UAB
Universitat Autònoma de Barcelona

Development of an open-source, comprehensive bioinformatics pipeline for the analysis of somatic NGS cancer panels

Raúl Marín Montes

UAB
Universitat Autònoma de Barcelona

Development of an open-source, comprehensive bioinformatics pipeline for the analysis of somatic NGS cancer panels

Raúl Marín Montes

Doctoral Program in Bioinformatics

Universitat Autònoma de Barcelona

Development of an open-source, comprehensive bioinformatics pipeline for the analysis of somatic NGS cancer panels

Thesis submitted by

Raúl Marín Montes

in pursuit of the

Doctoral degree from the Universitat Autònoma de Barcelona

Thesis conducted under the supervision of Dr. Xavier Solé Acha, Dr. Ernest Nadal Alforja, and Dr. Víctor Moreno Aguado at the Institut Català d'Oncologia (ICO) and Institut d'Investigació Biomèdica de Bellvitge (IDIBELL)

2020 – 2025

Xavier Solé Acha
Co-director

Ernest Nadal Alforja
Co-director

Raúl Marín Montes
Candidate

Víctor Moreno Aguado
Co-director

Juan Ramón González Ruiz
Tutor

Barcelona, September 2025

With funding of the Ministerio de Universidades, through the predoctoral fellowship number FPU19/01734 for the Formación de Profesorado Unviersitario (FPU). We thank CERCA Programme / Generalitat de Catalunya for institutional support.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

A	Adenine
AA	Amino Acid
AACR	American Association for Cancer Research
ACMG	American College of Medical Genetics and Genomics
AD	Allele Depth
AF	Allele Frequency
AGL	Agilent Custom Comprehensive Cancer Panel V2
ALL	Acute Lymphoblastic Leukemia
ALT	Alternate Allele
AML	Acute Myeloid Leukemia
AMP	Amplification // Association for Molecular Pathology
AR-V7	Androgen Receptor Splice Variant 7
ASCII	American Standard Code for Information Interchange
ASCO	American Society of Clinical Oncology
ASCETS	Arm-level Somatic Copy-Number Events in Targeted Sequencing
ATRT	Atypical Teratoid Rhabdoid Tumor
AWS	Amazon Web Services
BAF	B-Allele Frequency
BALSAMIC	Bioinformatic Analysis pipeLine for SomAtic MutatIons in Cancer
BAM	Binary Alignment Map
BCL	Binary Base Call
BED	Browser Extensible Data
Bp	Base Pair
BRP	Burning Rock DX Oncoscreen Plus
BWA-MEM2	Burrows-Wheeler Aligner - Maximum Exact Matches
C	Cytosine
CADD	Combined Annotation Dependent Depletion
CAP	College of American Pathologists
cDNA	Complementary DNA
CDx	Companion Diagnostics
CE-IVDR	European CE certification under the In Vitro Diagnostic Regulation
cfDNA	Circulating Cell-Free DNA
CGC	Cancer Genomics Consortium
CGI	Cancer Genome Interpreter

CGP	Comprehensive Genomic Profiling
chr	Chromosome
CIViC	Clinical Interpretations of Variants in Cancer
CLIA	Clinical Laboratory Improvement Amendments
ClinGen	Clinical Genome Resource
ClinVar	Clinical Variant Database
CLL	Chronic Lymphocytic Leukemia
CML	Chronic Myeloid Leukemia
CMML	Chronic Myelomonocytic Leukemia
CN	Copy Number
CNA	Copy-Number Alteration
COSMIC	Catalogue Of Somatic Mutations In Cancer
CRAM	Compressed Reference-Oriented Alignment Map
CRC	Colorectal Cancer
CSV	Comma-Separated Values
CTAT	Cancer Transcriptome Analysis Toolkit
CTR	Consensus Targeted Region
DDM	Data-Driven Medicine
DEL	Deletion
DFCI	Dana-Farber Cancer Institute
DOID	Disease Ontology Identifier
DP	Read Depth
DUP	Duplication
EGA	European Genome-Phenome Archive
ENCODE	Encyclopedia of DNA Elements
ER	Estrogen Receptor
ESCAT	ESMO Scale for Clinical Actionability of Molecular Targets
ESMO	European Society for Medical Oncology
FAIR	Findable, Accessible, Interoperable, and Reusable
FDA	Food and Drug Administration
FFPE	Formalin-Fixed Paraffin-Embedded
FFPM	Fusion Fragments per Million
FISH	Fluorescence *in situ* Hybridization
FN	False Negative
FP	False Positive
G	Guanine

GATK4	Genome Analysis Toolkit version 4
GA4GH	Global Alliance for Genomics and Health
GC	Guanine-Cytosine
GENIE	Genomics Evidence Neoplasia Information Exchange
GEP	Gene Expression Profiling
GIAB	Genome in a Bottle
GUI	Graphical User Interface
gnomAD	Genome Aggregation Database
GRC	Genome Reference Consortium
GRCh38	GRC Human Build 38
GRCm38	GRC Mouse build 38
GTF	Gene Transfer Format
hg	Human Genome
HGNC	Human Genome Organization Gene Nomenclature Committee
HGVS	Human Genome Variation Society
HighAmp	High-level AMP
HighDel	High-level DEL
HPC	High-Performance Computing
HRD	Homologous Recombination Deficiency
ICI	Immune Checkpoint Inhibitor
IDT	Integrated DNA Technologies xGen Pan-Cancer Panel
IGT	iGeneTech AIOnco-seq
IGV	Integrative Genomics Viewer
IHC	Immunohistochemistry
ILM	Illumina TruSight Tumor 170
InDel	Insertion and Deletion
ISO	International Organization for Standardization
IVD	*In vitro* Diagnostic
KB	Knowledgebase
KN	Known Negative
KP	Known Positive
LOH	Loss of Heterozygosity
LowAmp	Low-level AMP
LowDel	Low-level DEL
MANE	Matched Annotation from NCBI and EMBL-EBI
MAD	Median Absolute Deviation

MAPD	Median Absolute Pairwise Difference
MetaKB	Meta-Knowledgebase
METex14	MET exon 14 skipping
MSI	Microsatellite Instability
MMR	Mismatch Repair
MSigDB	Molecular Signatures Database
MSK-IMPACT	Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets
MTB	Molecular Tumor Board
Mut/Mb	Mutations per Megabase
NCG	Network of Cancer Genes
ND	No Data
NGS	Next-Generation Sequencing
NSCLC	Non-Small Cell Lung Cancer
OCA	Oncomine™ Comprehensive Assay
OncoKB	Oncology Knowledge Base
OM	Oncogenic Moderate
OP	Oncogenicity Prediction
OPA	Oncomine™ Precision Assay
OVS	Oncogenic Very Strong
PCR	Polymerase Chain Reaction
PCR-Amp	PCR Amplification
PDX	Patient-Derived Xenograft
PoN	Panel of Normals
pVAF	Population VAF
QC	Quality Control
QCI	QIAGEN Clinical Insight
q10	Mean mapping quality <10
q20	Quality score <20
REF	Reference Allele
REVEL	Rare Exome Variant Ensemble Learner
RNA-seq	RNA Sequencing
RT	Reverse Transcription
RUO	Research Use Only
SaaS	Software as a Service
SAM	Sequence Alignment Map

SB	Strand Bias
SBP	Somatic Benign Supporting
SBS	Sequencing-By-Synthesis // Somatic Benign Strong
SBVS	Somatic Benign Very Strong
SCHOOL	Software for Clinical Health Omics Onocolgy Laboratories
SEQC2	Sequencing Quality Control Phase II
SNP	Single-Nucleotide Polymorphism
SNV	Single-Nucleotide Variant
SOP	Standard Operating Procedure
STAR	Spliced Transcripts Alignment to a Reference
SV	Structural Variant
sWGS	Shallow Whole Genome Sequencing
T	Thymine
TCGA	The Cancer Genome Atlas
TFS	Thermo Fisher Oncomine Comprehensive Assay v3
TMB	Tumor Mutational Burden
TMAP	Torrent Mapping Alignment Program
TOSCA	Tumor Only Somatic CAlling
TP	True Positive // Tumor Purity
TSG	Tumor Suppressor Gene
TSO500	TruSight™ Oncology 500
TSV	Tab-Separated Values
TVC	Torrent Variant Caller
T2T- CHM13	Telomere-to-Telomere CHM13
uBAM	Unmapped Binary Alignment Map
UCSC	University of California Santa Cruz
UMI	Unique Molecular Identifier
VAF	Variant Allele Frequency
VCF	Variant Call Format
VEP	Variant Effect Predictor
VICC	Variant Interpretation for Cancer Consortium
VUS	Variant of Uncertain Significance
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing

SUMMARY

Next-generation sequencing (NGS) has revolutionized cancer genomics by enabling the detection of clinically relevant somatic alterations. While targeted NGS panels are widely used for tumor characterization, their effectiveness depends on bioinformatics pipelines capable of analyzing tumor-only data and producing accurate, reproducible, and interpretable results. Current solutions often lack flexibility, transparency, or full integration, underscoring the need for more adaptable and comprehensive alternatives.

This thesis presents the implementation and validation of ClinBioNGS, an open-source, comprehensive bioinformatics pipeline designed for the analysis of somatic NGS cancer panels. The project pursued two main objectives: (1) to design a flexible, reproducible pipeline for the analysis of tumor-only DNA and RNA panel data; and (2) to evaluate its performance using standardized reference datasets and retrospective real-world data from diverse NGS panels.

ClinBioNGS enables the detection of a wide range of somatic events—including small variants, copy-number alterations (CNAs), gene fusions, splice variants, and complex biomarkers such as tumor mutational burden (TMB) and microsatellite instability (MSI). Built with Nextflow and containerized environments, its panel-agnostic architecture ensures reproducibility and portability across computing infrastructures. The pipeline integrates consensus variant calling strategies, panel-specific CNA and MSI reference models, automated annotation and prioritization modules, internal quality control systems, and a variant database for longitudinal tracking. Results are presented through interactive, visual HTML reports tailored for interpretability and multidisciplinary review.

Validation using multi-panel reference datasets confirmed high accuracy in small variant detection. Benchmarking with real-world clinical samples from multiple institutions and panels demonstrated performance comparable to commercial solutions, while providing broader detection capabilities and improved interpretability in complex cases. The pipeline is freely available for non-commercial research use only at: https://github.com/raulmarinm/ClinBioNGS.

This work provides a robust, versatile, and openly accessible solution for somatic NGS panel analysis, contributing to the advancement of both precision oncology and cancer research.

RESUM (CATALÀ)

La seqüenciació de nova generació (NGS) ha revolucionat la genòmica del càncer en permetre la detecció d'alteracions somàtiques clínicament rellevants. Tot i que els panells dirigits de NGS s'utilitzen àmpliament per caracteritzar tumors, la seva eficàcia depèn de *pipelines* bioinformàtics capaços de analitzar dades tumorals sense teixit sa aparellat i generar resultats precisos, reproduïbles i interpretables. Les solucions actuals sovint presenten limitacions de flexibilitat, transparència o integració completa, fet que posa de manifest la necessitat d'alternatives més adaptables i integrals.

Aquesta tesi presenta la implementació i validació de ClinBioNGS, un *pipeline* bioinformàtic complet i de codi obert dissenyat per a l'anàlisi de panells NGS somàtics en càncer. El projecte aborda dos objectius principals: (1) dissenyar un *pipeline* flexible i reproduïble per a l'anàlisi de dades de panells d'ADN i ARN tumorals sense teixit sa associat, i (2) avaluar-ne el rendiment mitjançant conjunts de dades de referència estandarditzats i dades reals retrospectives procedents de panells NGS diversos.

ClinBioNGS permet la detecció d'un ampli ventall d'alteracions somàtiques, incloent petits canvis de nucleòtids, alteracions del nombre de còpies (CNAs), fusions gèniques, variants d'*splicing* i biomarcadors complexos com la càrrega mutacional tumoral (TMB) i la inestabilitat de microsatèl·lits (MSI). El seu disseny independent del panell, construït amb Nextflow i entorns contenidoritzats, garanteix la seva reproductibilitat i portabilitat entre infraestructures computacionals. També incorpora estratègies de consens per a la detecció de variants, referències específiques per a CNA i MSI, mòduls automatitzats per a l'anotació i priorització clínica, sistemes de control de qualitat interns, i una base de dades local per al seguiment longitudinal de variants. Els resultats es presenten mitjançant informes HTML interactius i visuals, optimitzats per a la seva interpretació i revisió multidisciplinària.

La validació amb conjunts de dades de referència multi panell va confirmar una alta precisió en la detecció de variants petites. L'avaluació comparativa amb dades clíniques reals de diverses institucions i panells comercials va demostrar un rendiment comparable a les solucions existents, tot oferint una major capacitat de detecció i millor interpretabilitat en casos complexos. El *pipeline* està disponible lliurement per a ús en recerca i finalitats no comercials a: https://github.com/raulmarinm/ClinBioNGS.

Aquest treball proporciona una solució sòlida, versàtil i accessible per a l'anàlisi de panells NGS somàtics, contribuint al progrés tant de l'oncologia de precisió com de la recerca translacional en càncer.

RESUMEN (CASTELLANO)

La secuenciación de nueva generación (NGS) ha revolucionado la genómica del cáncer al permitir la detección de alteraciones somáticas clínicamente relevantes. Aunque los paneles dirigidos de NGS se utilizan ampliamente para la caracterización tumoral, su eficacia depende de *pipelines* bioinformáticos capaces de analizar muestras tumorales sin tejido sano emparejado y de generar resultados precisos, reproducibles e interpretables. Las soluciones actuales a menudo carecen de flexibilidad, transparencia o integración completa, lo que pone de manifiesto la necesidad de alternativas más adaptables e integrales.

Esta tesis presenta la implementación y validación de ClinBioNGS, un *pipeline* bioinformático de código abierto y carácter integral, diseñado para el análisis de paneles de cáncer por NGS en muestras somáticas. El proyecto aborda dos objetivos principales: (1) diseñar un pipeline flexible y reproducible para el análisis de datos tumorales de ADN y ARN, y (2) evaluar su rendimiento utilizando conjuntos de referencia estandarizados y datos retrospectivos del mundo real obtenidos de distintos paneles comerciales.

ClinBioNGS permite la detección de una amplia variedad de eventos somáticos, incluyendo pequeños cambios de nucleótidos, alteraciones del número de copias (CNAs), fusiones génicas, variantes de *splicing* y biomarcadores complejos como la carga mutacional tumoral (TMB) y la inestabilidad de microsatélites (MSI). Su diseño independiente del panel, basado en Nextflow y entornos contenerizados, garantiza la reproducibilidad y portabilidad entre infraestructuras computacionales. El pipeline integra estrategias de detección de variantes por consenso, modelos de referencia específicos por panel para CNA y MSI, módulos automatizados de anotación y priorización clínica, sistemas internos de control de calidad y una base de datos de variantes para seguimiento longitudinal. Los resultados se presentan mediante informes HTML interactivos y visuales, optimizados para su interpretación y revisión multidisciplinar.

La validación con conjuntos de datos de referencia multi panel confirmó una alta precisión en la detección de variantes pequeñas. El análisis comparativo con muestras clínicas reales de múltiples instituciones y paneles demostró un rendimiento comparable al de soluciones comerciales, al tiempo que ofreció un mayor alcance de detección y mejor capacidad interpretativa en casos complejos. El pipeline está disponible libremente para uso en investigación y con fines no comerciales en: https://github.com/raulmarinm/ClinBioNGS.

Este trabajo proporciona una solución robusta, versátil y accesible para el análisis de paneles somáticos por NGS, contribuyendo al avance tanto de la oncología de precisión como de la investigación del cáncer.

1. INTRODUCTION

1.1. Current state of molecular profiling in precision oncology

1.1.1. Molecular basis of cancer

Cancer is fundamentally a genetic disease driven by alterations in functional regions of DNA, commonly referred to as genes, that disrupt normal cellular regulatory mechanisms[1,2]. These genetic alterations (also known as variants or mutations) can be inherited from parents or acquired over time due to intrinsic biological processes (e.g., DNA replication errors, oxidative damage, or cytosine deamination) and through extrinsic exposure to damaging agents (e.g., tobacco smoke, alcohol, radiation, chemical carcinogens, or viral infections)[1–3]. Such mutations can affect oncogenes, tumor suppressor genes (TSGs), and DNA repair genes, ultimately leading to uncontrolled proliferation, evasion of apoptosis, and genomic instability[1–4]. Unlike germline variants, which are inherited and present in all cells of the body, somatic variants arise spontaneously in non-germline cells during a person's lifetime and are not transmitted to offspring[1–4]. The progressive accumulation of somatic mutations enables tumor initiation, clonal evolution, and disease progression[1–3]. Consequently, the specific mutational landscape of a tumor profoundly influences its biological behavior, response to treatment, and clinical outcome, underscoring the critical importance of molecular characterization in oncology[2–6].

1.1.2. Emergence of precision oncology

In recent years, precision oncology has transformed the landscape of cancer treatment, shifting from a one-size-fits-all approach to a strategy guided by the molecular profile of each individual tumor, with the goal of maximizing treatment efficacy while minimizing toxicity for each patient [4–8] . Since the introduction of targeted therapy against estrogen receptor (ER) expression in breast cancer in the 1970s, precision oncology has evolved rapidly[5–8]. Parallel advances in technological innovation, notably the emergence of next-generation sequencing (NGS), and deeper understanding of tumorigenesis have driven the discovery of new actionable genomic alterations[4–9]. These developments have enabled the implementation of both alteration-specific, tumor-related therapies and biomarker-driven, tumor-agnostic treatments (**Figure 1**)[4–8].

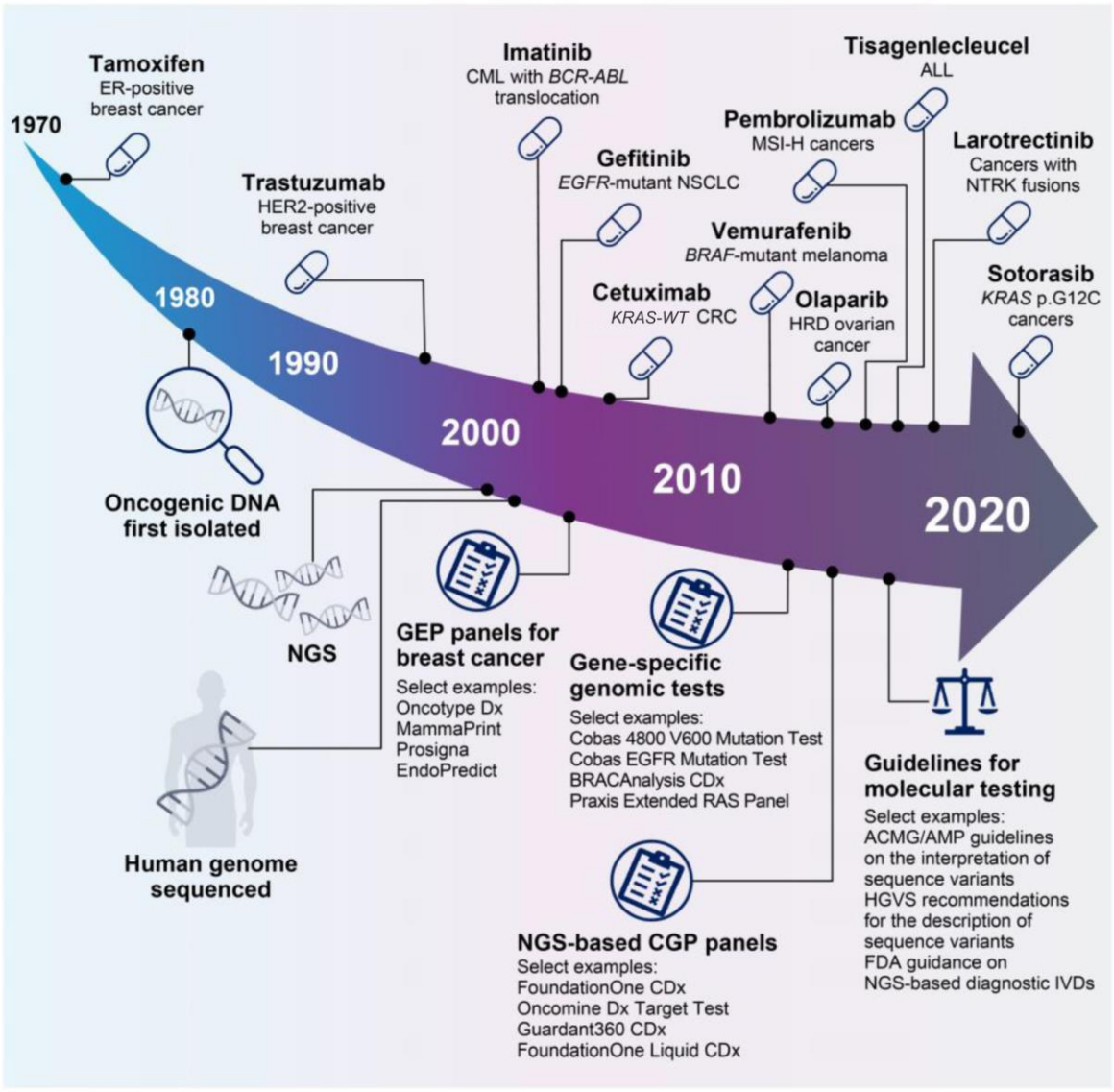


Figure 1. Timeline of key milestones in 50 years of precision oncology.
The upper section highlights landmark therapeutic advances and their associated molecular targets. The lower section summarizes major technological innovations, diagnostic tests, and guidelines that have shaped the development of molecular diagnostics. Adapted from *Rulten et al., 2023* [6].

At the core of this paradigm is molecular profiling, a comprehensive process that identifies a wide spectrum of genetic and molecular alterations implicated in tumorigenesis and provides clinically relevant information[4–13]. These alterations can occur at the DNA, RNA, or protein level and encompass diverse classes of biomarkers. Increasingly, they are used as predictive biomarkers to match patients with targeted therapies, immunotherapies, or clinical trials, as well as to refine tumor classification, complement pathological diagnosis, and guide prognostic stratification[4–17]. A summary of these alteration types and their clinical relevance is provided in **Table 1**.

Table 1. Types of molecular alterations and biomarkers relevant to precision oncology.
Each alteration or biomarker includes a brief description and representative examples of its clinical relevance.

Alteration / Biomarker	Description	Clinical relevance
Small variants	Single-nucleotide variants (SNVs) or small insertions and deletions (InDels) of nucleotides that can affect protein function.	Diagnostic (*NPM1* in AML), poor prognosis (*TP53* in CLL), drug response (*BRAF V600E* in melanoma/NSCLC) or resistance (*EGFR T790M* in NSCLC).
Copy-number alterations (CNAs)	Genomic amplifications (AMPs) or deletions (DELs) that affect gene dosage.	Diagnostic (*SMARCB1* loss in ATRT), poor prognosis (*CDKN2A* loss in low-grade glioma), drug response (*ERBB2* gain in breast/gastric cancer) or resistance (*PTEN* loss in bladder carcinoma).
Fusions	Structural rearrangements that result in gene fusions.	Diagnostic (*PDGFRA/B* in CMML), better outcome (*RUNX1-RUNX1T1* in AML), targeted therapy (*EML4-ALK* or *CD74-ROS1* in NSCLC).
Splice variants	Mutations in splice sites or splicing factors leading to alternative splicing isoforms.	Drug response (*MET* exon 14 skipping in NSCLC) or resistance (*AR-V7* in prostate cancer).
Epigenetic alterations	Genomic changes that modulate gene expression without altering DNA sequence	Better outcome and drug response (*MGMT* promoter methylation in glioblastoma).
Tumor mutational burden (TMB)	Total number of somatic mutations in tumor cells.	Predictive of response to immune checkpoint inhibitors (ICIs) (TMB-high solid tumors).
Microsatellite instability (MSI)	Genetic hypermutability caused by mismatch repair deficiency.	Predictive of response to ICIs (MSI-high/MMR deficient solid tumors)
HRD	Genomic instability due to homologous recombination deficiency (HRD).	Predictive of response to PARP inhibitors (HRD-high *BRCA1/2* mutations in breast/ovarian cancer)
Mutational signatures	Specific mutation patterns linked to mutagenic processes or exposures.	Drug response (UV/tobacco/APOBEC/POLE signatures to ICIs) or resistance (APOBEC to tyrosine kinase inhibitors).
Gene expression	Expression signatures of single genes or gene panels.	Molecular subtyping and risk stratification (PAM50 in breast cancer), drug response (BRCAness to PARP inhibitors).
Protein expression	Abnormal levels or activation of specific proteins	Therapy guidance (PD-L1 expression in tumor and/or immune cells for ICI eligibility).

1.1.3. The rise of NGS-based comprehensive genomic profiling

The detection of the molecular alterations described above relies on a variety of laboratory assays, each with distinct advantages and limitations. Traditionally, molecular testing in oncology has relied on single-gene methods such as polymerase chain reaction (PCR), immunohistochemistry (IHC), fluorescence *in situ* hybridization (FISH), or DNA Sanger sequencing[5–13]. These conventional techniques offer high sensitivity (i.e., ability to identify existing variants) and specificity (i.e., ability to avoid false variants) for predefined alterations and remain indispensable in many diagnostic workflows[5–10]. Moreover, each test has a distinct limit of detection, defined as the lowest variant allele frequency or minimal number of mutant copies that can be reliably identified as true variants. However, because each assay targets a specific biomarker, their scope is inherently limited and often lacks the ability to detect unexpected, rare, or complex genomic events[4–6,9].

In response to these limitations, improvements in NGS-based technologies have enabled more complex, scalable, and cost-effective analyses of biological molecules, significantly expanding the catalogue of clinically actionable alterations. As a result, there has been a progressive shift from single-target diagnostic assays to NGS-based comprehensive genomic profiling (CGP)[4–13,18]. This approach allows for the simultaneous interrogation of multiple genes and alteration types within a single experiment, providing a more efficient, cost-effective, and tissue-sparing alternative to sequential single-biomarker testing[4–13,18]. While gene-specific assays continue to play an important role in routine diagnostics and in resolving discordant results, the adoption of NGS-based CGP panels in clinical oncology has accelerated, driven by their ability to deliver a holistic view of the tumor genome and to inform precision oncology decision-making (**Figure 1**)[5–11,19]. The characteristics of representative gene-specific assays and NGS-based CGP panels are summarized in **Table 2**[7–13,18–20].

Table 2. Representative examples of gene-specific tests and NGS-based CGP panels in oncology.
This table summarizes Food and Drug Administration (FDA)-approved diagnostic assays ranging from single-gene tests to large-scale NGS-based panels. For each test, the table lists the number of genes interrogated, sample type, assay method, and molecular profiling scope.

Test	Genes	Sample	Method	Molecular profiling
Cobas 4800 BRAF V600 Mutation Test (Roche)	1	Formalin-fixed paraffin-embedded (FFPE) tissue	Real-time PCR	*BRAF V600E* (melanoma)
Cobas EGFR Mutation Test V2 (Roche)	1	FFPE	Real-time PCR	*EGFR* exons 18-21 mutations (NSCLC)
BRACAnalysis companion diagnostics (CDx) (Myriad Genetics)	2	Blood	PCR + Sanger Multiplex PCR	*BRCA1/2* SNVs, InDels, and AMP/DEL (ovarian and breast cancers)
Praxis Extended RAS Panel (Illumina)	2	FFPE	NGS	*K/N-ras* exons 2-4 mutations (colorectal cancer)
FoundationOne CDx (Foundation Medicine)	324	FFPE (DNA)	NGS	SNVs, InDels, CNAs (16 genes), rearrangements (36 genes), TMB, MSI
MSK IMPACT (Memorial Sloan Kettering)	468	FFPE (DNA)	NGS	SNVs, InDels, CNAs, rearrangements, TMB, MSI
Oncomine Dx Target Test (Thermo Fisher)	46	FFPE (DNA+RNA)	NGS	SNVs and InDels (42 genes), CNAs (10 genes), fusions and splice variants (17 genes)
Trusight Oncology 500 (Illumina)	523	FFPE (DNA+RNA)	NGS	SNVs, InDels, CNAs (59 genes), fusions and splice variants (55 genes), TMB, MSI
Oncomine Comprehensive Assay (Thermo Fisher)	161	FFPE (DNA+RNA)	NGS	SNVs and InDels (135 genes), CNAs (43 genes), fusions and splice variants (51 genes)
FoundationOne Liquid CDx (Foundation Medicine)	324	Plasma (circulating cell-free DNA [cfDNA])	NGS	SNVs and InDels (311 genes), CNAs (4 genes), rearrangements (4 genes), TMB, MSI
Guandant360 CDx (Guardant)	55	Plasma (cfDNA)	NGS	SNVs and InDels (55 genes), CNAs (2 genes), fusions (4 genes)

With the advent of large-scale genome sequencing initiatives (e.g., The Cancer Genome Atlas [TCGA][21], the International Cancer Genome Consortium[22], and the 1000 Genomes Project[23]) vast catalogs of tumor-associated variants have been identified and consolidated into public reference resources such as the Genome Aggregation Database (gnomAD)[24], the Catalogue Of Somatic Mutations In Cancer (COSMIC)[25], and the American Association for Cancer Research (AACR) Project Genomics Evidence Neoplasia Information Exchange (GENIE)[26]. These foundational efforts have, in turn, enabled the development of specialized clinical interpretation databases, including the Clinical Variant Database (ClinVar)[27], the Oncology Knowledge Base (OncoKB)[28], and the Clinical Interpretations of Variants in Cancer (CIViC)[29], that curate variant-specific evidence to support clinical reporting.

However, the proliferation of numerous independent resources, curated by different groups with varying scopes and methodologies, has also introduced significant challenges. Discordances in variant representation across platforms—ranging from inconsistent nomenclature to divergent evidence grading—complicate the accurate interpretation of molecular findings in clinical settings. Consequently, the standardization and harmonization of variant annotation, classification, and reporting have become critical to ensure equity, reproducibility, and consistency of genomic results across laboratories and institutions[4,5].

To address these challenges, multiple guidelines and recommendations for molecular testing have been developed (**Figure 1**), including frameworks for variant nomenclature (e.g., Human Genome Variation Society [HGVS] recommendations)[30] and variant interpretation (e.g., American College of Medical Genetics and Genomics [ACMG] and Association for Molecular Pathology [AMP] consensus recommendations[31]), and guidance documents published by scientific societies (e.g., European Society for Medical Oncology [ESMO][32]) and regulatory agencies (e.g., Food and Drug Administration [FDA])[33] to define requirements for NGS-based *in vitro* diagnostics (IVDs).

In order to fully leverage the potential of NGS-driven precision oncology, the molecular tumor board (MTB) has emerged as a central entity in many institutions. MTBs bring together multidisciplinary teams—including oncologists, pathologists, geneticists, research scientists, and bioinformaticians, among others—to collaboratively review each patient's data from clinical, pathological, and molecular perspectives and formulate evidence-based treatment recommendations[5,34].

However, as the number of patients requiring MTB evaluation continues to grow, there is an increasing need for automated systems that can efficiently integrate variant annotation, prioritization, and reporting to streamline this process[34–36]. Moreover, given the interdisciplinary composition of MTBs, the results must be presented in a visual, intuitive, and accessible format—without sacrificing scientific rigor—to support effective collaboration and improve clinical decision-making[34].

In summary, the adoption of CGP through NGS has rapidly expanded in routine oncology practice, driven by multiple converging factors:

- The growing number of clinically actionable biomarkers[4–9].
- The increasing availability of targeted therapies and biomarker-driven treatments[4–9].
- Continuous advances in NGS technologies and bioinformatics pipelines.
- Decreasing costs and shorter turnaround times, making high-throughput sequencing more feasible and accessible.
- Ongoing harmonization of genomic knowledge bases to support reliable variant interpretation.
- Strengthened regulatory frameworks and endorsement by clinical guidelines.

Together, these factors underscore the growing need for robust, standardized, and scalable bioinformatics solutions capable of transforming raw NGS data into clinically actionable insights.

1.2. The NGS-based panel workflow in molecular diagnostics

1.2.1. General considerations

Building on the advances described in the previous section, NGS-based CGP panels have become a cornerstone of precision oncology, enabling the simultaneous detection of multiple classes of clinically relevant alterations within a single assay[5–13,37]. In routine diagnostics, NGS panel workflows follow a standardized sequence of steps (**Figure 2**)—including sample processing, library preparation and target enrichment, sequencing, and bioinformatics analysis—to transform clinical specimens into structured, interpretable genomic data[7–13]. This integrated approach not only facilitates the identification of actionable biomarkers to guide patient management but also promotes consistency and transparency across laboratories and platforms. Each stage of the NGS workflow involves specific technical considerations that influence the overall performance and clinical utility of the assay.

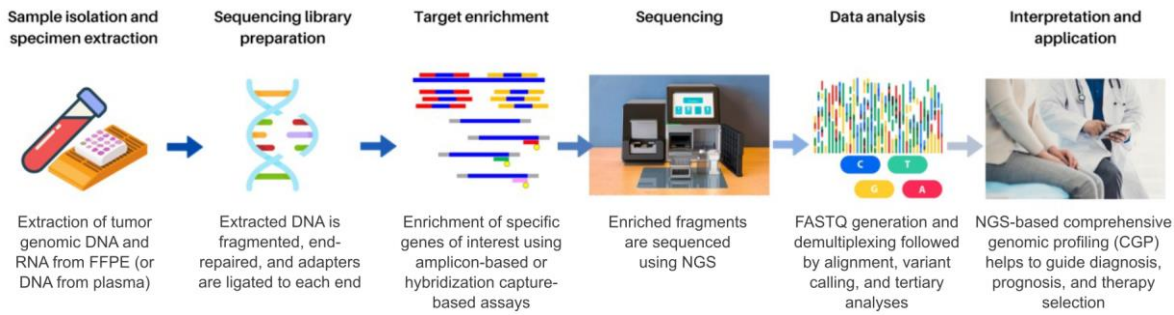


Figure 2. General workflow of NGS-based panel testing.
The process begins with nucleic acid extraction from the selected specimen type, followed by library preparation and target enrichment to capture the genomic regions of interest. Sequencing is then performed, and bioinformatics analysis pipelines are applied to generate structured variant data and facilitate clinical interpretation. Adapted from *Pei et al., 2023*[13].

Before introducing the NGS workflow in detail, it is important to highlight several key NGS-related features that are consistently evaluated across the workflow and influence both the quality of the data and the interpretation of the results[9–13,38]:

- Reads: Individual DNA or RNA fragments sequenced from the sample, and their number and quality directly affect downstream analyses.
- Read depth (DP): Number of times each base or nucleotide is sequenced, corresponding to the total number of reads overlapping a specific genomic position.
- Coverage: From a broader perspective than DP, it indicates how much of the genome or target region is sequenced. It can be expressed in percentage (e.g., 90% of the genome is sequenced at least once) or commonly as a fold of the genome size expressed with an "X" (e.g., 500X means that the target size is sequenced on average 500 times).
- Reference allele (REF): Nucleotide(s) present at defined locus of the reference genome used for comparison.
- Alternate allele (ALT): Nucleotide(s) observed in the sample corresponding to the detected variant that differs from the REF at the same position.
- Allele depth (AD): Number of reads supporting each allele (REF, ALT) at a specific locus.
- Allele frequency (AF): Proportion of reads supporting a specific allele (commonly referred to ALT) relative to DP. In the population context, it represents the proportion of each allele within a population.
- Variant allele frequency (VAF): Proportion of reads supporting the variant allele (i.e., ALT).
- Probe or capture bait: Oligonucleotides used during target enrichment to isolate specific genomic regions.
- Hotspots regions: Genomic loci that are frequently mutated in cancer may be associated with a specific tumor type or carry clinical relevance. Variants found within these loci are also referred to as hotspots mutations.

Targeted NGS panels have become a routine molecular diagnostic tool in both clinical and research settings, as they deliver reliable results at relatively low cost and turnaround time. Regarding the panel design, several aspects should be considered to ensure appropriate panel and optimal capture of the relevant genomic alterations[9–13].

The NGS panel choice depends primarily on the clinical purpose and the genes or biomarkers of interest. For example, germline testing often requires different target regions and analytical approaches compared to somatic profiling, and the requirements also differ between solid tumors and hematologic malignancies[9–13]. In this thesis, the focus is on somatic applications in solid tumors, mainly in the context of routine clinical diagnostics. For this purpose, commercially available pan-

cancer CGP panels (**Table 2**) are frequently adopted, as they allow batching of samples from diverse tumor types and clinical indications, ultimately saving time and reducing costs[11–13].

Depending on the type of alteration targeted (**Table 1**), the design of the capture regions within the panel may vary considerably. Small variants, including single-nucleotide variants (SNVs) and insertions and deletions (InDels), can be reliably detected by targeting specific hotspot regions, minimizing the overall size of the panel. In contrast, copy-number alteration (CNA) assessment requires multiple probes spanning the entire gene of interest to obtain accurate results[9–13].

Gene fusions can be detected using either DNA- or RNA-based approaches. Because most fusion breakpoints occur within intronic regions, DNA-based detection requires capture probes spanning introns—an approach that can be technically challenging due to the typically large size of these regions. Conversely, RNA-based methods target exon–exon junctions, facilitating the detection of both known and novel fusion breakpoints in a more efficient manner. For this reason, RNA sequencing (RNA-seq) is increasingly preferred for comprehensive fusion profiling, particularly when novel fusion partners are expected[9–13].

1.2.2. Wet-lab workflow: from sample preparation to sequencing

The wet-lab phase of NGS-based CGP encompasses all laboratory steps required to transform clinical specimens into sequencing-ready libraries. This process begins with specimen selection and nucleic acid extraction, continues through library preparation and target enrichment, and culminates in high-throughput sequencing. Each step must be carefully optimized to ensure the generation of high-quality, reproducible data suitable for downstream bioinformatics analysis. Variables such as the type and preservation of the input material, the enrichment strategy employed, and the chosen sequencing platform can all influence the sensitivity, specificity, and overall reliability of the assay[5,9–13]. The following sections describe the main considerations and methodological approaches involved in each stage of the wet-lab workflow.

1.2.2.1. Sample processing

The first step of any clinical NGS panel workflow is the preparation of input specimens, followed by the extraction and evaluation of nucleic acids (i.e., DNA and RNA). While these procedures are not strictly bioinformatic, they have significant downstream consequences for data quality and interpretation[9–13].

The choice of specimen depends on the clinical indication. Germline analyses typically require saliva or peripheral blood to isolate non-tumor cells. In contrast, somatic testing—usually performed after tumor diagnosis—commonly uses FFPE tumor tissue, fresh-frozen tumor tissue, or circulating cell-

free DNA (cfDNA), as shown in the assays listed in **Table 2**. The preservation method can strongly influence analysis outcomes. For example, DNA extracted from FFPE is prone to fixation-induced damage and artifacts (such as cytosine deamination and strand bias) that must be accounted for during variant calling[9–13].

Another important consideration is sample quantity, which is often limited in the clinical setting, particularly for solid tumors where tissue collection is invasive. Although single-gene tests generally require smaller input amounts, the use of NGS panels provides broader information with a single analysis, maximizing the yield from scarce specimens and increasing the likelihood of identifying actionable biomarkers[9–13].

Tumor-specific features also impact interpretation. For instance, estimating tumor purity (TP)—defined as the proportion of tumor cells within the total sample—is essential when evaluating VAFs of somatic variants and CNAs[9–12]. In solid tumor samples, pathologist-assessed tumor content can be improved by microdissection of tumor-rich regions to increment neoplastic cell fraction and improve sensitivity. However, pathology-based estimates are inherently subjective and can be influenced by factors such as interobserver variability, infiltrating non-tumor cells, inflammation, or necrosis[9,11]. Computational estimation of TP from sequencing data presents an alternative but is also affected by genomic features such as chromosomal instability. Combining microscopic assessment and *in silico* estimation could provide a more robust and reliable estimation of tumor content[9–11].

Following tissue selection, DNA and RNA are extracted and quantified. For RNA-based assays, reverse transcription (RT) is performed to generate complementary DNA (cDNA). The required input DNA quantity varies by panel and can range from 10 ng to 1000 ng. Additional parameters (e.g., quality, concentration, overall yield) are also evaluated to confirm sample suitability[9–13].

In research applications, a wider variety of specimen types are commonly encountered. Beyond tumor tissues, samples may include tumor-derived cell lines or tissues from patient-derived xenograft (PDX) models. These cases often require additional bioinformatic pre-processing steps to correctly attribute sequencing reads to their origin.

1.2.2.2. Library preparation

Library preparation is the process by which extracted DNA is converted into a form compatible with sequencing. In a typical workflow, genomic DNA is first fragmented, and short adapter sequences (which are complementary to the sequencing platform's flow cell) are ligated to each end of each fragment, creating what is known as an insert. This is followed by PCR amplification (PCR-Amp) to increase DNA yield. However, the details of this process vary depending on the enrichment strategy employed[9–13].

Depending on the chosen NGS approach, the scope of captured genomic regions differs considerably. Whole genome sequencing (WGS) is untargeted, covering the entire genome. Whole exome sequencing (WES) focuses on coding regions of all protein-coding genes, while targeted sequencing restricts capture to a defined set of genes or hotspots of known clinical relevance. The main characteristics of these three strategies are summarized in **Figure 3**. In clinical diagnostics, targeted gene panels are generally preferred, as prior knowledge of relevant genes enables greater sensitivity for detecting known alterations, while maintaining lower costs and shorter turnaround times[5,9–13].

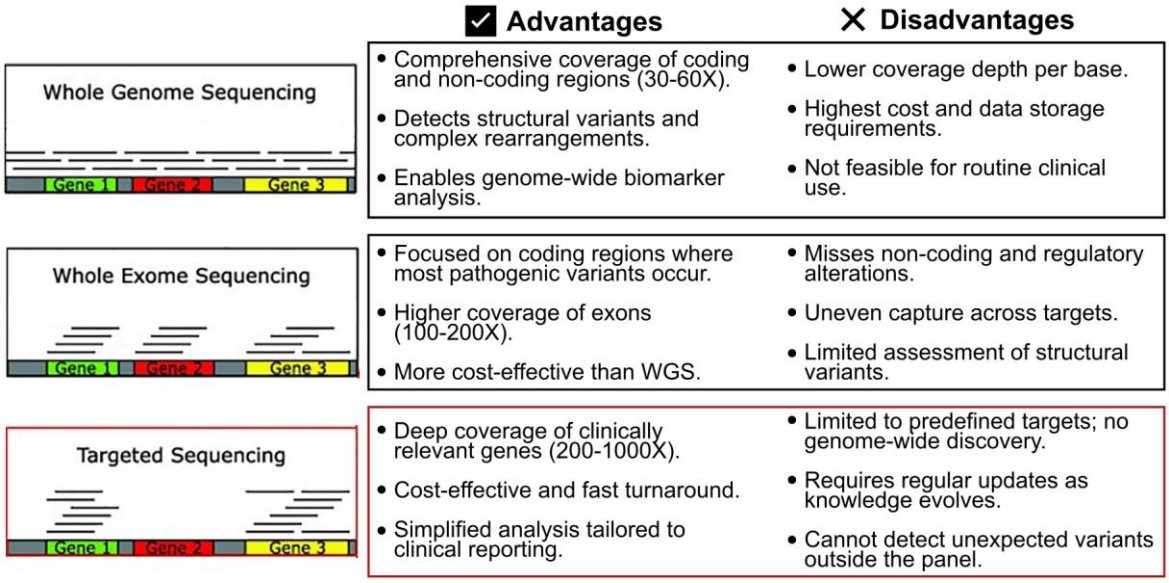


Figure 3. Overview of genomic NGS approaches.
Illustration of three main sequencing strategies—WGS, WES, and targeted sequencing—highlighting their genomic scope, advantages, and limitations. Adapted from *Bewicke-Copley et al., 2019*[12].

Two principal methods are used for target enrichment and library preparation (**Figure 4**)[9–13]:

- Amplicon-based approaches rely on PCR primers designed to bind to the flanking regions of targets and selectively amplify them (**Figure 4A**). In this method, adapters are incorporated during PCR-Amp itself. Often, multiple overlapping primers are included to ensure full coverage. Because all reads generated have the same start and end coordinates (defined by primer positions), it cannot distinguish true unique molecules from PCR duplicates by coordinates alone, complicating deduplication during analysis.
- Hybridization capture-based approaches use biotinylated probes (i.e., capture baits) that hybridize to target regions. These probe-bound fragments are captured via streptavidin-coated magnetic beads (**Figure 4B**). In this workflow, adapters are ligated before hybridization. The resulting reads start and end at variable positions, allowing accurate detection and removal of duplicates. In general, hybridization capture yields more uniform and accurate coverage, while amplicon-based methods are advantageous for smaller-scale experiments, limited DNA input, or resource-constrained clinical applications.

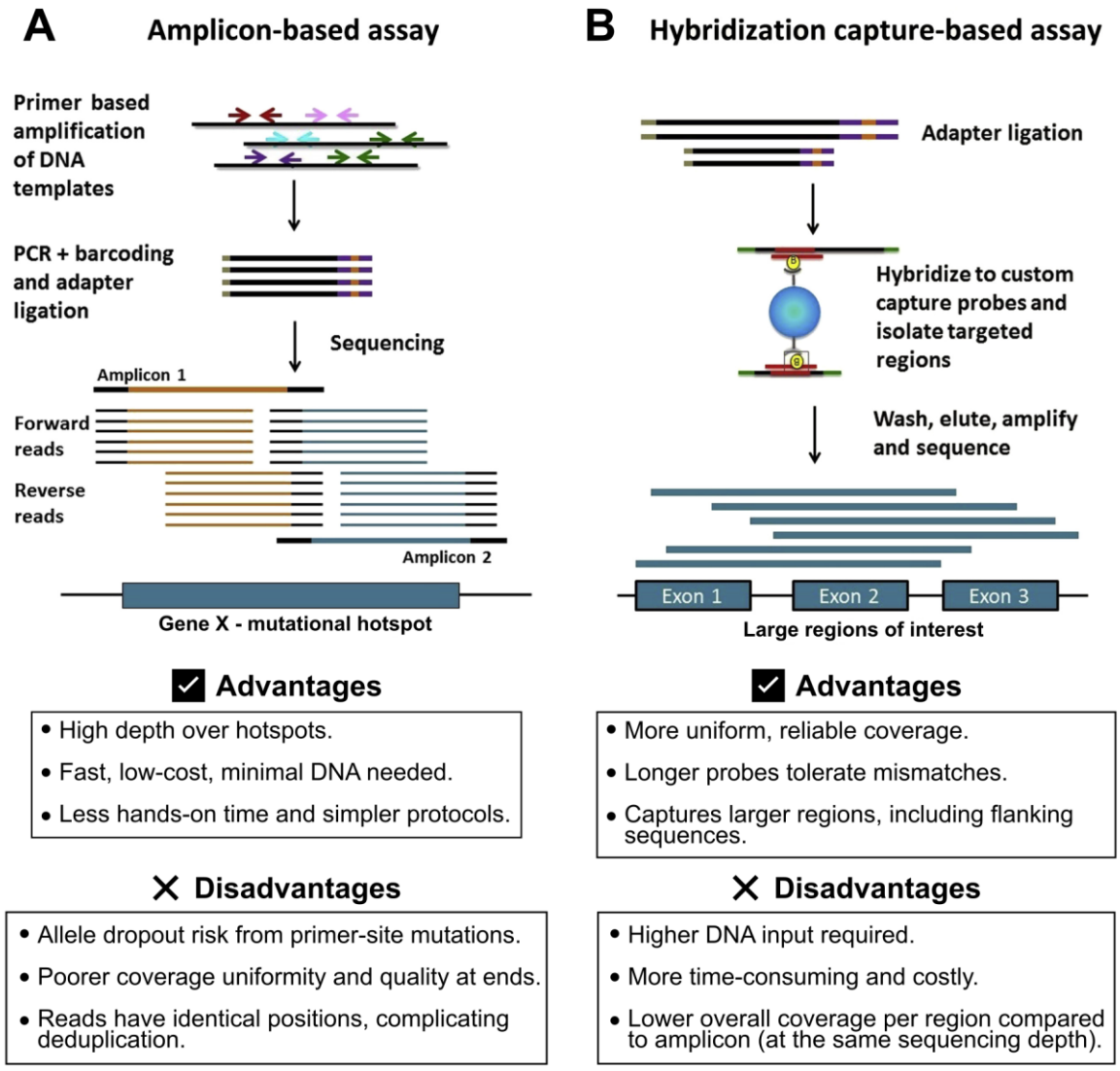


Figure 4. Overview of amplicon-based and hybridization capture-based enrichment protocols.
(**A**) Amplicon-based assay: enrichment achieved by PCR-Amp using primers targeting regions of interest. (**B**) Hybridization capture-based assay: enrichment performed using biotinylated probes complementary to target regions and isolation of captured fragments via streptavidin magnetic beads. Adapted from *Jennings et al., 2017* [11].

Because targeted panels require less sequencing depth compared to WES or WGS, it is common in clinical routine to pool multiple libraries together for sequencing, improving efficiency and reducing per-sample costs. This process (i.e., multiplexing) relies on the addition of sample-specific short sequences called barcodes or indexes (typically 8–12 base pairs [bp]), ligated to each end of the inserts. After sequencing, these barcodes are used during demultiplexing to assign reads back to their original samples[9,13].

Another element often incorporated into library preparation is the Unique Molecular Identifier (UMI)[11,12]. UMIs are short, random sequences ligated to each fragment before PCR-Amp. As shown in **Figure 5**, UMIs enable identification of unique original molecules, helping to distinguish true duplicates from unrelated reads with identical start and end positions[12]. This is particularly important for low-input or degraded DNA samples (such as from FFPE or cfDNA), where PCR duplicates and

sequencing artifacts are common. By leveraging UMIs, pipelines can achieve more confident variant calling at lower VAFs and perform accurate deduplication even for amplicon-based libraries, mitigating the risk of overestimating read coverage[11,12].

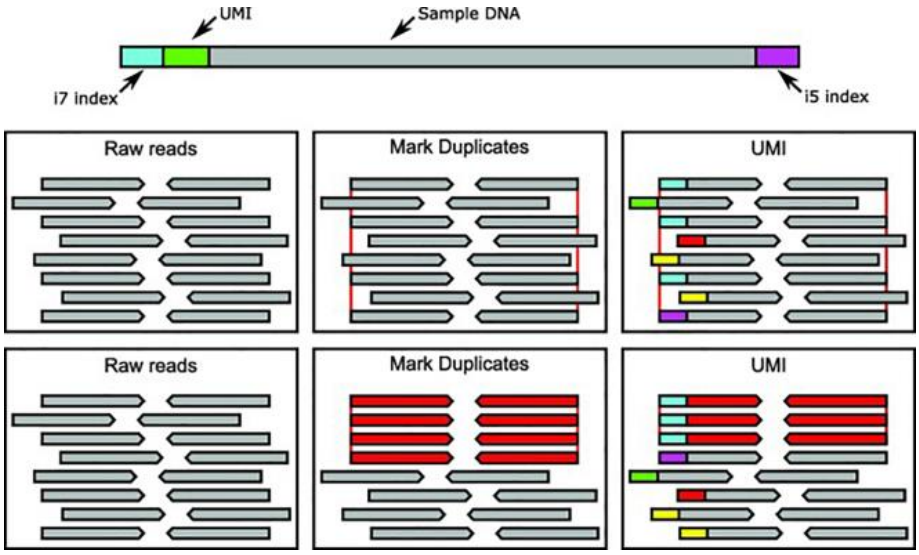


Figure 5. Use of UMIs to identify duplicates in NGS data.
UMIs are incorporated immediately before or after the DNA insert, while indexes (i7 and i5) enable sample identification. Duplicate reads appear similar to true unique reads (raw reads), but they represent technical noise that can inflate coverage estimates. Two deduplication strategies are illustrated: (i) detection using only start and end coordinates (red lines), and (ii) detection using coordinates plus UMI tags (colored segments), which allows more accurate identification of true duplicate molecules among reads sharing the same coordinates. Reads identified as duplicates are shown in red; reads retained as unique are shown in grey. Reproduced from *Bewicke-Copley et al., 2019*[12].

1.2.2.3. Sequencing

Once the sequencing library is prepared, the process of determining the order of nucleotides—adenine (A), thymine (T), guanine (G), and cytosine (C)—that make up the DNA molecule, known as sequencing, is performed[9–13]. Over the past decades, advances in sequencing technologies have unlocked the ability to interrogate molecular genetics at unprecedented depth. These technologies are commonly categorized into three generations (**Figure 6**)[9,39]:

- First-generation sequencing, primarily represented by Sanger sequencing, is based on sequencing individual DNA molecules.
- Second-generation sequencing (commonly referred to as NGS) enables massive parallel sequencing of millions of fragments, revolutionizing throughput and reducing costs.
- Third-generation sequencing allows the direct sequencing of native DNA molecules without PCR-Amp, generating much longer reads compared to previous technologies (typically <500 bp in NGS vs. >10,000 bp in long-read platforms).

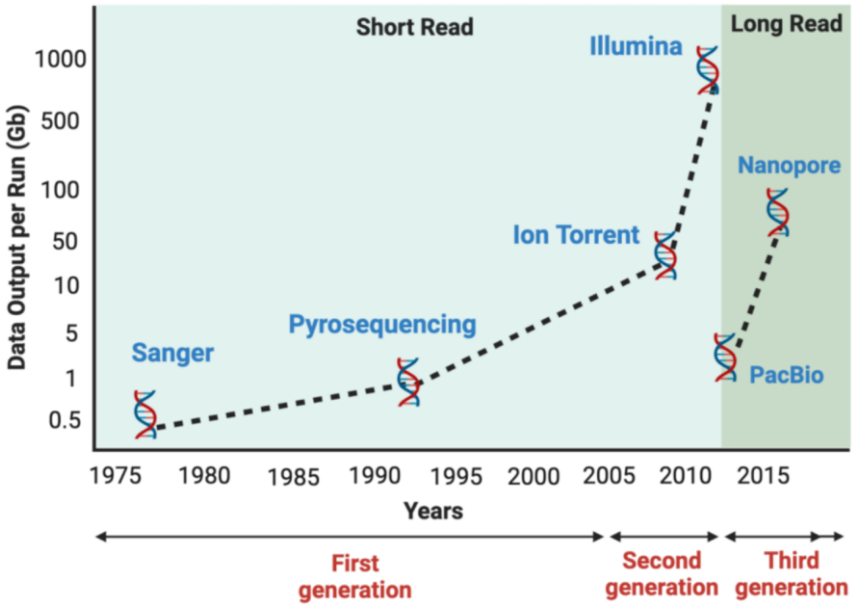


Figure 6. Evolution of NGS platforms.
Development of sequencing generations over time (x-axis). The y-axis indicates the amount of data generated per run in gigabases. Adapted from *Satam et al., 2023* [39].

The most widely used platforms—whose data are analyzed in this thesis—are Illumina and Ion Torrent:

- Illumina sequencing relies on sequencing-by-synthesis (SBS). As illustrated in **Figure 7**, this approach resembles Sanger sequencing: a denatured DNA template is extended by DNA polymerase using fluorescently labeled nucleotides that terminate synthesis. After detection of the incorporated base, Illumina SBS removes the terminator group, permitting continued extension and enabling base-by-base sequencing in cycles. Millions of DNA fragments are immobilized on a solid substrate via their ligated adapters and sequenced in parallel, with cycle number (typically 75–150) determining read length[9,39].
- Ion Torrent sequencing also performs real-time nucleotide incorporation but uses semiconductor technology. Each incorporated nucleotide releases a hydrogen ion, producing a detectable pH change that generates a voltage signal corresponding to the base identity[9,39].

Although both platforms yield similar performance, Ion Torrent has limitations in accurately resolving homopolymer tracts due to difficulties distinguishing voltage changes from multiple identical nucleotides incorporated in succession. Conversely, Illumina sequencing generally yields more reads per run but may be more expensive and slower, while Ion Torrent workflows are often faster and more cost-effective—especially for smaller targeted panels[11].

Modern sequencers feature multiple flowcell lanes, allowing independent processing of different samples or runs[9]. When a single sample's reads are distributed across lanes, the data must be merged during pre-processing.

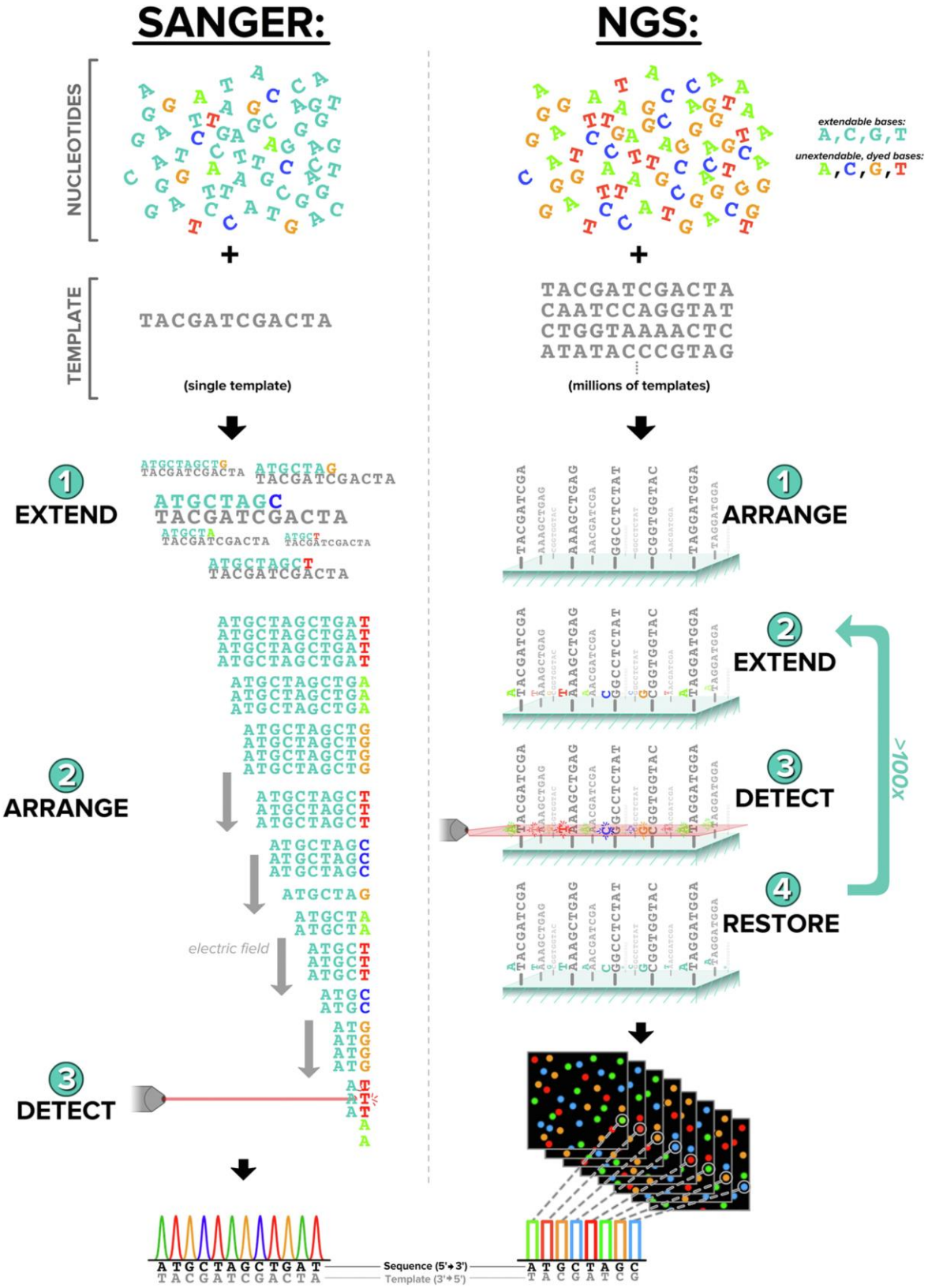


Figure 7. Overview of Sanger and Illumina sequencing processes.
The traditional Sanger sequencing workflow is depicted on the left. The Illumina-based NGS workflow is shown on the right. Both methods rely on the detection of fluorescently labeled nucleotides incorporated during DNA synthesis. Unlike Sanger sequencing, Illumina sequencing is reversible—allowing continuous base detection—and is massively parallelized across millions of templates. Reproduced from *Larson et al., 2023* [9].

Another technical consideration is the choice between single-end and paired-end sequencing. Single-end sequencing reads DNA from one end of the insert, whereas paired-end sequencing reads from both ends (forward direction usually labeled as “R1” and reverse as “R2”). Paired-end reads improve mapping accuracy and coverage and enhance sensitivity for detecting structural variants (SVs), but they can be more time- and resource-intensive[9].

In summary, each platform and assay have specific considerations—including target size, variant type and complexity, turnaround time, technical support, and bioinformatics requirements—that determine its suitability for clinical or research applications. For example, in this thesis, two commercial pan-cancer NGS panels were used with distinct focuses:

- Illumina: Hybrid-capture-based enrichment of a large gene set, sequenced as paired-end reads on the Illumina platform. This approach prioritizes comprehensiveness and accuracy over cost, input quantity, or turnaround time.
- Thermo Fisher (Ion Torrent): Amplicon-based capture targeting fewer genes, primarily hotspot regions, with single-end reads. This configuration emphasizes streamlined workflows, shorter turnaround, and lower input and cost—features well-suited to routine clinical diagnostics.

1.2.3. Bioinformatics workflow: from sequence generation to clinical insights

NGS assays generate massive volumes of complex, multidimensional data that require sophisticated computational methods to convert raw sequencing reads into clinically meaningful insights[9,34,40,41]. In this context, bioinformatics expertise is essential to manage, process, and interpret these datasets by applying specialized informatics techniques. A bioinformatic pipeline refers to a structured collection of algorithms and tools that are executed sequentially to analyze NGS data in a standardized and reproducible manner[36,42]. These pipelines are designed to handle specific data formats and associated metadata, systematically transforming them through a series of processing steps. While pipelines can be adapted to individual laboratory requirements and platform specifications, clinical NGS workflows generally follow a common structure composed of the following major stages: sequence generation and pre-processing, sequence alignment, variant calling, variant annotation, variant prioritization, and visualization and reporting[9,34,40,41]. The following sections will describe each of these components in detail, highlighting their objectives, methodologies, and implications for downstream clinical interpretation.

1.2.3.1. Sequence generation and pre-processing

Sequence generation—commonly referred to as base calling—is the process by which raw sensor data (e.g., optical or electrical signals) from the sequencing instrument are translated into a nucleotide sequence for each DNA fragment[9,42]. Each platform produces base call data in proprietary formats. For example, Illumina sequencers generate binary base call (BCL) files that store the raw fluorescence intensities, the interpreted nucleotide calls (A, T, G, C), and their associated quality metrics (Q-scores)[9,42]. These BCL files must be converted into a standardized format suitable for downstream bioinformatics processing. The most common output format is the FASTQ file, which is generally considered the starting point (i.e., raw sequencing data) for analysis[9,42]. In contrast, Ion Torrent platforms export base calls in unmapped Binary Alignment Map (uBAM) file format[43]. Unlike FASTQ, the Binary Alignment Map (BAM) format can also store platform-specific flow signal data, which some Ion Torrent pipelines use for downstream steps such as variant refinement. **Table 3** provides an overview of these formats and other commonly used files in NGS bioinformatics workflows[9,42,43].

Table 3. Common bioinformatics file formats.
This table summarizes widely used file formats in NGS bioinformatics workflows. Each format is listed along with its typical file extension(s), coordinate system (0-based or 1-based), and a brief description. Based on *Larson et al., 2023* [9].

File format	File extension	Coordinate system	Description
Binary base call (BCL)	.bcl	-	Binary files that store raw intensity measurements from Illumina sequencers. These files are demultiplexed and converted into FASTQ format before downstream analysis.
FASTQ	.fastq, .fq	-	Text-based format containing nucleotide sequences and their corresponding base quality scores. Commonly used as the starting point for read processing.
FASTA	.fasta, .fa	-	Text-based file format used to store reference genome sequences or other nucleotide sequences.
Browser Extensible Data (BED)	.bed	0-based	Tab-separated values (TSV) format specifying genomic intervals and optional annotations. Used for defining regions of interest.
Sequence Alignment Map (SAM)	.sam	1-based	TSV text file containing sequencing reads aligned to a reference genome.
Binary Alignment Map (BAM)	.bam	0-based	Binary compressed version of a SAM file. Standard format for storing and exchanging aligned read data.
Compressed Reference-oriented Alignment Map (CRAM)	.cram	0-based	Compressed format similar to BAM but optimized for storage by saving only differences between reads and the reference genome.
Variant Call Format (VCF)	.vcf	1-based	TSV format used to store identified variants and their annotations.
Gene Transfer Format (GTF)	.gtf	1-based	TSV format describing gene structure annotations (e.g., exons and transcripts).

FASTQ files are plain-text files in which each sequencing read is represented by a four-line structure[9,11,42] (**Figure 8**):

- Sequence identifier: A header line starting with the "@" symbol that uniquely labels each read and often includes information such as the instrument, flowcell, lane, and optionally barcodes like UMIs. In paired-end sequencing, two FASTQ files are generated with matching identifiers distinguished by a read direction suffix (e.g., /1 and /2). Additional metadata (e.g., sample name, index, or read length) may be appended.
- Nucleotide sequence: The called bases (A, T, G, C, or N for ambiguous calls).
- Separator: Usually a "+" sign marking the start of the quality string.
- Quality scores: A string of American Standard Code for Information Interchange (ASCII)-encoded Phred Q-scores representing the estimated error probability for each base. For example, Q=30 corresponds to a 0.1% chance of error (99.9% accuracy).

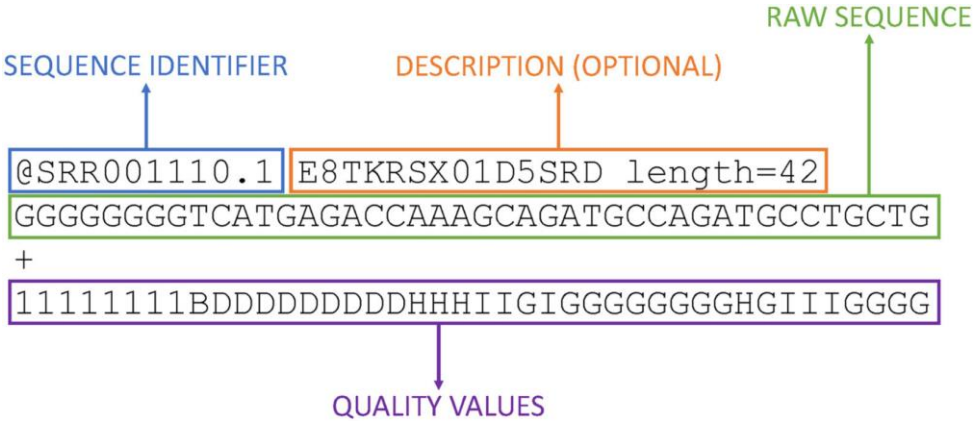


Figure 8. Example of a FASTQ read entry.
Illustration of a single read entry highlighting the four-line structure within a FASTQ file: the read identifier, nucleotide sequence, separator, and encoded base quality scores. Reproduced from *Larson et al., 2023* [9].

During this step, demultiplexing is typically performed to assign reads to their corresponding samples based on the index sequences ligated during library preparation. A potential issue is index hopping, where indexes are incorrectly assigned to the wrong sample[10,13]. This can be mitigated using dual indexing, in which two independent index sequences (e.g., i7 and i5) are applied (**Figure 5**)[10].

Once sample-specific FASTQ files have been generated, raw reads are pre-processed to retain only the relevant insert sequences by[12,36]:

- Trimming adapter sequences, UMIs, and low-quality bases (especially from the 3′ ends).
- Removing contaminant sequences, such as reads derived from non-human organisms.
- Filtering out excessively short reads that would map ambiguously.

Sequencing quality control (QC) is also conducted at this stage[10,12,36]. QC checks typically include:

- Assessing base quality score distributions to identify systematic biases or degradation.

- Inspecting per-base composition to detect contamination or technical artifacts (e.g., residual adapters, sample cross-talk).
- Estimating guanine-cytosine (GC) content to confirm that it matches the expected range for human DNA. Deviations (such as an unexpected secondary GC peak) can indicate contamination by non-human DNA.

1.2.3.2. Sequence alignment

The next step in the NGS workflow is the alignment of sequencing reads to a reference genome (i.e., read mapping), which aims to determine the most likely genomic position of each fragment, while accounting for natural genetic variation and sequencing errors[10,12,36]. Short-read aligners are designed to efficiently map millions of reads to the reference genome by using pre-built indexes that enable rapid pattern matching. For RNA data, where reads span joined exons, splice-aware aligners are required to accurately map reads crossing exon–exon boundaries[9,10]. As determining the origin of each read is critical for understanding the sequenced genetic information, this step must be highly accurate. However, it is also computationally intensive and time-consuming, as each read is compared to the entire genome[9,10,36].

The reference genome is typically stored in the FASTA file format (**Table 3**), which contains a series of entries with a header line (starting with ">") and the corresponding nucleotide sequence. The current standard in clinical applications is the Genome Reference Consortium (GRC) Human Build 38 (GRCh38)—also referred to as Human genome build 38 (hg38)—released by the GRC in 2013 (with the latest patch GRCh38.p14 from 2022)[9,10,40]. While this build includes significant improvements over its predecessor GRCh37/hg19 (2009), the latter remains widely used, necessitating compatibility through coordinate conversion, commonly referred to as liftover. Other alternatives include Telomere-to-Telomere CHM13 (T2T-CHM13)—the first gapless haploid genome assembly—which improves the representation of difficult regions, but does not reflect human population diversity[44]. To address this, the Human Pangenome Reference Consortium is developing a multi-reference genome derived from diverse individuals, aiming to better represent population-specific genomic variation[45].

Depending on the outcome of the alignment, reads can be classified as[46]:

- Mapped reads: Successfully aligned to a unique position in the genome. Most reads should be mapped, and the mapping rate is a common QC metric.
- Unmapped reads: Failed to align to any region. They may originate from novel or non-human sequences, repetitive regions, or reflect structural variations.

- Clipped reads: Partial alignments where one end of the read is not mapped. These can be soft-clipped (retained but not aligned) or hard-clipped (removed). Clipping may reflect low-quality bases or true biological alterations (e.g., InDels, SVs).
- Multi-mapping reads: Align equally well to multiple locations. These often originate from repetitive elements or paralogous genes.

Alignment outputs are saved in specific file formats[9,36,43,46] (**Table 3**):

- Sequence Alignment Map (SAM): A plain-text, tab-separated values (TSV) format that includes aligned and unaligned reads. It contains a multi-line header (starting with "@") with metadata (e.g., software, reference genome), and per-read information (**Table 4**).
- BAM: The binary version of SAM, optimized for storage and processing. BAM files are compressed, indexable, and compatible with most downstream tools and genome browsers for manual inspection.
- Compressed Reference-Oriented Alignment Map (CRAM): A more compressed format that stores only differences relative to the reference genome. It enables significant file size reduction, although it requires specialized tools and may involve lossy compression, potentially discarding non-variant reads.

Table 4. Summary of mandatory SAM file fields.
This table summarizes the eleven mandatory fields of SAM files. Adapted from the *Sequence Alignment/Map Format Specification* [46].

No.	Field	Type	Description
1	QNAME	String	Query name. Typically the read identifier, matching the one in the FASTQ file.
2	FLAG	Integer	Bitwise flag representing read attributes (e.g., read is paired, properly aligned, unmapped, etc.).
3	RNAME	String	Reference name (e.g., chromosome) where the read is aligned. Set to "*" for unmapped reads.
4	POS	Integer	1-based leftmost position of the aligned read. Set to 0 for unmapped reads.
5	MAPQ	Integer	Mapping quality. Phred-scaled score estimating the probability that the alignment is incorrect. Set to 255 when not available.
6	CIGAR	String	Encodes the alignment of the read to the reference using a sequence of operations (e.g., "M": alignment match, "I": insertion, "D": deletion, "S": soft-clipped, "H": hard-clipped).
7	RNEXT	Integer	Reference name of the mate/next read. Set to "*" if unavailable, or "=" if identical to RNAME.
8	PNEXT	Integer	1-based position of the mate/next read. Set to 0 if unavailable.
9	TLEN	String	Observed template length (i.e., insert size for paired-end reads). It can be negative depending on orientation.
10	SEQ	String	Read sequence. Set to "*" if sequence is not stored; can be "=" if identical to reference.
11	QUAL	String	ASCII-encoded Phred base quality scores for each base in SEQ. Set to "*" if unavailable.

Read alignment enables the coverage calculation which directly impacts the VAF and the sensitivity and reliability of variant detection. High coverage is essential for detecting low-VAF somatic variants, especially subclonal mutations present in only a subset of tumor cells. For this reason,

targeted NGS panels are ideal for clinical applications, as they provide deep coverage in predefined regions of interest[9,10].

Before aligned reads can be used for variant calling, several post-alignment processing steps are usually performed to generate analysis-ready BAM files[9,10,12,36] (some steps may be included within variant calling tools):

- Deduplication: PCR duplicates, originating from over-amplified DNA fragments (especially in FFPE or low-input samples), are removed. Duplicates are typically identified by identical start and end coordinates, although UMIs (**Figure 5**) can enhance accuracy by tagging original DNA fragments.
- Local realignment: Reads near potential InDels are locally realigned to improve alignment accuracy and facilitate InDel detection.
- Bias correction: Systematic errors in base quality scores can be recalibrated using known variant datasets, improving variant calling precision.
- BAM QC: In addition to sequencing QC, various metrics assess alignment quality (**Figure 9**). These include the mapping rate, on-target rate (for targeted panels), insert size distribution, read length, duplication rate, and coverage statistics (e.g., mean, median, or percentage of target bases above specific coverage thresholds).

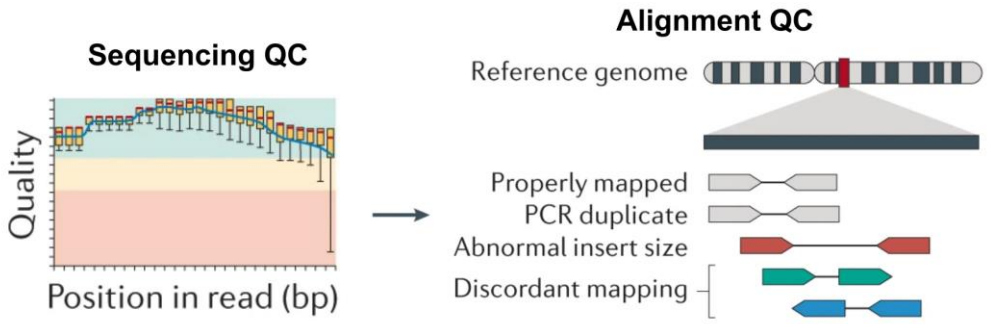


Figure 9. Sequencing and alignment QC of reads.
Read QC is assessed before and after the alignment to the reference genome. Based on *Cortés-Ciriano et al., 2022* [10].

Several biological and technical factors can compromise alignment, including genomic complexity (e.g., repeats, segmental duplications), sequencing errors, and limitations of the reference genome. Misalignments may lead to false positive (FP) variants, so it is essential to monitor these issues and apply corrective strategies during downstream analysis to ensure robust results[10].

1.2.3.3. Variant calling

Once the genomic positions of sequencing reads have been established, the next step is to identify genetic differences between the tumor sample and a reference genome—a process known as variant calling. This step enables the detection of various types of genomic alterations, such as small variants (SNVs, InDels), CNAs, and SVs[9,42].

The resulting data is typically stored in a Variant Call Format (VCF) file (**Table 3**), a plain-text, TSV file widely used for representing genomic variants. As shown in **Figure 10**, VCF files consist of a header section—beginning with a "#"—which includes metadata such as file format version, reference genome, and descriptions of quality metrics, followed by a series of variant entries with defined fields described in **Table 5**[9,38].

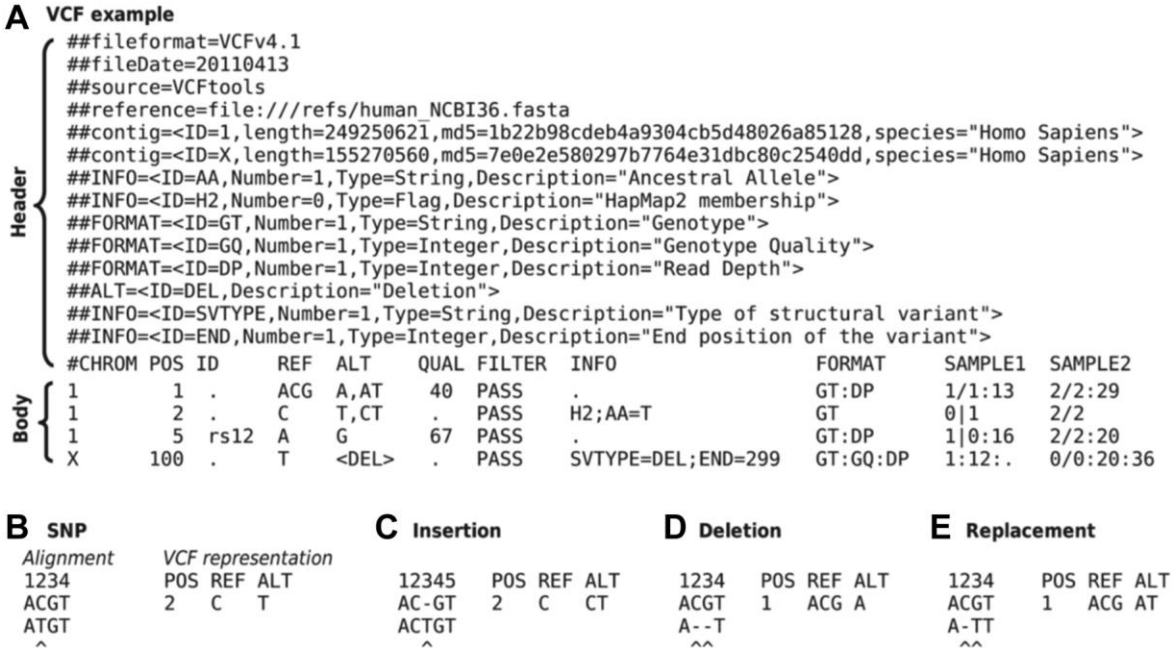


Figure 10. Structure of a VCF file.
(**A**) Example of a VCF file showing the header section and several variant records. The header contains metadata about the file and the reference genome, while each row in the records section corresponds to a detected variant with structured fields. (**B-E**) Illustrative examples of sequence alignments and their corresponding VCF representations for different types of small variants. Reproduced from *Larson et al., 2023* [9].

Table 5. Summary of fixed VCF file fields.
This table summarizes the mandatory fields in VCF files used to describe genomic variants. Missing values in any field are indicated by a dot. Adapted from the *Variant Call Format Specification* [38].

No.	Field	Type	Description
1	CHROM	String	Chromosome or contig name where the variant is located.
2	POS	Integer	1-based position of the variant on the chromosome.
3	ID	String	Variant identifier. Commonly includes dbSNP rsID if available.
4	REF	String	Reference allele(s). The base at POS is the first base in this string.
5	ALT	String	Alternate allele(s) or symbolic SVs (e.g., DEL, INS, DUP, INV). Multiple values separated by commas for multiallelic sites.
6	QUAL	Float	Phred-scaled quality score assigned to the variant call.
7	FILTER	String	Filter status of the variant. "PASS" if it passes all quality filters; otherwise, a semicolon-separated list of failed filters.
8	INFO	String	Semicolon-separated list of key-value pairs with additional variant annotations. Format defined in the VCF header.
9	FORMAT	String	Colon-separated list of fields describing sample-specific data in the next columns.

Small variants, including SNVs and InDels, are the most common types of somatic alterations found in tumors. These variant classes are typically identified simultaneously by specialized software tools known as variant callers, and the term variant calling often refers specifically to their detection[4,9,10,12,40–42]. This step is one of the most computationally intensive in the pipeline, as it

involves comparing each base in the aligned reads to the reference genome to identify deviations. To reduce the likelihood of reporting FPs, variant callers assign quality metrics to each call and apply filtering criteria—either automatically or as a post-processing step—to flag or exclude low-confidence variants. Common filtering parameters include base quality, mapping quality, strand bias, read position bias, and the presence of multiple nearby alternative alleles, among others. The resulting file may include both pass and filtered variants, typically annotated in the VCF using the FILTER field[42].

The core objective of the variant calling process is to distinguish true genetic variants from sequencing or alignment artifacts. An illustrative example of how sequencing reads are aligned to the reference genome and how small variants are detected in different NGS contexts is shown in **Figure 11**. In this context, SNV refers specifically to somatic variants, while single-nucleotide polymorphism (SNP) denotes germline origin. Several factors can influence the accuracy of small variant detection (see *1.3. Bioinformatics challenges in the analysis of somatic NGS panel* for further detail)[10]:

- DP and VAF: Both variant metrics are a key determinant of variant confidence. Germline SNPs typically exhibit VAFs near 100% (homozygous) or ~50% (heterozygous). In contrast, somatic SNVs often have lower VAFs, influenced by TP, ploidy, and intra-tumor heterogeneity. For example, WGS (**Figure 11A**) provides uniform coverage across the genome, which supports the reliable detection of clonal variants, but has limited sensitivity for subclonal mutations. In contrast, targeted NGS panels (**Figure 11B**) achieve higher sequencing depth over specific regions, enhancing sensitivity for low-frequency subclonal variants, although the coverage variability across targets can hinder accurate estimation of copy number (CN).
- Duplicate reads: PCR duplicates should be removed to avoid overestimating variant-supporting reads and reduce the influence of potential artifacts.
- Tumor-only calling: In paired tumor-normal analyses, reads from matched normal tissue help differentiate true somatic variants from germline alterations and sequencing artifacts. However, normal samples are often unavailable in routine clinical workflows, necessitating the use of alternative filtering strategies to suppress germline variants and technical noise.
- Systematic biases: Certain artifacts may display strand bias, where a variant is disproportionately supported by reads from one strand (forward or reverse). Such biases are commonly associated with FPs and should be carefully evaluated or filtered out.

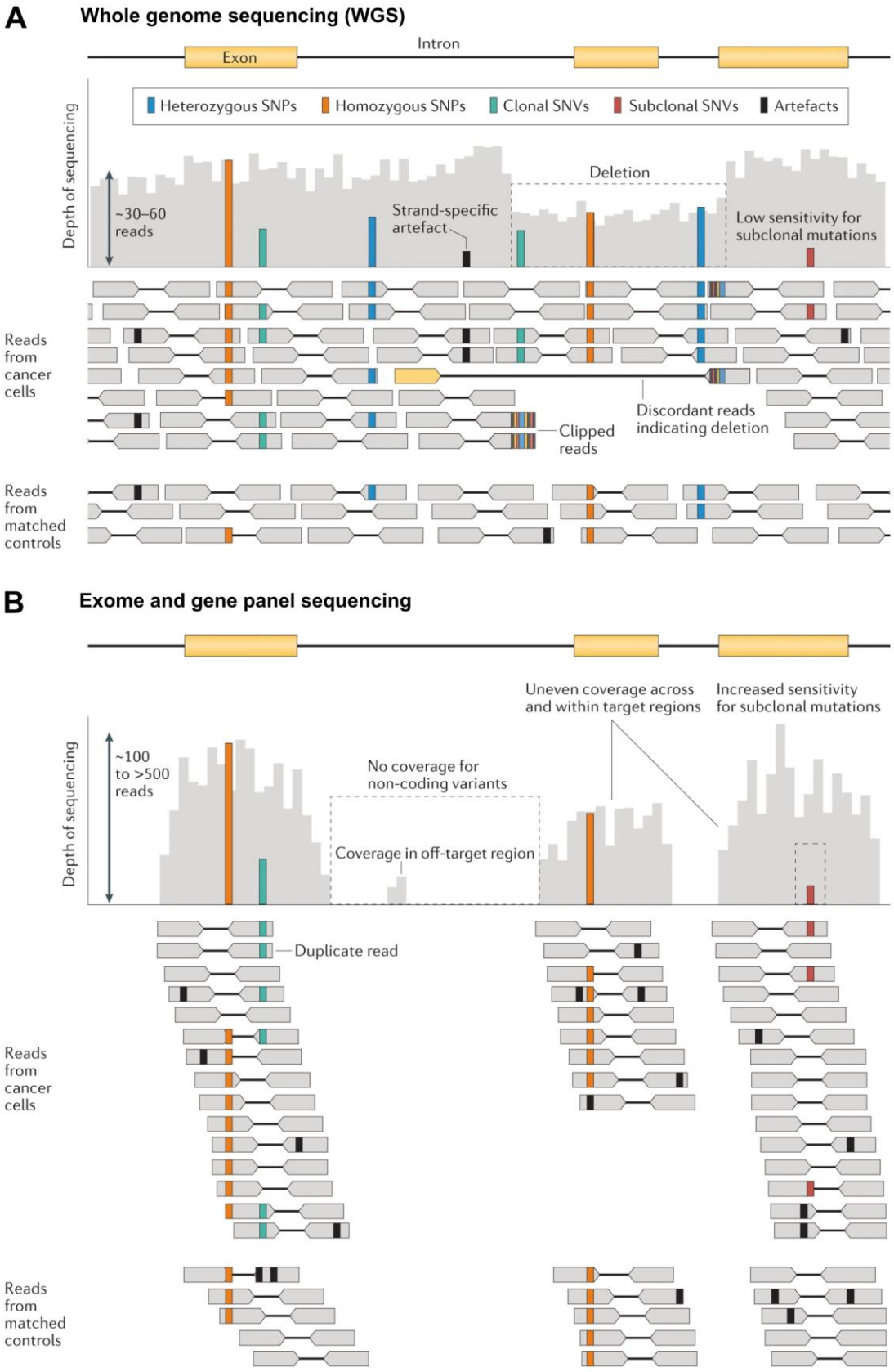


Figure 11. Illustration of variant identification in different NGS contexts.
Visual representation of aligned sequencing reads within a gene locus under (**A**) WGS and (**B**) exome or gene panel sequencing approaches. Variants and sequencing artifacts are highlighted with distinct colors, and annotations indicate key events. The figure illustrates how sequencing depth, coverage uniformity, and context affect variant detection sensitivity. Reproduced from *Cortés-Ciriano et al., 2022* [10].

CNAs are another important class of somatic events typically assessed in clinical bioinformatics workflows. Their identification is primarily based on coverage differences between a tumor sample and a reference, which may be a matched normal sample or a reference pool of samples. An increase or decrease in sequencing coverage across a genomic region is indicative of an amplification (AMP) or deletion (DEL), respectively (**Figure 12A**). Some advanced approaches integrate allelic imbalance information using the B-allele frequency (BAF) of heterozygous SNPs, which enables improved CNA resolution and the detection of copy-neutral loss of heterozygosity (LOH) events—situations where one allele is deleted and the other is amplified, leading to an unbalanced but diploid state (**Figure 12A**). However, the resolution of these methods is limited in targeted gene panels due to the reduced number of heterozygous SNPs and uneven coverage across the genome, which constrains their ability to robustly distinguish such events[10,36].

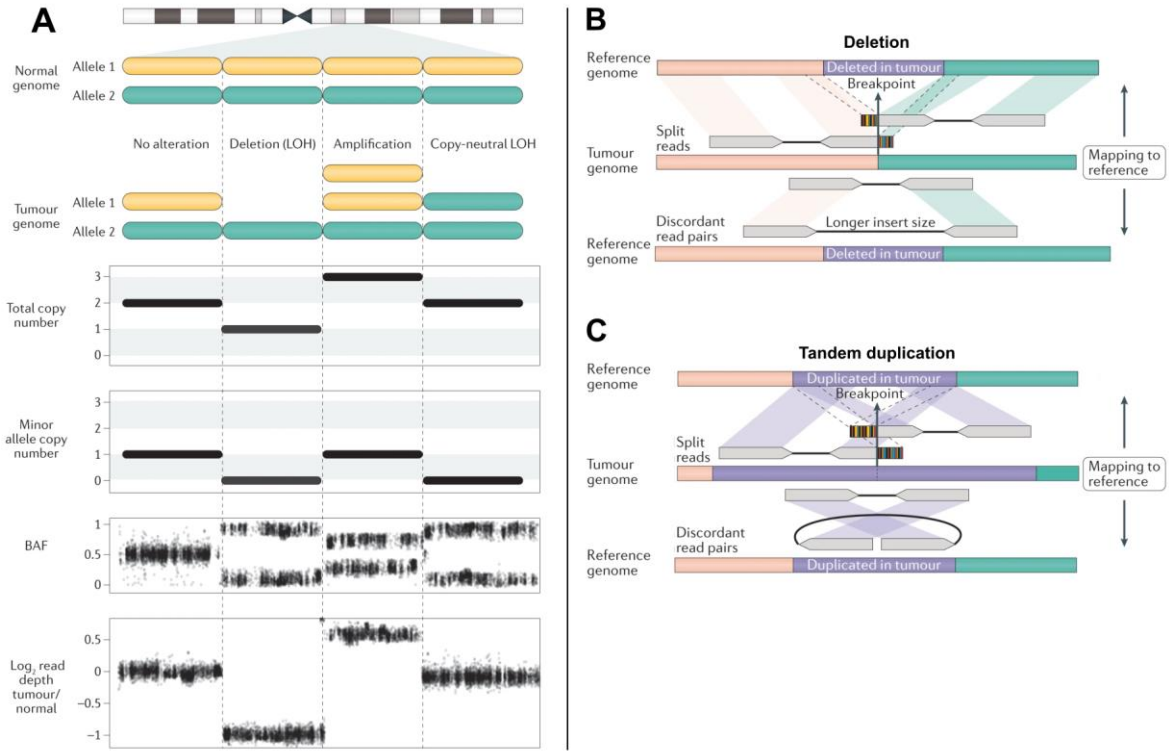


Figure 12. Detection of CNAs and SVs using different sequencing-based approaches.
(**A**) Schematic representation of how distinct CNAs affect depth and BAF profiles. (**B**) Example of a deletion detected by discordant read pairs and split reads spanning the deleted region. (**C**) Example of a tandem duplication identified by paired-end reads with unexpected orientation or insert size, and split reads aligning at the duplication junction. Based on *Cortés-Ciriano et al., 2022* [10].

There are several systematic biases introduced during library preparation and sequencing that can affect the accurate calculation of coverage and, consequently, impact the CNA detection in coverage-based analyses[10,47]:

- GC content: Regions with extremely high GC content tend to exhibit reduced efficiency in hybridization, PCR-Amp, and sequencing. This results in lower observed coverage compared to regions with balanced GC content.
- Repetitive sequences: Genomic regions containing repetitive elements—such as microsatellites or segmental duplications—pose challenges for short-read sequencing and accurate read mapping, leading to decreased mappability and artificially reduced coverage.
- Target density bias: Uneven coverage across target regions is a known limitation in targeted gene panels, particularly those using hybridization capture. Two edge-related effects can occur: (i) a negative bias at the borders of target regions due to incomplete probe hybridization—commonly referred to as the shoulder effect, and (ii) a positive bias in flanking regions when adjacent targets are close enough for their capture signals to overlap, causing inflated coverage values.

Other approaches for detecting CNAs—and more broadly, SVs—rely on the analysis of read mapping patterns, particularly split (or clipped) reads and discordant read pairs (**Figure 12B**). Split reads are individual sequencing reads that partially align to two distinct genomic regions, typically corresponding to breakpoints where structural changes occur. Similarly, discordant read pairs are paired-end reads whose mapping characteristics (e.g., orientation, insert size, or genomic distance) deviate from the expected pattern, indicating a potential structural rearrangement. These signals are particularly useful for identifying focal events (e.g., deletions, duplications, inversions, or translocations) which may not always result in clear coverage imbalances[10,36].

From a clinical perspective, the most relevant SVs are those that give rise to actionable oncogenic gene fusions or splicing aberrations, as they can drive tumorigenesis and represent therapeutic targets. In the context of clinical NGS panels, these events are more reliably detected using RNA-seq data. Unlike DNA sequencing, RNA-seq captures only the transcribed regions of genes, thereby skipping large intronic regions and improving the detection sensitivity for fusion transcripts. In this approach, reads are aligned to the transcriptome or to a genome-guided transcript model to identify two key signals: split reads that directly map to fusion junctions (i.e., across exons from different genes), and discordant read pairs that span fusion breakpoints but map to non-adjacent gene regions. This strategy enables the precise detection of both known and novel gene fusions, as well as splicing isoforms with potential clinical significance [9,10].

1.2.3.4. Variant annotation and prioritization

Once genomic variants have been identified, the next critical step is to interpret their biological and clinical relevance. Raw variant calls, regardless of their type, lack the contextual information required for clinical decision-making. Therefore, a comprehensive annotation process is essential to enrich each variant with relevant metadata, such as its genomic context, population frequency, known pathogenicity, and potential therapeutic associations. This annotated information forms the foundation for subsequent filtering and prioritization strategies that aim to highlight the most clinically relevant alterations for diagnosis, prognosis, and therapy selection[4,9,11].

Annotation of small variants (SNVs and InDels) typically includes[4]:

- Genomic context: Identification of the affected gene and genomic region (e.g., coding vs. non-coding), as well as specific features such as exons, introns, splice sites, or regulatory elements.
- Predicted functional consequences: Evaluation of the potential impact on protein function (e.g., synonymous, missense, nonsense, frameshift, or splice site variants), often supplemented by *in silico* prediction tools.
- Population frequency: Cross-referencing large-scale population databases (e.g., gnomAD) to distinguish common polymorphisms and infer germline origin.
- Clinical databases: Integration of clinical interpretation data from curated resources (e.g., ClinVar, COSMIC, OncoKB, and CIViC) which provide information on pathogenicity or drug sensitivity.

Annotation of CNAs typically involves[4,29]:

- Gene role in cancer: Identification of affected oncogenes or TSGs using curated cancer gene lists to infer potential biological relevance.
- Clinical interpretation: Evaluation of the functional consequence of focal or arm-level events (e.g., *ERBB2* AMP in breast cancer or *CDKN2A* DEL in glioma), often supported by known clinical associations.

Annotation of gene fusions and splicing variants includes[4,29]:

- Fusion structure: Assessment of the fusion partner genes, reading frame, and functional domains to evaluate whether the event is likely to be oncogenic.
- Database matching: Comparison with curated fusion databases to identify known oncogenic rearrangements.
- Therapeutic relevance: Identification of clinically actionable fusions (e.g., *ALK*, *ROS1*, *NTRK* genes) or splicing events (e.g., *MET* exon 14 skipping [*METex14*]) associated with targeted therapies.

Once annotated, variant prioritization becomes essential to identify the most relevant alterations, particularly in clinical settings where only a subset of variants is actionable. This process generally involves:

- Variant confidence: Evaluation of calling metrics (e.g., DP, AD, AF) and variant context (e.g., difficult regions) to estimate reliability.
- Biological and oncogenic relevance: Prioritization of variants affecting well-established cancer genes, functional domains, or hotspot regions, often guided by classification frameworks such as the Clinical Genome Resource (ClinGen)/Cancer Genomics Consortium (CGC)/Variant Interpretation for Cancer Consortium (VICC) Standard Operating Procedure (SOP) guidelines[48].
- Clinical significance: Assignment of clinical relevance tiers based on established guidelines such as the AMP/American Society of Clinical Oncology (ASCO)/College of American Pathologists (CAP)[4] consensus recommendations and the ESMO Scale for Clinical Actionability of Molecular Targets (ESCAT)[49] framework.

To facilitate this process, automated pipelines are used to apply dynamic filtering criteria and generate structured outputs such as tiered variant classifications, which can then be summarized in interactive reports. These outputs support downstream interpretation in multidisciplinary settings, such as MTB**s**, where clinical teams evaluate the potential diagnostic, prognostic, and therapeutic implications of each case[48].

1.2.3.5. Complex genomic biomarkers

Beyond individual genomic alterations, a subset of complex biomarkers derived from broader mutational patterns or genome-wide instability has emerged as highly relevant in precision oncology. These complex genomic biomarkers provide important insights into tumor pathophysiology and are increasingly used to predict therapeutic response, particularly to immune checkpoint inhibitors (ICIs), DNA-damaging agents, and targeted therapies. Unlike discrete variants, their detection requires integrative bioinformatic analyses across multiple genomic features, often demanding specific computational strategies and sufficient sequencing breadth or depth. However, despite their clinical potential, the standardization of methods for their accurate assessment, especially in targeted panel settings, remains an ongoing bioinformatic and clinical challenge[5,10,36,40].

Microsatellite instability (MSI) refers to a hypermutator phenotype caused by defective DNA mismatch repair (MMR), resulting in InDel errors at microsatellite regions—short tandem repeats scattered throughout the genome. Tumors with high MSI accumulate frameshift mutations that can generate neoantigens, making MSI-High status a predictive biomarker for immunotherapy efficacy. Traditionally assessed via PCR-based assays or IHC, MSI can also be inferred from NGS data by

examining either the length variability of specific microsatellite loci or characteristic mutation patterns[5,36].

Tumor mutational burden (TMB) quantifies the number of somatic mutations per megabase (Mut/Mb) of coding DNA and is also associated with response to ICIs. While WES remains the reference method, TMB estimation in clinical settings is commonly adapted to targeted panels. In this case, the TMB score is extrapolated from the count of somatic mutations within the panel's coding region, normalized by its effective size[50].

Mutational signatures represent specific patterns of somatic single-nucleotide substitutions that reflect distinct mutagenic processes, such as environmental exposures (e.g., UV light, tobacco), enzymatic activity (e.g., APOBEC), or defects in DNA repair pathways. These signatures are defined based on the frequency of each of the 96 possible trinucleotide substitution contexts. The associated mutational processes can be discovered *de novo* from a large cohort of cancer genomes, or in case of a small set or single sample, the relative contribution for a set of predefined signatures (i.e., refitting) is obtained. This concept is extensively applied to other variant types, such as InDels, CNAs, or other rearrangements[10,40].

Homologous recombination deficiency (HRD) reflects the inability of tumor cells to faithfully repair DNA double-strand breaks via the homologous recombination repair pathway. HRD is frequently caused by biallelic inactivation of *BRCA1/2* or other HR genes and is associated with increased sensitivity to platinum-based chemotherapy and PARP inhibitors. Computational methods have been developed to quantify HRD through so-called genomic scar scores, which measure the accumulation of large-scale genomic aberrations indicative of defective DNA repair, including LOH, large-scale state transitions (LST), and telomeric allelic imbalance (TAI)[51].

1.2.3.6. Visualization and reporting

The final step in the clinical bioinformatics workflow involves transforming the processed and interpreted variant data into a structured format that supports decision-making in a clinical context. This is typically achieved through intuitive visualizations and standardized reports that summarize relevant findings, including detected variants, affected genes, clinical annotations, and suggested therapies. Effective visualization tools, such as Integrative Genomics Viewer (IGV), allow the inspection of sequence alignments, facilitating manual validation of critical findings. Reports must be clear, concise, and adapted to multidisciplinary users (e.g., oncologists, pathologists, and geneticists). They often integrate clinical classifications, evidence levels, and therapeutic implications, and should highlight clinically actionable alterations, potential resistance markers, and relevant biomarker statuses. In the context of high-throughput settings, automated reporting systems

are essential to ensure reproducibility, scalability, and turnaround time, while enabling human oversight for critical cases[4,36].

1.2.3.7. Workflow management and containerization

The increasing complexity and volume of clinical NGS data has made manual execution of bioinformatics analyses impractical, error-prone, and difficult to reproduce. To address these challenges, workflow management systems and software containerization have become essential tools in modern clinical bioinformatics[36,52,53].

Workflow managers (e.g., Nextflow, Snakemake) enable the design and execution of complex pipelines by orchestrating a series of bioinformatics tasks in a modular, scalable, and reproducible manner. These systems handle dependencies, resource allocation, parallelization, and job scheduling, and are compatible with a wide range of computing environments, including local machines, high-performance computing (HPC) clusters, and cloud infrastructures. These tools help save time, reduce errors, and ensure accuracy and reliability of the analyses[36,52,53].

In parallel, containerization technologies (e.g., Docker, singularity) ensure consistent software environments across different computational platforms. Containers encapsulate all the software, libraries, and dependencies required for each step of the workflow, avoiding conflicts and simplifying deployment. In clinical bioinformatics, containerization is particularly relevant for ensuring version control, minimizing discrepancies across institutions, and facilitating regulatory compliance[52,53].

Together, workflow management and containerization provide the technical backbone for building reproducible, auditable, and scalable clinical bioinformatics pipelines. Their integration is now considered a best practice for implementing robust NGS workflows that support routine diagnostics, regulatory requirements, and large-scale genomic data processing[36,52,53].

1.3. Bioinformatics challenges in the analysis of somatic NGS panels

Despite the widespread adoption of NGS-based cancer panels in clinical practice, the bioinformatic analysis of somatic alterations remains a multifaceted and technically demanding task. Unlike germline testing, somatic variant analysis must contend with tumor-specific complexities—including variable TP, intra-tumor heterogeneity, and the frequent use of low-quality or low-input DNA from FFPE specimens[9–13]. The limited and uneven genomic coverage typical of targeted panels further complicates the detection of certain alterations, particularly InDels, SVs, CNAs, and complex biomarkers such as TMB, MSI, or HRD[5,10,36,40].

Additional challenges stem from the lack of matched normal samples, which hinders the accurate discrimination of somatic versus germline variants and requires the adoption of alternative filtering strategies[10,12,41]. RNA-based analyses introduce further layers of complexity, from expression-dependent detection limits to degraded input quality and the intricacies of fusion transcript interpretation[9,10,54]. Technical and computational variability across platforms—along with the absence of standardized, harmonized workflows—exacerbate inconsistency in variant calling and interpretation across laboratories[9,11,42,54]. Moreover, limited automation for variant prioritization, lack of user-friendly clinical reporting tools, and persistent barriers to data sharing and interoperability restrict the broader utility of these pipelines in routine oncology[4,34,35]. Finally, existing commercial and academic solutions often fall short of meeting clinical requirements for flexibility, portability, and end-to-end reporting[40]. This section reviews each of these critical challenges in detail and outlines current efforts and strategies to overcome them.

1.3.1. Low-quality and low-input DNA

The quality and quantity of input DNA are critical factors influencing the success of NGS assays, particularly in clinical oncology where samples are often scarce and derived from FFPE tissues. Although FFPE is the standard method for long-term tissue preservation in diagnostic pathology, the fixation process causes DNA fragmentation, crosslinking, and chemical modifications (e.g., cytosine deamination), all of which introduce technical artifacts and may compromise the reliability of downstream genomic analyses[9–11,54].

Degraded or damaged DNA typically results in shorter fragment lengths and higher levels of sequencing artifacts, such as nucleotide misincorporations, chimeric reads, and PCR duplicates. These issues can compromise the detection of true somatic variants, especially low-frequency mutations, by inflating FPs or reducing sensitivity. Additionally, chemical alterations to DNA bases may interfere with primer binding during PCR-Amp or hinder adapter ligation during library preparation, ultimately reducing library complexity and target region coverage[11,54].

Low-input DNA—frequently encountered in small biopsies or cytological specimens—limit the feasibility of high-depth sequencing or technical replicates, and often lead to overamplification during library preparation, resulting in elevated duplication rates and reduced effective (unique) coverage. These conditions can severely impact variant calling performance, particularly in applications requiring high sensitivity for subclonal or actionable alterations[11,54].

To address these issues, bioinformatics pipelines must integrate specific preprocessing and filtering strategies tailored to the limitations of compromised input material[12], including:

- Duplicate read filtering: High duplication rates are a hallmark of low-complexity libraries. While duplicate removal is necessary to avoid coverage overestimation, distinguishing technical duplicates from biological duplicates can be difficult, particularly in amplicon-based assays. The use of UMIs can help to resolve this ambiguity by tagging original DNA molecules before PCR-Amp.
- Error-aware variant calling: Tools designed for FFPE-derived samples include models to detect strand bias or sequencing artifacts, helping reduce false-positive calls from chemically damaged bases (e.g., C>T transitions).
- Quality-based trimming and filtering: Removing low-quality bases (commonly from 3' ends), filtering reads with low mapping quality, and excluding variants with weak support or high strand bias improve overall specificity and variant reliability.

Despite the use of correction strategies, certain genomic regions—particularly those with extreme GC content, repetitive elements, or high fragmentation—may remain poorly covered or inaccessible, limiting confidence in negative findings. Therefore, integrating sequencing QC metrics (e.g., fragment length distributions, read quality profiles, on-target rate, duplication rate, and coverage uniformity) into the bioinformatics workflow is essential for reliable downstream interpretation and for informing the clinical confidence of reported results[36].

In summary, low-quality and low-input DNA remain significant barriers to accurate somatic variant detection in NGS panel assays. Overcoming these limitations requires a combination of optimized wet-lab protocols and dedicated bioinformatic strategies to ensure data quality, analytical robustness, and clinical utility—particularly when working with suboptimal but routinely available clinical specimens such as FFPE[54].

1.3.2. Tumor heterogeneity and low-frequency variant detection

The sensitivity of somatic variant detection in NGS panel analysis is profoundly affected by different forms of tumor heterogeneity, which can dilute or obscure the signal of clinically relevant alterations[9–11,54]. These heterogeneity sources include:

- Tissue heterogeneity (purity): Clinical tumor specimens often contain a mixture of neoplastic and non-neoplastic cells, such as stromal, endothelial, or immune cells. This lowers the overall TP and dilutes the representation of somatic variants in the sequencing data.
- Tumor cell heterogeneity (intra-tumor heterogeneity): Tumors are composed of diverse cellular subpopulations or subclones, each harboring distinct genomic alterations. Subclonal variants may only be present in a fraction of the tumor cells, resulting in lower VAFs.

These combined effects can markedly limit the sensitivity of variant detection. Somatic variants present at low VAFs may fall below the detection threshold of variant callers or be mistakenly filtered as sequencing artifacts. This is particularly relevant in routine diagnostics where matched normal samples or orthogonal validation are rarely available[9–11,54].

To improve detection under these conditions, several strategies are applied[11,12,41]:

- High-depth sequencing: Deep sequencing—especially in hotspot regions—enhances the ability to detect variants at lower VAFs. Amplicon-based approaches are particularly useful in this context but must be coupled with error suppression strategies (e.g., UMIs, strand bias correction) to maintain specificity.
- Variant calling algorithms optimized for low-VAF detection: Some tools incorporate probabilistic models and artifact filters tailored for low-frequency variant calling, especially in tumor-only contexts.

CNA detection is also impacted by TP. In low-TP, the signal from AMPs or DELs is attenuated, reducing the log2 ratio shifts and BAF deviations. Some tools attempt to adjust CNA models based on estimated TP, but these require reliable estimation of TP, which is not always available or accurate[10,47].

In summary, both TP and intra-tumor heterogeneity are major confounding factors in somatic NGS analysis. Accurate detection of low-frequency events requires optimized panel design, robust error suppression, and bioinformatics tools specifically adapted to handle signal dilution and high-background conditions. These considerations are essential to avoid false negatives (FNs) and to capture clinically actionable subclonal events that may influence treatment resistance or tumor progression[54].

1.3.3. Tumor-only sequencing: lack of matched normal samples

In clinical oncology, most somatic NGS analyses are performed on tumor-only samples, without a matched normal (non-tumor) specimen from the same patient. While this approach simplifies logistics, reduces sequencing costs, and shortens turnaround time, it introduces key limitations for distinguishing true somatic variants from germline variants and technical artifacts[10,12].

A matched normal sample provides a personalized reference that enables accurate subtraction of germline variants and systematic sequencing noise. In its absence, variant interpretation in tumor-only workflows must rely on indirect filtering strategies, such as excluding variants present in large-scale population databases. However, this approach has important caveats[10,12,41]:

- Rare germline variants, especially those specific to underrepresented populations, may be incorrectly classified as somatic.

- Conversely, true somatic mutations that overlap with common polymorphisms or occur in hypermutable regions may be filtered out, reducing sensitivity.

To compensate for the lack of patient-specific germline data, many workflows employ a panel of normals (PoN)—a collection of normal samples sequenced and processed using the same protocols. The PoN is used to flag recurrent sequencing artifacts and systematic noise (e.g., oxidative damage, homopolymer-related errors), but it cannot substitute the matched normal for resolving patient-specific germline variants[9,10,41].

An additional strategy widely adopted in clinical labs is empirical artifact filtering based on routine experience: variants observed at high frequency across unrelated tumor samples—especially if not annotated as driver mutations or known hotspots—are likely artifacts and are filtered accordingly. This approach, while heuristic, provides a valuable internal QC mechanism that complements algorithmic and database-driven filters[12].

Further complicating tumor-only analysis are sample preparation artifacts, such as cytosine deamination (caused by formalin fixation in FFPE samples) or oxidative base damage, which can mimic true mutations. In these cases, read-level metrics—including strand bias, read orientation, or positional base quality—become essential for discriminating true variants from technical noise[10,12,36].

In summary, the absence of matched normal samples remains one of the main limitations of somatic NGS analysis in routine diagnostics. While current bioinformatic strategies provide workarounds to reduce FPs, none fully replicate the reliability of tumor–normal paired analyses. Therefore, the design of pipelines must integrate multi-layered filtering approaches and contextual annotations to mitigate this inherent limitation of tumor-only testing.

1.3.4. Complex genomic regions

Certain genomic regions possess intrinsic sequence features that complicate their analysis by short-read NGS technologies, leading to limitations in read alignment, variant detection, and interpretation. Despite the targeted design of clinical panels, some loci remain difficult to sequence or interpret due to sequence repetitiveness, low mappability, or high sequence homology[9,10,12]. Key problematic regions include[11,55]:

- Repetitive sequences: such as microsatellites, homopolymer runs, and transposable elements, which can cause ambiguous read alignments and complicate InDel detection.
- Segmental duplications: large regions of nearly identical sequence shared across multiple loci, which frequently lead to multi-mapping reads and uncertainty in variant localization.
- GC-rich regions: which impair hybridization efficiency, PCR-Amp, and sequencing fidelity, resulting in coverage dropouts and reduced sensitivity.

These factors increase the rate of clipped or misaligned reads and reduce effective coverage in affected regions. For example, InDels in tandem repeat regions may be misaligned or ambiguously represented[10,42].

Another significant challenge arises from pseudogenes and highly homologous gene families. Reads from genes that have nearby pseudogenes or paralogs with high sequence identity may align equally well to multiple loci, resulting in FP calls or missed true variants due to mapping uncertainty[11].

Although bioinformatic strategies such as GC bias correction and multi-mapping filtering can partially mitigate these issues, they do not fully restore confidence in affected loci. Consequently, regions with consistently poor coverage or ambiguous mapping should be flagged during analysis and interpretation. For high-impact variants in such areas, orthogonal validation methods—such as Sanger sequencing or long-read technologies—are strongly recommended to confirm or rule out candidate alterations[11,41,42].

In summary, the complexity of certain genomic regions imposes persistent limitations on somatic variant analysis with short-read NGS. Awareness of these challenges is essential in both pipeline development and clinical reporting, ensuring that uncertain regions are properly annotated and addressed in the diagnostic workflow.

1.3.5. Detection of complex genomic biomarkers

While targeted NGS panels have proven highly effective for the detection of individual alterations, their limited genomic scope imposes significant constraints on the detection of complex genomic biomarkers (e.g., MSI, TMB, mutational signatures, HRD). These biomarkers require integrative analysis of broad mutational or copy-number patterns, which are not easily captured in small, focused genomic assays.

MSI detection by NGS requires sufficient coverage of a representative set of microsatellite loci. However, most targeted panels contain only a small number of such regions, reducing sensitivity and increasing the likelihood of FNs. Furthermore, PCR slippage and sequencing artifacts—particularly prevalent in FFPE-derived DNA—can mimic the signal of instability, necessitating careful calibration, filtering, and the use of specialized algorithms to distinguish true MSI from technical noise[5,36].

TMB estimation is another challenging application in panel-based assays. Traditionally calculated from WES, TMB in targeted panels is extrapolated from the number of somatic mutations observed within the captured coding territory. This calculation is highly sensitive to the size of the panel, sequencing depth, variant filtering thresholds, and the presence of germline contamination—especially in tumor-only workflows. Moreover, differences in the inclusion of synonymous vs. non-

synonymous variants, and lack of standardized filtering pipelines, lead to inconsistencies across panels and laboratories, hindering clinical harmonization and benchmarking[50].

The detection of mutational signatures is even more constrained. These signatures are based on the trinucleotide context of mutations and typically require a large number of somatic variants—often hundreds—to be robustly inferred. Such mutation counts are rarely observed in targeted panels, and the biased representation of genomic regions further complicates signature extraction. Although signature deconvolution tools have been adapted for high-depth panels, the accuracy and interpretability of the results remain limited in this context[10,40].

HRD assessment typically relies on genome-wide analysis of copy-number patterns, including metrics such as LOH, LST, and TAI. These require dense SNP coverage across the genome—generally achievable only with WGS or high-resolution SNP arrays. While some targeted panels incorporate surrogate scores for HRD, they lack sufficient resolution to capture subtle allelic imbalance events. Alternatively, *BRCA1/2* mutation status and HRD-associated mutational signatures can serve as indirect markers, but they provide only partial insight into the HRD phenotype[51].

In summary, the accurate detection of complex genomic biomarkers in targeted panel assays is hindered by limited genomic representation, reduced mutation counts, and technical variability. While recent efforts have enabled approximation of some biomarkers, significant improvements in panel design, analytical methodology, and standardization are still required to fully support their clinical application in precision oncology[54].

1.3.6. RNA-seq–based somatic analysis

RNA-seq provides complementary insights to DNA-based profiling by enabling the detection of gene fusions, alternative splicing, and transcript expression changes—features that are critical in many cancer types. However, its implementation in somatic panel analysis introduces unique technical and bioinformatic challenges that must be addressed to ensure robust and clinically meaningful results[9,10,54].

A central limitation of RNA-seq is its dependence on gene expression. Transcripts must be expressed at sufficient levels for sequencing reads to adequately cover fusion breakpoints or splice junctions. Low expression, variability across tumor types or subclones, and stochastic transcriptional noise all contribute to uneven coverage, potentially resulting in missed alterations. Unlike DNA, RNA represents a dynamic transcriptional snapshot rather than a stable genomic baseline, complicating the interpretation of AF and clonality[10,54].

From a technical perspective, RNA derived from FFPE tissues—a common source in clinical settings—is frequently degraded and fragmented. This results in shorter reads with lower quality, hindering both transcript assembly and the sensitivity of fusion detection. RNA degradation also increases the likelihood of sequencing artifacts and mispriming during library preparation, especially in older samples[54,56,57].

Splice-aware alignment is another critical requirement of RNA-seq analysis. Short reads spanning exon-exon junctions must be accurately mapped to the transcriptome, particularly for detecting fusions or splicing aberrations. However, non-canonical junctions, complex rearrangements, and regions with high sequence homology (e.g., pseudogenes or paralogs) can lead to misalignments and FPs. The choice of aligner and its configuration are therefore essential for minimizing error propagation[9–11,57].

Fusion transcript interpretation is inherently complex. It requires determining whether the fusion is in-frame, assessing predicted protein products, and evaluating biological relevance. Supporting evidence—such as the number of split and spanning reads, expression of fusion partners, and recurrence in curated fusion databases—must be integrated to distinguish likely driver events from passengers or artifacts[10,57].

Despite its growing adoption, standardization of RNA-seq analysis pipelines in the clinical context remains limited. There is no consensus on best practices for alignment, fusion calling, filtering, or reporting. Furthermore, benchmarking datasets for evaluating RNA-based variant detection—particularly fusions—are still scarce, limiting tool validation and cross-platform reproducibility[42,57].

In summary, RNA-seq enhances the clinical utility of NGS panels by uncovering transcript-level alterations, but its application is challenged by RNA quality, expression variability, complex bioinformatics, and a lack of standardization. Addressing these barriers is essential for the reliable integration of RNA-based biomarkers into clinical oncology workflows[54].

1.3.7. Lack of automated and standardized systems for variant prioritization

Following variant detection and annotation, a critical bottleneck in somatic NGS panel analysis is the prioritization of clinically relevant alterations. This step is fundamental for guiding diagnostic, prognostic, and therapeutic decisions, yet it remains largely manual, time-consuming, and inconsistent across laboratories[4,34,35].

In current practice, variant prioritization often requires expert review across multiple layers of information, including functional impact, cancer gene relevance, known pathogenicity, updated predictive biomarker state of art, and drug sensitivity and resistance associations. Although many public and commercial databases provide curated knowledge, there is no universally accepted system

capable of automatically integrate and interpret this information into structured, clinically actionable outputs[34,35]. Several challenges contribute to this gap, which are also discussed in MTBs:

- Heterogeneity of variant types, including SNVs, InDels, CNAs, fusions, and splicing alterations, each requiring distinct interpretation frameworks.
- Lack of harmonized criteria across knowledge databases and inconsistent use of clinical evidence levels.
- Rapidly evolving clinical guidelines that are difficult to keep up-to-date in static pipelines.
- The frequent presence of variants of uncertain significance (VUS), which lack sufficient evidence for automated classification and require expert review.

Most bioinformatics workflows rely on custom filtering scripts or heuristic rules for prioritization—such as VAF thresholds, impact prediction, or known hotspot filters[10]. However, these rules are often hard-coded, panel-specific, and difficult to generalize or maintain. A few tools support semi-automated classification following standardized tiering systems (e.g., AMP/ASCO/CAP), but these are often not fully integrated into end-to-end workflows and rarely account for multi-variant or multi-omics context [9,34].

International efforts are actively addressing the lack of standardized variant prioritization frameworks. For example, the Cancer Genome Interpreter (CGI)-Clinics project (Horizon Europe) is transforming the CGI framework[58] into a clinical-grade, community-driven decision-support platform that enables automated variant tiering and integration of evolving clinical evidence for oncology workflows. Similarly, the VICC, under the Global Alliance for Genomics and Health (GA4GH) umbrella, harmonizes clinical interpretations across major knowledge bases through resources like meta-knowledgebase (MetaKB), promoting consensus-driven, scalable variant interpretation[35]. These initiatives exemplify the push toward reproducible and interoperable solutions for clinical genomics.

The absence of robust automated prioritization systems introduces subjectivity, inter-operator variability, and reporting delays, especially when dealing with complex or ambiguous findings. To address this limitation, future clinical pipelines should aim to incorporate[34,35]:

- Rule-based or machine-learning prioritization modules aligned with international guidelines.
- Dynamic evidence integration from updated knowledge bases and drug approvals.
- Multi-variant interpretation strategies capable of joint prioritization (e.g., co-occurring mutations, fusions, and CNAs).

Ultimately, automated and standardized prioritization frameworks would enhance reproducibility, scalability, and clinical confidence—key goals for the routine implementation of precision oncology.

1.3.8. Deficient visualization and reporting tools for clinical interpretation

A critical step in the clinical translation of NGS data is the effective communication of results to end users—primarily clinicians, molecular pathologists, and members of the MTBs—many of whom lack bioinformatics expertise. However, most existing pipelines, particularly academic or research-oriented ones, output raw data formats (e.g., BAM, VCF) that are difficult to interpret without specialized training. These formats typically lack accessible summaries, interactive dashboards, or clinically relevant contextualization, creating a communication gap between data producers and decision-makers[34,40,42].

Commercial solutions often aim to bridge this gap through simplified graphical interfaces and summary dashboards. However, they frequently suffer from limited interactivity, rigid designs, and superficial outputs, focusing on mutation tables without integrating important complementary data such as VAFs and read support, copy-number profiles, fusion diagrams or splicing illustrations, and relevant clinical guidelines or therapy associations[4,59].

Manual review remains essential for variant QC, particularly for ambiguous or borderline calls. Yet tools like IGV, which allow read-level inspection in BAM files, require bioinformatics expertise and are not scalable or user-friendly in high-throughput diagnostic workflows[42]. Similarly, cross-sample summaries, coverage statistics, or QC visualizations are rarely included in standard outputs, even though they are vital for interpreting negative results or validating complex findings.

To overcome these limitations, clinical-grade reporting systems must evolve to incorporate:

- Integrated, tiered variant summaries aligned with interpretation frameworks (e.g., AMP/ASCO/CAP, ESCAT).
- Interactive and customizable web-based visualizations, including mutation lollipop plots, CNV heatmaps, fusion gene maps, and variant filtering interfaces.
- Sample-level overviews, including key QC metrics and sequencing coverage benchmarks.
- Links to supporting clinical evidence, databases, and therapeutic annotations.

Ideally, these reports should be automatically generated at the end of the bioinformatics pipeline and exportable in clinician-friendly formats (e.g., PDF, HTML) to support streamlined decision-making in MTBs. Enhanced reporting not only improves interpretability and traceability but also increases the reproducibility and clinical utility of somatic NGS analyses[34,36,40].

1.3.9. Variability and lack of standardization across somatic NGS workflows

Despite the increasing adoption of NGS panel assays in clinical oncology, significant technical and computational variability persists across laboratories and platforms, undermining the reproducibility

and comparability of somatic variant analysis. This variability stems from differences in sequencing technologies, wet-lab protocols, bioinformatic pipelines, and result interpretation frameworks, each of which can introduce platform-specific biases or inconsistencies in downstream analysis[9,11,42,54].

At the technical level, sequencing platforms such as Illumina and Ion Torrent use distinct chemistries and detection methods, leading to different error profiles. Ion Torrent, for example, is known to struggle with homopolymers due to its pH-based detection, while Illumina typically offers higher base quality but may be susceptible to issues like index hopping. Read length, fragment size distribution, and depth of coverage further influence the detection of InDels, CNAs, and SVs[9,39].

From a computational standpoint, each variant caller applies different algorithms and thresholds, often optimized for specific variant types or sequencing depths. As a result, using different tools or pipelines—even on the same dataset—can yield divergent variant calls, particularly in challenging contexts such as low TP or noisy regions. Differences in alignment strategies, duplicate removal, and quality recalibration compound this variability. Additionally, tool-specific VCF formatting and variant representations (e.g., InDel coordinates and alleles) complicate harmonization of results[4,60].

To improve accuracy, ensemble calling strategies have been adopted in many cancer genomic projects. These strategies combine multiple callers (e.g., through majority voting or intersection rules) to reduce FPs and improve specificity[10,40,41]. However, they require careful post-processing to normalize variant representations, including left-alignment (shifting variants to the most leftward equivalent position) and parsimony (representing variants using the shortest allele strings possible), to ensure consistent interpretation[60]. This additional complexity presents a burden for clinical laboratories lacking specialized bioinformatics resources.

To address these challenges, there is growing momentum toward adopting standardized and reproducible workflows grounded in the Findable, Accessible, Interoperable, and Reusable (FAIR) principles. Initiatives like nf-core promote community-curated pipelines built on modular Nextflow scripts with version control, testing, and full containerization[53]. Similarly, platforms such as Dockstore and GA4GH Workflow Execution Services facilitate the deployment of standardized pipelines across institutions and cloud environments. These tools improve transparency, portability, and auditability—essential for clinical accreditation and external quality assurance[53,61].

Despite the emergence of community efforts and benchmarking initiatives—such as those led by GA4GH, Sequencing Quality Control Phase II (SEQC2), and International Organization for Standardization (ISO) working groups—that aim to define standardized file formats, performance metrics, and analysis protocols, widespread implementation of these standards in clinical laboratories remains limited. Many centers continue to rely on internally developed workflows that lack full

documentation, validation, or traceability, further complicating cross-study comparability and quality assurance[54,62].

To ensure the reliability and clinical utility of somatic NGS panel testing, it is essential to minimize both technical and computational variability through coordinated efforts that promote protocol standardization, validated pipelines, and QC frameworks. Ultimately, the adoption of transparent, reproducible, and community-vetted bioinformatics workflows—aligned with FAIR principles—will be pivotal for achieving consistent, interpretable, and clinically actionable results across laboratories and platforms[34,53,54,62].

1.3.10. Limited data sharing and interoperability in clinical genomics

Despite the increasing implementation of NGS in oncology, the reuse, integration, and exchange of genomic data across laboratories and institutions remain restricted. Multiple factors contribute to this limitation, undermining efforts to build collective knowledge and improve the reproducibility and scalability of clinical genomics.

A major barrier is the presence of strict privacy regulations and institutional policies, which restrict the sharing of genomic data—particularly when linked to sensitive clinical or personal information. While international efforts such as the GA4GH, the European Genome-phenome Archive (EGA), and the Beacon Project aim to promote secure, federated access to genomic datasets, their adoption in routine diagnostics remains limited[63–65]. The Beacon Project, for example, allows institutions to share the existence of specific genomic variants without disclosing identifiable data, providing a privacy-aware model for data discoverability[65].

A second issue is the lack of interoperability and standardization across data formats and analysis pipelines. Although core file types like FASTQ, BAM, and VCF are widely adopted, the structure, metadata fields, and variant representations often differ between tools and institutions. These discrepancies—ranging from inconsistent filtering tags to diverging nomenclature or alignment conventions—complicate cross-tool comparison, meta-analysis, and benchmarking[4,42,62].

Furthermore, proprietary pipelines and closed-source platforms commonly used in commercial diagnostics exacerbate this fragmentation by generating outputs in non-standard or locked formats. These barriers hinder downstream integration, prevent external QC, and limit the ability to contribute to shared datasets or collaborative research efforts[40].

Addressing these challenges will require a coordinated global effort to promote the use of open formats, interoperable standards, and federated infrastructures for secure data access and sharing. These initiatives provide valuable frameworks for fostering transparency, reproducibility, and responsible data stewardship in clinical genomics[53,54,62,64].

1.3.11. Limitations of existing solutions

In clinical diagnostics, commercial NGS panels are frequently accompanied by proprietary, "ready-to-use" software solutions tailored to their specific assays[20,37,66,67]. These platforms are designed for ease of use, requiring minimal bioinformatics expertise and allowing laboratories with limited computational resources to perform basic analyses. However, closed-source pipelines provide very limited flexibility. Users are unable to readily customize workflows, update individual tools, or incorporate emerging methodologies. Moreover, their output is often restricted to basic variant tables with minimal contextual information, lacking interactive visualizations, clinical annotations, or quality metrics that are essential for comprehensive interpretation. Consequently, additional post-processing steps (e.g., variant annotation, filtering, and manual review) are typically required before results can be effectively used in clinical decision-making settings such as MTBs[40].

Conversely, open-source academic pipelines provide transparency and flexibility but often address only isolated components of the workflow (e.g., small variant calling, fusion detection, or annotation). Although several integrated and portable tools have been developed for somatic variant analysis[68–72], they frequently fall short in clinical settings, especially when applied to targeted cancer panels. Key limitations include:

- Incomplete analytical scope: Many pipelines are designed for either DNA or RNA analysis, but not both, restricting their application to dual DNA-RNA clinical panels, which are increasingly employed for comprehensive tumor profiling.
- Dependency on matched normal samples: A significant number of workflows assume the availability of paired tumor-normal samples for somatic filtering, yet such specimens are rarely collected in routine clinical practice.
- Limited adaptability to panel-specific protocols: Variability in panel design, library preparation, and sequencing technology demands flexible pipelines, but most tools cannot readily accommodate to these variables.
- Insufficient clinical reporting: Academic tools often do not include modules for variant prioritization, clinical classification, or interactive reporting, which are essential for real-world diagnostics.
- Portability and reproducibility challenges: Many pipelines suffer from complex installation, dependency conflicts, or poor documentation, which hinder their deployment across different institutions and computing environments.

Together, these shortcomings highlight a critical gap in the current landscape: the lack of robust, adaptable, and clinically oriented bioinformatics workflows capable of supporting the full analysis and reporting of somatic NGS panels in precision oncology[10,40,41].

1.3.12. Concluding remarks

The bioinformatics analysis of somatic NGS panel data in oncology entails multiple technical, biological, and computational challenges. These include the absence of matched normal samples, intratumor heterogeneity, suboptimal DNA/RNA quality, and the complexity of detecting structural alterations and composite biomarkers. In addition, variability across platforms, non-standardized workflows, and limited interoperability still compromise the reproducibility and comparability of results across laboratories and institutions.

Although existing commercial and academic solutions address specific analytical needs, none provides a fully integrated, transparent, and adaptable framework that meets the broad requirements of somatic cancer panel analysis. Key limitations persist in areas such as variant prioritization, support for DNA-RNA panels, automated reporting, and workflow portability.

These gaps underscore the need for a comprehensive, modular, and open-source bioinformatics pipeline capable of supporting the end-to-end analysis of somatic NGS panel data, from raw sequencing files to interpretable, report-ready results. Addressing this unmet need motivates the work presented in this thesis.

2. HYPOTHESIS AND OBJECTIVES

2.1. Rationale

The adoption of NGS has revolutionized cancer genomics by enabling the simultaneous detection of multiple somatic alterations, such as mutations, CNAs, and gene fusions, using targeted panels with high resolution and efficiency. These NGS cancer panels are now routinely applied in both clinical and research settings to support diagnosis, prognosis, therapeutic selection, and biomarker discovery. However, the full value of these assays depends not only on sequencing technologies, but also on the availability of robust bioinformatics workflows capable of handling the complexity and diversity of tumor-derived data.

Current bioinformatics solutions present key limitations in flexibility, transparency, scalability, and analytical completeness. Commercial platforms often operate as closed systems with limited adaptability and superficial interpretability, while academic tools frequently lack full integration, support for dual DNA-RNA inputs, or visual reporting. Additional challenges, such as the absence of matched normal controls, low-quality input material, and a lack of standardized workflows, further complicate the reproducibility and utility of somatic NGS analyses.

The hypothesis of this thesis is that, through a research-driven process to define the most suitable analytical strategies for somatic NGS cancer panels, the development of a tailored, in-house bioinformatics pipeline can enhance the accuracy, reproducibility, and applicability of genomic analysis in both translational research and clinical oncology.

2.2. General objective

The main aim of this thesis is to design, implement, and evaluate an open-source, comprehensive bioinformatics pipeline for the analysis of somatic NGS cancer panels. The pipeline is intended to address the analytical complexity of tumor-derived data by enabling accurate variant detection, automated annotation and prioritization, and the generation of visual reports. Through this, it seeks to support both research and clinical applications by facilitating the interpretation of NGS panel results in diverse precision oncology contexts.

2.3. Specific objectives

1. To design and implement an open-source, comprehensive bioinformatics pipeline for the analysis of somatic NGS cancer panels.
    a) To develop a robust bioinformatics workflow capable of addressing the diverse scenarios encountered in somatic panel analysis, using state-of-the-art open-source tools, to ensure analytical reliability, reproducibility, and portability across diverse environments.
    b) To integrate an automated reporting system that generates interactive and user-friendly visual outputs to enhance the accessibility, interpretation, and communication of results.
2. To evaluate the performance and applicability of the implemented pipeline.
    a) To validate the accuracy of variant detection and assess the pipeline's panel-agnostic design using standardized public reference datasets.
    b) To benchmark the pipeline with retrospective real-world data from multiple tumor types and commercial panels, evaluating its analytical robustness and adaptability to routine diagnostic and research contexts.

3. METHODOLOGY

3.1. Implementation of the ClinBioNGS pipeline

3.1.1. General architecture

The developed bioinformatics pipeline, ClinBioNGS, is a modular and fully automated clinical bioinformatics pipeline designed for the analysis of somatic DNA and RNA sequencing data derived from targeted NGS cancer panels. The pipeline is implemented in Nextflow[73] (v24.10.1) and all required tools are encapsulated in Apptainer[74] (formerly Singularity; v1.4.1) containers to ensure portability, reproducibility, and ease of deployment. The selection of software and resources was guided by criteria prioritizing open-source availability, broad accessibility, active maintenance, and widespread adoption within the bioinformatics community.

The pipeline architecture follows a modular design in which each analytical step is implemented as an independent Nextflow process. This structure allows for clear separation of functional stages (e.g., pipeline set up, pre-processing, alignment, variant calling, annotation, reporting), facilitates debugging and maintenance, and supports customization and extension. Processes are connected through channels that coordinate input and output dependencies, while computational resources are dynamically assigned according to the specific requirements of each task. Multiple processes corresponding to the same analytical stage are grouped into subworkflows, adding an additional layer of modularization and enabling higher-level functional organization. Configuration is driven by user-defined parameters and profile-based (already defined) settings that adapt the pipeline to the sequencing platform, cancer panel, and computational environment.

ClinBioNGS supports multiple input data formats, including raw sequencing files in FASTQ, BCL, or uBAM. These files are automatically pre-processed prior to downstream analysis. The pipeline integrates QC at various stages and handles both DNA and RNA data processing for the detection of related alterations and genomic biomarkers. Analysis results are collected into structured outputs, including interactive reports, variant registries, and detailed logs for traceability and clinical review.

Overall, ClinBioNGS is designed to address the practical requirements of clinical genomics workflows, providing automation, transparency, and compatibility with real-world diagnostic and research environments. The pipeline has been validated on eight commercial panels and currently supports full analytical workflows for the Illumina TruSight™ Oncology 500 (TSO500), Thermo Fisher Oncomine™ Precision Assay (OPA), and Thermo Fisher Oncomine™ Comprehensive Assay (OCA). ClinBioNGS is freely available for non-commercial research use only (RUO) at: https://github.com/raulmarinm/ClinBioNGS.

3.1.2. Pipeline's resources preparation

ClinBioNGS includes a pre-analysis module that automatically downloads and prepares all the resources and containerized tools required for each functional stage of the pipeline. This ensures reproducibility, minimizes manual intervention, and standardizes the analysis across environments. **Supplementary Table 1** lists all software and **Supplementary Table 2** lists all resources used in the pipeline, including version and role in the pipeline[14,24,26,47,48,55,58,60,75–127].

3.1.2.1. Apptainer images

ClinBioNGS relies on containerized tools executed via Apptainer. Images are either downloaded directly or built from publicly available Docker repositories (e.g., BioContainers[128], Galaxy Project[129], Docker Hub). Due to compatibility issues in certain environments, custom Docker images were created for the Pisces[92] and Octopus[91] small variant callers. Additionally, a dedicated R[93] environment with all required packages was encapsulated in a single image. All three images are publicly available on Docker Hub.

3.1.2.2. User-defined metadata files

ClinBioNGS allows user customization through metadata files:

- *SampleInfo.csv*: A comma-separated values (CSV) file that provides sample-level metadata such as sex, age, tumor type, and estimated TP.
- *WhitelistGenes.csv*: A CSV file that defines tumor-specific or general whitelist genes for prioritization.
- *TumorNames.csv*: A CSV file that maps user-defined tumor names to Disease Ontology Identifiers (DOIDs)[121], top-level ontology nodes, and OncoTree[122] tumor codes, ensuring compatibility with clinical evidence annotations from CIViC[14].

It is recommended that the DOID values in *SampleInfo.csv* match those in *TumorNames.csv* to ensure accurate downstream mapping.

3.1.2.3. Reference genomes and genome resources

Reference genome files for GRCh38 and GRC Mouse build 38 (GRCm38) are automatically downloaded from the Illumina iGenomes Amazon Web Services (AWS) repository[77]. Index files are generated using Burrows-Wheeler Aligner - Maximum Exact Matches (BWA-MEM2)[81] , Torrent Mapping Alignment Program (TMAP)[96] (Ion Torrent platform), and Xengsort[102] (mouse).

Cytoband information is retrieved from the University of California Santa Cruz (UCSC) Genome Browser Database[98] for gene annotation and copy-number analyses. Chain files to convert hg19 and hg38 coordinates are also downloaded from UCSC and used for liftover operations.

3.1.2.4. MANE annotation files

To ensure transcript consistency, ClinBioNGS uses files from the Matched Annotation from NCBI and EMBL-EBI (MANE) collaboration v1.4[104]. These include Gene Transfer Format (GTF)-based exon, intron, and coding annotations, which are used in various downstream annotation steps.

3.1.2.5. Target region files

Panel manifest files are standardized into 4-column BED format with hg38 coordinates and gene names. When necessary, several processing steps are applied:

- Convert manifests from vendor-specific formats (e.g., Illumina, Ion Torrent).
- Perform coordinate liftover (if hg19).
- Annotate MANE genes.
- Remove non-primary chromosomes.
- Normalize gene symbols.
- Merge overlapping regions using Bedtools[79].

Final BED files are converted to interval lists using Genome Analysis Toolkit version 4 (GATK4)[86], and additional versions are generated for padded regions or clipping, as required.

A gene annotation table is also generated per panel, providing updated gene symbols from the Human Genome Organization Gene Nomenclature Committee (HGNC)[130], cytobands, RefSeq and Ensembl IDs, and full gene names from the MANE resource.

3.1.2.6. VCF headers

Predefined VCF header templates are provided for each alteration type to ensure standardized output formats. These templates are automatically appended when generating the results.

3.1.2.7. Gene role and oncogenicity resources

- Network of Cancer Genes (NCG)[106]: Used to annotate oncogenes and TSGs.
- Catalog of Validated Oncogenic Mutations[58]: Curated list of functionally validated variants from CGI resource.
- CIViC oncogenic evidence: Oncogenic variants from the CIViC database.
- ClinGen/CGC/VICC SOP[48] dataset: Set of previously classified oncogenic variants.

These datasets are harmonized (e.g., HGNC symbol updates, hg19 to hg38 liftover) and used for variant classification.

3.1.2.8. Variant annotation resources (VEP)

ClinBioNGS uses Ensembl Variant Effect Predictor (VEP)[83] for small variant annotation, supported by:

- VEP cache and GRCh38 FASTA reference.
- VEP plugins and data for Combined Annotation Dependent Depletion (CADD)[123], Rare Exome Variant Ensemble Learner (REVEL)[107], and AlphaMissense[108] pathogenicity predictors.
- Publicly available ClinVar and CIViC VCFs, integrated as custom annotation.

3.1.2.9. Cancer hotspot resources

Small variants are annotated against cancer hotspot evidence:

- Panel-specific hotspot BED files, either provided by the user or auto-generated from Ion Torrent output.
- AACR GENIE BED file of known somatic events (lifted over to hg38).
- Cancer Hotspots[110] list of amino acid (AA) changes and counts for each mutation.

3.1.2.10. Problematic and high-confidence regions

To support the interpretation of small variant results, ClinBioNGS incorporates the annotation of both low-confidence and high-confidence genomic regions:

- Low-confidence regions can be flagged using two types of BED files:
    - Panel-specific blacklists provided by the user, which identify regions known to be technically challenging or prone to artifacts in specific panels.
    - A comprehensive BED file generated by ClinBioNGS that merges multiple publicly available genomic stratification datasets, including:
        - UCSC resources: Encyclopedia of DNA Elements (ENCODE) blacklist, GRC exclusions, and regions with unusual mapping characteristics.
        - Genome in a Bottle (GIAB) stratification files[55]: homopolymers, tandem repeats, segmental duplications, low mappability regions, and highly polymorphic loci.
- In contrast, high-confidence regions are annotated using the Consensus Targeted Regions (CTRs) defined by the SEQC2 Consortium. These regions represent genomic intervals with robust sequencing reliability, making them suitable for confident small variant detection[111].

3.1.2.11. GENIE cancer registry

Raw data files from the AACR GENIE project were previously downloaded and processed for downstream annotation in ClinBioNGS. The resulting processed files, organized by alteration type, are stored within the pipeline's annotation directory.

For raw somatic mutations (small variants), the following steps were applied:

- Standardize gene symbols according to the HGNC nomenclature.
- Count the number of samples harboring each unique mutation.
- Calculate the maximum population VAF (pVAF) in non-bottlenecked populations from the gnomAD database.
- Annotate the gene role with NCG resource and identify hotspot mutations from Cancer Hotspots database.
- Lift over hg19 to hg38 coordinates.
- Classify by oncogenicity according to the ClinGen/CGC/VICC SOP guidelines.

For raw CNAs, processing steps included:

- Update gene symbols (HGNC).
- Count the number of samples profiled per gene.
- Select genes with AMP or homozygous DEL status.
- Calculate the sample count and frequency for each CNA status per gene.

For raw SVs, including gene fusions, the following steps were applied:

- Update gene symbols (HGNC) for both fusion partners.
- Keep only in-frame fusions with potential functional relevance.
- Standardize coordinate nomenclature ("chromosome:position") and lift over to hg38 for both fusion breakpoints.
- Count the number of samples in which each fusion event (e.g., "geneA::geneB") was detected.

The final output includes a curated list of oncogenic mutations and summary tables detailing the sample counts and frequencies for mutations, CNAs, and fusions, which are used throughout the ClinBioNGS annotation workflow.

3.1.2.12. Clinical evidence files (CIViC)

A curated list of predictive, prognostic, and diagnostic clinical evidence from the CIViC database is used to classify detected alterations according to the AMP/ASCO/CAP guidelines[4] for clinical significance. This resource is compiled from the raw variant, molecular profile, and clinical evidence

files provided by CIViC, in combination with a standardized list of tumor names (with DOIDs) used in ClinBioNGS.

Alterations including mutations, CNAs, and gene fusions are extracted and processed from the variant and molecular profile files. Clinical evidence entries are filtered to include only those with:

- Evidence types classified as predictive, prognostic, or diagnostic.
- Evidence levels rated “A” (validated association) to “D” (pre-clinical evidence).
- Evidence star ratings between 3 (convincing) and 5 (strong, well supported).
- Tumor-specific relevance (excluding germline associations).

Tumor names from CIViC are harmonized with the internal ClinBioNGS tumor list to ensure consistency with user-provided metadata. Finally, each alteration is annotated with its associated clinical evidence, distinguishing between tumor-specific and, for predictive evidence, drug-specific associations.

3.1.2.13. RNA resources

Several external and in-house resources are integrated into ClinBioNGS to support various aspects of RNA analysis:

- Trinity Cancer Transcriptome Analysis Toolkit (CTAT) library[82]: Provides RNA genome and annotation files necessary for fusion and splice variant detection.
- AACR GENIE registry: Supplies fusion event frequencies in cancer, based on the previously processed datasets used for benchmarking and annotation.
- Mitelman Database[112]: A comprehensive list of reported gene fusions was obtained via a full export of all entries listed under the "Gene Fusions" section of the database’s website.
- In-house whitelists: Curated collections of known gene fusions and splice variants, including variant names and genomic coordinates, were compiled from multiple peer-reviewed publications[113–119]. These support enhanced annotation and prioritization of clinically relevant RNA alterations.

3.1.2.14. Panel-recurrent small variants (TSO500, OPA, OCA)

ClinBioNGS supports the flagging of panel-recurrent small variants to help identify potential technical artifacts or common population-specific variants[41]. To enable this feature, users must provide a list of recurrent variants tailored to the panel being analyzed. Precompiled lists are available for the Illumina TSO500 and Thermo Fisher OPA and OCA commercial panels based on aggregated variant data from benchmarking cohorts (N_{TSO500} = 655, N_{OPA} = 621, and N_{OCA} = 537). Variants were considered recurrent if detected in at least 15% of the samples for each panel[12]. These files are available in the ClinBioNGS GitHub repository.

3.1.2.15. Panel-specific CNA baselines (TSO500, OPA, OCA)

CNA analysis in ClinBioNGS requires a panel-specific pooled reference cohort to serve as the baseline for assessing tumor CNAs. Dedicated CNA baselines have been generated for the TSO500, OPA, and OCA panels using large tumor cohorts (further details on their construction are provided in the *3.1.7. Analysis of CNAs* section). Each baseline file contains a list of genomic regions annotated with average coverage values and variability scores (i.e., spread) across the reference dataset. All baseline files are available in the ClinBioNGS GitHub repository.

3.1.2.16. Panel-specific MSI baseline (TSO500)

To assess MSI in tumor samples, a platform-specific pooled reference cohort is required. For the TSO500 panel, a baseline was generated using a cohort of microsatellite-stable (MSS) tumor samples (see *3.1.10. Analysis of genomic biomarkers* section for further details on its construction). The TSO500-specific MSI baseline is available in the GitHub repository.

3.1.3. Input data and pre-processing

ClinBioNGS supports various input data formats and includes multiple pre-processing steps to ensure compatibility across sequencing platforms, panel designs, and library preparation protocols.

3.1.3.1. FASTQ generation from raw sequencing data

Starting data for DNA and RNA libraries can be provided directly as FASTQ files or internally generated from BCL or uBAM files. All input files must be placed in the *--startingDataDir* path.

FASTQ files

By default, the pipeline expects DNA and RNA FASTQ files to be provided by the user. File naming must follow the format *<sample>_<DNA/RNA>*.fastq**, where "<sample>" matches the sample identifier specified in the sample sheet.

BCL files

FASTQ files can be generated from raw BCL files using Illumina BCL Convert™, which also supports adapter trimming, UMI processing, and sample demultiplexing based on parameters specified in the sample sheet. Sample- and lane-level QC metrics are generated, which are parsed and visualized in a report by MultiQC[90]. The following QC metrics are provided per sample and lane:

- Total number of clusters (read pairs) and bases (yield).
- Percentage of bases with a Phred quality score of 30 or higher.
- Percentage of reads with perfect sample index (0 mismatches) or one mismatch.
- Mean quality score of bases.

The number of undetermined reads—those not assigned to any sample—is also reported. To allow for the use of different processing parameters, BCL Convert is executed separately for DNA and RNA libraries. As a result, some reads may appear as undetermined simply because they belong to the other library type (DNA or RNA) processed in a separate run, rather than being truly unassigned. MultiQC recognizes that both DNA and RNA outputs originate from the same sequencing run and recalculates the undetermined read statistics, providing a more accurate representation of unassigned reads. These QC metrics help identify underperforming samples or lanes and allow pre-alignment quality checks before heavy compute steps.

BAM files

FASTQ files can also be derived from uBAM files using Samtools[78]. Supported input includes user-provided uBAMs (*<sample>_<DNA/RNA>*.bam*) or platform-specific files (e.g., Ion Torrent's directory with **rawlib.basecaller.bam*). If UMIs are encoded in a BAM tag, they can be extracted and appended to the FASTQ header using Samtools and awk. Tag and header information are preserved to support downstream tools that depend on them (e.g., alignment and variant calling for Ion Torrent platforms).

3.1.3.2. FASTQ pre-processing

Pre-processing steps are panel-aware and tailored based on factors such as capture protocol (hybridization vs. amplicon), library type (paired-end vs. single-end, UMI usage), sequencing platform (Illumina-like vs. Ion Torrent), and sample origin (e.g., tumor tissue, cell lines, PDX).

Lane merging

If input FASTQ files are split by sequencing lane, they are automatically merged. Paired-end sequences are combined into R1 and R2 files.

UMI transfer

In some panels (e.g., Agilent), UMI sequences are stored in a separate FASTQ file. The pipeline uses UMI-transfer[100] to append the UMI to the FASTQ header to enable downstream deduplication.

Host contamination filtering (PDX samples)

For sequencing data derived from PDX models, host-mouse contamination is filtered using Xengsort[102]. Only human-specific reads are retained for analysis.

UMI separator normalization

Certain sequencing platforms encode dual UMIs with custom symbols in the FASTQ header. If these separators are incompatible with downstream tools (e.g., Gencore), they are replaced using Bioawk[80]. For example, the "+" symbol in BCL Convert output FASTQs is converted to "_".

Trimming, filtering, and UMI processing

Comprehensive FASTQ pre-processing is performed using FastP[84], which supports adapter trimming, removal of low-quality or short reads, UMI extraction from reads or indices, and quality profiling before and after the filtering. This step generates the final processed FASTQ files used in QC and alignment. It is executed by default but can be skipped if clean FASTQs are already provided.

3.1.4. Alignment and deduplication

3.1.4.1. DNA workflow

Pre-processed DNA reads are initially aligned to the GRCh38 reference genome. By default, alignment is performed using BWA-MEM2[81] for Illumina and other non-Ion Torrent platforms. For Ion Torrent panels, reads are aligned using TMAP[96], following conversion of the processed FASTQ files to BAM format with Samtools. In both cases, the resulting BAM files are also sorted using Samtools. This initial alignment step serves to establish mapping positions necessary for subsequent deduplication.

Deduplication is then performed to eliminate PCR duplicates and reduce sequencing artifacts. In the default approach, GATK4 MarkDuplicates is used to identify duplicate reads based on identical start and end coordinates. When UMIs are present, UMI-aware deduplication is applied to more accurately distinguish true biological molecules from sequencing errors. In ClinBioNGS, Gencore is used for paired-end libraries because it supports UMI-aware deduplication with consensus read generation, enhancing specificity for somatic variant calling[87]. However, Gencore does not support single-end read processing. For single-end libraries, UMI-tools is employed due to its ability to deduplicate single-end reads effectively. While UMI-tools does not generate consensus reads, it selects the most representative read within each UMI group[99]. This specific use of each tool ensures optimal deduplication performance across different library types.

After deduplication, the resulting unique reads are realigned to improve mapping precision. For Ion Torrent data, realignment is performed using TMAP, without requiring BAM-to-FASTQ conversion, since the input remains in BAM format. For non-Ion Torrent data, the deduplicated BAM is first converted back to FASTQ format, then realigned to the reference genome using BWA-MEM2. To further enhance alignment accuracy—particularly around InDels and complex regions—the BWA-

MEM2-aligned BAM files (non-Ion Torrent) undergo local reassembly around target regions with Abra2[75]. The resulting realigned BAM file serves as the final output for all downstream analyses.

By default, the deduplication step is enabled, but it can be optionally skipped for specific panel types, such as amplicon-based assays without UMI support (e.g., OCA). In such cases, both the initial alignment and deduplication steps are omitted.

3.1.4.2. RNA workflow

Pre-processed RNA reads are aligned to the GRCh38 reference genome using Spliced Transcripts Alignment to a Reference (STAR)[95], a splice-aware aligner optimized for transcriptomic data. This step produces a sorted BAM file containing the initial mapped reads. STAR is executed with parameters recommended by the CTAT framework, which are specifically adapted to the accurate fusion detection and compatibility with STAR-Fusion[82]. Key settings include the two-pass mapping mode, where splice junctions identified during the first pass are used to refine the second pass, improving sensitivity and reducing spurious junctions[95]. Additionally, parameters such as *--chimSegmentMin,* which enables chimeric read detection, and *--chimOutType/--chimOutJunctionFormat*, which control the format and content of chimeric output, are applied to ensure the output contains all necessary metadata for downstream fusion analysis[95]. Further options are defined in the dedicated configuration file provided with the pipeline.

Following alignment, deduplication is applied using the same strategy as for DNA libraries, depending on whether the data are single-end or paired-end and whether UMIs are present. This step is used to eliminate PCR duplicates and reduce false-positive signals in downstream analyses.

The deduplicated RNA reads are subsequently realigned using STAR, generating the final set of alignment outputs required for downstream analyses. These include:

- A BAM file containing uniquely aligned, deduplicated reads.
- A chimeric junction file (*Chimeric.out.junction*) that reports chimeric alignments, where individual reads map to two distinct genomic loci. These intergenic junctions, defined by the first intronic base of the donor and acceptor sites[95], represent fusion-like events and are used as input for downstream fusion gene detection.
- A splice junction file (*SJ.out.tab*) containing high-confidence collapsed splice junctions, defined by the start and end positions of intronic regions within genes[95]. Each junction is accompanied by read support metrics and is used in the detection of splice variants.

If deduplication is disabled—such as in amplicon-based libraries without UMIs or in low-complexity samples—the same output files are generated based on all reads, without removing duplicates.

3.1.5. Quality metrics

3.1.5.1. FASTQ QC

Sequencing QC is performed on the processed FASTQ files using FastQC[85], which provides a comprehensive assessment of read quality and potential technical artifacts. FastQC outputs are aggregated and visualized using MultiQC, which provides a summary report of QC metrics.

The FastQC module in MultiQC compiles per-sample statistics and visualizations. General metrics include the total number of reads, GC content, read length, and sequence duplication levels. The main plots provided are sequence counts distribution, per base sequence quality, per sequence quality scores, per base sequence content, per sequence GC content, per base N content, sequence length distribution, sequence duplication levels, overrepresented sequences, and adapter content. These visualizations allow for the early detection of technical issues such as adapter contamination, low-complexity reads, or sequencing bias.

3.1.5.2. BAM QC

Alignment quality metrics are collected from both initial and final alignment steps for DNA and RNA libraries. For the initial alignment (prior to deduplication), GATK4 modules (CollectAlignmentSummaryMetrics and CollectInsertSizeMetrics) and samtools are used to calculate global and target-specific alignment statistics. These metrics reflect the full read set, including duplicates, and provide a baseline assessment of sequencing and alignment quality. After realignment, the same quality metrics are recalculated using the deduplicated BAM files, thereby reflecting only unique, high-confidence reads. In addition, coverage statistics over the target regions are computed using Mosdepth[88], based on the realigned BAM files and the corresponding panel-specific BED files. These coverage metrics support downstream evaluation of panel performance and region-level completeness.

BAM-related quality metrics for all DNA and RNA samples are aggregated and presented in separate visual reports for each data type using MultiQC. These reports integrate output from Picard's CollectAlignmentSummaryMetrics and CollectInsertSizeMetrics GATK4 modules, including mapped reads, read length, and insert size distribution plots. Additionally, Mosdepth provides per-sample metrics including coverage statistics (e.g., mean, median, minimum, maximum), fraction of target genome with at least X coverage, and library size (in bp), along with cumulative coverage distribution, coverage per chromosome plot, and XY coverage histogram[90]. These combined summaries provide a comprehensive view of alignment quality and target enrichment performance across all processed samples.

3.1.5.3. QC results

All DNA and RNA quality metrics collected during alignment and coverage analysis are compiled and summarized on a per-sample basis using a custom R script. This step produces a set of standardized QC tables and visual plots for evaluating sequencing and mapping performance for each sample.

Global QC metrics

A comprehensive table of global QC metrics is generated for each sample by integrating data from alignment summaries (GATK4, Samtools) and coverage outputs (Mosdepth). These metrics enable assessment of sequencing depth, alignment quality, target enrichment, and library complexity, which are essential for determining whether a sample meets quality thresholds for inclusion in downstream analyses. A summary of reported global QC metrics is shown in **Table 6**.

Table 6. Global QC metrics calculated for each DNA and RNA sample.
Each metric is accompanied by a description and its corresponding quality category.

Metric	Description	Category
TOTAL_READS	Total number of sequencing reads	Sequencing depth
ALIGNED_READS	Number of reads aligned to the reference genome	Alignment quality
PCT_ALIGNED	Percentage of aligned reads	Alignment quality
ONTARGET_READS	Number of reads mapped within target regions	Target enrichment
PCT_ONTARGET	Percentage of on-target reads	Target enrichment
HQ_ALIGNED_READS	Number of aligned reads with a mapping quality of ≥Q20 (≤1% error probability)	Alignment quality
PCT_HQ_ALIGNED	Percentage of high quality aligned reads	Alignment quality
MEDIAN_READ_LENGTH	Median length of sequencing reads	Read characteristics
MEDIAN_INSERT_SIZE	Median distance between paired-end reads	Library preparation
UNIQUE_READS	Number of deduplicated reads (non-redundant)	Library complexity
PCT_DUP	Percentage of duplicated reads	Library complexity
MEDIAN_COVERAGE	Median sequencing depth over target regions	Coverage statistics
MEAN_COVERAGE	Average sequencing depth across target regions	Coverage statistics
PCT_X_COV	Percentage of target bases with ≥X coverage	Coverage completeness
PCT_0.4X_MEAN	Percentage of target bases covered at ≥40% of the mean coverage	Coverage uniformity
MIN_COVERAGE	Minimum observed coverage across target regions	Coverage statistics
MAX_COVERAGE	Maximum observed coverage across target regions	Coverage statistics
PANEL_SIZE	Total size (bp) of the target regions (from BED file)	Panel information

Multi-level coverage assessment

Using the per-base coverage data generated by Mosdepth, ClinBioNGS performs a comprehensive multi-level coverage analysis. For each DNA and RNA sample, coverage metrics—including mean, minimum, and maximum coverage, various PCT_X_COV thresholds (e.g., percentage of bases covered at ≥5×, ≥10×, etc.), and panel size—are summarized in structured tables across multiple resolution levels:

- By chromosome: coverage across all target regions within each chromosome.
- By target region: individual coverage statistics for each defined interval in the panel BED file.
- By gene: coverage over specific loci, including target regions, coding regions, exons, and the entire gene body.
- By exon: per-exon coverage for each target gene.

Gene-based and exon-level coverage calculations are based on coordinates derived from the MANE SELECT transcript, ensuring consistency and clinical relevance in exon definitions and coding sequence boundaries.

For visualization of gene-level coverage, multiple plots are generated using the karyoploteR R package[125], based on the previously calculated coverage metrics. As with the coverage tables, visualizations are produced at various levels of resolution:

- Global and per-chromosome plots display the mean coverage of each target gene across its genomic location. These plots can highlight genes from a user-defined whitelist and can annotate genes with low coverage (below a user-defined threshold) in red for quick identification.
- For individual gene visualization, detailed plots represent the observed coverage along the MANE SELECT transcript structure, including labelled exons and distinguishing coding from non-coding regions. Coverage across the entire gene locus is plotted, and target regions from the panel manifest (BED file) are annotated to distinguish on-target from off-target areas.

Each coverage table and plot are saved as a separate file for inclusion in the final report and are also integrated into an Excel summary file to facilitate interactive review and sample-level consultation.

3.1.6. Small variant analysis

3.1.6.1. Small variant calling

Following DNA processing, high-quality, deduplicated reads are used to detect small variants (SNVs and InDels) by comparison to the reference genome. By default, ClinBioNGS applies a ±25 bp padding to the target regions to capture variants in flanking sequences. This padding is omitted for amplicon-based panels, where primer sequences may artificially extend coverage beyond the region of interest. In such cases, off-target read ends are clipped using Samtools to avoid their inclusion in variant calling. Notably, this trimming step is unnecessary for Ion Torrent panels, as their alignment process inherently accounts for primer trimming.

ClinBioNGS uses a multi-caller ensemble strategy for robust tumor-only small variant detection, integrating five independent variant callers: Mutect2[127], Pisces[92], VarDict[101], Octopus[91], and Torrent Variant Caller (TVC)[96] (specific to Ion Torrent data). To improve efficiency and reduce computation time, variant calling is parallelized by chromosome: the target BED file is split per chromosome, and each caller processes these partitions independently, producing raw VCF files per sample, chromosome, and caller.

Subsequently, all raw VCFs are merged. Variants are retained if they pass caller-specific filters and overlap with the defined target regions. This filtering is performed using Bcftools[78]. To ensure compatibility across callers, multiallelic variants are decomposed and InDels normalized using Vt[60].

A custom R script is used to generate a representative list of consensus unique variants and provide calling-related metrics. This process includes the following steps:

- Intra-caller variant consensus:
  1) Extract all variants and associated metrics from each processed VCF.
  2) Resolve multiallelic positions by selecting the variant with the highest VAF, ensuring only the most representative variant per locus is retained.
- Inter-caller comparison:
  1) Each variant is annotated with the number of matching callers (exact genomic change) and overlapping callers (based on genomic positions).
  2) All variants from all callers are consolidated into a single coordinate-sorted table.
- Inter-caller consensus:
  1) Identify overlapping variant groups. As variants are coordinate sorted, each variant position is compared with the next one until no overlap is found to identify those groups.
  2) For each group, select the most recurrent variant (i.e., supported by the highest number of match callers). In the case of ties, the variant with the highest VAF is chosen.
  3) Among matching callers, the variant with the highest VAF provides the primary set of metrics for the consensus output.
  4) All supporting metrics from other overlapping variants are retained and annotated for traceability (collapsed with commas if multiple).
- Variants are lifted over from hg38 to hg19 coordinates.
- Variants are annotated with panel-specific hotspots and blacklisted regions provided by the user.

The final output includes a summary table and a consensus VCF file containing all unique small variants along with their metrics. An overview of the key metrics captured during this process is provided in **Table 7**.

Table 7. Small variant calling metrics provided by ClinBioNGS.
Each metric is accompanied by a description. Metrics are listed in alphabetical order.

Metric	Description
AD_ALT	Number of reads supporting the alternate allele
AD_ALT_<CALLER>	Alternate read count from each caller
AD_REF	Number of reads supporting the reference allele
AD_REF_<CALLER>	Reference read count from each caller
AF	Allele frequency from the selected caller. It refers to the VAF.
AF_<CALLER>	Allele frequency reported by each individual caller. It refers to the VAF.
ALT	Alternate allele
CALLERS	List of overlapping callers (e.g., Mutect2, Pisces, VarDict, Octopus, TVC)
CHROM	Chromosome on which the variant is located
DP	Total read depth at the variant position
DP_<CALLER>	Read depth from each caller
END	End position of the variant (hg38 reference)
END_HG19	End position (hg19 reference)
FILTER	Primary flag assigned by ClinBioNGS
MATCH_CALLERS	Number of callers reporting the same variant
OVERLAP_CALLERS	Number of callers with overlapping variants
PANEL_BLACKLIST	Indicates if the variant overlaps a user-defined blacklisted region
PANEL_HOTSPOT	Indicates if the variant overlaps a user-defined hotspot region
REF	Reference allele
START	Start position of the variant (hg38 reference)
START_HG19	Start position (hg19 reference)
TYPE	Variant type (i.e., SNV, INDEL)
VAR	Variant representation (i.e., <chrom>:<pos>_<ref>/<alt>)
VAR_HG19	Variant representation in hg19 coordinates (lifted-over)
VAR_<CALLER>	Variant representation from each caller (comma-separated if multiple)

3.1.6.2. Small variant annotation

Small variants are comprehensively annotated using VEP (v113), supplemented with information from multiple external resources (see *3.1.2. Pipeline's resources preparation* section for details on resource preparation). Annotation is performed at the run level, whereby all consensus VCF files from individual samples are aggregated and the unique set of variants is annotated with VEP.

The resulting annotated VCF is further processed with a custom R script to integrate additional data from external databases and to structure the information for downstream interpretation. A run-specific annotation table is generated, linking each annotated variant back to the corresponding samples in which it was detected. An overview of the information provided in the run-specific annotation table is presented in **Table 8**. All this information is used in downstream analysis for flagging and prioritizing the small variant results.

Table 8. Small variant annotation provided by ClinBioNGS.
Each term is accompanied by a description and its corresponding source. Terms are listed in alphabetical order.

Term	Description	Source
AA	Amino acid change	VEP
AlphaMissense_SCORE	AlphaMissense score	AlphaMissense (VEP)
AlphaMissense_TERM	AlphaMissense classification (likely pathogenic: ≥0.5, default)	Custom
APPRIS	APPRIS principal isoform (e.g., P1-5, A1-2)	VEP
BIOTYPE	Transcript biotype (e.g., protein coding, ncRNA)	VEP
CANONICAL	Indicates if the transcript is canonical (Ensembl-based)	VEP
CANONICAL_DRIVER	Classified as a canonical driver gene by NCG	NCG
CADD_SCORE	CADD score	CADD (VEP)
CADD_TERM	CADD classification (likely pathogenic: ≥15, default)	Custom
CCDS	Consensus coding DNA sequence identifier	VEP
CDNA_POS	Position in cDNA (position/length)	VEP
CDS_POS	Position in coding sequence (position/length)	VEP
CIViC_<term>	CIViC's term (e.g., variant ID, alteration, evidence level, rating, type, effect, drug, tumor)	CIViC
CLASS	Variant class (e.g., SNV, insertion, deletion)	VEP
ClinVar_<term>	ClinVar's term (e.g., allele ID, clinical significance, review status, disease name, allele origin)	ClinVar (VEP)
CODING	Coding region (based on CDS_POS)	VEP
CODONS	Affected codons	VEP
CONSEQUENCE	Variant effect on transcript	VEP
CTR_REGION	Located in a high-confidence CTR region	SEQC2
dbSNP_ID	Identifier in dbSNP (based on EXISTING_VARIATION)	VEP
DRIVER	Indicates driver gene (found in NCG resource)	NCG
EXISTING_VARIATION	Known variant IDs (co-located)	VEP
EXON	Affected exon (number/total)	VEP
GENE_ENSEMBL	Ensembl gene identifier	VEP
GENE_HGNC	HGNC gene symbol	VEP
GENE_SYMBOL	Gene name	VEP
GENIE_CNT	Mutation count in GENIE cancer registry	GENIE
gnomAD_MAX_AF	Maximum pVAF in gnomAD (excluding AMI, ASJ, FIN, MID, and "Remaining Individuals")	gnomAD (VEP)
gnomADe_<POP>_AF	Exome allele frequency by population	gnomAD (VEP)
gnomADg_<POP>_AF	Genome allele frequency by population	gnomAD (VEP)
HGVSg	Genomic HGVS notation	VEP
HGVSc	Coding HGVS notation	VEP
HGVSp	Protein HGVS notation	VEP
HGVSp_SHORT	Short protein change	Custom
HOTSPOT_MUT_CNT	Mutation count in Cancer Hotspots	Cancer Hotspots
HOTSPOT_POS_CNT	AA position count in Cancer Hotspots	Cancer Hotspots
IMPACT	Predicted functional impact (e.g., High, Low, Moderate)	VEP
INTRON	Affected intron (number/total)	VEP
MANE_PLUS_CLINICAL	Transcript in the MANE Plus Clinical set	VEP
MANE_SELECT	Transcript in the MANE Select set	VEP
MMR_GENE	Mismatch repair gene	MSigDB
MUT_ID	Mutation ID (<GENE_SYMBOL>_<MUTATION>)	Custom
MUTATION	Human-readable mutation name (abbreviated AA change)	Custom
NMD_ESCAPE	Nonsense-mediated mRNA decay escaping variant	VEP
ONCOGENE	Indicates oncogene based on NCG	NCG
ONCOGENE_EVIDENCE	Supporting oncogene evidence in NCG	NCG
ONCOGENIC_SOP_MUT	Previously classified mutation as "oncogenic" using ClinGen/CGC/VICC SOP	ClinGen/CGC/VICC GENIE

ONCOGENIC_SOP_POS	AA position in a previously "oncogenic" mutation (ClinGen/CGC/VICC SOP)	ClinGen/CGC/VICC GENIE
ONCOGENIC_VALID_MUT	Mutation with oncogenic effect in functional studies	CGI/CIViC
ONCOGENIC_VALID_VAR	Variant with oncogenic effect in functional studies	CGI/CIViC
PROBLEMATIC_REGION	Located in a problematic region	UCSC/GIAB
PROTEIN_ENSEMBL	Ensembl protein identifier	VEP
PROTEIN_POS	AA position (position/length)	VEP
RECURRENT	Panel-specific recurrent variant	Custom
REVEL_SCORE	REVEL score	REVEL (VEP)
REVEL_TERM	REVEL classification (likely pathogenic: ≥0.5, default)	Custom
STRAND	Transcript strand	VEP
SOMATIC_WHITELIST	Located in a known somatic position (hotspot evidence)	GENIE (BED)
TRANSCRIPT_ENSEMBL	Ensembl transcript identifier	VEP
TRANSCRIPT_REFSEQ	RefSeq transcript identifier	VEP
TSG	Indicates TSG based on NCG	NCG
TSG_EVIDENCE	Supporting TSG evidence in NCG	NCG
TSL	Transcript support level (e.g., 1-5)	VEP

3.1.6.3. Small variant flagging

ClinBioNGS incorporates a systematic flagging system to distinguish high-confidence small variants from those with lower reliability. This classification is based on a series of predefined flags applied using information generated during both the variant calling (**Table 7**) and annotation (**Table 8**) stages.

Flags that assess the variant calling process are called primary flags. Variants may be flagged for:

- Low read support, based on metrics such as AD_ALT, VAF, and DP.
- Insufficient caller support assessed using the OVERLAP_CALLERS metric.
- Localization within user-defined blacklisted regions (PANEL_BLACKLIST). Notably, variants located in user-defined hotspot regions (PANEL_HOTSPOT) are exempt from blacklist flagging, and custom read support thresholds can be defined for such cases.

Flags that provide additional variant context from post-calling annotation to further refine confidence assessments are called secondary flags. These include:

- Non-hotspot germline variants based on maximum pVAF >0.05% (following GENIE germline filtering), or observed VAF >90%, suggesting potential homozygosity.
- Non-hotspot variants located outside high-confidence regions (CTR_REGION) or within problematic regions (PROBLEMATIC_REGION).
- Panel-specific recurrent variants (RECURRENT), suggesting systematic technical artifacts or common population variants.

Hotspot variants are defined as those meeting any of the following:

- Mapped to panel-specific hotspot regions (PANEL_HOTSPOT).
- Located in a known somatic position (SOMATIC_WHITELIST).
- Mutations recorded in the Cancer Hotspots database (HOTSPOT_MUT_CNT ≥1).

Variants that do not meet any low-confidence criteria are considered high-confidence and labeled as "OK".

A summary of small variant primary and secondary flags applied by ClinBioNGS is presented in **Table 9**. Thresholds associated with each flag are fully customizable, allowing users to adapt the stringency of the analysis to specific requirements.

Table 9. Description of flags used by ClinBioNGS to assess small variant confidence.
Each flag is defined along with its description and the corresponding pipeline step in which it is applied.

Pipeline step	Flag	Description
Small variant calling (primary flags)	Blacklist	Non-hotspot variant located within a panel-specific blacklisted region
	LowAD	AD_ALT < 5 reads
	LowCallers	Overlapping callers < 2 (< 3 for Ion Torrent panels)
	LowDP	Total read depth (DP) <10 reads
	LowVAF	VAF <1%
	PASS	Variant passed all calling-related quality filters
Small variant annotation (secondary flags)	Germline	Non-hotspot variant with gnomAD_MAX_AF >0.05% or AF >90%
	NoCall	Variant failed calling-related primary flags
	OutCTR	Non-hotspot variant located outside CTR region
	ProblematicRegion	Non-hotspot variant located within a problematic genomic region
	Recurrent	Variant identified as recurrent in panel background samples
	OK	High-confidence variant (passed all predefined flags)

3.1.6.4. Small variant prioritization

To assess the potential clinical and biological relevance of detected small variants, ClinBioNGS implements two complementary classification frameworks: oncogenicity and clinical significance.

Oncogenicity

Oncogenic potential is evaluated according to the SOP developed by the ClinGen/CGC/VICC consortium. Each variant is scored based on the strength of supporting evidence across multiple predefined categories. These scores are summed up using a point-based system to assign the variant to one of the following five categories:

- Oncogenic: ≥10 points.
- Likely oncogenic: 6 to 9 points.
- VUS: 0 to 5 points.
- Likely benign: -1 to -6 points.
- Benign: ≤ -7 points.

A detailed description of the scoring criteria and evidence types is provided in **Supplementary Table 3**. ClinBioNGS incorporates 14 of the 17 evidence categories described in the SOP. The remaining evidence types—OM1, OP2, and SBS2—are not currently implemented due to unavailability of the required input data.

Color coding is applied to facilitate interpretation: shades of red indicate pathogenicity, yellow corresponds to uncertain significance, and shades of green represent benign classifications. These colors are reflected in the visualizations and tabular entries within the final sample report.

Clinical significance

Clinical significance is classified according to the AMP/ASCO/CAP guidelines, based on curated data from the CIViC database—an open-access, expert-reviewed resource endorsed by the ClinGen Somatic Cancer Working Group. The classification integrates tumor-specific therapeutic, prognostic, and diagnostic evidence from CIViC with the variant's oncogenicity status, and assigns each variant to one of four clinical tiers:

- Tier I: Strong clinical significance (supported by high-level CIViC evidence).
- Tier II: Potential clinical significance (supported by CIViC evidence).
- Tier III: Unknown clinical significance (no CIViC evidence and not classified as benign/likely benign).
- Tier IV: Benign or likely benign (no CIViC evidence and classified as benign or likely benign based on oncogenicity).

A comprehensive description of classification rules and tier definitions is available in **Supplementary Table 4**.

3.1.6.5. Small variant results

For each sample, small variant results are provided in an annotated VCF file containing both variant- and sample-level information. In addition, ClinBioNGS generates summary tables and visualizations using custom R scripts to facilitate result interpretation. Two separate summary tables are produced: one for somatic variants and another for likely germline variants.

Variants are prioritized according to the following criteria:

1) Clinical significance: Variants classified as Tier I (strong) or Tier II (potential) are listed first.
2) Flagging status: High-confidence variants flagged as "OK" are prioritized, followed by those without any calling-related flags ("PASS").
3) Panel hotspot location: Variants located in panel-defined hotspot regions are given higher priority.
4) Oncogenicity: Variants are ranked by oncogenicity score, from most to least likely to be pathogenic.
5) Whitelist gene inclusion: Variants occurring in genes defined in the user-provided whitelist are prioritized.

For visualization purposes, only genes containing clinically or biologically relevant variants are plotted. These include genes with: (i) Tier I or Tier II variants, (ii) variants classified as oncogenic, or (iii) high-confidence ("OK") non-benign variants, particularly if located within a panel hotspot. Within each selected gene, relevant somatic variants are visualized using the karyoploteR R package. The gene structure (based on the MANE SELECT transcript) is displayed, and variants are highlighted with respect to their classification.

All summary tables and corresponding plots are saved as standalone files for inclusion in the final interactive report. Additionally, all results are consolidated into an Excel file for convenient review.

3.1.7. Analysis of CNAs

3.1.7.1. CNA calling

After DNA processing, high-quality, unique reads are used to identify CNAs. CNA calling is performed by comparing the observed coverage in each target genomic region to the expected coverage derived from a panel-specific pooled reference cohort.

Coverage differences are calculated across predefined, panel-specific target regions (i.e., bins) using CNVkit[47]. First, the "coverage" module computes coverage values for each region. Then, the "fix" module applies a two-step normalization and correction process to generate a table of corrected copy ratios (log2 fold changes) for each bin:

- Intra-sample normalization: Calculated coverage values are median-centered (i.e., subtracting the median), bias-corrected based on GC content, the fraction of masked repeats, and (for hybrid-capture panels) target density, and converted to log2 scale.
- Reference-based correction: The log2 coverage values from the reference baseline are subtracted from each bin. Additionally, each bin is assigned a weight based on its genomic size and coverage variability within the baseline cohort.

Subsequently, gene-level copy ratios are computed as the weighted mean of all associated bins, and absolute gene copy numbers (CNs) are estimated using fixed thresholds (**Table 10**) with a custom R script. Each gene is then assigned to a CNA status based on the inferred copy number:

- Neutral: Two copies (CN = 2).
- AMP: More than two copies (CN ≥ 3), further subclassified as:
    - "LowAmp": Low-level AMP with CN < 5.
    - "HighAmp": High-level AMP with CN ≥ 5.
- DEL: Fewer than two copies (CN < 2), further subclassified as:
    - "LowDel": Low-level DEL with CN = 1 (suggesting heterozygous DEL).
    - "HighDel": High-level DEL with CN = 0 (suggesting homozygous DEL).

Notably, CN values are not adjusted based on TP estimates, to maintain consistency across samples and ensure reproducibility in tumor-only analyses. An overview of the CNA classification thresholds is provided in **Table 10**.

Table 10. Classification criteria for gene-level CNAs.
CNA classification is based on fixed copy ratio thresholds and corresponding absolute CN values.

Log2 copy ratio thresholds	CN	CNA	CNA class
-Inf < log2 ≤ -1.20	0	DEL	HighDel
-1.20 < log2 ≤ -0.40	1	DEL	LowDel
-0.40 < log2 ≤ 0.40	2	NEUTRAL	Neutral
0.40 < log2 ≤ 0.80	3	AMP	LowAmp
0.80 < log2 ≤ 1.20	4	AMP	LowAmp
1.20 < log2 ≤ 1.50	5	AMP	HighAmp
1.50 < log2 ≤ 1.70	6	AMP	HighAmp
1.70 < log2 ≤ 1.91	7	AMP	HighAmp
1.91 < log2 ≤ 2.09	8	AMP	HighAmp
2.09 < log2 ≤ 2.25	9	AMP	HighAmp
2.25 < log2 ≤ 2.39	10	AMP	HighAmp
…	…	…	…
5.64 < log2 < Inf	100	AMP	HighAmp

CNA calling is performed across all autosomes and chromosome X (chrX). Chromosomal sex can be inferred from coverage patterns using CNVkit; however, if user-provided sex metadata is available, it will take precedence. For male samples, the chrX copy ratio is adjusted by applying a log2 + 1 transformation to account for haploidy. This adjustment is used solely for classification purposes—raw copy ratios are retained for transparency in both tables and visualizations.

In hybrid capture panels (e.g., Illumina TSO500), CNVkit also performs segmentation to infer genomic regions with consistent copy-number signals using the circular binary segmentation algorithm by default, recommended for mid-size panels and exomes[47]. These segments are further processed using Arm-level Somatic Copy-number Events in Targeted Sequencing (ASCETS)[76] tool to estimate arm-level CNAs, particularly useful in data with off-target reads that enhance segmentation resolution. Each chromosomal arm is assigned one of the following CNA statuses:

- NEUTRAL: Mean log2 copy ratio between ± 0.40.
- AMP: Mean log2 copy ratio ≥ 0.40.
- DEL: Mean log2 copy ratio ≤ -0.40.
- CONFLICT: Inconsistent signal; alteration fraction < 0.7 (default threshold).

Color coding of CNA categories is applied to facilitate interpretation: shades of blue indicate AMPs, while shades of red represent a DEL status. These colors are reflected in the visualizations and tabular entries within the final sample report.

3.1.7.2. CNA annotation

Gene-level CNAs identified by ClinBioNGS are annotated using a custom R script that integrates information from several external resources (**Supplementary Table 2**). The annotation process provides the following key features for each target gene:

- Genomic cytoband location derived from UCSC resources to facilitate chromosomal context.
- MANE transcript coverage determines the percentage of the gene or exons covered by targeted bins, based on the MANE SELECT transcript model.
- Gene role in cancer classification as an oncogene, TSG, or general cancer driver using NCG data.
- Frequency in the AACR GENIE registry reports how frequently CNAs in the gene occur across cancer types based on a large tumor cohort.
- Clinical evidence retrieved from the CIViC database, including annotations relevant to therapy response, prognosis, and diagnosis.

A summary of the annotations provided per gene is shown in **Table 11**. This information supports downstream interpretation by enabling evidence-based flagging and prioritization of clinically relevant CNA events.

Table 11. Gene-level CNA metrics and annotation provided by ClinBioNGS.
Each term is accompanied by a brief description. Terms are sorted alphabetically.

Term	Description
ALTERATION	Formatted name of the general CNA status (<GENE>_<CNA>)
BINS_TARGET	Number of on-target bins overlapping the gene
BINS_TOTAL	Total number of bins overlapping the gene (including off-target bins, if applicable)
CANONICAL_DRIVER	Indicates canonical driver gene based on NCG
CHROM	Chromosome where the gene is located
CIViC_<term>	CIViC's term (e.g., variant ID, alteration, evidence level, rating, type, effect, drug, tumor)
CLASS	Specific CNA class (e.g., HighAmp/Del, LowAmp/Del, Neutral)
CN	Estimated absolute copy number for the gene (e.g., 0-100)
CNA	Simplified CNA category (e.g., AMP, DEL, NEUTRAL)
CYTOBAND	Cytogenetic band location of the gene (UCSC)
DEPTH	Weighted mean read depth across on-target bins
DRIVER	Indicates driver gene based on NCG
END	End coordinate of the gene (last bin end position)
GENE	Gene symbol (HGNC)
GENE_ENSEMBL	Ensembl gene ID (based on MANE annotations)
GENIE_CNT	Number of samples with CNA in this gene from the GENIE registry
GENIE_FREQ	Frequency of CNA in this gene in the GENIE registry
LOG2	Weighted mean log2 copy ratio across overlapping bins
ONCOGENE	Indicates an oncogene based on NCG

PERC_COV_GENE_ALL	Percentage of total gene length covered by bins (based on MANE annotations)
PERC_COV_GENE_EXON	Percentage of exonic gene length covered by bins (based on MANE annotations)
START	Start coordinate of the gene (first bin start position)
STRAND	Strand orientation of the gene (e.g., + or −, based on MANE annotations)
TRANSCRIPT_ENSEMBL	Ensembl transcript ID for MANE transcript
TRANSCRIPT_REFSEQ	RefSeq transcript ID for MANE transcript
TSG	Indicates a tumor suppressor gene based on NCG
VAR	Formatted name of the specific CNA status (<GENE>_<CLASS>)
WEIGHT	Sum of weights assigned to each overlapping bin based on size and variability
WHITELIST	Indicates whether the gene is included in a user-defined whitelist

3.1.7.3. CNA flagging

ClinBioNGS implements a systematic flagging approach to differentiate high-confidence gene-level CNAs from those with lower reliability. This classification relies on a set of predefined flags applied to each gene-level CNA based on the previously collected metrics (**Table 11**).

Gene CNAs may be flagged under the following conditions:

- Neutral or low-level alterations.
- Insufficient bin support, determined by the number of on-target bins overlapping the gene.

Gene CNAs that do not meet any low-confidence criteria are considered high-confidence and are labeled as “OK”.

A summary of the CNA flags used by ClinBioNGS is provided below in **Table 12**. All thresholds used for flagging are fully customizable, allowing users to tailor the stringency of the analysis according to specific needs.

Table 12. Description of flags used by ClinBioNGS to assess CNA confidence.
Each flag includes a brief description of the criteria used to assess gene-level CNA confidence.

Flag	Description
LowAmp	Low-level AMP with estimated CN < 5
LowBins	Gene with < 4 on-target bins
LowDel	Low-level DEL with estimated CN = 1
Neutral	Gene has two copies (no CNA event is called)
OK	High-confidence CNA (passes all flag criteria)

3.1.7.4. CNA prioritization

Gene-level CNAs are prioritized based on clinical significance following the AMP/ASCO/CAP joint consensus guidelines. This classification incorporates tumor-specific therapeutic, prognostic, and diagnostic evidence from the CIViC database, as well as gene-level CNA frequencies from the AACR GENIE registry. Based on this information, CNAs are assigned to one of four clinical significance tiers (**Supplementary Table 4**):

- Tier I/II (strong/potential clinical significance): CNAs with curated clinical evidence in CIViC.
- Tier III (unknown clinical significance): CNAs without CIViC evidence but observed at a frequency ≥ 0.1% in the GENIE cancer registry.
- Tier IV (benign or likely benign): CNAs lacking both CIViC evidence and sufficient prevalence in GENIE.

This tier-based system enables the prioritization of CNAs with potential diagnostic, prognostic, or therapeutic relevance and helps to filter out alterations less likely to be clinically meaningful.

3.1.7.5. CNA results

For each sample, CNA results are delivered in both an annotated VCF file and accompanying summary tables and plots generated using a custom R script.

Gene-level CNAs are prioritized using the following criteria:

1) Clinical significance, with Tier I and Tier II CNAs listed first.
2) Flagging status, giving precedence to high-confidence CNAs labeled as “OK”.
3) Whitelist gene inclusion, prioritizing CNAs in user-defined genes of interest.

Visualizations are produced with the karyoploteR R package. Genes flagged as “LowBins” are excluded from plots unless they are clinically relevant. By default, only Tier I/II CNAs or CNAs affecting whitelist genes are labeled in the global overview and detailed gene-level plots. For small panels, an optional setting allows plotting of all targeted genes regardless of their prioritization.

All CNA-related outputs—including summary tables and plots—are saved as individual files for inclusion in the final interactive HTML report. Additionally, they are consolidated into an Excel file for convenient review and distribution.

3.1.7.6. Panel-specific CNA baseline construction

Panel-specific CNA baselines for the Illumina TSO500 and Thermo Fisher OPA and OCA panels were generated using multiple CNVkit modules. Because matched normal samples were unavailable, large tumor cohorts were leveraged to identify samples with low coverage variability, presumed to approximate normal-like profiles suitable for baseline construction.

First, a BED file defining the accessible regions of the GRCh38 genome was created with the CNVkit “access” module. This step excluded problematic loci, including centromeres, telomeres, long stretches of “N” bases, and difficult regions flagged by GIAB stratification files (e.g., false duplications, polymorphic sites, low-mappability regions). From these accessible regions and the panel-specific target BEDs, anti-target BED files were generated with the “antitarget” module to

define off-target regions. For amplicon-based panels (OPA and OCA), off-target regions were omitted because they are not sequenced.

Next, using these files, a flat reference model assuming a neutral CN (i.e., log2 = 0.0) for each region was built with the "reference" module. This model incorporates GC content and repeat-masked proportions to correct for systematic biases.

Subsequently, as described in the CNA calling section, bin-level coverage and log2 copy ratios were computed for large in-house tumor cohorts (N = 655 for TSO500, N = 623 for OPA, and N = 537 for OCA) using the flat reference. For hybrid-capture panels such as TSO500, segmentation and quality metrics were additionally computed with the "segment" and "metrics" modules, respectively.

To evaluate coverage variability and select appropriate samples for the reference cohort, a custom R script applied the following filtering steps:

1) Retain only autosomal target regions to avoid sex-related variability.
2) Exclude bins lacking coverage or showing extreme log2 values (≤ –5 or ≥ 5), following CNVkit recommendations.
3) For each bin, calculate variability thresholds as the median ± 1 median absolute deviation (MAD) across all samples, assuming most coverage values approximate those from normal samples.
4) For each sample, compute:
   - Percentage of bins falling within the defined variability thresholds (i.e., "normal" bins).
   - Absolute value of the global weighted mean log2 copy ratio.
5) For hybrid-capture data (e.g., TSO500), estimate a noisiness score as the product of the number of segmented regions and the biweight midvariance reported by CNVkit.

Samples were included in the final reference baseline if they met all of the following criteria, indicating the lowest coverage variability within the cohort (**Supplementary Figure 1**):

- ≥90% of bins classified as "normal".
- Global weighted mean log2 copy ratio ≤0.1 (absolute value).
- Noisiness score below the cohort median (criterion applied only to TSO500).

3.1.8. Analysis of gene fusions

3.1.8.1. Fusion calling

Following RNA processing, aligned reads (BAM files) and chimeric junctions are analyzed using STAR-Fusion[94], a component of the CTAT toolkit[82]. The fusion calling process involves the following key steps:

- Detection of candidate fusion transcripts by mapping junction and spanning reads to a curated reference annotation set from the CTAT library.
- *In silico* validation (*--FusionInspector validate*): The full set of input reads is realigned to a combined reference composed of the standard genome and a set of fusion-gene contigs—synthetic constructs that model candidate fusion partners in their proposed fused orientation. Reads that align more accurately to the fusion contigs than to the reference genome are identified and reported as supporting the fusion. Additionally, non-fused reads that misalign across the fusion junction are also identified and quantified.
- Prediction of fusion impact on coding sequences, with classification of the resulting chimeric proteins (e.g., in-frame, frameshift).

STAR-Fusion outputs a table listing candidate fusions along with relevant metrics. These results undergo additional processing with a custom R script, which performs the following operations:

- Filtering of invalid fusions, removing those involving atypical chromosomes or gene pairs outside the panel's target regions.
- Calculation of fusion-supporting reads, defined as the sum of:
    - Junction reads: split reads that span the predicted fusion breakpoint.
    - Spanning fragments: paired-end reads mapping to different fusion partners.
- Estimation of fusion depth, defined as the sum of fusion-supporting reads and non-fused partner reads (the latter obtained during *in silico* validation).
- Computation of fusion AF as the proportion of fusion-supporting reads relative to the total fusion DP.
- Coordinate liftover from hg38 to hg19 for fusion breakpoints.

3.1.8.2. Fusion annotation

Fusion candidates identified by STAR-Fusion are annotated using the CTAT resource, which provides detailed information on the genes involved, known fusion artifacts, and events commonly detected in normal tissues—helping to distinguish cancer-related fusions from non-relevant or technical artifacts.

Additional annotations are incorporated using a custom R script that integrates multiple external resources (**Supplementary Table 2**). The following steps are applied:

- Annotate the exon or intron affected at each fusion breakpoint based on MANE SELECT transcript annotations.
- If the breakpoint lies within an intron, calculate the distance (in bp) to the nearest exon.
- Define the fused genomic region for each partner gene (from the gene start to the breakpoint, or from the breakpoint to the gene end) based on MANE annotations.

- Calculate the mean exon coverage within and outside the fused regions using per-base coverage data from Mosdepth.
- Annotate the role of each gene in cancer (e.g., oncogene, TSG, general driver) using the NCG database.
- Report fusion frequencies in the AACR GENIE registry and Mitelman Database.
- Match against a whitelist of curated, known fusion events (**Supplementary Table 5**).
- Integrate curated clinical evidence from the CIViC database.

A summary of all fusion-related annotations is provided in **Table 13**. These annotations are used in downstream analyses to support the flagging and prioritization of fusion events based on their potential biological and clinical relevance.

Table 13. Fusion metrics and annotations provided by ClinBioNGS.
Each entry includes a description and its corresponding data source. Entries are listed in alphabetical order.

Entry	Description	Source
AD	Number of fusion-supporting reads (junction + spanning)	Custom
AD_NonFused_A/B	Number of non-fused reads for partner A/B	STAR-Fusion
AF	Allele frequency of the fusion, calculated as AD / DP	Custom
BASES_FROM_EXON_A/B	Distance (in bp) from intronic breakpoint to closest exon	Custom
BREAKPOINT_A/B	Genomic breakpoint coordinate	STAR-Fusion
BREAKPOINT_A/B_HG19	Lifted-over breakpoint coordinate in hg19	Custom
CANONICAL_DRIVER_A/B	Indicates whether the gene is a canonical cancer driver	NCG
CHROM_A/B	Chromosome of the fusion breakpoint	STAR-Fusion
CIViC_<term>	CIViC's term (e.g., variant ID, alteration, evidence level, rating, type, effect, drug, tumor)	CIViC
COV_IN_FUSION_A/B	Mean exon coverage within fused genomic region	Custom
COV_OUT_FUSION_A/B	Mean exon coverage outside fused region	Custom
DP	Total fusion depth: fusion-supporting reads + non-fused reads	Custom
DRIVER_A/B	Indicates whether the gene is cancer driver	NCG
EXON_A/B	Exon involved in the fusion breakpoint	MANE
FFPM	Fusion fragments per million, normalized fusion expression	STAR-Fusion
FUSION_NAME	Formatted name: <geneA>::<geneB> (<exonA>::<exonB>)	Custom
FUSION_RANGE_A/B	Genomic coordinates of fused region	Custom
FUSION_SHORT	Short representation of the fusion (<geneA >::< geneB >)	STAR-Fusion
FUSION_VARIANT	Fusion variant label based on fusion whitelist match	Whitelist
GENE_A/B	HGNC gene symbol	STAR-Fusion
GENE_ENSEMBL_A/B	Ensembl gene identifier	STAR-Fusion
GENIE_CNT	Number of samples with this fusion in the GENIE database	GENIE
INTRON_A/B	Intron involved in the fusion breakpoint	MANE
InSilicoValid	Indicates if fusion was validated by *in silico* realignment	STAR-Fusion
JunctionReadCount	Number of reads split across the fusion junction	STAR-Fusion
LargeAnchorSupport	Indicates whether there are split reads with almost 25bp aligned on both sides of breakpoint	STAR-Fusion
Left/RightBreakDinuc	Dinucleotide sequence at each breakpoint	STAR-Fusion
MitelmanDB_CNT	Number of samples with this fusion in the Mitelman Database	MitelmanDB
MODEL_CDS_A/B	Coding sequence identifier of fusion protein model	STAR-Fusion
MODEL_CDS_RANGE_A/B	Coding coordinates of fusion protein model	STAR-Fusion
MODEL_FUSION_TYPE	Predicted fusion protein type (e.g., in-frame, frameshift)	STAR-Fusion
Normal	Indicates if fusion is commonly observed in normal samples	STAR-Fusion
ONCOGENE_A/B	Indicates an oncogene partner	NCG

POS_A/B	Genomic position of the fusion breakpoint	STAR-Fusion
RefSpliceSite	Indicates if the fusion uses a canonical splice site	STAR-Fusion
RTartifact	Indicates if fusion is a known RT artifact	STAR-Fusion
SpanningFragCount	Number of paired-end reads mapping to different partners	STAR-Fusion
STRAND_A/B	Strand orientation of the transcript	STAR-Fusion
TRANSCRIPT_ENSEMBL_A/B	Ensembl transcript identifier	MANE
TRANSCRIPT_REFSEQ_A/B	RefSeq transcript identifier	MANE
TSG_A/B	Indicates a tumor suppressor gene	NCG
VAR	Formatted fusion identifier (<breakpointA>::(<breakpointB>)	Custom
VAR_HG19	Lifted-over version of fusion identifier in hg19	Custom
WHITELIST_FUSION	Indicates whether the fusion is found in the fusion whitelist	Whitelist
WHITELIST_GENE	Indicates whether either gene is in the gene whitelist	Whitelist

3.1.8.3. Fusion flagging

ClinBioNGS incorporates a systematic flagging framework to distinguish high- from low-confidence gene fusion candidates. Flags are assigned using metrics derived from both the fusion calling stage and the post-calling processing steps from **Table 13**.

During fusion calling, candidates can include primary flags for low support based on key metrics such as the number of junction and spanning reads, fusion fragments per million (FFPM) reads, and the presence or absence of large anchor support—particularly relevant when spanning reads are not detected (e.g., single-end data).

During post-calling processing, secondary flags include:

- Custom-calculated read metrics, such as the total of fused and non-fused ADs, fusion DP, and fusion VAF.
- *In silico* validation results: fusions not confirmed by STAR-Fusion's validation process are flagged accordingly.
- Known artifacts or fusions typically found in normal tissue.
- Absence in cancer-specific resources such as AACR GENIE or the Mitelman Database.

Fusions that do not trigger any predefined flag are considered high-confidence and labeled as “OK”.

A summary of fusion primary and secondary flags applied by ClinBioNGS is presented in **Table 14**. All thresholds associated with these flags are fully customizable, enabling users to adjust stringency to their specific analytical needs.

Table 14. Description of flags used by ClinBioNGS to assess fusion confidence.
Each flag is defined along with its description and the pipeline step in which it is applied.

Pipeline step	Flag	Description
Fusion calling (primary flags)	LowSupport	Junction reads < 5 FFPM < 1 No spanning reads and < 25 bp of anchor support in junction reads
	PASS	No calling-related quality issue detected
Post-calling processing (secondary flags)	LowAD	Fusion-supporting reads (junction + spanning) < 10
	LowDP	Total fusion depth (supporting + non-fused reads) < 20 (< 10 for Ion Torrent)
	LowNonFused	(AD_NonFused_A + AD_NonFused_B) < 5
	LowVAF	Fusion allele frequency (AF) < 3%
	NoCall	Flagged due to low quality at the calling stage
	NoInSilicoValid	The fusion was not confirmed by STAR-Fusion's in silico validation
	Normal	The fusion is commonly observed in normal tissue
	RTartifact	The fusion is identified as a known artifact
	Unknown	The fusion is not present in cancer databases (e.g., GENIE, MitelmanDB)
	OK	High-confidence fusion candidate; passed all flag criteria

3.1.8.4. Fusion prioritization

Gene fusions are prioritized according to clinical significance based on the AMP/ASCO/CAP joint consensus guidelines[4]. This classification integrates tumor-specific therapeutic, prognostic, and diagnostic evidence from the CIViC database, along with fusion presence in cancer-specific resources such as the AACR GENIE registry and the Mitelman Database.

Based on this information, each fusion is assigned to one of four clinical significance tiers (**Supplementary Table 4**):

- Tier I/II (strong or potential clinical significance): Fusions with curated clinical evidence in CIViC.
- Tier III (unknown clinical significance): Fusions not supported by CIViC but reported in cancer resources such as GENIE or MitelmanDB.
- Tier IV (benign or likely benign): Fusions absent from both CIViC and cancer databases.

3.1.8.5. Fusion results

Fusion results for each sample are delivered as an annotated VCF file, accompanied by summary tables and visualizations generated using a custom R script.

Fusion entries are ranked based on the following prioritization criteria:

1) Clinical significance, with Tier I and Tier II fusions listed first.
2) Presence in a whitelist of known or clinically relevant fusions.
3) Flagging status, prioritizing high-confidence fusions labeled as "OK".
4) Involvement of genes included in a user-defined whitelist.

For visualization, each fusion event is displayed alongside the gene structures of both fusion partners. These plots highlight the predicted breakpoint and fusion range, as well as sequencing coverage across the involved genes. Visualizations are created using the karyoploteR R package. Fusions flagged as "LowSupport" are excluded from plots unless they match a known fusion in the whitelist.

All summary tables and plots are saved as standalone files for inclusion in the final interactive report and are also consolidated into a comprehensive Excel file for convenient review.

3.1.9. Splice variant analysis

3.1.9.1. Splice variant calling

Following RNA processing, aligned reads (BAM files) and splice junctions are analyzed using CTAT-splicing, a component of the CTAT toolkit designed to detect potential cancer-associated splice variants. The tool maps junction reads to a curated reference annotation and outputs a table of candidate splice junctions along with their supporting read counts.

These results are further processed using a custom R script, applying the following steps:

- Exclude splice variants located in off-target genes.
- Calculate splice variant AD as the sum of supporting unique and multi-mapped reads.
- Estimate total DP at each junction breakpoint (start and end) by aggregating splice-supporting reads from all variants sharing the same coordinate.
- Compute the VAF for each splice variant as the proportion of splice-supporting reads relative to the maximum DP across both breakpoints.
- Convert junction coordinates from hg38 to hg19.

3.1.9.2. Splice variant annotation

Splice variants identified by CTAT-splicing are initially annotated using the CTAT resource, which highlights cancer-enriched junctions based on comparative analyses of tumor (TCGA) and normal (GTEx) tissues.

Additional annotation is performed using a custom R script and several external resources (**Supplementary Table 2**). The following steps are applied:

- Annotate the affected exon or intron using MANE SELECT transcript annotations.
- Calculate average coverage across affected and flanking exons using per-base coverage data from Mosdepth.
- Determine the gene's role in cancer (e.g., oncogene, TSG, general driver) using the NCG database.

- Identify known splice events by matching against a curated whitelist of cancer-associated splice variants (**Supplementary Table 6**).
- Integrate curated clinical evidence from the CIViC database.
- Cross-reference detected splice variants with small variants identified in the corresponding DNA sample to identify overlapping mutations at splice donor or acceptor sites.

A summary of the splicing-related annotations is presented in **Table 15**. These annotations are used in downstream analyses for the flagging and prioritization of splice variants.

Table 15. Splice variant metrics and annotations provided by ClinBioNGS.
Each entry includes a description and its corresponding data source. Entries are listed in alphabetical order.

Entry	Description	Source
AD	Number of splice-supporting reads (unique + multi-mapped)	Custom
AF	Allele frequency, calculated as AD / DP_MAX	Custom
CancerEnriched	Indicates if the splice variant is commonly observed in tumor tissues	CTAT
CANONICAL_DRIVER	Indicates whether the gene is a canonical cancer driver	NCG
CHROM	Chromosome on which the splice variant occurs	CTAT
CIViC_<term>	CIViC's term (e.g., variant ID, alteration, evidence level, rating, type, effect, drug, tumor)	CIViC
COV_EXONS_AFFECTED	Mean coverage across affected exons	Custom
COV_EXONS_FLANKING	Mean coverage across exons flanking the splice junction	Custom
DP_END	Read depth at the end position of the splice junction	Custom
DP_MAX	Maximum of DP_START and DP_END	Custom
DP_MEAN	Mean of DP_START and DP_END	Custom
DP_START	Read depth at the start position of the splice junction	Custom
DRIVER	Indicates whether the gene is cancer driver	NCG
END	End coordinate of the splice junction	CTAT
END_HG19	Lifted-over end coordinate in hg19	Custom
EXONS_AFFECTED	Exon(s) overlapping the splice junction	MANE
GENE	HGNC gene symbol	CTAT
GENE_ENSEMBL	Ensembl gene identifier	CTAT
GTEx	Number of samples with this splice event in normal tissues (GTEx)	CTAT
INTRONS_AFFECTED	Intron(s) overlapping the splice junction	MANE
MULTI_MAPPED_READS	Number of reads mapped to multiple genomic locations	CTAT
MUTATION	Small variant affecting donor or acceptor splice sites in DNA	Custom
ONCOGENE	Indicates whether the gene is an oncogene	NCG
REGION_AFFECTED	Affected region formatted as "Exon" or "Intron" followed by index	Custom
START	Start coordinate of the splice junction	CTAT
START_HG19	Lifted-over start coordinate in hg19	Custom
STRAND	Strand orientation of the transcript	CTAT
TCGA	Number of samples with this splice event in tumor tissues (TCGA)	CTAT
TRANSCRIPT_ENSEMBL	Ensembl transcript identifier	MANE
TRANSCRIPT_REFSEQ	RefSeq transcript identifier	MANE
TSG	Indicates whether the gene is a tumor suppressor gene	NCG
UNIQ_MAPPED_READS	Number of reads uniquely mapped to the genome	CTAT
VAR	Variant coordinates in the format "<chr>:<start>-<end>"	CTAT
VAR_GENE	Formatted name combining gene and affected region	Custom
VAR_HG19	Lifted-over variant coordinates in hg19	Custom
VAR_NAME	Specific splice variant name from the curated whitelist	Whitelist
WHITELIST_GENE	Indicates whether the gene is included in a user-defined whitelist	Whitelist
WHITELIST_SPLICING	Indicates whether the splice variant is found in a curated whitelist	Whitelist

3.1.9.3. Splice variant flagging

Splice variants are evaluated for confidence with a set of predefined flags using metrics derived from the calling and post-calling processing steps in **Table 15**.

Primary flags are based on a minimum number of total splice-supporting reads. Secondary flags are subsequently applied to refine confidence, including:

- Custom-calculated metrics such as a higher splice variant AD, DP, and VAF.
- Absence from both the CTAT database of cancer-enriched splice junctions and a curated whitelist of known variants.

Splice variants that do not trigger any of these flags are considered high-confidence and are labeled as “OK”.

A summary of splicing primary and secondary flags applied by ClinBioNGS is presented in **Table 16**. All flagging thresholds are fully customizable, allowing users to adjust the stringency of the analysis to meet specific requirements.

Table 16. Description of flags used by ClinBioNGS to assess splice variant confidence.
Each flag includes a description and the corresponding pipeline step where it is applied.

Pipeline step	Flag	Description
Splice variant calling (primary flags)	LowSupport	Splice-supporting reads (unique + multi-mapped) < 10
	PASS	No quality issue detected at the calling stage
Post-calling processing	LowAD	Splice-supporting reads < 100
	LowDP	Maximum read depth < 200
	LowVAF	Variant allele frequency < 3%
	NoCall	Variant flagged due to low support at the calling stage
	NoCancerEnriched	Variant not enriched in tumor tissues (absent from CTAT cancer database)
	OK	High-confidence variant candidate that passed all flag criteria

3.1.9.4. Splice variant prioritization

Splice variants are prioritized based on clinical significance following the AMP/ASCO/CAP tiered classification guidelines. Final tier assignments (**Supplementary Table 4**) are determined by the presence of curated CIViC evidence, inclusion in the CTAT splicing database, or matching to a known variant in the curated whitelist:

- Tier I/II (strong or potential clinical significance): Splice variants with curated clinical evidence in CIViC database.
- Tier III (unknown clinical significance): Variants without CIViC evidence but identified as cancer-enriched in the CTAT database or matched to the splicing whitelist.
- Tier IV (benign or likely benign): Variants not found in any cancer-specific resource or curated whitelist.

3.1.9.5. Splice variant results

Final splice variant results are provided per sample as annotated VCF files, along with summary tables and visualizations generated using a custom R script. Variants are prioritized based on the following criteria:

1) Clinical significance (Tier I and II).
2) Presence in the curated splicing whitelist.
3) High-confidence classification ("OK").
4) Occurrence in genes included in the user-defined whitelist.

Splice junctions are visualized on the MANE SELECT transcript using sashimi-style plots, which display splicing patterns and local read coverage. These plots are generated with the karyoploteR R package. By default, only high-confidence and cancer-enriched junctions are plotted, except for known variants from the whitelist, which are always included.

All tables and plots are saved as standalone files for inclusion in the final HTML report and are also consolidated into a single Excel file for a convenient review.

3.1.10. Analysis of genomic biomarkers

3.1.10.1. TMB

TMB is defined as the somatic Mut/Mb of interrogated genomic sequence. For each sample, the TMB score is calculated as the ratio of qualifying somatic small variants to the total length of eligible DNA target regions[50].

To ensure robust and accurate estimation, several filters are applied to both the annotated small variants and the DNA target regions used for calculation.

Target region filters (denominator): To define the length of high-confidence target regions, the following criteria are applied:

- Exclude regions with low coverage (<100 reads), based on per-base coverage data from Mosdepth.
- Remove non-coding regions that do not overlap with MANE coding regions.
- Discard regions that overlap with known problematic loci from UCSC and GIAB.

Variant filters (numerator): Only robust somatic small variants are considered, based on the following conditions:

- Located within the high-confidence regions defined above.
- Meet minimum read support thresholds (default values):
  - AD_ALT ≥ 5.
  - DP ≥ 100.
  - VAF ≥ 5%.
- Absent from the gnomAD database (gnomAD_MAX_AF = 0%) or with an observed AF ≤ 90%, to exclude likely germline variants. This population filter is stricter than in other steps (e.g., variant flagging), as it has been shown to better align with TMB estimates from WES, the current gold standard[50].
- Not flagged as panel-specific recurrent, to remove potential artifacts or population-specific germline events.
- Not classified as hotspot, oncogenic/likely oncogenic, or clinically relevant (Tier I/II), since these known pathogenic variants can artificially inflate TMB scores—particularly in panels enriched for cancer-related genes[50].

Additionally, ClinBioNGS also reports an alternative TMB score that includes only non-synonymous variants from the eligible set, for panels where this calculation is recommended. However, it has been shown that excluding synonymous variants has a minimal impact on the approximation of TMB to WES values[50].

Finally, two result tables are generated for each sample: one containing the calculated TMB scores along with associated metrics (e.g., number of eligible variants and effective region size), and another serving as a TMB trace table, listing all evaluated small variants with the corresponding evidence used for inclusion or exclusion. Both tables are incorporated into the final HTML report and are also compiled into a single Excel file for convenient review.

3.1.10.2. MSI

MSI is evaluated using MSIsensor-pro[89], which compares microsatellite lengths between each tumor sample and a panel-specific baseline reference.

Baseline construction (this process is performed once and not during per-sample analysis): An MSI baseline was generated for the TSO500 panel. Because matched normal samples were not available, we used 66 microsatellite stable (MSS) tumor samples identified with the TSO500 Local App, based on evidence that MSS tumors exhibit profiles comparable to normal tissue.[131] The baseline was constructed as follows:

1) A list of microsatellite sites from the GRCh38 reference genome was generated using the "scan" module of MSIsensor-pro, producing a table of genomic coordinates, repeat length and times, and flanking bases.
2) Homopolymer repeats 10–20 bp in length were selected using a custom R script, mirroring the length range used by the FoundationOne CDx panel[15].
3) The MSI baseline was built using the "baseline" module of MSIsensor-pro with the selected microsatellite sites and the DNA processed reads from the MSS cohort (minimum required: 20 samples). Only microsatellite loci with sufficient coverage (≥100 reads) were included. For each locus, an instability threshold (probability of deletion) was estimated from the reference cohort[89].

Sample analysis: After DNA processing, aligned reads (BAM) and the MSI baseline are used to quantify polymerase slippage events at microsatellite loci. Only loci with adequate coverage in the sample (≥100 reads by default) are assessed. MSIsensor-pro classifies each locus as unstable if the calculated probability exceeds the predefined baseline threshold. Finally, an MSI score per sample is calculated as the percentage of unstable loci among all evaluated microsatellites[89].

Because MSI is a hallmark of MMR deficiency[15], ClinBioNGS also includes a summary of small variants detected in MMR-related genes. These genes are defined using curated gene sets from the Molecular Signatures Database (MSigDB) collections[120] (**Supplementary Table 7**), providing complementary information to support MSI interpretation.

Both the MSI score and the list of MMR gene mutations are included in the final interactive HTML report.

3.1.11. Processing of final results

3.1.11.1. Generation of a variant registry

Upon completion of the analysis, ClinBioNGS generates two types of SQLite databases using the DBI R package[124] to systematically store and organize results.

First, a run-specific database is created. This includes a separate table for each type of alteration and compiles results from all samples within the run.

Second, a global variant registry is built by aggregating selected information from the run-specific databases, including:

- Sample metadata.
- DNA and RNA QC metrics.
- Small variants that passed the calling step (flagged as "PASS").

- Gene- and arm-level CNAs classified as non-neutral (i.e., AMPs and DELs).
- High-confidence RNA alterations (flagged as “OK”).
- TMB and MSI scores.

The registry also calculates the frequency of each variant across all analyzed samples, to track recurrence.

The file path of each run-specific database is recorded within the global registry. When a new run is processed, ClinBioNGS automatically retrieves existing paths, integrates newly detected variants with previously stored ones, recalculates variant frequencies, and updates the registry accordingly.

3.1.11.2. Generation of a comprehensive report of results

All processed tables and visual outputs are compiled into a self-contained, interactive HTML report using the flexdashboard R package[126]. An example report is available in the ClinBioNGS GitHub repository.

The report provides a user-friendly interface for exploring the results and is organized into multiple sections:

- Main results: A summary overview highlighting key findings, including Tier I–III variants and biomarker scores (TMB and MSI).
- Alteration-specific sections: For each type of variant (small variants, CNAs, gene fusions, splice variants), the report provides:
    - An overview with top-tier variants, variant statistics (e.g., counts in each flag and tier categories), and associated clinical evidence.
    - Visualization subsections with variant-specific plots.
    - Interactive tables with filtering and export options for detailed exploration.
- Biomarkers section: Includes dedicated subsections for TMB and MSI results:
    - TMB subsection: Displays the calculated TMB score along with the TMB trace table, listing all evaluated small variants.
    - MSI subsection: Shows MSI metrics and a table of small variants in MMR-related genes to support MSI interpretation.
- Sample QC section: Summarizes patient and sample metadata (e.g., tumor type, TP, sex, age), as well as global DNA and RNA QC metrics. Each sample is assigned a color-coded QC status. Additional subsections include coverage-specific tables and plots.

The HTML report is designed for seamless navigation, interpretation, and sharing of results, making it a key output for both clinical and research applications.

3.1.12. Installation, configuration, and structure of the pipeline

3.1.12.1 Installation

To run the pipeline, only Nextflow and Apptainer must be installed on the system. Please refer to the official documentation for instructions on installing these dependencies.

To install the pipeline, clone the public GitHub repository and make the executable scripts available:

```
nextflow clone raulmarinm/ClinBioNGS
cd ClinBioNGS
chmod +x bin/*
```

After cloning the repository, two setup modules are available to automatically download the required container images (**Supplementary Table 1**) and general resource files (**Supplementary Table 2**):

- To download Apptainer images:

```
nextflow run main.nf --prepareImages --runName setup
```

- To download general resource files only:

```
nextflow run main.nf --resourcesOnly --runName setup
rm -r work # optional: delete intermediate files to save space
```

This step prepares general resources (approximately 200 GB). Panel-specific resources (e.g., manifest files) can be automatically generated as needed during execution. To avoid delays during the first full analysis, it is advisable to first execute the pipeline with *--resourcesOnly* to pre-generate any panel-specific resources. Then, run the full analysis.

3.1.12.2. Configuration

The configuration of ClinBioNGS is modular and organized across multiple files:

- *nextflow.config*: It defines global defaults and Nextflow profiles.
- *base.config*: It specifies computational resource allocations.
- *modules.config*: It sets process-specific options.

Additionally, profile-based configurations (via *-profile*) allow customization for specific computational environments (e.g., “sge”, “slurm”) and NGS panels (e.g., “tso500”, “opa”, “oca”). For unsupported panels, users can define a *custom.config* file called by “custom” profile.

To initiate a run, users must define the following parameters:

- *--projectDir*: Path where output files will be saved.
- *--dataDir*: Directory to store processed data (e.g., FASTQ, BAM, or VCF).
- *--runName*: A unique identifier for the analysis.

- *--startingDataDir*: Directory containing the input sequencing data. Supported formats include:
    - uBAM or FASTQ files: the directory may contain files named *<sample>_<DNA/RNA>*.bam* or *<sample>_<DNA/RNA>*.fastq**. Symbolic links are supported. Set *--startingDataType* to "FASTQ" (default) or "BAM".
    - Illumina BCL directory (e.g., TSO500): contains raw BCL files and a *SampleSheet.csv*. Set *--startingDataType* to "BCL".
    - Ion Torrent results directory (e.g., OPA, OCA): should contain **.tar.xz* files under *Final_Results_Files/*, including uBAMs (**rawlib.basecaller.bam*) and auxiliary files. Set *--startingDataType* to "BAM" and add *--prepareIontorrentBam* if pre-processing is needed (this is pre-configured in panel profiles).
- *--sampleSheet*: A CSV file describing sample identifiers. Multiple options are supported:
    - Manual input: Users can prepare a sample sheet manually (examples provided in the GitHub repo). If it is saved at the default location *./resources/sampleSheets/SampleSheet_<runName>.csv*, it is automatically detected.
    - Illumina BCL directory: A standard *SampleSheet.csv* is typically included and can be used directly or placed in the default directory.
    - Ion Torrent (OPA/OCA): The sample sheet can be automatically generated from *--startingDataDir* using the *Info.csv* file. In this case, specify *--prepareIontorrentSamplesheet* (this is pre-configured in panel profiles).
- For custom panels, specify a custom *--manifestDir* to define the location of panel-specific manifest files. A recommended structure is *./resources/manifests/<seqPanel>/*, containing the required *--dnaManifest* and *--rnaManifest* files. Update pipeline parameters accordingly.

The pipeline uses a tag-based system to allocate computing resources (e.g., CPU, memory, execution time), with tags including "min", "low", "med", "high", and "extra". If resource limits are exceeded, the pipeline automatically retries the task with increased allocation.

In summary, ClinBioNGS provides a modular, controlled setup process with flexible profile options and user-defined metadata, enabling easy deployment in clinical and research settings.

Example: TSO500 analysis on a SLURM cluster

```
nextflow run main.nf -profile slurm,tso500 \
  --runName TSO500_RUN \
  --projectDir /mnt/projects/ClinBioNGS/output \
  --dataDir /mnt/projects/ClinBioNGS/data \
  --startingDataDir /mnt/illumina_runs/TSO500_Run/BclDirectory
```

Example: Custom panel analysis on an SGE cluster

```
nextflow run main.nf -profile sge,custom \
  --runName customPanel_RUN \
  --projectDir /mnt/projects/ClinBioNGS/output \
  --dataDir /mnt/projects/ClinBioNGS/data \
  --startingDataDir /mnt/data/custom_samples \
  --sampleSheet ./resources/sampleSheets/SampleSheet_customPanel_RUN.csv
```

3.1.12.3. Structure

Pipeline source

The ClinBioNGS pipeline is hosted in a public GitHub repository and is organized to be fully compatible with Nextflow. Its directory structure is modular and clearly separated into configuration files, executable scripts, and predefined resource folders. A summary of the pipeline's source structure is presented in **Table 17**.

Table 17. Overview of the ClinBioNGS source directory structure.
This table summarizes the main files and folders included in the repository, along with a brief description of their contents.

Files/Folders	Description
main.nf	Main Nextflow script that defines the overall workflow execution logic.
nextflow.config	Global configuration file that defines default parameters and loads specific configuration files.
bin/	Contains external scripts (e.g., R) that are executed as part of the workflow processes.
conf/	Configuration directory with global, module-specific, and panel-specific settings.
docker/	Custom Dockerfiles for selected tools (e.g., Octopus, Pisces, R).
modules/	Nextflow scripts organized into: • *process/*: individual pipeline processes. • *subworkflow/*: pipeline subworkflows.
resources/	Directory containing user-defined and predefined metadata or resource files, structured as follows: • User-defined metadata files: *SampleInfo.csv*, *TumorNames.csv*, and *WhitelistGenes.csv* • *annotation/*: variant annotation files (e.g., GENIE) • *cna/*: CNA-related files such as panel-specific baselines • *fusion/*: fusion-related files (e.g., VCF headers, curated whitelist) • *manifests/*: panel-specific manifest files • *msi/*: MSI-related files including the required baseline • *sampleSheets/*: run-specific sample sheets • *smallVariant/*: predefined files for small variant analysis (e.g., hotspot, recurrent, blacklist) • *splicing/*: splicing-related files (e.g., curated whitelist, VCF header)

Pipeline outputs

When the pipeline is executed, all output files are organized into structured directories based on their type and purpose. The key output locations and contents are described below:

- Temporary files: Each process generates files stored within its dedicated working directory under the automatically created *work/* folder (in the current working directory). Once the process is completed, only essential results are retained. The *work/* folder can be deleted after the run to conserve disk space.

- Resources files: Any newly generated resources (e.g., panel-specific target files, annotation data) are stored in the *resources/* directory within the main ClinBioNGS folder.
- Global variant registry: The global SQLite database that tracks all detected variants and metadata is stored at the root of the project directory (*<projectDir>/*).
- Run-specific output files: Each execution of the pipeline generates a structured output directory under the specified run name (*--runName*), typically inside the *<projectDir>/*. These outputs include:
    - *Analysis/*: Located at *<projectDir>/<runName>/*, this folder contains the main analytical results.
        - A subdirectory is created for each sample.
        - Within each sample folder, subfolders are organized by pipeline module (e.g., *01_FASTQ_PROCESSING/*, *02_FASTQ_QC/*, *03_ALIGNMENT/*).
        - Each subfolder includes outputs such as QC tables, variant files, and associated visualizations.
    - *Data/*: Located at *<dataDir>/<runName>/*, this directory contains processed data files in standard formats.
        - Subdirectories are created per sample.
        - Each sample folder contains separate folders by data type (e.g., *BAM/*, *FASTQ/*, *VCF/*), containing the final processed files.
    - *Logs/*: Located at *<projectDir>/<runName>/*, this folder stores all Nextflow-related log files.
        - Organized by subworkflow and then by DNA or RNA processes.
        - Each process-specific folder includes all log files (*.err*, *.log*, *.out*, *.run*, *.sh*) for every sample.
    - *Results/*: Located at *<projectDir>/<runName>/*, this folder contains the final, user-friendly output files.
        - MultiQC HTML reports summarizing DNA and RNA QC metrics.
        - Interactive HTML reports per sample.
        - The run-specific SQLite database.
        - Includes a folder for each sample, with Excel files summarizing all QC metrics and detected variants by alteration type (SNVs/Indels, CNAs, fusions, splice variants).

3.2. Cross-panel small variant validation on reference datasets

3.2.1. Dataset description

To assess the performance of ClinBioNGS in detecting small variants across multiple NGS panel technologies, publicly available reference data from the SEQC2 Consortium was analyzed[105,111]. This dataset includes multi-panel sequencing of a genomic reference sample ("Sample A") engineered to contain both known positive (KP) variants—introduced at various VAFs—and known negative (KN) positions, which are high-confidence wild-type sites within coding regions.

Each panel was sequenced independently by a different laboratory, with four technical replicates per panel. Of the eight initially considered panels, two were excluded due to either unclear UMI processing requirements or their non-commercial nature. The remaining six commercial panels included in the analysis were (**Supplementary Table 8**):

- Agilent Custom Comprehensive Cancer Panel v2 (AGL).
- Burning Rock DX OncoScreen Plus (BRP).
- Integrated DNA Technologies xGen Pan-Cancer Panel (IDT).
- iGeneTech AIOnco-seq (IGT).
- Illumina TruSight Tumor 170 (ILM).
- Thermo Fisher Oncomine Comprehensive Assay v3 (TFS).

The reference dataset included:

- A BED file (hg38) defining CTR regions, covering validated KP and KN sites.
- A VCF file listing over 40,000 KP variants for Sample A (originally in hg19, lifted to hg38).
- A BED file (hg38) containing over 10 million KN positions for Sample A.

Panel-specific resources consisted of:

- Raw sequencing files (FASTQ or uBAM) downloaded from NCBI BioProject PRJNA677997[132] (via AWS S3).
- Panel-specific variant call files (VCFs), lifted to hg38 coordinates.
- BED files defining panel target regions (hg19).

3.2.2. ClinBioNGS analysis

For each panel, custom configuration files were created to ensure accurate execution of ClinBioNGS. These configurations are available in the pipeline's GitHub repository. Key modifications included:

- Only the small variant analysis module was enabled. CNAs, TMB, MSI, and RNA-based analyses were disabled.

- Input BED files defining the panel regions were converted to hg38 coordinates using the *--liftoverManifest* option.
- For AGL panel, UMI sequences were extracted from an additional FASTQ file (R3) using *--fastqUmiTransfer*.
- For ILM, FASTQ files were merged across sequencing lanes via *--fastqMergeLanes*.
- For BRP and IGT, trimming and filtering parameters were aligned with those described in the SEQC2 supplementary methods[105].
- For TFS (Ion Torrent data), platform-specific options were applied, including handling of single-end uBAM files and the incorporation of blacklisted regions. Panel BED files were also lifted to hg38 (*--liftoverManifest*, *--liftoverVariantBlacklistBed*).

Each pipeline run required:

- A sample sheet listing the sample identifiers in the format "<sample>_DNA". These files are found in the pipeline's repository.
- A raw data directory (*--startingDataDir*) containing either FASTQ or uBAM files named accordingly (*<sample>_DNA*.fastq.gz* or *.bam*). Symbolic links were used to point to the original files for consistency and ease of access.

3.2.3. Output and performance evaluation

Following small variant detection, the variant calls produced by ClinBioNGS and the original commercial pipelines were compared against the SEQC2 reference dataset to assess performance.

The evaluation process involved the following steps:

- Extraction of KP variants from the reference VCF file.
- Compilation of detected variants and associated metrics from each analysis.
- Annotation of each variant with CTR and KN regions.

To ensure consistency and reliability, the following filters were applied:

- Only target regions overlapping CTRs and outside blacklisted regions were considered.
- KP variants had to meet panel-specific VAF thresholds:
    - AGL ≥1 %.
    - BRP ≥ 1%.
    - IDT ≥ 2%.
    - IGT ≥ 1%.
    - ILM ≥ 2.6%.
    - TFS ≥ 2.5%.

- Detected variants were required to
    - Have a “PASS” status (or no filter) in the VCF.
    - Meet minimum read support thresholds:
        - AD_ALT ≥ 5.
        - DP ≥ 10.
        - VAF ≥ panel-specific threshold.

The following metrics were computed:

- True positives (TPs):
    - SNVs: exact positional and allelic match with KP set.
    - InDels: matched by overlapping position.
- FNs: KP variants not detected by the pipeline.
- FPs: Variants identified within KN regions.

Based on these, the key performance indicators were calculated:

- Precision = TP / (TP + FP).
- Recall = TP / (TP + FN).
- F1 Score = 2 × (Precision × Recall) / (Precision + Recall).

These metrics allowed for a robust comparison of variant calling performance across multiple NGS panels and pipelines.

3.3. Cross-panel benchmarking in real-world clinical cohorts.

3.3.1. Dataset description

To evaluate the performance of ClinBioNGS in a real-world clinical setting, over 2,000 tumor samples sequenced using three commercial pan-cancer NGS panels (**Supplementary Table 8**) were analyzed. The data was provided by participating clinical institutions, each contributing raw sequencing files, manifest files, and the corresponding results from their established commercial analysis pipelines:

- *Institut Català d’Oncologia* and *Hospital Universitari de Bellvitge* contributed data for the Illumina TSO500 and Thermo Fisher OPA panels.
- *Hospital Clínic de Barcelona* provided data for the Thermo Fisher OCA panel.

All participants provided written informed consent for NGS testing as part of their clinical evaluation. The project was approved by the Ethical Committee of the participating hospitals and conducted in accordance with the Declaration of Helsinki.

For the TSO500 panel, raw data in BCL format, panel manifest files (in hg19 coordinates), and results from the TSO500 Local App (v2.2.0.12) were collected.

For the OPA and OCA panels (Ion Torrent platform), uBAM files and panel-specific resources were obtained from the Torrent Suite software (v6.6.2.1), using the following assay versions:

- OPA: Oncomine™ Precision GX5 DNA and Fusions v3.2.0.
- OCA: Oncomine™ Comprehensive v3 GX5 DNA and Fusions v5.0.2.

This diverse and clinically annotated dataset enabled robust benchmarking of the pipeline across different sequencing technologies and analysis environments.

3.3.2. ClinBioNGS analysis

Panel-specific configuration files were developed to ensure accurate analysis with ClinBioNGS. These configurations are publicly available in the pipeline's GitHub repository. Several adjustments were made to adapt the pipeline beyond its default settings:

- TSO500 (Illumina):
    - The BCL folder was provided as the input directory, from which FASTQ files and the sample sheet were automatically generated (*--startingDataDir <bclDir>*, *--startingDataType BCL*).
    - Manifest files were converted to BED files (*--manifestToBed*) and lifted over from hg19 to hg38 (*--liftoverManifest*).
    - UMIs were extracted from DNA samples during the BCL Convert step (*--umiDna*), with delimiter adjustments applied as needed (*--fastqChangeUmiSep*).
- OPA and OCA (Ion Torrent):
    - The Torrent Suite results directory was used as the input (*--startingDataDir*), from which all relevant resources were extracted, including manifest files (*--prepareIontorrentManifest*), sample sheet (*--prepareIontorrentSamplesheet*), hotspot BED (*--prepareIontorrentVariantHotspots*), and uBAM files (*--prepareIontorrentBam*, *--startingDataType BAM*).
    - Manifest files (BED) were lifted over to hg38 coordinates (*--liftoverManifest*).
    - As these panels rely on single-end (*--singleEnd*), amplicon-based (*--amplicon*) sequencing from Ion Torrent platform (*--seqPlatform IonTorrent*), adequate pre-processing steps were applied.
    - Alignments were performed using TMAP parameters from the commercial pipeline.
    - For OPA, which includes UMIs in DNA library (*--umiDna*), deduplication was performed using the extracted UMI information. This step was not applied to OCA, as it does not include them.

- A blacklist BED file, adapted from the SEQC2 Thermo Fisher panel, was used to flag small variants in problematic regions.
- TMB and MSI analyses were not performed on Ion Torrent data due to platform-specific limitations.

3.3.3. Output and comparative analysis

After pipeline execution, results from both ClinBioNGS and the commercial pipelines were collected for comparative analysis. Only samples that passed the QC criteria—generated by ClinBioNGS's QC module (**Table 6**)—were included. QC criteria are summarized in **Supplementary Table 9**.

For cases sequenced in multiple runs, the sample with the highest total read count was selected. Final cohort sizes were as follows:

- TSO500: 755 samples (N_{DNA} = 655, N_{RNA} = 687).
- OPA: 674 samples (N_{DNA} = 624, N_{RNA} = 588).
- OCA: 595 samples (N_{DNA} = 538, N_{RNA} = 508).

To harmonize the commercial pipeline results for comparison:

- Gene symbols were updated to match HGNC nomenclature.
- Coordinates were lifted over from hg19 to hg38.
- Duplicate variants per sample were removed.
- Hotspot and oncogenic variants were annotated.
- CIViC clinical evidence was linked where applicable.

Variants included in the comparison:

- ClinBioNGS
  - Small variants with primary "PASS" flag (i.e., no calling-related filter; **Table 9**).
  - Non-neutral CNAs in genes covered by ≥4 bins in the panel-specific CNA baseline.
  - Fusions and splice variants with primary "PASS" flag (i.e., no "LowSupport" flag; **Table 14** and **Table 16**). Splice variants with ≥1000 supporting reads in OCA panel.
- Commercial pipelines:
  - TSO500:
    - Reported small variants in *CombinedVariantOutput.tsv* (final results).
    - CNAs with ALT field not equal to ".".
    - Non-intragenic fusions labeled "KEEPFUSION".
    - Splice variants in target genes from the final results.
  - OPA/OCA:
    - Variants marked as "PRESENT".

Cancer-related subset (applied to the filtered variants above):

- Small variants annotated as oncogenic, hotspot, or with CIViC evidence.
- CNAs with ≥0.5% frequency in GENIE or annotated in CIViC.
- Gene fusions present in GENIE, MitelmanDB, or CIViC.
- Splice variants annotated in CIViC.

Comparison methodology:

- Small variants (SNVs/InDels): matched by genomic position or AA change.
- CNAs: matched by gene and AMP/DEL status.
- Fusions: matched by gene partners.
- Splice variants: matched by splice sites (and also by exon for TSO500).

Variant-level concordance between pipelines was calculated and summarized using alluvial plots, illustrating the overlap in detected variants.

Moreover, Pearson coefficients and linear regression lines were computed and plotted to assess the correlation of gene-level copy-number ratios and estimated CN values between pipelines. TMB-high (≥10 mut/Mb) and MSI-high (≥20% unstable loci) classifications were also compared in TSO500 samples.

4. RESULTS

4.1. ClinBioNGS enables end-to-end analysis of somatic NGS cancer panels

4.1.1. Workflow design enables comprehensive analysis

This thesis presents ClinBioNGS, an open-source, comprehensive bioinformatics pipeline designed for the analysis of somatic NGS cancer panels. It provides a fully automated, portable, and end-to-end solution—covering raw data pre-processing through to variant detection, annotation, prioritization, and reporting. The pipeline leverages state-of-the-art open-source tools and curated external resources (**Supplementary Table 1** and **Supplementary Table 2**). A schematic representation of the ClinBioNGS workflow is shown in **Figure 13**.

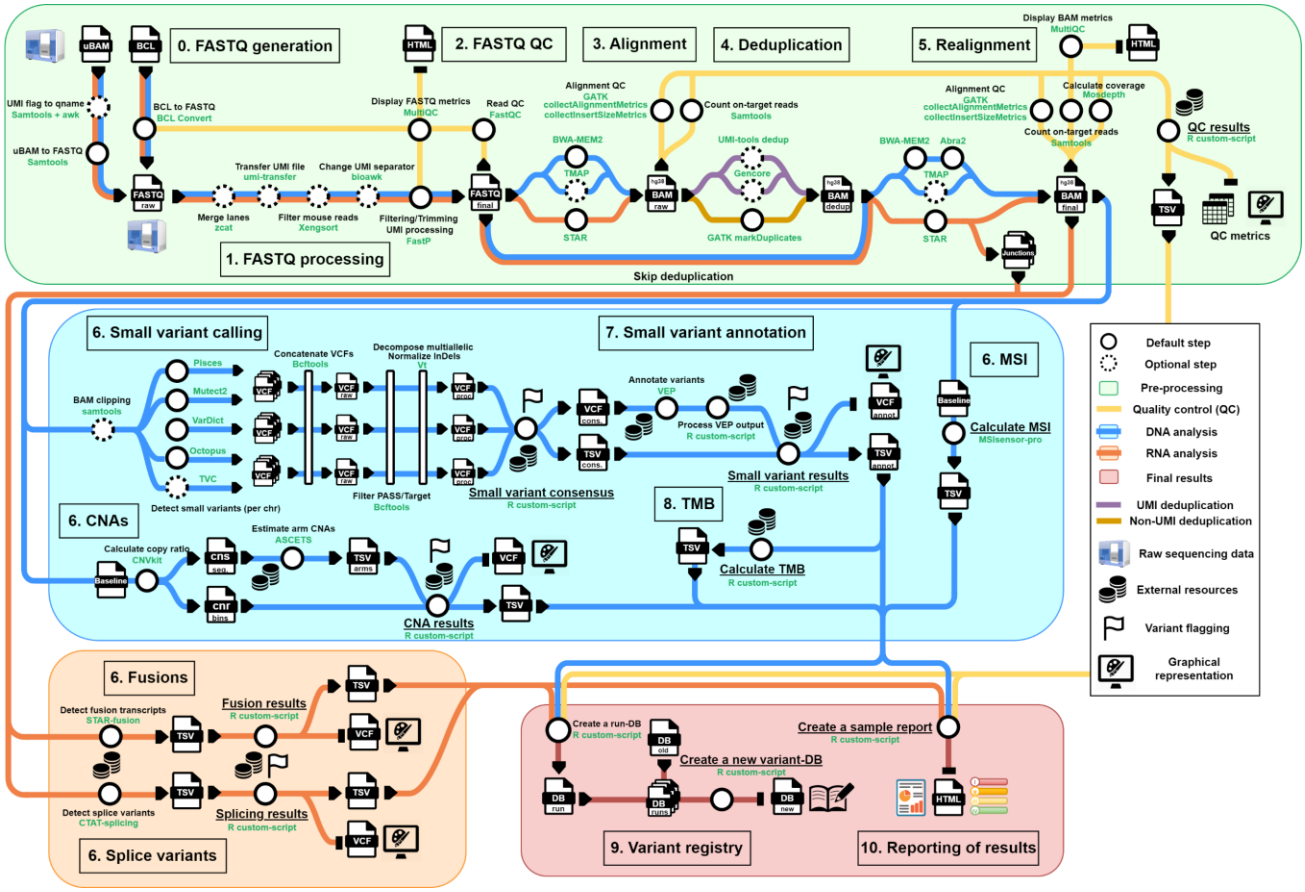


Figure 13. Overview of the ClinBioNGS workflow.
The top section (green box) illustrates the pre-processing of raw sequencing data (DNA and RNA), including FASTQ handling, alignment, deduplication, and QC. The middle section (blue box) shows the DNA analysis module, which detects small variants (SNVs/InDels), CNAs, MSI, and TMB. The bottom section (orange box) outlines the RNA analysis module, which includes the detection of gene fusions and splice variants. Outputs from QC, DNA, and RNA analyses are consolidated into a variant registry and a self-contained, interactive HTML report for each sample.

The pipeline is highly flexible and supports DNA and RNA pre-processing across a broad range of experimental settings, including variations in:

- Panel design: hybrid-capture or amplicon-based;
- Library preparation: paired-end or single-end; with or without UMIs;
- Sequencing platforms: Illumina or Ion Torrent technology;
- Sample types: tumor tissue, cell lines, or PDXs;
- Input formats: FASTQ, BCL, or uBAM files.

ClinBioNGS performs a comprehensive quality assessment by calculating global QC metrics and evaluating gene- and exon-level coverage. The pipeline supports full analysis of distinct somatic alterations—including small variants, CNAs, gene fusions, splice variants, and, when panel design allows, TMB and MSI. All findings are stored in a variant registry, enabling longitudinal tracking and knowledge reuse. Upon completion, the results for each sample are integrated into a self-contained, interactive HTML report optimized for clinical review.

4.1.2. Visualizations enhance interpretability of results

In addition to comprehensive QC metrics and variant outputs, ClinBioNGS generates diverse informative visualizations. The following sections present representative real-case examples, derived from Illumina TSO500 panel data, to illustrate how these plots help contextualize genomic alterations and facilitate clinical interpretability.

4.1.2.1. Coverage visualizations enable sequencing quality assessment

ClinBioNGS produces a suite of informative coverage plots at multiple levels of resolution. The following examples correspond to a non-small cell lung cancer (NSCLC) sample.

A genome-wide overview of gene-level coverage is provided for all targeted regions (**Figure 14**). This plot enables rapid evaluation of overall panel performance by displaying the mean coverage for each target gene across all chromosomes. Chromosome-specific plots (**Figure 15**) are provided to localize target genes within each chromosome. These representations allow users to assess capture uniformity across chromosomes. Poorly covered genes or those included in a user-defined whitelist are also highlighted to facilitate their evaluation.

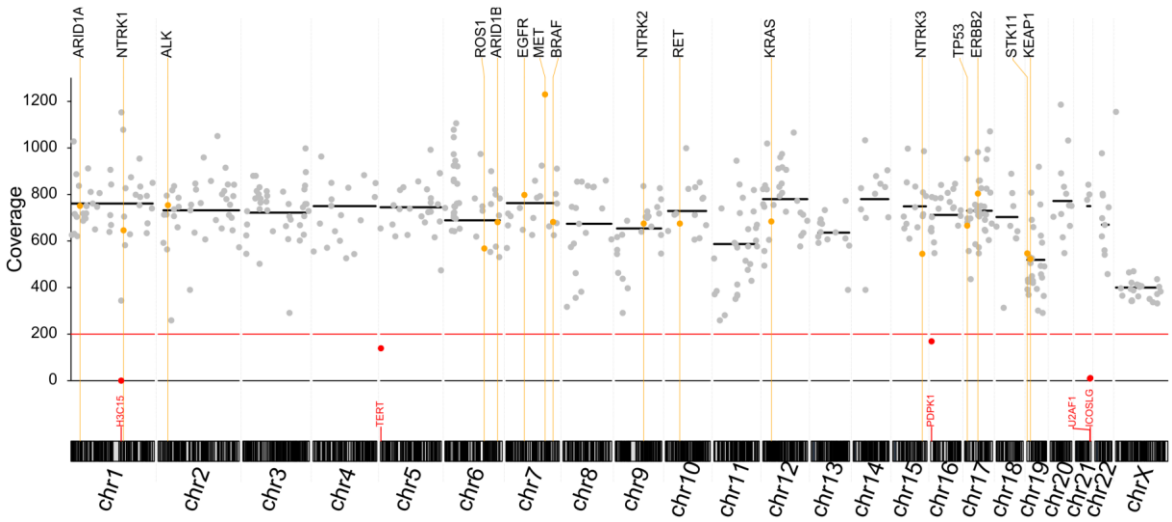


Figure 14. Genome-wide gene coverage visualization (TSO500 DNA data in NSCLC).
Each point represents the average coverage of an individual target gene. A dark grey line optionally shows the mean coverage per chromosome. Genes in a user-defined whitelist are highlighted in orange and labeled above the coverage track. Genes falling below a user-defined minimum coverage threshold (indicated by a red horizontal line) are labeled in red along the bottom of the plot.

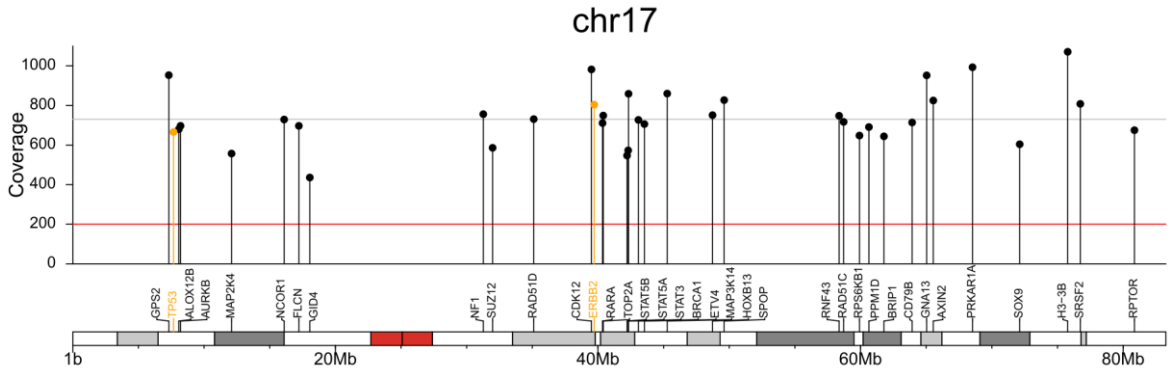


Figure 15. Chromosome-specific gene coverage plot for chr17 (TSO500 DNA data in NSCLC).
Each point represents the mean coverage of a targeted gene on the selected chromosome. A light grey line shows the average coverage of all targeted regions in that chromosome. Gene symbols are displayed along the bottom. Genes in the whitelist are highlighted in orange, and those below the coverage threshold are annotated in red.

For high-resolution assessment, ClinBioNGS generates single-gene coverage plots, which are especially useful for evaluating intragenic coverage variability. By default, these plots are created only for genes in a user-defined whitelist, though full-panel plotting can be enabled for smaller panels. Each plot shows gene structure, coverage across the genomic region, and target capture regions from the manifest. **Figure 16** presents an example for the *ERBB2* gene using both DNA and RNA data from the same sample. In DNA data, coverage often extends to intronic and off-target flanking regions, while RNA coverage presents narrower peaks reflecting transcript structure and exon-exon junctions.

These multi-resolution coverage visualizations—spanning genome, chromosome, and gene levels—can facilitate robust evaluation of sequencing performance, data quality, and capture efficiency of target genes.

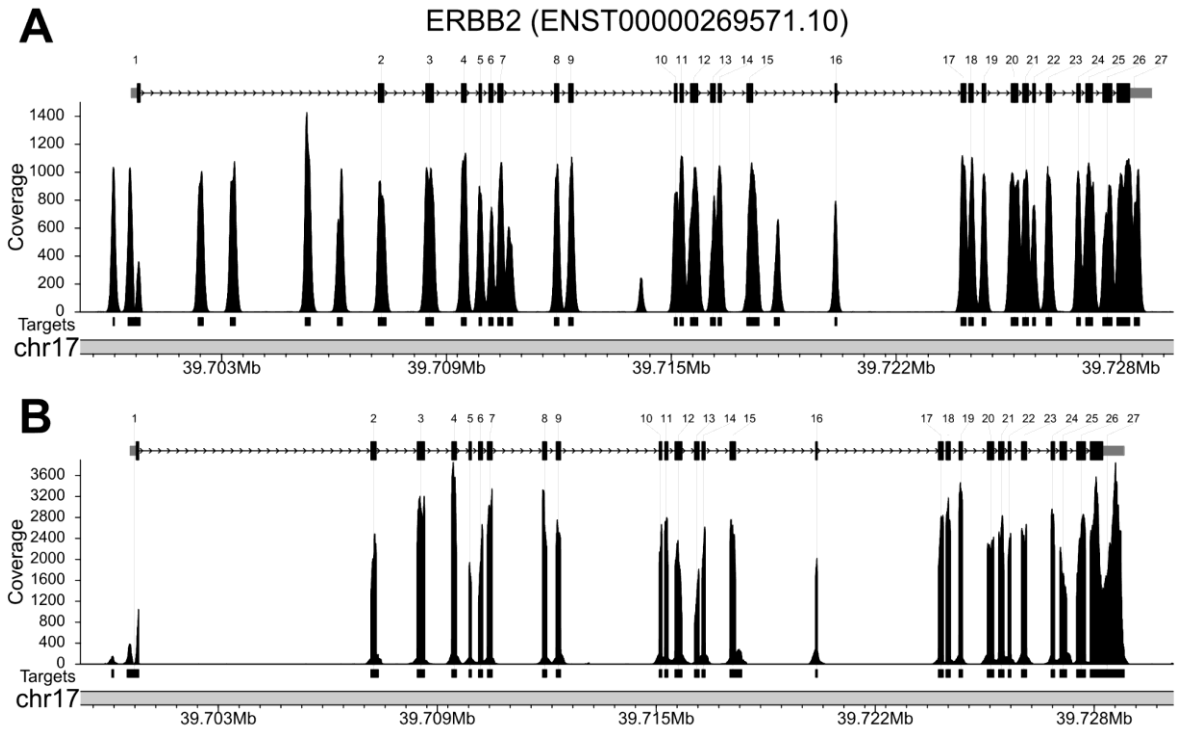


Figure 16. Single-gene coverage plots for *ERBB2* (TSO500 DNA and RNA data in NSCLC).
(**A**) DNA-based coverage. (**B**) RNA-based coverage. Gene structure (top) is based on the MANE SELECT transcript, with coding (black) and non-coding (grey) exons labeled. Per-base coverage across the gene is shown in the middle. Panel-defined target regions are shown below as black segments.

4.1.2.2. Gene-centric visualizations support contextual interpretation of small variants

ClinBioNGS produces small variant visualizations that position mutations within gene structures to aid interpretation. Each variant is provided with the following information: observed VAF, amino acid change, oncogenicity classification (ClinGen/CGC/VICC), clinical tier (AMP/ASCO/CAP), and occurrence in the GENIE cancer registry.

Representative examples are shown in **Figure 17**, illustrating a glioma case with four missense *TP53* variants (**Figure 17A**) and a NSCLC sample with a well-known *BRAF V600E* mutation (**Figure 17B**). These gene-centric maps help contextualize mutations within the full gene structure and mutational landscape:

- TSGs like *TP53* often show dispersed inactivating mutations clustered in functional regions.
- Oncogenes such as *BRAF* typically display hotspot activating mutations at specific loci.

Additionally, the inclusion of known oncogenic mutations from the GENIE registry enables users to assess the broader clinical and biological significance of observed variants.

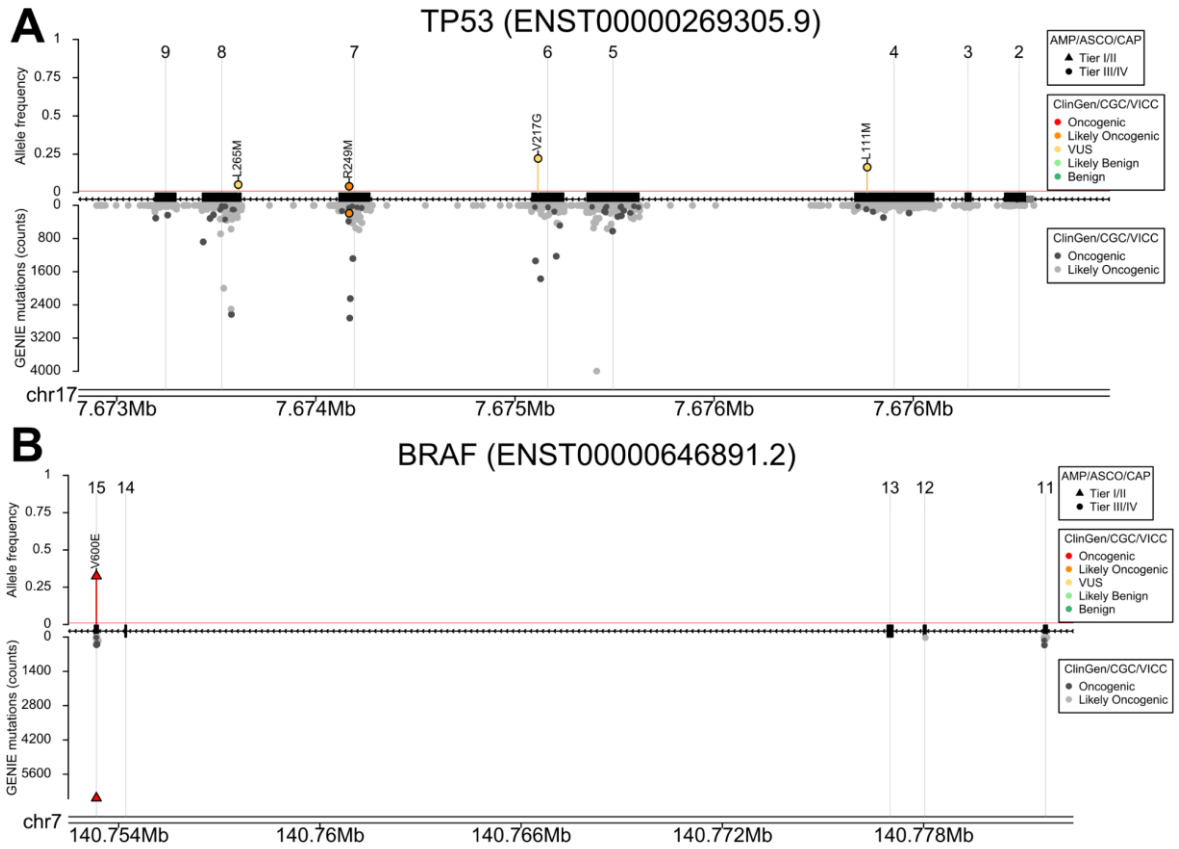


Figure 17. Visualization of small variants mapped to their gene locus (TSO500 DNA data).
(**A**) Glioma with four missense *TP53* variants and **(B)** NSCLC with *BRAF V600E* mutation. The central panel shows gene structure from the MANE SELECT transcript, with annotation of the covered exons. Detected variants are displayed in the upper section with their corresponding VAFs and AA changes, color-coded by predicted oncogenicity. AMP/ASCO/CAP classifications are indicated with symbols. A red line marks the minimum user-defined VAF threshold. Known oncogenic variants from the GENIE registry are shown below the gene structure, with the height reflecting cohort counts.

4.1.2.3. Multi-level CNA visualizations enhance comprehensive analysis

CNAs are visualized at multiple levels to support both global and fine-grained interpretation:

- **Figure 18** shows genome-wide CNA patterns, ideal for identifying broad events such as CNAs in short arms (i.e., "p") or long arms (i.e., "q") of chromosomes. The example features a uveal melanoma case, where canonical events in this tumor type—such as 3p/q, 6q, and 8p losses, and 8q gain[133]—are clearly visible.
- **Figure 19** presents a chromosome-specific view (chr8 in the same case), facilitating inspection of CNA status in specific target genes.
- **Figure 20** shows a high-resolution CNA profile for the *MET* gene, including copy ratios for individual bins and the observed read coverage, supporting detailed assessment of focal events and intra-gene variability. In this case we can observe that copy ratios are generally uniform along the genome.

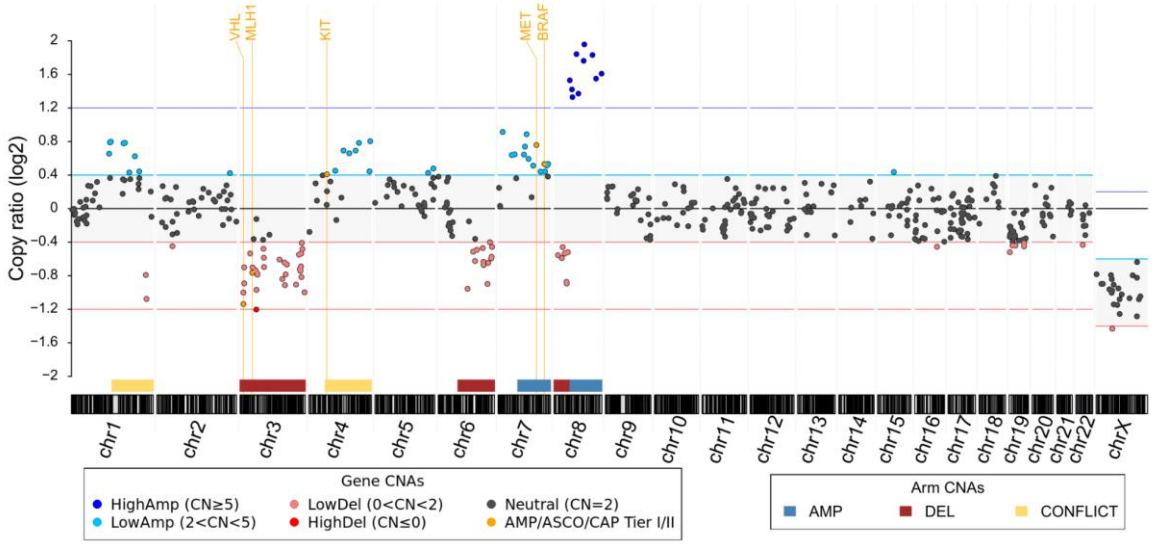


Figure 18. Genome-wide CNA visualization (TSO500 DNA data in uveal melanoma).
Each point represents a gene-level copy ratio, color-coded by CNA classification. Thresholds for CNA classification are shown as horizontal lines. Clinically relevant CNAs are labeled. Arm-level CNAs are displayed as colored segments along the x-axis.

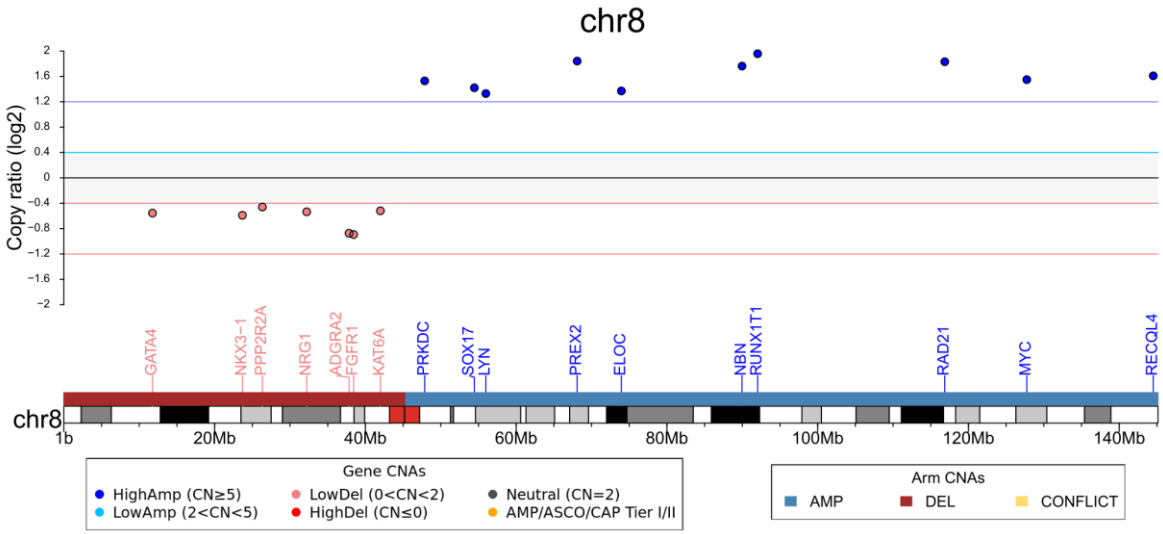


Figure 19. Chromosome-specific CNA results for chr8 (TSO500 DNA data in uveal melanoma).
Chromosome-level CNA plot for chr8. Each point corresponds to a gene's copy ratio, color-coded by classification. Gene symbols are displayed and colored by CNA status.

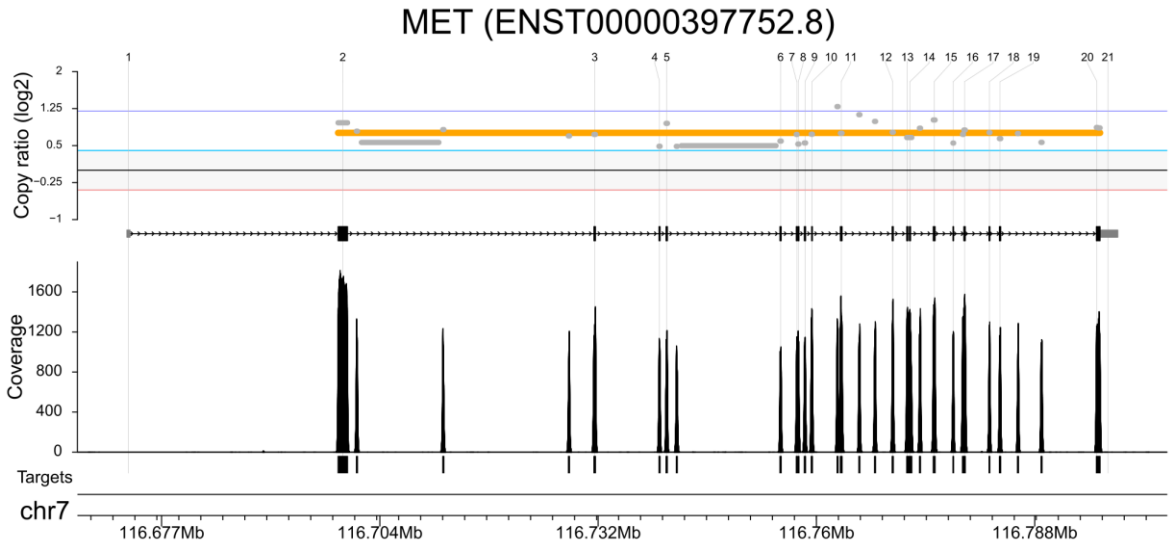


Figure 20. CNA profile of the *MET* gene (TSO500 DNA data in uveal melanoma).
The gene structure (middle panel) is annotated from the MANE SELECT transcript. Individual bin-level copy ratios are shown (top panel) with a weighted mean overlay. The lower panel shows per-base coverage with panel-defined target regions indicated in black.

4.1.2.4. RNA-based visualizations facilitate functional assessment of results

ClinBioNGS generates detailed visualizations for RNA-based alterations to support the interpretation of gene fusions and splicing events.

- **Figure 21** shows an example of the *EML4–ALK* fusion in a NSCLC sample. The plot displays predicted breakpoints and coverage profiles for both fusion partners. We can observe the selective expression in the 3′ region of *ALK*, which harbors the oncogenic kinase domain[114].

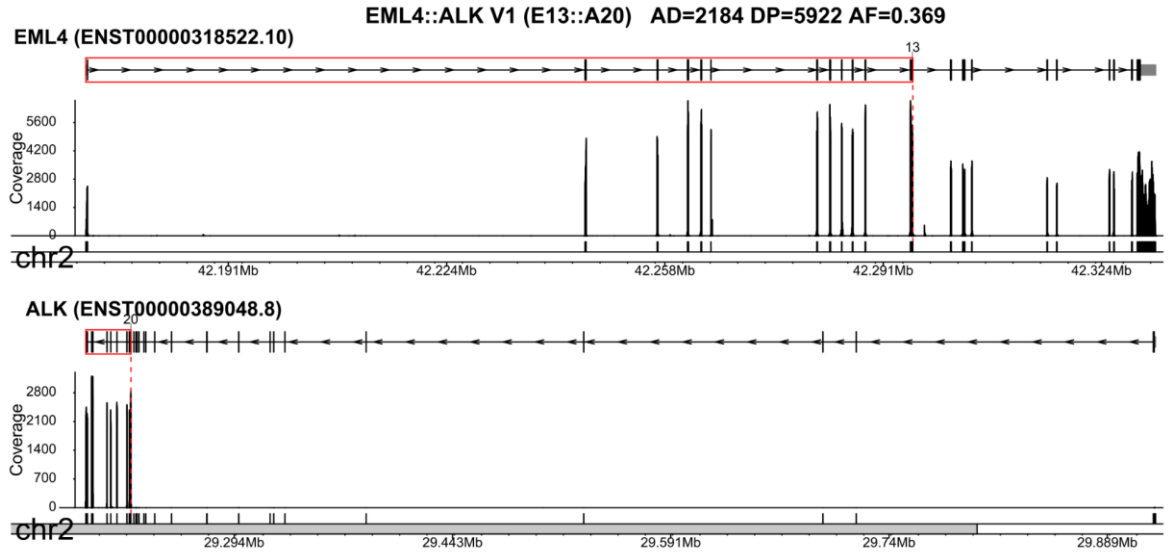


Figure 21. Visualization of *EML4–ALK* fusion (TSO500 RNA data in NSCLC).
Coverage profiles of both fusion partners are shown with gene structure based on MANE SELECT transcripts. The fusion breakpoint and fusion region are marked with a red dashed line. Key metrics and the variant name appear at the top. Target regions from the panel manifest are shown below the coverage track.

- **Figure 22** illustrates the *METex14* event in NSCLC, a known actionable alteration[134]. Splice variants are depicted with sashimi-style plots displaying splice junction reads and the observed coverage in the affected genomic locus. We can observe the drop in coverage over the skipped exon 14.

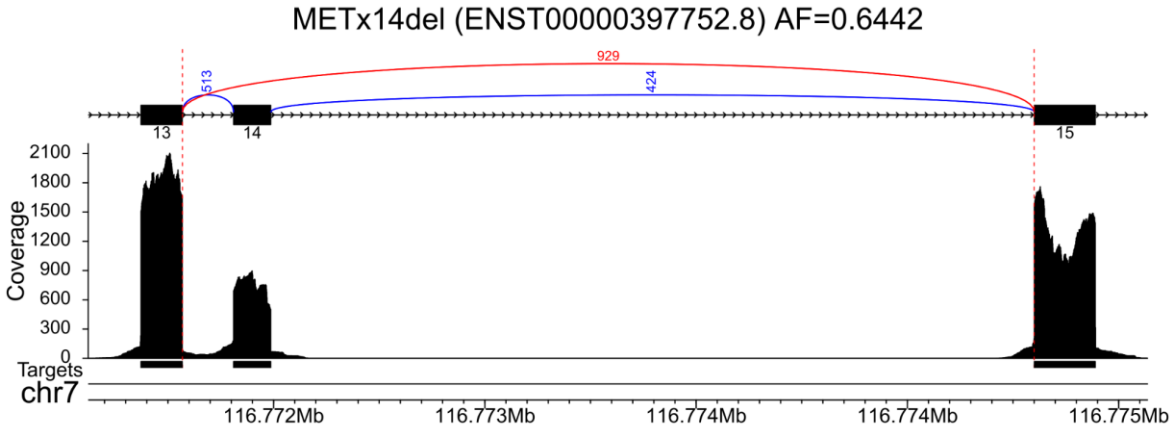


Figure 22. Visualization of *METex14* variant (TSO500 RNA data in NSCLC).
Exon structures, splice junctions, and supporting read counts are displayed. Blue arcs denote canonical junctions, while the red arc highlights the aberrant, cancer-associated splice event. The variant name and estimated VAF are noted at the top. The observed coverage across the gene locus and panel-defined target regions are also included below the transcript structure.

Together, these plots provide indirect evidence of gene expression and help contextualize the functional effect of RNA alterations detected by ClinBioNGS.

4.1.3. Interactive report supports exploration of results

ClinBioNGS consolidates prioritized findings, QC metrics, and visualizations in a self-contained HTML report. This interactive report serves as the main deliverable of the pipeline, offering a user-friendly and centralized interface for exploring complex genomic results on a per-sample basis.

The following subsections illustrate representative screenshots of the report's core features from the previous TSO500 cases. A composite HTML report is also available in the pipeline's repository.

4.1.3.1. Summary section highlights key results

The report opens with a summary section that highlights the most relevant findings at a glance (**Figure 23**). This overview is structured as a grid, where each column represents a specific variant type (e.g., SNV/Indel, CNA, fusion, splicing), and each row corresponds to the assigned clinical tier (Tier I–III). Additional biomarker results, such as TMB and MSI scores, are also displayed in a column, enabling quick assessment of genomic indicators associated with therapeutic response. Sample metadata—including run name, sample ID, tumor type, and panel—are displayed in the upper-left corner and remain visible during navigation.

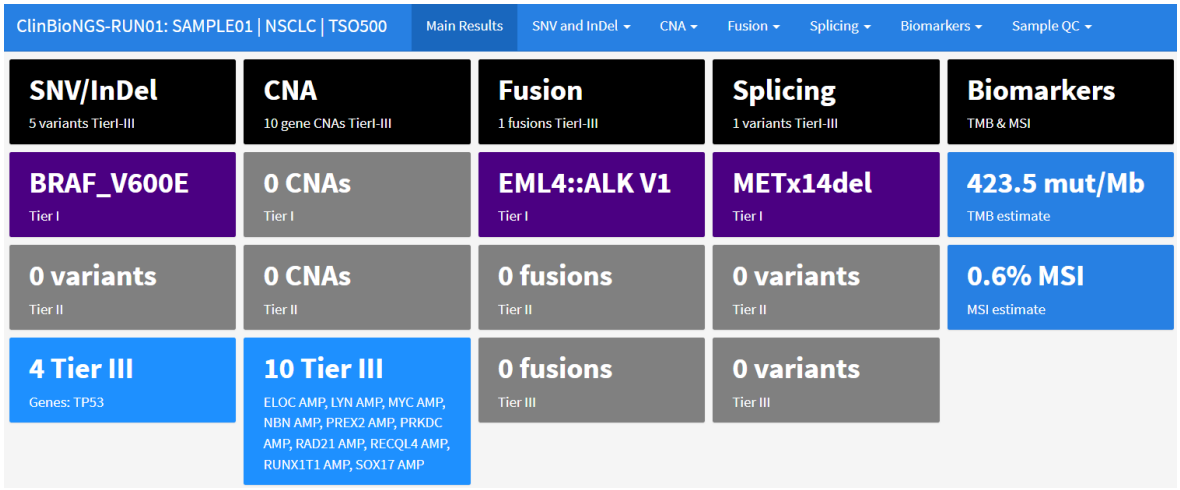


Figure 23. Summary section of the ClinBioNGS report.
Columns represent variant types, and rows correspond to variant tiers (Tier I–III). The final column summarizes TMB and MSI scores. Clickable boxes link to specific report sections. Sample metadata is displayed in a fixed panel at the top left, and navigation tabs provide access to all major sections.

Each colored box in the summary view is interactive, allowing users to click and access the corresponding visualizations or detailed tables. Additionally, navigation tabs at the top of the report enable direct access to each report section.

4.1.3.2. QC section supports sample assessment

The "Sample QC" section of the ClinBioNGS report provides a centralized view of key sample characteristics and sequencing quality metrics. It begins with an overview subsection that summarizes sample metadata (e.g., tumor type, tumor whitelist inclusion, TP, sex, age), alongside calculated DNA and RNA global QC metrics (**Figure 24**). Each sample is assigned a color-coded QC status for DNA and RNA to facilitate rapid assessment:

- Green indicates acceptable values across all metrics.
- Orange highlights potential issues (warning).
- Red indicates any failed QC metric.

These colors are determined based on user-defined thresholds configured within the pipeline and apply independently to DNA and RNA metrics. A summarized QC box at the top of the section reflects the overall QC status for each sample.

Additional tabs provide access to detailed coverage visualizations and tables. The coverage visualizations shown in **Figure 14** to **Figure 16** are fully integrated into specific subsections.

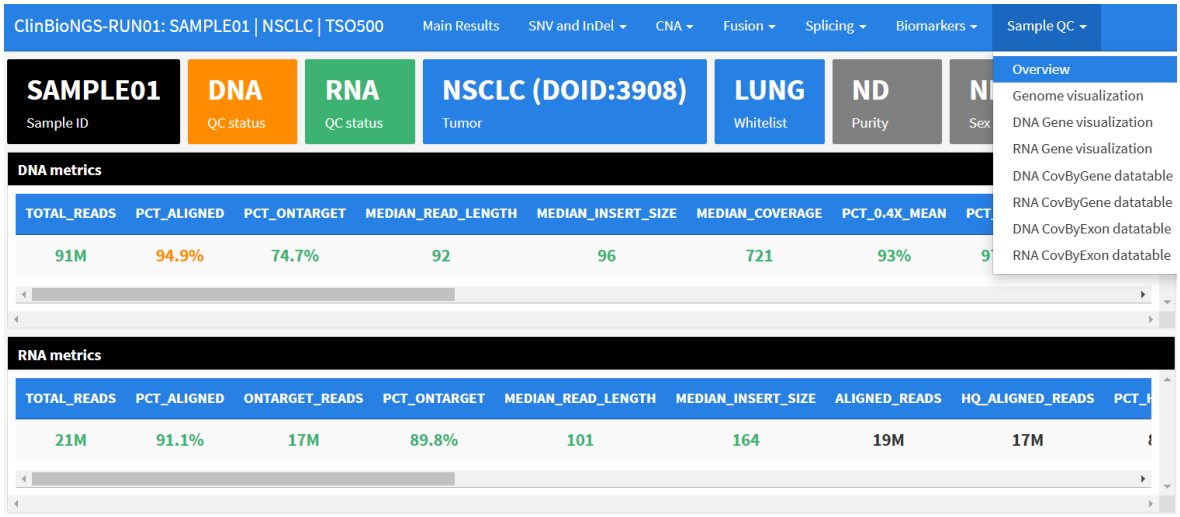


Figure 24. Overview of the Sample QC subsection in the ClinBioNGS report.
The top row displays collected sample metadata (e.g., sample ID, tumor type with DOID, tumor-specific whitelist inclusion, estimated TP, sex, and age). If a field is missing, "ND" (No Data) is displayed. Global QC metrics are shown in separate tables for DNA and RNA, with horizontally scrollable views. Metrics are colored based on configured thresholds: green (pass), orange (warning), red (fail). The most severe color across all metrics determines the final QC status shown at the top of the section. Tabs are available to access additional coverage-related content.

ClinBioNGS also generates interactive per-gene and per-exon coverage tables for both DNA and RNA data. **Figure 25** shows examples of these tables, which are equipped with powerful features to enhance usability:

- Column-based filtering (checkboxes, value entry, sliders for ranges).
- Full-text search bar.
- Column sorting and reordering.
- Row highlighting.
- Export of filtered tables to CSV or Excel.

Two examples are illustrated in **Figure 25**:

- **Figure 25A** shows a DNA per-gene coverage table filtered to display only genes included in a user-defined whitelist. The coverage for different loci categories (e.g., target regions, coding regions, exons) and coverage statistics are presented.
- **Figure 25B** shows a per-exon RNA coverage table for the *MET* gene, filtered to show exons 12–16 from the *METex14* case (**Figure 22**). The highlighted exon 14 row shows a visibly reduced mean coverage, supporting the splicing event identified in the sashimi plot.

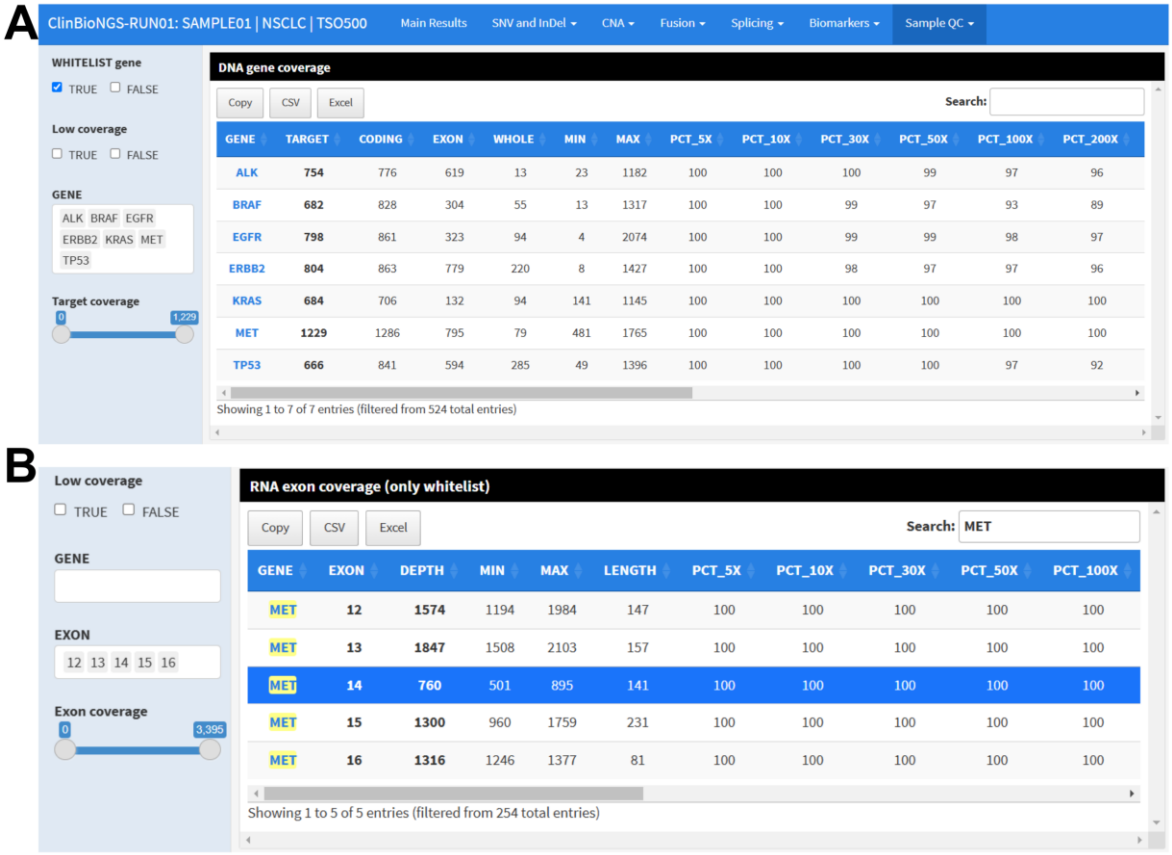


Figure 25. Interactive coverage tables in the ClinBioNGS report.
Users can apply filters via the left-hand panel or the top-right search bar, and export results in various formats. (**A**) DNA per-gene coverage table showing filtered whitelist genes and their mean coverage at various loci with coverage statistics. (**B**) RNA per-exon coverage table for the *MET* gene, highlighting exon 14 with reduced coverage indicative of exon skipping.

4.1.3.3. Alteration-specific sections facilitate tumor result exploration

ClinBioNGS organizes all tumor-related findings into dedicated sections within the interactive report, facilitating an intuitive exploration and review of detected somatic alterations and biomarkers. Each alteration type is presented through a common structure that includes an overview of key findings, informative visualizations (seen in the previous section), and dynamic result tables. Additionally, calculated biomarker scores for TMB and MSI are shown in separate subsections.

Small Variants (SNVs and InDels)

The following screenshots illustrate two core components of the "SNV and InDel" section using the previously shown case with a Tier I *BRAF V600E* and four Tier III *TP53* mutations (**Figure 17**).

- **Figure 26** shows the "Overview" subsection:
    - The top panel highlights the most relevant findings, also presented in the "Main Results" section (**Figure 23**).
    - Middle panel provides summary statistics (e.g., assigned flags, clinical and oncogenic classifications), using color-coded categories.

- Bottom panel displays collected tumor-specific clinical evidence from CIViC (e.g., predictive, prognostic, diagnostic). Actionable associations are detected for this tumor case (i.e., trametinib and dabrafenib for *BRAF V600E* in NSCLC).

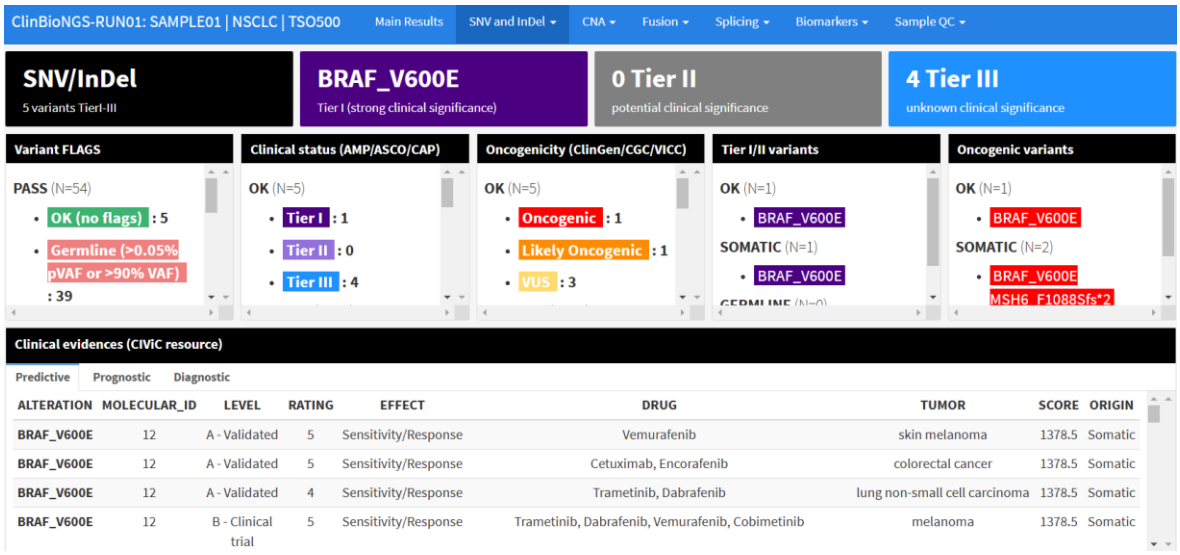


Figure 26. Overview of small variant results in the ClinBioNGS report.
Top findings, summary statistics, and CIViC clinical evidence are organized into distinct panels. Color coding is used for quick visual reference, and tumor-specific clinical evidence is displayed at the bottom.

- **Figure 27** presents the "Somatic Datatable" subsection:
  - In this example some filters have been applied to show SNVs with the "OK" flag.
  - Clinically relevant (Tier I/II) and oncogenic variants are prioritized and highlighted.
- Users can consult detailed annotation, apply custom filters, and export selected subsets.

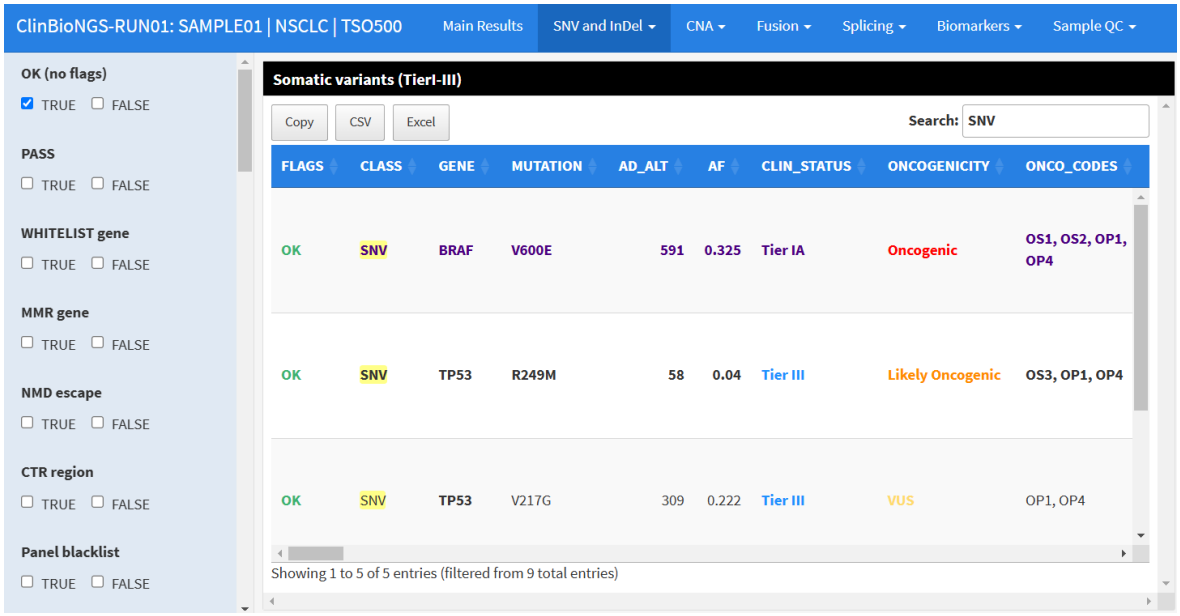


Figure 27. Interactive table of somatic small variants in the ClinBioNGS report.
Variants are color-coded by clinical relevance and oncogenicity. Rows are sortable and filterable. Top-tier variants appear at the top, and selected filters are applied to streamline review.

CNAs

The CNA "Overview" (similar to **Figure 26**) is provided (**Supplementary Figure 2A**), along with interactive CNA tables (**Figure 28**):

- **Figure 28A** shows gene-level CNA results:
  - Visual flags and CNA classifications are color-coded (e.g., red/blue for AMP/DEL).
  - Results are ranked by clinical tier and confidence flag for fast triaging.
- **Figure 28B** presents arm-level CNA results:
  - Only altered chromosomes are displayed based on filters.
  - Each row includes metrics such as mean copy ratio and frequency from the variant registry (AC_SAMPLES, AF_SAMPLES).

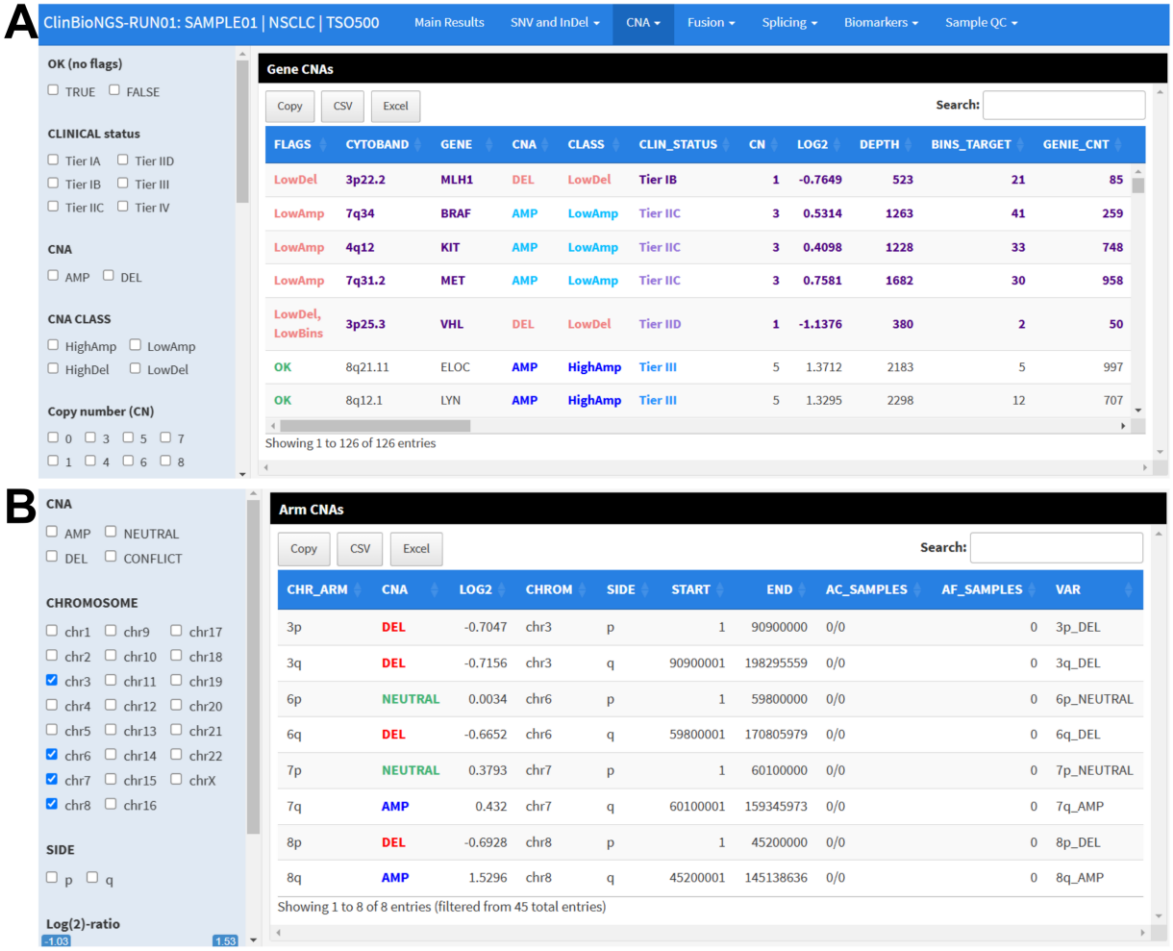


Figure 28. Interactive tables of CNA results in the ClinBioNGS report.
(**A**) Gene-level CNA results include color-coded CNA classification and QC flags, and the annotated metrics. (**B**) Arm-level CNA results filtered to show chromosomes with non-neutral events.

RNA alterations: fusions and splice variants

Each RNA variant type has its own section in the report, including visual summaries (**Supplementary Figure 2B-C**) and interactive tables of results that include variant names, clinical classification, QC flags, metrics, and a filtering column (**Figure 29**):

- **Figure 29A** displays gene fusion results (e.g., *EML4–ALK*).
- **Figure 29B** shows splice variants (e.g., *METex14*, plus a flagged known variant).

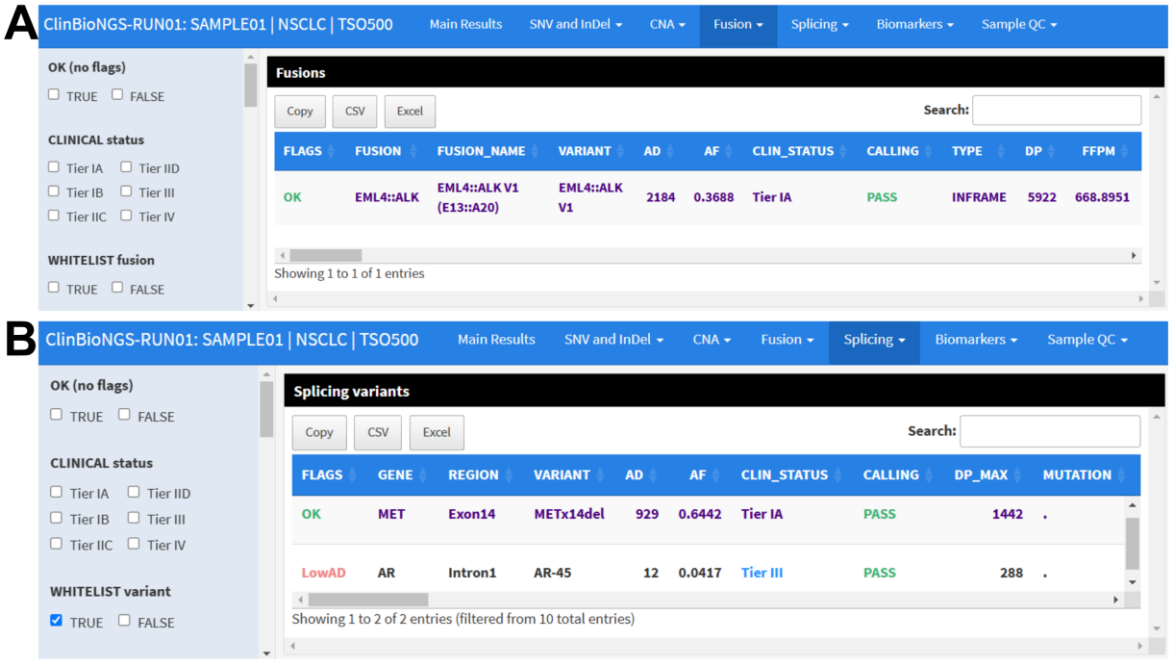


Figure 29. Interactive tables of RNA-based findings in the ClinBioNGS report.
(**A**) Gene fusions and (**B**) splice variants. Clinically relevant variants are highlighted in purple, low-confidence flags in red, and high-confidence flags in green ("OK"). Tables are filterable, scrollable, and exportable.

Genomic biomarkers: TMB and MSI

ClinBioNGS includes calculated metrics for TMB and MSI in dedicated tabs.

- **Figure 30** shows the "TMB" subsection:
  - Displays calculated scores (overall and non-synonymous).
  - Interactive table of eligible variants is provided below. Note that Tier I/II or oncogenic variants are excluded for TMB calculation.
- **Figure 31** shows the "MSI" subsection:
  - Presents calculated metrics (i.e., unstable microsatellites over total assessed)
  - Contains small variants in MMR genes, supporting MSI status beyond scores.
  - The example illustrates that only low-confidence variants are present, as no "OK" variants appear at the top. The observed oncogenic variant is found in the provided list of TSO500 recurrent mutations, suggesting a panel-specific recurrent artefact.

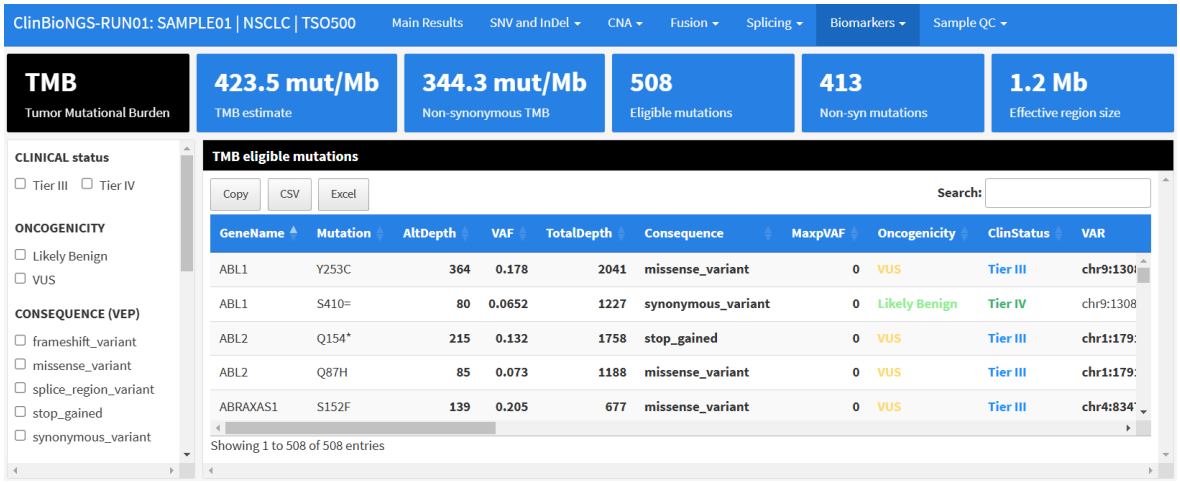


Figure 30. TMB results in the ClinBioNGS report.
Calculated metrics appear at the top. Annotated eligible variants are shown below. Filter options and color-coding help interpret the variant selection criteria. Rows have been sorted by gene name by clicking on the corresponding column.

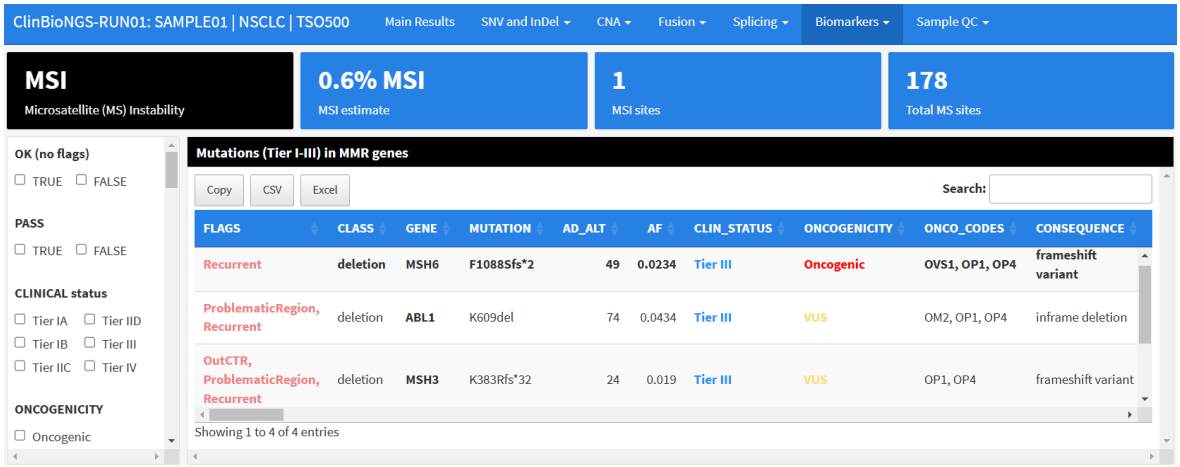


Figure 31. MSI results in the ClinBioNGS report.
Calculated metrics are shown above. Variants in MMR genes are displayed below. Flags, clinical status, and oncogenicity are highlighted.

4.2. Accurate detection of small variants across multiple NGS panels

ClinBioNGS achieved high accuracy for small variant detection using SEQC2 reference datasets across six commercial NGS panels. **Figure 32** summarizes the performance results:

- **Figure 32A** presents replicate-level precision and recall values for both ClinBioNGS and the commercial pipelines.
- **Figure 32B** shows the distribution of F1-scores per panel, comparing the overall performance of both pipelines.

Precision (0.987−1.000), recall (0.920−0.997), and F1-score (0.956−0.999) were consistently high in ClinBioNGS. These results were in line with, and in several cases slightly exceed, those obtained from commercial pipelines. Particularly, ClinBioNGS showed superior performance for the AGL panel, which included the most comprehensive set of known positive variants (n = 2,824). Complete benchmarking results, including replicate-level metrics, are provided in **Supplementary Table 10**.

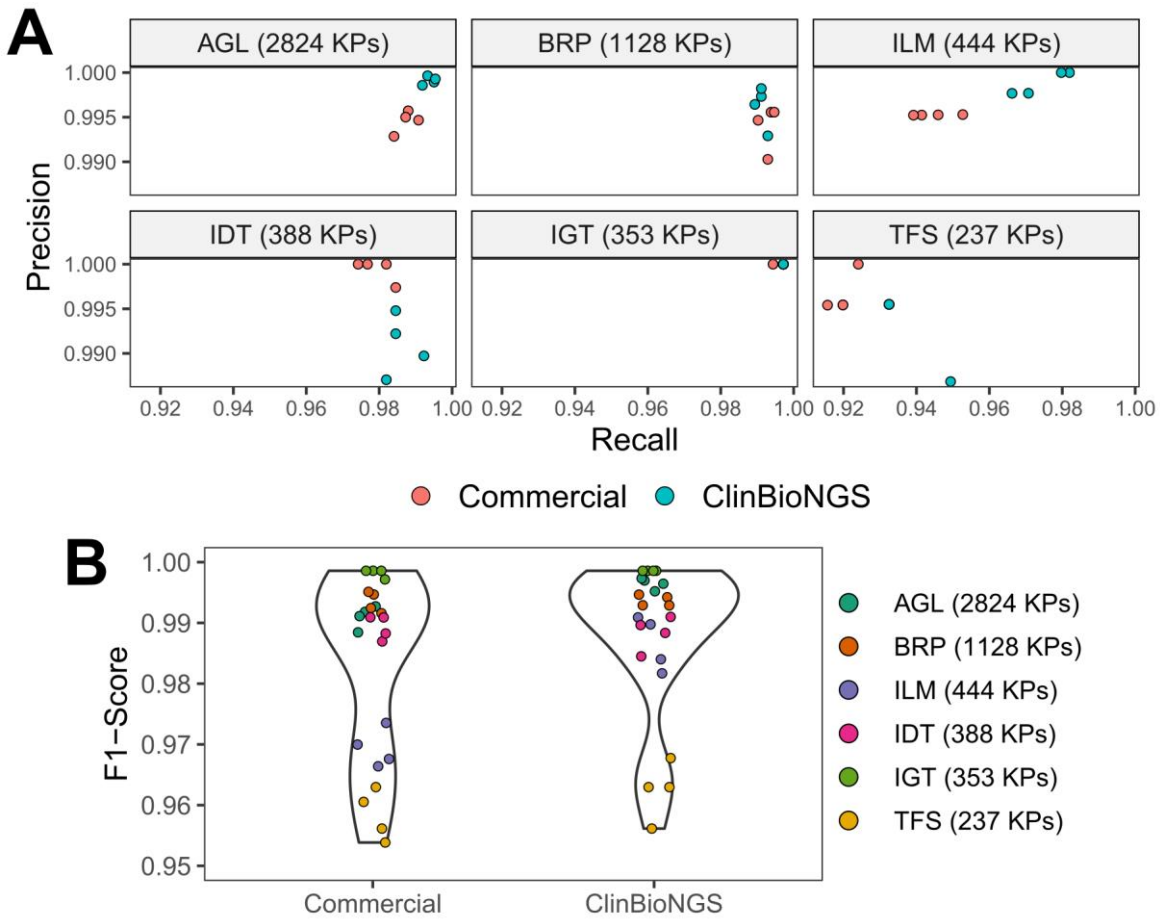


Figure 32. Cross-panel evaluation of ClinBioNGS small variant calling using SEQC2 datasets.
(**A**) Precision and recall metrics for each of four replicates per panel. Points are colored by pipeline (red for commercial, blue for ClinBioNGS). (**B**) Violin plots showing F1-score distributions across replicates for each panel, grouped by pipeline and colored by panel.

4.3. Real-world comparative analysis across commercial panels

4.3.1. High concordance for detecting cancer-related alterations

To evaluate the clinical utility of ClinBioNGS beyond controlled benchmarking, the pipeline was applied to 2,024 clinical tumor samples using three commercial pan-cancer NGS panels: Illumina TSO500 (n = 755), Ion Torrent OCA (n = 595), and Ion Torrent OPA (n = 674). Cohort-level characteristics are detailed in **Supplementary Table 11**.

Figure 33 shows the full comparative analysis of cancer-related alterations between ClinBioNGS and commercial pipelines. ClinBioNGS demonstrated high concordance with commercial pipelines, recapitulating 97% of small variants (3,502 of 3,606; **Figure 33A**), 89% of CNAs (2,083 of 2,339; **Figure 33B**), and 94% of RNA alterations (217 of 231; **Figure 33C**). Aggregate benchmarking metrics are summarized in **Supplementary Table 12**.

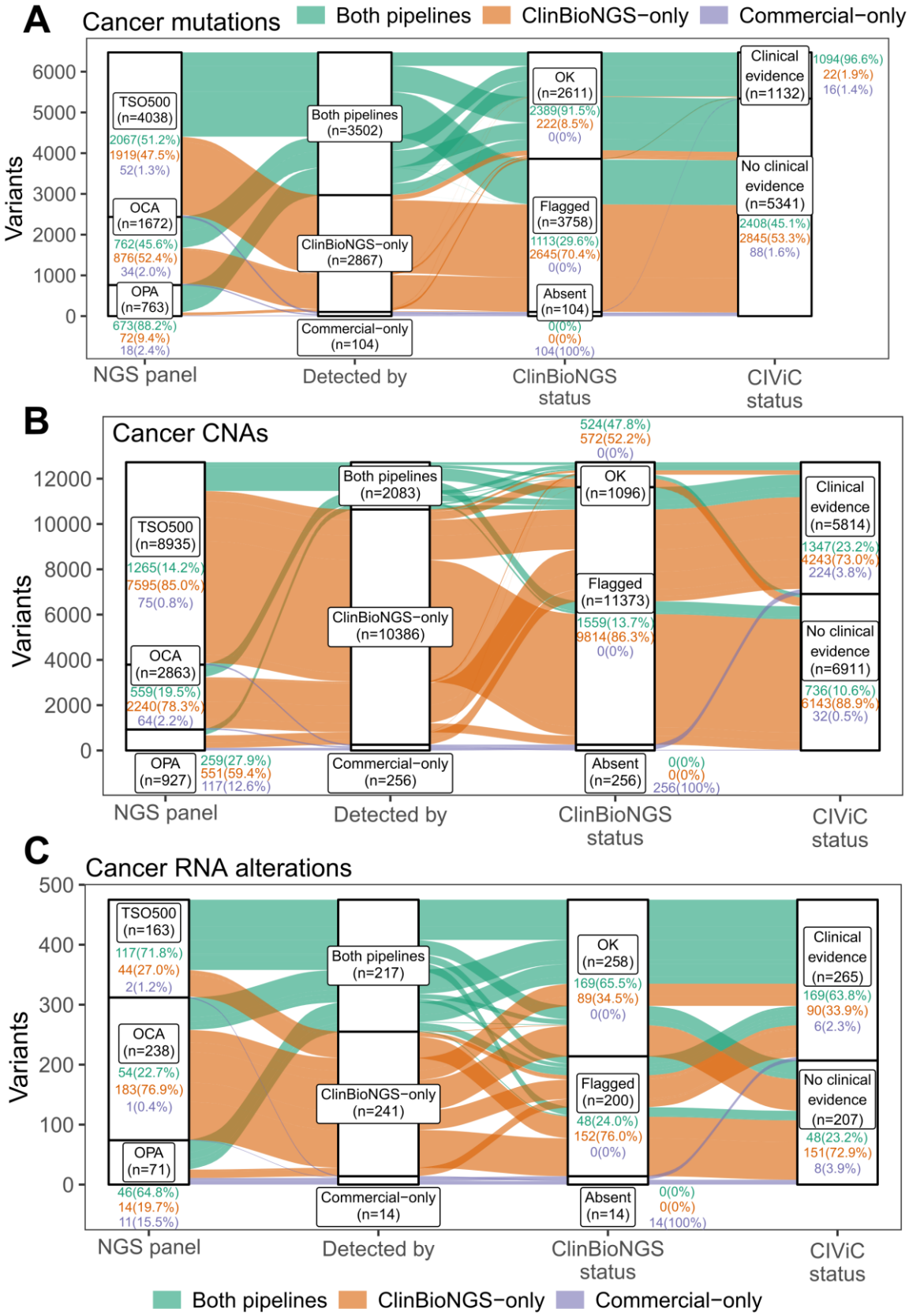


Figure 33. Real-world comparative analysis of all cancer-related alterations.
Alluvial plots showing concordance for cancer-related (**A**) mutations, (**B**) CNAs, and (**C**) RNA alterations. "OK" and flagged (i.e., secondary flags) ClinBioNGS variants are included. Each plot displays NGS panel, detection status, ClinBioNGS variant classification status, and clinical evidence status. Flows are colored by detection status and annotated with absolute counts and percentages.

Moreover, strong correlations were observed for normalized copy ratio values (R = 0.97 for TSO500, R = 0.96 for OPA; OCA excluded due to unavailable commercial values) and for absolute CN estimates (R = 0.74 for TSO500, R = 0.75 for OPA, R = 0.89 for OCA) (**Figure 34**).

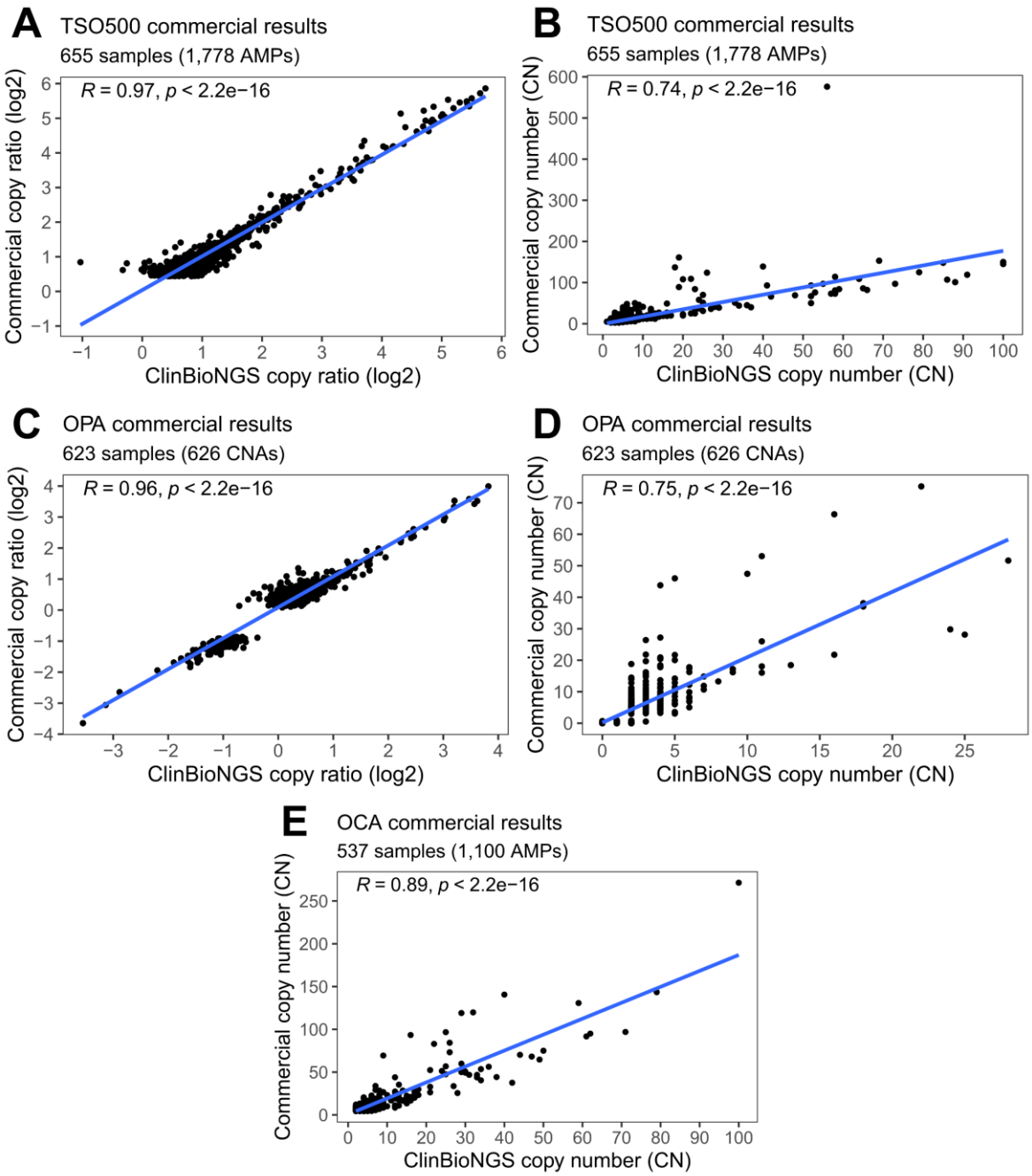


Figure 34. Correlation of copy ratios and CNs between ClinBioNGS and commercial pipelines.
Results are shown for TSO500 (**A-B**), OPA panel (**C-D**), and OCA (**E**) panels. The x-axis represents values from ClinBioNGS, and the y-axis represents those from the commercial pipeline. Pearson correlation coefficients (R) and linear regression lines are shown.

4.3.2. Discrepancies between ClinBioNGS and commercial solutions

Low-confidence ClinBioNGS calls—flagged due to limited support, location in problematic genomic regions, or recurrence in background samples (**Supplementary Figure 3**)—were mostly unique and lacked potential clinical relevance. Therefore, subsequent discrepancy analyses were restricted to high-confidence ("OK") variants to ensure greater interpretability and clinical value.

ClinBioNGS reported 222 additional mutations (**Figure 35**).

- Fourteen showed associated clinical evidence from CIViC (**Supplementary Table 13**):
  - Thirteen had low VAFs (<5%) and were likely filtered by commercial pipelines due to limited support or suboptimal quality metrics.
  - A *KRAS G12A* mutation (OPA) with 288 supporting reads and 17% VAF was not reported by the commercial pipeline.
  - An *EGFR S768I* mutation (TSO500) with 1.4% VAF was missed by the commercial solution but orthogonally confirmed using Roche Cobas EGFR Mutation Test v2.
- Among the remaining 208 variants without clinical evidence, 35 were classified as oncogenic and 33 as likely oncogenic (**Supplementary Table 14**), including:
  - Borderline calls that were filtered by commercial pipelines due to limited support and low-quality issues.
  - Well-supported variants that were either blacklisted by TSO500 commercial pipeline or omitted from predefined SNV/InDel lists used in OCA and OPA workflows.

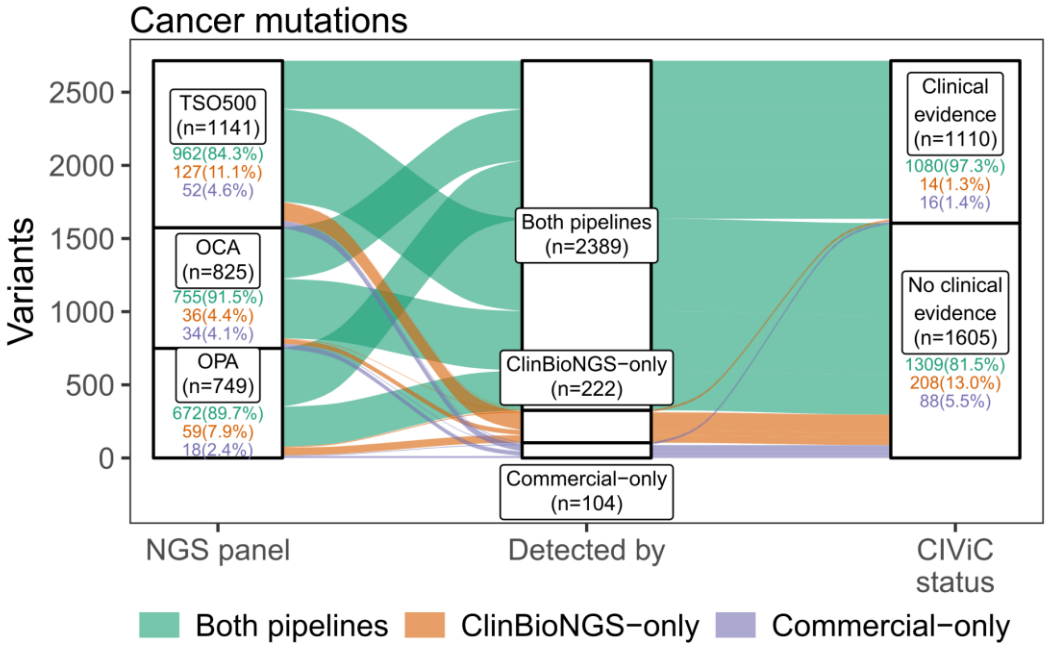


Figure 35. Real-world benchmarking of "OK" ClinBioNGS cancer-related mutations.
ClinBioNGS variants are restricted to high-confidence calls (i.e., those passing all internal flags). Alluvial plot showing concordance for cancer-related mutations. Each plot displays NGS panel, detection status, and clinical evidence status. Flows are colored by detection status and annotated with absolute counts and percentages.

Conversely, 104 commercial variants were not reported by ClinBioNGS (**Figure 35**).

- Sixteen had associated clinical evidence **(Supplementary Table 15)**:
    - These variants showed in ClinBioNGS variant callers specific filters related to poor quality or strand bias. That is the reason why they do not reach the minimum number of callers established for each panel (2/4 for TSO500 and 3/5 for Ion Torrent panels).
    - Although these variants were mostly flagged by ClinBioNGS as "LowCallers" (i.e., primary flag), which led to their exclusion from concordance counting, they would also be presented in the final results for the appropriate review.
    - In the absence of orthogonal validation, their clinical relevance remains uncertain.
- Among the remaining 88 variants (**Supplementary Table 16**), most were also filtered by ClinBioNGS callers due to poor quality and strand bias issues. Notably, *MST1 G673S* and *U2AF1 S34F* were recurrently missed by ClinBioNGS in the TSO500 panel.

ClinBioNGS reported 572 additional CNAs (**Figure 36**), including 191 with clinical evidence. Most involved genes were not assessed by the commercial pipelines, with *CDKN2A* deletion being the most frequent event (n = 107) across all panels (**Supplementary Table 17**).

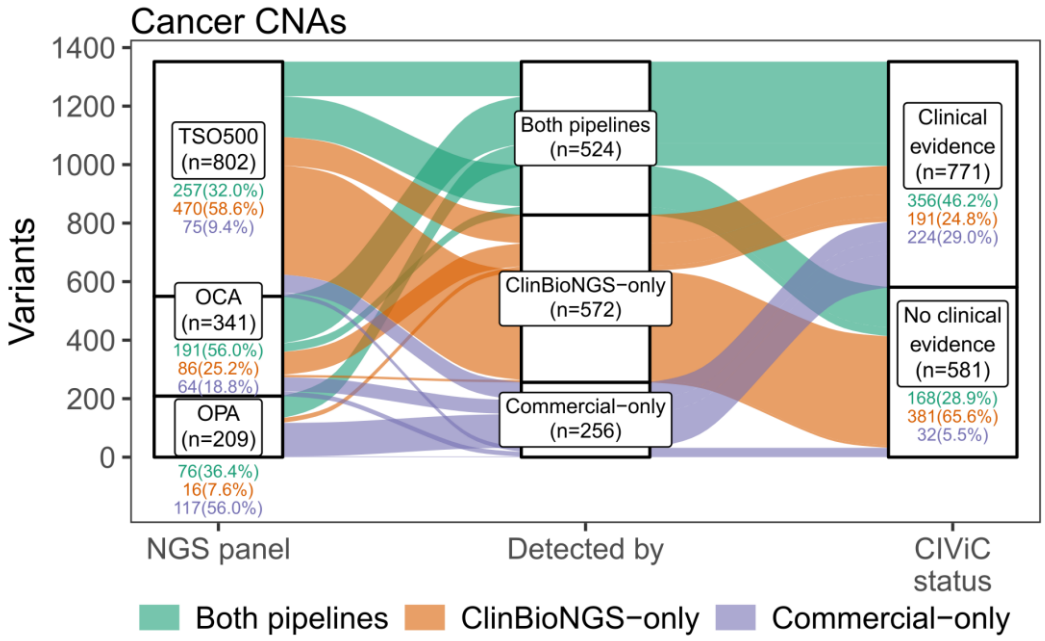


Figure 36. Real-world benchmarking of "OK" ClinBioNGS cancer-related CNAs.
ClinBioNGS variants are restricted to high-confidence calls (i.e., those passing all internal flags). Alluvial plot showing concordance for cancer-related CNAs. Each plot displays NGS panel, detection status, and clinical evidence status. Flows are colored by detection status and annotated with absolute counts and percentages.

Conversely, 256 CNAs were exclusively reported by commercial solutions (**Figure 36**), including 224 with clinical evidence (**Supplementary Table 18**). Most discrepancies were attributable to borderline events in TSO500 (median CN = 3) and TP-based CN corrections applied in OCA and OPA samples (median TP = 20%).

ClinBioNGS identified 89 additional RNA events (**Figure 37**), including 37 with clinical evidence from the OCA panel (**Supplementary Table 19**).

- Fusions (n = 21):
  - Most were supported by low number of reads (median = 21), often falling below the thresholds of commercial filters.
  - Notably, one *EML4-ALK* fusion (17 reads) was missed by the commercial pipeline due to overall sample QC failure.
- Splice variants (n = 16):
  - All were androgen receptor splice variant 7 (*AR-V7*) robustly supported by ClinBioNGS but not assessed by commercial analysis.

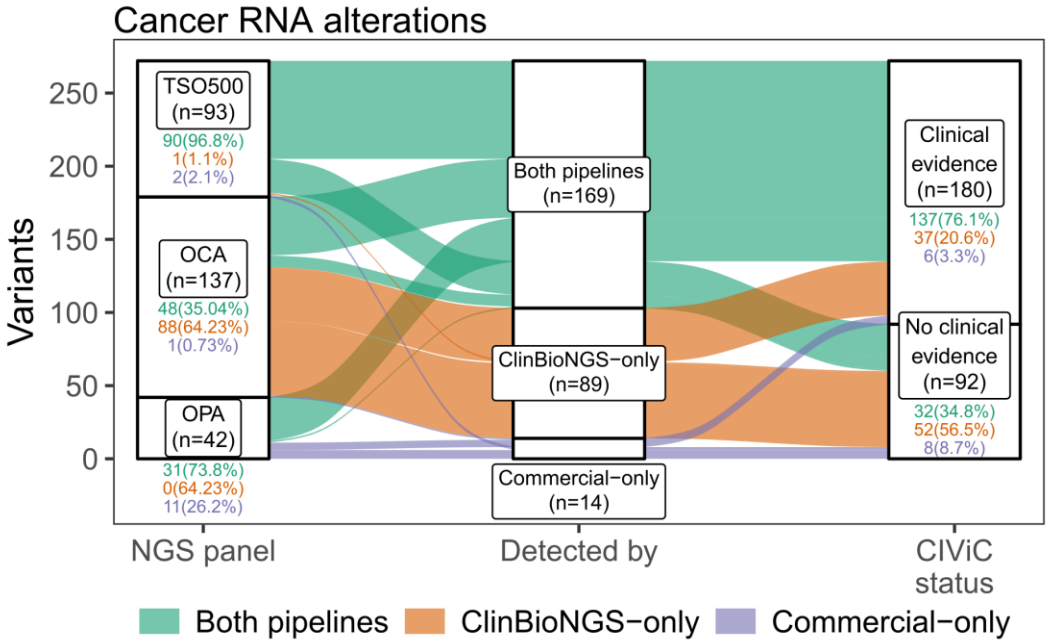


Figure 37. Real-world benchmarking of "OK" ClinBioNGS cancer-related RNA alterations.
ClinBioNGS variants are restricted to high-confidence calls (i.e., those passing all internal flags). Alluvial plot showing concordance for cancer-related RNA alterations. Each plot displays NGS panel, detection status, and clinical evidence status. Flows are colored by detection status and annotated with absolute counts and percentages.

Conversely, 14 RNA events were uniquely called by commercial pipelines (**Figure 37**), including 6 with clinical evidence (**Supplementary Table 20**).

- Three low-read *BRAF* fusions, lacking strong supporting evidence and likely clinically irrelevant without further validation.
- One *KIF5B-RET* fusion (OPA) with higher read support. Upon disabling deduplication, ClinBioNGS recovered this event. From now, OPA's deduplication is disabled by default.
- One *METex14* event (OPA) was also detected by ClinBioNGS, although flagged as "LowSupport" (8 supporting reads), which is comparable to the commercial call.

4.3.3. High concordance in biomarker classification (TMB and MSI)

ClinBioNGS demonstrated strong agreement with commercial TSO500 pipeline in the classification of TMB-High and MSI-High samples. Correlation was high for both metrics (R = 0.99 for TMB [**Figure 38A**] and R = 0.97 for MSI [**Figure 38C**]).

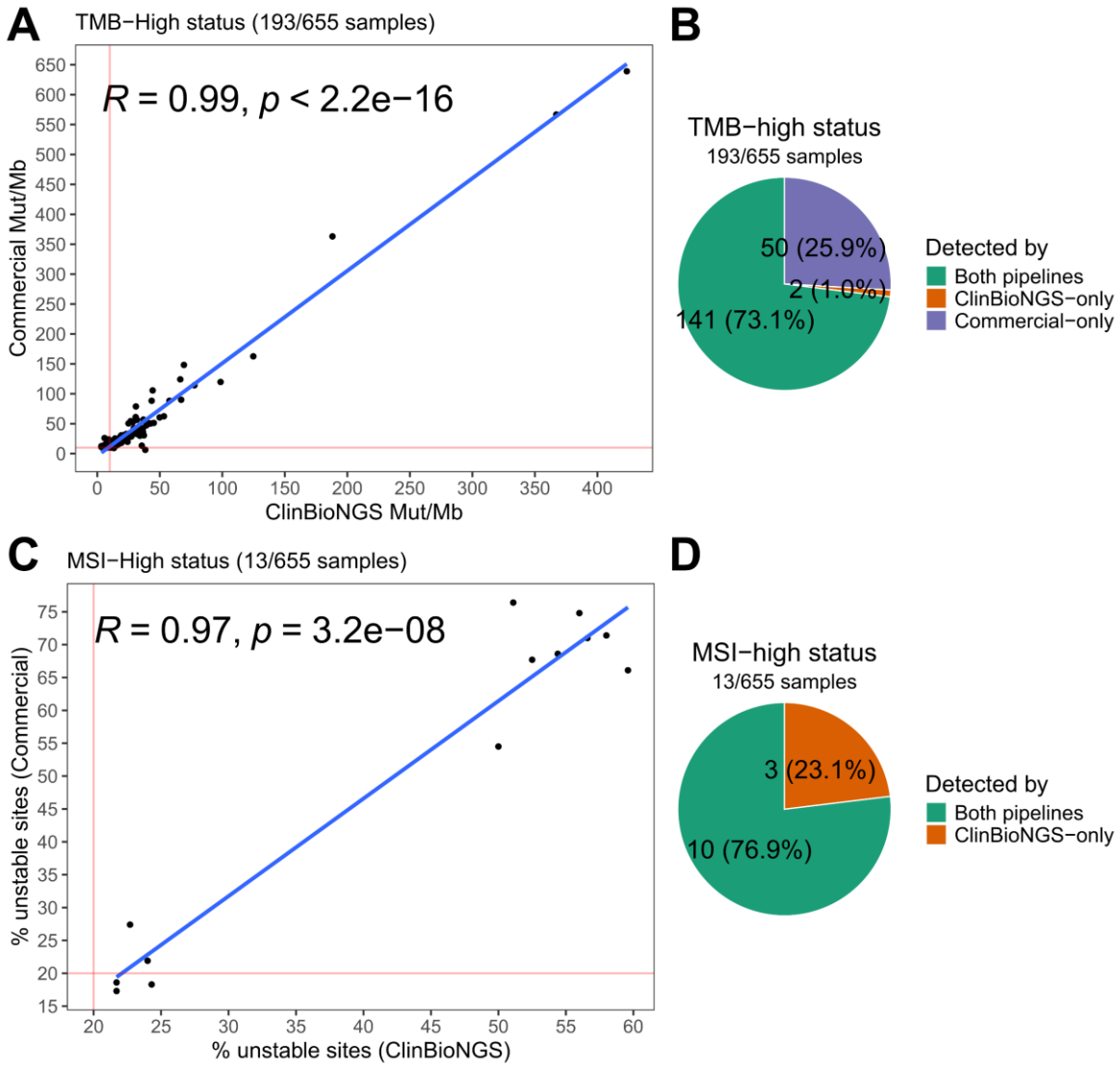


Figure 38. Biomarker agreement between ClinBioNGS and TSO500 commercial pipeline.
(**A**) Correlation plot of TMB-high values. The x-axis shows ClinBioNGS values, and the y-axis shows those from the commercial pipeline. Pearson correlation coefficient (R) and the linear regression line are shown. Red dashed lines at 10 mut/Mb indicate the threshold used for TMB-high classification. (**B**) Pie chart showing concordance of TMB-high classification between pipelines, with color-coded segments indicating agreement or discrepancy between pipelines. (**C**) Correlation plot of MSI-high values. Red dashed lines at 20% unstable loci indicate the MSI-high classification threshold. (**D**) Pie chart showing concordance of MSI-high classification.

For TMB-High status, 141 of 193 samples (73.1%) were concordantly classified by both pipelines (**Figure 38B**). Among the 52 discordant cases, 50 were classified as TMB-High only by the commercial pipeline (TSO500). Notably, most discordant cases had TMB values close to the high/low decision threshold, as shown in the scatter plot.

For MSI-High status, 10 of 13 samples (76.9%) were concordantly classified between pipelines (**Figure 38D**). In the remaining three discordant cases, ClinBioNGS classified the samples as MSI-High while the commercial pipeline did not. All three had MSI scores near the threshold, indicating potential classification ambiguity. Importantly, one discordant sample—initially classified as MSI-Low by the commercial pipeline (17.3% unstable loci) but as MSI-High by ClinBioNGS (21.7%)—was confirmed to be MSI-High (~75% instability) upon later re-sequencing, supporting the ClinBioNGS classification. No additional validation was available for the other discordant samples.

4.4. Case studies illustrating the extended capabilities of ClinBioNGS

Beyond controlled benchmarking, the real-world application of ClinBioNGS in clinical and translational settings underscores its value in handling diverse and challenging scenarios commonly encountered in somatic panel analysis. The following case studies illustrate how the pipeline delivers robust performance in routine practice, presenting critical improvements over commercial solutions.

These examples emphasize several key strengths of ClinBioNGS:

- Reliable detection and representation of complex alterations, including adjacent InDels and arm-level CNAs.
- Recovery of relevant variants that may be missed or filtered by vendor pipelines due to platform-specific limitations or filtering.
- Improved transparency and interpretability, thanks to comprehensive flagging, standardization, and intuitive visualizations.
- Adaptability to non-standard or research-oriented samples (e.g., xenografts).

4.4.1. Correction of TMB overestimation in pancreatic PDX samples

A set of 16 PDX pancreatic tumor samples analyzed with the TSO500 panel showed abnormally high numbers of small variants and inflated TMB scores. The likely cause was mouse DNA contamination, which is common in PDX models but not accounted for by the commercial pipeline.

To address this, a pre-processing step was integrated into ClinBioNGS to filter mouse reads prior to variant calling. This adjustment resulted in a substantial reduction in variant counts and normalization of TMB scores to expected ranges.

This case highlights the versatility of ClinBioNGS to accommodate complex experimental settings and adapt to research-specific requirements.

4.4.2. Refined detection of complex *EGFR* exon 19 deletions in Ion Torrent OPA samples

4.4.2.1. Case 1: Resolution of a multi-event complex InDel

In one sample analyzed with the Ion Torrent OPA panel, the commercial pipeline reported a complex *EGFR* exon 19 variant (chr7:55242468_ATTAAGAGAAGCAACATC/GCAACA, VAF 1.9%, hg19). The OPA viewer suggested multiple adjacent small deletions with higher VAFs (**Figure 39A**).

ClinBioNGS resolved this complex signal into three distinct deletions (hg19):

- chr7:55242465_GGAATTAAGA/G (VAF 26%).
- chr7:55242479_CA/C (VAF 26%).
- chr7:55242483_ATC/A (VAF 26%).

These calls aligned more closely with the read evidence and yielded more consistent and reliable VAF estimates, enhancing clinical interpretability. This illustrates the advantage of ClinBioNGS's multi-caller consensus approach in accurately deconstructing complex InDels.

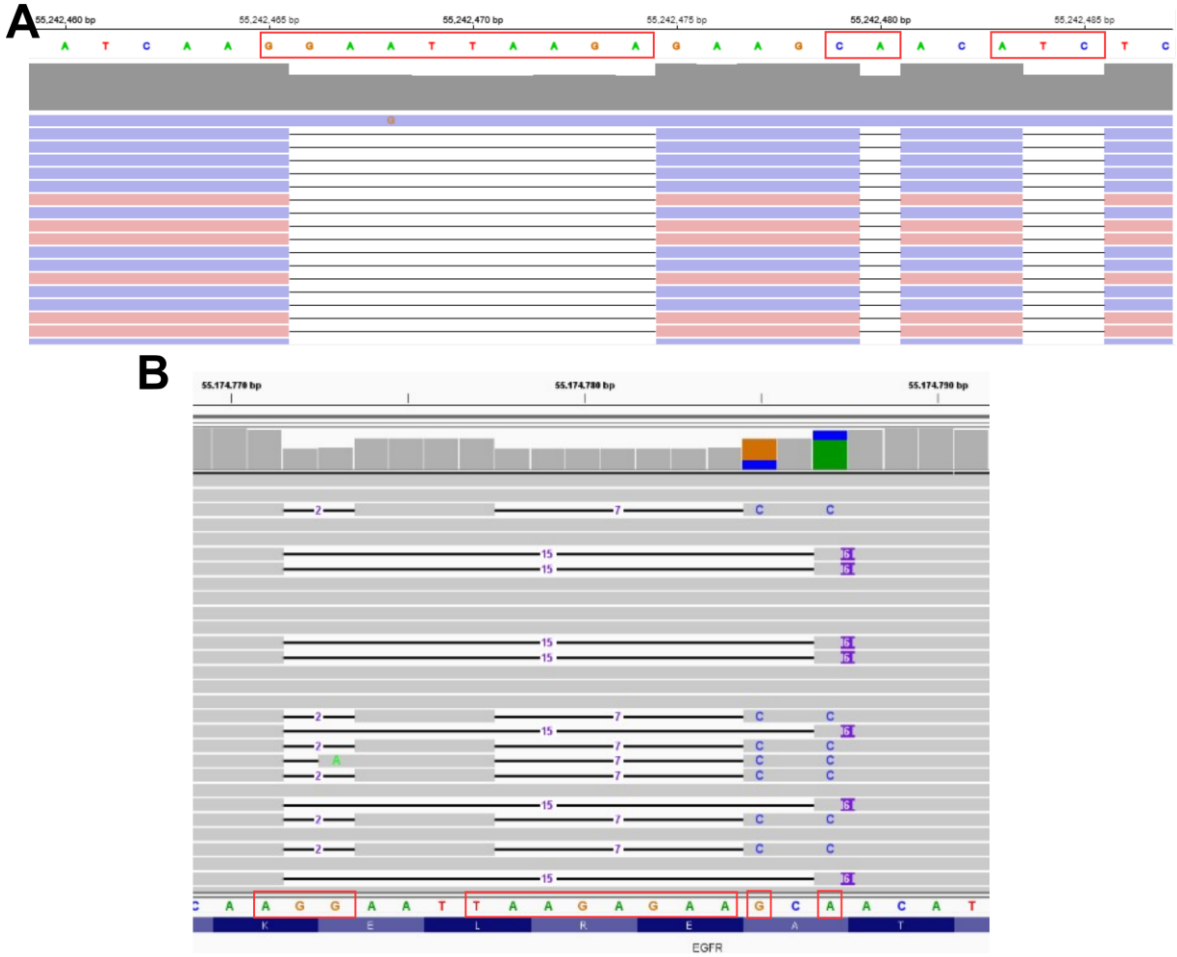


Figure 39. Detection of complex *EGFR* exon 19 variants in Ion Torrent OPA samples.
(**A**) Screenshot from the OPA viewer. (**B**) Screenshot from IGV visualizing the variants detected by ClinBioNGS.

4.4.2.2. Case 2: Recovery of a filtered complex variant

In another case, an *EGFR* exon 19 deletion was detected via cfDNA by a Foundation Liquid test. The sample was initially analyzed using the OPA panel, but no variant call was returned—the commercial pipeline flagged the event as "Complex" and filtered it out. A second analysis with relaxed parameters ("allowing complex variants") reported a composite variant without a clear VAF:

- c.2235_2250delinsAATTCCC (chr7:55174772_GGAATTAAGAGAAGCA/AATTCCC).

Re-analysis with ClinBioNGS identified the following four variants in that region (hg38):

- chr7:55174771_AGG/A; c.2235_2236del (VAF 47.5%).
- chr7:55174777_TAAGAGAA/T; c.2241_2247del (VAF 47.4%).
- chr7:55174785_G/C; c.2248G>C (VAF 47.5%).
- chr7:55174787_A/C; c.2250A>C (VAF 47.7%).

Inspection with IGV (**Figure 39B**) revealed two main read groups:

- One group showed two interspersed InDels followed by two SNVs, consistent with the calls made by both pipelines.
- The other group contained a full-length deletion spanning the affected region. This allele was not reported by any pipeline, likely due to lower read support. ClinBioNGS selects the most frequent representation across callers in multi-allelic contexts, so less frequent forms may be omitted from the final consensus.

The patient received first line osimertinib and achieved a partial response and significant clinical benefit, demonstrating the relevance of identifying complex variants that align with clinically actionable *EGFR* exon 19 deletion profiles.

4.4.2.3. Interpretation and implications

These cases illustrate both strengths and limitations of ClinBioNGS in handling complex InDels:

- Strengths: ClinBioNGS's multi-caller consensus approach facilitates the deconstruction of complex variant signals and yields clearer representations with reliable VAF estimates—key factors for proper clinical interpretation and to define whether patients harboring those complex genomic alterations may benefit from targeted therapies. Moreover, it can recover complex deletions that commercial pipelines may filter out due to variant complexity or flagging systems.
- Limitations: As individual variant calls are reported per caller, phasing information may be lost. When most callers report variants separately, they appear as isolated events in the consensus output. Thus, manual review (e.g., IGV) is still necessary in complex regions to evaluate allelic phasing and reconstruct the full mutational profile.

These examples underscore the multiallelic and complex nature of *EGFR* exon 19 deletions and reinforce their clinical relevance. While ClinBioNGS offers a powerful solution for resolving these events, future integration of phasing-aware algorithms—or complementary long-read sequencing technologies—may further improve the characterization of such complex alterations.

4.4.3. Recovery of a relevant germline *MSH6* mutation

A known pathogenic germline frameshift mutation in the *MSH6* gene (chr2:47803500_A/AC; p.Phe1088Leufs*5; ClinVar ID: 89364) was previously identified through a hereditary cancer susceptibility panel at our institution. This variant, known to be heterozygous and classified as pathogenic, was expected to be detected in a tumor sample subsequently analyzed with the Illumina TSO500 panel.

However, the TSO500 Local App failed to report the mutation. The reason was that the variant position lies within a blacklisted region defined by the commercial pipeline, likely to avoid reporting recurrent artifacts. In contrast, ClinBioNGS successfully detected the *MSH6* variant in this sample, as well as in three additional samples from the benchmarking cohort. In none of these cases was reported by the commercial pipeline.

ClinBioNGS classified the variant as oncogenic and assigned it an “OK” flag. The variant showed intermediate VAF values, consistent with expected heterozygous germline origin, and had no strong representation in general population databases, further supporting its pathogenic classification. For the three additional cases, the absence of matched normal tissue precluded confirmation of their germline or somatic origin.

The commercial pipeline’s blacklisting appears to be related to a different variant at the same position (chr2:47803500_AC/A; F1088Sfs*2), which is recurrently detected in TSO500 data (found in our TSO500-recurrent variant list). This variant was detected in 456 samples across the benchmarking cohort and is flagged as “Recurrent” by ClinBioNGS. This suggests that the position was blacklisted due to its association with this alternate, artifact-prone event.

This case underscores the strength of ClinBioNGS’s transparent and informative flagging system:

- The truly pathogenic variant was retained and reported with appropriate annotation.
- The potentially artifactual variant was detected and explicitly flagged as “Recurrent”, aiding interpretation without suppressing evidence.

In contrast to the commercial pipeline’s strategy of outright filtering, ClinBioNGS’s philosophy of reporting all potentially relevant variants—alongside contextual flags and quality indicators—minimizes the risk of missing potentially actionable or clinically relevant genomic alterations.

4.4.4. Accurate detection of typical arm-level CNAs in uveal melanoma

As shown previously (**Figure 18**), ClinBioNGS successfully identified canonical arm-level CNAs in a uveal melanoma case, including 8q AMP and 3p/q, 6q, and 8p DELs. These hallmark alterations were clearly visualized in the CNA plots, illustrating the pipeline's ability to accurately detect both focal and arm-level CNAs. No additional data was available regarding orthogonal validation of this case, and no other uveal melanoma samples were sequenced with the TSO500 panel during the study period, limiting the possibility of further cross-validation.

4.4.5. Detection of 1p/19q co-deletion in oligodendroglioma

In a TSO500 case of oligodendroglioma validated by FISH as harboring a hallmark 1p/19q co-deletion, the TSO500 Local App failed to report any CNA events. In contrast, ClinBioNGS clearly identified both 1p and 19q arm-level deletions, as visualized in **Figure 40**. FISH was performed on FFPE tumor tissue using dual-color probes for chromosomal regions 1p36/1q25 and 19q13/19p13 (Vysis). A total of 100 non-overlapping nuclei were evaluated, with allelic loss defined as a red/green signal ratio ≤ 0.8. The observed signal ratios were 0.65 for chr1 and 0.6 for chr19, consistent with allelic loss of both 1p and 19q. These results confirm the presence of a canonical 1p/19q co-deletion, in agreement with the arm-level events detected by ClinBioNGS (FISH image was not available).

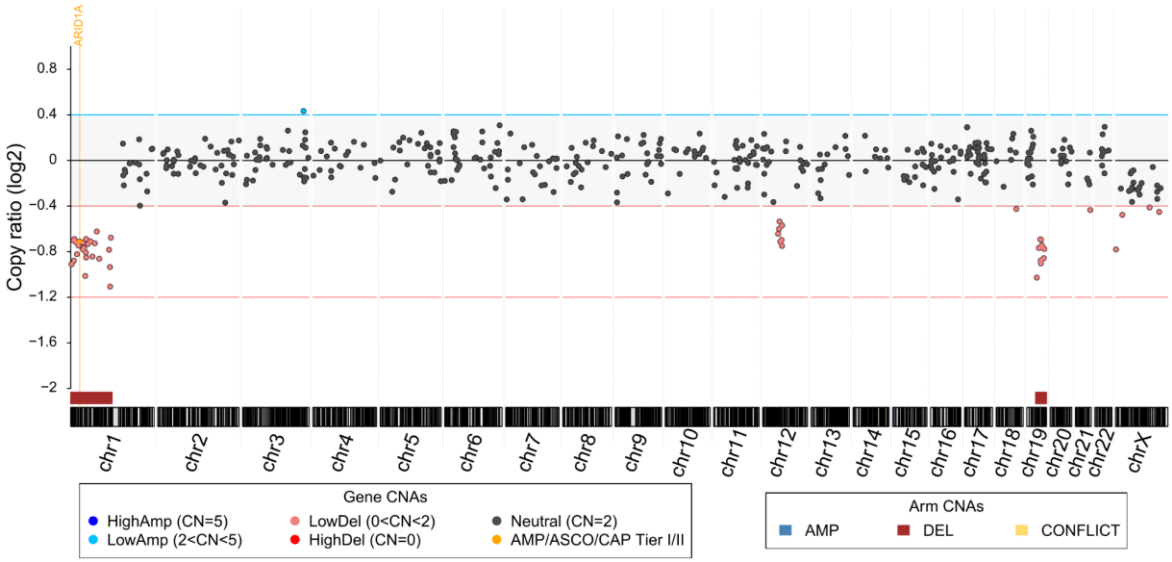


Figure 40. Visualization of 1p/19q co-deletion in oligodendroglioma detected by ClinBioNGS.

4.4.6. Cross-validation of arm-level CNAs in mesothelioma using sWGS

A case of malignant pleural mesothelioma was analyzed using both the Illumina TSO500 panel and shallow whole genome sequencing (sWGS) in a research setting, providing an opportunity to cross-validate CNA profiles obtained through targeted and genome-wide approaches.

ClinBioNGS successfully detected several arm-level CNAs in the TSO500 data that were concordant with those observed in the sWGS analysis, including 11q AMP and 12q, 14pq, and 22pq DELs.

These hallmark CNAs were clearly visualized in the CNA plots produced by ClinBioNGS and matched the copy-number patterns observed in the genome-wide sWGS profile (**Figure 41**).

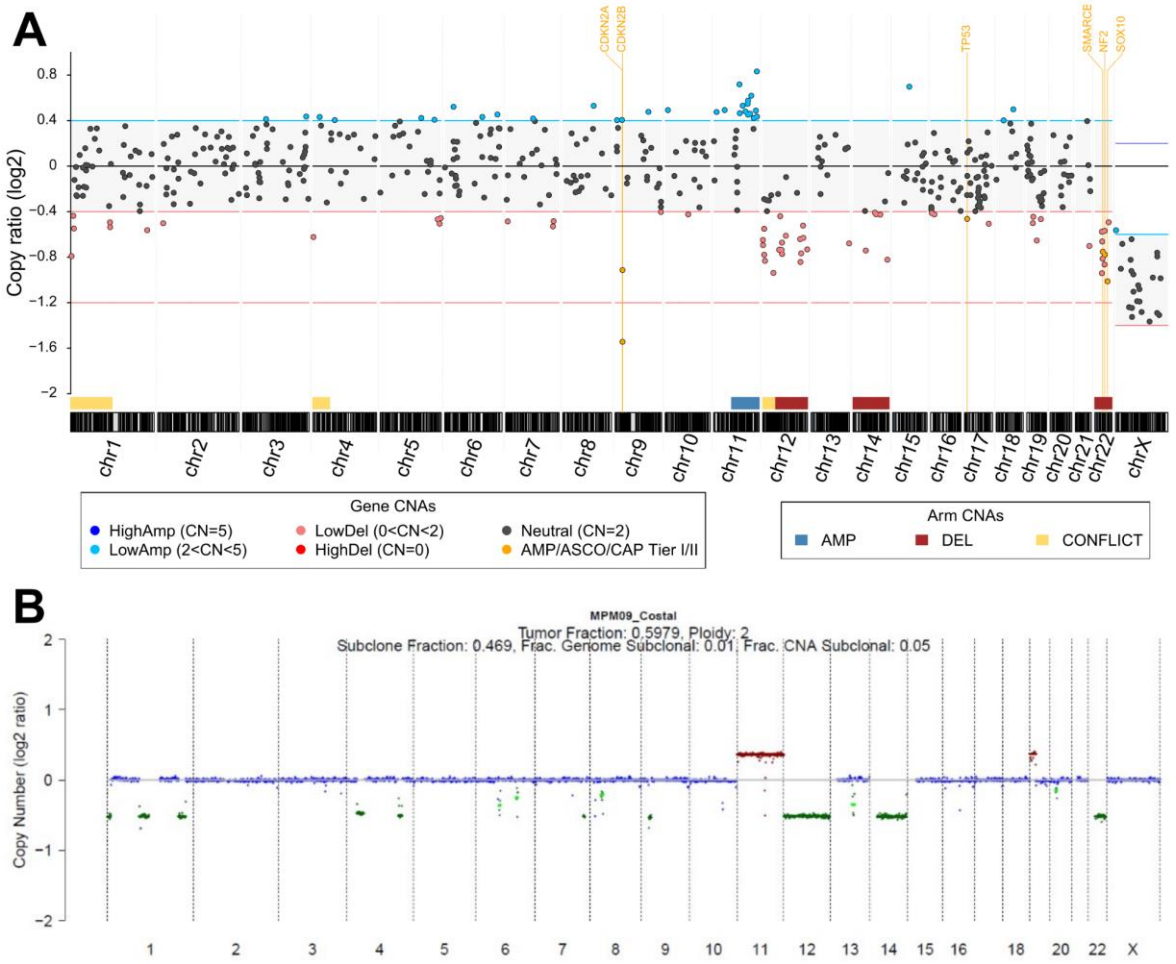


Figure 41. Arm-level CNA detection in mesothelioma.
(**A**) CNA profile from ClinBioNGS (TSO500 panel). (**B**) CNA profile from sWGS results.

This case illustrates the reliability of ClinBioNGS in capturing broad chromosomal alterations even when derived from targeted sequencing panels. It also highlights the pipeline's ability to identify biologically plausible, arm-level CNAs with performance comparable to low-pass genome-wide methods.

The consistency with sWGS reinforces the pipeline's potential utility for tumor types like mesothelioma, where chromosomal instability and arm-level alterations are clinically informative but may be under-reported in standard panel workflows.

5. DISCUSSION

5.1. Overview and contextualization of ClinBioNGS

In this work, ClinBioNGS is presented as a novel open-source bioinformatics pipeline specifically developed to address the analytical and interpretative challenges associated with somatic NGS cancer panels. The pipeline has been designed to enable the comprehensive processing, analysis, and interpretation of genomic alterations derived from tumor-only sequencing data. In addition, over the course of this thesis, ClinBioNGS has been extensively validated using both public standardized datasets and internal clinical cohorts, establishing ClinBioNGS as a reliable and versatile tool for precision oncology.

The development of ClinBioNGS was driven by the increasing adoption of somatic NGS panels in routine clinical oncology and the critical need for standardized, transparent, reproducible, and scalable bioinformatics workflows that fulfill clinical-grade requirements[4–13,18].

Existing pipelines for somatic NGS panel analysis present important limitations in scope, flexibility, and clinical utility. To contextualize the development of ClinBioNGS within the landscape of existing solutions, we performed a comparative analysis of representative workflows covering commercial, institutional, and open-source pipelines (**Supplementary Table 21**). Commercial platforms (e.g., Archer Analysis, FoundationOne CDx, Illumina TSO500, QCI Interpret, SOPHiA DDM, Thermo Ion Reporter, VarSome Clinical) offer standardized, ready-to-use environments with vendor support and interactive graphical interfaces, but they are proprietary, limited in flexibility, and often restricted to fixed assays. Institutional pipelines such as MSK-IMPACT or DFCI OncoPanel provide clinically validated frameworks, but remain panel-specific and typically inaccessible outside their home institutions.

By contrast, academic and open-source workflows (e.g., BALSAMIC, bcbio-nextgen, nf-core/sarek, DNAscan2, MIRACUM-Pipe, PipeIT2, SCHOOL, TOSCA) provide transparency and adaptability, yet they often lack end-to-end integration, multi-omics support, or clinically oriented annotation and reporting. Many rely on static outputs, provide limited visualization capacity, or are no longer actively maintained.

ClinBioNGS was designed to bridge this gap by combining the portability and transparency of open-source solutions with the comprehensiveness and interpretability demanded by clinical and translational contexts. Compared with other pipelines, it uniquely integrates DNA and RNA analyses within a single workflow, supports tumor-only data, provides a comprehensive annotation framework with clinical prioritization, generates informative QC and variant-specific visualizations, and delivers

interactive HTML reports of results. This positions ClinBioNGS as a robust, versatile, and actively maintained alternative that extends beyond the limitations of existing solutions.

5.2. Key innovations and strengths of the pipeline

ClinBioNGS introduces a series of methodological and practical innovations that distinguish it from existing solutions (**Supplementary Table 21**) for somatic NGS panel analysis. The pipeline was designed to achieve a robust balance between flexibility, reproducibility, and clinical applicability, addressing many of the recurring challenges in routine genomic diagnostics[34,36,40,42].

5.2.1. Integrated DNA and RNA analysis

A key strength of ClinBioNGS is its ability to integrate both DNA- and RNA-based analyses within a unified, standardized workflow—an uncommon feature among most existing pipelines that often focus exclusively on a single data type. This dual capability enables the detection of diverse genomic alterations, including small variants (SNVs and InDels), CNAs, gene fusions, splice variants, and complex genomic biomarkers such as TMB and MSI. This multi-layered integration enables a more comprehensive tumor profiling, enhancing the detection of complex and co-occurring genomic events with potential clinical relevance[10,36,40–42].

5.2.2. Standardized and comprehensive annotation framework

A central strength of ClinBioNGS lies in its unified and fully standardized annotation framework, specifically designed to support accurate and clinically meaningful interpretation of somatic alterations[4,35,36,48]. Key features include:

- Panel-agnostic results: ClinBioNGS applies a consistent analytical strategy across different commercial panels, producing harmonized output formats and variant classifications. This ensures interpretability and comparability of results regardless of panel design.
- Consistent genome reference usage: All analyses are performed using the updated GRCh38 reference genome, which improves alignment accuracy and compatibility with current annotation resources[9,10,40]. For backward compatibility with legacy systems, output files also include liftover coordinates to GRCh37.
- Standardized file formats: The entire pipeline is built around widely accepted bioinformatics standards[9,38,46] (e.g., FASTQ, BAM, VCF), with fully annotated VCF files generated for each variant type (SNVs/InDels, CNAs, fusions, splice variants). This ensures interoperability with downstream tools and promotes transparency in the variant review[9,42].

- Up-to-date gene and transcript annotations: Annotations are based on HGNC-approved gene symbols, MANE Select transcripts, and HGVS-compliant nomenclature, in line with current clinical reporting best practices.
- Integration of curated knowledge bases: Multiple public and expert-curated resources—including GENIE, CIViC, CTAT, and others—are incorporated to enrich variant annotations with up-to-date biological, functional, and clinical information.
- Implementation of international prioritization guidelines: ClinBioNGS integrates recommendations from ClinGen/CGC/VICC for oncogenicity classification and AMP/ASCO/CAP guidelines for assessing clinical significance, enabling standardized and evidence-based variant prioritization.

This comprehensive and transparent annotation strategy supports robust variant interpretation while maintaining full traceability of the underlying evidence[35]. All annotations are directly accessible through the interactive HTML report, empowering users to evaluate variants in detail within a clinically oriented context.

5.2.3. Internal flagging and prioritization system

ClinBioNGS incorporates a comprehensive internal flagging system designed to enhance transparency, interpretability, and clinical relevance in variant reporting[34,62]. Each detected alteration is systematically evaluated and assigned one or more flags based on calling- and context-based indicators to help distinguish between well-supported and borderline findings.

Rather than discarding borderline or lower-confidence alterations through hard filters without any explanation in the final output—as commonly done in commercial pipelines—ClinBioNGS retains and transparently flags such variants, allowing users to assess variant quality and relevance with full contextual awareness.

For example, a particularly valuable feature of this approach is the systematic flagging of potential germline variants, which are compiled in a dedicated section of the final report. This allows for focused clinical review of potentially heritable findings[31,135], while maintaining the somatic scope of the pipeline.

By reducing the risk of FNs (e.g., discarding rare but relevant variants) and FPs (e.g., reporting artifacts as real events), the flagging system supports nuanced interpretation and multidisciplinary oversight, which is particularly valuable in complex or borderline cases.

5.2.4. Tumor-only analytical strategies

In routine somatic testing, the lack of matched normal samples poses a significant challenge for accurately distinguishing somatic mutations from germline variants and technical artifacts[10,12,41]. ClinBioNGS addresses this limitation by implementing a set of tumor-only analytical strategies specifically optimized for this context:

- Multi-caller consensus strategy for small variant detection: ClinBioNGS integrates the output of multiple variant callers to increase sensitivity and robustness, particularly for low-VAF variants and complex InDels. This consensus approach mitigates the limitations of any single tool and reduces the impact of caller-specific artifacts, improving both confidence and reproducibility of variant calls[10,40,41].
- Panel-specific pooled reference baselines for CNA and MSI analysis: Instead of relying on matched normals, ClinBioNGS leverages curated tumor-only reference datasets specific to each panel. These pooled baselines enable effective estimation of CNAs and MSI, allowing accurate detection of somatic events even in tumor-only sequencing data. In benchmarking analyses, these methods achieved performance metrics comparable to, or exceeding, those of commercial pipelines.

Together with the pipeline's comprehensive annotation framework and internal flagging system, these tumor-only strategies ensure a more reliable and transparent classification of tumor-specific alterations in the absence of normal controls. This enables broader applicability of ClinBioNGS in real-world clinical settings, where matched normal samples are rarely available.

5.2.5. Generation of informative plots and interactive reports

One of the key features that distinguishes ClinBioNGS from other existing tools is its ability not only to perform comprehensive analyses, but also to generate dedicated visualizations for each type of result. All outputs are integrated into a fully interactive, self-contained HTML report, which greatly facilitates the interpretation and exploration of complex genomic data.

Unlike most academic pipelines—which often provide only static outputs and limited visualization capacity—ClinBioNGS delivers intuitive, interactive summaries that enhance accessibility. While some commercial platforms offer interactive graphical interfaces, they are typically restricted to proprietary environments with limited flexibility and transparency. By contrast, ClinBioNGS combines interactivity with openness and adaptability, ensuring reproducible, user-friendly reporting across diverse contexts (**Supplementary Table 21**).

In clinical practice, this functionality supports multidisciplinary communication, for instance in MTBs, and ultimately streamlines the clinical decision-making process for each patient.

5.2.6. Modular, portable, and open-source design

ClinBioNGS is built on a modular architecture implemented in Nextflow, with containerization through Singularity images to ensure reproducibility, scalability, and portability across environments ranging from local workstations to HPC systems[36,53,62,73,74]. Its flexible configuration system allows seamless adaptation to diverse experimental and clinical setups, as demonstrated by its ability to analyze multiple panels from both public datasets and routine diagnostic data across institutions.

The open-source availability of ClinBioNGS further enhances transparency and adaptability, enabling laboratories and research groups to inspect, customize, and extend the workflow to their specific needs. Version-controlled profiles and environment-independent execution guarantee consistent results regardless of infrastructure, while fostering collaborative development and alignment with best practices in clinical bioinformatics[34,53,62,65].

By integrating modularity, portability, and open access, ClinBioNGS stands as a robust, institution-independent solution suitable for a wide range of applications—from diagnostic workflows in clinical laboratories to exploratory cancer research—supporting the evolving needs of precision oncology[53].

5.3. Validation and benchmarking performance

The analytical performance of ClinBioNGS was comprehensively evaluated using both standardized public reference datasets and large-scale real-world clinical tumor samples. This dual benchmarking approach enabled a robust assessment of the pipeline's accuracy, reproducibility, and concordance across sequencing platforms, commercial panels, and variant classes. Results confirmed that ClinBioNGS delivers reliable and consistent outputs in both controlled benchmarking and routine clinical contexts, supporting its use in diverse diagnostic and research applications.

5.3.1. High analytical accuracy in SEQC2 reference datasets

Benchmarking with SEQC2 public reference datasets provided a rigorous framework for assessing small variant detection against established ground truth[54,105]. The evaluation included six commercial pan-cancer NGS panels from different vendors, representing a broad spectrum of panel designs and technical challenges. ClinBioNGS achieved consistently high accuracy across all datasets, with performance metrics comparable to vendor-provided pipelines. These results validate the robustness of its small variant calling strategy and demonstrate its capacity to operate as a truly panel-agnostic workflow under standardized conditions.

5.3.2. Robust performance across real-world clinical tumor samples

To complement benchmarking on public datasets, ClinBioNGS was evaluated on retrospective internal cohorts encompassing three commercial NGS panels analyzed across different institutions. This real-world benchmarking provided a comprehensive assessment of performance across variant types and clinical contexts. ClinBioNGS demonstrated high concordance with established commercial pipelines, underscoring its readiness to support routine use.

Observed discrepancies primarily involved borderline variants of limited clinical significance. Nonetheless, certain well-supported small variants required closer attention. Several ClinBioNGS-only calls were located in regions blacklisted by the TSO500 pipeline, while others were omitted by OCA and OPA tools due to their absence from predefined SNV/InDel reporting lists. Although not immediately actionable, some of these variants—including oncogenic or likely oncogenic findings—may acquire clinical value as new evidence emerges. Conversely, some TSO500 commercial-only calls can be attributed to differences in reference genome versions, such as the known failure to detect *U2AF1 S34F* in GRCh38[136], which is used by ClinBioNGS. More broadly, most commercial-only alterations were not truly missed by ClinBioNGS. They were excluded from the comparative because they were associated with calling-related flags (i.e., primary flags). However, these events would still be surfaced for expert review, allowing their borderline nature to be recognized and evaluated in the context of available clinical evidence.

ClinBioNGS identified a large number of CNAs not reported by commercial pipelines, reflecting its broader detection scope. For example, the TSO500 Local App reports amplifications in only 59 of 523 genes, and the OCA pipeline restricts analysis to amplifications, explaining most discrepancies. Among genes assessed by ClinBioNGS and commercial solutions, copy ratio values showed strong correlation, although slight differences in copy-number estimates were observed—likely due to TP adjustments applied by commercial pipelines. ClinBioNGS instead adopts a more conservative strategy, avoiding purity correction to ensure stable results when purity estimates are unavailable or unreliable. While this may lead to CNA underestimation in low-purity samples, users are encouraged to interpret CNA results in the context of tumor cellularity. Future versions will incorporate panel-specific models for purity-adjusted CNA estimation, leveraging our large tumor cohorts.

Regarding biomarker classification, ClinBioNGS demonstrated strong concordance with commercial pipelines for both TMB and MSI status. Most discordant cases occurred near classification thresholds, consistent with the pipeline's conservative mutation filtering and robust MSI estimation calibrated on a broad baseline. In one notable case, an MSI-High classification by ClinBioNGS was later confirmed by repeat testing, illustrating the pipeline's capacity to detect early or borderline biomarker signals not captured by commercial tools.

Overall, this multi-panel benchmarking shows that ClinBioNGS achieves performance comparable to commercial pipelines in routine analyses while providing enhanced sensitivity, a broader detection scope, and improved interpretability in complex or borderline scenarios. Its transparent reporting and robust flagging system ensure that even low-confidence or atypical events are highlighted for expert evaluation—an essential feature in clinical genomics, where seemingly marginal findings can hold diagnostic or therapeutic relevance[4,34,35,42,48].

5.3.3. Extended capabilities in real-world case studies

Beyond benchmarking and quantitative comparisons, a series of real-world clinical and research cases further illustrated the practical value of ClinBioNGS in diverse and challenging scenarios. These cases highlight the pipeline's ability to not only replicate the outputs of commercial pipelines but also to enhance the resolution, sensitivity, and interpretability of complex genomic alterations that may influence diagnostic, prognostic, or therapeutic assessments.

Several cases demonstrated the recovery of clinically relevant variants missed by commercial solutions, either due to overly strict quality filters, blacklisted regions, or incomplete algorithmic support. For example, ClinBioNGS successfully detected a known pathogenic germline mutation in *MSH6* previously validated by a dedicated hereditary panel, which had been filtered out by the commercial pipeline because it fell within a blacklisted region which accumulates recurrent artifacts. Unlike a binary reporting strategy, ClinBioNGS flagged this variant appropriately while retaining it for expert review, thereby demonstrating the advantage of transparent reporting over hard filtering.

In other cases, the multi-caller consensus strategy proved instrumental in resolving complex InDels, particularly *EGFR* exon 19 deletions, which were misclassified or filtered out by commercial software. The ability to disentangle multiallelic or compound InDel events using multiple callers enhanced variant representation and VAF estimation, improving interpretability for therapy selection[10,40,41]. Although the pipeline currently does not retain phasing information across consecutive variants, it provides all variant-level details and BAM files to facilitate manual inspection when required.

The broader CNA detection scope of ClinBioNGS also provided added clinical value. In tumors such as oligodendroglioma and uveal melanoma, ClinBioNGS successfully identified canonical arm-level CNAs that were not reported by vendor pipelines, enabling a more comprehensive molecular characterization. Several of these results were concordant with orthogonal analyses by FISH or shallow WGS, reinforcing the reliability of the CNA module, even for large-scale chromosomal alterations.

Taken together, these case studies showcase the real-world utility of ClinBioNGS in scenarios where data complexity, platform limitations, or algorithmic constraints hinder accurate variant reporting. They emphasize the value of combining high sensitivity with interpretive flexibility, allowing variants of uncertain or borderline confidence to be surfaced with sufficient context for expert review. This approach minimizes the risk of excluding potentially actionable findings and reinforces ClinBioNGS as a valuable asset for both clinical diagnostics and translational research.

5.4. Limitations and current challenges

5.4.1. Inherent challenges in tumor-only somatic NGS panel analysis

Despite the strengths demonstrated by ClinBioNGS, several intrinsic challenges persist in the analysis of tumor-only NGS panel data. These limitations are not specific to this pipeline but rather reflect broader obstacles commonly encountered in clinical bioinformatics and somatic variant interpretation:

- Lack of universal gold standards: The absence of universally accepted reference datasets and standardized workflows for tumor-only analysis complicates benchmarking and harmonization across laboratories[34,54,62]. The wide variety of available tools, databases, and analytical strategies further contributes to variability in implementation and interpretation[35].
- Absence of matched normal tissue and underrepresentation of population diversity: Without access to matched samples, it is difficult to definitively distinguish somatic mutations from germline variants and to effectively suppress panel- or platform-specific artifacts. While population databases assist in filtering common germline variants, rare germline alterations—particularly those in underrepresented populations—may be misclassified as somatic[10,12,41]. Conversely, somatic variants present in these databases can be incorrectly excluded. Such misclassifications may result in diagnostic inaccuracies or inappropriate therapeutic decisions, especially when germline mutations associated with hereditary cancer risk are missed[137].
- Impact of TP on CNA detection: The accuracy of CN estimation is influenced by tumor cellularity, especially for events near detection thresholds where diluted signals may lead to underestimation. This may impact both sensitivity and specificity for detecting low-level AMPs or DELs[10,47].
- Low-VAF variant detection (<1–2%): Identifying variants at very low VAFs remains inherently difficult[9–11,54]. While ClinBioNGS's multi-caller consensus strategy enhances detection robustness, tuning for sensitivity must be carefully balanced against the risk of FPs in a clinical setting.

- Challenges in RNA-based alteration detection: Detection sensitivity for gene fusions and splice variants depends heavily on sequencing depth, read quality, and pre-processing choices such as deduplication[9,10,54,56,57]. Subtle differences in thresholds across pipelines can result in discordant or missed events, particularly for low-abundance transcripts.

ClinBioNGS addresses many of these limitations through a comprehensive annotation and prioritization system, conservative yet informative flagging of uncertain events, and transparent reporting of all detected variants, including those near confidence thresholds. However, certain edge cases will inevitably require expert review within the appropriate clinical context to ensure accurate interpretation, underscoring the indispensable role of multidisciplinary evaluation in MTBs for precision oncology[4,34,35,42,48].

5.4.2. Limitations of the benchmarking approach

While the benchmarking efforts presented in this thesis provided valuable insights into the performance of ClinBioNGS, several important limitations must be acknowledged:

- Limited reference resources for CNA and RNA validation: Unlike small variants, for which high-quality reference datasets such as SEQC2 are available, benchmarking of CNAs and RNA-based events remains challenging due to the lack of publicly accessible gold-standard datasets for somatic NGS panels. This limits the ability to independently assess these genomic alterations under standardized conditions[54,62].
- Constraints in experimental validation: In the clinical routine, the limited availability of tumor tissue restricts opportunities for orthogonal validation of discordant calls[9–11,54].
- Differences in genome reference builds: Several observed discrepancies may also arise from differences in reference genome versions. While ClinBioNGS applies a consistent GRCh38-based reference, the commercial pipelines used for comparison were operating on the older GRCh37 build during the benchmarking period. Such differences can affect variant coordinates, coverage, and mapping quality, particularly in complex genomic regions[9,10,40].
- Pipeline versioning and evolving vendor tools: The benchmarking was conducted using commercial pipeline versions available at the start of this study. However, vendor platforms are continuously evolving, and more recent versions may now offer improved analytical performance. For example, Illumina's updated DRAGEN[138] commercial platform for TSO500 provides broader CNA coverage compared to the Local App version used in this study. Thus, the comparisons reported here represent a temporal snapshot of performance and should be interpreted in that context.

Overall, while these limitations introduce important caveats, the benchmarking results nonetheless provide strong evidence of the robustness and competitiveness of ClinBioNGS across distinct

genomic alterations and platforms. As both ClinBioNGS and commercial solutions continue to evolve, future benchmarking using updated datasets and tools will be essential to further refine performance comparisons and guide pipeline optimization.

5.5. Perspectives for future developments

The modular and open architecture of ClinBioNGS positions it for continued evolution in parallel with advances in sequencing technologies, data interpretation frameworks, and precision oncology workflows. Several promising directions for future development include:

- Expansion of multi-caller strategies to additional variant types: Extending the current consensus approach to CNA and RNA alterations could enhance sensitivity, robustness, and variant confidence scoring, particularly for low-abundance or borderline cases[40,41].
- Continuous enrichment of the annotation framework: Regular updates incorporating emerging curated resources (e.g., ClinGen[135], OncoKB[28], COSMIC[25]) will maintain the clinical and biological relevance of variant interpretations.
- Incorporation of additional scales of actionability: New frameworks, such as ESCAT[59], will provide more integrative clinical prioritization and evidence grading of alterations.
- Adoption of updated genome references: Integration of the T2T-CHM13[44] reference genome may improve alignment accuracy and variant detection in regions poorly represented in GRCh38, particularly for complex or repetitive loci.
- Adoption of new and improved standard file formats: Incorporating support for updated formats such as CRAM, a compressed alternative to BAM, will enable more efficient data storage and management without compromising compatibility with downstream tools[9,43].
- Enhanced somatic-germline discrimination: Adding support for matched tumor-normal pipelines or leveraging panel-specific PoN can improve the classification of rare germline variants and mitigate panel-related artifacts in tumor-only data[9,10,41].
- Automated TP and contamination assessment: Implementing modules for estimating TP and detecting sample contamination will improve QC metrics and support better interpretation of CNAs and low-frequency variants[9–11,54].
- TP-aware CNA recalibration: Providing CN values adjusted by estimated TP could facilitate the interpretation of borderline AMPs or DELs and enhance downstream clinical utility.
- Integration of RNA-based expression analysis: Adding transcript quantification and expression imbalance modules could aid in prioritizing fusions and splice variants based on functional impact and aberrant expression.

- Expanded biomarker profiling: Development of additional modules for emerging biomarkers—such as mutational signatures, chromosomal instability, HRD, and complex rearrangements—will further support translational and research applications[10,40,51,54].
- Development of interactive query tools for variant registries: Building a visual explorer to interactively query, filter, and review aggregated variant data across samples and panels would support institutional audits, cohort-level analyses, and the development of internal genomic knowledge bases.
- Support for additional data types and study designs: Broadening compatibility to WES and WGS sequencing approaches, cfDNA from liquid biopsies, and long-read sequencing platforms (e.g., Oxford Nanopore, PacBio) will extend the pipeline's applicability to a wider range of clinical and research scenarios[5–7,9,10,13,34,39].
- Improved data-sharing interoperability: Incorporation of established standards for secure data exchange (e.g., GA4GH Beacon[65] and related APIs) could facilitate regulated genomic data sharing across institutions, enhancing collaboration and multi-center integration[62,64].

These future directions highlight the capacity of ClinBioNGS to remain a dynamic and sustainable platform, supporting both clinical diagnostics and research-driven precision oncology in an evolving genomic landscape.

5.6. Final remarks

ClinBioNGS emerges as an open-source, comprehensive, and flexible, bioinformatics pipeline tailored to the analytical and interpretative challenges of somatic NGS cancer panels. By integrating DNA and RNA variant detection within a unified workflow—together with standardized annotation, detailed flagging, and informative reporting—it provides a robust solution to support precision oncology in both clinical and translational settings.

Its modular and transparent architecture, built on containerized environments and workflow management systems, ensures reproducibility, portability, and long-term sustainability. The capacity to adapt to diverse sequencing technologies, panel designs, and sample types makes ClinBioNGS widely applicable across institutions, fostering harmonization and standardization in somatic variant analysis.

Extensive validation using public benchmark datasets and large-scale real-world clinical tumor cohorts confirmed the pipeline’s analytical accuracy, reliability, and concordance with commercial solutions. Importantly, ClinBioNGS was able to recover relevant variants and biomarkers overlooked by vendor pipelines, underscoring its potential to enhance diagnostic sensitivity and broaden the detection of clinically meaningful alterations, particularly in complex or borderline cases.

As genomic medicine advances and the complexity of tumor sequencing data increases, tools like ClinBioNGS will be essential to ensure that results are not only accurately generated but also meaningfully interpreted and effectively communicated to clinicians. By bridging the gap between raw sequencing data and clinical insights, ClinBioNGS contributes a scalable and transparent bioinformatics solution to the evolving landscape of precision oncology.

6. CONCLUSIONS

This thesis presents the design, implementation, and evaluation of ClinBioNGS, an open-source, panel-agnostic, and comprehensive bioinformatics pipeline tailored for the analysis of somatic NGS cancer panels. The project was driven by in-depth research to identify and address the key computational challenges involved in tumor-only NGS panel analysis, culminating in a robust solution suitable for both routine diagnostics and translational research applications.

The main conclusions of this thesis are as follows:

1. ClinBioNGS was developed as a portable, open-source workflow that enables standardized and reproducible analysis of somatic NGS panel data across diverse settings.
2. Its modular architecture, implemented using Nextflow and containerized environments, ensures flexibility and portability, enabling panel-agnostic execution across heterogeneous computing infrastructures.
3. The pipeline supports the comprehensive detection of DNA and RNA somatic alterations—including small variants, CNAs, gene fusions, splice variants, and complex biomarkers such as TMB and MSI—within a single unified analytical framework.
4. ClinBioNGS incorporates a multi-caller consensus strategy for small variant detection and uses panel-specific pooled references for the analysis of CNAs and MSI, providing robust performance in tumor-only data.
5. It integrates automated modules for variant annotation and clinical prioritization, leveraging curated databases and established guidelines to support consistent and clinically relevant interpretation.
6. The workflow includes dedicated systems for QC, variant flagging, and custom filtering, ensuring transparency and supporting expert review of potentially actionable findings.
7. ClinBioNGS integrates a local variant database module to store and retrieve detected variants across analyses, facilitating longitudinal tracking and reuse of internal knowledge.
8. The pipeline results are delivered through self-contained, interactive HTML reports that combine intuitive visualizations and comprehensive tables, enhancing interpretability and facilitating use in multidisciplinary settings.
9. All annotations are directly accessible through the interactive HTML report, empowering users to evaluate variants in detail within a clinically oriented context.

10. Validation with SEQC2 multi-panel public reference datasets confirmed high accuracy in small variant detection, supporting the pipeline's analytical reliability.

11. Benchmarking on retrospective, real-world clinical datasets from multiple institutions and commercial NGS panels showed that ClinBioNGS performs comparably to existing commercial solutions, while offering broader detection capabilities and greater interpretability, particularly in complex or ambiguous cases.

12. Real-world case studies further demonstrated the pipeline's adaptability across both diagnostic and research applications, highlighting its added value in resolving challenging variants and improving the resolution to assess CNAs.

BIBLIOGRAPHY

1. The Genetics of Cancer - NCI. Accessed May 28, 2025. https://www.cancer.gov/about-cancer/causes-prevention/genetics
2. Swanton C, Bernard E, Abbosh C, et al. Embracing cancer complexity: Hallmarks of systemic disease. *Cell*. 2024;187(7):1589-1616. doi:10.1016/j.cell.2024.02.009
3. Solís-Moruno M, Batlle-Masó L, Bonet N, Aróstegui JI, Casals F. Somatic genetic variation in healthy tissue and non-cancer diseases. *European Journal of Human Genetics*. 2023;31(1):48-54. doi:10.1038/s41431-022-01213-8
4. Li MM, Datto M, Duncavage EJ, et al. Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer. *The Journal of Molecular Diagnostics*. 2017;19(1):4-23. doi:10.1016/j.jmoldx.2016.10.002
5. Malone ER, Oliva M, Sabatini PJB, Stockley TL, Siu LL. Molecular profiling for precision cancer therapies. *Genome Med*. 2020;12(1):8. doi:10.1186/s13073-019-0703-1
6. Rulten SL, Grose RP, Gatz SA, Jones JL, Cameron AJM. The Future of Precision Oncology. *Int J Mol Sci*. 2023;24(16):12613. doi:10.3390/ijms241612613
7. Nagahashi M, Shimada Y, Ichikawa H, et al. Next generation sequencing-based gene panel tests for the management of solid tumors. *Cancer Sci*. 2019;110(1):6-15. doi:10.1111/cas.13837
8. Colomer R, Mondejar R, Romero-Laorden N, Alfranca A, Sanchez-Madrid F, Quintela-Fandino M. When should we order a next generation sequencing test in a patient with cancer? *EClinicalMedicine*. 2020;25:100487. doi:10.1016/j.eclinm.2020.100487
9. Larson NB, Oberg AL, Adjei AA, Wang L. A Clinician's Guide to Bioinformatics for Next-Generation Sequencing. *Journal of Thoracic Oncology*. 2023;18(2):143-157. doi:10.1016/j.jtho.2022.11.006
10. Cortés-Ciriano I, Gulhan DC, Lee JJK, Melloni GEM, Park PJ. Computational analysis of cancer genome sequencing data. *Nat Rev Genet*. 2022;23(5):298-314. doi:10.1038/s41576-021-00431-y
11. Jennings LJ, Arcila ME, Corless C, et al. Guidelines for Validation of Next-Generation Sequencing–Based Oncology Panels. *The Journal of Molecular Diagnostics*. 2017;19(3):341-365. doi:10.1016/j.jmoldx.2017.01.011

12. Bewicke-Copley F, Arjun Kumar E, Palladino G, Korfi K, Wang J. Applications and analysis of targeted genomic sequencing in cancer studies. *Comput Struct Biotechnol J*. 2019;17:1348-1359. doi:10.1016/j.csbj.2019.10.004

13. Pei XM, Yeung MHY, Wong ANN, et al. Targeted Sequencing Approach and Its Clinical Applications for the Molecular Diagnosis of Human Diseases. *Cells*. 2023;12(3):493. doi:10.3390/cells12030493

14. Krysiak K, Danos AM, Saliba J, et al. CIViCdb 2022: evolution of an open-access cancer variant interpretation knowledgebase. *Nucleic Acids Res*. 2023;51(D1):D1230-D1241. doi:10.1093/nar/gkac979

15. Shimozaki K, Hayashi H, Tanishima S, et al. Concordance analysis of microsatellite instability status between polymerase chain reaction based testing and next generation sequencing for solid tumors. *Sci Rep*. 2021;11(1):20003. doi:10.1038/s41598-021-99364-z

16. Yaacov A, Ben Cohen G, Landau J, Hope T, Simon I, Rosenberg S. Cancer mutational signatures identification in clinical assays using neural embedding-based representations. *Cell Rep Med*. 2024;5(6):101608. doi:10.1016/J.XCRM.2024.101608

17. Bradley RK, Anczuków O. RNA splicing dysregulation and the hallmarks of cancer. *Nat Rev Cancer*. 2023;23(3):135-155. doi:10.1038/s41568-022-00541-7

18. Mosteiro M, Azuara D, Villatoro S, et al. Molecular profiling and feasibility using a comprehensive hybrid capture panel on a consecutive series of non-small-cell lung cancer patients from a single centre. *ESMO Open*. 2023;8(6):102197. doi:10.1016/j.esmoop.2023.102197

19. Takeda M, Takahama T, Sakai K, et al. Clinical Application of the FoundationOne CDx Assay to Therapeutic Decision-Making for Patients with Advanced Solid Tumors. *Oncologist*. 2021;26(4):e588-e596. doi:10.1002/onco.13639

20. Cheng DT, Mitchell TN, Zehir A, et al. Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT). *The Journal of Molecular Diagnostics*. 2015;17(3):251-264. doi:10.1016/j.jmoldx.2014.12.006

21. Weinstein JN, Collisson EA, Mills GB, et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet*. 2013;45(10):1113-1120. doi:10.1038/ng.2764

22. Hudson TJ, Anderson W, Aretz A, et al. International network of cancer genome projects. *Nature*. 2010;464(7291):993-998. doi:10.1038/nature08987

23. Auton A, Abecasis GR, Altshuler DM, et al. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74. doi:10.1038/nature15393

24. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 2020;581(7809):434-443. doi:10.1038/s41586-020-2308-7

25. Tate JG, Bamford S, Jubb HC, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res*. 2019;47(D1):D941-D947. doi:10.1093/nar/gky1015

26. Pugh TJ, Bell JL, Bruce JP, et al. AACR Project GENIE: 100,000 Cases and Beyond. *Cancer Discov*. 2022;12(9):2044-2057. doi:10.1158/2159-8290.CD-21-1547

27. Landrum MJ, Lee JM, Riley GR, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res*. 2014;42(D1):D980-D985. doi:10.1093/nar/gkt1113

28. Chakravarty D, Gao J, Phillips S, et al. OncoKB: A Precision Oncology Knowledge Base. *JCO Precis Oncol*. 2017;1(1):1-16. doi:10.1200/PO.17.00011

29. Griffith M, Spies NC, Krysiak K, et al. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nat Genet*. 2017;49(2):170-174. doi:10.1038/ng.3774

30. den Dunnen JT, Dalgleish R, Maglott DR, et al. HGVS Recommendations for the Description of Sequence Variants: 2016 Update. *Hum Mutat*. 2016;37(6):564-569. doi:10.1002/humu.22981

31. Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine*. 2015;17(5):405-424. doi:10.1038/gim.2015.30

32. van de Haar J, Roepman P, Andre F, et al. ESMO Recommendations on clinical reporting of genomic test results for solid cancers. *Annals of Oncology*. 2024;35(11):954-967. doi:10.1016/j.annonc.2024.06.018

33. Luh F, Yen Y. FDA guidance for next generation sequencing-based testing: balancing regulation and innovation in precision medicine. *NPJ Genom Med*. 2018;3(1):28. doi:10.1038/s41525-018-0067-2

34. Pallocca M, Betti M, Baldinelli S, et al. Clinical bioinformatics desiderata for molecular tumor boards. *Brief Bioinform*. 2024;25(5). doi:10.1093/bib/bbae447

35. Wagner AH, Walsh B, Mayfield G, et al. A harmonized meta-knowledgebase of clinical interpretations of somatic genomic variants in cancer. *Nat Genet*. 2020;52(4):448-457. doi:10.1038/s41588-020-0603-8

36. Cabello-Aguilar S, Vendrell JA, Solassol J. A Bioinformatics Toolkit for Next-Generation Sequencing in Clinical Oncology. *Curr Issues Mol Biol*. 2023;45(12):9737-9752. doi:10.3390/cimb45120608

37. Milbury CA, Creeden J, Yip WK, et al. Clinical and analytical validation of FoundationOne®CDx, a comprehensive genomic profiling assay for solid tumors. Vousden G, ed. *PLoS One*. 2022;17(3):e0264138. doi:10.1371/journal.pone.0264138

38. The Variant Call Format Specification. Published online 2024. https://github.com/samtools/hts-specs.

39. Satam H, Joshi K, Mangrolia U, et al. Next-Generation Sequencing Technology: Current Trends and Advancements. *Biology (Basel)*. 2023;12(7):997. doi:10.3390/biology12070997

40. Dotolo S, Esposito Abate R, Roma C, et al. Bioinformatics: From NGS Data to Biological Complexity in Variant Detection and Oncological Clinical Practice. *Biomedicines*. 2022;10(9):2074. doi:10.3390/biomedicines10092074

41. Koboldt DC. Best practices for variant calling in clinical sequencing. *Genome Med*. 2020;12(1):91. doi:10.1186/s13073-020-00791-w

42. Roy S, Coldren C, Karunamurthy A, et al. Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines. *The Journal of Molecular Diagnostics*. 2018;20(1):4-27. doi:10.1016/j.jmoldx.2017.11.003

43. Shokrof M, Abouelhoda M. IonCRAM: a reference-based compression tool for ion torrent sequence files. *BMC Bioinformatics*. 2020;21(1):397. doi:10.1186/s12859-020-03726-9

44. Nurk S, Koren S, Rhie A, et al. The complete sequence of a human genome. *Science (1979)*. 2022;376(6588):44-53. doi:10.1126/science.abj6987

45. Liao WW, Asri M, Ebler J, et al. A draft human pangenome reference. *Nature*. 2023;617(7960):312-324. doi:10.1038/s41586-023-05896-x

46. Sequence Alignment/Map Format Specification. Published online 2024. https://github.com/samtools/hts-specs.

47. Talevich E, Shain AH, Botton T, Bastian BC. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS Comput Biol*. 2016;12(4):e1004873. doi:10.1371/journal.pcbi.1004873

48. Horak P, Griffith M, Danos AM, et al. Standards for the classification of pathogenicity of somatic variants in cancer (oncogenicity): Joint recommendations of Clinical Genome Resource (ClinGen), Cancer Genomics Consortium (CGC), and Variant Interpretation for

Cancer Consortium (VICC). *Genetics in Medicine*. 2022;24(5):986-998. doi:10.1016/j.gim.2022.01.001

49. Mateo J, Chakravarty D, Dienstmann R, et al. A framework to rank genomic alterations as targets for cancer precision medicine: the ESMO Scale for Clinical Actionability of molecular Targets (ESCAT). *Annals of Oncology*. 2018;29(9):1895-1902. doi:10.1093/annonc/mdy263

50. Vega DM, Yee LM, McShane LM, et al. Aligning tumor mutational burden (TMB) quantification across diagnostic platforms: phase II of the Friends of Cancer Research TMB Harmonization Project. *Annals of Oncology*. 2021;32(12):1626-1636. doi:10.1016/j.annonc.2021.09.016

51. Witz A, Dardare J, Betz M, et al. Homologous recombination deficiency (HRD) testing landscape: clinical applications and technical validation for routine diagnostics. *Biomark Res*. 2025;13(1):31. doi:10.1186/s40364-025-00740-y

52. Reiter T, Brooks† PT, Irber† L, et al. Streamlining data-intensive biology with workflow systems. *Gigascience*. 2021;10(1). doi:10.1093/gigascience/giaa140

53. Ewels PA, Peltzer A, Fillinger S, et al. The nf-core framework for community-curated bioinformatics pipelines. *Nat Biotechnol*. 2020;38(3):276-278. doi:10.1038/s41587-020-0439-x

54. Mercer TR, Xu J, Mason CE, Tong W. The Sequencing Quality Control 2 study: establishing community standards for sequencing in precision medicine. *Genome Biol*. 2021;22(1):306. doi:10.1186/s13059-021-02528-3

55. Dwarshuis N, Kalra D, McDaniel J, et al. The GIAB genomic stratifications resource for human reference genomes. *Nat Commun*. 2024;15(1):9029. doi:10.1038/s41467-024-53260-y

56. Dube S, Al-Mannai S, Liu L, et al. Systematic comparison of quantity and quality of RNA recovered with commercial FFPE tissue extraction kits. *J Transl Med*. 2025;23(1):1-11. doi:10.1186/S12967-024-05890-5

57. Haas BJ, Dobin A, Ghandi M, et al. Targeted in silico characterization of fusion transcripts in tumor and normal tissues via FusionInspector. *Cell Reports Methods*. 2023;3(5):100467. doi:10.1016/j.crmeth.2023.100467

58. Tamborero D, Rubio-Perez C, Deu-Pons J, et al. Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Med*. 2018;10(1):25. doi:10.1186/s13073-018-0531-8

59. Wolff L, Kiesewetter B. Applicability of ESMO-MCBS and ESCAT for molecular tumor boards. *memo - Magazine of European Medical Oncology*. 2022;15(3):190-195. doi:10.1007/s12254-022-00800-1

60. Tan A, Abecasis GR, Kang HM. Unified representation of genetic variants. *Bioinformatics*. 2015;31(13):2202-2204. doi:10.1093/bioinformatics/btv112

61. Yuen D, Cabansay L, Duncan A, et al. The Dockstore: enhancing a community platform for sharing reproducible and accessible computational protocols. *Nucleic Acids Res*. 2021;49(W1):W624-W632. doi:10.1093/nar/gkab346

62. Rehm HL, Page AJH, Smith L, et al. GA4GH: International policies and standards for data sharing across genomic research and healthcare. *Cell Genomics*. 2021;1(2):100029. doi:10.1016/j.xgen.2021.100029

63. Lappalainen I, Almeida-King J, Kumanduri V, et al. The European Genome-phenome Archive of human data consented for biomedical research. *Nat Genet*. 2015;47(7):692-695. doi:10.1038/ng.3312

64. Thorogood A, Rehm HL, Goodhand P, et al. International federation of genomic medicine databases using GA4GH standards. *Cell Genomics*. 2021;1(2):100032. doi:10.1016/j.xgen.2021.100032

65. Fiume M, Cupak M, Keenan S, et al. Federated discovery and sharing of genomic data using Beacons. *Nat Biotechnol*. 2019;37(3):220-224. doi:10.1038/s41587-019-0046-x

66. Garcia EP, Minkovsky A, Jia Y, et al. Validation of OncoPanel: A Targeted Next-Generation Sequencing Assay for the Detection of Somatic Variants in Cancer. *Arch Pathol Lab Med*. 2017;141(6):751-758. doi:10.5858/arpa.2016-0527-OA

67. Al-Kateb H, Knight SM, Sivasankaran G, et al. Clinical Validation of the TruSight Oncology 500 Assay for the Detection and Reporting of Pan-Cancer Biomarkers. *The Journal of Molecular Diagnostics*. 2025;27(4):292-305. doi:10.1016/j.jmoldx.2025.01.002

68. Hanssen F, Garcia MU, Folkersen L, et al. Scalable and efficient DNA sequencing analysis on different compute infrastructures aiding variant discovery. *NAR Genom Bioinform*. 2024;6(2):31. doi:10.1093/nargab/lqae031

69. Del Corvo M, Mazzara S, Pileri SA. TOSCA: an automated Tumor Only Somatic CAlling workflow for somatic mutation detection without matched normal samples. Stamatakis A, ed. *Bioinformatics Advances*. 2022;2(1):vbac070. doi:10.1093/bioadv/vbac070

70. Marriott H, Kabiljo R, Al Khleifat A, Dobson RJ, Al-Chalabi A, Iacoangeli A. DNAscan2: a versatile, scalable, and user-friendly analysis pipeline for human next-generation sequencing data. Alkan C, ed. *Bioinformatics*. 2023;39(4). doi:10.1093/bioinformatics/btad152

71. Raulerson CK, Villa EC, Mathews JA, et al. SCHOOL: Software for Clinical Health in Oncology for Omics Laboratories. *J Pathol Inform*. 2022;13(1):100163. doi:10.4103/jpi.jpi_20_21

72. Metzger P, Hess ME, Blaumeiser A, et al. MIRACUM-Pipe: An Adaptable Pipeline for Next-Generation Sequencing Analysis, Reporting, and Visualization for Clinical Decision Making. *Cancers (Basel)*. 2023;15(13):3456. doi:10.3390/cancers15133456

73. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. *Nat Biotechnol*. 2017;35(4):316-319. doi:10.1038/nbt.3820

74. Kurtzer GM, Sochat V, Bauer MW. Singularity: Scientific containers for mobility of compute. Gursoy A, ed. *PLoS One*. 2017;12(5):e0177459. doi:10.1371/journal.pone.0177459

75. Mose LE, Perou CM, Parker JS. Improved indel detection in DNA and RNA via realignment with ABRA2. Stegle O, ed. *Bioinformatics*. 2019;35(17):2966-2973. doi:10.1093/bioinformatics/btz033

76. Spurr LF, Touat M, Taylor AM, et al. Quantification of aneuploidy in targeted sequencing data using ASCETS. Robinson P, ed. *Bioinformatics*. 2021;37(16):2461-2463. doi:10.1093/bioinformatics/btaa980

77. iGenomes: Ready-To-Use Reference Sequences and Annotations. Accessed May 28, 2025. https://emea.support.illumina.com/sequencing/sequencing_software/igenome.html

78. Danecek P, Bonfield JK, Liddle J, et al. Twelve years of SAMtools and BCFtools. *Gigascience*. 2021;10(2):1-4. doi:10.1093/gigascience/giab008

79. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26(6):841-842. doi:10.1093/bioinformatics/btq033

80. lh3/bioawk: BWK awk modified for biological data. Accessed May 28, 2025. https://github.com/lh3/bioawk

81. Md V, Misra S, Li H, Aluru S. Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. *IEEE Parallel and Distributed Processing Symposium (IPDPS)*. Published online July 27, 2019. doi:10.48550/arXiv.1907.12931

82. TrinityCTAT/Trinity_CTAT: Trinity Cancer Transcriptome Analysis Toolkit (CTAT). https://github.com/TrinityCTAT/Trinity_CTAT

83. McLaren W, Gil L, Hunt SE, et al. The Ensembl Variant Effect Predictor. *Genome Biol*. 2016;17(1):122. doi:10.1186/s13059-016-0974-4

84. Chen S. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *iMeta*. 2023;2(2):e107. doi:10.1002/imt2.107

85. Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010. http://www.bioinformatics.babraham.ac.uk/projects/fastqc

86. Van der Auwera G, O'Connor B. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. O'Reilly Media; 2020.

87. Chen S, Zhou Y, Chen Y, et al. Gencore: an efficient tool to generate consensus reads for error suppressing and duplicate removing of NGS data. *BMC Bioinformatics*. 2019;20(S23):606. doi:10.1186/s12859-019-3280-9

88. Pedersen BS, Quinlan AR. Mosdepth: quick coverage calculation for genomes and exomes. Hancock J, ed. *Bioinformatics*. 2018;34(5):867-868. doi:10.1093/bioinformatics/btx699

89. Jia P, Yang X, Guo L, et al. MSIsensor-Pro: Fast, Accurate, and Matched-Normal-Sample-Free Detection of Microsatellite Instability. *Genomics Proteomics Bioinformatics*. 2020;18(1):65-71. doi:10.1016/j.gpb.2020.02.001

90. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*. 2016;32(19):3047-3048. doi:10.1093/bioinformatics/btw354

91. Cooke DP, Wedge DC, Lunter G. A unified haplotype-based method for accurate and comprehensive variant calling. *Nat Biotechnol*. 2021;39(7):885-892. doi:10.1038/s41587-021-00861-3

92. Dunn T, Berry G, Emig-Agius D, et al. Pisces: an accurate and versatile variant caller for somatic and germline next-generation sequencing data. Schwartz R, ed. *Bioinformatics*. 2019;35(9):1579-1581. doi:10.1093/bioinformatics/bty849

93. R Core Team. R: A Language and Environment for Statistical Computing. Preprint posted online 2024. https://www.R-project.org/

94. Haas BJ, Dobin A, Stransky N, et al. STAR-Fusion: Fast and Accurate Fusion Transcript Detection from RNA-Seq. *bioRxiv.Cold Spring Harbor Laboratory*. Preprint posted online March 24, 2017:120295. doi:10.1101/120295

95. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15-21. doi:10.1093/bioinformatics/bts635

96. iontorrent/TS: Torrent Suite. https://github.com/iontorrent/TS

97. Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics*. 2010;26(17):2204-2207. doi:10.1093/bioinformatics/btq351

98. Perez G, Barber GP, Benet-Pages A, et al. The UCSC Genome Browser database: 2025 update. *Nucleic Acids Res*. 2025;53(D1):D1243-D1249. doi:10.1093/nar/gkae974

99. Smith T, Heger A, Sudbery I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res*. 2017;27(3):491-499. doi:10.1101/gr.209601.116

100. SciLifeLab/umi-transfer: A Rust software to extract Unique Molecular Identifiers in FastQ files and embedd them into the ID of the corresponding reads. https://github.com/SciLifeLab/umi-transfer

101. Lai Z, Markovets A, Ahdesmaki M, et al. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res*. 2016;44(11):e108-e108. doi:10.1093/nar/gkw227

102. Zentgraf J, Rahmann S. Fast lightweight accurate xenograft sorting. *Algorithms for Molecular Biology*. 2021;16(1):2. doi:10.1186/s13015-021-00181-w

103. Karolchik D, Hinricks AS, Furey TS, et al. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res*. 2004;32(Database issue):D493. doi:10.1093/NAR/GKH103

104. Morales J, Pujar S, Loveland JE, et al. A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature*. 2022;604(7905):310-315. doi:10.1038/s41586-022-04558-8

105. Gong B, Li D, Kusko R, et al. Cross-oncopanel study reveals high sensitivity and accuracy with overall analytical performance depending on genomic regions. *Genome Biol*. 2021;22(1):109. doi:10.1186/s13059-021-02315-0

106. Dressler L, Bortolomeazzi M, Keddar MR, et al. Comparative assessment of genes driving cancer and somatic evolution in non-cancer tissues: an update of the Network of Cancer Genes (NCG) resource. *Genome Biol*. 2022;23(1):35. doi:10.1186/s13059-022-02607-z

107. Ioannidis NM, Rothstein JH, Pejaver V, et al. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *The American Journal of Human Genetics*. 2016;99(4):877-885. doi:10.1016/j.ajhg.2016.08.016

108. Cheng J, Novati G, Pan J, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science (1979)*. 2023;381(6664). doi:10.1126/science.adg7492

109. Landrum MJ, Lee JM, Benson M, et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res*. 2018;46(D1):D1062-D1067. doi:10.1093/nar/gkx1153

110. Chang MT, Bhattarai TS, Schram AM, et al. Accelerating Discovery of Functional Mutant Alleles in Cancer. *Cancer Discov*. 2018;8(2):174-183. doi:10.1158/2159-8290.CD-17-0321

111. Jones W, Gong B, Novoradovskaya N, et al. A verified genomic reference sample for assessing performance of cancer panels detecting small variants of low allele frequency. *Genome Biol*. 2021;22(1):111. doi:10.1186/s13059-021-02316-z

112. Mitelman F, Johansson B, Mertens F. Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer (2025). https://mitelmandatabase.isb-cgc.org

113. Sabir S, Yeoh S, Jackson G, Bayliss R. EML4-ALK Variants: Biological and Molecular Properties, and the Implications for Patients. *Cancers (Basel)*. 2017;9(9):118. doi:10.3390/cancers9090118

114. Sasaki T, Rodig SJ, Chirieac LR, Jänne PA. The biology and treatment of EML4-ALK non-small cell lung cancer. *Eur J Cancer*. 2010;46(10):1773-1780. doi:10.1016/j.ejca.2010.04.002

115. Lin JJ, Zhu VW, Yoda S, et al. Impact of *EML4-ALK* Variant on Resistance Mechanisms and Clinical Outcomes in *ALK*-Positive Lung Cancer. *Journal of Clinical Oncology*. 2018;36(12):1199-1206. doi:10.1200/JCO.2017.76.2294

116. Deng K, Yao J, Huang J, Ding Y, Zuo J. Abnormal alternative splicing promotes tumor resistance in targeted therapy and immunotherapy. *Transl Oncol*. 2021;14(6):101077. doi:10.1016/j.tranon.2021.101077

117. Clark ME, Rizos H, Pereira MR, et al. Detection of *BRAF* splicing variants in plasma-derived cell-free nucleic acids and extracellular vesicles of melanoma patients failing targeted therapy therapies. *Oncotarget*. 2020;11(44):4016-4027. doi:10.18632/oncotarget.27790

118. Sowalsky AG, Figueiredo I, Lis RT, et al. Assessment of Androgen Receptor Splice Variant-7 as a Biomarker of Clinical Response in Castration-Sensitive Prostate Cancer. *Clinical Cancer Research*. 2022;28(16):3509-3525. doi:10.1158/1078-0432.CCR-22-0851

119. Kanayama M, Lu C, Luo J, Antonarakis ES. AR Splicing Variants and Resistance to AR Targeting Agents. *Cancers (Basel)*. 2021;13(11):2563. doi:10.3390/cancers13112563

120. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*. 2011;27(12):1739-1740. doi:10.1093/bioinformatics/btr260

121. Schriml LM, Mitraka E, Munro J, et al. Human Disease Ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids Res*. 2019;47(D1):D955-D962. doi:10.1093/nar/gky1032

122. Kundra R, Zhang H, Sheridan R, et al. OncoTree: A Cancer Classification System for Precision Oncology. *JCO Clin Cancer Inform*. 2021;(5):221-230. doi:10.1200/CCI.20.00108

123. Schubach M, Maass T, Nazaretyan L, Röner S, Kircher M. CADD v1.7: using protein language models, regulatory CNNs and other nucleotide-level scores to improve genome-wide variant predictions. *Nucleic Acids Res*. 2024;52(D1):D1143-D1154. doi:10.1093/nar/gkad989

124. R Special Interest Group on Databases (R-SIG-DB), Wickham H, Müller K. DBI: R Database Interface. Preprint posted online 2024. https://dbi.r-dbi.org

125. Gel B, Serra E. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. Hancock J, ed. *Bioinformatics*. 2017;33(19):3088-3090. doi:10.1093/bioinformatics/btx346

126. Garrick Aden-Buie, Carson Sievert, Richard Iannone, JJ Allaire, Barbara Borges. flexdashboard: R Markdown Format for Flexible Dashboards. Preprint posted online 2024. https://pkgs.rstudio.com/flexdashboard/

127. Benjamin D, Sato T, Cibulskis K, Getz G, Stewart C, Lichtenstein L. Calling Somatic SNVs and Indels with Mutect2. *bioRxiv*.*bioRxiv*. Preprint posted online December 2, 2019:861054. doi:10.1101/861054

128. da Veiga Leprevost F, Grüning BA, Alves Aflitos S, et al. BioContainers: an open-source and community-driven framework for software standardization. Valencia A, ed. *Bioinformatics*. 2017;33(16):2580-2582. doi:10.1093/bioinformatics/btx192

129. Abueg LAL, Afgan E, Allart O, et al. The Galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update. *Nucleic Acids Res*. 2024;52(W1):W83-W94. doi:10.1093/nar/gkae410

130. Seal RL, Braschi B, Gray K, et al. Genenames.org: the HGNC resources in 2023. *Nucleic Acids Res*. 2023;51(D1):D1003-D1009. doi:10.1093/nar/gkac888

131. Kroeze LI, de Voer RM, Kamping EJ, et al. Evaluation of a Hybrid Capture–Based Pan-Cancer Panel for Analysis of Treatment Stratifying Oncogenic Aberrations and Processes. *The Journal of Molecular Diagnostics*. 2020;22(6):757-769. doi:10.1016/j.jmoldx.2020.02.009

132. Homo sapiens (ID 677997) - BioProject - NCBI. https://www.ncbi.nlm.nih.gov/bioproject/PRJNA677997

133. Johansson PA, Brooks K, Newell F, et al. Whole genome landscapes of uveal melanoma show an ultraviolet radiation signature in iris tumours. *Nat Commun*. 2020;11(1):2408. doi:10.1038/s41467-020-16276-8

134. Paik PK, Felip E, Veillon R, et al. Tepotinib in Non–Small-Cell Lung Cancer with *MET* Exon 14 Skipping Mutations. *New England Journal of Medicine*. 2020;383(10):931-943. doi:10.1056/NEJMoa2004407

135. Rehm HL, Berg JS, Brooks LD, et al. ClinGen — The Clinical Genome Resource. *New England Journal of Medicine*. 2015;372(23):2235-2242. doi:10.1056/NEJMsr1406261

136. Miller CA, Walker JR, Jensen TL, et al. Failure to Detect Mutations in U2AF1 due to Changes in the GRCh38 Reference Sequence. *The Journal of Molecular Diagnostics*. 2022;24(3):219-223. doi:10.1016/j.jmoldx.2021.10.013

137. Pollard RD, Wilkerson MD, Rajagopal PS. Identification of germline population variants misclassified as cancer-associated somatic variants. *Front Med (Lausanne)*. 2024;11:1361317. doi:10.3389/fmed.2024.1361317

138. Behera S, Catreux S, Rossi M, et al. Comprehensive genome analysis and variant detection at scale using DRAGEN. *Nat Biotechnol*. 2025;43(7):1177-1191. doi:10.1038/s41587-024-02382-1

APPENDIX

A1. Supplementary Tables

Supplementary Table 1. Software tools required by ClinBioNGS.
Each entry includes the tool name, version, and its role within the pipeline. Resources are listed in alphabetical order.

Tool	Version	Role in ClinBioNGS
ABRA2[75]	2.24	Realignment of DNA reads around target regions
ASCETS[76]	1.1.2	Estimate arm-level CNAs from inferred segment copy ratio
AWS CLI	2.15.32	Download of reference genomes from the AWS iGenomes repository[77]
Bcftools[78]	1.16	VCF processing and filtering
BCL Convert™	4.0.3	Conversion of BCL to FASTQ
Bedtools[79]	2.31.0	Processing and manipulation of BED files
Bioawk[80]	1	Text manipulation for biological data
BWA-MEM2[81]	2.2.1	Alignment of DNA reads to the reference genome
CNVkit[47]	0.9.9	CNA analysis from DNA reads
CTAT-splicing[82]	0.0.2	Detection and annotation of splicing variants from RNA data
Ensembl VEP[83]	113	Annotation of small variants
FastP[84]	0.23.4	Pre-processing and quality filtering of FASTQ files
FastQC[85]	0.12.1	FASTQ QC
GATK4[86]	4.3.0.0	File manipulation, deduplication (MarkDuplicates), small variant calling (Mutect2[127]), and BAM QC
Gencore[87]	0.17.2	UMI-aware deduplication for paired-end reads
Mosdepth[88]	0.3.3	Per-base and region-level read coverage calculation
MSIsensor-pro[89]	1.2.0	MSI detection from DNA reads
MultiQC[90]	1.22.3	Aggregation and visualization of QC metrics
Octopus[91]	0.7.4	Small variant calling from DNA reads
Pisces[92]	5.3.0.0	Small variant calling from DNA reads
R[93]	4.0.4	Data manipulation, statistical analysis, and generation of tables, plots, and reports
Samtools[78]	1.18	Processing and manipulation of BAM and FASTQ files
STAR-Fusion[94,95]	1.13.0	Detection of fusion transcripts from RNA data
TMAP / TVC[96]	5.12.1	Alignment and small variant calling for Ion Torrent DNA reads
UCSC bigBedToBed[97]	377	Conversion of bigBed files to BED format
UCSC liftOver[98]	377	Conversion of genomic coordinates between genome builds (e.g., hg19 to hg38)
UMI-tools[99]	1.1.2	UMI-aware deduplication for single-end reads
UMI-transfer[100]	1.0.0	Transfer of UMI information from separate FASTQ files into read headers
VarDict[101]	1.8.3	Small variant calling from DNA reads
Vt[60]	0.57721	VCF decomposition and normalization of indels
Xengsort[102]	1.1.0	Filtering of mouse-derived reads in xenograft sequencing data

Supplementary Table 2. External resources required by ClinBioNGS.
Each entry includes the resource name, version (if applicable), and its role within the pipeline. Resources are grouped by category for clarity.

Category	Resource	Version	Role in ClinBioNGS
User-defined metadata files	Sample sheet	-	Specify DNA and RNA samples
	SampleInfo.csv	-	Provide metadata (e.g., sex, tumor, purity, DOID)
	TumorNames.csv	-	Define specific tumor names and associated DOIDs[121,122]
	WhitelistGenes.csv	-	Define tumor-specific gene lists for prioritization
Genome resources	Human reference genome[77]	GRCh38	Define reference genome for DNA analysis
	BWA-MEM2 indexed genome	-	Enable DNA alignment via BWA-MEM2[81]
	TMAP indexed genome	-	Enable DNA alignment via TMAP[96]
	Mouse reference genome[77]	GRCm38	Enable mouse read filtering
	Xengsort indexed genome	-	Support Xengsort-based mouse filtering[102]
	UCSC cytoband[103]	hg38	Annotate gene cytobands
	UCSC arm coordinates	hg38	Support CNA arm-level annotations
	UCSC hg38 / hg19 liftover chain[98]	-	Convert genomic coordinates
MANE[104] annotation files	MANE GTF	1.4	Define MANE transcript structures
	MANE summary	-	List MANE transcript-gene associations
	MANE genes / exons / coding / introns	-	Annotate regions based on MANE
Target region files	TSO500 / SEQC2[105] raw manifests	hg19	Define panel target regions
	Target 4-column BED	hg38	Define final analysis regions
	Target genes	-	Annotate panel-specific gene content
	Extended BED / Interval list	hg38	Apply padding for small variant calling
	Off-target BED	hg38	Support BAM clipping in amplicon panels
	Target chromosomes	-	Enable per-chromosome variant calling
VCF headers	VCF headers for consensus, annotation, CNA, fusion, splicing	4.2	Standardize output formatting
Gene role and oncogenicity	NCG[106] resource file	1.7	Annotate oncogenes and tumor suppressors
	Catalog of Validated Oncogenic Mutations[58]	20180130	List of validated oncogenic variants
	CIViC[14] oncogenic evidence	01-Nov-2024	Support for oncogenic classification
	GENIE[26] oncogenic mutations	16.1	Previously classified oncogenic variants
	ClinGen/CGC/VICC[48] oncogenic	-	Previously classified oncogenic variants
VEP-related resources	VEP cache[83]	113	Reference data for VEP annotation
	gnomAD[24]	4.1	Population allele frequencies
	CADD SNVs/InDels[123]	1.7	Pathogenicity prediction
	REVEL[107]	1.3	Pathogenicity prediction
	AlphaMissense[108]	hg38	Pathogenicity prediction
	ClinVar[109] VCF	20241103	Pathogenicity annotations
	CIViC[14] accepted VCF	01-Nov-2024	Clinical evidence annotations
Cancer hotspots	Panel-specific hotspot BED	hg38	Define user-specific hotspot regions
	GENIE[26] whitelist BED	hg19	Define known somatic hotspots
	Cancer Hotspots[110]	V2	Define statistically enriched mutations

Problematic and high-confidence regions	Panel-specific blacklist BED	hg38	User-defined problematic regions
	UCSC problematic regions[103]	20240606	Define problematic regions
	GIAB stratification BED[55]	3.5	Define problematic regions
	CTR regions[26,111]	hg38	Define high-confidence callable regions
GENIE[26]	GENIE mutation, CNA, fusion data	16.1	Annotate variant recurrence in cancer
Clinical evidence (CIViC)[14]	CIViC variant summaries (raw)	01-Nov-2024	List of CIViC variants
	CIViC molecular profiles (raw)	01-Nov-2024	Variant-to-evidence mapping
	CIViC clinical evidence (raw)	01-Nov-2024	Clinical evidence
	CIViC evidence (processed)	01-Nov-2024	Tumor-specific curated clinical evidence
RNA resources	CTAT library[82]	Oct292023	Fusion/splicing detection reference
	CTAT splicing database[82]	Jun232020	Annotate cancer-associated splice events
	Mitelman Database[112]	20241105	Annotate fusion recurrence in cancer
	Fusion/Splicing whitelists	-	Curated list of known fusion/splice variants[113–119]
Panel-specific files	TSO500 / OPA / OCA recurrent variants	2024XX	Flag panel-specific recurrent small variants
	TSO500 / OPA / OCA CNA baseline	2024XX	Enable copy number analysis
	TSO500 MSI baseline	20230124	Define baseline for MSI analysis
Other resources	MSigDB[120] MMR gene sets	2024.1	MSI-related annotations
	TVC parameters file[96]	-	TVC configuration for Ion Torrent
	Sequence-accessible regions[47]	hg38	Define callable genome for CNA baseline
	CNA problematic regions (GIAB)[55]	hg38	Exclude regions with unreliable coverage
	Microsatellite loci (10–20bp)[89]	hg38	Identify MSI loci

Supplementary Table 3. Standards for oncogenicity classification of somatic variants based on ClinGen/CGC/VICC SOP recommendations.
This table summarizes the evidence types used by ClinBioNGS to classify small variants by oncogenic potential. Each evidence type is assigned a weight (in points) and linked to specific criteria implemented in the pipeline.

Type	Category	Points	Evidence	Description	Criteria
Oncogenicity	Very Strong	+8	OVS1	Null variant in a validated TSG	VEP "HIGH" impact (excluding "stop_lost") in a TSG (≥2/3 evidence in NCG)
	Strong	+4	OS1	Same AA change as a known "oncogenic" variant (using this standard).	Same AA change as variant labeled "Oncogenic" in GENIE or ClinGen/CGC/VICC datasets
			OS2	Functional studies support an oncogenic effect. AA change is not compatible with OS1.	Variant in CIViC (oncogenic evidence) or Catalog of Validated Oncogenic Mutations at nucleotide or AA (not OS1) level
			OS3	Same AA change as a Cancer Hotspot mutation (≥50 position count and ≥10 mutation count). Not compatible with OS1.	Cancer Hotspots: position count ≥50 and mutation count ≥10 (not OS1)
	Moderate	+2	OM1	Located in a critical functional domain.	Not implemented due to lack of data
			OM2	In-frame InDel in oncogene/TSG or stop-loss in TSG. Not compatible with OVS1.	VEP "inframe_deletion" or "inframe_insertion" in oncogene/TSG, or "stop_lost" in TSG
			OM3	Same AA change as a Cancer Hotspot mutation (<50 position count and ≥10 mutation count). Not compatible with OM1 or OM4.	Cancer Hotspots: position count <50 and mutation count ≥10 (not OM1 or OM4)
			OM4	Same AA position, different change than a known "oncogenic" variant. Not compatible with OS1, OS3 or OM1.	VEP "missense_variant" and AA position in GENIE/ClinGen "Oncogenic" dataset (not OS1 or OS3)
	Supporting	+1	OP1	All *in silico* predictors support oncogenicity.	Unique predictor term is "likely_pathogenic"
			OP2	Somatic variant in cancer with a single genetic etiology.	Not implemented due to lack of data
			OP3	Same AA change as Cancer Hotspot mutation with <10 samples.	Cancer Hotspots: mutation count <10
			OP4	Absent or extremely rare in population controls (gnomAD).	gnomAD pVAF ≤0.05%
Benignity	Supporting	-1	SBP1	All *in silico* predictors suggest no impact.	Unique predictor term is "likely_benign"
			SBP2	Synonymous variant with no predicted effect.	VEP impact is "LOW" or "MODIFIER" (not OP1)
	Strong	-4	SBS1	Minor AF between 1% and 5% in gnomAD.	1% < gnomAD pVAF ≤ 5%
			SBS2	Functional studies show no oncogenic effects.	Not implemented due to lack of data
	Very Strong	-8	SBVS1	Minor AF >5% in gnomAD.	gnomAD pVAF > 5%

Supplementary Table 4. Standards for clinical variant prioritization based on AMP/ASCO/CAP guidelines.
This table outlines the evidence categories used to classify variants according to clinical significance, as implemented in ClinBioNGS. Each category includes a description, and the specific evidence-based criteria applied within the pipeline. Classification follows tier-based recommendations from AMP/ASCO/CAP guidelines.

Clinical significance	Category	Description	Criteria
Strong	Tier IA	Therapeutic, prognostic, or diagnostic evidence from FDA-approved therapies or professional guidelines.	CIViC evidence (therapeutic, prognostic, diagnostic) with 3–5 stars and level "A" for the same tumor type.
	Tier IB	Evidence from well-powered studies with expert consensus.	CIViC evidence (therapeutic, prognostic, diagnostic) with 3–5 stars and level "B" for the same tumor type.
Potential	Tier IIC	Evidence includes FDA-approved therapies in other tumor types or investigational therapies; multiple smaller studies with some consensus.	CIViC evidence with 3–5 stars and level "A" or "B" in other tumors, or level "C" (any tumor).
	Tier IID	Evidence from preclinical trials or a few case reports without consensus.	CIViC evidence with 3–5 stars and level "D".
Unknown	Tier III	Variant not observed in control population databases; found in cancer-specific databases without definitive clinical evidence.	Small variants: Classified as "Oncogenic", "Likely Oncogenic" or "VUS" (ClinGen/CGC/VICC).
			CNAs: Frequency ≥ 0.1% in GENIE.
			Fusions: Found in GENIE or MitelmanDB.
			Splice variants: Annotated as cancer-enriched in CTAT.
Benign or Likely Benign	Tier IV	Variant observed at significant frequency in population datasets or absent from cancer-specific resources.	Small variants: Classified as "Benign" or "Likely Benign" (ClinGen/CGC/VICC).
			CNAs: Frequency <0.1% in GENIE.
			Fusions: Not found in GENIE or MitelmanDB.
			Splice variants: Not cancer-enriched.

Supplementary Table 5. Reference list of known variants for gene fusions used by ClinBioNGS.
Each fusion contains the breakpoints, name, variant, and fused genomic region for each partner gene.

Fusion breakpoints (A::B)	Fusion name	Variant	Fusion range A	Fusion range B
chr2:42295516::chr2:29223528	EML4::ALK (E13::A20)	V1	chr2:42169350-42295516	chr2:29192774-29223528
chr2:42325554::chr2:29223528	EML4::ALK (E20::A20)	V2	chr2:42169350-42325554	chr2:29192774-29223528
chr2:42264731::chr2:29223528	EML4::ALK (E6::A20)	V3a	chr2:42169350-42264731	chr2:29192774-29223528
chr2:42264764::chr2:29223528	EML4::ALK (E6ins33::A20)	V3b	chr2:42169350-42301392	chr2:29192774-29223528
chr2:42264731::chr2:29223546	EML4::ALK (E6::ins18A20)	V3c	chr2:42169350-42264731	chr2:29192774-29223546
chr2:42301392::chr2:29223479	EML4::ALK (E14::del49A20)	V4	chr2:42169350-42301392	chr2:29192774-29223479
chr2:42245687::chr2:29223528	EML4::ALK (E2::A20)	V5a	chr2:42169350-42245687	chr2:29192774-29223528
chr2:42245687::chr2:29223645	EML4::ALK (E2::ins117A20)	V5b	chr2:42169350-42245687	chr2:29192774-29223645
chr2:42295516::chr2:29223597	EML4::ALK (E13::ins69A20)	V6	chr2:42169350-42295516	chr2:29192774-29223597
chr2:42301392::chr2:29223516	EML4::ALK (E14del12::A20)	V7	chr2:42169350-42301392	chr2:29192774-29223516
chr2:42304551::chr2:29223558	EML4::ALK (E17::ins30A20)	V8a	chr2:42169350-42304551	chr2:29192774-29223558
chr2:42304581::chr2:29223593	EML4::ALK (E17ins30::ins65A20)	V8b	chr2:42169350-42304581	chr2:29192774-29223593

Supplementary Table 6. Reference list of known splice variants used by ClinBioNGS.
Each variant contains the genomic coordinates and the associated variant name, gene, and transcript.

Splice range	Variant	Gene	Transcript	Splice range (hg19)
chr7:116771655-116774880	METx14del	MET	ENST00000397752.8	chr7:116411709-116414934
chr7:55161632-55171174	EGFRvII	EGFR	ENST00000275493.7	chr7:55229325-55238867
chr7:55161632-55170306	EGFRvIIb	EGFR	ENST00000275493.7	chr7:55229325-55237999
chr7:55019366-55155829	EGFRvIII	EGFR	ENST00000275493.7	chr7:55087059-55223522
chr7:55109959-55155829	EGFRvIIIb	EGFR	ENST00000275493.7	chr7:55177652-55223522
chr7:55200414-55205255	EGFRvIVa	EGFR	ENST00000275493.7	chr7:55268107-55272948
chr7:55200414-55202516	EGFRvIVb	EGFR	ENST00000275493.7	chr7:55268107-55270209
chrX:67686127-67689555	AR-V1/AR-V2/ AR-V3/AR-V4	AR	ENST00000374690.9	chrX:66905969-66909397
chrX:67643408-67680719	AR-V3	AR	ENST00000374690.9	chrX:66863250-66900561
chrX:67681004-67686009	AR-V3/AR-V4	AR	ENST00000374690.9	chrX:66900846-66905851
chrX:67686127-67692267	AR-V5	AR	ENST00000374690.9	chrX:66905969-66912109
chrX:67686127-67692187	AR-V6	AR	ENST00000374690.9	chrX:66905969-66912029
chrX:67686127-67694672	AR-V7	AR	ENST00000374690.9	chrX:66905969-66914514
chrX:67686127-67690281	AR-V8	AR	ENST00000374690.9	chrX:66905969-66910123
chrX:67686127-67693569	AR-V9	AR	ENST00000374690.9	chrX:66905969-66913411
chrX:67686127-67694878	AR-V10	AR	ENST00000374690.9	chrX:66905969-66914720
chrX:67711690-67722826	AR-V12	AR	ENST00000374690.9	chrX:66931532-66942668
chrX:67721964-67728673	AR-V13	AR	ENST00000374690.9	chrX:66941806-66948515
chrX:67722985-67728673	AR-V14	AR	ENST00000374690.9	chrX:66942827-66948515
chrX:67643408-67685940	AR-V23	AR	ENST00000374690.9	chrX:66863250-66905782
chrX:67546763-67568835	AR-45	AR	ENST00000374690.9	chrX:66766605-66788677
chrX:67569023-67643255	AR-45	AR	ENST00000374690.9	chrX:66788865-66863097
chrX:67546763-67685940	AR8	AR	ENST00000374690.9	chrX:66766605-66905782
chr7:140787585-140924565	BRAFx2-8del	BRAF	ENST00000646891.2	chr7:140487385-140624365
chr7:140783158-140924565	BRAFx2-9del	BRAF	ENST00000646891.2	chr7:140482958-140624365
chr7:140781694-140924565	BRAFx2-10del	BRAF	ENST00000646891.2	chr7:140481494-140624365
chr7:140778076-140924565	BRAFx2-11del	BRAF	ENST00000646891.2	chr7:140477876-140624365
chr7:140777089-140924565	BRAFx2-12del	BRAF	ENST00000646891.2	chr7:140476889-140624365
chr7:140754234-140924565	BRAFx2-13del	BRAF	ENST00000646891.2	chr7:140454034-140624365
chr7:140753394-140924565	BRAFx2-14del	BRAF	ENST00000646891.2	chr7:140453194-140624365
chr7:140749419-140924565	BRAFx2-15del	BRAF	ENST00000646891.2	chr7:140449219-140624365
chr7:140739947-140924565	BRAFx2-16del	BRAF	ENST00000646891.2	chr7:140439747-140624365
chr7:140734771-140924565	BRAFx2-17del	BRAF	ENST00000646891.2	chr7:140434571-140624365
chr7:140787585-140850110	BRAFx3-8del	BRAF	ENST00000646891.2	chr7:140487385-140549910
chr7:140783158-140850110	BRAFx3-9del	BRAF	ENST00000646891.2	chr7:140482958-140549910
chr7:140781694-140850110	BRAFx3-10del	BRAF	ENST00000646891.2	chr7:140481494-140549910
chr7:140778076-140850110	BRAFx3-11del	BRAF	ENST00000646891.2	chr7:140477876-140624365
chr7:140777089-140850110	BRAFx3-12del	BRAF	ENST00000646891.2	chr7:140476889-140624365
chr7:140754234-140850110	BRAFx3-13del	BRAF	ENST00000646891.2	chr7:140454034-140624365
chr7:140753394-140850110	BRAFx3-14del	BRAF	ENST00000646891.2	chr7:140453194-140624365
chr7:140749419-140850110	BRAFx3-15del	BRAF	ENST00000646891.2	chr7:140449219-140624365
chr7:140739947-140850110	BRAFx3-16del	BRAF	ENST00000646891.2	chr7:140439747-140624365
chr7:140734771-140850110	BRAFx3-17del	BRAF	ENST00000646891.2	chr7:140434571-140624365
chr7:140787585-140834608	BRAFx4-8del	BRAF	ENST00000646891.2	chr7:140487385-140534408
chr7:140783158-140834608	BRAFx4-9del	BRAF	ENST00000646891.2	chr7:140482958-140534408

chr7:140781694-140834608	BRAFx4-10del	BRAF	ENST00000646891.2	chr7:140481494-140534408
chr7:140778076-140834608	BRAFx4-11del	BRAF	ENST00000646891.2	chr7:140477876-140624365
chr7:140777089-140834608	BRAFx4-12del	BRAF	ENST00000646891.2	chr7:140476889-140624365
chr7:140754234-140834608	BRAFx4-13del	BRAF	ENST00000646891.2	chr7:140454034-140624365
chr7:140753394-140834608	BRAFx4-14del	BRAF	ENST00000646891.2	chr7:140453194-140624365
chr7:140749419-140834608	BRAFx4-15del	BRAF	ENST00000646891.2	chr7:140449219-140624365
chr7:140739947-140834608	BRAFx4-16del	BRAF	ENST00000646891.2	chr7:140439747-140624365
chr7:140734771-140834608	BRAFx4-17del	BRAF	ENST00000646891.2	chr7:140434571-140624365

Supplementary Table 7. Mismatch repair pathway genes used by ClinBioNGS.
This table lists the MSigDB (v2024.1) gene set collections and their corresponding gene symbols employed by ClinBioNGS to annotate small variants involved in the MMR pathway. These genes provide complementary evidence in the assessment of MSI. Collections and gene lists are shown in alphabetical order.

MSigDB collection	Gene list
GOBP_MISMATCH_REPAIR	ABL1, AXIN2, EXO1, HDAC10, HMGB1, LIG1, MCM8, MCM9, MLH1, MLH3, MSH2, MSH3, MSH4, MSH5, MSH6, MUTYH, PCNA, PMS1, PMS2, PMS2P1, PMS2P3, PMS2P5, PMS2P6, POLD3, PRKCG, RNASEH2A, RNASEH2B, RNASEH2C, RPA1, RPA2, RPA3, SETD2, TP73, TREX1, XPC
GOCC_MISMATCH_REPAIR_COMPLEX	MLH1, MLH3, MSH2, MSH3, MSH6, PMS1, PMS2, PMS2P1, PMS2P3, PMS2P5, PMS2P6
GOMF_MISMATCH_REPAIR_COMPLEX_BINDING	ATR, MCM8, MCM9, MLH1, MSH2, MSH6, MUTYH, PCNA, PMS2, TREX1, WRN
GOMF_MISMATCHED_DNA_BINDING	APTX, MLH1, MLH3, MSH2, MSH3, MSH4, MSH5, MSH6, MUTYH, PCNA, PMS1, PMS2, TDG, XPC
WP_DNA_MISMATCH_REPAIR	EXO1, LIG1, MLH1, MSH2, MSH6, PCNA, PMS2, POLD1, POLD2, POLD3, POLD4, POLE, POLE2, POLE3, POLE4, RFC1, RFC2, RFC3, RFC4, RFC5, RPA1, RPA2, RPA3

Supplementary Table 8. Overview of the pan-cancer NGS panels assessed in the SEQC2 validation study and benchmarking in a real clinical setting.
The table summarizes key technical parameters for each panel, including target region size, input DNA, enrichment strategy, library layout, selection method, and sequencing platform.

Study	Panel	Full name	Target size (Kb)	Input DNA (ng)	Enrichment assay	Library layout	Library selection	Sequencing platform
SEQC2	AGL	Agilent Custom Comprehensive Cancer Panel v2	7,625	30	Capture	Paired-end	Hybrid	Illumina NovaSeq 6000
	BRP	Burning Rock DX OncoScreen Plus	1,631	100	Capture	Paired-end	Hybrid	Illumina NovaSeq 6000
	IDT	Integrated DNA Technologies xGen Pan-Cancer Panel	780	100	Capture	Paired-end	Hybrid	Illumina NovaSeq 6000
	IGT	iGeneTech AIOnco-seq	944	100	Capture	Paired-end	Hybrid	Illumina HiSeq 2500
	ILM	Illumina TruSight Tumor 170	527	50	Capture	Paired-end	Hybrid	Illumina NextSeq 550
	TFS	Thermo Fisher Oncomine Comprehensive Assay v3	349	20	Amplicon	Single-end	PCR	Ion Torrent S5 XL
Clinical setting	OCA	Oncomine Comprehensive v3 GX5 DNA and Fusions	349	10	Amplicon	Single-end	PCR	Ion Torrent Genexus
	OPA	Oncomine Precision GX5 DNA and Fusions	14	10	Amplicon	Single-end	PCR	Ion Torrent Genexus
	TSO500	Illumina TruSight Oncology 500	1,940	40	Capture	Paired-end	Hybrid	Illumina NextSeq550

Supplementary Table 9. Summary table with the QC criteria used to select tumor samples for benchmarking in the TSO500, OPA, and OCA panels.
It includes the type of panel and sample, followed by each metric.

Panel	Nucleic acid	Total reads	Aligned reads (%)	On-target reads (%)	Median read length	Median insert size	Median target coverage	Target bases ≥100X (%)	Target bases ≥0.4xMean (%)
OCA	DNA	≥2M	≥95	≥90	≥80	-	≥500	≥90	≥50
	RNA	≥500K	≥75	≥50	≥80	-	-	-	-
OPA	DNA	≥500K	≥90	≥85	≥80	-	≥500	≥95	≥80
	RNA	≥250K	≥85	≥30	≥60	-	-	-	-
TSO500	DNA	≥10M	≥90	≥70	≥80	≥70	≥200	≥90	≥80
	RNA	≥10M	≥80	≥80	≥70	≥70	-	-	-

Supplementary Table 10. Performance metrics from the multi-panel validation of ClinBioNGS small variant detection using SEQC2 reference data.
For each analyzed sample, variant calling performance was assessed for both ClinBioNGS and the corresponding commercial pipeline. The following metrics were calculated: TP, FN, FP, precision, recall, and F1-score.

NGS panel	KP variants	ID	Pipeline	TP variants	FN variants	FP variants	Precision	Recall	F1-score
AGL	2824	SampleA_AGL1_ST01_30ng_LIB4	Commercial	2790	34	12	0.995717345	0.98796034	0.991823676
			ClinBioNGS	2810	14	3	0.998933523	0.995042493	0.996984211
		SampleA_AGL1_ST01_30ng_LIB3	Commercial	2779	45	20	0.992854591	0.984065156	0.988440334
			ClinBioNGS	2801	23	4	0.998573975	0.991855524	0.995203411
		SampleA_AGL1_ST01_30ng_LIB2	Commercial	2788	36	14	0.995003569	0.987252125	0.991112691
			ClinBioNGS	2805	19	1	0.999643621	0.993271955	0.996447602
		SampleA_AGL1_ST01_30ng_LIB1	Commercial	2798	26	15	0.994667615	0.990793201	0.992726628
			ClinBioNGS	2811	13	2	0.999289015	0.995396601	0.99733901
BRP	1128	SampleA_BRP1_ST27_100ng_LIB2	Commercial	1120	8	11	0.990274094	0.992907801	0.991589199
			ClinBioNGS	1120	8	8	0.992907801	0.992907801	0.992907801
		SampleA_BRP1_ST27_100ng_LIB1	Commercial	1117	11	6	0.994657168	0.990248227	0.992447801
			ClinBioNGS	1118	10	3	0.997323818	0.991134752	0.994219653
		SampleA_BRP1_ST27_100ng_LIB4	Commercial	1121	7	5	0.995559503	0.993794326	0.994676131
			ClinBioNGS	1116	12	4	0.996428571	0.989361702	0.992882562
		SampleA_BRP1_ST27_100ng_LIB3	Commercial	1122	6	5	0.995563443	0.994680851	0.995121951
			ClinBioNGS	1118	10	2	0.998214286	0.991134752	0.994661922

IDT	388	SampleA_IDT1_ST06_100ng_LIB1	Commercial	378	10	0	1	0.974226804	0.98694517
			ClinBioNGS	381	7	5	0.987046632	0.981958763	0.984496124
		SampleA_IDT1_ST06_100ng_LIB2	Commercial	379	9	0	1	0.976804124	0.988265971
			ClinBioNGS	382	6	3	0.992207792	0.984536082	0.98835705
		SampleA_IDT1_ST06_100ng_LIB3	Commercial	381	7	0	1	0.981958763	0.990897269
			ClinBioNGS	382	6	2	0.994791667	0.984536082	0.989637306
		SampleA_IDT1_ST06_100ng_LIB4	Commercial	382	6	1	0.997389034	0.984536082	0.990920882
			ClinBioNGS	385	3	4	0.989717224	0.992268041	0.990990991
IGT	353	SampleA_IGT1_ST08_100ng_LIB4	Commercial	352	1	0	1	0.997167139	0.99858156
			ClinBioNGS	352	1	0	1	0.997167139	0.99858156
		SampleA_IGT1_ST08_100ng_LIB3	Commercial	351	2	0	1	0.994334278	0.997159091
			ClinBioNGS	352	1	0	1	0.997167139	0.99858156
		SampleA_IGT1_ST08_100ng_LIB2	Commercial	352	1	0	1	0.997167139	0.99858156
			ClinBioNGS	352	1	0	1	0.997167139	0.99858156
		SampleA_IGT1_ST08_100ng_LIB1	Commercial	352	1	0	1	0.997167139	0.99858156
			ClinBioNGS	352	1	0	1	0.997167139	0.99858156
ILM	444	SampleA_ILM1_ST10_50ng_LIB4	Commercial	420	24	2	0.995260664	0.945945946	0.969976905
			ClinBioNGS	431	13	1	0.997685185	0.970720721	0.984018265
		SampleA_ILM1_ST10_50ng_LIB3	Commercial	423	21	2	0.995294118	0.952702703	0.973532796
			ClinBioNGS	436	8	0	1	0.981981982	0.990909091
		SampleA_ILM1_ST10_50ng_LIB2	Commercial	418	26	2	0.995238095	0.941441441	0.967592593
			ClinBioNGS	435	9	0	1	0.97972973	0.989761092
		SampleA_ILM1_ST10_50ng_LIB1	Commercial	417	27	2	0.99522673	0.939189189	0.966396292
			ClinBioNGS	429	15	1	0.997674419	0.966216216	0.981693364
TFS	237	SampleA_TFS1_ST24_20ng_LIB4	Commercial	221	16	1	0.995495495	0.932489451	0.962962963
			ClinBioNGS	225	12	3	0.986842105	0.949367089	0.967741935
		SampleA_TFS1_ST24_20ng_LIB3	Commercial	217	20	1	0.995412844	0.915611814	0.953846154
			ClinBioNGS	218	19	1	0.99543379	0.919831224	0.956140351
		SampleA_TFS1_ST24_20ng_LIB2	Commercial	218	19	1	0.99543379	0.919831224	0.956140351
			ClinBioNGS	221	16	1	0.995495495	0.932489451	0.962962963
		SampleA_TFS1_ST24_20ng_LIB1	Commercial	219	18	0	1	0.924050633	0.960526316
			ClinBioNGS	221	16	1	0.995495495	0.932489451	0.962962963

Supplementary Table 11. Patient characteristics in the clinical benchmarking cohort.
Age, sex, sample type, TP, and tumor type characteristics are summarized across TSO500, OPA, and OCA panels. Most represented tumor types (≥10 samples) in any panel are specified.

Characteristic	TSO500 (N = 755)	OCA (N = 595)	OPA (N = 674)
Age, years, No. (%)			
<50	103 (13.6)	33 (5.5)	41 (6.1)
50-60	177 (23.5)	83 (14.0)	104 (15.4)
61-70	176 (23.3)	127 (21.3)	200 (29.7)
71-80	170 (22.5)	114 (19.2)	184 (27.3)
>80	18 (2.4)	32 (5.4)	45 (6.7)
Missing	111 (14.7)	206 (34.6)	100 (14.8)
Sex, No. (%)			
Male	443 (58.7)	311 (52.3)	378 (56.1)
Female	306 (40.5)	242 (40.7)	191 (28.3)
Missing	6 (0.8)	42 (7.0)	105 (15.6)
Sample type, No. (%)			
DNA & RNA	587 (77.8)	449 (75.5)	536 (79.5)
DNA-only	68 (9.0)	88 (14.8)	87 (12.9)
RNA-only	100 (13.2)	58 (9.7)	51 (7.6)
TP (%), No. (%)			
<25	112 (14.8)	58 (9.7)	119 (17.7)
25-50	241 (31.9)	212 (35.6)	226 (33.5)
51-75	239 (31.7)	197 (33.1)	137 (20.3)
>75	163 (21.6)	105 (17.7)	88 (13.1)
Missing	0 (0)	23 (3.9)	104 (15.4)
Tumor type, No. (%)			
Biliary tract	5 (0.7)	1 (0.2)	21 (3.1)
Breast	1 (0.1)	29 (4.9)	1 (0.1)
Central nervous system	97 (12.8)	51 (8.6)	-
Connective tissue	54 (7.2)	1 (0.2)	2 (0.3)
Head and neck	17 (2.3)	8 (1.3)	-
Large intestine	57 (7.5)	79 (13.3)	112 (16.6)
Lung	340 (45.0)	207 (34.8)	403 (59.8)
Ovarian	12 (1.6)	15 (2.5)	-
Pancreatic	10 (1.3)	19 (3.2)	3 (0.5)
Prostate	13 (1.7)	16 (2.7)	-
Skin	13 (1.7)	46 (7.7)	-
Thyroid	13 (1.7)	5 (0.8)	-
Urinary bladder	12 (1.6)	26 (4.4)	-
Uterine	28 (3.7)	21 (3.5)	1 (0.1)
Other	81 (10.7)	49 (8.2)	27 (4.0)
Missing	2 (0.3)	22 (3.7)	104 (15.4)

Supplementary Table 12. Comparative analysis of ClinBioNGS and commercial pipeline results across three pan-cancer NGS panels.

This table summarizes the number of cancer-related alterations detected by both pipelines or uniquely by either ClinBioNGS or the commercial pipeline. Results are categorized by panel and variant type. For each group, the ClinBioNGS classification status and the presence of each variant in the CIViC database are considered.

NGS panel	Variant type	Detection status	ClinBioNGS status	CIViC	Variants
TSO500	Small variant	Both pipelines	OK	TRUE	330
				FALSE	632
			Flagged	TRUE	12
				FALSE	1093
		ClinBioNGS-only	OK	TRUE	3
				FALSE	124
			Flagged	FALSE	1792
		Commercial-only	Absent	TRUE	6
				FALSE	46
OCA	Small variant	Both pipelines	OK	TRUE	350
				FALSE	405
			Flagged	TRUE	2
				FALSE	5
		ClinBioNGS-only	OK	TRUE	4
				FALSE	32
			Flagged	TRUE	8
				FALSE	832
		Commercial-only	Absent	TRUE	5
				FALSE	29
OPA	Small variant	Both pipelines	OK	TRUE	400
				FALSE	272
			Flagged	FALSE	1
		ClinBioNGS-only	OK	TRUE	7
				FALSE	52
			Flagged	FALSE	13
		Commercial-only	Absent	TRUE	5
				FALSE	13
TSO500	AMP	Both pipelines	OK	TRUE	118
				FALSE	139
			Flagged	TRUE	545
				FALSE	463
		ClinBioNGS-only	OK	TRUE	11
				FALSE	348
			Flagged	TRUE	689
				FALSE	4996
		Commercial-only	Absent	TRUE	61
				FALSE	14
	DEL	ClinBioNGS-only	OK	TRUE	86
				FALSE	25
			Flagged	TRUE	1068
				FALSE	372
OCA	AMP	Both pipelines	OK	TRUE	162

				FALSE	29
			Flagged	TRUE	264
				FALSE	104
		ClinBioNGS-only	OK	TRUE	12
				FALSE	5
			Flagged	TRUE	985
				FALSE	356
		Commercial-only	Absent	TRUE	48
				FALSE	16
	DEL	ClinBioNGS-only	OK	TRUE	66
				FALSE	3
			Flagged	TRUE	775
				FALSE	38
OPA	AMP	Both pipelines	OK	TRUE	48
			Flagged	TRUE	167
				FALSE	1
		Commercial-only	OK	TRUE	5
			Flagged	TRUE	309
			Absent	TRUE	115
				FALSE	2
	DEL	Both pipelines	OK	TRUE	28
			Flagged	TRUE	15
		ClinBioNGS-only	OK	TRUE	11
			Flagged	TRUE	226
TSO500	Fusion	Both pipelines	OK	TRUE	45
				FALSE	23
			Flagged	TRUE	3
				FALSE	14
		ClinBioNGS-only	OK	FALSE	1
			Flagged	TRUE	5
				FALSE	36
		Commercial-only	Absent	FALSE	2
	Splice variant	Both pipelines	OK	TRUE	22
			Flagged	TRUE	10
		ClinBioNGS-only	Flagged	TRUE	2
OCA	Fusion	Both pipelines	OK	TRUE	33
				FALSE	8
			Flagged	TRUE	6
		ClinBioNGS-only	OK	TRUE	21
				FALSE	51
			Flagged	TRUE	32
				FALSE	63
		Commercial-only	Absent	TRUE	1
	Splice variant	Both pipelines	OK	TRUE	7
		ClinBioNGS-only	OK	TRUE	16
OPA	Fusion	Both pipelines	OK	TRUE	21

			FALSE	1
		Flagged	TRUE	5
			FALSE	2
	ClinBioNGS-only	Flagged	TRUE	3
	Commercial-only	Absent	TRUE	4
			FALSE	6
Splice variant	Both pipelines	OK	TRUE	9
		Flagged	TRUE	8
	ClinBioNGS-only	Flagged	TRUE	11
	Commercial-only	Absent	TRUE	1

Supplementary Table 13. ClinBioNGS-only "OK" cancer mutations with clinical evidence.
For each small variant uniquely detected by ClinBioNGS, the following information is provided: variant identifier, mutation name, AD, VAF, assigned flags, CIViC evidence type and level, oncogenicity classification, and number of supporting callers. The last column includes the detection status in the commercial pipeline. Results are grouped by NGS panel.

Panel	Variant ID	Mutation	AD	VAF (%)	CIViC Evidence type	CIViC Evidence level	Oncogenicity	Callers	Commercial status
TSO500	chr11:108365359_C/T	ATM_R3008C	10	2.42	Predictive	D - Preclinical	VUS	2	LowSupport, AD=4, VAF=1.8%
	chr7:55181312_G/T	EGFR_S768I	18	1.4	Predictive	C - Case study	Likely Oncogenic	2	LowSupport, AD=10, VAF=1.2% Confirmed with single-gene test
	chr17:7674252_C/T	TP53_M237I	11	1.4	Predictive	D - Preclinical	Likely Oncogenic	2	LowSupport, AD=49, VAF=1.2%
OCA	chr12:25245350_C/T	KRAS_G12D	113	1.63	Prognostic	B - Clinical trial	Oncogenic	3	QualityScore<8, AD=36, VAF=1.8%
	chr12:25245351_C/T	KRAS_G12S	117	1.29	Predictive	B - Clinical trial	Oncogenic	3	ABSENT, AD=24, VAF=1.2%
	chr3:38141150_T/C	MYD88_L252P	9	1.9	Predictive	B - Clinical trial	Likely Oncogenic	3	QualityScore<8, AD=11, VAF=1.8%
	chr17:7674221_G/A	TP53_R248W	11	2.4	Prognostic	B - Clinical trial	Oncogenic	3	QualityScore<8, AD=12, VAF=1.8%
OPA	chr7:55191822_T/G	EGFR_L858R	62	2.54	Predictive	A - Validated	Oncogenic	3	QualityScore<6, AD=103, VAF=2.4%
	chr12:25225627_G/A	KRAS_A146V	59	2.4	Predictive	D - Preclinical	Oncogenic	3	QualityScore<6, AD=124, VAF=2.4%
	chr12:25245350_C/G	KRAS_G12A	288	17	Predictive	B - Clinical trial	Oncogenic	3	NO CALL, AD=207, VAF=10.7%
	chr12:25245350_C/T	KRAS_G12D	42	2.5	Prognostic	B - Clinical trial	Oncogenic	3	QualityScore<6, AD=62, VAF=2.4%
	chr17:7674220_C/T	TP53_R248Q	56	2.24	Prognostic	B - Clinical trial	Oncogenic	3	QualityScore<6, AD=182, VAF=2.4%
	chr17:7673802_C/T	TP53_R273H	5	3.5	Prognostic	B - Clinical trial	Oncogenic	3	QualityScore<6, AD=21, VAF=2.2%
	chr17:7673781_C/T	TP53_R280K	23	1.75	Predictive	D - Preclinical	Likely Oncogenic	3	ABSENT, AD=46, VAF=2%

Supplementary Table 14. ClinBioNGS-only "OK" cancer mutations with no clinical evidence.
Only "Oncogenic" and "Likely Oncogenic" small variants are shown. For each variant uniquely detected by ClinBioNGS, the following information is provided: variant identifier, mutation name, alternate read count, allele frequency, and number of supporting callers. Some variants are condensed into one row, and a range of minimum and maximum values are represented. The last column includes the detection status in the commercial pipeline. Results are grouped by panel.

Panel	Oncogenicity	Variant ID	Mutation	AD	VAF (%)	Callers	Commercial status
TSO500	Oncogenic	chr20:32434485_C/A	ASXL1_Y591*	14	1.3	2	LowSupport
		chr15:44711549_G/C	B2M_M1?	10	1.7	2	LowSupport
		chr4:152323033_G/A	FBXW7_R658*	17	1.9	2	LowSupport
		chr4:152329731_G/A	FBXW7_R393*	19	1.8	2	LowSupport
		chr7:152358671_G/A	KMT2C_R56*	6	2.1	2	Blacklist;LowSupport
		4 x chr2:47803500_A/AC	**4 x** MSH6_F1088Lfs*5	[26,686]	[3.8,60.4]	[2,4]	Blacklist
		chr5:68293123_C/T	PIK3R1_R348*	21	1.5	2	LowSupport
		2 x chr10:87960914_G/A	**2 x** PTEN_W274*	[10, 40]	[4.9,10.3]	[3,4]	Blacklist
		chr13:48368549_C/T	RB1_R358*	13	1.8	2	LowSupport
		chr13:48379594_C/T	RB1_R445*	5	1.9	2	LowSupport
		chr13:48379594_C/T	RB1_R445*	8	1.5	2	LowSupport
		chr3:10149804_C/T	VHL_R161*	6	2.0	2	LowSupport
	Likely Oncogenic	chr1:26773716_CGGT…TATA/C	ARID1A_c.4004_4005-2del	6	3.8	2	Not found in the VCF
		7 x chr20:32434638_A/AG	**7 x** ASXL1_G646Wfs*12	[12,127]	[2.1,30.8]	[2,3]	Blacklist
		chr7:140781678_G/A	BRAF_R444W	10	2.4	3	LowSupport
		chr9:21971037_C/A	CDKN2A_D108Y	10	3.9	3	LowSupport
		chr2:25234373_C/T	DNMT3A_R882H	19	1.1	2	LowSupport
		chr17:7673534_CCTG…AAAG/C	TP53_c.920-31_993del	93	28.4	4	Not found in the VCF
OCA	Oncogenic	chr9:21971209_C/T	CDKN2A_c.151-1G>A	14	100	5	Not found in the SNV/InDel list
		chr12:25245348_C/A	KRAS_G13C	49	1.9	3	QualityScore<8
		chr10:87961095_C/T	PTEN_R335*	23	1.1	3	Not found in the SNV/InDel list
		chr13:48362847_C/T	RB1_R251*	20	1.7	3	QualityScore<8
		chr13:48362859_C/T	RB1_R255*	37	2.4	3	Not found in the SNV/InDel list
		chr13:48465238_C/T	RB1_R787*	22	1.5	3	Not found in the SNV/InDel list
	Likely Oncogenic	**2 x** chr7:55174737_G/A	**2 x** EGFR_E734K	[18,26]	[1.3,1.8]	3	Not found in the SNV/InDel list

		2 x chr7:55174729_G/A	**2 x** EGFR_W731*	[25,48]	1.4	3	Not found in the SNV/InDel list
		chr15:90088703_G/A	IDH2_R140W	16	2.5	3	QualityScore<8
		chr12:25227349_C/T	KRAS_A59T	36	1.3	3	QualityScore<8
		chr12:25245345_C/T	KRAS_V14I	25	1.7	3	Not found in the SNV/InDel list
		chr15:66436825_C/T	MAP2K1_P124L	22	1.2	3	ABSENT
		chr3:179234296_C/T	PIK3CA_H1047Y	20	1.9	3	QualityScore<8
OPA	Oncogenic	chr9:21971209_C/A	CDKN2A_c.151-1G>T	382	37.8	5	Not found in the SNV/InDel list
		chr20:58909366_G/A	GNAS_R201H	21	2.7	3	QualityScore<6
		3 x chr17:7675052_C/T	**3 x** TP53_c.559+1G>A	[44,417]	[2.2,42.3]	[3,5]	Not found in the SNV/InDel list
		chr17:7675052_C/G	TP53_c.559+1G>C	623	34.3	4	Not found in the SNV/InDel list
		chr17:7674858_C/G	TP53_c.672+1G>C	1089	72.1	4	Not found in the SNV/InDel list
		chr17:7674858_C/A	TP53_c.672+1G>T	80	8.1	4	Not found in the SNV/InDel list
		chr17:7673767_C/A	TP53_E285*	383	25.6	5	Not found in the SNV/InDel list
		chr17:7673824_CC/AA	TP53_G266*	390	38.3	5	Not found in the SNV/InDel list
		chr17:7673824_C/A	TP53_G266*	292	39.0	5	Not found in the SNV/InDel list
		chr17:7675206_G/A	TP53_Q136*	142	11.9	4	Not found in the SNV/InDel list
		chr17:7675182_G/A	TP53_Q144*	611	54.5	5	Not found in the SNV/InDel list
	Likely Oncogenic	chr9:21971186_CG/C	CDKN2A_R58Efs*88	5	1.2	3	Not found in the SNV/InDel list
		chr9:21971184_CTCGG/C	CDKN2A_R58Wfs*87	66	15.4	3	Not found in the SNV/InDel list
		chr10:87952142_C/T	PTEN_R173C	41	1.5	3	ABSENT
		chr17:7674239_A/AG	TP53_C242Lfs*22	128	16.0	5	Not found in the SNV/InDel list
		2 x chr17:7673809_C/T	**2 x** TP53_E271K	[67,1386]	[5.1, 49.1]	4	Not found in the SNV/InDel list
		chr17:7675204_TTG/T	TP53_Q136Tfs*12	665	20.8	4	Not found in the SNV/InDel list
		chr17:7675167_AATC…CTGC/A	TP53_Q144Pfs*21	563	31.7	4	Not found in the SNV/InDel list
		chr17:7673802_C/A	TP53_R273L	8	2.1	3	ABSENT
		3 x chr17:7674887_C/A	**3 x** TP53_S215I	[62,527]	[12.5,47.6]	[4,5]	Not found in the SNV/InDel list

Supplementary Table 15. Commercial-only cancer mutations with clinical evidence.
For each small variant uniquely reported by the commercial pipelines, the following information is provided: variant identifier, mutation name, AD, VAF, CIViC evidence type and level, and its detection status in ClinBioNGS. Results are grouped by NGS panel.

Panel	Variant ID	Mutation	AD	VAF (%)	CIViC Evidence type	CIViC Evidence level	ClinBioNGS status
TSO500	chr7:140753334_T/C	BRAF_K601E	10	0.87	Predictive	B - Clinical trial	Primary flags=LowVAF;LowCallers (Mutect2), Pisces=q20;SB
	chr7:140753336_A/T	BRAF_V600E	13	0.9	Predictive	A - Validated	Not reported, Pisces=q20;SB
	chr7:55174772_GGAATTAAGAGAAGCA/G	EGFR_E746_A750del	3	0.32	Predictive	C - Case study	Not reported, Pisces=q20;SB, Mutect2=weak_evidence
	chr17:39724728_A/AGCATACGTGATG	ERBB2_Y772_A775dup	22	1.13	Predictive	B - Clinical trial	Not reported, Pisces=q20;SB, Mutect2=slippage
	chr12:25245350_C/A	KRAS_G12V	5	1.14	Predictive	B - Clinical trial	Not reported, Pisces=q20;SB
	chr17:7675139_C/A	TP53_R158L	14	1.53	Predictive	D - Preclinical	Primary flags=LowCallers (VarDict), Pisces=q20;SB
OCA	chr7:55191822_T/G	EGFR_L858R	53	2.7	Predictive	A - Validated	Primary flags=LowCallers (VarDict, TVC), Pisces=q20;SB, Mutect2=clustered_events
	chr7:55181378_C/T	EGFR_T790M	347	17.4	Predictive	A - Validated	Primary flags=LowCallers (Pisces, TVC), Mutect2=base_qual;haplotype
	chr12:25245350_C/T	KRAS_G12D	80	4	Prognostic	B - Clinical trial	Primary flags=LowCallers (VarDict, TVC), Pisces=q20;SB, Mutect2=clustered_events
	chr3:179218303_G/A	PIK3CA_E545K	53	2.7	Predictive	C - Case study	Primary flags=LowCallers (VarDict, TVC), Pisces=q20;SB
	chr3:179234297_A/G	PIK3CA_H1047R	71	3.5	Predictive	C - Case study	Primary flags=LowCallers (VarDict, TVC), Pisces=q20, Mutect2=clustered_events
OPA	chr7:140753354_T/C	BRAF_D594G	160	2.9	Predictive	C - Case study	Primary flags=LowCallers (VarDict, TVC), Pisces=q20;SB
	chr7:55174771_AGGAATTAAGAGAAGC/A	EGFR_E746_A750del	234	3.1	Predictive	C - Case study	Primary flags=LowCallers (VarDict, TVC), Pisces=q20;SB
	chr7:55181378_C/T	EGFR_T790M	4352	38.9	Predictive	A - Validated	Primary flags=LowCallers (Pisces, TVC), Mutect2=orientation
	chr7:55181378_C/T	EGFR_T790M	440	11.7	Predictive	A - Validated	Primary flags=LowCallers (TVC), Pisces=SB, Mutect2=haplotype;orientation
	chr17:7673802_C/T	TP53_R273H	36	3.2	Prognostic	B - Clinical trial	Not reported, Pisces=q20;SB

Supplementary Table 16. Commercial-only cancer mutations with no clinical evidence.
For each small variant uniquely reported by the commercial pipelines, the following information is provided: variant identifier, mutation name, AD, VAF, and its detection status in ClinBioNGS. Some variants are condensed into one row, and a range of minimum and maximum values are represented. Results are grouped by NGS panel.

Panel	Variant ID	Mutation	AD	VAF (%)	ClinBioNGS status
TSO500	chr1:26761012_C/T	ARID1A_R693*	9	2.2	Primary flags=LowCallers (VarDict), Pisces=q20
	chr9:21971097_C/A	CDKN2A_E88*	17	1.4	Not reported, Pisces=q20
	chr2:25234374_G/A	DNMT3A_R882C	7	1.0	Primary flags=LowCallers (Mutect2), Pisces=q20;SB
	chr2:25234373_C/T	DNMT3A_R882H	11	1.2	Primary flags=LowCallers (Mutect2), Pisces=q20;SB
	chr20:58909365_C/T	GNAS_R844C	12	1.9	Primary flags=LowCallers (VarDict), Pisces=q20, Mutect2=orientation
	chr12:120978847_A/C	HNF1A_I27L	12	1.2	Not reported, Pisces=q20;SB
	29 x chr3:49684189_C/T	**29 x** MST1_G673S	[6,59]	[2.7,11.3]	Not reported, Pisces=q20;SB, VarDict=q10
	chr3:179199088_G/A	PIK3CA_R88Q	11	1.1	Primary flags=LowCallers (Mutect2), Pisces=q20;SB
	chr17:7673743_C/T	TP53_G293R	8	1.7	Not reported, Pisces=q20;SB
	chr17:7675077_G/T	TP53_H179N	3	1.7	Primary flags=LowCallers (VarDict), Pisces=q20
	chr17:7673802_C/G	TP53_R273P	9	0.9	Not reported, Pisces=q20;SB
	7 x chr21:43104346_G/A	**7 x** U2AF1_S34F	[10,234]	[3.7,27.4]	Not reported, VarDict=q10
OCA	chr1:26779059_C/T	ARID1A_R1721*	17	3.1	Primary flags=LowCallers (TVC; VarDict), Pisces=q20;SB
	chr7:140753332_T/G	BRAF_K601N	12	2.9	Primary flags=LowCallers (TVC; VarDict), Pisces=q20;SB
	chr7:140753339_G/A	BRAF_T599I	7	3.7	Primary flags=LowCallers (TVC; VarDict), Pisces=q20;SB
	chr9:21971139_C/A	CDKN2A_D74Y	233	59	Primary flags=Blacklist
	chr9:21970969_CA/C	CDKN2A_L130Rfs*16	32	26.7	Primary flags=Blacklist
	chr9:21974726_CGCCTCCAGCAGCGCCCGC/C	CDKN2A_R29_A34del	541	29.1	Primary flags=Blacklist
	chr9:21971187_G/A	CDKN2A_R58*	2	4.3	Primary flags=LowAD; LowCallers (TVC), Pisces=q20

chr9:21974793_G/T	CDKN2A_S12*	157	72.7	Primary flags=Blacklist
chr9:21974732_CAGCAGCGCCCGCACCTCC/C	CDKN2A_V28_E33del	258	13.8	Primary flags=Blacklist
chr5:68295418_A/T	-	598	46.6	Primary flags=Blacklist
chr9:21971210_T/G	-	9	45.0	Primary flags=Blacklist
chr3:41224607_A/T	CTNNB1_D32V	23	2.6	Primary flags=LowCallers (TVC; VarDict), Pisces=q20;SB
chr17:39725187_C/T	ERBB2_H878Y	11	3.0	Primary flags=LowCallers (TVC; VarDict); Pisces=q20;SB
chr4:1801837_C/T	FGFR3_R248C	4	4.2	Primary flags=LowAD;LowCallers (TVC; VarDict), Pisces=q20;SB, Mutect2=clustered_events
chr5:177095550_G/T	FGFR4_V550L	3	3.0	Not reported, Pisces=q20;SB
chr12:120988846_C/T	HNF1A_R114C	4	3.0	Primary flags=LowAD;LowCallers (TVC; VarDict), Pisces=q20;SB
chr12:120988846_C/T	HNF1A_R114C	11	6.9	Primary flags=LowCallers (TVC; VarDict), Pisces=q20;SB, Mutect2=orientation
chr3:179221146_G/A	PIK3CA_E726K	163	8.2	Primary flags=LowCallers (Pisces;TVC), VarDict=q10
chr3:179218307_A/G	PIK3CA_Q546R	72	3.6	Primary flags=LowCallers (TVC; VarDict), Pisces=q20;SB
chr12:112489106_G/A	PTPN11_Q510=	9	2.8	Primary flags=LowCallers (TVC; VarDict), Pisces=q20;SB
chr13:48345108_G/T	RB1_E137*	740	37.2	Primary flags=Blacklist
chr10:43114501_G/A	RET_C634Y	6	3.3	Primary flags=LowCallers (TVC; VarDict), Pisces=q20;SB
chr3:49375540_C/T	RHOA_G17E	28	2.5	Primary flags=LowCallers (TVC; VarDict), Pisces=q20;SB, Mutect2=orientation
chr22:23834143_G/A	SMARCB1_R374Q	5	3.4	Primary flags=LowCallers (TVC; VarDict), Pisces=q20;SB
chr19:1220429_A/G	STK11_H174R	97	20.0	Primary flags=Blacklist
chr17:7673795_A/ACAAA	TP53_A276Lfs*31	961	48.3	PASS (4 callers) but different insertion representation
chr17:7674908_T/C	TP53_D208G	77	24.5	Primary flags=Blacklist
chr17:7674233_C/A	TP53_G244C	13	2.4	Primary flags=LowCallers (TVC; VarDict), Pisces=q20;SB

	chr17:7674893_C/T	TP53_R213Q	9	2.2	Primary flags=LowCallers (TVC; VarDict), Pisces=q20;SB
OPA	chr3:41224613_G/A	CTNNB1_G34E	325	4.5	Primary flags=LowCallers (TVC; VarDict), Pisces=SB, Mutect2=haplotype;orientation
	chr3:41224622_C/T	CTNNB1_S37F	224	3.5	Primary flags=LowCallers (TVC; VarDict), Pisces=q20;SB
	chr4:1804392_G/A	FGFR3_G380R	26	5.6	Primary flags=LowAD;LowCallers (TVC), Pisces=q20;SB
	chr5:177095550_G/T	FGFR4_V550L	14	5.1	Primary flags=LowCallers (TVC; VarDict), Pisces=q20;SB
	chr5:177095550_G/A	FGFR4_V550M	24	5.0	Primary flags=LowAD, Pisces=q20;SB
	chr15:90088703_G/A	IDH2_R140W	78	3.0	Primary flags=LowCallers (TVC; VarDict), Pisces=q20;SB
	chr10:87894057_C/T	PTEN_P38S	1870	41.3	Primary flags=LowCallers (TVC; VarDict), Pisces=SB, Mutect2=haplotype;orientation
	chr17:7675085_C/G	TP53_C176S	1808	29.1	Primary flags=LowCallers (TVC; VarDict), Pisces=SB, Mutect2=base_qual;orientation
	chr17:7674248_T/C	TP53_N239D	240	5.9	Primary flags=LowCallers (TVC; VarDict), Pisces=q20, Mutect2=orientation
	chr17:7675071_G/A	TP53_R181C	472	15.3	Primary flags=LowCallers (Pisces; VarDict), Mutect2=clustered_events
	chr17:7673782_T/C	TP53_R280G	29	3.6	Primary flags=LowCallers (TVC; VarDict), Pisces=q20;SB
	2 x chr17:7673776_G/A	**2 x** TP53_R282W	[20,38]	[2.9,3.4]	Primary flags=LowCallers (TVC; VarDict), Pisces=q20;SB
	chr17:7673776_G/A	TP53_R282W	20	3.4	Primary flags=LowCallers (TVC; VarDict), Pisces=q20;SB

Supplementary Table 17. ClinBioNGS-only "OK" cancer CNAs with clinical evidence.
For each gene, the following information is provided: number of affected samples, estimated absolute CN, CIViC evidence type and level, and detection status in the commercial pipeline. Results are grouped by panel and CNA status.

Panel	CNA	Gene	Samples	CN	CIViC Evidence type	CIViC Evidence level	Commercial status
TSO500	AMP	FOXP1	1	8	Diagnostic	B - Clinical trial	No CNA gene
		MAPK1	1	11	Predictive	D - Preclinical	No CNA gene
		MYB	1	5	Prognostic	B - Clinical trial	No CNA gene
		NCOA3	4	5	Prognostic	B - Clinical trial	No CNA gene
		REL	1	8	Diagnostic	B - Clinical trial	No CNA gene
		TOP1	3	5	Predictive	B - Clinical trial	No CNA gene
	DEL	ATRX	3	0	Predictive	D - Preclinical	No CNA gene
		CDKN2A	55	0	Prognostic	B - Clinical trial	No CNA gene
		IKZF1	3	0	Prognostic	A - Validated	No CNA gene
		KMT2C	1	0	Predictive	B - Clinical trial	No CNA gene
		NF1	1	0	Predictive	D - Preclinical	No CNA gene
		PTEN	10	0	Predictive	B - Clinical trial	LowValidation (deletions are not reported)
		SMAD4	3	0	Predictive	B - Clinical trial	No CNA gene
		SMARCA4	3	0	Predictive	D - Preclinical	No CNA gene
		SMARCB1	1	0	Diagnostic	B - Clinical trial	No CNA gene
		STK11	6	0	Predictive	D - Preclinical	No CNA gene
OCA	AMP	BRAF	1	5	Predictive	C - Case study	ABSENT (CN=4.6)
		CCNE1	1	5	Prognostic	B - Clinical trial	ABSENT (CN=4.73)
		CDK4	1	10	Predictive	B - Clinical trial	NO CALL (DIFFERENT_MEAN_SIGNAL; CN=10)
		EGFR	1	5	Predictive	B - Clinical trial	NO CALL (MAPD>0.5)
		KIT	2	5	Predictive	B - Clinical trial	NO CALL (MAPD>0.5)
		MDM2	1	5	Prognostic	B - Clinical trial	NO CALL (MAPD>0.5)
		MYC	2	6, 5	Predictive	D - Preclinical	NO CALL (MAPD>0.5)
		PIK3CA	2	5	Predictive	B - Clinical trial	ABSENT (CN=4.7); NO CALL (MAPD>0.5)
		TERT	1	5	Prognostic	B - Clinical trial	ABSENT (CN=5.38)
	DEL	ATRX	2	0	Prognostic	B - Clinical trial	ABSENT (CN≈0)
		CDKN2A	43	0	Prognostic	B - Clinical trial	ABSENT (CN≈0)
		PTEN	7	0	Predictive	B - Clinical trial	ABSENT (CN≈0)
		SMARCA4	2	0	Predictive	B - Clinical trial	NO CALL (SEVERE_GRADIENT)
		STK11	6	0	Predictive	D - Preclinical	ABSENT (CN≈0)

		TP53	6	0	Predictive	D - Preclinical	ABSENT (CN≈0)
OPA	AMP	KIT	3	8, 7, 5	Predictive	B - Clinical trial	No CNA gene
		PDGFRA	2	8, 7	Predictive	D - Preclinical	No CNA gene
	DEL	CDKN2A	9	0	Prognostic	B - Clinical trial	ABSENT (CN=0.6-3.13)
		PTEN	1	0	Predictive	B - Clinical trial	ABSENT (CN=0.35)
		TP53	1	0	Predictive	D - Preclinical	No CNA gene

Supplementary Table 18. Commercial-only cancer CNAs with clinical evidence.
For each gene, the following information is provided: number of affected samples, estimated absolute CN, CIViC evidence type and level, and sample's TP. CN and TP values include the number of samples in parentheses when it is necessary. Results are grouped by panel and CNA status.

Panel	CNA	Gene	Samples	CN	CIViC Evidence type	CIViC Evidence level	TP (%)
TSO500	AMP	ALK	23	3 (n=13), 4 (n=7), 5 (n=3)	Predictive	C - Case study	Median=70
		BRAF	3	3	Predictive	C - Case study	80 (n=2), 10 (n=1)
		CDK4	2	4, 13	Predictive	B - Clinical trial	60, 10
		EGFR	21	3 (n=11), 4 (n=9), 7 (n=1;TP=20%)	Predictive	B - Clinical trial	Median=60
		ERBB2	8	3 (n=6), 4 (n=1), 7 (n=1;TP=20%)	Predictive	A - Validated	Median=65
		FGFR1	1	3	Prognostic	B - Clinical trial	95
		MET	1	4	Predictive	C - Case study	50
		MYCN	1	3	Prognostic	A - Validated	90
		PDGFRA	1	3	Predictive	D - Preclinical	70
OCA	AMP	ALK	1	6	Predictive	C - Case study	20
		BRAF	4	7 (n=2), 8 (n=2)	Predictive	C - Case study	20 (n=3), 8 (n=1)
		EGFR	3	6 (n=2), 7 (n=1)	Predictive	B - Clinical trial	20 (n=2), 15 (n=1)
		FGFR2	9	6 (n=4), 7 (n=3), 9 (n=2)	Predictive	B - Clinical trial	Median=20
		KIT	10	5 (n=1), 6 (n=2), 7 (n=5), 8 (n=1), 12 (n=1; TP=8%)	Predictive	B - Clinical trial	Median=20
		KRAS	16	6 (n=8), 7 (n=6), 11 (n=2)	Prognostic	B - Clinical trial	Median=25
		MET	3	5 (n=1), 6 (n=2)	Predictive	C - Case study	20 (n=2), 30 (n=1)
		PDGFRA	1	8	Predictive	D - Preclinical	20
		PIK3CA	1	9	Predictive	B - Clinical trial	10
OPA	AMP	ALK	2	4	Predictive	C - Case study	30, 25
		EGFR	17	4 (n=8), 5 (n=7), 8 (n=1), 14 (n=1)	Predictive	B - Clinical trial	Median=12.5
		ERBB2	24	4 (n=12), 5 (n=7), 6 (n=2), 7 (n=2), 11 (n=1)	Predictive	A - Validated	Median=20
		FGFR1	3	4	Prognostic	B - Clinical trial	20
		KRAS	27	4 (n=9), 5 (n=11), 6 (n=4), 7 (n=1), 8 (n=1)	Prognostic	B - Clinical trial	Median=15
		MET	11	4 (n=6), 5 (n=2), 6 (n=2), 7 (n=1)	Predictive	C - Case study	Median=15
		PIK3CA	31	4 (n=12), 5 (n=11), 6 (n=2), 7(n=2), 9 (n=1), 10 (n=3)	Predictive	B - Clinical trial	Median=15

Supplementary Table 19. ClinBioNGS-only "OK" cancer RNA events with clinical evidence.
For each RNA alteration uniquely detected by ClinBioNGS, the following information is provided: number of samples, number of supporting reads, CIViC evidence type and level, and detection status in the commercial pipeline. Results are grouped by panel and variant type.

Panel	Variant type	Variant name	Samples	Supporting reads	CIViC Evidence type	CIViC Evidence level	Commercial status
OCA	Fusion	AKAP13::NTRK3	1	48	Predictive	B - Clinical trial	ABSENT
		BAG4::FGFR1	1	12	Diagnostic	A - Validated	ABSENT
		CDC27::BRAF	1	12	Predictive	A - Validated	ABSENT
		EML4::ALK V2	1	17	Predictive	B - Clinical trial	57 reads, NO CALL (Sample QC FAIL)
		FGFR3::TACC3	3	902, 27, 11	Predictive	C - Case study	ABSENT (Read Count ≤40)
		FIP1L1::PDGFRA	4	99, 55, 14, 13	Diagnostic	A - Validated	ABSENT (Read Count ≤40)
		LMNA::NTRK1	1	64	Predictive	B - Clinical trial	ABSENT
		NSD3::FGFR1	7	61, 33, 24, 21, 20, 16, 15	Diagnostic	A - Validated	ABSENT (Read Count ≤1000)
		SLC34A2::ROS1	1	11	Predictive	A - Validated	ABSENT (Read Count ≤40)
		SND1::BRAF	1	33	Predictive	A - Validated	ABSENT
	Splicing	AR-V7	16	Median: 2164	Predictive	B - Clinical trial	No calling status

Supplementary Table 20. Commercial-only cancer RNA alterations with clinical evidence.
For each RNA alteration uniquely detected by commercial pipelines, the following information is provided: number of samples, number of supporting reads, CIViC evidence type and level, and detection status in ClinBioNGS. Results are grouped by panel and variant type.

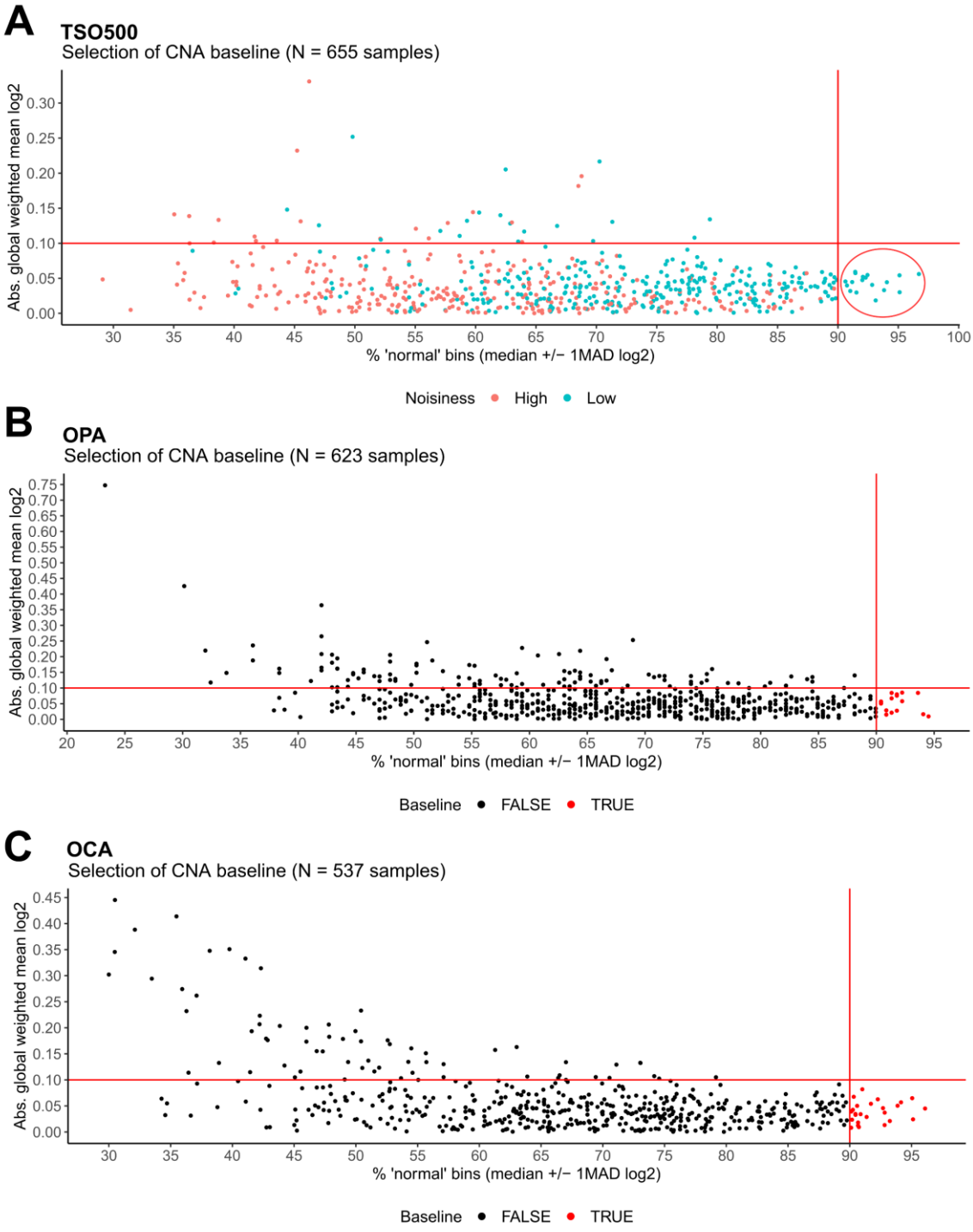
Panel	Variant type	Variant name	Samples	Supporting reads	CIViC Evidence type	CIViC Evidence level	ClinBioNGS status
OCA	Fusion	SND1::BRAF	1	187	Predictive	A - Validated	Not detected
OPA	Fusion	FGFR2::CCDC6	1	3	Predictive	C - Case study	Not detected
		KIF5B::RET	1	90	Predictive	A - Validated	Detected after reanalysis (no deduplication)
		MKRN1::BRAF	1	4	Predictive	A - Validated	Not detected
		SND1::BRAF	1	3	Predictive	A - Validated	Not detected
	Splicing	METx14del	1	7	Predictive	A - Validated	Primary flags=LowSupport (8 reads)

Supplementary Table 21. Comparative overview of representative bioinformatics workflows for the analysis of somatic NGS cancer panels.
Workflows are grouped by their availability (i.e., commercial, institutional, or open-source) with clarifying notes on their status (e.g., SaaS, FDA-approved, CE-IVDR, CLIA, RUO). Columns summarize the following features: support for tumor-only analysis, DNA–RNA integration, type of genomic profiling, annotation and prioritization strategies, generation of informative plots, reporting format, adaptability, deployment complexity, and support status. The last row corresponds to ClinBioNGS, the pipeline developed in this thesis.

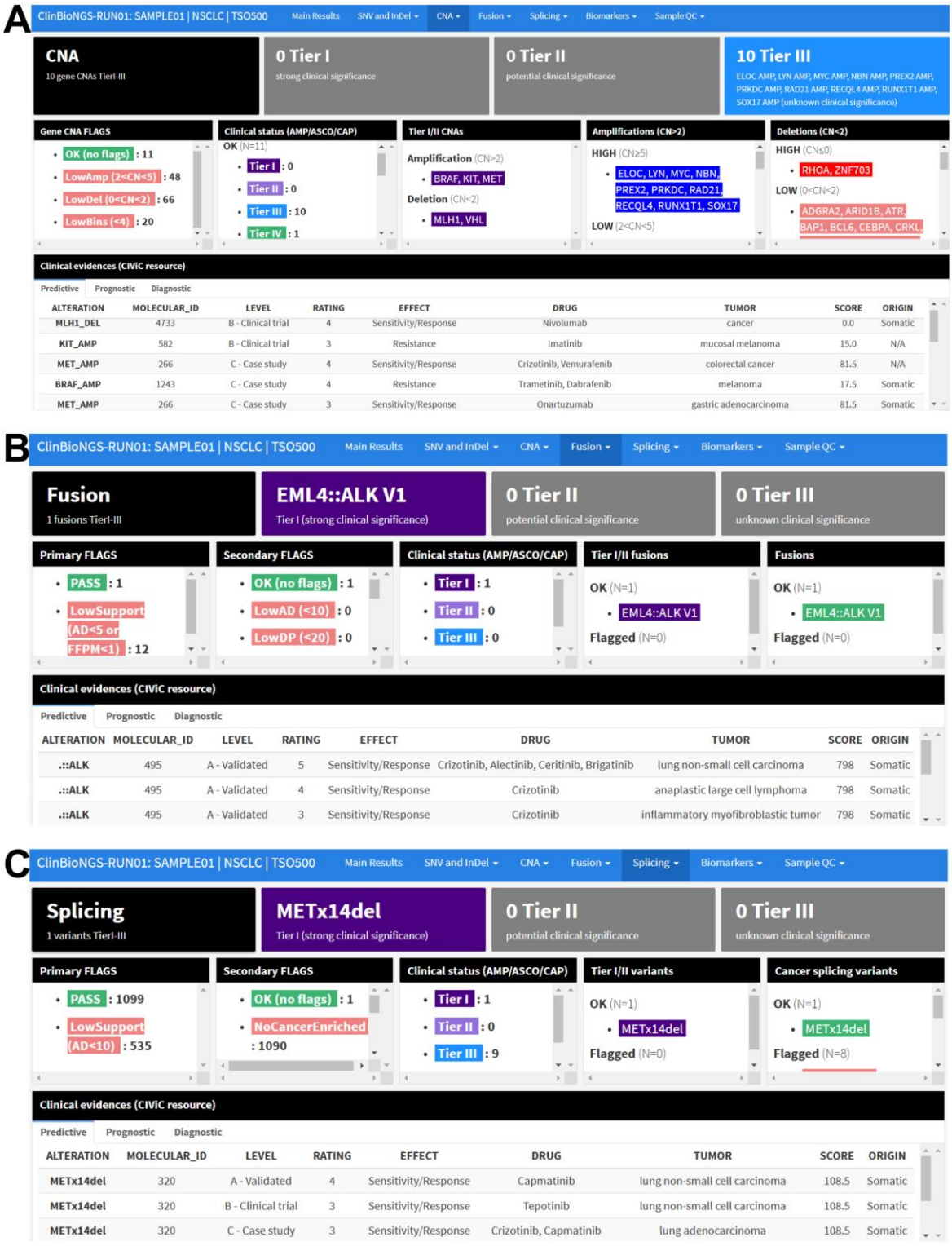
Workflow / Platform	Availability	Tumor-only	DNA+RNA integration	CGP	Annotation / Prioritization	Informative plots	Reporting	Adaptability	Deployment	Support status
Archer™ Analysis	Commercial (SaaS)	✓	✓	✓	Basic annotation; no clinical tiering	Genome browser, CNA profile	Interactive (GUI web explorer)	✗ Low (proprietary, fixed assays)	✓ Easy (ready-to-use)	✓ Active
FoundationOne CDx	Commercial (FDA-approved)	✓	✗	✓	FDA-approved AMP/ASCO/CAP-based tiering	✗	⚠ Static only	✗ Low (proprietary, fixed panels)	✓ Easy (sample send-out)	✓ Active
Illumina TSO500	Commercial (RUO/IVD)	✓	✓	✓	Basic annotation; no clinical tiering	QC, CNA profile	⚠ Static only	✗ Low (proprietary, fixed panel)	✓ Easy (ready-to-use)	✓ Active
QCI Interpret	Commercial (SaaS, CE-IVDR)	✓	✓	✓	Comprehensive annotation; AMP/ASCO/CAP tiering	QC, genome browser, CNA profile	Interactive (GUI web explorer)	⚠ Moderate (panel-agnostic but proprietary)	✓ Easy (ready-to-use)	✓ Active
SOPHiA DDM™	Commercial (SaaS, CE-IVDR)	✓	✓	✓	Comprehensive, proprietary KB; AMP/ASCO/CAP, ESCAT tiering	QC, genome browser, CNA profile	Interactive (GUI web explorer)	⚠ Moderate (panel-agnostic but proprietary)	✓ Easy (ready-to-use)	✓ Active
Thermo Ion Reporter	Commercial (RUO/IVD)	✓	✓	✓	Basic annotation; no clinical tiering	Genome browser, CNA profile	Interactive (GUI web explorer)	✗ Low (proprietary, fixed panel)	✓ Easy (ready-to-use)	✓ Active
VarSome Clinical	Commercial (SaaS, CE-IVDR)	✓	✗	✓	Comprehensive, proprietary KB; AMP/ASCO/CAP, tiering	QC, genome browser, “Lollipop” graph, CNA profile	Interactive (GUI web explorer)	⚠ Moderate (panel-agnostic but proprietary)	✓ Easy (ready-to-use)	✓ Active
DFCI OncoPanel	Institutional (CLIA)	✓	✗	✓	Internal AMP/ASCO/CAP-based tiering	✗	⚠ Static only	✗ Low (institutional fixed panel)	✗ Institutional use only	✓ Active
MSK-IMPACT	Institutional (FDA-approved)	✗	✗	✓	FDA-approved AMP/ASCO/CAP-based tiering	✗	⚠ Static only	✗ Low (institutional fixed panel)	✗ Institutional use only	✓ Active
BALSAMIC	Open-source (RUO)	✓	✗	✓	Basic annotation; no clinical tiering	QC (MultiQC)	⚠ Static only	✓ High (open-source,	⚠ Moderate (containerized	✓ Active

								panel-agnostic)	but expert requirement)	
bcbio-nextgen	Open-source (RUO)	✓	✗	⚠ Limited (no biomarkers)	Basic annotation; no clinical tiering	✗	⚠ Static only	✓ High (open-source, panel-agnostic)	✗ Hard (legacy, complex install)	✗ Inactive since 2024
DNAscan2	Open-source (RUO)	⚠ Unclear	✗	⚠ Limited (no biomarkers)	Basic annotation; no clinical tiering	✗	⚠ Static only	✓ High (open-source, panel-agnostic)	⚠ Moderate (containerized but expert requirement)	⚠ Unclear (last update 2023)
MIRACUM-Pipe	Open-source (RUO)	✓	✓	✓	Basic annotation; no clinical tiering	CNA profile, circos graph, cBioPortal integration	⚠ Static only	✓ High (open-source, panel-agnostic)	✗ Hard (low portability, some manual installation)	⚠ Unclear (last update 2023)
nf-core/sarek	Open-source (RUO)	✓	✓	✓	Basic annotation; no clinical tiering	QC (MultiQC)	⚠ Static only	✓ High (open-source, panel-agnostic)	⚠ Moderate (containerized but expert requirement)	✓ Active
PipeIT2	Open-source (RUO)	✓	✗	⚠ Limited (SNVs/InDels only)	Basic annotation; no clinical tiering	✗	⚠ Static only	⚠ Moderate (open-source, fixed assay)	⚠ Moderate (containerized but expert requirement)	⚠ Unclear (last update 2023)
SCHOOL	Open-source (RUO/CLIA)	⚠ Unclear	✓	✓	Basic annotation; no clinical tiering	✗	⚠ Static only	✓ High (open-source, panel-agnostic)	⚠ Moderate (containerized but expert requirement)	⚠ Unclear (last update 2022)
TOSCA	Open-source (RUO)	✓	✗	⚠ Limited (SNVs/InDels only)	Basic annotation; no clinical tiering	QC (MultiQC)	⚠ Static only	✓ High (open-source, panel-agnostic)	⚠ Moderate (containerized but expert requirement)	⚠ Unclear (last update 2022)
ClinBioNGS	Open-source (RUO)	✓	✓	✓	Comprehensive annotation; internal flagging; AMP/ASCO/CAP	QC, coverage, and variant-specific visualizations	Interactive (HTML)	✓ High (open-source, panel-agnostic)	⚠ Moderate (containerized but expert requirement)	✓ Active

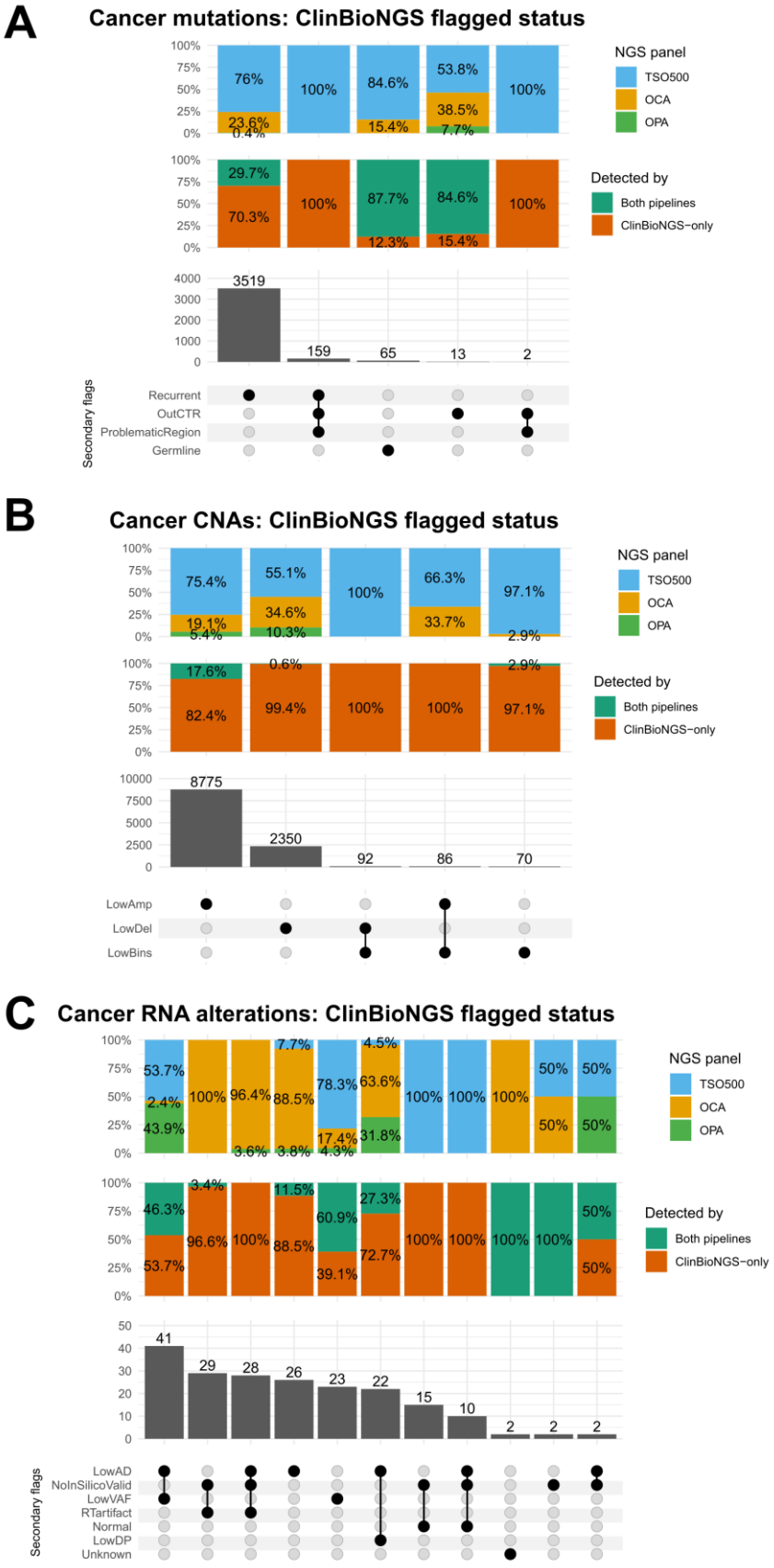
A2. Supplementary Figures



Supplementary Figure 1. Selection of CNA reference samples based on coverage variability.
(**A**) TSO500, (**B**), OPA, and (**C**) OCA panels. Scatter plots display individual samples, with the percentage of "normal" bins (log2 within median ± 1 MAD) on the x-axis and the absolute global weighted mean log2 copy ratio on the y-axis. For TSO500 (A), samples are highlighted based on their noisiness status (high or low). Samples meeting the reference selection criteria—≥90% "normal" bins and absolute weighted mean log2 ≥ 0.1—are marked as selected reference samples: red circle in (**A**) and red-colored points in (**B**) and (**C**).



Supplementary Figure 2. Overview of other results in the ClinBioNGS report.
(**A**) CNA, (**B**) fusion, and (**C**) splicing results. Top findings, summary statistics, and CIViC clinical evidence are organized into distinct panels. Color coding is used for quick visual reference, and tumor-specific clinical evidence is displayed at the bottom.



Supplementary Figure 3. ClinBioNGS flagged status in the real-world benchmarking.
UpSet plots illustrate the recurrence and combination of secondary flags assigned by ClinBioNGS across cancer-related (**A**) mutations, (**B**) CNAs, and (**C**) RNA alterations. Each plot includes two accompanying bar charts: one showing the distribution of detection status, and another showing the distribution by NGS panel for each intersection group of flags.

Universitat Autònoma
de Barcelona