

# PopHuman: the human population genomics browser

Sònia Casillas<sup>1,\*</sup>, Roger Mulet<sup>1,†</sup>, Pablo Villegas-Mirón<sup>2</sup>, Sergi Hervás<sup>1</sup>, Esteve Sanz<sup>3</sup>, Daniel Velasco<sup>1</sup>, Jaume Bertranpetit<sup>2</sup>, Hafid Laayouni<sup>2,4</sup> and Antonio Barbadilla<sup>1,3,\*</sup>

<sup>1</sup>Institut de Biotecnologia i de Biomedicina and Department de Genètica i de Microbiologia, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain, <sup>2</sup>Institute of Evolutionary Biology (UPF-CSIC), Universitat Pompeu Fabra, Doctor Aiguader 88 (PRBB), 08003 Barcelona, Catalonia, Spain, <sup>3</sup>Servei de Genòmica i Bioinformàtica, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain and <sup>4</sup>Bioinformatics Studies, ESCI-UPF, Pg. Pujades 1, 08003 Barcelona, Spain

Received August 11, 2017; Revised September 18, 2017; Editorial Decision October 02, 2017; Accepted October 04, 2017

## ABSTRACT

**The 1000 Genomes Project (1000GP) represents the most comprehensive world-wide nucleotide variation data set so far in humans, providing the sequencing and analysis of 2504 genomes from 26 populations and reporting >84 million variants. The availability of this sequence data provides the human lineage with an invaluable resource for population genomics studies, allowing the testing of molecular population genetics hypotheses and eventually the understanding of the evolutionary dynamics of genetic variation in human populations. Here we present PopHuman, a new population genomics-oriented genome browser based on JBrowse that allows the interactive visualization and retrieval of an extensive inventory of population genetics metrics. Efficient and reliable parameter estimates have been computed using a novel pipeline that faces the unique features and limitations of the 1000GP data, and include a battery of nucleotide variation measures, divergence and linkage disequilibrium parameters, as well as different tests of neutrality, estimated in non-overlapping windows along the chromosomes and in annotated genes for all 26 populations of the 1000GP. PopHuman is open and freely available at <http://pophuman.uab.cat>.**

## INTRODUCTION

Soon after the elucidation of the entire human genome (1–3), the description of genetic variation in human populations and the identification of those variants that affect health and disease became the next challenges of genomics research (4). The International HapMap Consortium built the first genome-wide catalog of common human

genetic variation in diverse populations (4–6), charting haplotype maps of 1.6 million single nucleotide polymorphisms (SNPs) in 1184 reference individuals from 11 global populations. In addition to numerous genome-wide association studies (GWAS) (7), the HapMap data allowed the detection of positive natural selection across the human genome (8,9), as well as the development of new tests to infer recent episodes of selective sweeps based on the length of haplotypes, such as the Long-Range Haplotype (LRH) (10), the integrated Haplotype Score (iHS) (11), and the Cross Population Extended Haplotype Homozygosity (XP-EHH) (8).

During the last decade, the development of next generation sequencing (NGS) technologies (12,13) has allowed the deciphering of complete genome sequences of thousands of human individuals, and the 1000 Genomes Project (1000GP) has become the reference data set for population genetics and genomics (14,15). With the aim of providing a deep characterization of human genome sequence variation, the most recent version of the 1000GP (Phase III) completes the sequencing and analysis of 2504 genomes from 26 populations and describes most variants with frequencies as low as 1%. Due to its higher resolution and smaller SNP ascertainment bias compared to HapMap genotyping data, the availability of the 1000GP data provides the human lineage with an invaluable resource on which to test molecular population genetics hypotheses and eventually understand the evolutionary dynamics of genetic variation in human populations (16).

Regions of the genome that are (or have been) subject to natural selection show distinctive patterns of genetic variation in the DNA sequence (17). The signature of long-range haplotypes persists for a relatively short period of time (<30 000 years), and related statistics can detect very recent selection only. However, other signatures persist longer in the genome: differentiation between populations (<50 000–<75 000 years), high frequency derived alleles (<80 000 years), reduction in genetic diversity and excess of rare al-

\*To whom correspondence should be addressed. Sònia Casillas. Tel: +34 93 5868958; Fax: +34 93 5812011; Email: [sonia.casillas@uab.cat](mailto:sonia.casillas@uab.cat)

Correspondence may also be addressed to Antonio Barbadilla. Email: [antonio.barbadilla@uab.cat](mailto:antonio.barbadilla@uab.cat)

†These authors contributed equally to this work as first authors.

Present address: Roger Mulet, Department of Hematology, Erasmus University Medical Center, Rotterdam, The Netherlands.

leles (<250 000 years), and high proportion of function-altering substitutions between species (many millions of years) (17).

Population genomics analyses of the 1000GP data set can be largely facilitated by (i) making an inventory of parameter values along the chromosomes that capture the evolutionary properties of the available sequences, and (ii) allowing the query and visualization of these estimates in a genome browser designed specifically for this data. As far as we are concerned, the 1000 Genomes Selection Browser 1.0 (18) is the only previous database that allows the interactive visualization and retrieval of population genetics metrics for the 1000GP data. It was published when the 1000GP was still in its first phase (1,092 individuals, 14 populations, 38 million SNPs) (14), and analyzed within-species polymorphism data for three populations in 30 kb sliding windows (18). Here, we present PopHuman, a new population genomics-oriented genome browser. PopHuman represents not only an update to the 1000GP Phase III (2504 individuals, 26 populations, 84.7 million SNPs), but also dramatic improvements in the amount of data analyzed and browser performance, compared to the 1000 Genomes Selection Browser 1.0. Furthermore, PopHuman analyzes between-species divergence, which allows the implementation of statistical tests to detect the signature of recurrent natural selection acting over prolonged periods of time, such as the McDonald and Kreitman test (MKT) (19), instead of recent selective sweeps only. Supplementary Table S1 details the differences between the two databases.

## POPHUMAN ANALYSIS PIPELINE

We have designed and implemented a custom pipeline (Figure 1) facing the unique features and limitations of the 1000GP Phase III data (15). The pipeline discards reportedly inbred individuals (20) and non-accessible nucleotides (15), incorporates the genomic sequence of the chimpanzee (21) as outgroup, and estimates a battery of nucleotide variation, divergence and linkage disequilibrium parameters, as well as different tests of neutrality, on the filtered data. Several metrics have been computed both in non-overlapping sliding windows along the chromosomes and in annotated protein coding genes for 26 populations of distinct geographical origin (15).

### Pre-processing of the 1000GP Phase III data

We retrieved human genome variation data generated by the 1000GP Phase III (15) from <http://www.internationalgenome.org/data> in Variant Call Format (VCF). This included 84.4 million variants detected across 2504 individuals from 26 different populations, mapped to the human reference genome version GRCh37/hg19. We want to warn the user that four of the analyzed populations present admixture (corresponding to the Admixed American metapopulation), so special care should be taken while interpreting PopHuman results in those cases.

*Inbred individuals.* The initial VCF files were filtered to exclude 243 individuals with inbreeding coefficients similar or

greater than the ones expected for first-cousin offspring, according to Gazal *et al.* (20).

*Genome accessibility mask.* Due to the nature of short-read sequencing, sequencing depth varies along the length of the genome. The 1000GP provides an ‘accessibility mask’, a Browser Extensible Data (BED) file that indicates which sites of the genome were accessible to the sequencing techniques and have power for variant discovery (15). Two definitions were used in the Phase III, of which we selected the ‘pilot-style’ mask. This definition is less conservative than the ‘strict’ mask while being still adequate for population genomics analyses, and was chosen to maximize the amount of genomic sequence to be analyzed. It excludes the portion of the genome where depth of coverage (summed across all samples) was higher or lower than the average depth by a factor of 2-fold, as well as sites where >20% of overlapping reads had mapping quality of zero. Overall, 89.4% of the genome is considered reliable (95.9% of the non-N bases). Specifically, we placed 10 kb non-overlapping sliding windows in accessible regions of the genome (i.e. windows do not overlap any non-accessible nucleotide) to focus on high quality genomic regions only. Table 1 summarizes the total amount of data analyzed by PopHuman by following this methodology. In addition, we analyzed longer non-overlapping sliding windows of 100 kb placed all along the genome (i.e. windows might overlap non-accessible nucleotides, although these positions were discarded for the population genomics analyses) to focus on broader scale patterns of diversity across the genome.

*Ancestral states.* The ancestral states of human segregating sites were taken from the 1000GP Phase III (15), which were obtained by using the 6-way EPO alignments available in Ensembl v71 (22).

*Outgroup species.* To compute divergence metrics and neutrality tests based on the comparison of polymorphism and divergence, we added differences between humans and chimpanzees to the VCF files, as identified from a pre-computed hg19 => panTro4 alignment obtained from the VISTA browser (23) in multi-FASTA format (MFA). Specifically, the pairwise alignment was converted to VCF using custom scripts and merged with the 1000GP VCF files using *bcftools merge*.

### Recombination

The most recent human genetic sex-specific maps were obtained from Bhérier *et al.* (24), based on a total of 104 246 informative meioses from six recent studies of human pedigrees.

### Estimation of population genomics statistics

*Windows-based.* Several windows-based variation statistics and tests of neutrality (Table 2) were computed for each population separately using the R package PopGenome (25) and custom functions, considering biallelic SNPs as within-species variation data. Haplotype-based statistics (iHS and XP-EHH) were computed in a multithreaded



**Figure 1.** PopHuman pipeline. Cited references in the figure: <sup>1</sup>1000GP Phase III (15); <sup>2</sup>Inbred individuals in the 1000GP (20); <sup>3</sup>VISTA Genome Browser (23); <sup>4</sup>Human genetic maps (24); <sup>5</sup>PopGenome software (25); <sup>6</sup>UCSC Genome Browser (35); <sup>7</sup>JBrowse software (34).

**Table 1.** Summary of the amount of data analyzed in PopHuman

Chromosome		Windows-based analysis		Genes-based analysis	
Chromosome number	Chromosome size (millions of bases) <sup>a</sup>	Number of windows <sup>b</sup>	Number of bases (millions)	Percentage of analyzed bases	Number of RefSeq <sup>c</sup> genes analyzed
1	249.25	14 741	147.41	59.14	2328
2	243.20	16 270	162.70	66.90	1464
3	198.02	13 575	135.75	68.55	1274
4	191.15	12 512	125.12	65.45	879
5	180.92	12073	120.73	66.73	1022
6	171.12	11 433	114.33	66.81	1206
7	159.14	9919	99.19	62.33	1108
8	146.36	9783	97.83	66.84	818
9	141.21	7358	73.58	52.11	944
10	135.53	8760	87.60	64.63	903
11	135.01	8877	88.77	65.75	1439
12	133.85	8773	87.73	65.54	1175
13	115.17	6481	64.81	56.27	449
14	107.35	5948	59.48	55.41	779
15	102.53	5334	53.34	52.02	791
16	90.35	4688	46.88	51.88	938
17	81.20	4556	45.56	56.11	1358
18	78.08	5164	51.64	66.14	341
19	59.13	2681	26.81	45.34	1609
20	63.03	4091	40.91	64.91	647
21	48.13	2211	22.11	45.94	296
22	51.30	2009	20.09	39.16	535
X	155.27	9312	93.12	59.97	918
Y	59.37	622	6.22	10.48	53
<b>TOTAL</b>	<b>3095.68</b>	<b>187 171</b>	<b>1871.71</b>	<b>60.46</b>	<b>23 274</b>

<sup>a</sup>Chromosome sizes are according to version GRCh37/hg19 of the human genome.

<sup>b</sup>Non-overlapping sliding windows of 10 kb have been defined such that they do not include non-accessible bases according to the Pilot-style Accessibility Mask of the 1000GP (15).

<sup>c</sup>RefSeq genes provided by the NCBI Entrez Gene database (33).

framework implemented by the program *selscan* (26), considering biallelic SNPs with Minor Allele Frequency (MAF) > 0.05 and a maximum gap of 20 kb between two consecutive SNPs. Then, whole chromosome per-SNP scores were summarized by calculating the mean of the absolute value of these scores for all SNPs in a window (27). Sexual chromosomes were not analyzed in these cases.

*Genes-based.* Comparisons of DNA polymorphism within populations and divergence to an outgroup species using the MKT (19) have been extensively used to detect the signature of natural selection at the molecular level (28). The MKT can be generalized to any two types of sites provided that one of them is assumed to evolve neutrally and that both types of sites are closely linked in the genome (29–31). Furthermore, Mackay *et al.* (32) developed an integrative new framework for the MKT by incorporating information on the MAF of the segregating sites, which allows estimating the fraction of new mutations that are strongly deleterious (and therefore not segregating), slightly deleterious (segregating at low frequency), old neutral (neutral before the split of humans and chimpanzees), and recently neutral (since the split of humans and chimpanzees), as well as the fraction of adaptive fixations. The standard and integrative MKTs (Table 3) were applied to all annotated human protein coding genes in RefSeq (33) and for different types of sites (i.e. 0-fold nonsynonymous coding sites, 5'UTR, 3'UTR, introns, and  $\pm 500$  bp intergenic flanking regions, compared to 4-fold synonymous coding sites), for each population separately, using custom functions build within PopGenome (25).

## OVERVIEW OF THE POPHUMAN GENOME BROWSER

PopHuman is a new population genomics-oriented genome browser based on JBrowse (34) that allows the interactive visualization and retrieval of several metrics estimated in non-overlapping sliding windows along the chromosomes and in annotated genes for all 26 populations of the 1000GP. It also includes a number of utilities and support resources.

### JBrowse implementation

PopHuman is built on JBrowse (34) and is currently running under Apache on a CentOS 7.2 Linux x64 server with 16 Intel Xeon 2.4 GHz processors and 32 GB RAM.

### Browser tracks

*Variation statistics.* Windows-based variation statistics and tests of neutrality (Table 2) are classified into: (i) frequency-based nucleotide variation; (ii) divergence-based metrics; (iii) linkage disequilibrium; (iv) recombination; (v) selection tests based on the Site Frequency Spectrum (SFS) and/or variability and (vi) selection tests based on the MKT. They are displayed for each population separately as histogram plots, with a yellow line showing the mean, and two shaded bands showing  $\pm 1$  and  $\pm 2$  standard deviations from the mean. Visualization style can be customized using the 'Edit config' option for each track.

*Reference tracks.* Several tracks have been imported from the UCSC Genome Browser (35) (Supplementary Table S2) and can be visualized along with variation statistics. They are classified into: (i) sequencing and annotation; (ii) regulation; (iii) comparative genomics; (iv) variation and (v) repeats.

### Utilities and support resources

*Tracks selector.* PopHuman contains more than a thousand tracks, including both variation statistics (Table 2) and reference tracks (Supplementary Table S2). Given the large number of tracks available, these can be filtered and selected using the 'Select tracks' tool, which can be accessed from the top left corner, below the navigation bar. The filtering process is normally performed by first narrowing the search using the menu on the left, and then selecting the tracks of interest from the main panel on the right. This process can be done several times in order to finally get all the desired tracks selected.

*Downloading raw data.* Variation statistics for a given region can be conveniently downloaded in bedGraph, Wiggle or GFF3 formats using the 'Save track data' option for each track. In addition, bulk downloads of full variation tracks are available in BigWig format from the Resources menu. Finally, variant calls for the analyzed individuals can also be downloaded in VCF format using the PopHuman utility 'Download sequences', which can be accessed either from the Resources menu, or directly from the navigation bar.

*Integrative MKT.* Gene-based MKTs (Table 3) can be retrieved by right-clicking a gene and selecting the option 'Integrative MKT'.

*Help section.* The Help section contains exhaustive documentation about the 1000GP Phase III data analyzed by PopHuman and details about the browser tracks. Interestingly, it contains a comprehensive tutorial introducing to the usage of the database and to the testing of evolutionary hypotheses from a population genetics perspective. The tutorial works out, in different sequential steps, the visualization and analysis of a genomic region of around 20 kb in chromosome 7 that includes the *TRPV6* gene. *TRPV6* is a well-studied protein coding gene involved in the absorption of calcium from the diet that has experienced parallel selective sweeps in non-African populations, coinciding with the establishment of agriculture first in Europe around 10 000 years ago, and later in Asia. The tutorial contains several step-by-step guides to facilitate reproducing the results that are shown both in the form of figures and descriptive text.

### Availability

All data, tools and support resources provided by PopHuman, as well as reference tracks downloaded from the UCSC Genome Browser (35), are open and freely available at <http://pophuman.uab.cat>.

## COMPARISON TO OTHER DATABASES

While the PopHuman analysis pipeline presented here is completely novel, the genome browser is based on a similar

**Table 2.** List of major windows-based variation statistics and tests of neutrality in PopHuman, computed for each population separately

Category	Track name	Track description	Reference
<b>Frequency-based nucleotide variation</b>	S	Number of segregating sites per site	(42)
	Pi	Nucleotide diversity: average number of nucleotide differences per site between any two sequences	(42–44)
	theta	Nucleotide polymorphism: proportion of nucleotide sites that are expected to be polymorphic in any suitable sample	(45–47)
<b>Divergence-based metrics</b>	hap_diversity_within	Haplotype diversity within the population	(48)
	Divsites	Number of divergent sites	
	K	Nucleotide divergence per base pair, corrected by Jukes-Cantor	(43)
<b>Linkage disequilibrium</b>	Kelly_ZnS	Average pairwise $r^2$ value	(49)
	Rozas_ZA	Average of $r^2$ only between adjacent polymorphic sites	(50)
	Rozas_ZZ	Rozas_ZA minus Kelly_ZnS	(50)
	Wall_B; Wall_Q	Proportion of pairs of adjacent segregating sites that are congruent, with values approaching 1 indicating extensive congruence among adjacent segregating sites	(51)
	iHS	Integrated haplotype score, based on the frequency of alleles in regions of high LD (computed for the autosomes)	(11)
	XP_EHH	Long-range haplotype method to detect recent selective sweeps (computed for the autosomes, between the major continental populations CEU, CHB and YRI, taken in pairs)	(8)
<b>Recombination</b>	recomb_Bherer2017_females/males/sexavg	Recombination estimates (cM/Mb) from the refined genetic map by Bh�erer <i>et al.</i> (2017), which collects recombination events from six recent studies of human pedigrees, pertaining to a total of 104 246 informative meioses. Maps are available in three separate tracks: females, males and sexavg	(24)
	recomb_deCODE_females/males/sexavg	deCODE genetic map based on 5136 microsatellite markers for 146 families with a total of 1257 meiotic events.	(52)
	recomb_Marshfield_females/males/sexavg	Marshfield genetic map based on 8325 short tandem repeat polymorphisms (STRPs) for 8 CEPH families consisting of 134 individuals with 186 meioses.	(53)
	recomb_Genethon_females/males/sexavg	Genethon genetic map based on 5264 microsatellites for 8 CEPH families consisting of 134 individuals with 186 meioses.	(54)
<b>Selection tests based on SFS and/or variability</b>	FayWu_H	Number of derived nucleotide variants at low and high frequencies with the number of variants at intermediate frequencies	(55)
	FuLi_D	Number of derived nucleotide variants observed only once in a sample with the total number of derived nucleotide variants	(29)
	FuLi_F	Number of derived nucleotide variants observed only once in a sample with the mean pairwise difference between sequences	(29)
	Tajima_D	Difference between the number of segregating sites and the average number of nucleotide differences.	(56)
	Zeng_E	Difference between $\theta_L$ and $\theta_W$ , sensitive to changes in high-frequency variants.	(57)
<b>Selection tests based on the MKT</b>	DoS	Direction of Selection: difference between the proportion of nonsynonymous divergence and nonsynonymous polymorphism	(58)
	NI	Neutrality Index: summarizes the four values in a McDonald and Kreitman test table as a ratio of ratios	(19,59)
	alpha; alpha_cor	Proportion of substitutions that are adaptive. The second is calculated after removing slightly deleterious mutations	(19,32,60,61)

A complete list is available under the section Help → Tracks Description of PopHuman.

**Table 3.** List of major gene-based variation statistics in PopHuman, computed for each population separately and for different types of sites

Category	Estimate	Description	Reference	Types of sites analyzed
<b>Descriptive statistics</b>	$\pi$	Nucleotide diversity: average number of nucleotide differences per site between any two sequences	(42–44)	Whole gene region $\pm 500$ bp
	K	Nucleotide divergence per base pair, corrected by Jukes-Cantor	(43)	
	$\pi_a/\pi_s$	Ratio of nonsynonymous to synonymous nucleotide polymorphism ( $\omega$ )	(44,62)	Ratio: 0-fold divided by 4-fold
	$K_a/K_s$	Ratio of nonsynonymous to synonymous nucleotide divergence ( $\omega$ )	(44,62)	
	DAF	Derived Allele Frequency: distribution of allele frequencies of segregating sites	(63)	Whole gene region $\pm 500$ bp
<b>Recombination (Bh�erer et al. 2017), cM/Mb Standard MKT</b>	cM/Mb	Recombination estimates (cM/Mb) from the refined genetic map by Bh�erer et al. 2017	(24)	Whole gene region $\pm 500$ bp
	P	Number of segregating sites	(42)	Separately: 4-fold; 0-fold; 5'UTR; 3'UTR; intron; intergenic ( $\pm 500$ bp)
	D	Number of divergent sites		
	$\pi$	Nucleotide diversity: average number of nucleotide differences per site between any two sequences	(42–44)	
<b>Integrative MKT</b>	K	Nucleotide divergence per base pair, corrected by Jukes-Cantor	(43)	
	$\alpha$	Proportion of substitutions that are adaptive. It is calculated both from P and D, and from $\pi$ and K	(19,32,60,61)	
	$d$	Fraction of new mutations that are strongly deleterious and do not segregate in the population	(32)	Separately: 0-fold; 5'UTR; 3'UTR; intron; intergenic ( $\pm 500$ bp)
	$b$	Fraction of new mutations that are slightly deleterious and segregate at minor allele frequency (MAF) $< 5\%$		
	$f-\gamma$	Fraction of new mutations that are neutral since before the split of humans and chimpanzees, calculated after removing the excess of sites at MAF $< 5\%$ due to slightly deleterious mutations		
	$\gamma$	Fraction of new mutations that have become neutral recently, after the split of humans and chimpanzees, calculated after removing the excess of sites at MAF $< 5\%$ due to slightly deleterious mutations		
	$\alpha$	Proportion of substitutions that are adaptive, calculated after removing slightly deleterious mutations	(19,32,60,61)	
	DoS	Direction of Selection: difference between the proportion of nonsynonymous divergence and nonsynonymous polymorphism	(58)	

A comprehensive explanation is available under the section Help  $\rightarrow$  Integrative MKT of PopHuman.

instance previously developed by our group that hosts population genomics statistics for 30 *Drosophila melanogaster* populations (36). Novel features that have been implemented in PopHuman include the utility to retrieve gene-based integrative MKT metrics.

Compared to the 1000 Genomes Selection Browser 1.0 (18), PopHuman presents three significant advantages. First, PopHuman analyzes the 1000GP Phase III data, which included 2.29 times more sampled sequences (2504 versus 1092) compared to the Phase I, and used an improved variant calling pipeline. Specifically, Phase III implemented an expanded set of variant callers, including some that use haplotype information and others that rely on *de novo* assembly, it considered low coverage and exome sequencing data jointly rather than independently, and used a different genotype calling that allowed the integration of multi-allelic

variants and complex events (15). Second, PopHuman analyzes 26 instead of just three populations. This allows detecting very recent selective sweeps that have occurred in a single population and that can only be detected by analyzing data for this specific population; or older selective sweeps shared among a few related populations, whose detection gives a reinforcement of the time depth and biology underlying the specific selection signal. Three illustrative examples are provided: (i) a recent selective sweep related to skin pigmentation (37) in the region comprising the genes *SLC24A5*, *MYEF2*, *SLC12A1* and *CTXN2* in European (EUR) and South Asian (SAS) populations but not in East Asian (EAS) populations (Supplementary Figure S1); (ii) the presence of high frequency derived alleles in the gene *TRPV6* in all non-African populations, with a stronger signature in EAS populations, intermediate in SAS popula-

tions, and weaker in EUR populations, reflecting the time frame in which the establishment of agriculture, and thus the corresponding selective sweeps, occurred in those populations (stronger signatures in more recent sweeps; Supplementary Figure S2) and (iii) the presence of high frequency derived alleles in the Duffy red cell antigen gene (*DARC*, *FY*, *ACKR1*) in sub-Saharan Africa, thought to be the result of selection for resistance to *P. vivax* malaria (38,39), which is also seen in EAS populations (Supplementary Figure S3). Finally, PopHuman, contrary to the 1000 Genomes Selection Browser 1.0, implements selection tests based on the comparison of polymorphism and divergence, which are the only ones able to reveal the fixation of adaptive variants and other signatures of recurrent selection occurring over the last millions of years. One extreme example is found in the gene *PRMI*, which encodes a sperm-specific protein that compacts sperm DNA and shows a clear excess of function-altering substitutions between humans and chimpanzees compared to synonymous substitutions, indicative of positive Darwinian selection (40,41) (Supplementary Figure S4).

## CONCLUSION

The PopHuman database and browser go a step forward in the description and analysis of the most comprehensive human diversity data to date from a population genomics perspective. We aim PopHuman to be extended to incorporate novel metrics of transcriptomic and epigenomic variation, not only across individuals and species but also during the lifetime of an individual and/or in different parts of the body. In this way, PopHuman will become a pioneer population multi-omics browser advancing the upcoming population –omics synthesis (16).

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

## ACKNOWLEDGEMENTS

The authors thank Daniel Rigden and two anonymous referees for helpful comments on the PopHuman implementation and manuscript. We also thank Oscar Conchillo for helpful discussions about the informatics infrastructure in which PopHuman is implemented.

## FUNDING

Ministerio de Economía y Competitividad/European Regional Development Fund [grant numbers BFU2013-42649-P to A.B., BFU2016-77961-P to J.B.]; Generalitat de Catalunya [2014-SGR-1346, 2014-SGR-866]; Departament de Genètica i de Microbiologia of the Universitat Autònoma de Barcelona [12<sup>a</sup> PIPF to S.H.]; Youth Employment Initiative and European Social Fund [PEJ-2014 to E.S]. Funding for open access charge: Ministerio de Economía y Competitividad/European Regional Development Fund [BFU2013-42649-P to A.B., BFU2016-77961-P to J.B.].

*Conflict of interest statement.* None declared.

## REFERENCES

- Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
- Consortium, I.H.G.S. (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931–945.
- 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. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.
- International HapMap 3 Consortium, Altshuler, D.M., Gibbs, R.A., Peltonen, L., Altshuler, D.M., Gibbs, R.A., Peltonen, L., Dermitzakis, E., Schaffner, S.F., Yu, F. *et al.* (2010) Integrating common and rare genetic variation in diverse human populations. *Nature*, **467**, 52–58.
- International HapMap Consortium (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.
- International HapMap Consortium, 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. *et al.* (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–861.
- Manolio, T.A. and Collins, F.S. (2009) The HapMap and genome-wide association studies in diagnosis and therapy. *Annu. Rev. Med.*, **60**, 443–456.
- Sabeti, P.C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E.H., McCarroll, S.A., Gaudet, R. *et al.* (2007) Genome-wide detection and characterization of positive selection in human populations. *Nature*, **449**, 913–918.
- Akey, J.M. (2009) Constructing genomic maps of positive selection in humans: where do we go from here? *Genome Res.*, **19**, 711–722.
- Sabeti, P.C., Reich, D.E., Higgins, J.M., Levine, H.Z., Richter, D.J., Schaffner, S.F., Gabriel, S.B., Platko, J.V., Patterson, N.J., McDonald, G.J. *et al.* (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature*, **419**, 832–837.
- Voight, B.F., Kudaravalli, S., Wen, X. and Pritchard, J.K. (2006) A map of recent positive selection in the human genome. *PLoS Biol.*, **4**, e72.
- Metzker, M.L. (2010) Sequencing technologies - the next generation. *Nat. Rev. Genet.*, **11**, 31–46.
- Goodwin, S., McPherson, J.D. and McCombie, W.R. (2016) Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.*, **17**, 333–351.
- 1000 Genomes Project Consortium, Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T. and McVean, G.A. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.
- The 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.
- Casillas, S. and Barbadilla, A. (2017) Molecular population genetics. *Genetics*, **205**, 1003–1035.
- Sabeti, P.C., Schaffner, S.F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., Palma, A., Mikkelsen, T.S., Altshuler, D. and Lander, E.S. (2006) Positive natural selection in the human lineage. *Science*, **312**, 1614–1620.
- Pybus, M., Dall’Olio, G.M., Luisi, P., Uzkudun, M., Carreño-Torres, A., Pavlidis, P., Laayouni, H., Bertranpetit, J. and Engelken, J. (2014) 1000 Genomes Selection Browser 1.0: a genome browser dedicated to signatures of natural selection in modern humans. *Nucleic Acids Res.*, **42**, D903–D909.
- McDonald, J.H. and Kreitman, M. (1991) Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature*, **351**, 652–654.
- Gazal, S., Sahbatou, M., Babron, M.-C., Génin, E. and Leutenegger, A.-L. (2015) High level of inbreeding in final phase of 1000 Genomes Project. *Sci. Rep.*, **5**, srep17453.
- Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, **437**, 69–87.
- Aken, B.L., Achuthan, P., Akanni, W., Amode, M.R., Bernsdorff, F., Bhai, J., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P. *et al.* (2017) Ensembl 2017. *Nucleic Acids Res.*, **45**, D635–D642.
- Poliakov, A., Foong, J., Brudno, M. and Dubchak, I. (2014) GenomeVISTA—an integrated software package for whole-genome alignment and visualization. *Bioinform. Oxf. Engl.*, **30**, 2654–2655.

24. Bhérer, C., Campbell, C.L. and Auton, A. (2017) Refined genetic maps reveal sexual dimorphism in human meiotic recombination at multiple scales. *Nat. Commun.*, **8**, 14994.
25. Pfeifer, B., Wittelsbürger, U., Ramos-Onsins, S.E. and Lercher, M.J. (2014) PopGenome: an efficient Swiss army knife for population genomic analyses in R. *Mol. Biol. Evol.*, **31**, 1929–1936.
26. Szpiech, Z.A. and Hernandez, R.D. (2014) selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. *Mol. Biol. Evol.*, **31**, 2824–2827.
27. Pybus, M., Luisi, P., Dall’Olio, G.M., Uzkudun, M., Laayouni, H., Bertranpetit, J. and Engelken, J. (2015) Hierarchical boosting: a machine-learning framework to detect and classify hard selective sweeps in human populations. *Bioinformatics*, **31**, 3946–3952.
28. Haas, R.J. and Payseur, B.A. (2016) Fifteen years of genomewide scans for selection: trends, lessons and unaddressed genetic sources of complication. *Mol. Ecol.*, **25**, 5–23.
29. Fu, Y.X. and Li, W.H. (1993) Statistical tests of neutrality of mutations. *Genetics*, **133**, 693–709.
30. Andolfatto, P. (2005) Adaptive evolution of non-coding DNA in *Drosophila*. *Nature*, **437**, 1149–1152.
31. Egea, R., Casillas, S. and Barbadilla, A. (2008) Standard and generalized McDonald-Kreitman test: a website to detect selection by comparing different classes of DNA sites. *Nucleic Acids Res.*, **36**, W157–W162.
32. Mackay, T.F.C., Richards, S., Stone, E.A., Barbadilla, A., Ayroles, J.F., Zhu, D., Casillas, S., Han, Y., Magwire, M.M., Cridland, J.M. *et al.* (2012) The *Drosophila melanogaster* Genetic Reference Panel. *Nature*, **482**, 173–178.
33. O’Leary, N.A., Wright, M.W., Brister, J.R., Ciufu, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.
34. Buels, R., Yao, E., Diesh, C.M., Hayes, R.D., Munoz-Torres, M., Helt, G., Goodstein, D.M., Elsik, C.G., Lewis, S.E., Stein, L. *et al.* (2016) JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol.*, **17**, 66.
35. Tyner, C., Barber, G.P., Casper, J., Clawson, H., Diekhans, M., Eisenhart, C., Fischer, C.M., Gibson, D., Gonzalez, J.N., Guruvadoo, L. *et al.* (2017) The UCSC Genome Browser database: 2017 update. *Nucleic Acids Res.*, **45**, D626–D634.
36. Hervas, S., Sanz, E., Casillas, S., Pool, J.E. and Barbadilla, A. (2017) PopFly: the *Drosophila* population genomics browser. *Bioinformatics*, **33**, 2779–2780.
37. Lamason, R.L., Mohideen, M.-A.P.K., Mest, J.R., Wong, A.C., Norton, H.L., Aros, M.C., Juryne, M.J., Mao, X., Humphreville, V.R., Humbert, J.E. *et al.* (2005) SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science*, **310**, 1782–1786.
38. Escalante, A.A., Cornejo, O.E., Freeland, D.E., Poe, A.C., Durrego, E., Collins, W.E. and Lal, A.A. (2005) A monkey’s tale: the origin of *Plasmodium vivax* as a human malaria parasite. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 1980–1985.
39. Hamblin, M.T., Thompson, E.E. and Di Rienzo, A. (2002) Complex signatures of natural selection at the Duffy blood group locus. *Am. J. Hum. Genet.*, **70**, 369–383.
40. Wyckoff, G.J., Wang, W. and Wu, C.I. (2000) Rapid evolution of male reproductive genes in the descent of man. *Nature*, **403**, 304–309.
41. Rooney, A.P. and Zhang, J. (1999) Rapid evolution of a primate sperm protein: relaxation of functional constraint or positive Darwinian selection? *Mol. Biol. Evol.*, **16**, 706–710.
42. Nei, M. (1987) *Molecular Evolutionary Genetics*. Columbia University Press, NY.
43. Jukes, T.H. and Cantor, C.R. (1969) Evolution of protein molecules. In: *Mammalian Protein Metabolism*. Academic Press, NY, pp. 21–32.
44. Nei, M. and Gojobori, T. (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.*, **3**, 418–426.
45. Watterson, G.A. (1975) On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.*, **7**, 256–276.
46. Tajima, F. (1993) Measurement of DNA polymorphism. In: *Mechanisms of Molecular Evolution: Introduction to Molecular Paleopopulation Biology*. Sinauer Associates Inc., Sunderland, Massachusetts.
47. Tajima, F. (1996) The amount of DNA polymorphism maintained in a finite population when the neutral mutation rate varies among sites. *Genetics*, **143**, 1457–1465.
48. Hudson, R.R., Slatkin, M. and Maddison, W.P. (1992) Estimation of levels of gene flow from DNA sequence data. *Genetics*, **132**, 583–589.
49. Kelly, J.K. (1997) A test of neutrality based on interlocus associations. *Genetics*, **146**, 1197–1206.
50. Rozas, J., Gullaud, M., Blandin, G. and Aguadé, M. (2001) DNA variation at the rp49 gene region of *Drosophila simulans*: evolutionary inferences from an unusual haplotype structure. *Genetics*, **158**, 1147–1155.
51. Wall, J.D. (1999) Recombination and the power of statistical tests of neutrality. *Genet Res.*, **74**, 65–79.
52. Kong, A., Gudbjartsson, D.F., Sainz, J., Jonsson, G.M., Gudjonsson, S.A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G. *et al.* (2002) A high-resolution recombination map of the human genome. *Nat. Genet.*, **31**, 241–247.
53. Broman, K.W., Murray, J.C., Sheffield, V.C., White, R.L. and Weber, J.L. (1998) Comprehensive human genetic maps: individual and sex-specific variation in recombination. *Am. J. Hum. Genet.*, **63**, 861–869.
54. Dib, C., Fauré, S., Fizames, C., Samson, D., Drouot, N., Vignal, A., Millasseau, P., Marc, S., Kazan, J., Seboun, E. *et al.* (1996) A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature*, **380**, 152–154.
55. Fay, J.C. and Wu, C.I. (2000) Hitchhiking under positive Darwinian selection. *Genetics*, **155**, 1405–1413.
56. Tajima, F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.
57. Zeng, K., Fu, Y.-X., Shi, S. and Wu, C.-I. (2006) Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics*, **174**, 1431–1439.
58. Stoletzki, N. and Eyre-Walker, A. (2011) Estimation of the Neutrality Index. *Mol. Biol. Evol.*, **28**, 63–70.
59. Rand, D.M. and Kann, L.M. (1996) Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from *Drosophila*, mice, and humans. *Mol. Biol. Evol.*, **13**, 735–748.
60. Charlesworth, B. (1994) The effect of background selection against deleterious mutations on weakly selected, linked variants. *Genet Res.*, **63**, 213–227.
61. Smith, N.G. and Eyre-Walker, A. (2002) Adaptive protein evolution in *Drosophila*. *Nature*, **415**, 1022–1024.
62. Li, W.H., Wu, C.I. and Luo, C.C. (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol. Biol. Evol.*, **2**, 150–174.
63. Ronen, R., Udpa, N., Halperin, E. and Bafna, V. (2013) Learning natural selection from the site frequency spectrum. *Genetics*, **195**, 181–193.