... Kim, S.-J. (2009). The first Korean genome sequence and analysis: Full genome sequencing for a socio-ethnic group. *Genome Research*, 19(9), pp. 1622–1629. <https://doi.org/10.1101/gr.092197.109>
- Ahsan, M. U., Liu, Q., Fang, L., & Wang, K. (2021). NanoCaller for accurate detection of SNPs and indels in difficult-to-map regions from long-read sequencing by haplotype-aware deep neural networks. *Genome Biology*, 22(1), pp. 261. <https://doi.org/10.1186/s13059-021-02472-2>
- Alipanahi, B., Kuhnle, A., Puglisi, S. J., Salmela, L., & Boucher, C. (2021). Succinct dynamic de Bruijn graphs. *Bioinformatics*, 37(14), pp. 1946–1952. <https://doi.org/10.1093/bioinformatics/btaa546>
- Alkan, C., Coe, B. P., & Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nature Reviews Genetics*, 12(5), pp. 363–376. <https://doi.org/10.1038/nrg2958>
- Altshuler, D. M., Gibbs, R. A., Peltonen, L., Altshuler, D. M., Gibbs, R. A., Peltonen, L., Dermitzakis, E., Schaffner, S. F., Yu, F., Peltonen, L., Dermitzakis, E., Bonnen, P. E., Altshuler, D. M., Gibbs, R. A., de Bakker, P. I. W., Deloukas, P., Gabriel, S. B., Gwilliam, R., Hunt, S., ... Scientific management. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311), pp. 52–58. <https://doi.org/10.1038/nature09298>

- Alzu'bi, A. A., Zhou, L., & Watzlaf, V. J. M. (2019). Genetic Variations and Precision Medicine. *Perspectives in Health Information Management*, 16, pp. 1a.
- Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., & Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biology*, 21(1), pp. 30. <https://doi.org/10.1186/s13059-020-1935-5>
- Ameur, A., Dahlberg, J., Olason, P., Vezzi, F., Karlsson, R., Martin, M., Viklund, J., Kähäri, A. K., Lundin, P., Che, H., Thutkawkorapin, J., Eisfeldt, J., Lampa, S., Dahlberg, M., Hagberg, J., Jareborg, N., Liljedahl, U., Jonasson, I., Johansson, Å., ... Gyllensten, U. (2017). SweGen: A whole-genome data resource of genetic variability in a cross-section of the Swedish population. *European Journal of Human Genetics*, 25(11), pp. 1253–1260. <https://doi.org/10.1038/ejhg.2017.130>
- Anderson, A. R., Hoffmann, A. A., McKechnie, S. W., Umina, P. A., & Weeks, A. R. (2005). The latitudinal cline in the In(3R)Payne inversion polymorphism has shifted in the last 20 years in Australian *Drosophila melanogaster* populations. *Molecular Ecology*, 14(3), pp. 851–858. <https://doi.org/10.1111/j.1365-294X.2005.02445.x>
- Antonacci, F., Kidd, J. M., Marques-Bonet, T., Ventura, M., Siswara, P., Jiang, Z., & Eichler, E. E. (2009). Characterization of six human disease-associated inversion polymorphisms. *Human Molecular Genetics*, 18(14), pp. 2555–2566. <https://doi.org/10.1093/hmg/ddp187>
- Antonarakis, S. E., Rossiter, J. P., Young, M., Horst, J., Moerlose, P. de, Sommer, S. S., Ketterling, R. P., Kazazian, H. H., Negrier, C., Vinciguerra, C., Gitschier, J., Goossens, M., Girodon, E., Ghanem, N., Plassa, F., Lavergne, J. M., Vidaud, M., Costa, J. M., Laurian, Y., ... Inaba, H. (1995). Factor VIII Gene Inversions in Severe Hemophilia A: Results of an International Consortium Study. *Blood*, 86(6), pp. 2206–2212. <https://doi.org/10.1182/blood.V86.6.2206.bloodjournal8662206>
- Audano, P. A., Sulovari, A., Graves-Lindsay, T. A., Cantsilieris, S., Sorensen, M., Welch, A. E., Dougherty, M. L., Nelson, B. J., Shah, A., Dutcher, S. K., Warren, W. C., Magrini, V., McGrath, S. D., Li, Y. I., Wilson, R. K., &

- Eichler, E. E. (2019). Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell*, 176(3), pp. 663–675.e19. <https://doi.org/10.1016/j.cell.2018.12.019>
- Audano, P. A., Paisie, C., Consortium (HGSVC), T. H. G. S. V., & Beck, C. R. (2024, December 22). *Large complex structural rearrangements in human genomes harbor cryptic structures*. <https://doi.org/10.1101/2024.12.19.629504>
- Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., ... National Eye Institute, N. I. H. (2015). A global reference for human genetic variation. *Nature*, 526(7571), pp. 68–74. <https://doi.org/10.1038/nature15393>
- Auwerx, C., Kutalik, Z., & Reymond, A. (2024). The pleiotropic spectrum of proximal 16p11.2 CNVs. *The American Journal of Human Genetics*, 111(11), pp. 2309–2346. <https://doi.org/10.1016/j.ajhg.2024.08.015>
- Ayala, D., Acevedo, P., Pombi, M., Dia, I., Boccolini, D., Costantini, C., Simard, F., & Fontenille, D. (2017). Chromosome inversions and ecological plasticity in the main African malaria mosquitoes. *Evolution; international journal of organic evolution*, 71(3), pp. 686–701. <https://doi.org/10.1111/evo.13176>
- Bachtrog, D. (2013). Y-chromosome evolution: Emerging insights into processes of Y-chromosome degeneration. *Nature Reviews Genetics*, 14(2), pp. 113–124. <https://doi.org/10.1038/nrg3366>
- Bagnall, R. D., Giannelli, F., & Green, P. M. (2005). Polymorphism and hemophilia A causing inversions in distal Xq28: A complex picture. *Journal of Thrombosis and Haemostasis*, 3(11), pp. 2598–2599. <https://doi.org/10.1111/j.1538-7836.2005.01566.x>
- Bagnall, R. D., Waseem, N., Green, P. M., & Giannelli, F. (2002). Recurrent inversion breaking intron 1 of the factor VIII gene is a frequent cause of severe hemophilia A. *Blood*, 99(1), pp. 168–174. <https://doi.org/10.1182/blood.V99.1.168>

- Bai, X., Chen, Z., Chen, K., Wu, Z., Wang, R., Liu, J., Chang, L., Wen, L., & Tang, F. (2024). Simultaneous de novo calling and phasing of genetic variants at chromosome-scale using NanoStrand-seq. *Cell Discovery*, 10(1), pp. 74. <https://doi.org/10.1038/s41421-024-00694-9>
- Bailey, J. A., Gu, Z., Clark, R. A., Reinert, K., Samonte, R. V., Schwartz, S., Adams, M. D., Myers, E. W., Li, P. W., & Eichler, E. E. (2002). Recent Segmental Duplications in the Human Genome. *Science*, 297(5583), pp. 1003–1007. <https://doi.org/10.1126/science.1072047>
- Balagué-Dobón, L., Cáceres, A., & González, J. R. (2022). Fully exploiting SNP arrays: A systematic review on the tools to extract underlying genomic structure. *Briefings in Bioinformatics*, 23(2), pp. bbac043. <https://doi.org/10.1093/bib/bbac043>
- Balajee, A. S., & Hande, M. P. (2018). History and evolution of cytogenetic techniques: Current and future applications in basic and clinical research. *Mutation Research/Genetic Toxicology and Environmental Mutagenesis*, 836, pp. 3–12. <https://doi.org/10.1016/j.mrgentox.2018.08.008>
- Balanyá, J., Oller, J. M., Huey, R. B., Gilchrist, G. W., & Serra, L. (2006). Global Genetic Change Tracks Global Climate Warming in *Drosophila subobscura*. *Science*, 313(5794), pp. 1773–1775. <https://doi.org/10.1126/science.1131002>
- Ballif, B. C., Theisen, A., Coppinger, J., Gowans, G. C., Hersh, J. H., Madan-Khetarpal, S., Schmidt, K. R., Tervo, R., Escobar, L. F., Friedrich, C. A., McDonald, M., Campbell, L., Ming, J. E., Zackai, E. H., Bejjani, B. A., & Shaffer, L. G. (2008). Expanding the clinical phenotype of the 3q29 microdeletion syndrome and characterization of the reciprocal microduplication. *Molecular Cytogenetics*, 1(1), pp. 8. <https://doi.org/10.1186/1755-8166-1-8>
- Bansal, V., Bashir, A., & Bafna, V. (2007). Evidence for large inversion polymorphisms in the human genome from HapMap data. *Genome Research*, 17(2), pp. 219–230. <https://doi.org/10.1101/gr.5774507>
- Bashirova, A. A., Zheng, W., Akdag, M., Augusto, D. G., Vince, N., Dong, K. L., O’hUigin, C., & Carrington, M. (2021). Population-specific diversity of the

- immunoglobulin constant heavy G chain (IGHG) genes. *Genes & Immunity*, 22(7), pp. 327–334. <https://doi.org/10.1038/s41435-021-00156-2>
- Bayés, M., Magano, L. F., Rivera, N., Flores, R., & Pérez Jurado, L. A. (2003). Mutational mechanisms of Williams-Beuren syndrome deletions. *American Journal of Human Genetics*, 73(1), pp. 131–151. <https://doi.org/10.1086/376565>
- Beau, M. M. L., Larson, R. A., Bitter, M. A., Vardiman, J. W., Golomb, H. M., & Rowley, J. D. (1983). Association of an Inversion of Chromosome 16 with Abnormal Marrow Eosinophils in Acute Myelomonocytic Leukemia. *New England Journal of Medicine*, 309(11), pp. 630–636. <https://doi.org/10.1056/NEJM198309153091103>
- Beck, C. R., Carvalho, C. M. B., Banser, L., Gambin, T., Stubbolo, D., Yuan, B., Sperle, K., McCahan, S. M., Henneke, M., Seeman, P., Garbern, J. Y., Hobson, G. M., & Lupski, J. R. (2015). Complex Genomic Rearrangements at the PLP1 Locus Include Triplication and Quadruplication. *PLOS Genetics*, 11(3), pp. e1005050. <https://doi.org/10.1371/journal.pgen.1005050>
- Berntorp, E., Fischer, K., Hart, D. P., Mancuso, M. E., Stephensen, D., Shapiro, A. D., & Blanchette, V. (2021). Haemophilia. *Nature Reviews Disease Primers*, 7(1), pp. 45. <https://doi.org/10.1038/s41572-021-00278-x>
- Besenbacher, S., Liu, S., Izarzugaza, J. M. G., Grove, J., Belling, K., Bork-Jensen, J., Huang, S., Als, T. D., Li, S., Yadav, R., Rubio-García, A., Lescai, F., Demontis, D., Rao, J., Ye, W., Mailund, T., Friberg, R. M., Pedersen, C. N. S., Xu, R., . . . Rasmussen, S. (2015). Novel variation and de novo mutation rates in population-wide de novo assembled Danish trios. *Nature Communications*, 6(1), pp. 5969. <https://doi.org/10.1038/ncomms6969>
- Beyter, D., Ingimundardottir, H., Oddsson, A., Eggertsson, H. P., Bjornsson, E., Jonsson, H., Atlason, B. A., Kristmundsdottir, S., Mehringer, S., Hardarson, M. T., Gudjonsson, S. A., Magnusdottir, D. N., Jonasdottir, A., Jonasdottir, A., Kristjansson, R. P., Sverrisson, S. T., Holley, G., Palsson, G., Stefansson, O. A., . . . Stefansson, K. (2021). Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and

- other traits. *Nature Genetics*, 53(6), pp. 779–786. <https://doi.org/10.1038/s41588-021-00865-4>
- Biesecker, L. G., & Spinner, N. B. (2013). A genomic view of mosaicism and human disease. *Nature Reviews Genetics*, 14(5), pp. 307–320. <https://doi.org/10.1038/nrg3424>
- Boerma, E. G., Siebert, R., Kluin, P. M., & Baudis, M. (2009). Translocations involving 8q24 in Burkitt lymphoma and other malignant lymphomas: A historical review of cytogenetics in the light of today's knowledge. *Leukemia*, 23(2), pp. 225–234. <https://doi.org/10.1038/leu.2008.281>
- Boettger, L. M., Handsaker, R. E., Zody, M. C., & McCarroll, S. A. (2012). Structural haplotypes and recent evolution of the human 17q21.31 region. *Nature Genetics*, 44(8), pp. 881–885. <https://doi.org/10.1038/ng.2334>
- Bolognini, D., Halgren, A., Lou, R. N., Raveane, A., Rocha, J. L., Guarracino, A., Soranzo, N., Chin, C.-S., Garrison, E., & Sudmant, P. H. (2024). Recurrent evolution and selection shape structural diversity at the amylase locus. *Nature*, 634(8034), pp. 617–625. <https://doi.org/10.1038/s41586-024-07911-1>
- Bondeson, M.-L., Dahl, N., Malmgren, H., Kleijer, W. J., Tönnesen, T., Carlberg, B.-M., & Pettersson, U. (1995). Inversion of the IDS gene resulting from recombination with IDS-related sequences in a common cause of the Hunter syndrome. *Human Molecular Genetics*, 4(4), pp. 615–621. <https://doi.org/10.1093/hmg/4.4.615>
- Boratyn, G. M., Camacho, C., Cooper, P. S., Coulouris, G., Fong, A., Ma, N., Madden, T. L., Matten, W. T., McGinnis, S. D., Merezuk, Y., Raytselis, Y., Sayers, E. W., Tao, T., Ye, J., & Zaretskaya, I. (2013). BLAST: A more efficient report with usability improvements. *Nucleic Acids Research*, 41, pp. W29–W33. <https://doi.org/10.1093/nar/gkt282>
- Bosch, N., Morell, M., Ponsa, I., Mercader, J. M., Armengol, L., & Estivill, X. (2009). Nucleotide, Cytogenetic and Expression Impact of the Human Chromosome 8p23.1 Inversion Polymorphism. *PLOS ONE*, 4(12), pp. e8269. <https://doi.org/10.1371/journal.pone.0008269>

- Botstein, D., White, R. L., Skolnick, M., & Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics*, 32(3), pp. 314–331.
- Bozkurt-Yozgatli, T., Lun, M. Y., Bengtsson, J. D., Sezerman, U., Chinn, I. K., Coban-Akdemir, Z., & Carvalho, C. M. B. (2025). Investigation of a pathogenic inversion in UNC13D and comprehensive analysis of chromosomal inversions across diverse datasets. *European Journal of Human Genetics*, 33(7), pp. 887–895. <https://doi.org/10.1038/s41431-025-01817-w>
- Brehm, A., & Krimbas, C. B. (1991). Inversion Polymorphism in *Drosophila obscura*. *Journal of Heredity*, 82(2), pp. 110–117. <https://doi.org/10.1093/oxfordjournals.jhered.a111044>
- Bursted, B., Zamariolli, M., Bellucco, F. T., & Melaragno, M. I. (2022). Mechanisms of structural chromosomal rearrangement formation. *Molecular Cytogenetics*, 15(1), pp. 23. <https://doi.org/10.1186/s13039-022-00600-6>
- Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O’Connell, J., Cortes, A., Welsh, S., Young, A., Effingham, M., McVean, G., Leslie, S., Allen, N., Donnelly, P., & Marchini, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726), pp. 203–209. <https://doi.org/10.1038/s41586-018-0579-z>
- Byrska-Bishop, M., Evani, U. S., Zhao, X., Basile, A. O., Abel, H. J., Regier, A. A., Corvelo, A., Clarke, W. E., Musunuri, R., Nagulapalli, K., Fairley, S., Runnels, A., Winterkorn, L., Lowy, E., Eichler, E. E., Korb, J. O., Lee, C., Marschall, T., Devine, S. E., ... Zody, M. C. (2022). High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell*, 185(18), pp. 3426–3440.e19. <https://doi.org/10.1016/j.cell.2022.08.004>
- Cáceres, A., & González, J. R. (2015). Following the footprints of polymorphic inversions on SNP data: From detection to association tests. *Nucleic Acids Research*, 43(8), pp. e53. <https://doi.org/10.1093/nar/gkv073>
- Cáceres, M., National Institutes of Health Intramural Sequencing Center Comparative Sequencing Program, Sullivan, R. T., & Thomas, J. W. (2007). A recur-

- rent inversion on the eutherian X chromosome. *Proceedings of the National Academy of Sciences*, 104(47), pp. 18571–18576. <https://doi.org/10.1073/pnas.0706604104>
- Campoy, E. (2024, January 5). *Bioinformatic analysis of the functional effects of human polymorphic inversions on gene expression, epigenetics and phenotypic traits*.
- Campoy, E., Puig, M., Yakymenko, I., Lerga-Jaso, J., & Cáceres, M. (2022). Genomic architecture and functional effects of potential human inversion supergenes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 377(1856), pp. 20210209. <https://doi.org/10.1098/rstb.2021.0209>
- Carrara, M., Gaillard, A.-L., Brion, A., Duvernois-Berthet, E., Giovannangeli, C., Concordet, J.-P., & Pézeron, G. (2025). Dynamic interplay of cNHEJ and MMEJ pathways of DNA double-strand break repair during embryonic development in zebrafish. *Scientific Reports*, 15(1), pp. 4886. <https://doi.org/10.1038/s41598-025-88564-6>
- Carvalho, C. M. B., & Lupski, J. R. (2016). Mechanisms underlying structural variant formation in genomic disorders. *Nature Reviews. Genetics*, 17(4), pp. 224–238. <https://doi.org/10.1038/nrg.2015.25>
- Carver, T., Harris, S. R., Berriman, M., Parkhill, J., & McQuillan, J. A. (2012). Artemis: An integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics*, 28(4), pp. 464–469. <https://doi.org/10.1093/bioinformatics/btr703>
- Chaisson, M. J. P., Huddleston, J., Dennis, M. Y., Sudmant, P. H., Malig, M., Hormozdiari, F., Antonacci, F., Surti, U., Sandstrom, R., Boitano, M., Landolin, J. M., Stamatoyannopoulos, J. A., Hunkapiller, M. W., Korlach, J., & Eichler, E. E. (2015). Resolving the complexity of the human genome using single-molecule sequencing. *Nature*, 517(7536), pp. 608–611. <https://doi.org/10.1038/nature13907>
- Chaisson, M. J. P., Sanders, A. D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., Gardner, E. J., Rodriguez, O. L., Guo, L., Collins, R. L., Fan, X., Wen, J., Handsaker, R. E., Fairley, S., Kronenberg, Z. N., Kong, X., Hormozdiari,

- F., Lee, D., Wenger, A. M., ... Lee, C. (2019). Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nature Communications*, 10(1), pp. 1784. <https://doi.org/10.1038/s41467-018-08148-z>
- Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*, 4(1), pp. s13742-015-0047–8. <https://doi.org/10.1186/s13742-015-0047-8>
- Chen, J.-M., Chuzhanova, N., Stenson, P. D., Férec, C., & Cooper, D. N. (2005). Intrachromosomal serial replication slippage in trans gives rise to diverse genomic rearrangements involving inversions. *Human Mutation*, 26(4), pp. 362–373. <https://doi.org/10.1002/humu.20230>
- Chen, X., & Zhang, J. (2016). The Genomic Landscape of Position Effects on Protein Expression Level and Noise in Yeast. *Cell Systems*, 2(5), pp. 347–354. <https://doi.org/10.1016/j.cels.2016.03.009>
- Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., & Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature methods*, 18(2), pp. 170–175. <https://doi.org/10.1038/s41592-020-01056-5>
- Chevrier, S., Richard, C., Mille, M., Bertrand, D., & Boidot, R. (2025). Nanopore adaptive sampling accurately detects nucleotide variants and improves the characterization of large-scale rearrangement for the diagnosis of cancer predisposition. *Clinical and Translational Medicine*, 15(1), pp. e70138. <https://doi.org/10.1002/ctm2.70138>
- Chheda, H., Palta, P., Pirinen, M., McCarthy, S., Walter, K., Koskinen, S., Salomaa, V., Daly, M., Durbin, R., Palotie, A., Aittokallio, T., & Ripatti, S. (2017). Whole-genome view of the consequences of a population bottleneck using 2926 genome sequences from Finland and United Kingdom. *European Journal of Human Genetics*, 25(4), pp. 477–484. <https://doi.org/10.1038/ejhg.2016.205>
- Chin, C.-S., Behera, S., Khalak, A., Sedlazeck, F. J., Sudmant, P. H., Wagner, J., & Zook, J. M. (2023). Multiscale analysis of pangenomes enables improved representation of genomic diversity for repetitive and clinically relevant genes.

- Nature Methods*, 20(8), pp. 1213–1221. <https://doi.org/10.1038/s41592-023-01914-y>
- Choudhury, A., Ramsay, M., Hazelhurst, S., Aron, S., Bardien, S., Botha, G., Chimusa, E. R., Christoffels, A., Gamielien, J., Sefid-Dashti, M. J., Joubert, F., Meintjes, A., Mulder, N., Ramesar, R., Rees, J., Scholtz, K., Sengupta, D., Soodyall, H., Venter, P., ... Pepper, M. S. (2017). Whole-genome sequencing for an enhanced understanding of genetic variation among South Africans. *Nature Communications*, 8(1), pp. 2062. <https://doi.org/10.1038/s41467-017-00663-9>
- Collins, R. L., Brand, H., Karczewski, K. J., Zhao, X., Alföldi, J., Francioli, L. C., Khera, A. V., Lowther, C., Gauthier, L. D., Wang, H., Watts, N. A., Solomonson, M., O'Donnell-Luria, A., Baumann, A., Munshi, R., Walker, M., Whelan, C. W., Huang, Y., Brookings, T., ... Talkowski, M. E. (2020). A structural variation reference for medical and population genetics. *Nature*, 581(7809), pp. 444–451. <https://doi.org/10.1038/s41586-020-2287-8>
- Cook, G. W., Benton, M. G., Akerley, W., Mayhew, G. F., Moehlenkamp, C., Raterman, D., Burgess, D. L., Rowell, W. J., Lambert, C., Eng, K., Gu, J., Baybayan, P., Fussell, J. T., Herbold, H. D., O'Shea, J. M., Varghese, T. K., & Emerson, L. L. (2020). Structural variation and its potential impact on genome instability: Novel discoveries in the EGFR landscape by long-read sequencing. *PloS One*, 15(1), pp. e0226340. <https://doi.org/10.1371/journal.pone.0226340>
- Cooper, G. M., Coe, B. P., Girirajan, S., Rosenfeld, J. A., Vu, T. H., Baker, C., Williams, C., Stalker, H., Hamid, R., Hannig, V., Abdel-Hamid, H., Bader, P., McCracken, E., Niyazov, D., Leppig, K., Thiese, H., Hummel, M., Alexander, N., Gorski, J., ... Eichler, E. E. (2011). A copy number variation morbidity map of developmental delay. *Nature Genetics*, 43(9), pp. 838–846. <https://doi.org/10.1038/ng.909>
- Cuenca-Guardiola, J., de la Morena-Barrio, B., García, J. L., Sanchis-Juan, A., Corral, J., & Fernández-Breis, J. T. (2023). Improvement of large copy number

- variant detection by whole genome nanopore sequencing. *Journal of Advanced Research*, 50, pp. 145–158. <https://doi.org/10.1016/j.jare.2022.10.012>
- Dana, M., & Stoian, V. (2012). Association of Pericentric Inversion of Chromosome 9 and Infertility in Romanian Population. *Mædica*, 7(1), pp. 25–29.
- Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., & Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, 27(15), pp. 2156–2158. <https://doi.org/10.1093/bioinformatics/btr330>
- Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, 10(2), pp. giab008. <https://doi.org/10.1093/gigascience/giab008>
- De Coster, W., Weissensteiner, M. H., & Sedlazeck, F. J. (2021). Towards population-scale long-read sequencing. *Nature Reviews Genetics*, 22(9), pp. 572–587. <https://doi.org/10.1038/s41576-021-00367-3>
- De Coster, W., & Rademakers, R. (2023). NanoPack2: Population-scale evaluation of long-read sequencing data. *Bioinformatics*, 39(5). <https://doi.org/10.1093/bioinformatics/btad311>
- Delaneau, O., Zagury, J.-F., Robinson, M. R., Marchini, J. L., & Dermitzakis, E. T. (2019). Accurate, scalable and integrative haplotype estimation. *Nature Communications*, 10(1), pp. 5436. <https://doi.org/10.1038/s41467-019-13225-y>
- Di Genova, A., Buena-Atienza, E., Ossowski, S., & Sagot, M.-F. (2021). Efficient hybrid de novo assembly of human genomes with WENGAN. *Nature Biotechnology*, 39(4), pp. 422–430. <https://doi.org/10.1038/s41587-020-00747-w>
- Dijk, E. L. van, Jaszczyszyn, Y., Naquin, D., & Thermes, C. (2018). The Third Revolution in Sequencing Technology. *Trends in Genetics*, 34(9), pp. 666–681. <https://doi.org/10.1016/j.tig.2018.05.008>
- Dishuck, P. C., Munson, K. M., Lewis, A. P., Dougherty, M. L., Underwood, J. G., Harvey, W. T., Hsieh, P., Pastinen, T., & Eichler, E. E. (2025). Structural variation, selection, and diversification of the *NP1P* gene family from the

- human pangenome. *Cell Genomics*, 5(10), pp. 100977. <https://doi.org/10.1016/j.xgen.2025.100977>
- Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., & Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398), pp. 376–380. <https://doi.org/10.1038/nature11082>
- Dobzhansky, T., & Sturtevant, A. H. (1938). Inversions in the chromosomes of *Drosophila pseudoobscura*. *Genetics*, 23(1), pp. 28–64. <https://doi.org/10.1093/genetics/23.1.28>
- Durbin, R. M., Altshuler, D., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Collins, F. S., De La Vega, F. M., Donnelly, P., Egholm, M., Flicek, P., Gabriel, S. B., Gibbs, R. A., Knoppers, B. M., Lander, E. S., Lehrach, H., Mardis, E. R., McVean, G. A., ... The Translational Genomics Research Institute. (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319), pp. 1061–1073. <https://doi.org/10.1038/nature09534>
- Durmaz, A. A., Karaca, E., Demkow, U., Toruner, G., Schoumans, J., & Cogulu, O. (2015). Evolution of Genetic Techniques: Past, Present, and Beyond. *BioMed Research International*, 2015(1), pp. 461524. <https://doi.org/10.1155/2015/461524>
- Ebert, P., Audano, P. A., Zhu, Q., Rodriguez-Martin, B., Porubsky, D., Bonder, M. J., Sulovari, A., Ebler, J., Zhou, W., Serra Mari, R., Yilmaz, F., Zhao, X., Hsieh, P., Lee, J., Kumar, S., Lin, J., Rausch, T., Chen, Y., Ren, J., ... Eichler, E. E. (2021). Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science*, 372(6537), pp. eabf7117. <https://doi.org/10.1126/science.abf7117>
- Eichler, E. E. (2001). Recent duplication, domain accretion and the dynamic mutation of the human genome. *Trends in Genetics*, 17(11), pp. 661–669. [https://doi.org/10.1016/S0168-9525\(01\)02492-1](https://doi.org/10.1016/S0168-9525(01)02492-1)
- Escaramís, G., Docampo, E., & Rabionet, R. (2015). A decade of structural variants: Description, history and methods to detect structural variation. *Briefings in*

- Functional Genomics*, 14(5), pp. 305–314. <https://doi.org/10.1093/bfgp/elv014>
- Espinosa, E., Bautista, R., Larrosa, R., & Plata, O. (2024). Advancements in long-read genome sequencing technologies and algorithms. *Genomics*, 116(3), pp. 110842. <https://doi.org/10.1016/j.ygeno.2024.110842>
- Falconer, E., Hills, M., Naumann, U., Poon, S. S. S., Chavez, E. A., Sanders, A. D., Zhao, Y., Hirst, M., & Lansdorp, P. M. (2012). DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution. *Nature Methods*, 9(11), pp. 1107–1112. <https://doi.org/10.1038/nmeth.2206>
- Fang, Z., Pyhäjärvi, T., Weber, A. L., Dawe, R. K., Glaubitz, J. C., González, J. d. J. S., Ross-Ibarra, C., Doebley, J., Morrell, P. L., & Ross-Ibarra, J. (2012). Megabase-Scale Inversion Polymorphism in the Wild Ancestor of Maize. *Genetics*, 191(3), pp. 883–894. <https://doi.org/10.1534/genetics.112.138578>
- Ferguson, S., McLay, T., Andrew, R. L., Bruhl, J. J., Schwessinger, B., Borevitz, J., & Jones, A. (2022). Species-specific basecallers improve actual accuracy of nanopore sequencing in plants. *Plant Methods*, 18, pp. 137. <https://doi.org/10.1186/s13007-022-00971-2>
- Feuk, L. (2010). Inversion variants in the human genome: Role in disease and genome architecture. *Genome Medicine*, 2(2), pp. 11. <https://doi.org/10.1186/gm132>
- Feuk, L., MacDonald, J. R., Tang, T., Carson, A. R., Li, M., Rao, G., Khaja, R., & Scherer, S. W. (2005). Discovery of Human Inversion Polymorphisms by Comparative Analysis of Human and Chimpanzee DNA Sequence Assemblies. *PLOS Genetics*, 1(4), pp. e56. <https://doi.org/10.1371/journal.pgen.0010056>
- Fickelscher, I., Liehr, T., Watts, K., Bryant, V., Barber, J. C. K., Heidemann, S., Siebert, R., Hertz, J. M., Tümer, Z., & Thomas, N. S. (2007). The Variant inv(2)(p11.2q13) Is a Genuinely Recurrent Rearrangement but Displays Some Breakpoint Heterogeneity. *The American Journal of Human Genetics*, 81(4), pp. 847–856. <https://doi.org/10.1086/521226>
- Flavahan, W. A., Drier, Y., Liao, B. B., Gillespie, S. M., Venteicher, A. S., Stemmer-Rachamimov, A. O., Suvà, M. L., & Bernstein, B. E. (2016). Insulator dys-

- function and oncogene activation in IDH mutant gliomas. *Nature*, 529(7584), pp. 110–114. <https://doi.org/10.1038/nature16490>
- Foreman, J., Perrett, D., Mazaika, E., Hunt, S. E., Ware, J. S., & Firth, H. V. (2023). DECIPHER: Improving Genetic Diagnosis Through Dynamic Integration of Genomic and Clinical Data. *Annual Review of Genomics and Human Genetics*, 24, pp. 151–176. <https://doi.org/10.1146/annurev-genom-102822-100509>
- Fornezza, S., Delvecchio, V. S., Harvey, W. T., Dishuck, P. C., Eichler, E. E., & Giannuzzi, G. (2024). AGAP duplicons associate with structural diversity at Chromosome 10q11.22. *Genome Research*, 34(10), pp. 1487–1499. <https://doi.org/10.1101/gr.279454.124>
- Franke, M., Ibrahim, D. M., Andrey, G., Schwarzer, W., Heinrich, V., Schöpflin, R., Kraft, K., Kempfer, R., Jerković, I., Chan, W.-L., Spielmann, M., Timmermann, B., Wittler, L., Kurth, I., Cambiaso, P., Zuffardi, O., Houge, G., Lambie, L., Brancati, F., ... Mundlos, S. (2016). Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature*, 538(7624), pp. 265–269. <https://doi.org/10.1038/nature19800>
- Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., Belmont, J. W., Boudreau, A., Hardenbol, P., Leal, S. M., Pasternak, S., Wheeler, D. A., Willis, T. D., Yu, F., Yang, H., Zeng, C., Gao, Y., Hu, H., & Hu, W. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164), pp. 851–861. <https://doi.org/10.1038/nature06258>
- Fuller, Z. L., Koury, S. A., Phadnis, N., & Schaeffer, S. W. (2019). How chromosomal rearrangements shape adaptation and speciation: Case studies in *Drosophila pseudoobscura* and its sibling species *Drosophila persimilis*. *Molecular Ecology*, 28(6), pp. 1283–1301. <https://doi.org/10.1111/mec.14923>
- Gao, Y., Yang, X., Chen, H., Tan, X., Yang, Z., Deng, L., Wang, B., Kong, S., Li, S., Cui, Y., Lei, C., Wang, Y., Pan, Y., Ma, S., Sun, H., Zhao, X., Shi, Y., Yang, Z., Wu, D., ... Xu, S. (2023). A pangenome reference of 36 Chinese populations. *Nature*, 619(7968), pp. 112–121. <https://doi.org/10.1038/s41586-023-06173-7>

- Geng, K., Merino, L. G., Wedemann, L., Martens, A., Sobota, M., Sanchez, Y. P., Søndergaard, J. N., White, R. J., & Kutter, C. (2022). Target-enriched nanopore sequencing and de novo assembly reveals co-occurrences of complex on-target genomic rearrangements induced by CRISPR-Cas9 in human cells. *Genome Research*. <https://doi.org/10.1101/gr.276901.122>
- Genome of the Netherlands Consortium. (2014). Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nature Genetics*, 46(8), pp. 818–825. <https://doi.org/10.1038/ng.3021>
- Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Yang, & The International HapMap Consortium. (2003). The International HapMap Project. *Nature*, 426(6968), pp. 789–796. <https://doi.org/10.1038/nature02168>
- Gimelli, G., Pujana, M. A., Patricelli, M. G., Russo, S., Giardino, D., Larizza, L., Cheung, J., Armengol, L., Schinzel, A., Estivill, X., & Zuffardi, O. (2003). Genomic inversions of human chromosome 15q11–q13 in mothers of Angelman syndrome patients with class II (BP2/3) deletions. *Human Molecular Genetics*, 12(8), pp. 849–858. <https://doi.org/10.1093/hmg/ddg101>
- Giner-Delgado, C., Villatoro, S., Lerga-Jaso, J., Gayà-Vidal, M., Oliva, M., Castellano, D., Pantano, L., Bitarello, B. D., Izquierdo, D., Noguera, I., Olalde, I., Delprat, A., Blancher, A., Lalueza-Fox, C., Esko, T., O'Reilly, P. F., Andrés, A. M., Ferretti, L., Puig, M., & Cáceres, M. (2019). Evolutionary and functional impact of common polymorphic inversions in the human genome. *Nature Communications*, 10(1), pp. 4222. <https://doi.org/10.1038/s41467-019-12173-x>
- Ginsburg, G. S., Denny, J. C., & Schully, S. D. (2023). Data-Driven Science and Diversity in the All of Us Research Program. *Science translational medicine*, 15(726), pp. eade9214. <https://doi.org/10.1126/scitranslmed.ade9214>
- González, J. R., Cáceres, A., Esko, T., Cuscó, I., Puig, M., Esnaola, M., Reina, J., Siroux, V., Bouzigon, E., Nadif, R., Reinmaa, E., Milani, L., Bustamante, M., Jarvis, D., Antó, J. M., Sunyer, J., Demenais, F., Kogevinas, M., Metspalu, A., ... Pérez-Jurado, L. A. (2014a). A Common 16p11.2 Inversion Underlies the Joint Susceptibility to Asthma and Obesity. *The American Journal of*

- Human Genetics*, 94(3), pp. 361–372. <https://doi.org/10.1016/j.ajhg.2014.01.015>
- González, J. R., Cáceres, A., Esko, T., Cuscó, I., Puig, M., Esnaola, M., Reina, J., Siroux, V., Bouzigon, E., Nadif, R., Reinmaa, E., Milani, L., Bustamante, M., Jarvis, D., Antó, J. M., Sunyer, J., Demenais, F., Kogevinas, M., Metspalu, A., ... Pérez-Jurado, L. A. (2014b). A Common 16p11.2 Inversion Underlies the Joint Susceptibility to Asthma and Obesity. *The American Journal of Human Genetics*, 94(3), pp. 361–372. <https://doi.org/10.1016/j.ajhg.2014.01.015>
- González, J. R., Ruiz-Arenas, C., Cáceres, A., Morán, I., López-Sánchez, M., Alonso, L., Tolosana, I., Guindo-Martínez, M., Mercader, J. M., Esko, T., Torrents, D., González, J., & Pérez-Jurado, L. A. (2020). Polymorphic Inversions Underlie the Shared Genetic Susceptibility of Obesity-Related Diseases. *The American Journal of Human Genetics*, 106(6), pp. 846–858. <https://doi.org/10.1016/j.ajhg.2020.04.017>
- Gorello, P., Nofrini, V., Brandimarte, L., Pierini, V., Crescenzi, B., Nozza, F., Daniele, G., Storlazzi, C. T., Di Giacomo, D., Matteucci, C., La Starza, R., & Mecucci, C. (2013). Inv(11)(p15q22)/NUP98-DDX10 fusion and isoforms in a new case of de novo acute myeloid leukemia. *Cancer Genetics*, 206(3), pp. 92–96. <https://doi.org/10.1016/j.cancergen.2013.02.001>
- Gouw, S. C., van den Berg, H. M., Oldenburg, J., Astermark, J., de Groot, P. G., Margaglione, M., Thompson, A. R., van Heerde, W., Boekhorst, J., Miller, C. H., le Cessie, S., & van der Bom, J. G. (2012). F8 gene mutation type and inhibitor development in patients with severe hemophilia A: Systematic review and meta-analysis. *Blood*, 119(12), pp. 2922–2934. <https://doi.org/10.1182/blood-2011-09-379453>
- Graffelman, J. (2015). Exploring Diallelic Genetic Markers: The HardyWeinberg Package. *Journal of Statistical Software*, 64, pp. 1–23. <https://doi.org/10.18637/jss.v064.i03>
- Grimes, K., Jeong, H., Amoah, A., Xu, N., Niemann, J., Raeder, B., Hasenfeld, P., Stober, C., Rausch, T., Benito, E., Jann, J.-C., Nowak, D., Emini, R.,

- Hoenicka, M., Liebold, A., Ho, A., Shuai, S., Geiger, H., Sanders, A. D., & Korbel, J. O. (2024). Cell-type-specific consequences of mosaic structural variants in hematopoietic stem and progenitor cells. *Nature Genetics*, 56(6), pp. 1134–1146. <https://doi.org/10.1038/s41588-024-01754-2>
- Gruber, T. A., Larson Gedman, A., Zhang, J., Koss, C. S., Marada, S., Ta, H. Q., Chen, S.-C., Su, X., Ogden, S. K., Dang, J., Wu, G., Gupta, V., Andersson, A. K., Pounds, S., Shi, L., Easton, J., Barbato, M. I., Mulder, H. L., Manne, J., ... Downing, J. R. (2012). An Inv(16)(p13.3q24.3)-Encoded CBFA2T3-GLIS2 Fusion Protein Defines an Aggressive Subtype of Pediatric Acute Megakaryoblastic Leukemia. *Cancer Cell*, 22(5), pp. 683–697. <https://doi.org/10.1016/j.ccr.2012.10.007>
- Gudmundsson, S., Singer-Berk, M., Watts, N. A., Phu, W., Goodrich, J. K., Solomonson, M., Consortium, G. A. D., Rehm, H. L., MacArthur, D. G., & O'Donnell-Luria, A. (2022). Variant interpretation using population databases: Lessons from gnomAD. *Human Mutation*, 43(8), pp. 1012–1030. <https://doi.org/10.1002/humu.24309>
- Guitart, X., Porubsky, D., Yoo, D., Dougherty, M. L., Dishuck, P. C., Munson, K. M., Lewis, A. P., Hoekzema, K., Knuth, J., Chang, S., Pastinen, T., & Eichler, E. E. (2024). Independent expansion, selection, and hypervariability of the TBC1D3 gene family in humans. *Genome Research*, 34(11), pp. 1798–1810. <https://doi.org/10.1101/gr.279299.124>
- Gurdasani, D., Carstensen, T., Tekola-Ayele, F., Pagani, L., Tachmazidou, I., Hatzikotoulas, K., Karthikeyan, S., Iles, L., Pollard, M. O., Choudhury, A., Ritchie, G. R. S., Xue, Y., Asimit, J., Nsubuga, R. N., Young, E. H., Pomilla, C., Kivinen, K., Rockett, K., Kamali, A., ... Sandhu, M. S. (2015). The African Genome Variation Project shapes medical genetics in Africa. *Nature*, 517(7534), pp. 327–332. <https://doi.org/10.1038/nature13997>
- Gustafson, J. A., Gibson, S. B., Damaraju, N., Zalusky, M. P. G., Hoekzema, K., Twesigomwe, D., Yang, L., Snead, A. A., Richmond, P. A., Coster, W. D., Olson, N. D., Guarracino, A., Li, Q., Miller, A. L., Goffena, J., Anderson, Z. B., Storz, S. H. R., Ward, S. A., Sinha, M., ... Miller, D. E. (2024). High-

- coverage nanopore sequencing of samples from the 1000 Genomes Project to build a comprehensive catalog of human genetic variation. *Genome Research*, 34(11), pp. 2061–2073. <https://doi.org/10.1101/gr.279273.124>
- Gutiérrez-Valencia, J., Hughes, P. W., Berdan, E. L., & Slotte, T. (2021). The Genomic Architecture and Evolutionary Fates of Supergenes. *Genome Biology and Evolution*, 13(5), pp. evab057. <https://doi.org/10.1093/gbe/evab057>
- Hach, F., Sarrafi, I., Hormozdiari, F., Alkan, C., Eichler, E. E., & Sahinalp, S. C. (2014). mrsFAST-Ultra: A compact, SNP-aware mapper for high performance sequencing applications. *Nucleic Acids Research*, 42, pp. W494–W500. <https://doi.org/10.1093/nar/gku370>
- Hallast, P., Ebert, P., Loftus, M., Yilmaz, F., Audano, P. A., Logsdon, G. A., Bonder, M. J., Zhou, W., Höps, W., Kim, K., Li, C., Hoyt, S. J., Dishuck, P. C., Porubsky, D., Tsetsos, F., Kwon, J. Y., Zhu, Q., Munson, K. M., Hasenfeld, P., ... Lee, C. (2023). Assembly of 43 human Y chromosomes reveals extensive complexity and variation. *Nature*, 621(7978), pp. 355–364. <https://doi.org/10.1038/s41586-023-06425-6>
- Hanlon, V. C. T., Mattsson, C.-A., Spierings, D. C. J., Guryev, V., & Lansdorp, P. M. (2021). InvertypeR: Bayesian inversion genotyping with Strand-seq data. *BMC Genomics*, 22(1), pp. 582. <https://doi.org/10.1186/s12864-021-07892-9>
- Hanlon, V. C. T., Lansdorp, P. M., & Guryev, V. (2022). A survey of current methods to detect and genotype inversions. *Human Mutation*, 43(11), pp. 1576–1589. <https://doi.org/10.1002/humu.24458>
- Hanlon, V. C. T., Porubsky, D., & Lansdorp, P. M. (2023). Chromosome-Length Haplotypes with StrandPhaseR and Strand-seq. In B. A. Peters & R. Drmanac (Eds.), *Haplotyping: Methods and Protocols* (pp. 183–200). Springer US. https://doi.org/10.1007/978-1-0716-2819-5_12
- Hastings, P. J., Lupski, J. R., Rosenberg, S. M., & Ira, G. (2009). Mechanisms of change in gene copy number. *Nature Reviews Genetics*, 10(8), pp. 551–564. <https://doi.org/10.1038/nrg2593>

- Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, *107*(1), pp. 1–8. <https://doi.org/10.1016/j.ygeno.2015.11.003>
- Hehir-Kwa, J. Y., Marschall, T., Kloosterman, W. P., Francioli, L. C., Baaijens, J. A., Dijkstra, L. J., Abdellaoui, A., Koval, V., Thung, D. T., Wardenaar, R., Renkens, I., Coe, B. P., Deelen, P., de Ligt, J., Lameijer, E.-W., van Dijk, F., Hormozdiari, F., Uitterlinden, A. G., van Duijn, C. M., ... Guryev, V. (2016). A high-quality human reference panel reveals the complexity and distribution of genomic structural variants. *Nature Communications*, *7*(1), pp. 12989. <https://doi.org/10.1038/ncomms12989>
- Heinz, J. M., Meyerson, M., & Li, H. (2025). Detecting Foldback Artifacts in Long-reads. *bioRxiv*, pp. 2025.07.15.664946. <https://doi.org/10.1101/2025.07.15.664946>
- Helal, A. A., Saad, B. T., Saad, M. T., Mosaad, G. S., & Aboshanab, K. M. (2024). Benchmarking long-read aligners and SV callers for structural variation detection in Oxford nanopore sequencing data. *Scientific Reports*, *14*(1), pp. 6160. <https://doi.org/10.1038/s41598-024-56604-2>
- Hinrichs, A. S., Karolchik, D., Baertsch, R., Barber, G. P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T. S., Harte, R. A., Hsu, F., Hillman-Jackson, J., Kuhn, R. M., Pedersen, J. S., Pohl, A., Raney, B. J., Rosenbloom, K. R., Siepel, A., Smith, K. E., Sugnet, C. W., ... Kent, W. J. (2006). The UCSC Genome Browser Database: Update 2006. *Nucleic Acids Research*, *34*, pp. D590–D598. <https://doi.org/10.1093/nar/gkj144>
- Hnisz, D., Weintraub, A. S., Day, D. S., Valton, A.-L., Bak, R. O., Li, C. H., Goldmann, J., Lajoie, B. R., Fan, Z. P., Sigova, A. A., Reddy, J., Borges-Rivera, D., Lee, T. I., Jaenisch, R., Porteus, M. H., Dekker, J., & Young, R. A. (2016). Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science*, *351*(6280), pp. 1454–1458. <https://doi.org/10.1126/science.aad9024>
- Ho, S. S., Urban, A. E., & Mills, R. E. (2020). Structural variation in the sequencing era. *Nature Reviews Genetics*, *21*(3), pp. 171–189. <https://doi.org/10.1038/s41576-019-0180-9>

- Hobart, H. H., Morris, C. A., Mervis, C. B., Pani, A. M., Kistler, D. J., Rios, C. M., Kimberley, K. W., Gregg, R. G., & Bray-Ward, P. (2010). Inversion of the Williams syndrome region is a common polymorphism found more frequently in parents of children with Williams syndrome. *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*, 154C(2), pp. 220–228. <https://doi.org/10.1002/ajmg.c.30258>
- Hoffmann, A. A., & Rieseberg, L. H. (2008). Revisiting the Impact of Inversions in Evolution: From Population Genetic Markers to Drivers of Adaptive Shifts and Speciation? *Annual review of ecology, evolution, and systematics*, 39, pp. 21–42. <https://doi.org/10.1146/annurev.ecolsys.39.110707.173532>
- Hollox, E. J., Zuccherato, L. W., & Tucci, S. (2022). Genome structural variation in human evolution. *Trends in Genetics*, 38(1), pp. 45–58. <https://doi.org/10.1016/j.tig.2021.06.015>
- Holmes, M. J., Mahjour, B., Castro, C. P., Farnum, G. A., Diehl, A. G., & Boyle, A. P. (2023, January 20). *LRphase: An efficient method for assigning haplotype identity to long reads*. <https://doi.org/10.1101/2023.01.18.524565>
- Holsinger, K. E., & Weir, B. S. (2009). Genetics in geographically structured populations: Defining, estimating and interpreting F_{ST} . *Nature Reviews Genetics*, 10(9), pp. 639–650. <https://doi.org/10.1038/nrg2611>
- Howie, B. N., Donnelly, P., & Marchini, J. (2009). A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLOS Genetics*, 5(6), pp. e1000529. <https://doi.org/10.1371/journal.pgen.1000529>
- Hsieh, P., Vollger, M. R., Dang, V., Porubsky, D., Baker, C., Cantsilieris, S., Hoekzema, K., Lewis, A. P., Munson, K. M., Sorensen, M., Kronenberg, Z. N., Murali, S., Nelson, B. J., Chiatante, G., Maggiolini, F. A. M., Blanché, H., Underwood, J. G., Antonacci, F., Deleuze, J.-F., & Eichler, E. E. (2019). Adaptive archaic introgression of copy number variants and the discovery of previously unknown human genes. *Science*, 366(6463), pp. eaax2083. <https://doi.org/10.1126/science.aax2083>

- Hsu, L. Y., Benn, P. A., Tannenbaum, H. L., Perlis, T. E., & Carlson, A. D. (1987). Chromosomal polymorphisms of 1, 9, 16, and Y in 4 major ethnic groups: A large prenatal study. *American Journal of Medical Genetics*, 26(1), pp. 95–101. <https://doi.org/10.1002/ajmg.1320260116>
- Huang, K., & Rieseberg, L. H. (2020). Frequency, Origins, and Evolutionary Role of Chromosomal Inversions in Plants. *Frontiers in Plant Science*, 11. <https://doi.org/10.3389/fpls.2020.00296>
- Hung, T.-K., Liu, W.-C., Lai, S.-K., Chuang, H.-W., Lee, Y.-C., Lin, H.-Y., Hsu, C.-L., Chen, C.-Y., Yang, Y.-C., Hsu, J. S., & Chen, P.-L. (2024). Genetic complexity of killer-cell immunoglobulin-like receptor genes in human pangenome assemblies. *Genome Research*, 34(8), pp. 1211–1223. <https://doi.org/10.1101/gr.278358.123>
- International HapMap Consortium. (2005). A haplotype map of the human genome. *Nature*, 437(7063), pp. 1299–1320. <https://doi.org/10.1038/nature04226>
- International Human Genome Sequencing Consortium. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011), pp. 931–945. <https://doi.org/10.1038/nature03001>
- Iqbal, M. A., Broeckel, U., Levy, B., Skinner, S., Sahajpal, N. S., Rodriguez, V., Stence, A., Awayda, K., Scharer, G., Skinner, C., Stevenson, R., Bossler, A., Nagy, P. L., & Kolhe, R. (2023). Multisite Assessment of Optical Genome Mapping for Analysis of Structural Variants in Constitutional Postnatal Cases. *The Journal of Molecular Diagnostics*, 25(3), pp. 175–188. <https://doi.org/10.1016/j.jmoldx.2022.12.005>
- Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., Tyson, J. R., Beggs, A. D., Dilthey, A. T., Fiddes, I. T., Malla, S., Marriott, H., Nieto, T., O’Grady, J., Olsen, H. E., Pedersen, B. S., Rhie, A., Richardson, H., Quinlan, A. R., . . . Loose, M. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology*, 36(4), pp. 338–345. <https://doi.org/10.1038/nbt.4060>
- Jain, M., Olsen, H. E., Turner, D. J., Stoddart, D., Bulazel, K. V., Paten, B., Hausler, D., Willard, H. F., Akeson, M., & Miga, K. H. (2018). Linear assembly

- of a human centromere on the Y chromosome. *Nature Biotechnology*, 36(4), pp. 321–323. <https://doi.org/10.1038/nbt.4109>
- Jana, U., Rodriguez, O. L., Lees, W., Engelbrecht, E., Vanwinkle, Z., Peres, A., Gibson, W. S., Shields, K., Schultze, S., Dorgham, A., Emery, M., Deikus, G., Sebra, R., Eichler, E. E., Yaari, G., Smith, M. L., & Watson, C. T. (2025). The human IG heavy chain constant gene locus is enriched for large structural variants and coding polymorphisms that vary among human populations. *Cell Genomics*, pp. 101058. <https://doi.org/10.1016/j.xgen.2025.101058>
- Jiang, Y. (2022). Contribution of Microhomology to Genome Instability: Connection between DNA Repair and Replication Stress. *International Journal of Molecular Sciences*, 23(21), pp. 12937. <https://doi.org/10.3390/ijms232112937>
- Jobling, M. A., Williams, G., Schiebel, K., Pandya, A., McElreavey, K., Salas, L., Rappold, G. A., Affara, N. A., & Tyler-Smith, C. (1998). A selective difference between human Y-chromosomal DNA haplotypes. *Current Biology*, 8(25), pp. 1391–1394. [https://doi.org/10.1016/S0960-9822\(98\)00020-7](https://doi.org/10.1016/S0960-9822(98)00020-7)
- John, A., Vinayan, K., & Varghese, J. (2016). Achiasmy: Male Fruit Flies Are Not Ready to Mix. *Frontiers in Cell and Developmental Biology*, 4, pp. 75. <https://doi.org/10.3389/fcell.2016.00075>
- John, H. A., Birnstiel, M. L., & Jones, K. W. (1969). RNA-DNA Hybrids at the Cytological Level. *Nature*, 223(5206), pp. 582–587. <https://doi.org/10.1038/223582a0>
- Jónsson, H., Sulem, P., Arnadottir, G. A., Pálsson, G., Eggertsson, H. P., Kristmundsdottir, S., Zink, F., Kehr, B., Hjorleifsson, K. E., Jensson, B. Ö., Jonsdottir, I., Marelsson, S. E., Gudjonsson, S. A., Gylfason, A., Jonasdottir, A., Jonasdottir, A., Stacey, S. N., Magnusson, O. T., Thorsteinsdottir, U., ... Stefansson, K. (2018). Multiple transmissions of de novo mutations in families. *Nature Genetics*, 50(12), pp. 1674–1680. <https://doi.org/10.1038/s41588-018-0259-9>
- Jumaa, M. G., & Ramadhan, M. A.-K. (2021). Cancer testis antigen XAGE-1 is a promising marker for the diagnosis and treatment of ovarian cancer. *Journal*

- of Medicine and Life*, 14(5), pp. 710–715. <https://doi.org/10.25122/jml-2021-0304>
- Kallenborn, F., Chacon, A., Hundt, C., Sirelkhatim, H., Didi, K., Cha, S., Dallago, C., Mirdita, M., Schmidt, B., & Steinegger, M. (2025). GPU-accelerated homology search with MMseqs2. *Nature Methods*, 22(10), pp. 2024–2027. <https://doi.org/10.1038/s41592-025-02819-8>
- Kallioniemi, A., Kallioniemi, O. P., Sudar, D., Rutovitz, D., Gray, J. W., Waldman, F., & Pinkel, D. (1992). Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science (New York, N.Y.)*, 258(5083), pp. 818–821. <https://doi.org/10.1126/science.1359641>
- Kaneko, H., Shimura, K., Kuwahara, S., Ohshiro, M., Tsutsumi, Y., Iwai, T., Horiike, S., Yokota, S., Ohkawara, Y., & Taniwaki, M. (2014). Inversion of chromosome 7q22 and q36 as a sole abnormality presenting in myelodysplastic syndrome: A case report. *Journal of Medical Case Reports*, 8(1), pp. 268. <https://doi.org/10.1186/1752-1947-8-268>
- Kapun, M., & Flatt, T. (2019). The adaptive significance of chromosomal inversion polymorphisms in *Drosophila melanogaster*. *Molecular Ecology*, 28(6), pp. 1263–1282. <https://doi.org/10.1111/mec.14871>
- Karakurt, H. U., Pekcan, H. A., Kahraman, A., Jihad, M., Akgün, B., Oksuz, C., & Onay, B. (2023). Evaluation of Differences of Fast and High Accuracy Base Calling Models of Guppy on Variant Calling Using Low Coverage WGS Data. *International Journal of Life Sciences and Biotechnology*, 6(3), pp. 276–287. <https://doi.org/10.38001/ijlsb.1308355>
- Kidd, J. M., Cooper, G. M., Donahue, W. F., Hayden, H. S., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., Antonacci, F., Haugen, E., Zerr, T., Yamada, N. A., Tsang, P., Newman, T. L., Tüzün, E., Cheng, Z., Ebling, H. M., Tusneem, N., ... Eichler, E. E. (2008). Mapping and sequencing of structural variation from eight human genomes. *Nature*, 453(7191), pp. 56–64. <https://doi.org/10.1038/nature06862>
- Kidd, J. M., Graves, T., Newman, T. L., Fulton, R., Hayden, H. S., Malig, M., Kallicki, J., Kaul, R., Wilson, R. K., & Eichler, E. E. (2010). A Human

- Genome Structural Variation Sequencing Resource Reveals Insights into Mutational Mechanisms. *Cell*, 143(5), pp. 837–847. <https://doi.org/10.1016/j.cell.2010.10.027>
- Kim, J., Weber, J. A., Jho, S., Jang, J., Jun, J., Cho, Y. S., Kim, H.-M., Kim, H., Kim, Y., Chung, O., Kim, C. G., Lee, H., Kim, B. C., Han, K., Koh, I., Chae, K. S., Lee, S., Edwards, J. S., & Bhak, J. (2018). KoVariome: Korean National Standard Reference Variome database of whole genomes with comprehensive SNV, indel, CNV, and SV analyses. *Scientific Reports*, 8(1), pp. 5677. <https://doi.org/10.1038/s41598-018-23837-x>
- King, D. A., Jones, W. D., Crow, Y. J., Dominiczak, A. F., Foster, N. A., Gaunt, T. R., Harris, J., Hellens, S. W., Homfray, T., Innes, J., Jones, E. A., Joss, S., Kulkarni, A., Mansour, S., Morris, A. D., Parker, M. J., Porteous, D. J., Shihab, H. A., Smith, B. H., . . . Hurles, M. E. (2015). Mosaic structural variation in children with developmental disorders. *Human Molecular Genetics*, 24(10), pp. 2733–2745. <https://doi.org/10.1093/hmg/ddv033>
- King, D. A., Sifrim, A., Fitzgerald, T. W., Rahbari, R., Hobson, E., Homfray, T., Mansour, S., Mehta, S. G., Shehla, M., Tomkins, S. E., Vasudevan, P. C., Hurles, M. E., & Study, T. D. D. D. (2017). Detection of structural mosaicism from targeted and whole-genome sequencing data. *Genome Research*. <https://doi.org/10.1101/gr.212373.116>
- Kirkpatrick, M. (2010). How and Why Chromosome Inversions Evolve. *PLOS Biology*, 8(9), pp. e1000501. <https://doi.org/10.1371/journal.pbio.1000501>
- Kleinjan, D.-J., & van Heyningen, V. (1998). Position Effect in Human Genetic Disease. *Human Molecular Genetics*, 7(10), pp. 1611–1618. <https://doi.org/10.1093/hmg/7.10.1611>
- Kolmogorov, M., Yuan, J., Lin, Y., & Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology*, 37(5), pp. 540–546. <https://doi.org/10.1038/s41587-019-0072-8>
- Komata, S., Yoda, S., KonDo, Y., Shinozaki, S., Tamai, K., & Fujiwara, H. (2022). Functional unit of supergene in female-limited Batesian mimicry of *Papilio*

- polytes. *Genetics*, 223(2), pp. iyac177. <https://doi.org/10.1093/genetics/iyac177>
- Koning, A. P. J. de, Gu, W., Castoe, T. A., Batzer, M. A., & Pollock, D. D. (2011). Repetitive Elements May Comprise Over Two-Thirds of the Human Genome. *PLOS Genetics*, 7(12), pp. e1002384. <https://doi.org/10.1371/journal.pgen.1002384>
- Koolen, D. A., Sharp, A. J., Hurst, J. A., Firth, H. V., Knight, S. J. L., Goldenberg, A., Saugier-Verber, P., Pfundt, R., Vissers, L. E. L. M., Destrée, A., Grisart, B., Rooms, L., Aa, N. V. der, Field, M., Hackett, A., Bell, K., Nowaczyk, M. J. M., Mancini, G. M. S., Poddighe, P. J., ... Vries, B. B. A. de. (2008). Clinical and molecular delineation of the 17q21.31 microdeletion syndrome. *Journal of Medical Genetics*, 45(11), pp. 710–720. <https://doi.org/10.1136/jmg.2008.058701>
- Koolen, D. A., Vissers, L. E. L. M., Pfundt, R., de Leeuw, N., Knight, S. J., Regan, R., Kooy, R. F., Reyniers, E., Romano, C., Fichera, M., Schinzel, A., Baumer, A., Anderlid, B.-M., Schoumans, J., Knoers, N. V., van Kessel, A. G., Sistermans, E. A., Veltman, J. A., Brunner, H. G., & de Vries, B. B. A. (2006). A new chromosome 17q21.31 microdeletion syndrome associated with a common inversion polymorphism. *Nature Genetics*, 38(9), pp. 999–1001. <https://doi.org/10.1038/ng1853>
- Koolen, D. A., Pfundt, R., Linda, K., Beunders, G., Veenstra-Knol, H. E., Conta, J. H., Fortuna, A. M., Gillessen-Kaesbach, G., Dugan, S., Halbach, S., Abdul-Rahman, O. A., Winesett, H. M., Chung, W. K., Dalton, M., Dimova, P. S., Mattina, T., Prescott, K., Zhang, H. Z., Saal, H. M., ... de Vries, B. B. (2016). The Koolen-de Vries syndrome: A phenotypic comparison of patients with a 17q21.31 microdeletion versus a KANSL1 sequence variant. *European Journal of Human Genetics*, 24(5), pp. 652–659. <https://doi.org/10.1038/ejhg.2015.178>
- Korbel, J. O., Urban, A. E., Affourtit, J. P., Godwin, B., Grubert, F., Simons, J. F., Kim, P. M., Palejev, D., Carriero, N. J., Du, L., Taillon, B. E., Chen, Z., Tanzer, A., Saunders, A. C. E., Chi, J., Yang, F., Carter, N. P., Hurles,

- M. E., Weissman, S. M., . . . Snyder, M. (2007). Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome. *Science*, *318*(5849), pp. 420–426. <https://doi.org/10.1126/science.1149504>
- Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: Scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, *27*(5), pp. 722–736. <https://doi.org/10.1101/gr.215087.116>
- Kosuthova, K., & Solc, R. (2023). Inversions on human chromosomes. *American Journal of Medical Genetics Part A*, *191*(3), pp. 672–683. <https://doi.org/10.1002/ajmg.a.63063>
- Kosyakova, N., Grigorian, A., Liehr, T., Manvelyan, M., Simonyan, I., Mkrtchyan, H., Aroutiounian, R., Polityko, A. D., Kulpanovich, A. I., Egorova, T., Jaroshevich, E., Frolova, A., Shorokh, N., Naumchik, I. V., Volleth, M., Schreyer, I., Nelle, H., Stumm, M., Wegner, R.-D., . . . Ergul, E. (2013). Heteromorphic variants of chromosome 9. *Molecular Cytogenetics*, *6*, pp. 14. <https://doi.org/10.1186/1755-8166-6-14>
- Kraft, K., Magg, A., Heinrich, V., Riemenschneider, C., Schöpflin, R., Markowski, J., Ibrahim, D. M., Acuna-Hidalgo, R., Despang, A., Andrey, G., Wittler, L., Timmermann, B., Vingron, M., & Mundlos, S. (2019). Serial genomic inversions induce tissue-specific architectural stripes, gene misexpression and congenital malformations. *Nature Cell Biology*, *21*(3), pp. 305–310. <https://doi.org/10.1038/s41556-019-0273-x>
- Krusche, P., Trigg, L., Boutros, P. C., Mason, C. E., De La Vega, F. M., Moore, B. L., Gonzalez-Porta, M., Eberle, M. A., Tezak, Z., Lababidi, S., Truty, R., Asimenos, G., Funke, B., Fleharty, M., Chapman, B. A., Salit, M., & Zook, J. M. (2019). Best practices for benchmarking germline small-variant calls in human genomes. *Nature Biotechnology*, *37*(5), pp. 555–560. <https://doi.org/10.1038/s41587-019-0054-x>
- Kulski, J. K., Suzuki, S., & Shiina, T. (2022). Human leukocyte antigen super-locus: Nexus of genomic supergenes, SNPs, indels, transcripts, and haplotypes. *Hu-*

- man Genome Variation*, 9(1), pp. 49. <https://doi.org/10.1038/s41439-022-00226-5>
- Kurki, M. I., Karjalainen, J., Palta, P., Sipilä, T. P., Kristiansson, K., Donner, K. M., Reeve, M. P., Laivuori, H., Aavikko, M., Kaunisto, M. A., Loukola, A., Lahtela, E., Mattsson, H., Laiho, P., Della Briotta Parolo, P., Lehisto, A. A., Kanai, M., Mars, N., Rämö, J., ... Palotie, A. (2023). FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature*, 613(7944), pp. 508–518. <https://doi.org/10.1038/s41586-022-05473-8>
- Lakich, D., Kazazian, H. H., Antonarakis, S. E., & Gitschier, J. (1993). Inversions disrupting the factor VIII gene are a common cause of severe haemophilia A. *Nature Genetics*, 5(3), pp. 236–241. <https://doi.org/10.1038/ng1193-236>
- Lam, H. Y. K., Mu, X. J., Stütz, A. M., Tanzer, A., Cayting, P. D., Snyder, M., Kim, P. M., Korbel, J. O., & Gerstein, M. B. (2010). Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nature Biotechnology*, 28(1), pp. 47–55. <https://doi.org/10.1038/nbt.1600>
- Lande, R., & Wilkinson, G. S. (1999). Models of sex-ratio meiotic drive and sexual selection in stalk-eyed flies. *Genetical Research*, 74(3), pp. 245–253. <https://doi.org/10.1017/S0016672399004218>
- Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., ... International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822), pp. 860–921. <https://doi.org/10.1038/35057062>
- Langer-Safer, P. R., Levine, M., & Ward, D. (1982). Immunological method for mapping genes on *Drosophila* polytene chromosomes. *Proceedings of the National Academy of Sciences of the United States of America*, 79(14), pp. 4381–4385. <https://doi.org/10.1073/pnas.79.14.4381>
- Lappalainen, I., Lopez, J., Skipper, L., Hefferon, T., Spalding, J. D., Garner, J., Chen, C., Maguire, M., Corbett, M., Zhou, G., Paschall, J., Ananiev, V., Flicek, P., & Church, D. M. (2013). dbVar and DGVa: Public archives for

- genomic structural variation. *Nucleic Acids Research*, 41(D1), pp. D936–D941. <https://doi.org/10.1093/nar/gks1213>
- Lee, C.-R., Wang, B., Mojica, J. P., Mandáková, T., Prasad, K. V. S. K., Goicoechea, J. L., Perera, N., Hellsten, U., Hundley, H. N., Johnson, J., Grimwood, J., Barry, K., Fairclough, S., Jenkins, J. W., Yu, Y., Kudrna, D., Zhang, J., Talag, J., Golser, W., ... Mitchell-Olds, T. (2017). Young inversion with multiple linked QTLs under selection in a hybrid zone. *Nature Ecology & Evolution*, 1(5), pp. 0119. <https://doi.org/10.1038/s41559-017-0119>
- Lee, J. A., Carvalho, C. M. B., & Lupski, J. R. (2007). A DNA Replication Mechanism for Generating Nonrecurrent Rearrangements Associated with Genomic Disorders. *Cell*, 131(7), pp. 1235–1247. <https://doi.org/10.1016/j.cell.2007.11.037>
- Lejeune, J., Gautier, M., & Turpin, R. (1959). [Study of somatic chromosomes from 9 mongoloid children]. *Comptes Rendus Hebdomadaires Des Seances De l'Academie Des Sciences*, 248(11), pp. 1721–1722.
- Lerga Jaso, J. (2020). *Integrative analysis of the functional consequences of inversions in the human genome*.
- Levy, S., Sutton, G., Ng, P. C., Feuk, L., Halpern, A. L., Walenz, B. P., Axelrod, N., Huang, J., Kirkness, E. F., Denisov, G., Lin, Y., MacDonald, J. R., Pang, A. W. C., Shago, M., Stockwell, T. B., Tsiamouri, A., Bafna, V., Bansal, V., Kravitz, S. A., ... Venter, J. C. (2007). The Diploid Genome Sequence of an Individual Human. *PLOS Biology*, 5(10), pp. e254. <https://doi.org/10.1371/journal.pbio.0050254>
- Levy-Sakin, M., Pastor, S., Mostovoy, Y., Li, L., Leung, A. K. Y., McCaffrey, J., Young, E., Lam, E. T., Hastie, A. R., Wong, K. H. Y., Chung, C. Y. L., Ma, W., Sibert, J., Rajagopalan, R., Jin, N., Chow, E. Y. C., Chu, C., Poon, A., Lin, C., ... Kwok, P.-Y. (2019). Genome maps across 26 human populations reveal population-specific patterns of structural variation. *Nature Communications*, 10(1), pp. 1025. <https://doi.org/10.1038/s41467-019-08992-7>
- Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18), pp. 3094–3100. <https://doi.org/10.1093/bioinformatics/bty191>

- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), pp. 2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>
- Li, H., Bloom, J. M., Farjoun, Y., Fleharty, M., Gauthier, L., Neale, B., & MacArthur, D. (2018). A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nature Methods*, *15*(8), pp. 595–597. <https://doi.org/10.1038/s41592-018-0054-7>
- Li, L., Leung, A. K.-Y., Kwok, T.-P., Lai, Y. Y. Y., Pang, I. K., Chung, G. T.-Y., Mak, A. C. Y., Poon, A., Chu, C., Li, M., Wu, J. J. K., Lam, E. T., Cao, H., Lin, C., Sibert, J., Yiu, S.-M., Xiao, M., Lo, K.-W., Kwok, P.-Y., ... Yip, K. Y. (2017). OMSV enables accurate and comprehensive identification of large structural variations from nanochannel-based single-molecule optical maps. *Genome Biology*, *18*(1), pp. 230. <https://doi.org/10.1186/s13059-017-1356-2>
- Liao, W.-W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., Lu, S., Lucas, J. K., Monlong, J., Abel, H. J., Buonaiuto, S., Chang, X. H., Cheng, H., Chu, J., Colonna, V., Eizenga, J. M., Feng, X., Fischer, C., Fulton, R. S., ... Paten, B. (2023). A draft human pangenome reference. *Nature*, *617*(7960), pp. 312–324. <https://doi.org/10.1038/s41586-023-05896-x>
- Lieber, M. R., Gu, J., Lu, H., Shimazaki, N., & Tsai, A. G. (2010). Nonhomologous DNA End Joining (NHEJ) and Chromosomal Translocations in Humans. *Sub-cellular biochemistry*, *50*, pp. 279–296. https://doi.org/10.1007/978-90-481-3471-7_14
- Liu, B., Zhou, L., Cao, L., Liu, H., Tong, L., Zhang, Z., Shen, H., Ruan, C., Zhou, Y., & Jiang, M. (2025). Optical genome mapping identified deletions, inversions, and insertions in hemophilia. *Blood Advances*, *9*(2), pp. 360–364. <https://doi.org/10.1182/bloodadvances.2024014762>
- Lledó, J. I. L., & Cáceres, M. (2013). On the Power and the Systematic Biases of the Detection of Chromosomal Inversions by Paired-End Genome Sequencing. *PLOS ONE*, *8*(4), pp. e61292. <https://doi.org/10.1371/journal.pone.0061292>

- Lodato, M. A., & Walsh, C. A. (2019). Genome aging: Somatic mutation in the brain links age-related decline with disease and nominates pathogenic mechanisms. *Human Molecular Genetics*, *28*(R2), pp. R197–R206. <https://doi.org/10.1093/hmg/ddz191>
- Logsdon, G. A., Vollger, M. R., & Eichler, E. E. (2020). Long-read human genome sequencing and its applications. *Nature Reviews Genetics*, *21*(10), pp. 597–614. <https://doi.org/10.1038/s41576-020-0236-x>
- Logsdon, G. A., Ebert, P., Audano, P. A., Loftus, M., Porubsky, D., Ebler, J., Yilmaz, F., Hallast, P., Prodanov, T., Yoo, D., Paisie, C. A., Harvey, W. T., Zhao, X., Martino, G. V., Henglin, M., Munson, K. M., Rabbani, K., Chin, C.-S., Gu, B., ... Marschall, T. (2025). Complex genetic variation in nearly complete human genomes. *Nature*, *644*(8076), pp. 430–441. <https://doi.org/10.1038/s41586-025-09140-6>
- Lowry, D. B., & Willis, J. H. (2010). A Widespread Chromosomal Inversion Polymorphism Contributes to a Major Life-History Transition, Local Adaptation, and Reproductive Isolation. *PLOS Biology*, *8*(9), pp. e1000500. <https://doi.org/10.1371/journal.pbio.1000500>
- Lühmann, J. L., Schmidt, G., Auber, B., Bergmann, A. K., Brandau, O., Louis, A., Hentze, S., Eisfeld, K., Schlegelberger, B., Klaes, R., & Steinemann, D. (2023). Parallel deletion and duplication at 7q11.23 in a silent carrier for two reciprocal syndromic disorders. *American Journal of Medical Genetics Part A*, *191*(7), pp. 1849–1857. <https://doi.org/10.1002/ajmg.a.63215>
- Lupiáñez, D. G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J. M., Laxova, R., Santos-Simarro, F., Gilbert-Dussardier, B., Wittler, L., Borschiwer, M., Haas, S. A., Osterwalder, M., Franke, M., Timmermann, B., Hecht, J., ... Mundlos, S. (2015). Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. *Cell*, *161*(5), pp. 1012–1025. <https://doi.org/10.1016/j.cell.2015.04.004>
- Lv, L., Yu, J., & Qi, Z. (2020). Acute myeloid leukemia with inv(16)(p13.1q22) and deletion of the 5'MYH11/3'CBFB gene fusion: A report of two cases and

- literature review. *Molecular Cytogenetics*, 13, pp. 4. <https://doi.org/10.1186/s13039-020-0474-9>
- MacDonald, J. R., Ziman, R., Yuen, R. K. C., Feuk, L., & Scherer, S. W. (2014). The Database of Genomic Variants: A curated collection of structural variation in the human genome. *Nucleic Acids Research*, 42(D1), pp. D986–D992. <https://doi.org/10.1093/nar/gkt958>
- Machiela, M. J. (2019). Mosaicism, aging and cancer. *Current Opinion in Oncology*, 31(2), pp. 108–113. <https://doi.org/10.1097/CCO.0000000000000500>
- Mahmoud, M., Gobet, N., Cruz-Dávalos, D. I., Mounier, N., Dessimoz, C., & Sedlazeck, F. J. (2019). Structural variant calling: The long and the short of it. *Genome Biology*, 20(1), pp. 246. <https://doi.org/10.1186/s13059-019-1828-7>
- Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A., Skoglund, P., Lazaridis, I., Sankararaman, S., Fu, Q., Rohland, N., Renaud, G., Erlich, Y., Willems, T., Gallo, C., ... Reich, D. (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*, 538(7624), pp. 201–206. <https://doi.org/10.1038/nature18964>
- Mandáková, T., Singh, V., Krämer, U., & Lysak, M. A. (2015). Genome Structure of the Heavy Metal Hyperaccumulator *Noccaea caerulescens* and Its Stability on Metalliferous and Nonmetalliferous Soils. *Plant Physiology*, 169(1), pp. 674–689. <https://doi.org/10.1104/pp.15.00619>
- Maney, D. L., Merritt, J. R., Prichard, M. R., Horton, B. M., & Yi, S. V. (2020). Inside the supergene of the bird with four sexes. *Hormones and Behavior*, 126, pp. 104850. <https://doi.org/10.1016/j.yhbeh.2020.104850>
- Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttman, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., ... Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265), pp. 747–753. <https://doi.org/10.1038/nature08494>

- Mantere, T., Kersten, S., & Hoischen, A. (2019). Long-Read Sequencing Emerging in Medical Genetics. *Frontiers in Genetics*, 10. <https://doi.org/10.3389/fgene.2019.00426>
- Marçais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, 27(6), pp. 764–770. <https://doi.org/10.1093/bioinformatics/btr011>
- Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bembien, L. A., Berka, J., Braverman, M. S., Chen, Y.-J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., . . . Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057), pp. 376–380. <https://doi.org/10.1038/nature03959>
- Marshall, C. R., Howrigan, D. P., Merico, D., Thiruvahindrapuram, B., Wu, W., Greer, D. S., Antaki, D., Shetty, A., Holmans, P. A., Pinto, D., Gujral, M., Brandler, W. M., Malhotra, D., Wang, Z., Fajardo, K. V. F., Maile, M. S., Ripke, S., Agartz, I., Albus, M., . . . Sebat, J. (2017). Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nature Genetics*, 49(1), pp. 27–35. <https://doi.org/10.1038/ng.3725>
- Martin, M., Patterson, M., Garg, S., Fischer, S. O., Pisanti, N., Klau, G. W., Schöenhuth, A., & Marschall, T. (2016, November 14). *WhatsHap: Fast and accurate read-based phasing*. <https://doi.org/10.1101/085050>
- Martínez-Fundichely, A., Casillas, S., Egea, R., Ràmia, M., Barbadilla, A., Pantano, L., Puig, M., & Cáceres, M. (2014). InvFEST, a database integrating information of polymorphic inversions in the human genome. *Nucleic Acids Research*, 42(D1), pp. D1027–D1032. <https://doi.org/10.1093/nar/gkt1122>
- McDaniel, J. H., Patel, V., Olson, N. D., He, H.-J., He, Z., Cole, K. D., Schmitt, A., Sikkink, K., Sedlazeck, F. J., Doddapaneni, H., Jhangiani, S. N., Muzny, D. M., Gingras, M.-C., Mehta, H., Paulin, L. F., Hastie, A. R., Yu, H.-C., Weigman, V., Rojas, A., . . . Zook, J. M. (2024, October 15). *Extensive sequencing of a broadly-consented Genome in a Bottle matched tumor-normal*

- pair towards development of somatic benchmarks*. <https://doi.org/10.1101/2024.09.18.613544>
- McVean, G. A., Altshuler (Co-Chair), D. M., Durbin (Co-Chair), R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., Lehrach, H., ... University of Geneva. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, *491*(7422), pp. 56–65. <https://doi.org/10.1038/nature11632>
- McVey, M., & Lee, S. E. (2008). MMEJ repair of double-strand breaks (director's cut): Deleted sequences and alternative endings. *Trends in Genetics*, *24*(11), pp. 529–538. <https://doi.org/10.1016/j.tig.2008.08.007>
- Mefford, H. C., Cooper, G. M., Zerr, T., Smith, J. D., Baker, C., Shafer, N., Thorland, E. C., Skinner, C., Schwartz, C. E., Nickerson, D. A., & Eichler, E. E. (2009). A method for rapid, targeted CNV genotyping identifies rare variants associated with neurocognitive disease. *Genome Research*, *19*(9), pp. 1579–1585. <https://doi.org/10.1101/gr.094987.109>
- Merla, G., Brunetti-Pierri, N., Micale, L., & Fusco, C. (2010). Copy number variants at Williams–Beuren syndrome 7q11.23 region. *Human Genetics*, *128*(1), pp. 3–26. <https://doi.org/10.1007/s00439-010-0827-2>
- Miga, K. H., Koren, S., Rhie, A., Vollger, M. R., Gershman, A., Bzikadze, A., Brooks, S., Howe, E., Porubsky, D., Logsdon, G. A., Schneider, V. A., Potapova, T., Wood, J., Chow, W., Armstrong, J., Fredrickson, J., Pak, E., Tigyi, K., Kremitzki, M., ... Phillippy, A. M. (2020). Telomere-to-telomere assembly of a complete human X chromosome. *Nature*, *585*(7823), pp. 79–84. <https://doi.org/10.1038/s41586-020-2547-7>
- Mosbrugger, T. L., Dinou, A., Duke, J. L., Ferriola, D., Li, Y., Hayeck, T. J., & Monos, D. S. (2025). Haplotypic resolution of the challenging genomic regions of MHC and KIR using a combination of targeted sequencing and a novel assembly pipeline. *Nucleic Acids Research*, *53*(10), pp. gkaf441. <https://doi.org/10.1093/nar/gkaf441>

- Moustakli, E., Christopoulos, P., Potiris, A., Zikopoulos, A., Mavrogianni, D., Karampas, G., Kathopoulis, N., Anagnostaki, I., Domali, E., Tzallas, A. T., Drakakis, P., & Stavros, S. (2025). Long-Read Sequencing and Structural Variant Detection: Unlocking the Hidden Genome in Rare Genetic Disorders. *Diagnostics*, 15(14), pp. 1803. <https://doi.org/10.3390/diagnostics15141803>
- Mozdarani, H., Meybodi, A. M., & Karimi, H. (2007). Impact of pericentric inversion of Chromosome 9 [inv (9) (p11q12)] on infertility. *Indian Journal of Human Genetics*, 13(1), pp. 26–29. <https://doi.org/10.4103/0971-6866.32031>
- Nagasaki, M., Yasuda, J., Katsuoka, F., Nariai, N., Kojima, K., Kawai, Y., Yamaguchi-Kabata, Y., Yokozawa, J., Danjoh, I., Saito, S., Sato, Y., Mimori, T., Tsuda, K., Saito, R., Pan, X., Nishikawa, S., Ito, S., Kuroki, Y., Tanabe, O., . . . Yamamoto, M. (2015). Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. *Nature Communications*, 6(1), pp. 8018. <https://doi.org/10.1038/ncomms9018>
- Nanoporetech/dorado*. (2025, September 17).
- Nassir, N., Almarri, M. A., Kumail, M., Mohamed, N., Balan, B., Hanif, S., AlObathani, M., Jamalalail, B., Elsokary, H., Kondaramage, D., Shiyas, S., Kosaji, N., Satsangi, D., Abdelmotagali, M. H. S., Abou Tayoun, A., Ahmed, O. Z. S., Youssef, D. F., Suwaidi, H. A., Albanna, A., . . . Uddin, M. (2025). A draft UAE-based Arab pangenome reference. *Nature Communications*, 16(1), pp. 6747. <https://doi.org/10.1038/s41467-025-61645-w>
- NCBI. (2025). *Overview of dbVar* [Accessed September 24, 2025].
- Nevo, E. (1978). Genetic variation in natural populations: Patterns and theory. *Theoretical Population Biology*, 13(1), pp. 121–177. [https://doi.org/10.1016/0040-5809\(78\)90039-4](https://doi.org/10.1016/0040-5809(78)90039-4)
- Nomura, T., Suzuki, S., Miyauchi, T., Takeda, M., Shinkuma, S., Fujita, Y., Nishie, W., Akiyama, M., & Shimizu, H. (2018). Chromosomal inversions as a hidden disease-modifying factor for somatic recombination phenotypes. *JCI Insight*, 3(6). <https://doi.org/10.1172/jci.insight.97595>
- Nonaka, T., Takahashi, M., Nonaka, C., Enomoto, T., & Takakuwa, K. (2019). The analysis of chromosomal abnormalities in patients with recurrent pregnancy

- loss, focusing on the prognosis of patients with inversion of chromosome (9). *Reproductive Medicine and Biology*, 18(3), pp. 296–301. <https://doi.org/10.1002/rmb2.12281>
- Norman, P. J., Norberg, S. J., Guethlein, L. A., Nemat-Gorgani, N., Royce, T., Wroblewski, E. E., Dunn, T., Mann, T., Alicata, C., Hollenbach, J. A., Chang, W., Won, M. S., Gunderson, K. L., Abi-Rached, L., Ronaghi, M., & Parham, P. (2017). Sequences of 95 human MHC haplotypes reveal extreme coding variation in genes other than highly polymorphic HLA class I and II. *Genome Research*, 27(5), pp. 813–823. <https://doi.org/10.1101/gr.213538.116>
- Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S. Y., . . . Phillippy, A. M. (2022). The complete sequence of a human genome. *Science*, 376(6588), pp. 44–53. <https://doi.org/10.1126/science.abj6987>
- Okuda, H., & Kajii, E. (2002). The evolution and formation of *RH* genes. *Legal Medicine*, 4(3), pp. 139–155. [https://doi.org/10.1016/S1344-6223\(02\)00031-7](https://doi.org/10.1016/S1344-6223(02)00031-7)
- Oldenburg, J., Pezeshkpoor, B., & Pavlova, A. (2014). Historical Review on Genetic Analysis in Hemophilia A. *Seminars in Thrombosis and Hemostasis*, 40(08), pp. 895–902. <https://doi.org/10.1055/s-0034-1395161>
- Oliveira, R., Dória, S., Madureira, C., Lima, V., Almeida, C., Pinho, M. J., Ramalho, C., Matoso, E., Barros, A., Carreira, I. M., & Moura, C. P. (2013). Inv21p12q22del21q22 and intellectual disability. *Gene*, 517(1), pp. 120–124. <https://doi.org/10.1016/j.gene.2012.12.045>
- Olson, N. D., Wagner, J., Dwarshuis, N., Miga, K. H., Sedlazeck, F. J., Salit, M., & Zook, J. M. (2023). Variant calling and benchmarking in an era of complete human genome sequences. *Nature Reviews Genetics*, 24(7), pp. 464–483. <https://doi.org/10.1038/s41576-023-00590-0>
- Osborne, L. R., Li, M., Pober, B., Chitayat, D., Bodurtha, J., Mandel, A., Costa, T., Grebe, T., Cox, S., Tsui, L.-C., & Scherer, S. W. (2001). A 1.5 million–base

- pair inversion polymorphism in families with Williams-Beuren syndrome. *Nature Genetics*, 29(3), pp. 321–325. <https://doi.org/10.1038/ng753>
- Osborne, L. R., & Mervis, C. B. (2007). Rearrangements of the Williams-Beuren syndrome locus: Molecular basis and implications for speech and language development. *Expert reviews in molecular medicine*, 9(15), pp. 1–16. <https://doi.org/10.1017/S146239940700035X>
- Pagnamenta, A. T., Yu, J., Evans, J., Twiss, P., Consortium, G. E. R., Mdt, M. G., Offiah, A. C., Wafik, M., Mehta, S. G., Javaid, M. K., Smithson, S. F., & Taylor, J. C. (2023). Conclusion of diagnostic odysseys due to inversions disrupting *GLI3* and *FBN1*. *Journal of Medical Genetics*, 60(5), pp. 505–510. <https://doi.org/10.1136/jmg-2022-108753>
- Pagnamenta, A. T., Yu, J., Walker, S., Noble, A. J., Lord, J., Dutta, P., Hashim, M., Camps, C., Green, H., Devaiah, S., Nashef, L., Parr, J., Fratter, C., Ibnouf Hussein, R., Lindsay, S. J., Lalloo, F., Banos-Pinero, B., Evans, D., Mallin, L., ... Taylor, J. C. (2024). The impact of inversions across 33,924 families with rare disease from a national genome sequencing project. *American Journal of Human Genetics*, 111(6), pp. 1140–1164. <https://doi.org/10.1016/j.ajhg.2024.04.018>
- Pande, S., Dawood, M., & Grochowski, C. M. (2025). Structural Variants: Mechanisms, Mapping, and Interpretation in Human Genetics. *Genes*, 16(8), pp. 905. <https://doi.org/10.3390/genes16080905>
- Pang, A. W., MacDonald, J. R., Pinto, D., Wei, J., Rafiq, M. A., Conrad, D. F., Park, H., Hurles, M. E., Lee, C., Venter, J. C., Kirkness, E. F., Levy, S., Feuk, L., & Scherer, S. W. (2010). Towards a comprehensive structural variation map of an individual human genome. *Genome Biology*, 11(5), pp. R52. <https://doi.org/10.1186/gb-2010-11-5-r52>
- Pang, A. W. C., Migita, O., MacDonald, J. R., Feuk, L., & Scherer, S. W. (2013). Mechanisms of Formation of Structural Variation in a Fully Sequenced Human Genome. *Human Mutation*, 34(2), pp. 345–354. <https://doi.org/10.1002/humu.22240>

- Paparella, A., L'Abbate, A., Palmisano, D., Chirico, G., Porubsky, D., Catacchio, C. R., Ventura, M., Eichler, E. E., Maggiolini, F. A. M., & Antonacci, F. (2023). Structural Variation Evolution at the 15q11-q13 Disease-Associated Locus. *International Journal of Molecular Sciences*, *24*(21), pp. 15818. <https://doi.org/10.3390/ijms242115818>
- Pardue, M. L., & Gall, J. G. (1969). Molecular hybridization of radioactive DNA to the DNA of cytological preparations. *Proceedings of the National Academy of Sciences of the United States of America*, *64*(2), pp. 600–604. <https://doi.org/10.1073/pnas.64.2.600>
- Partida-Pérez, M., Domínguez, M. G., Neira, V. A., Figuera, L. E., & Rivera, H. (2012). De novo inv(17)(p11.2q21.3) in an intellectually disabled girl: Appraisal of 21 inv(17) constitutional instances. *Journal of Genetics*, *91*(2), pp. 241–244. <https://doi.org/10.1007/s12041-012-0172-4>
- Pinto, D., Pagnamenta, A. T., Klei, L., Anney, R., Merico, D., Regan, R., Conroy, J., Magalhaes, T. R., Correia, C., Abrahams, B. S., Almeida, J., Bacchelli, E., Bader, G. D., Bailey, A. J., Baird, G., Battaglia, A., Berney, T., Bolshakova, N., Bölte, S., ... Betancur, C. (2010). Functional impact of global rare copy number variation in autism spectrum disorders. *Nature*, *466*(7304), pp. 368–372. <https://doi.org/10.1038/nature09146>
- Pollard, M. O., Gurdasani, D., Mentzer, A. J., Porter, T., & Sandhu, M. S. (2018). Long reads: Their purpose and place. *Human Molecular Genetics*, *27*(R2), pp. R234–R241. <https://doi.org/10.1093/hmg/ddy177>
- Porubsky, D., Sanders, A. D., Höps, W., Hsieh, P., Sulovari, A., Li, R., Mercuri, L., Sorensen, M., Murali, S. C., Gordon, D., Cantsilieris, S., Pollen, A. A., Ventura, M., Antonacci, F., Marschall, T., Korb, J. O., & Eichler, E. E. (2020). Recurrent inversion toggling and great ape genome evolution. *Nature Genetics*, *52*(8), pp. 849–858. <https://doi.org/10.1038/s41588-020-0646-x>
- Porubsky, D., Sanders, A. D., Tautdt, A., Colomé-Tatché, M., Lansdorp, P. M., & Guryev, V. (2020). breakpointR: An R/Bioconductor package to localize strand state changes in Strand-seq data. *Bioinformatics*, *36*(4), pp. 1260–1261. <https://doi.org/10.1093/bioinformatics/btz681>

- Porubsky, D., Ebert, P., Audano, P. A., Vollger, M. R., Harvey, W. T., Marijon, P., Ebler, J., Munson, K. M., Sorensen, M., Sulovari, A., Haukness, M., Ghareghani, M., Lansdorp, P. M., Paten, B., Devine, S. E., Sanders, A. D., Lee, C., Chaisson, M. J. P., Korbel, J. O., ... Marschall, T. (2021). Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads. *Nature Biotechnology*, *39*(3), pp. 302–308. <https://doi.org/10.1038/s41587-020-0719-5>
- Porubsky, D., Höps, W., Ashraf, H., Hsieh, P., Rodriguez-Martin, B., Yilmaz, F., Ebler, J., Hallast, P., Maggiolini, F. A. M., Harvey, W. T., Henning, B., Audano, P. A., Gordon, D. S., Ebert, P., Hasenfeld, P., Benito, E., Zhu, Q., Lee, C., Antonacci, F., ... Korbel, J. O. (2022). Recurrent inversion polymorphisms in humans associate with genetic instability and genomic disorders. *Cell*, *185*(11), pp. 1986–2005.e26. <https://doi.org/10.1016/j.cell.2022.04.017>
- Porubsky, D., Harvey, W. T., Rozanski, A. N., Ebler, J., Höps, W., Ashraf, H., Hasenfeld, P., Paten, B., Sanders, A. D., Marschall, T., Korbel, J. O., Eichler, E. E., Human Pangenome Reference Consortium (HPRC), & Human Genome Structural Variation Consortium (HGSVC). (2023). Inversion polymorphism in a complete human genome assembly. *Genome Biology*, *24*(1), pp. 100. <https://doi.org/10.1186/s13059-023-02919-8>
- Porubsky, D., & Eichler, E. E. (2024). A 25-year odyssey of genomic technology advances and structural variant discovery. *Cell*, *187*(5), pp. 1024–1037. <https://doi.org/10.1016/j.cell.2024.01.002>
- Porubsky, D., Dashnow, H., Sasani, T. A., Logsdon, G. A., Hallast, P., Noyes, M. D., Kronenberg, Z. N., Mokveld, T., Koundinya, N., Nolan, C., Steely, C. J., Guarracino, A., Dolzhenko, E., Harvey, W. T., Rowell, W. J., Grigorev, K., Nicholas, T. J., Goldberg, M. E., Oshima, K. K., ... Eichler, E. E. (2025). Human de novo mutation rates from a four-generation pedigree reference. *Nature*, *643*(8071), pp. 427–436. <https://doi.org/10.1038/s41586-025-08922-2>

- Puig, M., Casillas, S., Villatoro, S., & Cáceres, M. (2015). Human inversions and their functional consequences. *Briefings in Functional Genomics*, 14(5), pp. 369–379. <https://doi.org/10.1093/bfgp/elv020>
- Puig, M., Castellano, D., Pantano, L., Giner-Delgado, C., Izquierdo, D., Gayà-Vidal, M., Lucas-Lledó, J. I., Esko, T., Terao, C., Matsuda, F., & Cáceres, M. (2015). Functional Impact and Evolution of a Novel Human Polymorphic Inversion That Disrupts a Gene and Creates a Fusion Transcript. *PLOS Genetics*, 11(10), pp. e1005495. <https://doi.org/10.1371/journal.pgen.1005495>
- Puig, M., Lerga-Jaso, J., Giner-Delgado, C., Pacheco, S., Izquierdo, D., Delprat, A., Gayà-Vidal, M., Regan, J. F., Karlin-Neumann, G., & Cáceres, M. (2020). Determining the impact of uncharacterized inversions in the human genome by droplet digital PCR. *Genome Research*, 30(5), pp. 724–735. <https://doi.org/10.1101/gr.255273.119>
- QIAGEN Digital Insights. (2025). *Bioinformatics software*.
- Quan, C., Li, Y., Liu, X., Wang, Y., Ping, J., Lu, Y., & Zhou, G. (2021). Characterization of structural variation in Tibetans reveals new evidence of high-altitude adaptation and introgression. *Genome Biology*, 22(1), pp. 159. <https://doi.org/10.1186/s13059-021-02382-3>
- Quigley, S., Damas, J., Larkin, D. M., & Farré, M. (2023). syntenyPlotteR: A user-friendly R package to visualize genome synteny, ideal for both experienced and novice bioinformaticians. *Bioinformatics Advances*, 3(1), pp. vbad161. <https://doi.org/10.1093/bioadv/vbad161>
- Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6), pp. 841–842. <https://doi.org/10.1093/bioinformatics/btq033>
- Quinlan, A. R., & Hall, I. M. (2012). Characterizing complex structural variation in germline and somatic genomes. *Trends in genetics: TIG*, 28(1), pp. 43–53. <https://doi.org/10.1016/j.tig.2011.10.002>
- R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria.

- Rajderkar, S., Barozzi, I., Zhu, Y., Hu, R., Zhang, Y., Li, B., Alcaina Caro, A., Fukuda-Yuzawa, Y., Kelman, G., Akeza, A., Blow, M. J., Pham, Q., Harrington, A. N., Godoy, J., Meky, E. M., von Maydell, K., Hunter, R. D., Akiyama, J. A., Novak, C. S., ... Pennacchio, L. A. (2023). Topologically associating domain boundaries are required for normal genome function. *Communications Biology*, 6(1), pp. 435. <https://doi.org/10.1038/s42003-023-04819-w>
- Rautiainen, M., Nurk, S., Walenz, B. P., Logsdon, G. A., Porubsky, D., Rhie, A., Eichler, E. E., Phillippy, A. M., & Koren, S. (2023). Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nature Biotechnology*, 41(10), pp. 1474–1482. <https://doi.org/10.1038/s41587-023-01662-6>
- Real, T. D., Hebbar, P., Yoo, D., Antonacci, F., Pačar, I., Diekhans, M., Mikol, G. J., Popoola, O. G., Mallory, B. J., Vollger, M. R., Dishuck, P. C., Guitart, X., Rozanski, A. N., Munson, K. M., Hoekzema, K., Ranchalis, J. E., Neph, S. J., Sedeño-Cortes, A. E., Paten, B., ... Eichler, E. E. (2025, March 17). *Genetic diversity and regulatory features of human-specific NOTCH2NL duplications*. <https://doi.org/10.1101/2025.03.14.643395>
- Rech, G. E., Sanz-Martín, J. M., Anisimova, M., Sukno, S. A., & Thon, M. R. (2014). Natural Selection on Coding and Noncoding DNA Sequences Is Associated with Virulence Genes in a Plant Pathogenic Fungus. *Genome Biology and Evolution*, 6(9), pp. 2368–2379. <https://doi.org/10.1093/gbe/evu192>
- Rhie, A., Nurk, S., Cechova, M., Hoyt, S. J., Taylor, D. J., Altemose, N., Hook, P. W., Koren, S., Rautiainen, M., Alexandrov, I. A., Allen, J., Asri, M., Bzikadze, A. V., Chen, N.-C., Chin, C.-S., Diekhans, M., Flicek, P., Formenti, G., Functammasan, A., ... Phillippy, A. M. (2023). The complete sequence of a human Y chromosome. *Nature*, 621(7978), pp. 344–354. <https://doi.org/10.1038/s41586-023-06457-y>
- Robinson, J. T., Thorvaldsdottir, H., Turner, D., & Mesirov, J. P. (2023). Igv.js: An embeddable JavaScript implementation of the Integrative Genomics Viewer (IGV). *Bioinformatics*, 39(1), pp. btac830. <https://doi.org/10.1093/bioinformatics/btac830>

- Rodríguez-Trelles, F., & Tarrío, R. (2024). Acceleration of *Drosophila subobscura* evolutionary response to global warming in Europe. *Nature Climate Change*, *14*(10), pp. 1101–1106. <https://doi.org/10.1038/s41558-024-02128-6>
- Rosendahl, J., Kirsten, H., Hegyi, E., Kovacs, P., Weiss, F. U., Laumen, H., Lichtner, P., Ruffert, C., Chen, J.-M., Masson, E., Beer, S., Zimmer, C., Seltsam, K., Algül, H., Bühler, F., Bruno, M. J., Bugert, P., Burkhardt, R., Cavestro, G. M., ... Sahin-Tóth, M. (2018). Genome-wide association study identifies inversion in the CTRB1-CTRB2 locus to modify risk for alcoholic and non-alcoholic chronic pancreatitis. *Gut*, *67*(10), pp. 1855–1863. <https://doi.org/10.1136/gutjnl-2017-314454>
- Rossetti, L. C., Radic, C. P., Larripa, I. B., & De Brasi, C. D. (2005). Genotyping the Hemophilia Inversion Hotspot by Use of Inverse PCR. *Clinical Chemistry*, *51*(7), pp. 1154–1158. <https://doi.org/10.1373/clinchem.2004.046490>
- Rubinacci, S., Delaneau, O., & Marchini, J. (2020). Genotype imputation using the Positional Burrows Wheeler Transform. *PLOS Genetics*, *16*(11), pp. e1009049. <https://doi.org/10.1371/journal.pgen.1009049>
- Ruiz-Arenas, C., Cáceres, A., López-Sánchez, M., Tolosana, I., Pérez-Jurado, L., & González, J. R. (2019). scoreInvHap: Inversion genotyping for genome-wide association studies. *PLOS Genetics*, *15*(7), pp. e1008203. <https://doi.org/10.1371/journal.pgen.1008203>
- Saiki, R. K., Scharf, S., Faloona, F., Mullis, K. B., Horn, G. T., Erlich, H. A., & Arnheim, N. (1985). Enzymatic Amplification of β -Globin Genomic Sequences and Restriction Site Analysis for Diagnosis of Sickle Cell Anemia. *Science*, *230*(4732), pp. 1350–1354. <https://doi.org/10.1126/science.2999980>
- Salm, M. P. A., Horswell, S. D., Hutchison, C. E., Speedy, H. E., Yang, X., Liang, L., Schadt, E. E., Cookson, W. O., Wierzbicki, A. S., Naoumova, R. P., & Shoulders, C. C. (2012). The origin, global distribution, and functional impact of the human 8p23 inversion polymorphism. *Genome Research*, *22*(6), pp. 1144–1153. <https://doi.org/10.1101/gr.126037.111>

- Samuel, G. N., & Farsides, B. (2017). The UK's 100,000 Genomes Project: Manifesting policymakers' expectations. *New Genetics and Society*, 36(4), pp. 336–353. <https://doi.org/10.1080/14636778.2017.1370671>
- Sanders, A. D., Hills, M., Porubský, D., Guryev, V., Falconer, E., & Lansdorp, P. M. (2016). Characterizing polymorphic inversions in human genomes by single-cell sequencing. *Genome Research*, 26(11), pp. 1575–1587. <https://doi.org/10.1101/gr.201160.115>
- Sanders, A. D., Falconer, E., Hills, M., Spierings, D. C. J., & Lansdorp, P. M. (2017). Single-cell template strand sequencing by Strand-seq enables the characterization of individual homologs. *Nature Protocols*, 12(6), pp. 1151–1176. <https://doi.org/10.1038/nprot.2017.029>
- Schloissnig, S., Pani, S., Ebler, J., Hain, C., Tsapalou, V., Söylev, A., Hüther, P., Ashraf, H., Prodanov, T., Asparuhova, M., Magalhães, H., Höps, W., Sotelo-Fonseca, J. E., Fitzgerald, T., Santana-Garcia, W., Moreira-Pinhal, R., Hunt, S., Pérez-Llanos, F. J., Wollenweber, T. E., ... Korb, J. O. (2025). Structural variation in 1,019 diverse humans based on long-read sequencing. *Nature*, pp. 1–11. <https://doi.org/10.1038/s41586-025-09290-7>
- Schneider, V. A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H.-C., Kitts, P. A., Murphy, T. D., Pruitt, K. D., Thibaud-Nissen, F., Albracht, D., Fulton, R. S., Kremitzki, M., Magrini, V., Markovic, C., McGrath, S., Steinberg, K. M., Auger, K., Chow, W., Collins, J., ... Church, D. M. (2017). Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Research*, 27(5), pp. 849–864. <https://doi.org/10.1101/gr.213611.116>
- Schouten, J. P., McElgunn, C. J., Waaijer, R., Zwiijnenburg, D., Diepvens, F., & Pals, G. (2002). Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic Acids Research*, 30(12), pp. e57. <https://doi.org/10.1093/nar/gnf056>
- Schuy, J., Grochowski, C. M., Carvalho, C. M., & Lindstrand, A. (2022). Complex genomic rearrangements: An underestimated cause of rare diseases. *Trends*

- in genetics : TIG*, 38(11), pp. 1134–1146. <https://doi.org/10.1016/j.tig.2022.06.003>
- Schwander, T., Libbrecht, R., & Keller, L. (2014). Supergenes and Complex Phenotypes. *Current Biology*, 24(7), pp. R288–R294. <https://doi.org/10.1016/j.cub.2014.01.056>
- Schwartz, D. C., Li, X., Hernandez, L. I., Ramnarain, S. P., Huff, E. J., & Wang, Y.-K. (1993). Ordered Restriction Maps of *Saccharomyces cerevisiae* Chromosomes Constructed by Optical Mapping. *Science*, 262(5130), pp. 110–114. <https://doi.org/10.1126/science.8211116>
- Scigliuolo, G. M., Boukouaci, W., Cappelli, B., Volt, F., Rivera Franco, M. M., Dhédin, N., de Latour, R. P., Devalck, C., Dalle, J.-H., Castelle, M., Hermine, O., Chardin, M. O., Poiré, X., Brichard, B., Paillard, C., Rafii, H., Kenzey, C., Wu, C.-L., Bouassida, J., . . . Tamouza, R. (2023). HLA haplotype frequencies and diversity in patients with hemoglobinopathies. *EJHaem*, 4(4), pp. 963–969. <https://doi.org/10.1002/jha2.763>
- Seltsam, K., Pentner, C., Weigl, F., Sutedjo, S., Zimmer, C., Beer, S., Bugert, P., Ewers, M., Ruffert, C., Michl, P., Laumen, H., Witt, H., & Rosendahl, J. (2020). Sequencing of the complex *CTRB1-CTRB2* locus in chronic pancreatitis. *Pancreatology*, 20(8), pp. 1598–1603. <https://doi.org/10.1016/j.pan.2020.09.017>
- Seol, J.-H., Shim, E. Y., & Lee, S. E. (2018). Microhomology-mediated end joining: Good, bad and ugly. *Mutation research*, 809, pp. 81–87. <https://doi.org/10.1016/j.mrfmmm.2017.07.002>
- Sereika, M., Kirkegaard, R. H., Karst, S. M., Michaelsen, T. Y., Sørensen, E. A., Wollenberg, R. D., & Albertsen, M. (2022). Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. *Nature Methods*, 19(7), pp. 823–826. <https://doi.org/10.1038/s41592-022-01539-7>

- Shagin, D. A., Lukyanov, K. A., Vagner, L. L., & Matz, M. V. (1999). Regulation of average length of complex PCR product. *Nucleic Acids Research*, 27(18), pp. e23–i. <https://doi.org/10.1093/nar/27.18.e23-i>
- Shakoori, A. R. (2017). Fluorescence In Situ Hybridization (FISH) and Its Applications. *Chromosome Structure and Aberrations*, pp. 343–367. https://doi.org/10.1007/978-81-322-3673-3_16
- Shao, H., Ganesamoorthy, D., Duarte, T., Cao, M. D., Hoggart, C. J., & Coin, L. J. M. (2018). npInv: Accurate detection and genotyping of inversions using long read sub-alignment. *BMC Bioinformatics*, 19, pp. 261. <https://doi.org/10.1186/s12859-018-2252-9>
- Sharp, A. J., Locke, D. P., McGrath, S. D., Cheng, Z., Bailey, J. A., Vallente, R. U., Pertz, L. M., Clark, R. A., Schwartz, S., Segreaves, R., Oseroff, V. V., Albertson, D. G., Pinkel, D., & Eichler, E. E. (2005). Segmental duplications and copy-number variation in the human genome. *American Journal of Human Genetics*, 77(1), pp. 78–88. <https://doi.org/10.1086/431652>
- Sharp, A. J., Cheng, Z., & Eichler, E. E. (2006). Structural Variation of the Human Genome. *Annual Review of Genomics and Human Genetics*, 7, pp. 407–442. <https://doi.org/10.1146/annurev.genom.7.080505.115618>
- Sharp, P. M., Averof, M., Lloyd, A. T., Matassi, G., & Peden, J. F. (1995). DNA sequence evolution: The sounds of silence. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 349(1329), pp. 241–247. <https://doi.org/10.1098/rstb.1995.0108>
- Sheen, C. R., McDonald, M. A., George, P. M., Smith, M. P., & Morris, C. M. (2011). Genotyping the factor VIII intron 22 inversion locus using fluorescent in situ hybridization. *Blood Cells, Molecules, and Diseases*, 46(2), pp. 151–158. <https://doi.org/10.1016/j.bcmed.2010.12.002>
- Sindi, S. S., & Raphael, B. J. (2010). Identification and Frequency Estimation of Inversion Polymorphisms from Haplotype Data. *Journal of Computational Biology*, 17(3), pp. 517–531. <https://doi.org/10.1089/cmb.2009.0185>

- Singh, V., Pandey, S., & Bhardwaj, A. (2022). From the reference human genome to human pangenome: Premise, promise and challenge. *Frontiers in Genetics*, 13. <https://doi.org/10.3389/fgene.2022.1042550>
- Small, K., Iber, J., & Warren, S. T. (1997). Emerin deletion reveals a common X-chromosome inversion mediated by inverted repeats. *Nature Genetics*, 16(1), pp. 96–99. <https://doi.org/10.1038/ng0597-96>
- Soda, M., Choi, Y. L., Enomoto, M., Takada, S., Yamashita, Y., Ishikawa, S., Fujiiwara, S.-i., Watanabe, H., Kurashina, K., Hatanaka, H., Bando, M., Ohno, S., Ishikawa, Y., Aburatani, H., Niki, T., Sohara, Y., Sugiyama, Y., & Mano, H. (2007). Identification of the transforming EML4–ALK fusion gene in non-small-cell lung cancer. *Nature*, 448(7153), pp. 561–566. <https://doi.org/10.1038/nature05945>
- Sotoudeh, A., Rostami, P., Nakhaeimoghadam, M., Mohsenipour, R., & Rezaei, N. (2017). Pericentric Inversion of Chromosome 9 in an Infant With Ambiguous Genitalia. *Acta Medica Iranica*, pp. 655–657.
- Spinner, N. B., & Conlin, L. K. (2014). Mosaicism and clinical genetics. *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*, 166(4), pp. 397–405. <https://doi.org/10.1002/ajmg.c.31421>
- Spracklen, C. N., Horikoshi, M., Kim, Y. J., Lin, K., Bragg, F., Moon, S., Suzuki, K., Tam, C. H. T., Tabara, Y., Kwak, S.-H., Takeuchi, F., Long, J., Lim, V. J. Y., Chai, J.-F., Chen, C.-H., Nakatochi, M., Yao, J., Choi, H. S., Iyengar, A. K., ... Sim, X. (2020). Identification of type 2 diabetes loci in 433,540 East Asian individuals. *Nature*, 582(7811), pp. 240–245. <https://doi.org/10.1038/s41586-020-2263-3>
- Srivastava, K., Stiles, D. A., Wagner, F. F., & Flegel, W. A. (2018). Two large deletions extending beyond either end of the RHD gene and their red cell phenotypes. *Journal of human genetics*, 63(1), pp. 27–35. <https://doi.org/10.1038/s10038-017-0345-3>
- Startek, M., Szafranski, P., Gambin, T., Campbell, I. M., Hixson, P., Shaw, C. A., Stankiewicz, P., & Gambin, A. (2015). Genome-wide analyses of LINE–LINE-

- mediated nonallelic homologous recombination. *Nucleic Acids Research*, 43(4), pp. 2188–2198. <https://doi.org/10.1093/nar/gku1394>
- Stefansson, H., Helgason, A., Thorleifsson, G., Steinthorsdottir, V., Masson, G., Barnard, J., Baker, A., Jonasdottir, A., Ingason, A., Gudnadottir, V. G., Desnica, N., Hicks, A., Gylfason, A., Gudbjartsson, D. F., Jonsdottir, G. M., Sainz, J., Agnarsson, K., Birgisdottir, B., Ghosh, S., ... Stefansson, K. (2005). A common inversion under selection in Europeans. *Nature Genetics*, 37(2), pp. 129–137. <https://doi.org/10.1038/ng1508>
- Steinberg, K. M., Antonacci, F., Sudmant, P. H., Kidd, J. M., Campbell, C. D., Vives, L., Malig, M., Scheinfeldt, L., Beggs, W., Ibrahim, M., Lema, G., Nyambo, T. B., Omar, S. A., Bodo, J.-M., Froment, A., Donnelly, M. P., Kidd, K. K., Tishkoff, S. A., & Eichler, E. E. (2012). Structural diversity and African origin of the 17q21.31 inversion polymorphism. *Nature Genetics*, 44(8), pp. 872–880. <https://doi.org/10.1038/ng.2335>
- Sturtevant, A. H. (1913). The linear arrangement of six sex-linked factors in *Drosophila*, as shown by their mode of association. *Journal of Experimental Zoology*, 14(1), pp. 43–59. <https://doi.org/10.1002/jez.1400140104>
- Sturtevant, A. H. (1917). Genetic Factors Affecting the Strength of Linkage in *Drosophila*. *Proceedings of the National Academy of Sciences*, 3(9), pp. 555–558. <https://doi.org/10.1073/pnas.3.9.555>
- Sturtevant, A. H. (1921a). A Case of Rearrangement of Genes in *Drosophila*1. *Proceedings of the National Academy of Sciences*, 7(8), pp. 235–237. <https://doi.org/10.1073/pnas.7.8.235>
- Sturtevant, A. H. (1921b). Genetic studies on *drosophila simulans*. ii. sex-linked group of genes. *Genetics*, 6(1), pp. 43–64. <https://doi.org/10.1093/genetics/6.1.43>
- Sturtevant, A. H., & Mather, K. (1938). The Interrelations of Inversions, Heterosis and Recombination. *The American Naturalist*, 72(742), pp. 447–452. <https://doi.org/10.1086/280797>
- Su, J., Zheng, Z., Ahmed, S. S., Lam, T.-W., & Luo, R. (2022). Clair3-trio: High-performance Nanopore long-read variant calling in family trios with trio-to-

- trio deep neural networks. *Briefings in Bioinformatics*, 23(5), pp. bbac301. <https://doi.org/10.1093/bib/bbac301>
- Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Hsi-Yang Fritz, M., Konkel, M. K., Malhotra, A., Stütz, A. M., Shi, X., Paolo Casale, F., Chen, J., Hormozdiari, F., Dayama, G., Chen, K., ... Korb, J. O. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, 526(7571), pp. 75–81. <https://doi.org/10.1038/nature15394>
- Tafazoli, A., Hemmati, M., Rafigh, M., Alimardani, M., Khaghani, F., Korostyński, M., & Karnes, J. H. (2025). Leveraging long-read sequencing technologies for pharmacogenomic testing: Applications, analytical strategies, challenges, and future perspectives. *Frontiers in Genetics*, 16, pp. 1435416. <https://doi.org/10.3389/fgene.2025.1435416>
- Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres, R., Taliun, S. A. G., Corvelo, A., Gogarten, S. M., Kang, H. M., Pitsillides, A. N., LeFaive, J., Lee, S.-b., Tian, X., Browning, B. L., Das, S., Emde, A.-K., Clarke, W. E., Loesch, D. P., ... Abecasis, G. R. (2021). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*, 590(7845), pp. 290–299. <https://doi.org/10.1038/s41586-021-03205-y>
- Talseth-Palmer, B. A., & Scott, R. J. (2011). Genetic Variation and its Role in Malignancy. *International Journal of Biomedical Science : IJBS*, 7(3), pp. 158–171.
- Taylor, C. M., Smith, R., Lehman, C., Mitchel, M. W., Singer, K., Weaver, W. C., & Chung, W. (1993). 16p11.2 Recurrent Deletion. In M. P. Adam, S. Bick, G. M. Mirzaa, R. A. Pagon, S. E. Wallace, & A. Amemiya (Eds.), *GeneReviews*. University of Washington, Seattle.
- The International HapMap Consortium. (2005). A haplotype map of the human genome. *Nature*, 437(7063), pp. 1299–1320. <https://doi.org/10.1038/nature04226>

- Thomas, N. S., Bryant, V., Maloney, V., Cockwell, A. E., & Jacobs, P. A. (2008). Investigation of the origins of human autosomal inversions. *Human Genetics*, 123(6), pp. 607–616. <https://doi.org/10.1007/s00439-008-0510-z>
- Thompson, M. J., & Jiggins, C. D. (2014). Supergenes and their role in evolution. *Heredity*, 113(1), pp. 1–8. <https://doi.org/10.1038/hdy.2014.20>
- Thorpe, J., Osei-Owusu, I. A., Avigdor, B. E., Tupler, R., & Pevsner, J. (2020). Mosaicism in Human Health and Disease. *Annual Review of Genetics*, 54, pp. 487–510. <https://doi.org/10.1146/annurev-genet-041720-093403>
- Tosca, L., Feraud, O., Magniez, A., Bas, C., Griscelli, F., Bennaceur-Griscelli, A., & Tachdjian, G. (2015). Genomic instability of human embryonic stem cell lines using different passaging culture methods. *Molecular Cytogenetics*, 8(1), pp. 30. <https://doi.org/10.1186/s13039-015-0133-8>
- Truty, R., Rojahn, S., Ouyang, K., Kautzer, C., Kennemer, M., Pineda-Alvarez, D., Johnson, B., Stafford, A., Basel-Salmon, L., Saitta, S., Slavotinek, A., Chandrasekharappa, S. C., Suarez, C. J., Burnett, L., Nussbaum, R. L., & Aradhya, S. (2023). Patterns of mosaicism for sequence and copy-number variants discovered through clinical deep sequencing of disease-related genes in one million individuals. *The American Journal of Human Genetics*, 110(4), pp. 551–564. <https://doi.org/10.1016/j.ajhg.2023.02.013>
- Turner, D. J., Shendure, J., Porreca, G., Church, G., Green, P., Tyler-Smith, C., & Hurles, M. E. (2006). Assaying chromosomal inversions by single-molecule haplotyping. *Nature Methods*, 3(6), pp. 439–445. <https://doi.org/10.1038/nmeth881>
- Turner, D. J., Tyler-Smith, C., & Hurles, M. E. (2008). Long-range, high-throughput haplotype determination via haplotype-fusion PCR and ligation haplotyping. *Nucleic Acids Research*, 36(13), pp. e82. <https://doi.org/10.1093/nar/gkn373>
- Tuttle, E. M., Bergland, A. O., Korody, M. L., Brewer, M. S., Newhouse, D. J., Minx, P., Stager, M., Betuel, A., Cheviron, Z. A., Warren, W. C., Gonser, R. A., & Balakrishnan, C. N. (2016). Divergence and Functional Degradation of a Sex Chromosome-like Supergene. *Current Biology*, 26(3), pp. 344–350. <https://doi.org/10.1016/j.cub.2015.11.069>

- Tuzun, E., Sharp, A. J., Bailey, J. A., Kaul, R., Morrison, V. A., Pertz, L. M., Haugen, E., Hayden, H., Albertson, D., Pinkel, D., Olson, M. V., & Eichler, E. E. (2005). Fine-scale structural variation of the human genome. *Nature Genetics*, 37(7), pp. 727–732. <https://doi.org/10.1038/ng1562>
- Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T., & Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews Methods Primers*, 1(1), pp. 59. <https://doi.org/10.1038/s43586-021-00056-9>
- UK10K Consortium, Walter, K., Min, J. L., Huang, J., Crooks, L., Memari, Y., McCarthy, S., Perry, J. R. B., Xu, C., Futema, M., Lawson, D., Iotchkova, V., Schiffels, S., Hendricks, A. E., Danecek, P., Li, R., Floyd, J., Wain, L. V., Barroso, I., . . . Soranzo, N. (2015). The UK10K project identifies rare variants in health and disease. *Nature*, 526(7571), pp. 82–90. <https://doi.org/10.1038/nature14962>
- Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amaratides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., . . . Zhu, X. (2001). The sequence of the human genome. *Science (New York, N.Y.)*, 291(5507), pp. 1304–1351. <https://doi.org/10.1126/science.1058040>
- Ventura, M., Antonacci, F., Cardone, M. F., Stanyon, R., D’Addabbo, P., Cellamare, A., Sprague, L. J., Eichler, E. E., Archidiacono, N., & Rocchi, M. (2007). Evolutionary Formation of New Centromeres in Macaque. *Science*, 316(5822), pp. 243–246. <https://doi.org/10.1126/science.1140615>
- Verbiest, M., Maksimov, M., Jin, Y., Anisimova, M., Gymrek, M., & Bilgin Sonay, T. (2023). Mutation and selection processes regulating short tandem repeats give rise to genetic and phenotypic diversity across species. *Journal of Evolutionary Biology*, 36(2), pp. 321–336. <https://doi.org/10.1111/jeb.14106>
- Vermeulen, S. J., Speleman, F., Vanransbeeck, L., Verspeet, J., Menten, B., Verschraegen-Spae, M.-R., Wilde, P. D., Messiaen, L., Michaelis, R. C., & Leroy, J. G. (2005). Familial pericentric inversion of chromosome 18: Behavioral abnor-

- malities in patients heterozygous for either the dup(18p)/del(18q) or dup(18q)/del(18p) recombinant chromosome. *European Journal of Human Genetics*, 13(1), pp. 52–58. <https://doi.org/10.1038/sj.ejhg.5201281>
- Vicente-Salvador, D., Puig, M., Gayà-Vidal, M., Pacheco, S., Giner-Delgado, C., Noguera, I., Izquierdo, D., Martínez-Fundichely, A., Ruiz-Herrera, A., Estivill, X., Aguado, C., Lucas-Lledó, J. I., & Cáceres, M. (2017). Detailed analysis of inversions predicted between two human genomes: Errors, real polymorphisms, and their origin and population distribution. *Human Molecular Genetics*, 26(3), pp. 567–581. <https://doi.org/10.1093/hmg/ddw415>
- Villoutreix, R., Ayala, D., Joron, M., Gompert, Z., Feder, J. L., & Nosil, P. (2021). Inversion breakpoints and the evolution of supergenes. *Molecular Ecology*, 30(12), pp. 2738–2755. <https://doi.org/10.1111/mec.15907>
- Visser, R., Shimokawa, O., Harada, N., Kinoshita, A., Ohta, T., Niikawa, N., & Matsumoto, N. (2005). Identification of a 3.0-kb Major Recombination Hotspot in Patients with Sotos Syndrome Who Carry a Common 1.9-Mb Microdeletion. *The American Journal of Human Genetics*, 76(1), pp. 52–67. <https://doi.org/10.1086/426950>
- Vogel, M. J., Pagie, L., Talhout, W., Nieuwland, M., Kerkhoven, R. M., & van Steensel, B. (2009). High-resolution mapping of heterochromatin redistribution in a *Drosophila* position-effect variegation model. *Epigenetics & Chromatin*, 2(1), pp. 1. <https://doi.org/10.1186/1756-8935-2-1>
- Vollger, M. R., Dishuck, P. C., Sorensen, M., Welch, A. E., Dang, V., Dougherty, M. L., Graves-Lindsay, T. A., Wilson, R. K., Chaisson, M. J. P., & Eichler, E. E. (2019). Long-read sequence and assembly of segmental duplications. *Nature Methods*, 16(1), pp. 88–94. <https://doi.org/10.1038/s41592-018-0236-3>
- Vos, P., Hogers, R., Bleeker, M., Reijans, M., Lee, T. van de, Hornes, M., Friters, A., Pot, J., Paleman, J., Kuiper, M., & Zabeau, M. (1995). AFLP: A new technique for DNA fingerprinting. *Nucleic Acids Research*, 23(21), pp. 4407–4414. <https://doi.org/10.1093/nar/23.21.4407>

- Wan, T., & Ma, E. (2012). Molecular cytogenetics : An indispensable tool for cancer diagnosis. *Biomedical Journal*, 35(2), pp. 96. <https://doi.org/10.4103/2319-4170.106161>
- Watson, J. D., & Crick, F. H. (1953). The structure of DNA. *Cold Spring Harbor Symposia on Quantitative Biology*, 18, pp. 123–131. <https://doi.org/10.1101/sqb.1953.018.01.020>
- Weber, T., Cosenza, M. R., & Korbel, J. (2023). MosaiCatcher v2: A single-cell structural variations detection and analysis reference framework based on Strand-seq. *Bioinformatics*, 39(11), pp. btad633. <https://doi.org/10.1093/bioinformatics/btad633>
- Weischenfeldt, J., Symmons, O., Spitz, F., & Korbel, J. O. (2013). Phenotypic impact of genomic structural variation: Insights from and for human disease. *Nature Reviews Genetics*, 14(2), pp. 125–138. <https://doi.org/10.1038/nrg3373>
- Wellenreuther, M., Rosenquist, H., Jaksons, P., & Larson, K. W. (2017). Local adaptation along an environmental cline in a species with an inversion polymorphism. *Journal of Evolutionary Biology*, 30(6), pp. 1068–1077. <https://doi.org/10.1111/jeb.13064>
- Wellenreuther, M., & Bernatchez, L. (2018). Eco-Evolutionary Genomics of Chromosomal Inversions. *Trends in Ecology & Evolution*, 33(6), pp. 427–440. <https://doi.org/10.1016/j.tree.2018.04.002>
- Wellenreuther, M., Mérot, C., Berdan, E., & Bernatchez, L. (2019). Going beyond SNPs: The role of structural genomic variants in adaptive evolution and species diversification. *Molecular Ecology*, 28(6), pp. 1203–1209. <https://doi.org/10.1111/mec.15066>
- Wheeler, M. M., Lannert, K. W., Huston, H., Fletcher, S. N., Harris, S., Teramura, G., Maki, H. J., Frazar, C., Underwood, J. G., Shaffer, T., Correa, A., Delaney, M., Reiner, A. P., Wilson, J. G., Nickerson, D. A., & Johnsen, J. M. (2019). Genomic characterization of the RH locus detects complex and novel structural variation in multi-ethnic cohorts. *Genetics in medicine : official journal of the American College of Medical Genetics*, 21(2), pp. 477–486. <https://doi.org/10.1038/s41436-018-0074-9>

- Wong, L.-P., Ong, R. T.-H., Poh, W.-T., Liu, X., Chen, P., Li, R., Lam, K. K.-Y., Pillai, N. E., Sim, K.-S., Xu, H., Sim, N.-L., Teo, S.-M., Foo, J.-N., Tan, L. W.-L., Lim, Y., Koo, S.-H., Gan, L. S.-H., Cheng, C.-Y., Wee, S., ... Teo, Y.-Y. (2013). Deep Whole-Genome Sequencing of 100 Southeast Asian Malays. *The American Journal of Human Genetics*, 92(1), pp. 52–66. <https://doi.org/10.1016/j.ajhg.2012.12.005>
- Wu, Y., Tian, L., Pirastu, M., Stambolian, D., & Li, H. (2013). MATCHCLIP: Locate precise breakpoints for copy number variation using CIGAR string by matching soft clipped reads. *Frontiers in Genetics*, 4. <https://doi.org/10.3389/fgene.2013.00157>
- Yahaya, T. O., Oladele, E. O., Anyebe, D., Obi, C., Bunza, M. D. A., Sulaiman, R., & Liman, U. U. (2021). Chromosomal abnormalities predisposing to infertility, testing, and management: A narrative review. *Bulletin of the National Research Centre*, 45(1), pp. 65. <https://doi.org/10.1186/s42269-021-00523-z>
- Yamada, K. (1992). Population studies of inv(9) chromosomes in 4,300 japanese: Incidence, sex difference and clinical significance. *Japanese journal of human genetics*, 37(4), pp. 293–301. <https://doi.org/10.1007/BF01883320>
- Yang, L., Lin, Y., Li, P., Wang, K., Li, J., Liu, Y., Bo, X., Ni, M., Li, P., & Song, H. (2025). A comprehensive benchmarking of adaptive sampling tools for nanopore sequencing. *Genome Biology*, 26(1), pp. 281. <https://doi.org/10.1186/s13059-025-03729-w>
- Yilmaz, F., Karageorgiou, C., Kim, K., Pajic, P., Scheer, K., Consortium, H. G. S. V., Beck, C. R., Torregrossa, A.-M., Lee, C., Gokcumen, O., Audano, P. A., Austine-Orimoloye, O., Beck, C. R., Eichler, E. E., Hallast, P., Harvey, W. T., Hastie, A. R., Hoekzema, K., Hunt, S., ... Yilmaz, F. (2024). Reconstruction of the human amylase locus reveals ancient duplications seeding modern-day variation. *Science*. <https://doi.org/10.1126/science.adn0609>
- Young, A. I., Wauthier, F., & Donnelly, P. (2016). Multiple novel gene-by-environment interactions modify the effect of FTO variants on body mass index. *Nature Communications*, 7(1), pp. 12724. <https://doi.org/10.1038/ncomms12724>

- Young, J. M., Endicott, R. M., Parghi, S. S., Walker, M., Kidd, J. M., & Trask, B. J. (2008). Extensive Copy-Number Variation of the Human Olfactory Receptor Gene Family. *American Journal of Human Genetics*, 83(2), pp. 228–242. <https://doi.org/10.1016/j.ajhg.2008.07.005>
- Yuan, Y., Chung, C. Y.-L., & Chan, T.-F. (2020). Advances in optical mapping for genomic research. *Computational and Structural Biotechnology Journal*, 18, pp. 2051–2062. <https://doi.org/10.1016/j.csbj.2020.07.018>
- Yunis, J. J. (1976). High Resolution of Human Chromosomes. *Science*, 191(4233), pp. 1268–1270. <https://doi.org/10.1126/science.1257746>
- Zhang, F., Carvalho, C. M. B., & Lupski, J. R. (2009). Complex human chromosomal and genomic rearrangements. *Trends in Genetics*, 25(7), pp. 298–307. <https://doi.org/10.1016/j.tig.2009.05.005>
- Zhang, T., Li, H., Jiang, M., Hou, H., Gao, Y., Li, Y., Wang, F., Wang, J., Peng, K., & Liu, Y.-X. (2024). Nanopore sequencing: Flourishing in its teenage years. *Journal of Genetics and Genomics*, 51(12), pp. 1361–1374. <https://doi.org/10.1016/j.jgg.2024.09.007>
- Zheng, Z., Yu, X., Chen, L., Lee, Y.-L., Xin, C., Wong, A. O. K., Jain, M., Kesharwani, R. K., Sedlazeck, F. J., & Luo, R. (2025). Clair3-RNA: A deep learning-based small variant caller for long-read RNA sequencing data. *Nature Communications*, 16(1), pp. 11553. <https://doi.org/10.1038/s41467-025-67237-y>
- Zody, M. C., Jiang, Z., Fung, H.-C., Antonacci, F., Hillier, L. W., Cardone, M. F., Graves, T. A., Kidd, J. M., Cheng, Z., Abouelleil, A., Chen, L., Wallis, J., Glasscock, J., Wilson, R. K., Reily, A. D., Duckworth, J., Ventura, M., Hardy, J., Warren, W. C., & Eichler, E. E. (2008). Evolutionary toggling of the MAPT 17q21.31 inversion region. *Nature Genetics*, 40(9), pp. 1076–1083. <https://doi.org/10.1038/ng.193>
- Zook, J. M., Chapman, B., Wang, J., Mittelman, D., Hofmann, O., Hide, W., & Salit, M. (2014). Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nature Biotechnology*, 32(3), pp. 246–251. <https://doi.org/10.1038/nbt.2835>

- Zook, J. M., Hansen, N. F., Olson, N. D., Chapman, L., Mullikin, J. C., Xiao, C., Sherry, S., Koren, S., Phillippy, A. M., Boutros, P. C., Sahraeian, S. M. E., Huang, V., Rouette, A., Alexander, N., Mason, C. E., Hajirasouliha, I., Rick-
etts, C., Lee, J., Tearle, R., ... Salit, M. (2020). A robust benchmark for
detection of germline large deletions and insertions. *Nature Biotechnology*,
38(11), pp. 1347–1355. <https://doi.org/10.1038/s41587-020-0538-8>

Appendix A

Structural variation in 1,019 diverse humans based on long-read sequencing

<https://doi.org/10.1038/s41586-025-09290-7>

Received: 22 April 2024

Accepted: 16 June 2025

Published online: 23 July 2025

Open access

 Check for updates

Siegfried Schloissnig^{1,19}, Samarendra Pani^{2,3,19}, Jana Ebler^{2,3}, Carsten Hain⁴, Vasiliki Tsapalou⁴, Arda Söylev^{2,3}, Patrick Hüther^{1,5}, Hufsah Ashraf^{2,3}, Timofey Prodanov^{2,3}, Mila Asparuhova^{1,6}, Hugo Magalhães^{2,3}, Wolfram Höps⁷, Jesus Emiliano Sotelo-Fonseca^{8,9}, Tomas Fitzgerald¹⁰, Walter Santana-Garcia¹⁰, Ricardo Moreira-Pinhal^{11,12}, Sarah Hunt¹⁰, Francy J. Pérez-Llanos^{13,14}, Tassilo Erik Wollenweber¹³, Sugirthan Sivalingam¹⁵, Dagmar Wieczorek¹⁵, Mario Cáceres^{11,12,16}, Christian Gilissen⁷, Ewan Birney¹⁰, Zhihao Ding¹⁷, Jan Nygaard Jensen¹⁷, Nikhil Podduturi¹⁷, Jan Stutzki¹⁸, Bernardo Rodriguez-Martin^{4,8,9}, Tobias Rausch⁴, Tobias Marshall^{2,3} & Jan O. Korbel^{4,10}

Genomic structural variants (SVs) contribute substantially to genetic diversity and human diseases^{1–4}, yet remain under-characterized in population-scale cohorts⁵. Here we conducted long-read sequencing⁶ in 1,019 humans to construct an intermediate-coverage resource covering 26 populations from the 1000 Genomes Project. Integrating linear and graph genome-based analyses, we uncover over 100,000 sequence-resolved biallelic SVs and we genotype 300,000 multiallelic variable number of tandem repeats⁷, advancing SV characterization over short-read-based population-scale surveys^{3,4}. We characterize deletions, duplications, insertions and inversions in distinct populations. Long interspersed nuclear element-1 (L1) and SINE-VNTR-Alu (SVA) retrotransposition activities mediate the transduction^{8,9} of unique sequence stretches in 5' or 3', depending on source mobile element class and locus. SV breakpoint analyses point to a spectrum of homology-mediated processes contributing to SV formation and recurrent deletion events. Our open-access resource underscores the value of long-read sequencing in advancing SV characterization and enables guiding variant prioritization in patient genomes.

SVs make up most polymorphic base pairs in the genome⁴, and are causally implicated in numerous common and rare diseases^{1,2}. A subset of SVs exhibit strong population stratification^{10,11}. Recently, the Human Pangenome Reference Consortium (HPRC) released a draft pangenome from 44 diploid long-read sequencing (LRS) assemblies generated from multiple genomic platforms, and demonstrated how this graph-based reference enhances SV discovery¹². Although LRS is increasingly used in disease research and diagnostics^{13–16}, population-scale datasets with comprehensive global representation remain limited⁵. Addressing this gap would be critical to facilitate widespread community access to rare and ancestry-specific genetic variation, especially for variants such as insertions, which are typically underrepresented in short-read sequencing datasets¹⁷. These data are expected to be instrumental to facilitate human diversity and disease research, including variant prioritization in genomic medicine.

Here we applied Oxford Nanopore Technologies (ONT) LRS to analyse SVs from the 1000 Genomes Project (1kGP) sample collection¹⁸, which allows open and unrestricted public data access, data sharing and reuse. To allow inclusion of a wide diversity of haplotypes, we performed ONT-based genomic sequencing of over a thousand human samples to intermediate coverage. To benefit from recent advances in interpreting genetic variation in a pangenome¹², we established methods that leverage linear and graph-based pangenomic references for LRS-based SV discovery, and engineered a computational framework using graph augmentation for SV genotyping. Our resulting pangenomic dataset constitutes a comprehensive collection of DNA sequence-resolved SV alleles, encompassing a variant frequency spectrum from common to rare across 26 diverse human populations. This resource provides a global reference of genetic variation using LRS technology—facilitating studies of SV biology and human disease, and yielding insight into allelic architecture, mechanistic

¹Research Institute of Molecular Pathology (IMP), Vienna BioCenter (VBC), Vienna, Austria. ²Institute for Medical Biometry and Bioinformatics, Medical Faculty and University Hospital Düsseldorf, Heinrich Heine University Düsseldorf, Düsseldorf, Germany. ³Center for Digital Medicine, Heinrich Heine University Düsseldorf, Düsseldorf, Germany. ⁴European Molecular Biology Laboratory (EMBL), Genome Biology Unit, Heidelberg, Germany. ⁵Institute of Molecular Biology (IMB), Mainz, Germany. ⁶Institute of Molecular Biotechnology of the Austrian Academy of Sciences (IMBA), Vienna BioCenter (VBC), Vienna, Austria. ⁷Department of Human Genetics, Radboud University Medical Center, Nijmegen, The Netherlands. ⁸Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain. ⁹Universitat Pompeu Fabra (UPF), Barcelona, Spain. ¹⁰European Molecular Biology Laboratory, European Bioinformatics Institute, Cambridge, UK. ¹¹Institut de Biociències i de Biomedicina, Universitat Autònoma de Barcelona, Bellaterra, Spain. ¹²Research Programme on Biomedical Informatics (GRIB), Hospital del Mar Research Institute, Barcelona, Spain. ¹³Biological and Medical Research Center (BMFZ), Medical Faculty and University Hospital Düsseldorf, Heinrich Heine University Düsseldorf, Düsseldorf, Germany. ¹⁴West German Genome Center (WGGC), Medical Faculty and University Hospital Düsseldorf, Heinrich Heine University Düsseldorf, Düsseldorf, Germany. ¹⁵Institute of Human Genetics, Medical Faculty and University Hospital Düsseldorf, Heinrich Heine University Düsseldorf, Düsseldorf, Germany. ¹⁶ICREA, Barcelona, Spain. ¹⁷Global Computational Biology and Digital Sciences (gCBDS), Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach an der Riss, Germany. ¹⁸BI X GmbH, Ingelheim am Rhein, Germany. ¹⁹These authors contributed equally: Siegfried Schloissnig, Samarendra Pani. ✉e-mail: bernardo.rodriguez@crgeu; tobias.rausch@embl.de; tobias.marshall@hhu.de; jan.korbel@embl.de

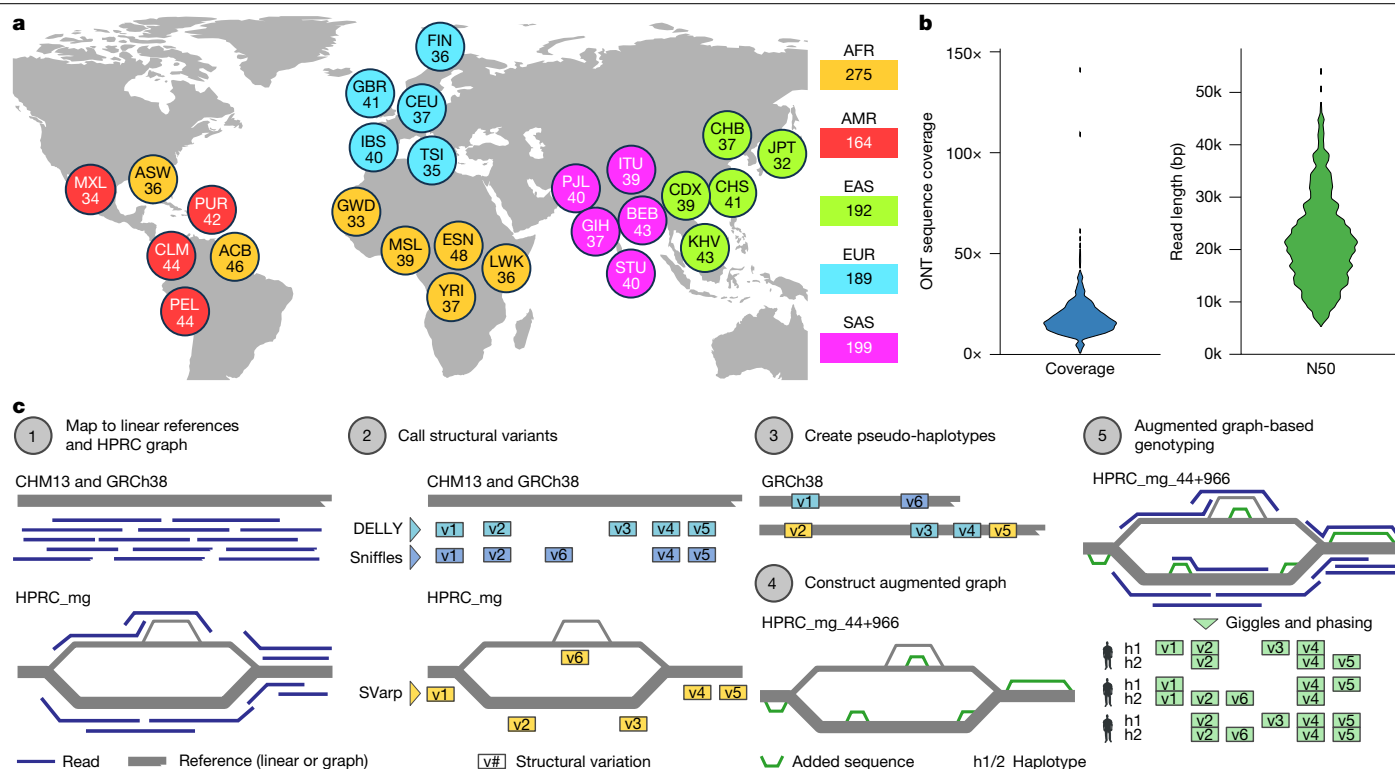


Fig. 1 | LRS and SAGA. a, Breakdown of self-identified geographical ancestries for 1,019 long-read genomes representing 26 geographies (that is, populations) from 5 continental regions. The three-letter codes used are equivalent to those used in the 1kGP phase III¹⁸ and are resolved in Supplementary Table 2. **b**, ONT sequence coverage per sample, expressed as fold-coverage (left), and N50

read length in base pairs (right). **c**, Schematic of the SAGA framework for graph-aware discovery and genotyping of SVs using a pangenome graph augmentation approach. Basemap in **a** from Natural Earth data (<https://www.naturalearthdata.com>).

origin, mutational recurrence and population distribution of SV classes.

LRS and graph-based SV discovery

We selected samples from the 1kGP collection, subjecting size-selected more than or equal to 25-kilobase pair (kb) DNA fragments to LRS (Supplementary Fig. 1 and Supplementary Table 1). After quality filtering (Methods), our study cohort comprises 1,019 genomes (Fig. 1a), sequenced to a median coverage of 16.9× with a median N50 read length of 20.3 kb (Fig. 1b and Supplementary Fig. 2). These samples are from 26 self-reported population groups and span five continental areas, including 189 from European (EUR) donors, 192 from East Asia (EAS), 199 from South Asia (SAS), 275 from Africa (AFR) and 164 from the Americas (AMR).

To allow for graph-aware SV discovery and genotyping in these data, we devised the SV analysis by graph augmentation (SAGA) framework (Fig. 1c). Outlined below, SAGA integrates read mapping to both linear and graph references, followed by graph-aware SV discovery and genotyping at population scale.

Long-read alignment

In the initial step of SAGA, we performed read alignment against both linear (GRCh38, and the Telomere-to-Telomere (T2T) reference, denoted CHM13) and graph (the minigraph HPRC reference graph, denoted HPRC_mg) genomic references (Methods). Comparative analyses of the alignments show that HPRC_mg and CHM13 outperform GRCh38 (with average mapping identities increased by more than 0.5%), whereas the size distribution of alignment gaps indicates that HPRC_mg comprises a more comprehensive collection of mobile element insertions (MEIs) and deletions (Supplementary Fig. 3). Using CHM13, we find that on

average 93.6% of the genome exhibits a coverage of 5× or more per sample (Supplementary Fig. 4). Furthermore, using WhatsHap (Methods) we haplotype-phased single nucleotide polymorphisms (SNPs) using our ONT dataset, and find that these show excellent concordance with phased SNPs from a short-read-based 1kGP study³, as evidenced by median switch error rates of only 0.69% in children from the six parent-offspring trios included in our dataset, and 1.32% for unrelated (and parental) samples (Supplementary Fig. 5 and Supplementary Note 1).

SV discovery

To allow for comprehensive SV discovery, we used LRS-based SV callers tailored to linear reference genomes. We used Sniffles and DELLY (Methods), applying both to GRCh38 and CHM13. Callset integration across both linear references yielded an average number of 15,301 and 21,529 SVs per sample for Sniffles and DELLY, respectively.

To complement these classical SV discovery methods, we additionally harnessed the graph-aware SVarp algorithm (Methods), which allows SV discovery in haplotype contexts represented in HPRC_mg yet absent from a linear reference. SVarp searches for SV patterns from graph-aligned reads, and then performs local long-read assembly to reconstruct ‘SV sequence contigs’ (svtigs). To maximize the accuracy of svtig assembly, we applied SVarp to a subset of our sample set ($n = 967$ genomes) previously sequenced at high coverage with short reads³. This allowed efficient haplotype assignment (‘haplo-tagging’) of 69.9% of the ONT reads using phased SNPs. Using this haplo-tagged read information, SVarp constructed on average 1,145 svtigs per genome not previously contained in the HPRC_mg graph (Supplementary Fig. 6).

Graph augmentation

To enable non-redundant SV callset integration across all SV callers, we augmented the HPRC_mg graph by incorporating further bubbles¹²

representing new SV allelic sequences. To achieve this, SAGA encompasses a pseudo-haplotype construction and a graph augmentation step (Fig. 1c). Pseudo-haplotype construction produces chromosome-wide, haplotype-like sequences that incorporate the discovered SV alleles in a non-overlapping manner (Methods). Graph augmentation then uses the minigraph tool (Methods) to integrate these pseudo-haplotypes into the graph. Our study augments the HPRC_mg graph (originally representing 44 samples) with SVs from all 967 LRS samples considered for comprehensive SV discovery. This process yields 'HPRC_mg_44+966', a pangenome representing SVs from 1,010 individuals (considering one sample in our resource (HG01258) is part of the HPRC_mg graph already; Extended Data Fig. 1 and Supplementary Fig. 7). HPRC_mg_44+966 comprises 220,168 bubbles altogether, compared with 102,371 bubbles present in HPRC_mg (Supplementary Figs. 8 and 9).

Of the 117,797 new bubbles represented in HPRC_mg_44+966, 105,744 (90%) are at least 1 kb away from the nearest bubble in the original HPRC_mg graph. These are presumed to reflect SVs not previously represented in HPRC_mg (Supplementary Fig. 10). To evaluate the quality of HPRC_mg_44+966, we mapped ONT reads from the HG00513 sample onto this graph. We observe improved alignment metrics, with a gain of 33,208 aligned reads and a further 152.5 megabases (Mb) of aligned bases compared with alignment onto HPRC_mg (Supplementary Table 3), suggesting the augmented graph provides enhanced SV analysis capabilities.

SV genotyping and phasing

Unified SV genotypes are a prerequisite for relevant downstream analyses with population-scale variant datasets, including population genetic and disease studies. As the final step of SAGA, we therefore use Giggles, a genotyping tool that harnesses graph-aligned long reads for SV genotyping (Methods). We genotyped the set of 967 samples of HPRC_mg_44+966 with available LRS data, yielding genotypes for 167,291 primary SV sites.

Consistent haplotype phasing elevates the value of SV resources, facilitating allele-specific analyses and investigation of haplotype structures. We used a recently generated CHM13 haplotype reference panel¹⁹ constructed from short-read SNP calls to carry out statistical SV phasing using SHAPEIT5 (Methods). We find that 164,571 (98.4%) of the genotyped SV sites are successfully phased. These comprise our final SAGA-based SV callset, and include 65,075 deletions, 74,125 insertions and 25,371 'putatively complex' sites, for which both the reference and alternative allele are larger than 1 base pair (bp) (Supplementary Note 2). On the basis of analysing the respective graph bubble structures (Methods), 107,005 SVs in the phased SV callset are classified as biallelic, and the remaining ($n = 57,566$) multiallelic.

Resource quality assessment

Comparison of this SV callset with SVs discovered from multi-platform genome assemblies recently constructed by the Human Genome Structural Variation Consortium (HGSVC)²⁰ suggests a genome-wide false discovery rate (FDR) of 15.55% for deletions and 15.89% for insertions. The FDR varies by SV size: SVs ≥ 250 bp show considerably lower FDR (deletions: 6.91%, insertions: 8.12%) than SVs < 250 bp (deletions: 19.14%, insertions: 19.57%; Extended Data Fig. 2 and Supplementary Figs. 11 and 12). The smallest SVs largely comprise tandem repeat variation, whose divergent representation between graph- and assembly-based callsets can complicate both SV discovery and comparative analysis²¹. MEIs—an SV class exhibiting well-defined allele architectures—exhibit a particularly low FDR (0.85–6.75%; Supplementary Fig. 13), whereas mobile element deletions show 1.94% FDR (Supplementary Note 3). Comparison with multi-platform assemblies²⁰ also yielded autosomal sensitivity estimates, which vary by SV class, allele frequency and size (Extended Data Fig. 2 and Supplementary Figs. 14 and 15). Sensitivity is particularly high for MEIs (ranging from 84.3% to 90.6%). The genome-wide true positive rate is 64.36% for deletions and 67.33%

for insertions after genotyping (Extended Data Fig. 2 and Methods). Reflecting our intermediate-coverage study design, more common SVs exhibit an improved true positive rate over rarer SVs, with 79.59% for deletions and 83.24% for insertions seen for minor allele frequency (MAF) ≥ 0.1 .

We additionally examined the primary SV genotypes. SV allele frequencies after phasing are in excellent agreement with the primary SV genotypes (Supplementary Figs. 16 and 17). We used data from the six parent–offspring trios to identify 'Mendelian inconsistencies', which could derive from de novo SV formation events or genotyping errors. The average rate of such inconsistencies for biallelic SVs is 3.87% for deletions, 4.44% for insertions and 4.10% for putatively complex sites, implying high genotype accuracy (Supplementary Tables 4–12). For multiallelic sites the average Mendelian inconsistency is 15.1%, reflecting previously reported challenges in genotyping such variant sites^{4,18}. Our estimates align with strong performance against Hardy–Weinberg equilibrium checks (Extended Data Fig. 3 and Supplementary Figs. 18 and 19) and genotypes are largely robust to coverage and read length variations (Supplementary Note 4). Moreover, our SV sites intersect 69.5% of insertions and 64.9% of deletions called from short-read sequencing in the same samples³ (Supplementary Fig. 20). At these intersecting sites, the genotype concordance is 98.7% for deletions (non-reference genotype concordance: 77.6%) and 96.8% for insertions (non-reference concordance: 79.0%) compared with Illumina-based genotypes³, in support of high genotype accuracy.

The SV landscape across 26 ancestries

We analysed the degree to which this SV dataset grows cumulatively after each new genome is added (Fig. 2a and Supplementary Fig. 9). We observe pronounced saturation effects²² for common SVs when adding new samples to the pangenome graph, with AFR samples disproportionately increasing SV yield consistent with higher levels of genetic variation in these populations^{18,22}.

When compared with previous population-scale genome sequencing studies in 1kGP samples^{3,4}, our resource captures more variants across the SV size spectrum, with SVs ranging from 50 to several hundred base pairs poorly captured using short reads¹⁷. The median SV count per sample is 23,969 for AFR samples (19,297 for other ancestries, non-AFR; Supplementary Note 4), compared with 9,963 and 8,540 for AFR and non-AFR, respectively, detected through short reads when subsetting to the same samples³. The proportionally largest gain is seen for insertions, which our resource captures with increased sensitivity across insertion sizes (Fig. 2b and Supplementary Fig. 21). By comparison, large deletions are detected with increased abundance using short reads¹⁷ (Fig. 2b). Furthermore, facilitated by the read length in our resource, we observe a more than tenfold increase in insertion sites with sequenced-resolved SV alleles (Fig. 2b). For deletions, a considerable fraction of which can be resolved through split-read analysis using short reads^{3,4,23}, our resource increases the nucleotide-resolved SV count by 40% (from 46,895 to 65,812). These analysis results are consistent with the improved characteristics of LRS in resolving SVs^{17,22,24}.

Rare SVs from diverse ancestries

Analysing the allele frequency spectrum from our augmented pangenome shows a broad spectrum of SV alleles, from common to rare (Fig. 2c, Extended Data Fig. 3 and Supplementary Fig. 22). We observe a considerable enhancement in capturing SVs with an allele frequency $\leq 2\%$ over HPRC_mg, with the vast majority of rare SVs stemming from samples newly incorporated into HPRC_mg_44+966 (Supplementary Fig. 23). Across our resource, most SVs are rare (59.3% have an MAF $< 1\%$), and, with the current sample size, are typically detected in a single continental cohort (AFR, AMR, EAS, EUR or SAS). By comparison, at an allele frequency $\geq 2.5\%$, most SVs are seen in at least two continental groups (Extended Data Fig. 3). Multiallelic SV loci generally

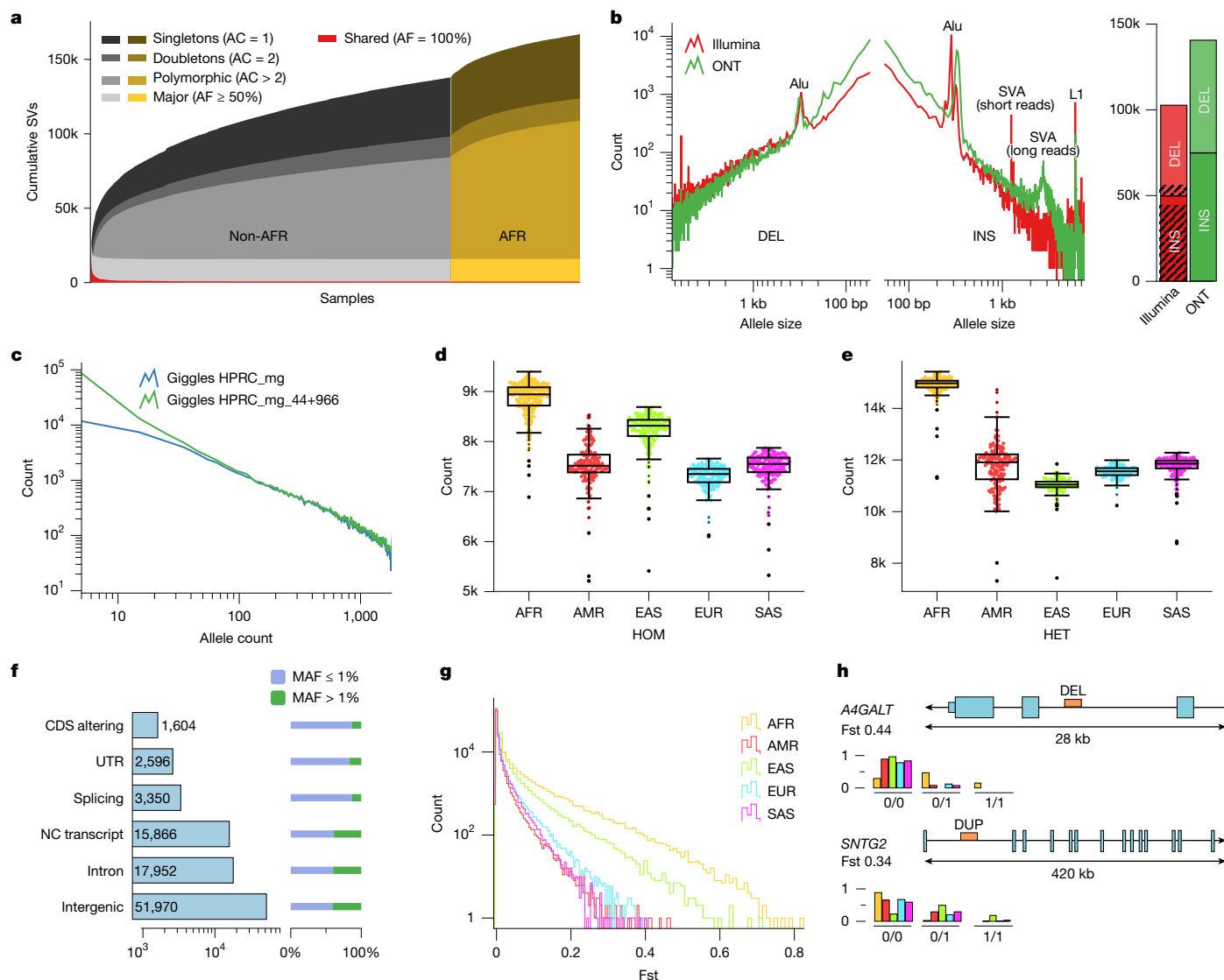


Fig. 2 | Callset properties and SV landscape in different geographical ancestries. **a**, Cumulative number of unique SVs from the SAGA framework when adding individual long-read genomes, from left to right. The rate of SV discovery slows with each new sample added. Colours denote singletons, doubletons, SVs seen with an allele count of more than 2 (polymorphic), as well as major and shared alleles. **b**, Left, SV length distributions in population-scale 1kGP SV callsets, versus the cumulative count of SV sites, with a comparison of ONT sequencing (ONT; $n = 967$ samples) with the short-read-based analysis of the 1kGP cohort³ (Illumina) subsetted to the same ($n = 967$) samples. Right, the 1kGP ONT callset is outnumbering short-read-based SV calls both for deletions (DEL, upper sections) and insertions (INS, lower sections). SVs previously unresolved by their sequence, using short reads³, are depicted by shaded areas. Notably, LRS provides exact MEI lengths, in contrast to the length approximations reported in short-read-based studies³ (Supplementary Fig. 29). **c**, SV allele count

(x-axis) relative to the count of SV sites (y-axis), constructed by genotyping the original HPRC graph (HPRC_mg) and the augmented graph (HPRC_mg_44+966) using Giggles. Consistent with the considerably smaller panel size used, HPRC_mg under-represents rare alleles. **d, e**, Homozygous (HOM) SV count (**d**) and heterozygous (HET) SV count (**e**) per sample stratified into self-identified geographies ($n = 967$; whiskers extend to points that are within $1.5 \times$ interquartile range from the upper or the lower quartiles). **f**, Impact of SVs on distinct genomic features. SVs affecting sequences of genes occur at lower MAF compared with non-coding regions. **g**, Frequency polygons of SV Fst values per continental population (histograms stratifying Fst for bi- and multiallelic are in Supplementary Fig. 28). **h**, Deletion (DEL) and duplication (DUP) in the intragenic space of two medically relevant genes exhibiting differentiation in AFR and EAS samples, respectively. AC, allele count; AF, allele frequency; CDS, coding sequence; NC, non-coding; UTR, untranslated region.

have a higher propensity to be shared across continents than biallelic SVs, with most multiallelic SVs with allele frequency $\geq 1.5\%$ being shared (Supplementary Fig. 24), which is potentially explained by recurrent rearrangements at these loci. Furthermore, the relative increase in SVs in AFR samples is more pronounced for heterozygous than for homozygous SVs, reflecting the greater genetic diversity in AFR samples²² (Fig. 2a,d,e and Supplementary Fig. 25).

Population characteristics by SV class

Unlike previous 1kGP studies that have placed a focus on characterizing sequence-resolved deletions^{3,4}, we find comparable population

characteristics between SVs primarily identified as deletions and insertions (Supplementary Fig. 19). Measuring the degree to which SVs are in linkage disequilibrium with nearby SNPs, we find that 62.6% of deletions and 62.9% of insertions with at least 1% MAF are in linkage disequilibrium with nearby SNPs ($r^2 \geq 0.5$) (Extended Data Fig. 3 and Supplementary Fig. 26). These metrics increase to 89.8% (deletions) and 91.7% (insertions), respectively, in Genome-in-a-Bottle (GiB) 'high-confidence regions' (Methods), which are more depleted of repeats and thus less likely subject to recurrent SV or gene conversion events²⁵. Both larger insertions and deletions are rarer in the population than smaller SVs (Supplementary Fig. 27), possibly explained by stronger negative

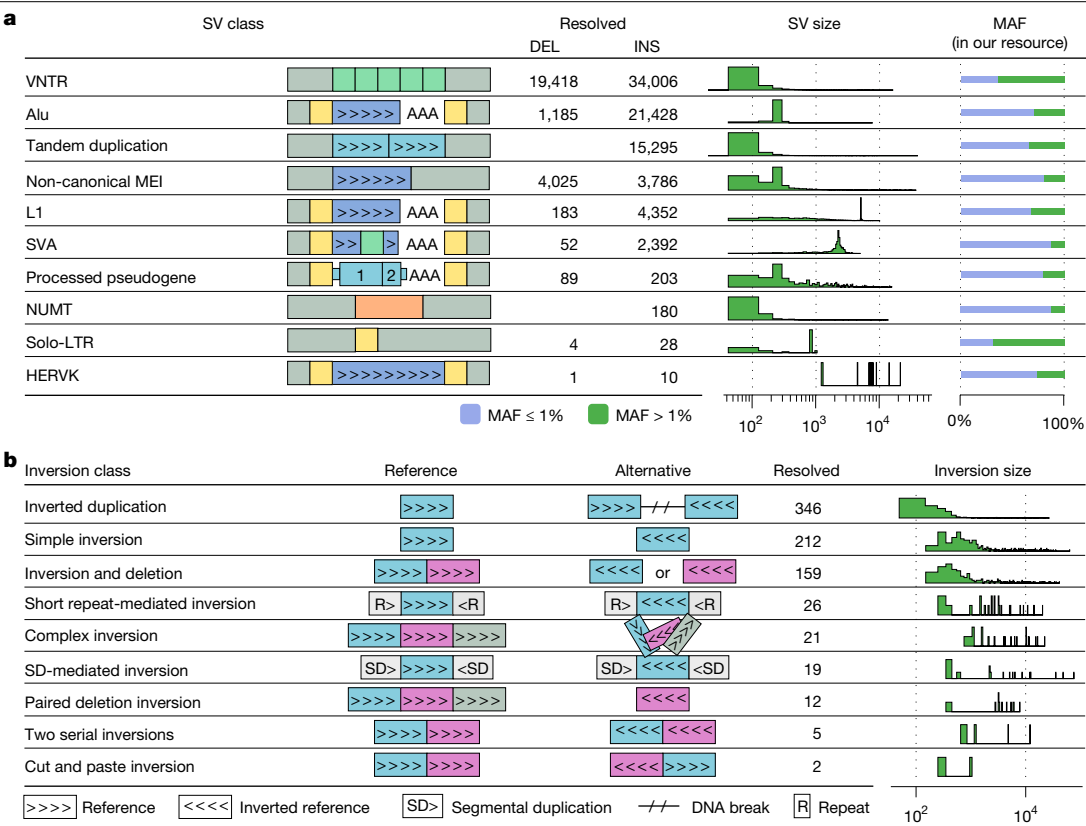


Fig. 3 | Prevalence of distinct SV classes in our SAGA-based data resource. **a**, VNTRs (here shown as identified with the SAGA framework), duplications and insertions, as classified by SVAN. The SV class is denoted by label and ideogram, and the number of resolved class members is given for insertions (INS; insertion relative to the reference) and deletions (DEL; deletion relative to reference), in addition to size distribution and percentage of low- and high-frequency alleles. Repeat units for VNTRs and internal VNTR sequence for SVA insertions are depicted using green boxes. TSDs flanking retrotransposition insertion events

and LTRs flanking HERVK are represented as yellow boxes. Numbers for interspersed and complex duplications are shown in Supplementary Table 17. Alu, L1 and SVA refer to canonical retrotransposition events. **b**, Inversion classes identified (twin priming events excluded from this display). For each class, the reference structure is contrasted with the alternative allele, the number of resolved inversions is shown and the inversion size distribution is shown (in bp).

selection against variants affecting functional sequence. Consistent with this notion, SVs affecting functional elements show significantly reduced MAF ($P < 7.4 \times 10^{-235}$; Kolmogorov–Smirnov test; Fig. 2f).

Geographical stratification of SVs

A principal component analysis of SV calls, as well as an SV-based admixture analysis, show groupings of samples corresponding well with donor ancestry⁴ (Extended Data Fig. 3). Using fixation indices (Fst) to quantify SV differentiation by continent, the allelic diversity of AFR (followed by EAS) is clearly observable (Fig. 2g and Supplementary Fig. 28). We find evidence for differentiation ($F_{st} > 0.2$) for 8,597 SVs (Supplementary Table 13), of which 105 encompass a GiaB-classified medically relevant gene²⁶, and thus are of potential interest for disease studies (Supplementary Table 14). Examples are a deletion and duplication affecting the intragenic regions of *A4GALT* (MAF 14%) and *SNTG2* (MAF 18%), enriched in AFR and EAS, respectively (Fig. 2h). We further note a complex SV near *LAMBI* exhibiting enrichment in AMR, and particularly among samples with self-identified ancestry of Peruvians in Lima (PEL) (Supplementary Fig. 29). Additionally, we observe a strong correlation between 1-Mb-window averaged SV-based and SNP-based Fst values (Pearson’s $P < 4.0 \times 10^{-16}$), deviations of which potentially serve as indicators for genomic areas with SV-driven differentiation. Requiring a confidence level of more than 5 s.d. ($P < 0.6 \times 10^{-6}$), we find 11 regions in which the differentiation is likely to be SV-driven (Supplementary Table 15), including a deleted region near gene *ARHGAP24* with the highest MAF of 27.3% ($F_{st} = 0.31$) seen in the Mende in Sierra Leone (MSL).

Resolved spectrum of SV classes

We devised the SV annotator (SVAN), an algorithm that leverages allelic representations and genomic annotations to classify SVs into distinct classes (Methods). SVAN classifies 72,346 insertions (96.0%) of our pangenomic resource (Fig. 3a and Extended Data Fig. 3), offering a detailed perspective of this variant class. SVAN additionally classifies 21,301 (32.2%) deletions, and 14,681 (57.1%) of all putatively complex sites (Fig. 3a), the latter of which resolve into 11,030 further insertions and 3,651 further deletions.

Duplications and variable number tandem repeats

In total, 17,029 (19.5%) of the insertions in our pangenomic resource are duplications, of which most (15,295, 89.8%) represent tandem duplications (Fig. 3). Additionally, 34,006 (39.0%) insertions and 19,418 (26.2%) deletions are classified as variable number tandem repeats (VNTRs), an SV class relevant to human diversity and disease²⁷ (Fig. 3a). Unlike for other SV classes, most VNTRs (50.4%) in our SAGA-based SV callset represent multiallelic SV sites, which poses a challenge to graph construction tools, as finding the correct alignment can be difficult if several SV alleles at the same locus differ in their representation by only a few base pairs²¹. Therefore, to allow us to systematically capture VNTR complexity, we additionally genotyped VNTR sites across our resource using the *vamos* tool (Methods). A detailed analysis of these genotypes shows 739 biallelic and 369,685 multiallelic VNTRs, considerably surpassing estimates derived from

graph-based analysis, and thus providing a refined estimate of allelic VNTR diversity. In-depth comparison against multi-platform whole-genome assemblies²⁰ shows a strong concordance in polymorphic repeat unit counts, supporting the robustness of the *vamos*-based VNTR genotypes (Supplementary Fig. 30). Unlike genome assemblies at present limited to small sample sizes, our resource enables population-scale analysis of extreme repeat unit counts, illustrated by the 1% and 99% percentiles of the repeat unit distribution (Extended Data Fig. 4). The utility of these genotypes is exemplified by our analysis of repeat unit variation at the *PLIN4* and *ABCA7* loci implicated in late-onset diseases (Extended Data Fig. 4, Supplementary Table 16 and Supplementary Note 5).

Mobile elements and nuclear mitochondrial DNA segments

We next focused on the distinct classes of insertions classified by SVAN, most of which are the product of the activity of mobile elements (Fig. 3a). These include 31,302 non-reference MEIs, including 23,212 Alu, 4,851 L1 and 3,239 SVA insertions (Extended Data Fig. 5 and Supplementary Table 18), which reflects an increase of 20% (for Alu), 166% (L1) and 179% (SVA), respectively, over short-read-based analysis in the same samples³. We further identify 3,813 reference MEIs primarily called as deletions (classified into 3,122 Alu, 460 L1 and 231 SVA events, respectively). Using SVAN, we classify MEIs as canonical (84.3%; 29,592) and non-canonical (15.73%; 5,523). Thereby, canonical MEIs show a conformation consistent with target-primed reverse transcription²⁸ or twin priming²⁹, whereas non-canonical ones are 3' truncated (34.4%; 1,901), lack poly(A) tails (66.4%; 3,665) or have internal rearrangements (52.8%; 2,916). Non-canonical MEIs may have arisen through alternative integration mechanisms^{30–32}, or be the product of SVs occurring after the integration event. Canonical MEIs typically show distinctive hallmarks of retrotransposition³³, including target site duplications (TSDs, median: 21 bp) and poly(A) tracts (median: 30 bp) seen for 98.7% and 100% of the events (Extended Data Fig. 5), respectively. Consistent with previous studies^{22,34–36}, most (72.5%; 3,288) of the canonical L1 insertions represent 'dead-on-arrival' copies owing to either truncation or inversion at their 5' ends (Extended Data Fig. 5). This contrasts with canonical Alu and SVA for which 74.4% and 72.9% of the insertions, respectively, are full-length. We find that full-length SVAs show a broad size distribution, ranging from 1 to 6 kb, largely owing to the extensive variability observed in their hexameric repeat and VNTR regions (Extended Data Fig. 5).

We also identify 203 non-reference processed pseudogenes, as well as 89 reference polymorphic processed pseudogenes^{37,38} (Extended Data Fig. 5). Most of these (73.9%; 216) are monoexonic, with the remaining pseudogenes comprising between 2 and 21 exons. We further identify evidence for human endogenous retrovirus (HERV) activity, with 10 insertions classified as HERVK and 28 solo-long terminal repeats (LTRs) (Extended Data Fig. 6). Manual inspection of these events shows that 82% (23 of 28) of solo-LTRs correspond to complete LTR5_Hs events. With respect to HERVK, five of the insertions encompass the proviral sequence flanked by two LTRs, four have a single LTR with the second residing on the human reference (CHM13) and one is heavily 5'-truncated. Comparisons with previous reports^{35,39} show that although only around 22% (5 of 23) of the solo-LTRs in our study were not previously reported, most (70%; 7 of 10) of the HERVK insertions were not described previously.

Finally, SVAN classifies 180 insertions as nuclear mitochondrial DNA segments (NUMTs)⁴⁰. The median length of these insertions is 126 bp, with 15 NUMTs larger than 1 kb, including a 14.8-kb insertion comprising 90% of the mitochondrial length identified in one carrier (Supplementary Fig. 31d).

Inversions

Inversions pose considerable challenges for detection by sequencing owing to their balanced copy-number state, which limits their

indirect identification through coverage depth analysis. Additionally, difficulties in sequence alignment, particularly arising from repetitive sequences flanking inversion breakpoints, further complicate their accurate ascertainment³⁶. The notable N50 read length of this ONT dataset led us to explore its potential for inversion discovery, for which we noted that performing read re-alignment at regions exhibiting clustered mismatches enhances inversion detection accuracy (Supplementary Note 6 and Supplementary Figs. 32 and 33). We devised an inversion discovery workflow leveraging these insights (Methods), identifying 491 inversions which were confirmed by semi-manual (dotplot-based) analysis (Supplementary Fig. 34). We added to these inversions further SVs that were primarily defined as insertions, yet reclassified by SVAN—including 311 inverted duplications and 1,047 inversions at the 5'-ends of L1 insertions probably resulting from twin priming²⁹ (Extended Data Fig. 5)—yielding 1,849 inversion calls overall (Fig. 3b). After omitting SVs previously genotyped with Giggles as well as SVs annotated as inverted duplications, the GeONTipe tool (Methods) successfully determined genotypes for 78% of the inversions.

We next conducted detailed analyses of the set of 733 inversions outside MEI events, focusing on their size distribution, allelic structure and complexity (Methods). Inverted duplications represent the most common class, with 277 events detected showing a median length of 284 bp (Fig. 3b and Supplementary Fig. 35). Additionally, we find 257 balanced (or 'simple') inversions with a median length of 1,565 bp, which are further broken down into 45 inversions bordered by various repeat classes in inverted orientation, implying formation through homology-directed repair (HDR) processes or non-allelic homologous recombination (NAHR)⁴¹.

We further find an abundance of complex inversions with multiple breakpoints^{4,36} (Fig. 3b). These include 159 inversions with an adjacent deletion, and 40 more complex structures, comprising five SVs characterized by two serial inversions; two instances of 'cut-and-paste' inversions with the locus structure characterized by an excised, inverted and inserted segment into a nearby location; 12 inversion sites exhibiting two flanking deletions; and 21 particularly complex inversions with a large number of breakpoints and in part incompletely resolved allelic structure. These events are likely to originate from a DNA replication-associated process, such as microhomology-mediated break-induced replication⁴².

Polymorphic L1 and SVA transductions

MEI transductions represent an SV generation process in which a full-length mobile element, denoted as source or progenitor locus, mobilizes flanking DNA sequences during retrotransposition⁴³. This process can result in the integration of a mobile element and a companion non-repetitive sequence situated in 5' or 3' relative to the progenitor copy^{8,9}. Harnessing our pangenomic resource, we analysed polymorphic transduction events among all sequence-resolved MEI sequences. We annotate 878 transductions, with 466 (8.8%) of the 5,311 L1 insertions and 412 (11.9%) of the 3,470 SVA insertions exhibiting a transduction (Supplementary Tables 19 and 20). In addition to these transduction events containing a companion mobile element segment (denoted 'partnered transductions'), a further set of 48 transductions are truncated resulting in the integration of the transduced sequence alone ('orphan transductions').

Consistent with previous studies^{8,9,22}, we find that the relative proportion of 5' or 3' partnered transductions differs between MEI families. Most (82.2%; 350) L1-mediated transductions are in 3', whereas SVAs generate 5' and 3' transductions at similar ratios, with 218 (53.7%) detected in 5' and 188 (46.3%) in 3' (Supplementary Table 20). Although SVA-mediated 3' transductions are significantly longer in size than 5' transductions ($P = 7.2 \times 10^{-6}$; two-tailed Mann–Whitney U test), we observe no significant length difference for L1s (Extended Data Fig. 5). Furthermore, SVA-mediated 3' transductions are significantly longer

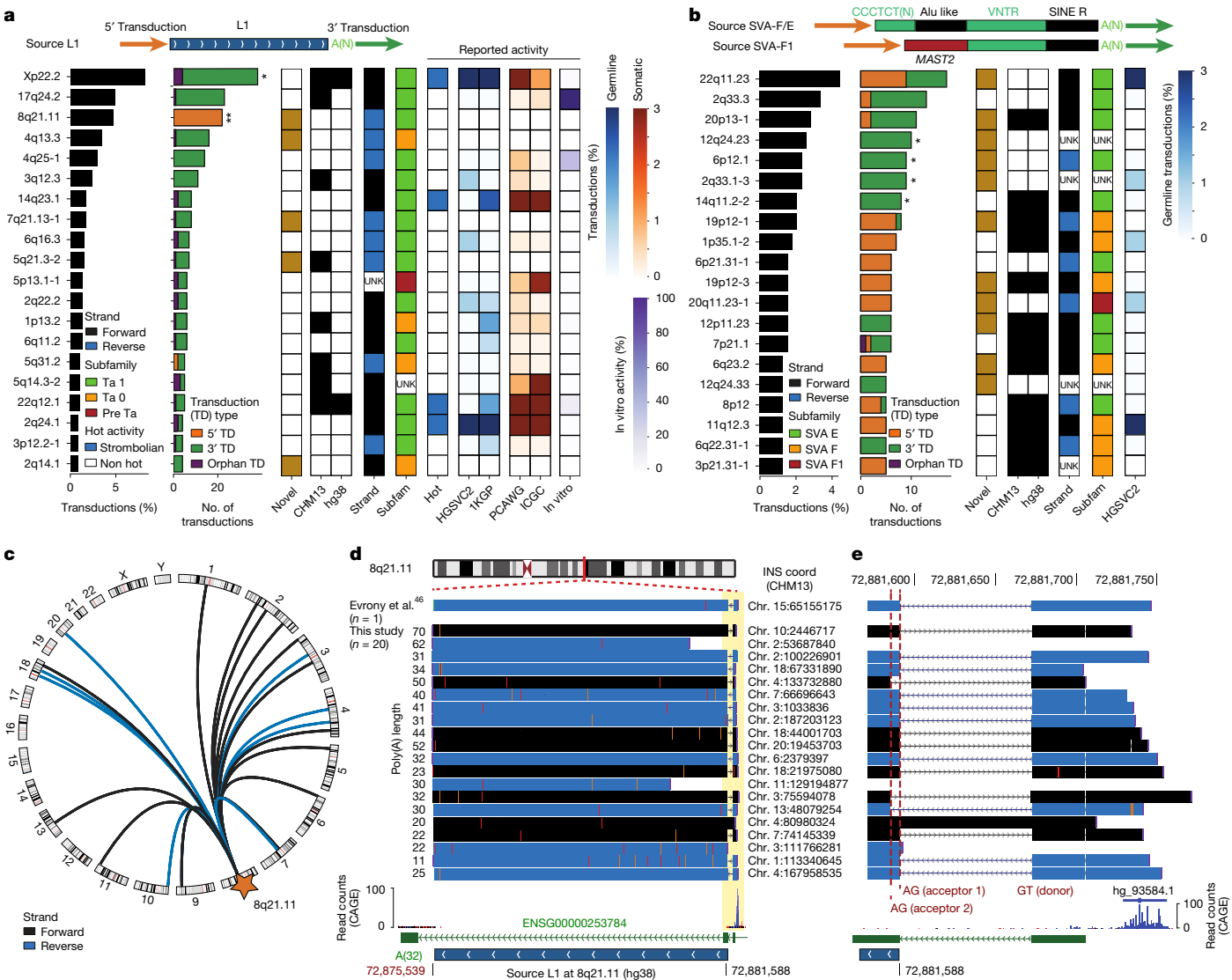


Fig. 4 | Polymorphic landscape of L1 and SVA transductions. **a, b.** Contribution relative to the total number of transductions (5', 3' and orphan) for the 20 most active L1 (**a**) and SVA (**b**) progenitors. Source elements are annotated by their presence/absence in the reference, orientation and subfamily. Source element novelty, hot activity status and previously reported activity estimates are based on in vitro assays and transduction tracing (Methods) and are shown as heat maps. Transduction 5' and 3' bias was assessed using a two-sided exact binomial test followed by multiple testing correction with Benjamini–Hochberg. Xp22.2 ($P_{\text{adj}} = 0.04$) and 8q21.11 ($P_{\text{adj}} = 3.9 \times 10^{-15}$) source L1s exhibit a significant 3' and 5' bias, respectively. All significant SVAs are 3' biased, namely 12q24.23 ($P_{\text{adj}} = 0.01$), 6p12.1 ($P_{\text{adj}} = 0.02$), 2q33.1-3 ($P_{\text{adj}} = 0.02$) and 14q11.2-2 ($P_{\text{adj}} = 0.04$). Adjusted P values for significantly biased loci are represented adjacent to each

bar as follows: $*P_{\text{adj}} < 0.05$; $**P_{\text{adj}} < 0.005$. **c.** Circos plot showing the integration positions for the 22 instances of 5' transductions mediated by the 8q21.11 element. **d.** Alignment of inserts containing 5' transductions at the source L1 region, including a single somatic transduction event reported in ref. 46 in the brain. Inserts are coloured according to whether they align in forward (black) or reverse (blue). Splicing between the full-length L1 and an upstream exon leading to 5' transductions, highlighted in yellow. **e.** Magnification showing that the 5' transductions initiate at a strong promoter located upstream, followed by canonical splicing between the first and second exon of *ENSG00000253784*, in addition to a second acceptor splice site within the L1 body. Transcription initiation is supported by an annotated transcription start site (hg_93584.1) and CAGE read counts⁶⁰.

compared with L1 ($P = 0.005$; Extended Data Fig. 5). This suggests the existence of family-specific determinants for 5' and 3' transductions. We leveraged the transduced sequences as barcodes to identify their progenitor loci, showing that L1 transductions originate from a limited set of 208 L1 source elements. Among these, 20 highly active source L1s are responsible for 43.9% (205) of all L1 transductions identified (Fig. 4a), with most (14 of 20) of them belonging to the youngest Ta-1 subfamily. By comparison, we detect 176 source SVA elements, with 20 loci alone mediating 36.9% (152) of all SVA transductions (Fig. 4b). All these highly active SVAs belong to the recent human-specific SVA-E and SVA-F subfamilies, with the exception of the source element at 20q11.23-1, which contains an exonic sequence from the *MAST2* gene—a signature of the SVA-F1 subfamily⁴⁴. Only a small subset—24% and 17% of

the source L1s and SVAs, respectively—were previously reported to be active on the basis of genomic or in vitro studies (Fig. 4ab and Methods). Of the 208 L1 progenitors, 156 exhibit only 3' transductions. This includes the most active source L1 locus, a Ta-1 element residing at Xp22.2 on the reference that mediates 38 transductions, all in 3' (Supplementary Tables 19 and 20). By comparison, the third most active L1 element at 8q21.11 exhibits 22 transductions, with all of them in its 5' resulting in a 5' transduction bias (adjusted P value ($P_{\text{adj}} = 3.9 \times 10^{-15}$, FDR-adjusted two-tailed binomial test; Fig. 4a,c). A previous report indicates the presence of a strong promoter located upstream of this source L1 (ref. 45), which has been associated with aberrant splicing leading to a somatic 5' transduction in the embryonic brain⁴⁶. Our analyses of 22 independent germline 5' transduction events arising from this

source locus suggest that all probably originate from the same mechanism, involving splicing of the first exon of the *ENSG00000253784* long non-coding RNA gene with two alternative acceptor splice sites located within the source L1 (Fig. 4d,e). This provides a mechanistic explanation for the 5' transduction bias seen for this source element, and indicates that L1 progenitors can hijack flanking regulatory elements leading to transduction bias in the germline.

By contrast, SVA progenitors show a pronounced locus-specific pattern of transduction activity, with 14 among the 20 most active source SVs showing solely 5' or 3' transductions (Fig. 4b), respectively. Of these, four SVA source elements exhibit significant bias towards 3' transductions (5% FDR; Methods)—including progenitors at 12q24.23, 2q33.1-3, 6p12.1 and 14q11.2-2. In summary, detailed analyses of our resource show family- and locus-specific patterns for L1- and SVA-mediated transductions in the germline.

Genomic breakpoint homology landscape

Prompted by the comprehensive set of nucleotide-resolved SVs, we comprehensively investigated SV breakpoint junctions, examining 66,198 deletions and 75,238 insertions from our SAGA resource (Fig. 5a and Methods). In examining insertions by class, we find that VNTRs and tandem duplications present extensive breakpoint homology (49.7% and 89.8%, respectively), with the homologous sequences flanking the SV frequently mirroring the inserted element in length (Extended Data Fig. 7 and Supplementary Fig. 36). VNTRs typically form by processes such as replication slippage, HDR and NAHR, which involve DNA sequence homology⁴⁷. But similar to simple tandem duplications, the allelic structure of VNTRs generates homologies at the SV flanks independently of the mechanism of formation, thus necessitating separate consideration of these two SV classes when analysing breakpoint junctions

Leveraging the annotations provided by SVAN, our analysis of breakpoint junctions verifies that most MEs exhibit TSDs of 10–20 bp at the respective insertion site (Fig. 5b,c). When analysing SVs not annotated as VNTR, tandem duplication or MEI, 35.0% of deletions and 28.7% of insertions exhibit homologous flanks exceeding 50 bp, indicative of SV formation through HDR processes. A considerable subset thereof—10.8% of the deletions and 6.7% of the insertions—are flanked by more than or equal to 200 bp of homology, and are probably mediated by NAHR⁴⁸.

We further identify several clusters of SVs flanked by Alu, L1 and LTR elements annotated in the reference genome (Fig. 5b,c,e), the formation of which is likely to be mediated by HDR- and NAHR-driven transposable element-mediated rearrangement (TEMR)⁴⁹. Among these, we find Alu-flanked SVs to be much more common in deletions ($n = 3,260$) than in insertions ($n = 80$). This group of SVs predominantly harbours pairs of full-length Alu elements at their flanks (89.3%), visible as a breakpoint homology length peak of 295 bp (Fig. 5b). Notably, these presumably Alu element-mediated SVs show a wide distribution of SV lengths, ranging from about 300 bp up to 20.4 kb for deletions and 9.5 kb for insertions. AluY and AluX elements in all combinations constitute 23.5% of all SV-flanking Alu pairs. The use of the different Alu subfamilies at breakpoint junctions highly correlates with their counts in the reference genome (Supplementary Fig. 37)—with the exception of the AluJ family, one of the oldest Alu families, whose members appear several times less frequently at the flanks of SVs than expected by their reference genome count (for example, with a 24-fold reduction seen for AluJb).

We also find some L1-flanked SVs, which similar to Alu-flanked SVs are much more common in deletions ($n = 219$) than in insertions ($n = 1$). These L1–L1 pairs mediate relatively large SVs up to 62.9 kb in size (Fig. 5c). Pairs of L1PA family members are most often represented, partaking in 76.4% of L1–L1-mediated SVs. L1 elements at deletion flanks are typically truncated (median length: 1.7 kb), with only 5.0% of L1 sequence homology-mediated deletions flanked by full-length L1

elements. We additionally observe three distinct clusters corresponding to LTR-flanked deletions and insertions, reflecting SVs involving members of the HERVK (deletion: 39, insertion: 3), HERVH (deletion: 111, insertion: 9) and mammalian apparent LTR retrotransposon (deletion: 46, insertion: 4) families. Apart from 30 events subjected to truncation, the homology lengths correspond to the length of the LTRs flanking the respective integration site, with the SV length corresponding to the size of the viral integration plus one LTR, indicating LTR recombination⁵⁰. Determination of the ancestral state of SV alleles through sequence alignment to the chimpanzee genome confirms that the vast majority of TEMRs indeed represent deletions (Supplementary Note 7).

Moreover, we find 59 segmental duplication (SD)-mediated deletions (Fig. 5f), which are up to 31.8 kb in size and exhibit a homology length of up to 14.9 kb. We note that our approach is expected to exhibit decreased sensitivity in identifying SVs flanked by SDs of a similar length or larger than the N50 read length, suggesting that larger SD-mediated SVs are underrepresented.

We next analysed 30,449 deletions not categorized into a particular subclass by SVAN and lacking homology of at least 50 bp at their breakpoints. Of these, 2,812 (9.2%) exhibit blunt-ended breakpoint junctions whereas 25,311 (83.1%) exhibit microhomology of 1–15 bp in size, indicative of SV formation by non-homologous end-joining, alternative end-joining or microhomology-mediated break-induced replication⁵¹. The remaining 2,326 (7.6%) events show microhomology of 16–49 bp in size, and are likely to have formed through homology-independent processes as well. We further find that inverted duplications rarely show homology at their breakpoints (8.8%), with most exhibiting microhomology of 1–15 bp at their respective breakpoint junction, suggesting a replicative formation origin⁵¹ of this SV class.

Last, we analysed the distribution of SV breakpoint homologies with increasing SV allelic length. Notably, this distribution lacks appreciable inflection points, notwithstanding peaks in homology length corresponding to the sizes of mobile elements at the deletion flanks (Fig. 5c,d). This implies that homology-mediated SV formation processes use a wide variety of flanking repeat lengths, rather than reflecting distinctive peaks as might be expected from a minimal processing length for homologous recombination⁵².

Recurrence of TEMR-mediated deletions

Given the abundance of mobile element sequences in the genome sharing sequence homology, we postulated that occasionally SVs may have formed recurrently in humans, facilitated by the same flanking element pair. Although precise mutation rate estimates will require larger cohorts sequenced by LRS, we systematically screened for short- to intermediate-length homology-mediated deletions, for which the flanking homology length (200–9,000 bp) would be compatible with flanking mobile elements (that is, TEMR) or small- to intermediate-sized SDs, shortlisting 42 potentially recurrent candidates (Methods). We searched for SNPs for which haplotypes with all four combinations of both SNP alleles with the deletion being absent/present are observed (termed recurrence-indicating SNPs). On the basis of the presence of such SNPs on both sides of the deletion for further quality control, we selected six sites with robust evidence for recurrence, all of which represent TEMRs. One of these involves L2/LINE sequences, whereas the remaining five involve paired *Alu* elements (Fig. 5g, Extended Data Fig. 8 and Supplementary Figs. 38–43).

We illustrate this phenomenon for an 806-bp deletion at 12p13.3 mediated by an AluX–AluY pair (Fig. 5g). After clustering haplotypes using SNPs in a 100-kb window centred around the deletion (Methods and Extended Data Fig. 8), the four major emerging groups recapitulate the geographical ancestries—with Group 1 mostly comprising EAS samples, Groups 2 and 4 mostly AFR samples and Group 3 mostly SAS samples. The SV allele is present in different haplotype backgrounds of Groups 2 and 3, as well as in an outlier haplotype that clusters with Group 4 which

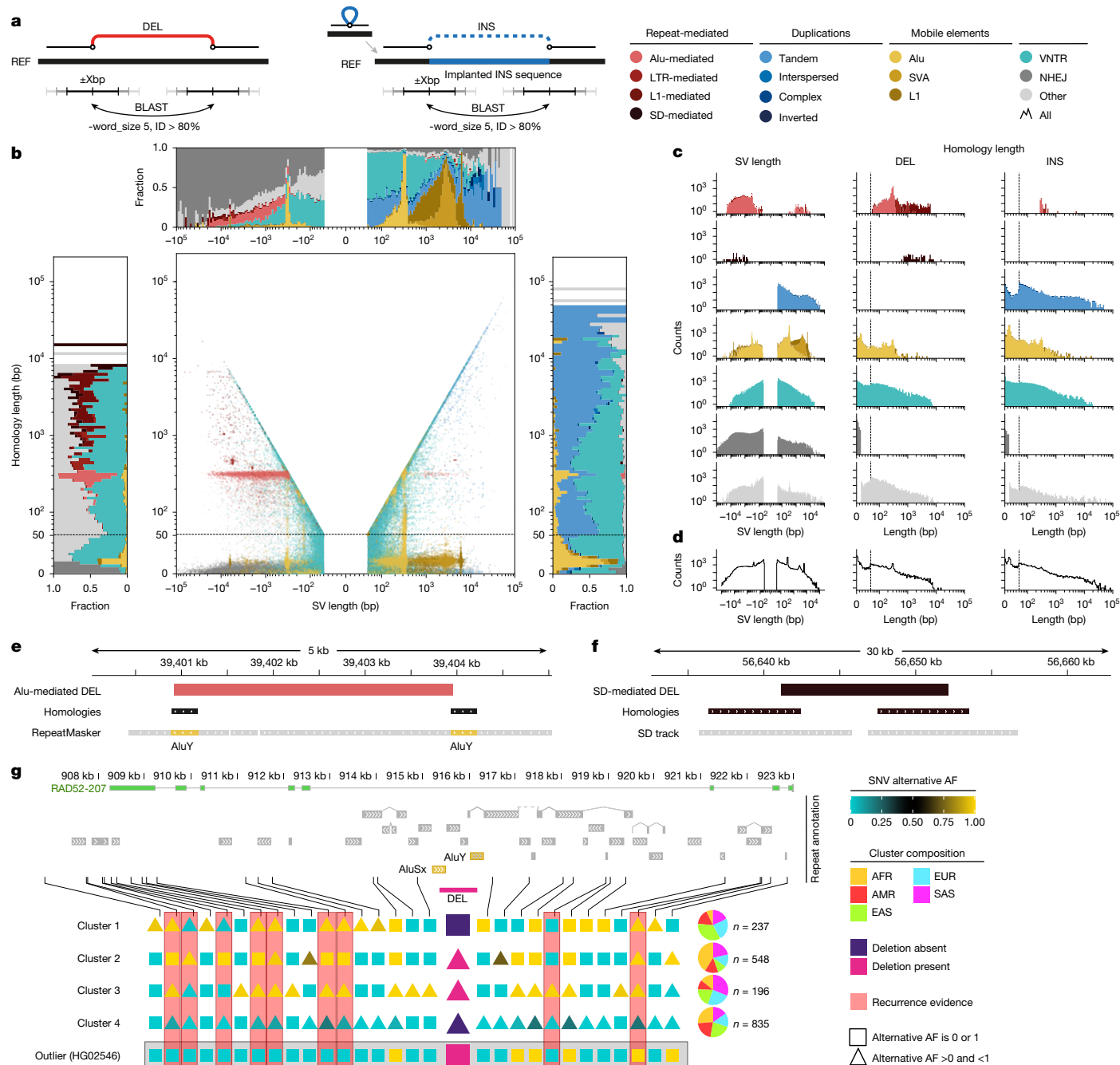


Fig. 5 | Breakpoint homology landscape and deletion recurrence. **a**, Approach for SV breakpoint junction analysis, in the case of primary insertions (INS) achieved by implanting SV sequences into CHM13 (denoted REF). DEL, primary deletion. **b**, SV length versus homology length (DELs depicted with negative length; INS with positive length). Marginal plots show the size-binned fraction of SV classes perpendicular to both axes, depicting DEL and INS at the left and right, respectively. SVs flanked by repeats are shown in different shades of red (mobile element) and black (SD). SVs exhibiting less than or equal to 15 bp of microhomology or blunt-ended breakpoints are in dark grey. Further colouring denotes: duplications (shades of blue), mobile elements (shades of yellow) and VNTRs (cyan). SVs not classified are in light grey. For visualization purposes, scales of scatter plot axes in **a** and x-axes in **b** are linear up to 50 bp (representing microhomology) and logarithmic afterwards (representing homology); this

split is denoted by the dashed line. **c**, SV and homology length distribution for distinct SV classes. **d**, Homology length trend lines for INS and DEL classes combined. **e, f**, Schematic view showing two repeat-mediated DELs, an Alu-mediated DEL (**e**) and an SD-mediated DEL (**f**). **g**, An 806-bp DEL at 12p13.3 mediated by an AluSx–AluY pair with evidence for recurrence. For visualization purposes, a consensus haplotype in a 20-kb window centred around the DEL is represented for each cluster. Clusters 1–4 were obtained from SNP-based clustering of the haplotypes in a 100-kb window centred around the DEL. Pie charts represent continental ancestries. Squares are used to represent an allele frequency of 0 (yellow) or 1 (cyan) within a cluster, whereas triangles represent allele frequency values in between. A haplotype from HG02546, grouping with cluster 4, is shown as an outlier. Red bars: SNPs indicating deletion recurrence. NHEJ, non-homologous end-joining.

is otherwise composed of haplotypes not carrying the deletion. We find nine recurrence-indicating SNPs (Fig. 5g, vertical red bars) characteristic of either recurrent deletions or extensive local recombination near the deletion. Given their proximity and association with specific haplotype

groups, the most plausible explanation is locally recurrent TEMRs, with scenarios based on homologous recombination on either side of the event seeming less likely. These data imply a relevant contribution of TEMRs to recurrent SV formation in the human genome.

Discussion

We show that intermediate-coverage LRS enables the generation of comprehensive population-scale SV catalogues comprising common and rare alleles. We predict our resource to serve as benchmarking data for tool development, supporting the continued development of graph-based genetic variant characterization methods^{53–55}.

Mobile elements are accountable for 1.32–3.14 Mb (median: 2.26 Mb) of the sequence variation detected per individual genome, with MEIs, TEMRs, transductions and processed pseudogenes contributing. This represents a considerable portion of the 4.78–9.88 Mb (median: 7.35 Mb) of sequence variation per genome attributed to SVs on the basis of the SAGA framework (Supplementary Table 21 and Supplementary Figs. 44 and 45). We show a subset of these have arisen recurrently in humans, opening up possibilities to study linkage disequilibrium patterns surrounding these events, and SV recurrence rates, in the future.

With respect to L1 and SVA transductions, whether the transduction occurs at the 5' or 3' end is influenced by mobile element class and locus. For instance, although L1s primarily mediate 3' transductions⁹, the L1 progenitor at 8q21.11 exclusively produces 5' transductions in our population-scale dataset and in the embryonic brain⁴⁶—a phenomenon probably driven by promoter hijacking⁴⁵. Similarly, the 13q12.3 source L1, previously reported to mediate 3' transductions in brain tissue⁴⁵, shows consistent behaviour in germline genomes, which implies shared mechanisms across cell types. A recent report suggests that strong polyadenylation sites downstream to source L1 loci can mediate 3' transduction in cancer⁵⁶. Altogether, these observations support the notion that transduction is influenced by locus-specific sequence determinants, such as regulatory elements.

Looking forward, we anticipate that our resource will support research on the phenotypic implications of SVs. A comparison with the Genome Aggregation Database (gnomAD)⁵⁷, a genetic variation repository with population-level allele frequencies, shows that most (50.9%) of the insertions from our resource were not previously represented in gnomAD. These include 8,077 Alu (34.8%), 2,586 L1 (52.9%) and 1,269 SVA element (47.1%) insertions. For deletions, 14.5% were not previously reported, reflecting the capabilities of short reads in deletion discovery^{3,4,23} (Supplementary Note 3). To explore the potential utility of our resource for variant prioritization, we analysed deep PacBio HiFi LRS data generated from four patients with rare diseases (Supplementary Note 8). Initial filtering using multi-platform whole-genome assemblies²⁰ left an average of 386 candidate SVs per patient. Incorporating our SAGA resource reduces these candidate SV lists by 54.7–56.4%, resulting in 159–187 candidate SVs remaining per patient (Extended Data Fig. 9 and Supplementary Figs. 46–48). We also evaluated the extent by which previously validated disease-causing SVs are filtered out through our resource, by analysing 31 published PacBio-sequenced rare disease genomes, comprising 35 validated causal SVs¹⁴. Of these validated causal SVs, only two were filtered out using our resource (Supplementary Note 9). These analyses show effective filtering of candidate SVs in patients, with causal SVs sensitively retained, using our LRS population-scale resource.

To further illustrate practical applications of our resource, we additionally explored the genotyping of medically relevant genes in complex loci of the genome. This was facilitated by the use of Locityper, a tool designed to address the challenges of accurately genotyping regions difficult to ascertain. We analysed 270 genes thought to represent challenging loci yet crucial for understanding genetic underpinnings of disease (Methods and Supplementary Table 22). Leveraging our resource, we find a high level of genotyping accuracy in the 1,019 samples of our study, particularly in regions fraught with structural complexity, considerably surpassing genotyping with short reads (Extended Data Fig. 10 and Supplementary Note 10).

The haplotype data derived from our ONT resource are also anticipated to facilitate long-read-based DNA methylation analyses

(Supplementary Fig. 45 and Supplementary Note 11), as well as to enhance variant imputation efforts, including in genomic areas previously difficult to analyse. To explore the latter aspect, a companion paper to this study demonstrates the usage of our ONT data resource for imputation and genome-wide association studies⁵⁸. In summary, these observations suggest broader value of our dataset for ascertaining disease-relevant genetic variation.

Although the data and computational methods developed in our study mark a step forward in cost-effective population-scale SV characterization, remaining challenges include increasing discovery power and sequence consensus quality for extremely rare SVs. Large inversions, centromeric regions, high sequence identity SDs and multiallelic VNTRs represent a particular challenge to graph-based approaches²¹ (Supplementary Fig. 12 and Supplementary Note 3). Resolving these complex regions comprehensively at population scale will necessitate expanding multi-platform assembly approaches²⁰ to more samples. Future improvements in graph algorithms should address needs for scalable analyses while considering all polymorphisms, including SNPs, indels and short tandem repeats.

As SAGA facilitates the integration of intermediate-coverage sequence data, we foresee broader adoption of our approach in large cohorts, which could accelerate advancements in population and disease research. Thereby, a coupled approach enhancing sample size at intermediate coverage and increasing sequencing depth on sample subsets seems particularly promising. With this in mind, our group has coordinated sample set design, as well as our public data release, with efforts piloting genome sequence assembly in smaller 1kGP sample subsets^{12,22,59}. This approach to open data sharing is guided by the principles of the 1kGP and is inspired by the potential to combine intermediate- and high-coverage techniques to advance the completion of the catalogue of human genomic variation encompassing the entire 1kGP cohort in the near future. Moving forward, it will be crucial to embed these findings into efforts aimed at understanding the genetic underpinnings of diseases, promoting equitable and global progress in precision medicine.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-025-09290-7>.

1. Spielmann, M., Lupiáñez, D. G. & Mundlos, S. Structural variation in the 3D genome. *Nat. Rev. Genet.* **19**, 453–467 (2018).
2. Weischenfeldt, J., Symmons, O., Spitz, F. & Korbel, J. O. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat. Rev. Genet.* **14**, 125–138 (2013).
3. Byrka-Bishop, M. et al. High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**, 3426–3440.e19 (2022).
4. Sudmant, P. H. et al. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
5. Rausch, T., Marschall, T. & Korbel, J. O. The impact of long-read sequencing on human population-scale genomics. *Genome Res.* **35**, 593–598 (2025).
6. Kasianowicz, J. J., Brandin, E., Branton, D. & Deamer, D. W. Characterization of individual polynucleotide molecules using a membrane channel. *Proc. Natl Acad. Sci. USA* **93**, 13770–13773 (1996).
7. Ren, J., Gu, B. & Chaisson, M. J. P. Varnish: variable-number tandem repeats annotation using efficient motif sets. *Genome Biol.* **24**, 175 (2023).
8. Damert, A. et al. 5'-Transducing SVA retrotransposon groups spread efficiently throughout the human genome. *Genome Res.* **19**, 1992–2008 (2009).
9. Goodier, J. L., Ostertag, E. M. & Kazazian, H. H. Jr. Transduction of 3'-flanking sequences is common in L1 retrotransposition. *Hum. Mol. Genet.* **9**, 653–657 (2000).
10. Kidd, J. M., Newman, T. L., Tuzun, E., Kaul, R. & Eichler, E. E. Population stratification of a common APOBEC gene deletion polymorphism. *PLoS Genet.* **3**, e63 (2007).
11. Middlebrooks, C. D. et al. Association of germline variants in the APOBEC3 region with cancer risk and enrichment with APOBEC-signature mutations in tumors. *Nat. Genet.* **48**, 1330–1338 (2016).
12. Liao, W.-W. et al. A draft human pangenome reference. *Nature* **617**, 312–324 (2023).
13. Noyes, M. D. et al. Familial long-read sequencing increases yield of de novo mutations. *Am. J. Hum. Genet.* **109**, 631–646 (2022).

14. Höps, W. et al. HiFi long-read genomes for difficult-to-detect, clinically relevant variants. *Am. J. Hum. Genet.* **112**, 450–456 (2025).
15. Beyter, D. et al. Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *Nat. Genet.* **53**, 779–786 (2021).
16. Gong, J. et al. Long-read sequencing of 945 Han individuals identifies structural variants associated with phenotypic diversity and disease susceptibility. *Nat. Commun.* **16**, 1494 (2025).
17. Zhao, X. et al. Expectations and blind spots for structural variation detection from long-read assemblies and short-read genome sequencing technologies. *Am. J. Hum. Genet.* **108**, 919–928 (2021).
18. 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
19. Lalli, J. L., Bortvin, A. N., McCoy, R. C. & Werling, D. M. A T2T-CHM13 recombination map and globally diverse haplotype reference panel improves phasing and imputation. Preprint at *bioRxiv* <https://doi.org/10.1101/2025.02.24.639687> (2025).
20. Logsdon, G. A. et al. Complex genetic variation in nearly complete human genomes. *Nature* <https://doi.org/10.1038/s41586-025-09140-6> (2025).
21. Li, H., Feng, X. & Chu, C. The design and construction of reference pangenome graphs with minigraph. *Genome Biol.* **21**, 265 (2020).
22. Ebert, P. et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, eabf7117 (2021).
23. Mills, R. E. et al. Mapping copy number variation by population-scale genome sequencing. *Nature* **470**, 59–65 (2011).
24. Chaisson, M. J. P. et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* **10**, 1784 (2019).
25. Read, L. R., Raynard, S. J., Rukšć, A. & Baker, M. D. Gene repeat expansion and contraction by spontaneous intrachromosomal homologous recombination in mammalian cells. *Nucleic Acids Res.* **32**, 1184–1196 (2004).
26. Wagner, J. et al. Curated variation benchmarks for challenging medically relevant autosomal genes. *Nat. Biotechnol.* **40**, 672–680 (2022).
27. Cui, Y. et al. A genome-wide spectrum of tandem repeat expansions in 338,963 humans. *Cell* **187**, 2336–2341 (2024).
28. Cost, G. J., Feng, Q., Jacquier, A. & Boeke, J. D. Human L1 element target-primed reverse transcription in vitro. *EMBO J.* **21**, 5899–5910 (2002).
29. Ostertag, E. M. & Kazazian, H. H. Jr. Twin priming: a proposed mechanism for the creation of inversions in L1 retrotransposition. *Genome Res.* **11**, 2059–2065 (2001).
30. Morrish, T. A. et al. DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. *Nat. Genet.* **31**, 159–165 (2002).
31. Srikanta, D. et al. An alternative pathway for Alu retrotransposition suggests a role in DNA double-strand break repair. *Genomics* **93**, 205–212 (2009).
32. Sen, S. K., Huang, C. T., Han, K. & Batzer, M. A. Endonuclease-independent insertion provides an alternative pathway for L1 retrotransposition in the human genome. *Nucleic Acids Res.* **35**, 3741–3751 (2007).
33. Cordaux, R. & Batzer, M. A. The impact of retrotransposons on human genome evolution. *Nat. Rev. Genet.* **10**, 691–703 (2009).
34. Zhou, W. et al. Identification and characterization of occult human-specific LINE-1 insertions using long-read sequencing technology. *Nucleic Acids Res.* **48**, 1146–1163 (2020).
35. Ewing, A. D. et al. Nanopore sequencing enables comprehensive transposable element epigenomic profiling. *Mol. Cell* **80**, 915–928.e5 (2020).
36. Porubsky, D. et al. Recurrent inversion polymorphisms in humans associate with genetic instability and genomic disorders. *Cell* **185**, 1986–2005.e26 (2022).
37. Kazazian, H. H. Jr. Processed pseudogene insertions in somatic cells. *Mob. DNA* **5**, 20 (2014).
38. Ewing, A. D. et al. Retrotransposition of gene transcripts leads to structural variation in mammalian genomes. *Genome Biol.* **14**, R22 (2013).
39. Wildschutte, J. H. et al. Discovery of unfixed endogenous retrovirus insertions in diverse human populations. *Proc. Natl Acad. Sci. USA* **113**, E2326–E2334 (2016).
40. Wei, W. et al. Nuclear-embedded mitochondrial DNA sequences in 66,083 human genomes. *Nature* **611**, 105–114 (2022).
41. Bailey, J. A. & Eichler, E. E. Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat. Rev. Genet.* **7**, 552–564 (2006).
42. Zhang, F. et al. The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. *Nat. Genet.* **41**, 849–853 (2009).
43. Beck, C. R., Garcia-Perez, J. L., Badge, R. M. & Moran, J. V. LINE-1 elements in structural variation and disease. *Annu. Rev. Genomics Hum. Genet.* **12**, 187–215 (2011).
44. Hanks, D. C. & Kazazian, H. H. Jr. SVA retrotransposons: evolution and genetic instability. *Semin. Cancer Biol.* **20**, 234–245 (2010).
45. Sanchez-Luque, F. J. et al. LINE-1 evasion of epigenetic repression in humans. *Mol. Cell* **75**, 590–604.e12 (2019).
46. Evrony, G. D. et al. Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell* **151**, 483–496 (2012).
47. Sulovari, A. et al. Human-specific tandem repeat expansion and differential gene expression during primate evolution. *Proc. Natl Acad. Sci. USA* **116**, 23243–23253 (2019).
48. Inoue, K. & Lupski, J. R. Molecular mechanisms for genomic disorders. *Annu. Rev. Genomics Hum. Genet.* **3**, 199–242 (2002).
49. Balachandran, P. et al. Transposable element-mediated rearrangements are prevalent in human genomes. *Nat. Commun.* **13**, 7115 (2022).
50. Hughes, J. F. & Coffin, J. M. Human endogenous retrovirus K solo-LTR formation and insertional polymorphisms: implications for human and viral evolution. *Proc. Natl Acad. Sci. USA* **101**, 1668–1672 (2004).
51. Liu, P., Carvalho, C. M. B., Hastings, P. J. & Lupski, J. R. Mechanisms for recurrent and complex human genomic rearrangements. *Curr. Opin. Genet. Dev.* **22**, 211–220 (2012).
52. Rubnitz, J. & Subramani, S. The minimum amount of homology required for homologous recombination in mammalian cells. *Mol. Cell. Biol.* **4**, 2253–2258 (1984).
53. The Computational Pan-Genomics Consortium et al. Computational pan-genomics: status, promises and challenges. *Brief. Bioinformatics* **19**, 118–135 (2016).
54. Wang, T. et al. The Human Pangenome Project: a global resource to map genomic diversity. *Nature* **604**, 437–446 (2022).
55. Eizenga, J. M. et al. Pangenome graphs. *Annu. Rev. Genomics Hum. Genet.* **21**, 139–162 (2020).
56. Zumalave, S. et al. Synchronous L1 retrotransposition events promote chromosomal crossover early in human tumorigenesis. Preprint at *bioRxiv* <https://doi.org/10.1101/2024.08.27.596794> (2024).
57. Gudmundsson, S. et al. Variant interpretation using population databases: lessons from gnomAD. *Hum. Mutat.* **43**, 1012–1030 (2022).
58. Noyvert, B. et al. Imputation of structural variants using a multi-ancestry long-read sequencing panel enables identification of disease associations. Preprint at *medRxiv* <https://doi.org/10.1101/2023.12.20.23300308> (2023).
59. Gustafson, J. A. et al. High-coverage nanopore sequencing of samples from the 1000 Genomes Project to build a comprehensive catalog of human genetic variation. *Genome Res.* **34**, 2061–2073 (2024).
60. Noguchi, S. et al. FANTOM5 CAGE profiles of human and mouse samples. *Sci. Data* **4**, 170112 (2017).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2025

Methods

Samples and DNA extraction

We initially selected 1,064 samples from the 1kGP collection, sourcing genomic DNA from Coriell and discarding low-quality DNA. The study complied with all relevant regulations for work with human participants. DNA was extracted from lymphoblastoid cell lines and resuspended in TE buffer (10 mM Tris, pH 8.0, 1 mM EDTA).

Quality control before sequencing

DNA concentrations were determined with the Quant-iT dsDNA Broad-Range assay Kit (Q33130), using Varioskan LUX multimode microplate reader (VL0000D0). The purity evaluation of genomic DNA was performed on a DeNovix DS-11 Series Spectrophotometer. Optical density 260/280 and 260/230 ratio of 1.8 and 2 was maintained, respectively, for sequencing. Fragment length was measured on the Femto Pulse system (Agilent, M5330AA) using the Genomic DNA 165 kb Ladder Fast Separation assay with a separation time of 70 min (Agilent, FP-1002-0275).

Size selection

All DNA samples were size selected using the Circulomics Short Read Eliminator Kit (Circulomics, SS-100-101-01). According to supplier information, the kit uses size-selective precipitation to reduce the amount of DNA fragments below 25 kb in length. The kit was used according to the manufacturer's recommendations (handbook v.2.0, 07/2019). Briefly, 60 µl of Buffer SRE was added to the sample tube (60-µl volume), gently mixed and the tube centrifuged at 10,000g for 30 min at room temperature. After supernatant removal, two washing steps were performed with 200 µl of 70% ethanol and a centrifugation at 10,000g for 2 min at room temperature. Finally, 50 µl Buffer EB was added and the tube was incubated at 50 °C for 30 min, followed by overnight incubation at room temperature to ensure efficient DNA elution.

LRS

Sequencing library preparation was carried out following the general guidelines from ONT, with modifications proposed by New England Biolabs (NEB) to ensure high-yield data generation and long-fragment sequencing. For library preparation, the following reagents were used: Ligation Sequencing Kit (ONT, SQK-LSK110), NEBNext Companion Module for ONT Ligation Sequencing (NEB, E7180S) and AMPure XP beads (made in-house by the Molecular Biology Service, Research Institute of Molecular Pathology). A DNA amount of 3 µg as input material was transferred into a 0.2-ml thin-walled PCR tube and the total volume was adjusted to 48 µl with nuclease-free water (Thermo Fisher, AM9937). DNA fragments were repaired and end-prepped as follows: 3.5 µl of NEBNext FFPE DNA Repair Buffer, 2 µl of NEBNext FFPE DNA Repair Mix, 3.5 µl of NEBNext Ultra II End Prep Reaction Buffer and 3 µl NEBNext Ultra II End Prep Enzyme Mix were added to each tube. After mixing and spinning down, the samples were incubated at 20 °C for 30 min, followed by a second incubation at 65 °C for 5 min. The prolonged incubation time allowed recovery of longer fragments. The solution from each tube was then transferred to a clean 1.5-ml Eppendorf DNA LoBind tube (Eppendorf) for clean-up. First, 60 µl of AMPure XP Beads was added to each tube. The samples were then incubated on a HulaMixer sample mixer (Thermo Fisher Scientific, 15920D) for 5 min at room temperature. Bead clean-up was performed with two washing steps on a magnetic rack, each time pipetting off the supernatant and adding 200 µl of freshly prepared 70% ethanol. The pellet was resuspended in 61 µl of nuclease-free water and incubated for 5 min at room temperature. Tubes were placed on a magnetic rack to collect the final eluate (1 µl was then taken out for quantification). For adaptor ligation and clean-up, 60 µl of DNA from the previous step was combined with 25 µl of Ligation Buffer, 10 µl of NEBNext Quick T4 DNA Ligase and 5 µl of Adapter Mix in a 1.5-ml Eppendorf DNA LoBind tube. The reaction was then incubated for 20 min at

room temperature. A second AMPure bead clean-up step was carried out by adding 40 µl of bead solution to each tube, followed by incubation on a HulaMixer for 5 min at room temperature. After pipetting off the supernatant on a magnet rack, the beads were washed twice with 250 µl of Long Fragment Buffer. Finally, the supernatant was discarded, and the pellet was resuspended in 25 µl of Elution Buffer EB and incubated for 10 min at 37 °C to collect the final library. Samples were quantified using a Qubit fluorimeter and diluted appropriately before loading onto the flow cells. The final mass loaded on the flow cells was determined on the basis of the molarity, dependent on average fragment size. Sequencing was carried out using FLO-PRO002 (R9.4.1) flow cells from ONT on the PromethION 48. The sequencing run was stopped after 24 h, the flow cell was washed using the Flow cell wash kit XL (ONT, EXP_WSH004-XL) and then the library was reloaded.

Base calling and adaptor trimming

Guppy v.6.2.1 was used for base calling the Fast5 input files in 'sup' accuracy mode with adaptor trimming and read splitting disabled and the output was converted to FASTA. For adaptor trimming and read splitting, we subsequently used Porechop⁶¹ v.0.2.4 on the generated FASTA files in chunks of 300,000 reads with default parameters.

Reference genome alignments

Reads were aligned to the GRCh38 (ref. 62) and CHM13 (ref. 63) linear reference genomes as well as the prebuilt human genome graph, denoted HPRC_mg, which was previously constructed by the HPRC¹² (from HPRC year-1 samples; <https://doi.org/10.5281/zenodo.6983934>). We selected HPRC_mg as the pangenomic reference as it represents SVs while omitting SNPs and less than 50-bp indels¹², thus comprising a compact graph structure facilitating analyses at the scale of a thousand long-read genomes. For the GRCh38 and CHM13, we used minimap2 (ref. 64) to map the ONT reads using the options '-a -x map-ont --rmq=yes --MD --cs -L'. Samtools⁶⁵ was used to sort the alignments and convert to CRAM. Multiple ONT runs for the same sample were tagged using different read-groups. Minigraph²¹ v.0.20-r559 was used to map the ONT reads against the HPRC_mg graph genome using the options '--vc -cx lr'. During alignment, the '--vc' flag enables the output of the alignments in vertex coordinates and the 'lr' option enables long-read mapping. The resulting graph alignments in GAF format²¹ were sorted using gaftools⁶⁶ and compressed using bgzip.

Sample and alignment quality control

Using SNP calls of the high-coverage short-read data from the 1kGP cohort³, we implemented rigorous quality control measures to effectively eliminate sample swaps and cross-contaminations. We first genotyped all SNPs from the short-read haplotype reference panel with an allele count greater than or equal to 6 in all samples using bcftools⁶⁷. Individual VCF files were merged using bcftools⁶⁷ into a multi-sample VCF file that was then combined with the short-read haplotype reference panel. We then used VCFtools⁶⁸ to calculate a relatedness statistic of the long-read sequenced samples compared with the short-read sequenced samples using the '--relatedness2' option. This analysis identified a sample swap between HG01951 and HG01983, which we relabelled afterwards. We excluded all samples that seemed to be cross-contaminated during library preparation, namely HG00138, HG02807, HG02813, HG02870, HG02888, HG02890, HG03804 and HG03778. Alignments were analysed using NanoPack⁶⁹ to determine median and N50 read length and genome coverage (Supplementary Figs. 2, 4 and 49). To compare linear and graph genome alignments in terms of percentage identity, number of aligned reads (bases) and largest Cigar I and D event, we used lorax⁷⁰ (Supplementary Fig. 3).

SV discovery using linear references

We used Sniffles⁷¹ v.2.0.7 and an LRS-optimized version of DELLY⁷² (v.1.1.7) to discover SVs using linear reference genomes. For Sniffles,

Article

we converted CRAM files to BAM format using samtools and then calculated for each sample candidate SVs and the associated SNF file using Sniffles. We then used Sniffles population-calling mode on all SNF files to generate two multi-sample VCF files, one for GRCh38 and one for CHM13.

Similarly, we also used DELLY's population-calling mode which first calls SVs by sample using the new long-read (lr) subcommand. We then merged all candidate SV sites using DELLY merge with the options '-p-a 0.05-v 3-c' to select PASS sites that are precise at single-nucleotide resolution with a minimum variant allele frequency of 5% and a minimum coverage of 3×. We then genotyped this SV site list in all samples using delly lr and merged the results by id with bcftools merge using the option '-m id'. We then applied sansa (<https://github.com/dellytools/sansa>), a newly developed, multi-sample SV annotation method that detects SV duplicates on the basis of SV allele and genotype concordance, to remove redundant SV sites. The parameters for the sansa markdup subcommand were '-y 0 -b 500 -s 0.5 -d 0.3 -c 0.1' to mark SV duplicates for sites that show an SV size ratio greater than 0.5, a maximum SV allele divergence of 30%, a maximum SV breakpoint offset of 500 bp and a minimum fraction of shared SV carriers of 10%. After removing duplicates, we generated the final multi-sample VCFs for GRCh38 and CHM13.

To ensure specificity and facilitate genome graph augmentation, we further generated for Sniffles and DELLY separately a consensus callset of SVs shared between the GRCh38 and CHM13 reference genomes. We therefore lifted the GRCh38 callsets to CHM13 using bedtools⁷³ and the liftOver tool⁷⁴ with the GRCh38 to CHM13 chain file. We then compared the lifted VCF with the original CHM13 VCF file to identify shared SVs using sansa's compvcf subcommand with the options '-m 0 -b 50 -s 0.8 -d 0.1' to identify SVs that have a size ratio greater than or equal to 0.8, a maximum SV breakpoint offset of 50 bp and an SV allele divergence of at most 10%. As the SV allele divergence filter requires a local assembly (for example, DELLY's consensus SV allele sequence), we did not apply this filter to Sniffles. For genome graph augmentation, we subset the final VCFs to deletions and insertions only. All inversion-type SVs called by DELLY and Sniffles using the minimap2 alignments were integrated separately in the inversion analysis.

Inversion analysis

We developed a multi-tiered analytical pipeline to comprehensively ascertain inversions on the basis of LRS data. By inspecting previously known inversions in our dataset, along with simulating a range of small inversions (less than 1 kb) with a coverage closely mirroring our dataset (median 17×), we discovered that minimap2 frequently misaligned reads in small inversion regions, leading to increasing error rates in those genomic locations. To allow capturing of these inversions, we examined strategies for inversion discovery by simulating inversions of varied sizes (Supplementary Note 6). We generated simulated ONT reads from these augmented genomes, using SURVIVOR⁷⁵, mimicking the sequencing coverage in our resource. Using pysamstats (<https://github.com/alimanfoo/pysamstats>), we first calculated the mismatch rate per base pair in 50-bp intervals, initially in the simulated datasets to tune parameters and subsequently in the real data to identify candidate inverted regions for re-alignment. This analytical process showed that most genomic regions maintained a mismatch rate below 20%; regions surpassing this rate were identified as having an unexpectedly high mismatch rate and selected for re-alignment with NGMLR⁷¹ (Supplementary Fig. 50). Post-exclusion of telomeric and centromeric regions, as well as of misaligned regions exceeding 1 kb in size, was then conducted to restrict the number of regions requiring remapping, thereby enhancing computational efficiency. The selected genomic segments underwent re-alignment using NGMLR and we then interrogated all realigned regions with DELLY⁷² to discover previously missed inversions. The final inversion calls in the remapped regions from all samples were consolidated using SURVIVOR merge⁷⁵. For the rest of the

regions not requiring re-alignment, inversion calling was conducted using both Sniffles and DELLY. Manual verification of true versus false positive calls was performed by examining dot plots and Integrative Genomics Viewer (IGV)-like plots generated with wally⁷⁰ for each candidate inversion location for the largest ten reads per candidate region, ensuring the accuracy of our findings. Ultimately, we generated a final comprehensive inversion callset by merging all unique instances from each dataset with 'bedtools merge' (v.2.31.1)⁷³. As the inversion analysis was conducted on the GRCh38 reference genome, regions were subsequently lifted over to the T2T-CHM13+Y genome, applying a 90% base remapping threshold to retain a region.

To evaluate our inversion dataset against two previous studies on 1kGP samples^{3,36}, we used 'bedtools intersect' (v.2.31.1)⁷³, defining inversions as 'known' if they exhibited a minimum of 50% reciprocal overlap with inversions from either previous dataset. Analysis against a previous study delineating inversions through whole-genome assembly³⁶ shows the efficacy of our methodology in detecting a diverse range of inversions, both repeat-mediated and non-repeat-mediated: our results showed that 65% of non-repeat-associated inversions and 41% of SD-mediated inversions were successfully identified. Furthermore, we refined our comparison to a 1kGP-derived short-read inversion dataset³, for which we included only those inversions from the comparison dataset with quality scores of 30 or higher, to ensure the accuracy of our comparative analysis. This approach showed an overlap of 289 inversions, or 36.5% concordance (median size of 530 bp).

Regarding the flanking repeats in repeat-rich inverted regions, we conducted a detailed analysis by manually inspecting the repeat types and their orientations at inversion breakpoints. Repeat data were acquired from the RepeatMasker track and the SD annotations of the CHM13 reference (obtained from <https://github.com/marbl/CHM13>); an inversion was classified as repeat-mediated if it was bracketed by repeats in reverse orientation relative to each other, detected through dotplot analysis.

Inversion genotyping was conducted using the GeONTIpe pipeline (commit: 1b5db07) (<https://github.com/RMoreiraP/GeONTIpe>), which identifies reads spanning inversion breakpoints and determines sequence order and orientation using probes positioned on both sides of the breakpoints. We focused on those inversions not genotyped by Giggles, by excluding inverted duplications and twin priming events. Multiple probe sets were tested, and for each inversion, a validated set of probes was generated and verified on dotplot-confirmed inversion carrier samples. Once the expected orientation was confirmed, the validated set of probes was applied to genotype the rest of the samples. All regions classified as 'inverted duplications' were excluded because of pipeline limitations, yielding a final dataset of 520 inversions. Of these, 407 inversions (78.3%) were successfully genotyped, with both haplotypes identified in 393 cases.

Notably, five inversions exclusively exhibited the inverted haplotype in all samples, suggesting either an inversion in the reference genome or a potential assembly error. In four cases, samples expected to have the inverted haplotype were classified as low confidence, with no other samples exhibiting the inversion. For a further five cases, manual analysis showed either a misclassification as an inverted duplication or the presence of complex structural rearrangements in the putative reads with the inversion.

Phasing with the ONT reads

To conduct phasing, we first pursued phasing experiments with the ONT reads to check how well they compare to the statistical phasing done previously in high-coverage 1kGP short-read data generated by the New York Genome Center (NYGC)³. The NYGC phased VCFs and the NYGC raw genotypes were used. Using the NYGC raw genotypes, the phasing was done by WhatsHap⁷⁶ (v.2.0) in three different ways: phasing with only the ONT reads (from hereon referred to as long-read phasing), trio phasing and trio phasing with the ONT reads (from hereon referred

to as long-read-trio phasing). The trio phasing and long-read-trio phasing was conducted for the six complete family trios (family IDs: 2418, CLM16, SH006, Y077, 1463 (paternal side), 1463 (maternal side)) for which our resource has long-read data. The long-read phasing was conducted for all of the 967 samples in the intersection of our 1,019 sample set and the NYGC sample set of 3,202 (Supplementary Fig. 7 and Supplementary Table 23). The phasing was pursued for all autosomes, and each chromosome was phased separately to allow parallel processing. The commands used are as follows:

- Long-read phasing: `whatshap phase -o <output phased vcf> --chromosome <chromosome ID> --sample <sample name> -r <reference fasta> <NYGC raw genotypes for sample> <ONT CRAM for sample>`
- Trio phasing: `whatshap phase -o <output phased vcf> --chromosome <chromosome ID> -r <reference fasta> --ped <pedigree data for the NYGC samples> <NYGC raw genotypes for samples in a family>`
- Long-read-trio phasing: `whatshap phase -o <output phased vcf> --chromosome <chromosome ID> -r <reference fasta> --ped <pedigree data for the NYGC samples> <NYGC raw genotypes for samples in a family> <ONT CRAM for samples in a family>`.

Comparison of the WhatsHap phased VCFs against NYGC statistical phasing

The phased VCFs produced by WhatsHap were compared against the NYGC statistical phasing using WhatsHap's `compare` function. The commands are as follows:

- For the samples without trio data: `whatshap compare --sample <sample name> --names longread,nygc --tsv-pairwise <pairwise tsv file> --tsv-multiway <multiway tsv name> <input longread phased vcf> <input nygc statistical phasing vcf>`
- For the samples with trio data: `compare --sample <sample name> --names trio,longread,trio-longread,nygc --tsv-pairwise <pairwise tsv file> --tsv-multiway <multiway tsv name> <input trio phased vcf> <input longread phased vcf> <input longread trio phased vcf> <input nygc statistical phasing vcf>`.

Haplotype tagging of ONT reads

The ONT reads were haplotype-tagged (or haplo-tagged) using WhatsHap⁷⁶ (v.2.0). The NYGC phased VCF³ was used as the reference for tagging the reads. The command used to tag the reads was: `whatshap haplotag --skip-missing-contigs --reference <reference fasta> --sample <sample name> --output-haplotag-list <output file> --output /dev/null <NYGC phased VCF> <ONT CRAM>`.

Although the main output of `whatshap haplotag` is a tagged alignment file, downstream tools used in this study required only a file containing the tag for each read which is given in `--output-haplotag-list`. Owing to the presence of pseudo-autosomal regions (PARs) in the phased VCF, the command was altered for the male samples. Instead of providing the entire NYGC phased VCF, the non-PAR records were removed and the haplo-tagging was performed. After haplo-tagging, the list of reads aligning to the non-PAR on chr. X were extracted, assigned manually as the maternal haplotype and added to the haplotag list.

SV discovery from the graph

The aim of SVarp⁷⁷ is to discover SVs on graph genomes, including for haplotypes missing in a linear reference. SVarp (commit: 0acba75) calls novel phased variant sequences, called svtigs, rather than a variant callset, which we later use in the graph augmentation step. To discover phased SV assemblies (svtigs) on top of the pangenome graph, we used haplotag read information and the alignment file (that is, GAF alignments) as input to the SVarp algorithm using `<svarp -a GAF-FILE -s 5 -d 500 -g GFA-FILE --fasta READS-FASTA-FILE -i SAMPLE_NAME --phase HAPLOTAG-FILE>` command. With 967 samples, we found a total of 1,108,850 variants (approximately 1,145 per sample).

To find specific SV breakpoint loci with respect to a linear reference genome, we used the PAV tool²² to call SV breakpoints, using svtigs that SVarp generated as input. This yielded 1,258,880 and 1,241,252 SVs relative to the CHM13 and GRCh38 linear genomes, respectively, that are more than 50 bp. However, we realized that some svtigs give rise to multiple SVs in the output of PAV. To ensure that variants called from the same svtig end up on the same pseudo-haplotypes in the graph augmentation step, we generated a script to combine records arising from the same svtig into a single VCF record. This is achieved by connecting multiple smaller such SVs into a single SV record through adding reference sequence in-between. This yielded 564,661 and 562,311 SVs relative to CHM13 and GRCh38, respectively.

The single-sample VCFs (relative to GRCh38) generated with PAV from the SVarp svtigs were merged into a multi-sample VCF using `bcftools merge`⁶⁷ (v.1.18) and then post-processed using `truvari collapse`⁷⁸ (v.4.1.0). The latter step merges SV records probably representing the same event into a single record, removing redundancy. This reduced the number of SVs from 451,942 to 215,209. Finally, we filtered the resulting VCF by keeping only records present in at least two samples. This filtered set contained 70,932 SVs.

Graph augmentation

We developed a pipeline to add extra variants found by DELLY, Sniffles and SVarp across the 1kGP ONT samples to the minigraph graph so that they can be genotyped by Giggles (v.1.0)⁷⁹. The main idea is to construct so-called 'pseudo-haplotypes' by implanting sets of non-overlapping variant calls into a reference genome and then adding them to the graph using minigraph²¹ (Extended Data Fig. 1). Our pipeline consists of the following steps. At first, we remove variant calls that fall into the centromere regions and mark the respective region in the reference genome by masking the sequence by Ns using the tool `bcftools maskfasta` (v.1.18). We used the GRCh38 reference genome and centromere annotations obtained from the UCSC genome browser. In the next step, we generate the pseudo-haplotypes as follows. Each of the SV discovery callsets (DELLY, Sniffles, SVarp) contains variants overlapping across samples. Thus, inserting all of them into one reference genome will fail. Therefore, we first group variants of each callset into sets of non-overlapping variants, and then generate a consensus sequence for each of these sets by implanting the variants into the reference genome using `bcftools consensus` (v.1.18). As a result, we obtain a whole-genome consensus sequence for each of these sets, which we call the pseudo-haplotypes. For the DELLY calls, we obtained 26 such pseudo-haplotypes, for Sniffles we got 69 and for SVarp we generated 117 pseudo-haplotypes. In total, these pseudo-haplotypes carry 154,319 DELLY SVs, 128,688 Sniffles SVs and 70,813 SVs detected by SVarp. In the last step, we insert all of these newly constructed genome sequences into the graph using the minigraph tool (v.0.20). Thereby, the minigraph algorithm incorporates a new SV allele only if it is sufficiently different (that is, shifted by more than or equal to 50 bp) from SV alleles already represented in the graph²¹. We first inserted all SVarp haplotypes, then all DELLY haplotypes and finally all Sniffles haplotypes into the graph using the command: `minigraph -cxggs -t32 <minigraph> <SVarp Pseudo-Haplotype FASTAs> <DELLY Pseudo-Haplotype FASTAs> <Sniffles Pseudo-Haplotype FASTAs> > augmented-graph.gfa`.

In the augmented graph, the number of bubbles (Supplementary Fig. 8) increases to 220,168 (102,371 in the original graph) and the total sequence represented in the graph increases from 3,297,884,175 bases to 3,477,266,061 bases. To identify bubbles in the augmented graph representing variation that was previously not represented in the original graph, we created BED files with coordinates of bubbles present in both graphs using `gfatools bubble`. Then we used `bedtools` closest to compute the distance between each bubble in the augmented graph and their respective closest bubble in the original graph (Supplementary Fig. 10). We defined all bubbles in the augmented graph

Article

whose distance to the closest bubble in the original graph is at least 1 kb as ‘new’ bubbles, representing novel SV sites.

To evaluate our augmented graph, we aligned ONT reads of human sample HG00513 (a sample not part of HPRC_mg) to HPRC_mg_44+966 using the command: `minigraph -cx lr -t24 augmented-graph.gfa HG00513-reads.fa --vc2> alignments.gaf`. For comparison, we aligned the same set of reads to the original graph (HPRC_mg) using the same command. We then computed alignment statistics using `gaftools stat` (Supplementary Table 3). We observed better alignment statistics when aligning reads to the augmented instead of the original graph. The number of aligned reads increased by 33,208, and the number of aligned bases by 152,454,715 bp.

Preparing phased VCF panel

We developed a pipeline that can reconstruct the alleles of the samples in the graph using the graph and the assemblies for the samples. First, the bubbles in the graphs are identified using `gaftools`⁶⁶ (commit ID: feaf7f4). The function `order_gfa` tags the nodes of the graph to identify whether the nodes are bubble nodes (nodes inside a bubble) or scaffold nodes (nodes outside a bubble). The phased panel is created by aligning the HPRC assemblies back to the tagged HPRC graph using `minigraph`²¹ (v.0.20-r559). The resulting alignments are processed to identify the alleles using the node paths between scaffold nodes. The allele information from the haplotype assemblies is then converted to a phased panel VCF. For the augmented graph, the same pipeline discussed above works where the pseudo-haplotypes are considered as assemblies and aligned to the augmented graph. On the basis of the tagging of the augmented graph, the alleles on the pseudo-haplotypes are identified and separate columns are created in the phased panel VCF corresponding to the alleles of the pseudo-haplotypes. This pipeline creates a multi-sample phased VCF, containing multiple alternative alleles for each record of the VCF.

The phased panel VCF is processed for its inclusion during Giggles⁷⁹ genotyping. Internal filter tags are set during the VCF creation step which allows for filtering of the bubbles for which allele information is not present for at least 80% of the samples. Additionally, bubbles are filtered out that do not have any alternative alleles in any of the haplotypes. From the phased panel VCF, a ‘Giggles-ready’ version is generated for which the reference alleles in the pseudo-haplotypes are masked with dots, because internally Giggles accounts for the allele frequency in the panels and the pseudo-haplotypes heavily bias the genotyping towards the reference used in the `bcftools` consensus step. The VCF records describe bubble structures in the graph. These bubbles often contain many nested and overlapping variant alleles, that is, a bubble does not necessarily represent a single variation event. To identify variant alleles represented in these bubbles, we applied the same bubble decomposition approach as previously described by the HPRC¹². In short, the idea is to compare node traversals of the reference and alternative paths through bubbles to identify nested alleles. As in the previous HPRC study¹², we then annotate the bubbles in our VCF to encode nested variants, so that we can translate genotypes computed for bubbles to genotypes for the variants represented in the graph.

Graph-aware genotyping

Graph-aware genotyping was conducted using Giggles⁷⁹ (v.1.0), a pangenome-based genome inference tool that leverages LRS data. Giggles serves as a long-read alternative to PanGenie’s short-read-based approach⁸⁰. It works with a Hidden Markov Model framework for which each variant position contains states corresponding to all possible genotypes at that position. The transition probabilities are based on the Li–Stephens model⁸¹ and the emission probabilities are based on the alignment of reads around an interval window of the variant position. Forward–backward computation of the Hidden Markov Model gives the posterior likelihoods of each state at each variant position which can be used to compute the likelihood of genotypes across the genome.

Giggles requires a phased VCF panel, sorted graph alignments, the input long reads, tagged graph and the haplotype tagging of the reads. We use the command: `giggles genotype --read-fasta <input reads> --sample <sample name> -o <output vcf> --rgfa <tagged gfa> --haplotag-tsv <haplotype tags> <phased panel VCF> <sorted alignments>`. This outputs an unphased VCF with the genotypes for the given sample.

The VCFs were further filtered using `bcftools`⁶⁷ to create a high-quality genotype set. This was achieved by masking genotypes of samples having a genotyping quality of less than 10 and dropping variant positions for which more than 5% of genotypes are missing. The commands are: `bcftools view --min-ac 1 -m2 -M2 <vcf> | bcftools +setGT --- -t q -n -i 'FMT/GQ<10' | bcftools +fill-tags --- -t all | bcftools filter -O z -o <filtered-vcf> -i 'INFO/AC>= 1 && INFO/F_MISSING <= 0.05'`.

We stratify the SVs as biallelic or multiallelic on the basis of the bubble these SVs originate from. If the bubble has more than two alleles defined in the phased VCF panel given to Giggles as an input, then the SV is defined to be multiallelic. Otherwise, if only one alternative allele is defined, the SV is defined to be biallelic.

Mendelian inconsistency statistics

For the six complete family trios in our dataset, we calculated Mendelian inconsistency statistics (Supplementary Fig. 51). We provide the confusion matrices and various statistics for each family, genotyped against HPRC_mg and HPRC_mg_44+966, and also report for various variant types depending on whether the variant came from a biallelic or multiallelic bubble (Supplementary Tables 5–12). The definitions for each statistic can be found in Supplementary Table 4.

Statistical phasing using high-coverage short-read data of 1kGP samples

Using SHAPEIT5 (ref. 82), we statistically phased the multi-sample VCF file outputted by Giggles using a recently constructed CHM13 haplotype reference panel¹⁹. This panel uses SNP and InDel calls generated from 1kGP short-read data^{3,83} for CHM13. We first subsetted the unphased input VCF from Giggles with 167,291 SVs to the 908 unrelated samples present in the haplotype reference panel, set genotypes with a quality less than 10 to ‘missing’ and subsequently dropped all SV sites with allele count zero. We then used the SHAPEIT5 common variant phasing mode to incorporate the new SV alleles into the SNP and InDel haplotype scaffold, yielding a phased VCF with 164,571 SVs. Before the phasing we split multiallelic variants into biallelic variants using `bcftools norm`⁶⁷ with the option ‘-m -any’. After phasing, we joined multiallelic variants back into the original state using `bcftools norm` with the option ‘-m +any’.

Using `plink`⁸⁴, we assessed linkage disequilibrium to nearby SNPs in a window of 1 Mb. We first converted the VCF to `plink` input files and then used `plink` to calculate r^2 values of all SVs to nearby SNPs. For the linkage disequilibrium analysis, we further subdivided the CHM13 genome into GiaB ‘difficult’ regions and ‘high-confidence’ regions⁸⁵ (using the CHM13_notinalldifficultregions.bed.gz file; Extended Data Fig. 3 and Supplementary Figs. 26 and 52).

Comparison with previous SV callsets

For the comparison with SV calls derived from high-quality genome assemblies from the HGSC we used `Truvari`⁷⁸, using the CHM13 reference genome (Supplementary Note 3). To enable further analyses with previous callsets, we additionally lifted our SV calls to GRCh38 using `liftOver`⁷⁴ with the CHM13 to GRCh38 chain file. As extra baseline callsets (GRCh38-based), we included the callsets produced in ref. 3, ref. 22 and ref. 4, labelled as `nygc`, `pangenie` and `phase3`, respectively (Supplementary Fig. 20). The comparison of SVs was restricted to autosomes and chromosome X for deletions and insertions separately. We used `sansa compvcf` to compare the SV VCF files using default parameters. We altered the base and comparison VCFs to identify SVs that are distinct

for one or the other callset and to identify potential 1:many SV overlaps among the shared SVs. We used the presence of INSSEQ and the absence of IMPRECISE in the INFO field as criteria for sequence-resolved insertions in previous short-read SV callsets³ (Fig. 1).

Population differentiation

We used ADMIXTURE v.1.3.0 (<https://github.com/NovembreLab/admixture>) to compute admixture for $K = 5$. For performing SV-based principal component analysis we used EIGENSOFT 8.0.0 (<https://github.com/DReichLab/EIG>). Fst values were calculated using VCFtools for each continental population (against the remaining samples) and each population (against the remaining samples) per site and for 1-Mb windows.

SV impact estimation on genomic features

We used Ensembl VEP with annotation from the CHM13 rapid release of Ensembl (107) to estimate the impact of the SVs on genomic features. Processing was performed using the command: `vep --assembly T2T-CHM13v2.0 --regulatory --offline -i final-vcf.unphased.vcf`.

We observe a large difference in the mean allele frequencies of SVs affecting coding (mean MAF 0.009) and non-coding (mean MAF 0.061) regions. This observation and the P value of 0.0 reported by both t -test and Kolmogorov–Smirnov test thereby support the hypothesis that the two underlying distributions are different.

Targeted genotyping of challenging loci

Locityper⁸⁶ (v.0.13.3) is a targeted whole-genome sequencing-based genotyper designed for challenging loci. Our initial set of target genes was preprocessed using locityper add -e 300k. Our analysis based on Locityper focused on 270 loci identified by GiaB as challenging, yet crucial for understanding genetic underpinnings of various diseases²⁶, along with 20 polymorphic *MUC* family genes and the *LPA* gene, which are of particular interest because of their association with various structural haplotypes and health conditions^{87,88}. ONT datasets were preprocessed and genotyped with locityper preprocess --tech ONT and locityper genotype, respectively. A database of locus haplotypes was constructed on the basis of the GRCh38 and CHM13 reference genomes, as well as on the basis of the 44 diploid HPRC whole-genome assemblies. To evaluate genotyping accuracy, we aligned haplotypes to each other using the Wavefront alignment algorithm⁸⁹, and calculated sequence divergence as a ratio between edit distance and alignment size. Local haplotypes were extracted from the NYGC callset using bcftools consensus.

SV annotation with SVAN

SV calls generated with our SAGA framework were processed with the SVAN tool (commit: 7b97325, v.1.3) (<https://github.com/REPBIO-LAB/SVAN>) to classify them in distinct classes on the basis of distinctive sequence features:

Retrotranspositions. Canonical retrotransposition events were identified by searching for poly(A) and poly(T) tails at the 5′ end and 3′ ends of the deleted and inserted sequences. Poly(A/T) tracts were required to be at least 10 bp in size, have a minimum purity of 90% and be at a maximum of 50 bp from the insert end or beginning, for poly(A) and poly(T), respectively. The sequence corresponding to the TSD was detected and trimmed from either the 5′ or 3′ end of the L1 insert through the identification of exact matches with the genomic sequence at the integration position. To have all candidate retrotransposed inserts in forward orientation, the reverse complement sequence for every trimmed insert occurring in the minus strand was obtained.

Candidate 3′ partnered transductions were detected by searching for a second poly(A) stretch at the 3′ end of the trimmed sequences using the same criteria outlined above. Integrants with a secondary tail were annotated as ‘3′ transduction’ candidates with the sequence in-between the poly(A) stretches corresponding to the transduced bit of DNA,

whereas those with a single tail were classified as ‘solo’ candidates. To trace transductions to their source loci, transduced sequences were aligned onto CHM13 using BWA-MEM⁹⁰. To maximize sensitivity for particularly short transduction events, a minimum seed length (-k) of 8 bp and a minimum score (-T) of 0 were used. Alignment hits were filtered by requiring a minimum mapping quality of 10. Transduction candidates with less than 80% of the transduced sequences aligning on the reference were further filtered out.

The retrotransposition candidates were further trimmed by removing the poly(A) tails and transduced sequences. To identify processed pseudogenes, orphan transductions, 5′ partnered transductions and retroelement sequences, the resulting trimmed inserts were aligned onto CHM13 using minimap 2.1, as well as BWA-MEM 0.7.17-r1188, by using the parameters described above and with the alignment conducted onto a database of consensus L1, Alu and SVA repeats. The SVA repeats database contains separate sequences for each SVA component, including the Alu-like region, VNTR, SINE-R and exon 1 of the *MAST2* gene, which is characteristic of the SVA-F1 subfamily.

Alignment hits were filtered by requiring a minimum mapping quality of 10, and chained on the basis of complementarity to identify the minimum set of non-overlapping alignments that span the maximum percentage of the trimmed insert. On the basis of the alignment chains, retrotranspositions are classified as processed pseudogenes (more than or equal to 75% aligns on the reference over single or multiple Gencode 35 annotated exons⁹¹), orphan transductions (more than or equal to 75% aligns on the reference outside exons), 5′ transductions (more than or equal to 75% aligns both into the reference on its 5′ and into the consensus of a retrotransposon) and solo (more than or equal to 75% aligns into a single consensus retrotransposable element). Alignment chains are further analysed to decompose each MEI into its sequence components, including inversions at the 5′ end of L1 and processed pseudogene insertions or the set of individual repeats composing SVA inserts.

Non-canonical MEIs. Insertions and deletions without a poly(A) tail detected were aligned into a database of consensus L1, Alu and SVA repeats as described above to identify non-canonical MEIs. Alignment hits were filtered and chained and non-canonical MEI calls made if more than or equal to 75% of the sequence aligned into one or multiple repeat classes.

Endogenous retroviruses. Inserted and deleted sequences were aligned with BWA-MEM 0.7.17-r1188 into a database containing consensus ERV and LTR sequences. Alignment hits were filtered and chained as described above. SVs with hits spanning more than or equal to 75% of their sequence on the retrovirus database were classified as Solo-LTR and ERVK, on the basis of the presence of an LTR alone or LTR plus retroviral sequence, respectively.

VNTRs. Expansions and contractions of VNTR loci were annotated by processing inserted and deleted sequences with TRF (Tandem Repeats Finder) v.4.04 (ref. 92). TRF hits were processed and chained on the basis of complementarity to determine the minimum set of non-overlapping hits that span the maximum fraction of the sequence. On the basis of the hit chains, SVAN classified VNTRs as simple (more than or equal to 75% of the sequence corresponds to a single repeated motif) and complex (more than or equal to 75% corresponds to more than one repeated motif).

Duplications. Diverse classes of duplications, including tandem, inverted and complex duplication events, were annotated by realigning the inserted sequences onto the reference using minimap 2.1. Alignments were filtered to select only those located within a 2-kb window around the insertion breakpoint. These were further chained on the basis of complementarity to determine the minimum set of non-overlapping

Article

hits spanning the maximum percentage of the sequence. On the basis of the alignment chains, duplications are classified as tandem (more than or equal to 75% of the sequence aligns at the insertion breakpoint in forward orientation), inverted (more than or equal to 75% of the sequence aligns at the insertion breakpoint in reverse orientation) and complex (more than or equal to 75% of the sequence aligns at the insertion breakpoint both in forward and reverse orientation).

NUMTs. Insertions with more than or equal to 75% of their sequence having one or multiple minimap 2.1 alignments on the mitochondrial reference are classified as NUMT.

Transduction analysis

L1 and SVA transductions were clustered on the basis of the alignment position on the reference of their corresponding transduced sequences to identify their source regions. A buffer of 10 kb was applied to their start and end alignment positions before clustering. Source regions were further intersected with a database of full-length loci to determine the progenitor repeat. This database was generated by aggregating all the full-length L1s (more than 5.9 kb in size) and SVAs (more than 1 kb) in CHM13 detected as non-reference insertions in our SV callset. Progenitor repeats were further annotated by determining their insertion orientation and subfamily. For L1 events, subfamily assignment was performed through the identification of subfamily diagnostic nucleotide positions on their 3' end. L1 progenitors bearing the diagnostic 'ACG' or 'ACA' triplet at the 5,929–5,931 position were classified as 'pre-Ta' and 'Ta', respectively. Ta elements were subclassified into 'Ta-0' or 'Ta-1' according to diagnostic bases at the 5,535 and 5,538 positions (Ta-0: G and C; Ta-1: T and G). Elements that did not show any of these diagnostic profiles could not be assigned to a particular category, and their subfamily status remained undetermined. For SVAs, the inserts were processed with RepeatMasker v.4.0.7 to determine the subfamily. If multiple RepeatMasker hits were obtained, the one with the highest Smith–Waterman score was selected as representative. To assess a bias in source elements to generate 5' or 3' transduction events, respectively, we conducted two-tailed binomial tests followed by controlling for the FDR according to Benjamini and Hochberg⁹³. We included all source elements with at least five non-orphan transductions when conducting statistical testing for 5' versus 3' bias.

To investigate source element novelty, we compared active source L1s and SVAs identified in this study with source elements previously reported to be active on the basis of transduction traces seen in germline^{22,94} as well as cancer genomes^{95,96}, and additionally compared our dataset with previously reported *in vitro* activity measurements^{97,98}.

SV breakpoint junction analysis

The detection of homologous sequences and microhomologies flanking deletions and insertions was conducted on the primary callset, after removing calls less than 50 bp in length, using two approaches: Microhomology was quantified using the available homology output from SV calls generated by DELLY⁷² (v.1.2.6). We systematically generated DELLY calls for each SV by building synthetic reads carrying the respective SV allele, mapping those synthetic reads with minimap2 (ref. 64) (v.2.2.26) onto the genomic reference of the SV with 4-kb padding and calling the respective SV with DELLY. The synthetic reads consist of a 2-kb reference sequence upstream of the SV start, the inserted sequence (only for insertions and modified for larger insertions) and a 2-kb reference sequence downstream of the SV end coordinate. For insertions longer than 100 bp, we used at most the first and last 50 bp of the inserted sequence to avoid aberrant mapping of the insert, which was especially observed in the case of long insertions. Owing to the use of truncated inserts for insertions, only homologies with a length of 50 bp were considered, and larger values were set to 50 bp. The second approach aimed to capture longer stretches of homology using BLAST⁹⁹-based detection of homologous stretches

of DNA. For this, various pairs of search windows around both breakpoints were defined. Search windows were either symmetric with a padding of 50 bp, 100 bp, 200 bp, 400 bp, 1 kb, 2 kb, 5 kb, 10 kb, 50 kb and 100 kb, or asymmetric with a window size of 'SV length', which is shifted regularly along the breakpoint (0 bp upstream/'SV length' bp downstream; 1/6 'SV length' upstream/5/6 'SV length' downstream; ...; 'SV length' upstream/0 bp downstream). From the predefined windows, only those leading to no overlap between windows were used. Potential homologies were detected using blastn 2.12.0 with -perc_identity 80 and -word_size 5, for which the sequences inside the search windows were passed using the -subject and -query parameters. To allow for equivalent analyses between deletions and insertions, insertions were implanted into the CHM13 reference genome and the search windows were defined on the modified reference sequence. BLAST results were filtered to ensure flanking homology stretches show the same directionality, span the respective SV breakpoints and contain the respective breakpoint at the relative same position inside the homology segment.

For annotation of repeat-mediated SVs, the overlap of the homologous sequences with RepeatMasker (v.4.1.2) annotations was used. For deletions, overlap of the homologous regions with the elements of the RepeatMasker track of the CHM13 reference^{12,22,63} (obtained from <https://github.com/marbl/CHM13>) was calculated with bedtools⁷³ (v.2.31.1) intersect. Deletions were classified as repeat-mediated if at least 85% of the bases of the homologous regions on both sides intersect a RepeatMasker annotation of the same class and at least one homologous region spans 85% of the RepeatMasker element. For insertions, the homologous sequences outputted by blastn were annotated with RepeatMasker and insertions for which both homologous flanks show a reciprocal overlap of at least 85% with a RepeatMasker annotation of the same class were deemed repeat-mediated. To analyse SD-mediated deletions, all deletions were intersected with the SD track of CHM13. Hits spanning more than 85% of the homologous region and spanning more than 200 bases or 50% of the SD were defined as putatively SD-mediated.

VNTR genotyping using vamos

We performed VNTR genotyping on the long-read data from our resource using vamos (v.2.1.3)⁷. We used the VNTR motif annotations provided by the authors of vamos on the CHM13 reference¹⁰⁰ to genotype VNTRs on all the samples of our dataset. We performed quality control on the VNTR calls by comparing the calls produced from these ONT reads with the VNTR calls produced from recently generated multi-platform whole-genome assemblies²⁰, focusing on the 16 overlapping sampling between the two callsets (Supplementary Fig. 7). For each VNTR position described by the motif annotations, we compared the number of repeat units that were present in the VNTR allele called using vamos. As the VNTRs called on our dataset are read-based and the HGSC dataset²⁰ is genome assembly-based, the VNTR alleles of our dataset are unphased whereas the HGSC VNTRs are phased. For allelic comparison, we paired the ONT 1kGP-based alleles with the HGSC alleles having the smallest difference in repeat unit counts. We removed the VNTR alleles for which the HGSC assemblies showed the VNTR to be close to the reference VNTR (when the VNTR allele from the assemblies had a base pair length between 90% and 110% of the reference VNTR allele). We compared the rest of the VNTR alleles to assess the concordance between calls on the basis of our resource and the HGSC assembly-based calls. We also conducted a comparison for mismatched samples, verifying that random VNTR matching does not produce similar results.

Deletion recurrence analysis

To detect potential NAHR-mediated deletion recurrence, we analysed deletions shorter than 5 kb in length exhibiting a flanking sequence homology of 200–9,000 bp. The lower bound of 200 bp

was introduced to account for the minimal processing length of flanking repeat sequences previously thought to be necessary for NAHR to occur^{52,101}. The upper bound of 9,000 bp was chosen to systematically assess the role of the most abundant repeat elements in the human genome, including small- to intermediate-sized SDs as well as mobile elements, which we consider to be ideal candidate sites for SV recurrence owing to their pervasive presence as repeats in the genome providing substrates for NAHR. Additionally, to focus the analysis on potentially recurrent events, we limited our analysis to deletions with allele frequency of 40–60%. Furthermore, we used Hardy–Weinberg equilibrium and Mendelian consistency statistics to exclude events with potential genotyping errors. Last, we filtered out deletions for which phasing information was unavailable. Using the above-mentioned criteria, we retained 42 deletions for our analysis. To find evidence of recurrence, we used an approach similar to the approach we previously devised to detect recurrent inversions in the genome³⁶. We screened for SNPs with at least 10% allele frequency lying within a 20-kb window centred around the deletion in search for positions for which we observe SNP alleles in haplotypes both with and without the deletion—which we denote recurrence-indicating SNPs. Additionally, we used centroid hierarchical clustering to cluster the SNP haplotypes in a 100-kb window centred around the deletion, in an effort to assess whether the haplotypes with and without the deletion appear together in similar clusters, a phenomenon that indicates that the event recurred in humans. For all 42 shortlisted potentially recurrent candidates, we manually inspected these visualizations of SNP haplotypes as well as read alignments from carriers of deletion and reference alleles, respectively, to confirm that genotyping was accurate, that the clustering signal was not driven by recombination and that recurrence-indicating SNPs were present on both sides of the deletion. We selected six candidates with the strongest evidence for SV recurrence upon inspection and included the visualizations as Supplementary Figs. 38–43 and as Extended Data Fig. 8.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Our open data resource is fully available to the community for download at the IGSR. Data access FTP link: https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1KG_ONT_VIENNA/. The data have additionally been made available at the Amazon Web Services (AWS) cloud computing platform at s3://1kg-ont-vienna and at the European Nucleotide Archive (ENA) under accession PRJEB89727. These repositories contain the graph and linear reference genomes, alignments, input SV callsets, the augmented genome graph, genotyped and phased SVs, as well as auxiliary data used for evaluating SV callset accuracy. Raw and basecalled sequencing data are also available via the OpnMe initiative of Boehringer Ingelheim: <https://opnme.com/genomiclens>. Further sequencing data are available from the European Genome-Phenome Archive under accession EGAS00001008170.

Code availability

General information about this project is available at <https://github.com/1kg-ont-vienna/sv-analysis>. A pipeline to create a multi-sample SVarp VCF from single-sample SVarp calls (used as one of the callsets going into graph augmentation) and the graph augmentation pipeline is available at <https://github.com/eblerjana/long-read-1kg> (commit: 44a1752). The Giggles genotyping algorithm is available at <https://github.com/samarendra-pani/giggles> (commit: 5226884). The pipelines and scripts available at <https://github.com/marschall-lab/project-ont-1kg> (commit: 32d896d) are: haplo-tagging of aligned

reads, phasing experiments with WhatsHap, SV genotyping with Giggles on HPRC_mg and HPRC_mg_44_966, SV discovery with SVarp, chimpanzee ancestral allele annotation, VNTR genotyping with vamos and some pipelines for quality assessment. The repository also hosts analysis scripts for the genotypes and the SVAN annotations. Curation work for VNTR-associated diseases has also been provided. The sequence-to-graph alignment sorting and processing of GAF files was done using gaftools. Gaftools is available at <https://github.com/marschall-lab/gaftools> (commit: feaf7f4). The Structural Variants ANnotator (SVAN) tool is available at <https://github.com/REPPIO-LAB/SVAN> (commit: 7b97325, v.1.3). DELLY pangenome filtering mode is available at <https://github.com/dellytools/delly> (commit: 4984ff2) and sansa for SV comparison at <https://github.com/dellytools/sansa> (commit: 198be12). The pangenome-based SV discovery tool SVarp is available at <https://github.com/asylvz/SVarp> (commit: 0acba75). Scripts for finding and annotating homologous flanks at SV breakpoints are available at https://github.com/carstenhain/SV_homology (commit: 534b216). A pipeline to simulate ONT data on a genome with small inversions for benchmarking genome mappers and callers is available at https://github.com/celiatsapalou/Simulations_ONT_Data (commit: cf0513b). A pipeline to export high-mismatch regions, perform remapping and carry out calling inversions is available at https://github.com/celiatsapalou/Small_Inversions_Remap (commit: f9d8a44). The pipeline to genotype inversions using ONT reads is available at <https://github.com/RMoreiraP/GeONTIpe> (commit: 1b5db07). The pipeline for filtering rare disease-associated SVs using our resource is available at <https://github.com/hugocarmaga/rare-disease-analysis> (commit: add69e9). A pipeline producing structural variation VCF files from HiFi reads is available at <https://github.com/hugocarmaga/variant-calling> (commit: 5062a67). DNA methylome analysis scripts are available at https://github.com/santanaw/1kGP_mods (commit: 8c7ba71).

61. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Completing bacterial genome assemblies with multiplex MinION sequencing. *Microb. Genom.* **3**, e000132 (2017).
62. Schneider, V. A. et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* **27**, 849–864 (2017).
63. Nurk, S. et al. The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
64. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
65. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
66. Pani, S., Dabbaghie, F., Marschall, T. & Söylev, A. Gaftools: a toolkit for analyzing and manipulating pangenome alignments. Preprint at *medRxiv* <https://doi.org/10.1101/2024.12.10.627813> (2024).
67. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
68. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
69. De Coster, W., D’Hert, S., Schultz, D. T., Cruts, M. & Van Broeckhoven, C. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* **34**, 2666–2669 (2018).
70. Rausch, T. et al. Long-read sequencing of diagnosis and post-therapy medulloblastoma reveals complex rearrangement patterns and epigenetic signatures. *Cell Genom.* **3**, 100281 (2023).
71. Sedlazeck, F. J. et al. Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).
72. Rausch, T. et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).
73. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
74. Kent, W. J. et al. The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
75. Jeffares, D. C. et al. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat. Commun.* **8**, 14061 (2017).
76. Martin, M. et al. WhatsHap: fast and accurate read-based phasing. Preprint at *bioRxiv* <https://doi.org/10.1101/085050> (2016).
77. Söylev, A. et al. SVarp: pangenome-based structural variant discovery. Preprint at *bioRxiv* <https://doi.org/10.1101/2024.02.18.580171> (2024).
78. English, A. C., Menon, V. K., Gibbs, R. A., Metcalf, G. A. & Sedlazeck, F. J. Truvari: refined structural variant comparison preserves allelic diversity. *Genome Biol.* **23**, 271 (2022).
79. Pani, S. Giggles. *GitHub* <https://github.com/samarendra-pani/giggles> (2024).
80. Ebler, J. et al. Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. *Nat. Genet.* **54**, 518–525 (2022).
81. Li, N. & Stephens, M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**, 2213–2233 (2003).

82. Hofmeister, R. J., Ribeiro, D. M., Rubinacci, S. & Delaneau, O. Accurate rare variant phasing of whole-genome and whole-exome sequencing data in the UK Biobank. *Nat. Genet.* **55**, 1243–1249 (2023).

83. Rhie, A. et al. The complete sequence of a human Y chromosome. *Nature* **621**, 344–354 (2023).

84. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

85. Dwarshuis, N. et al. The GIAB genomic stratifications resource for human reference genomes. *Nat. Commun.* **15**, 9029 (2024).

86. Prodanov, T. et al. Locityper: targeted genotyping of complex polymorphic genes. Preprint at *bioRxiv* <https://doi.org/10.1101/2024.05.03.592358> (2024).

87. Ringel, J. & Löhr, M. The MUC gene family: their role in diagnosis and early detection of pancreatic cancer. *Mol. Cancer* **2**, 9 (2003).

88. Schmidt, K., Noureen, A., Kronenberg, F. & Utermann, G. Structure, function, and genetics of lipoprotein (a). *J. Lipid Res.* **57**, 1339–1359 (2016).

89. Marco-Sola, S., Moure, J. C., Moreto, M. & Espinosa, A. Fast gap-affine pairwise alignment using the wavefront algorithm. *Bioinformatics* **37**, 456–463 (2021).

90. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

91. Frankish, A. et al. GENCODE: reference annotation for the human and mouse genomes in 2023. *Nucleic Acids Res.* **51**, D942–D949 (2023).

92. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).

93. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).

94. Gardner, E. J. et al. The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. *Genome Res.* **27**, 1916–1929 (2017).

95. Tubio, J. M. C. et al. Mobile DNA in cancer. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. *Science* **345**, 1251343 (2014).

96. Rodriguez-Martin, B. et al. Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition. *Nat. Genet.* **52**, 306–319 (2020).

97. Brouha, B. et al. Hot L1s account for the bulk of retrotransposition in the human population. *Proc. Natl Acad. Sci. USA* **100**, 5280–5285 (2003).

98. Beck, C. R. et al. LINE-1 retrotransposition activity in human genomes. *Cell* **141**, 1159–1170 (2010).

99. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

100. Gu, B. & Chaisson, M. TRcompDB v1.0, a global reference of tandem repeat variation. *Zenodo* <https://doi.org/10.5281/zenodo.13263615> (2024).

101. Waldman, A. S. & Liskay, R. M. Dependence of intrachromosomal recombination in mammalian cells on uninterrupted homology. *Mol. Cell. Biol.* **8**, 5350–5357 (1988).

102. De Roeck, A. et al. An intronic VNTR affects splicing of ABCA7 and increases risk of Alzheimer’s disease. *Acta Neuropathol.* **135**, 827–837 (2018).

103. Ruggieri, A. et al. Multiomic elucidation of a coding 99-mer repeat-expansion skeletal muscle disease. *Acta Neuropathol.* **140**, 231–235 (2020).

104. Willemssen, R., Levenga, J. & Oostra, B. A. CGG repeat in the FMR1 gene: size matters. *Clin. Genet.* **80**, 214–225 (2011).

Acknowledgements We thank all participants of the 1000 Genomes Project, without whom this project would not have been possible. We thank M. Zody, X. Zhao and other members of the HGSC for valuable feedback, P. Ebert for assistance with data management, and J. Charest and the Vienna BioCenter Core Facilities for assistance with long-read DNA sequencing. We thank P. Contzen for preparing DNA for long-read sequencing of rare disease samples. We thank L. Vissers and A. Hoischen for generously allowing us to use published patient data and providing oversight for their interpretation. Moreover, we acknowledge the EMBL IT services, the Centre for Information and Media Technology at Heinrich Heine University Düsseldorf, the

IT services at the IMP, as well as the CRG Core Technologies Programme for providing resources for data processing and analysis. Finally, we thank members of the International Genome Sample Resource (IGSR) for assistance with providing the open data releases for this study. Funding for sequence data production was provided by the MARVL initiative, a collaboration between the IMP, BI X and Boehringer Ingelheim. Additional funding came from the following sources: National Institutes of Health (NIH) (to J.O.K., T.M., grant no. U24HG007497), the Ministry of Culture and Science of the State of North Rhine-Westphalia (to T.M., grant no. PROFILNRW-2020-107-A), the German Research Foundation (to T.M., grant no. 525152594) and the GraphGenomes project funded by the BMBF (to T.M., grant no. 031L0184A, and J.O.K., grant no. 031L0184C). W.H. received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 101150006. B.R.-M. was supported by a Bridging Excellence Fellowship provided by the Life Science Alliance. The project also received the support of a fellowship from the ‘la Caixa’ Foundation (ID 100010434), with fellowship code ‘LCF/BQ/DI24/12070028’ (to J.E.S.-F.). Additional funding came from the EU Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 713673. We acknowledge support of the Spanish Ministry of Science and Innovation through the Centro de Excelencia Severo Ochoa (grant no. CEX2020-001049-S, MCIN/AEI /10.13039/501100011033), and the Generalitat de Catalunya through the CERCA programme. We received support with data management from the EMBL throughout this project. We thank the DFG Research Infrastructure West German Genome Center (grant no. 407493903) as part of the Next Generation Sequencing Competence Network (project 423957469) for the LRS support.

Author contributions M.A. and S. Schloissnig were responsible for data acquisition; P.H. and T.R. were responsible for base calling, read alignment and primary SV discovery versus linear references; S.P. performed read-based haplotype phasing, haplotype tagging of reads, SV genotyping, VNTR calling and chimpanzee ancestral allele annotation; T.R. performed statistical phasing of the SV genotypes; A.S. performed graph-based SV discovery with SVarp; J.E. was responsible for implementation and execution of the graph augmentation pipeline; S.P. and T.R. provided the final SV callset construction and SV callset benchmarking; S. Schloissnig performed the analysis of geographic stratification of SV alleles; C.H. performed breakpoint homology analysis; B.R.-M. and J.E.S.-F. performed SV class annotation and mobile element analysis; H.A. performed the deletion recurrence analysis; T.P. performed the complex locus genotyping; T.F., W.S.-G. and E.B. performed explorative analyses of DNA methylation; V.T., R.M.-P. and M.C. performed the inversion analysis; S.H. provided data management assistance; T.E.W. and F.J.P.-L. performed PacBio sequencing of patients with rare diseases; H.M., W.H., S. Sivalingam, D.W., C.G. and T.M. analysed the genomes of patients with rare diseases; S. Schloissnig, Z.D., J.N.J., N.P. and J.S. provided data production planning; T.R., T.M. and J.O.K. jointly supervised the project; H.A., C.H., S.P., B.R.-M., T.M., T.R., S. Schloissnig, V.T. and J.E.S.-F. prepared the main figures; H.A., J.E., C.H., J.O.K., T.M., S.P., T.P., B.R.-M., A.S., S. Schloissnig, V.T. and T.R. wrote the manuscript, with input from all authors.

Funding Open access funding provided by Heinrich-Heine-Universität Düsseldorf.

Competing interests Z.D., J.N.J. and N.P. are employees of Boehringer Ingelheim Pharma GmbH & Co. KG. J.S. is an employee of BI X GmbH. E.B. is a consultant for and shareholder of Oxford Nanopore Technologies. The other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41586-025-09290-7>.

Correspondence and requests for materials should be addressed to Bernardo Rodriguez-Martin, Tobias Rausch, Tobias Marschall or Jan O. Korbel.

Peer review information *Nature* thanks Geoffrey Faulkner, Alexander Hoischen and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at <http://www.nature.com/reprints>.