PLOS Computational Biology

EDUCATION

Tutorial for variant interrogation in tumor samples

Riley J. Arseneau[1,2☯], **Leah K. MacLean**[1,2☯], **Jeanette E. Boudreau**[1,2,3*], **Daniel Gaston**[1,2,4*]

1 Department of Pathology, Dalhousie University, Halifax, Nova Scotia, Canada, **2** Beatrice Hunter Cancer Research Institute, Halifax, Nova Scotia, Canada, **3** Department of Microbiology and Immunology, Dalhousie University, Halifax, Nova Scotia, Canada, **4** Pathology & Laboratory Medicine, Nova Scotia Health, Halifax, Nova Scotia, Canada

☯ These authors contributed equally to this work.
* jeanette.boudreau@dal.ca (JEB); dan.gaston@nshealth.ca (DG)

Abstract

The increasing accessibility of next-generation sequencing has empowered researchers to investigate somatic mutations in cancer. The complexity of variant analysis pipelines, terminology, and tool selection remains a major barrier, especially for those new to the field or working in translational settings. To address this challenge, we present a practical framework that guides researchers through the critical steps of variant interrogation in tumor samples. This guide is broken into four phases: *Planning*—laying the foundation for thoughtful experimental design and a clear understanding of sequencing outputs; *Gathering Resources*—assembling the tools, reference data, and variant annotation sets required for analysis; *Filtering and Validation*—executing a systematic approach to prioritize meaningful variants; and *Dissemination and Storage*—ensuring findings are reproducible and accessible through transparent reporting and data sharing. Developed with an emphasis on accessibility, reproducibility, and clinical relevance, this framework equips researchers with the guidance to navigate variant analysis with confidence and rigor.

Citation: Arseneau RJ, MacLean LK, Boudreau JE, Gaston D (2026) Tutorial for variant interrogation in tumor samples. PLoS Comput Biol 22(2): e1013924. https://doi.org/10.1371/journal.pcbi.1013924

Editor: Patricia M. Palagi, SIB: Swiss Institute of Bioinformatics, SWITZERLAND

Published: February 17, 2026

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Next-generation sequencing (NGS) enables the investigation of somatic mutations in cancer. However, the concurrent proliferation of analysis pipelines, plugins, and programs [1] can be overwhelming for beginners. This *tutorial for variant interrogation in tumor samples* (Fig 1 provides a practical roadmap for analyzing sequencing data and disseminating findings. It is intended for researchers new to NGS or seeking greater confidence in variant analysis workflows.

We aim to empower researchers to navigate variant interrogation in tumor samples using the tools available publicly. Inspired by clinical guidelines but adapted for translational research, this guide excludes clinical decision making, which requires clinical training, licensure, and stringent criteria [2,3]. Our glossary of key terms and

concepts should be reviewed before reading the tutorial (Table 1). Most sequencing analysis, including many of the tools discussed in this article, requires familiarity with the command line interface (CLI); resources are available elsewhere [4]. CLI code examples are provided throughout this manuscript, and working through our demonstration dataset will reinforce key principles (S1 Fig, S1 Data).

Phase 1: Planning and pre-processing

Tailor the sequencing approach or selection of existing datasets to the research question

Whether generating new data or analyzing existing datasets, the research question(s) determines the most appropriate sequencing type, as methods differ in variant detection [1,5]. An ill-suited method risks poor data [5], while the right approach maximizes relevant variant detection [5]. Key guiding questions include:

- Are you characterizing alterations across the genome, or focusing on specific genes or mutation types (e.g., **single nucleotide variants** (SNVs), **insertions/deletions** (indels), **structural variants** (SVs), or **copy number variations** (CNVs))?
- Are you seeking novel or low-frequency variants?
- Are you interested in coding regions, non-coding regions, or both?

These questions clarify whether **whole genome sequencing** (WGS), **whole exome sequencing** (WES), or **targeted sequencing** (TS) best suits the study. Each involves trade-offs in breadth and depth of coverage, variant detection capability, and cost (Table 2). Higher **depth of coverage**, or read depth, increases confidence in detecting low-frequency variants [6], while **breadth of coverage** reflects how much of the genome is sequenced [7]. Depth is particularly relevant in highly heterogenous samples like tumors, where high variability and low-frequency mutations are expected [6]. Higher depth increases cost and computational requirements; WES/TS offer higher depth over smaller regions, while WGS provides broad coverage at lower per-region depth [1]. While each approach can detect the types of variants listed in Table 2, their sensitivity varies (e.g., CNV and SV detection with WES and TS is limited by capture biases and uneven coverage) [1,8,9].

Long read sequencing technologies are increasingly incorporated into cancer genomic studies [10]. While they offer improved SV detection and resolution of complex regions, they have higher error rates, lower throughput, and greater cost compared to short-read platforms [10,11].

Even with a clear strategy, practical constraints may necessitate compromise. Considerations include:

- Tumor content: Samples with <30% tumor cells require greater sequencing depth due to reduced sensitivity [14].
- Sample quality: Fresh frozen samples generally yield high-quality DNA, while **formalin-fixed paraffin-embedded** (FFPE) tissues often require higher depth to account for artifacts [15].

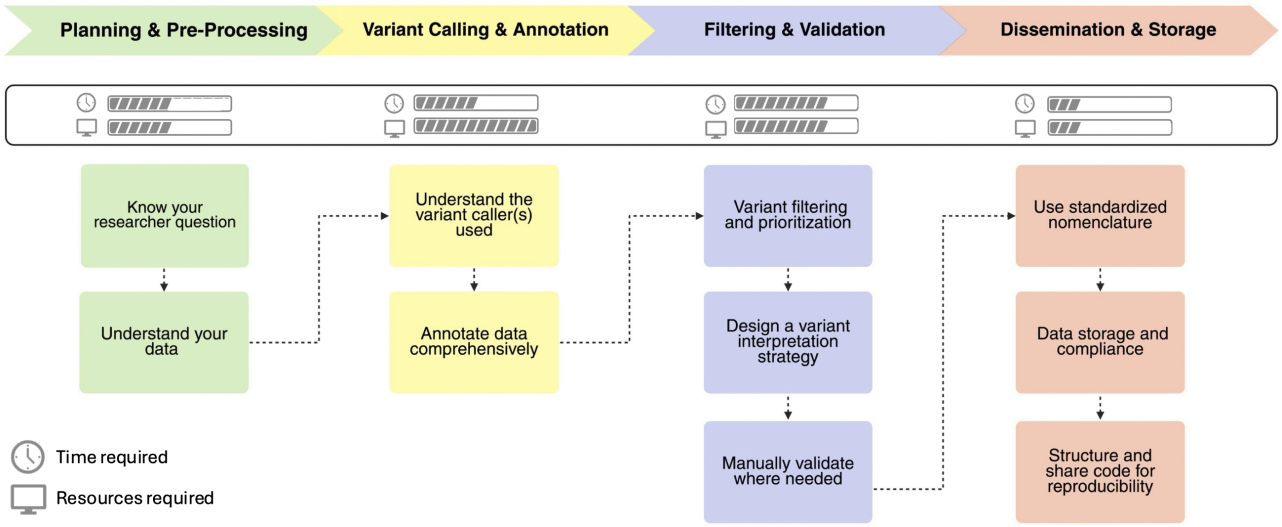

Fig 1. General workflow of variant interrogation in tumor samples. Please note that the steps are intended to be taken sequentially, and each should be completed before moving on to the next; however, depending on how data has been processed prior to your work, it may be necessary to start later in the pipeline. Created in BioRender. Arseneau, R. (2026) https://BioRender.com/armq9mn.

https://doi.org/10.1371/journal.pcbi.1013924.g001

- Budget: WGS costs most per sample, despite having the lowest cost per base pair.
- Computational resources: requisite processing power is directly proportional to the amount of data; narrowing breadth of coverage reduces data burden.
- Control samples: Control samples (e.g., commercial samples [16], panel of normals, germline samples) [17–19] improve confidence of variant calling [20].
- Public dataset (e.g., The Cancer Genome Atlas [21]) availability may necessitate adaption of the research objectives.

Understand the capabilities of the sequencing data used

Understand the sequencing data type and its processing history. Data may be received at any stage in the processing pipeline, with each step involving different files, tools, and assumptions. Incomplete or overprocessed data can lead to false positives, missed variants, and irreproducible results [22,23].

The typical workflow includes pre-processing/alignment, variant calling, annotation, filtering, and prioritization (Fig 2). Table 3 outlines common genomic file type structures (e.g., **FASTQ**, **sequencing alignment map** (SAM)/ **binary alignment map** (BAM), **variant call format** (VCF), and **annotated VCFs**).

Questions to understand previous processing:

- FASTQ: Are the reads raw or processed (e.g., adaptor trimmed)? Which tools and/or thresholds were used?
- SAM/BAM: Has alignment been performed? Were they aligned to a modern reference genome?
- VCF: Which variant caller(s) was used? Were any filtering parameters applied?
- Annotated VCF: What annotations were applied? Are they suitable for your goals? Were any filters applied that could limit variant output?

Table 1. Key terms and concepts.

Concept	Acronym	Definition / Description
Sequencing approaches		
Targeted sequencing	TS	Sequencing approach focused on targeted regions of the genome. Targeted sequencing covers highly curated combinations of genomic regions that can include coding, non-coding regions, and/or genes associated with a particular pathology.
Whole genome sequencing	WGS	Yields sequencing data across the entire genome, including both coding and non-coding regions. WGS can identify indels, CNVs, and SNVs. WGS can identify variants that impact splicing patterns or non-coding RNAs.
Whole exome sequencing	WES	Generates sequencing data for all areas of the genome that encode proteins (the exome).
Variant types		
Copy number variant	CNV	A type of SV in which a duplication or deletion changes the total number of copies of a DNA region within the genome.
Germline variant		Variants present in germ cells that are passed to all cells within the organism and passed from parent to offspring.
Insertion-deletion variant	Indel	Small insertion or deletion (<50 base pairs) variants.
Single-nucleotide variant	SNV	A variant in which a single nucleotide is substituted for another. SNVs within a coding region can be synonymous (no change in amino acid) or nonsynonymous (causes an amino acid substitution). SNVs can be rare or common within populations of humans of different ancestry.
Somatic variant		Acquired variants are only present in a subset of cells within an organism.
Structural variant	SV	A large (>50 base pairs) rearrangement of part of the genome.
File formats		
Annotated variant call format		A variant call format file with additional information about each variant, typically after processing by tools that provide functional annotations, clinical significance, or other details.
Binary alignment map	BAM	The compressed binary format of the SAM file format. Faster to read and write than SAM files. BAM is lossless compression, therefore can convert back from BAM to SAM.
FASTQ		A text-based file format that is an extension of the FASTA file format and stores the sequence ID, the sequence itself, and sequence quality data. Most common format for storing raw sequencing data.
Sequencing alignment map	SAM	Text-based file format for storing sequencing data aligned to a reference genome. SAMs are typically large files that are converted to a BAM file (lossless compressed file).
Variant call format	VCF	Text-based file format that stores SNV, indel, and structural variation calls. VCFs have two main components: the header, which stores information about the dataset and necessary reference files, and the variant call records, which store information about each variant called.
Sequencing concepts		
Base quality score		A score used to indicate the probability of an error in a specific base call. This score is represented on a Phred scale.
Phred scaling/scores		A logarithmic score expressing the confidence of a base call or variant call in sequencing. A higher Q indicates a higher confidence. Q10: 1/10 error probability (90% accuracy). Q20: 1/100 error probability (99% accuracy). Q30: 1/1000 error probability (99.9% accuracy). Q40: 1/10000 error probability (99.99% accuracy).
Sequence breadth of coverage (genome coverage)		The proportion of the genome (or targeted region) that has been covered by at least one sequencing read. Ensures the entirety of the targeted region has been sequenced.
Sequence depth of coverage (read depth)		The number of times a specific nucleotide in the genome is read during sequencing OR the average number of times the genome is read during sequencing. Also referred to as read depth. Higher depth of coverage increases the confidence of a called variant at a specific location.
Variant allele frequency	VAF	The frequency at which the variant is detected within the specimen estimated by dividing the number of variant sequencing reads at a particular location by the total number of sequencing reads at that location (reference and variant reads).
Variant functional prioritization		Variant interrogation is used to identify variants with potential or clinical significance. Functional prioritization is also used to identify variants that require manual validation.

(Continued)

Table 1. (Continued)

Concept	Acronym	Definition / Description
Variant quality filtering		Variant filtering is used to distinguish true positive variants from false positive variants.
Variant quality score		A score that reflects the confidence that a detect variant truly differs from the reference genome. This score is represented on a Phred scale.
Variant analysis concepts and tools		
Catalogue of Somatic Mutations in Cancer	COSMIC	Comprehensive resource cataloguing the occurrence of somatic mutations in human cancers.
ClinVar		Public archive reporting the relationships among human genomic variations and phenotypes hosted by the National Center for Biotechnology Information (NCBI).
Flag		A label in the filter field of a VCF that indicates if a variant has passed or failed quality control filters.
Genome aggregation database	gnomAD	Database of exome and genome sequencing data to catalog the frequency of variants within human populations.
Indexing		The process of creating an auxiliary file that records the positions of data within a large genomic file, allowing rapid access to specific genomic regions without scanning the entire file.
Integrative genomic viewer	IGV	A visualization tool for interactive exploration of variants and genomic alignments.
LiftOver		The process of converting genomic coordinates from one reference genome to another.
Pathogenicity predictors		A subset of variant annotators that interrogate identified variants to predict if a given variant is deleterious or associated with disease.
Population allele frequency		The proportion of chromosomes in a population that carry a specific allele. This represents how common a particular variant is within a population.
Variant annotators		Any tool that provides information at the variant-level.
Variant effect predictor	VEP	A toolset curated and maintained by Ensembl is used for the analysis, annotation, and prioritization of genomic variants.
Variant callers		Algorithmic tools are used to detect differences between the aligned reads of a sample and the corresponding reference genome.
Organizations with variant analysis and/or reporting guidelines		
American College of Medical Genetics Criteria	ACMG	Categorize variants in Mendelian disorders into five categories based on multiple lines of evidence; Pathogenic—Strong evidence that the variant causes disease Likely pathogenic—High likelihood of being disease-causing but not definitive Uncertain significance—Insufficient or conflicting evidence Likely Benign—Likely harmless but not fully confirmed Benign—Strong evidence the variant does not cause disease
Association of Molecular Pathology and American Society of Clinical Oncology Criteria	AMP ASCO	Classify somatic mutations into four tiers based on clinical relevance; Tier 1: Strong clinical significance—FDA-approved therapies Tier 2: Potential clinical significance—Clinical trials, emerging data Tier 3: Unknown Clinical significance Tier 4: Likely benign or neutral
Hugo Gene Nomenclature Committee	HGNC	A working group under the Human Genome Organization, with the aim to define the standard for the description of all DNA, RNA, and protein variants.
Human Genome Variation Society	HGVS	International guidelines for the standard description of DNA, RNA, and protein-level sequencing variants. These guidelines are managed by the HGNC.
Additional definitions		
Lossless compression		A form of data compression that reduces file sizes without sacrificing any information in the process.
Lossy compression		A type of data compression used when a file can afford to lose some data. Information is lost during lossy compression that cannot be retrieved when decompressing.
Formalin-fixed paraffin-embedded	FFPE	A common way of preserving tissue samples from which DNA or RNA is often extracted for sequencing.

https://doi.org/10.1371/journal.pcbi.1013924.t001

Table 2. Summary of sequencing approaches for variant detection.

Sequencing Approach	Average cost per sample*	Breadth of Coverage	Minimum Read Depth (Min-Recommended)	Detectable Variants
Whole genome sequencing	$$$	~95–98% of the genome	30–40× (germline) 80–100× (tumor)	SNVs, Indels, CNVs, SVs
Whole exome sequencing	$$	~1–2% of genome; ~85–95% of exome	100–200× (tumor)	SNVs, Indels CNVs
Targeted sequencing	$	<<1% of genome; ~100% of targeted regions	500–1000× (tumor)	SNVs, Indels, CNVs

Comparison of sequencing approaches used in tumor genomic profiling. Each approach offers trade-offs in cost, resolution, and variant detection capability. Values are derived from Illumina [12] and Yu and colleagues, 2023 [13] and will vary with adaptations to best practices as cost of sequencing decreases.

https://doi.org/10.1371/journal.pcbi.1013924.t002

Fig 2. Flowchart of common sequencing file types and analysis stages. The diagram illustrates the typical file types encountered throughout the sequencing and analysis pipeline, progressing from raw data (left) to results (right). File types are grouped by processing phase: green (Phase 1: Planning and pre-processing), yellow (Phase 2: Variant calling and annotation), and blue (Phase 3: Filtering and validation). Solid arrows indicate the standard forward progression of file generation, while dotted arrows represent steps where data can be reverted to a previous file type. The objective is to complete the pipeline, transforming raw reads into interpretable variants. Created in BioRender. Arseneau, R. (2026) https://BioRender.com/wx6ql68.

https://doi.org/10.1371/journal.pcbi.1013924.g002

CLI Example Code 1. FASTQ pre-processing.

```
# Quality check raw FASTQ files using FastQC
# INPUT: sample_R1.fastq.gz and sample_R2.fastq.gz (paired-end reads)
# OUTPUT: HTML and.zip reports in qc_reports/ directory
fastqc sample_R1.fastq.gz sample_R2.fastq.gz -o qc_reports/
# Align reads to GRCh38 reference genome using BWA-MEM
# INPUT: FASTQ files (R1 and R2), reference genome GRCh38.fa
# OUTPUT: unsorted SAM stream
bwa mem GRCh38.fa sample_R1.fastq.gz sample_R2.fastq.gz -o sample.sam
# Convert SAM to BAM using samtools view
# INPUT: sample.sam
# OUTPUT: sample.bam
samtools view -@ 8 -bS -o sample.bam sample.sam
# Sort BAM file by genomic coordinates
# INPUT: sample.bam
# OUTPUT: sample_sorted.bam
samtools sort -@ 8 -o sample_sorted.bam sample.bam
# Index BAM for fast retrieval in downstream tools
# INPUT: sample_sorted.bam
# OUTPUT: sample_sorted.bam.bai (index file)
samtools index sample_sorted.bam
```

Phase 2: Variant calling and annotation

Understand the variant caller(s) used

Variant calling transforms sequencing data into a list of genetic changes and is typically the most computationally demanding step [27]. While DRAGEN [28], MuTect2 [29], and GATK HaplotypeCaller [30] are widely used callers for

Table 3. Common sequencing file types and their structure.

File Type	Structure
FASTQ [24]	Each read contains: 1. Sequence identifier 2. Raw sequence nucleotides 3. Optional comment line(s) 4. Quality score strings
SAM [25]	Contains a header followed by alignment reads: 1. QNAME (query name or read identifier) 2. FLAG (bitwise flag for alignment information) 3. RNAME (reference sequence name) 4. POS (1-based leftmost position of the alignment) 5. Other options fields
BAM [25]	Same structure as SAM but compressed in binary format. BAM files are indexed to allow quick retrieval of alignments overlapping specific genomic regions.
VCF/Annotate VCF [26]	Contains a header (metadata and description of each column) followed by data rows for each variant called: 1. CHROM (chromosome) 2. POS (position) 3. ID (variant identifier) 4. REF (reference allele) 5. ALT (alternate allele) 6. QUAL (quality score) 7. Other optional fields If annotated, additional INFO fields are added for annotation data.

While the typical structure of each file type is described, there may be variation in the presence or extent of included metadata.

https://doi.org/10.1371/journal.pcbi.1013924.t003

detecting SNVs and Indels [22,31], they are not optimal for all sample types or goals. Challenging samples or specialized analyses may require adjusting thresholds or selecting alternate callers [32]. Adjustable caller settings, such as minimum read depth, **base quality score** thresholds, **variant allele frequency (VAF)** cutoffs, and **variant quality scores**, should match the study's goals; inappropriate thresholds risk false negatives or false positives [33]. Table 4 outlines several variant callers and their typical use cases.

CLI Example Code 2. SNV calling with MuTect2

```
# Call somatic variants using GATK Mutect2
# INPUT: tumor BAM, reference genome GRCh38.fa
# OUTPUT: somatic.vcf.gz (compressed VCF of called variants)
gatk Mutect2 \
    -R GRCh38.fa \            # reference genome
    -I tumor.bam \              # tumor sample BAM
    -O somatic.vcf.gz       # output VCF file
```

Key considerations when selecting and configuring variant caller(s):

- Variant type: Tools vary in sensitivity to detect different types of variants. Comparative studies [44,45] and documentation can guide selection. Note that SNV and CNV calling can be inconsistent between tools, so validation and cross-caller consensus may be necessary [22]. Consensus calling reduces false positives but may exclude true variants, so it is generally best used for validation or high-confidence reporting, rather than exploratory analyses.
- Sample heterogeneity: Highly heterogenous tumors may require lowering VAF or read depth thresholds to capture subclonal variants [46].

Table 4. Variant Caller information.

Variant Type	Specializations Suggested for Variant Caller	Variant Callers
SNV and small indels (<50 bp)	• High-sensitivity tools required for low-VAF detection, especially in somatic contexts [28,29,34]. • Precise base-level resolution and variant quality annotation [28,29,34,35].	Mutect2 [29], VarScan2 [34], FreeBayes [35], DRAGEN [28]
CNVs	• Use of read-depth modeling and/or segmentation algorithms [28,36–38]. • Sensitivity to large-scale copy number changes [28,36–39]. • Ability to integrate matched normal or population controls to reduce false positives [28].	DRAGEN-CNV [28], CNVKit [36], Delly [40], Lumpy [39], CNVnator [37], Canvas [38]
SVs	• Use of split-read and discordant paired-end read signals for breakpoint detection [28,40–43]. • Ability to detect large insertions, deletions, translocations, inversions, and complex rearrangements [28,36,40–42]. • Integration of read depth and mapping anomalies [37,40,41,43].	Pindel [41], DRAGEN-SV [28], Manta [42], Delly [40], CNVnator [37], TIDDIT [43]

Summary of variant types, key considerations, and example callers.

https://doi.org/10.1371/journal.pcbi.1013924.t004

- Sample quality: FFPE DNA is prone to artifacts (e.g., cytosine deamination (C>T) transitions) [15]. Minimize false positives by increasing quality thresholds [15,46].
- Discovery vs. validation: For exploratory analysis, relaxing filtering parameters and/or using multiple variant callers can maximize sensitivity. For validation, stricter filters may be warranted.

Variant callers require alignment to an up-to-date reference genome (e.g., National Library of Medicine [47] or Ensembl [48,49]. If SAM/BAM or FASTQ files are available, it's best practice to re-align to a modern reference genome. If only VCF files are available, coordinates can be converted between assemblies with **LiftOver** tools (e.g., BCFtools/liftover, CrossMap [50,51]) for better annotation.

CLI Example Code 3. Cross Caller Consensus Using BCFtools

```
# Intersect variants from two callers using bcftools isec
# INPUT: VCF from caller 1 (c1.vcf) and caller 2 (c2.vcf)
# Use bcftools isec to find variants detected by BOTH tools.
# OUTPUT: consensus_output/ directory containing:
#     0000.vcf ->intersection of both callers
#     0001.vcf ->unique to first file (c1.vcf)
#     0002.vcf ->unique to second file (c2.vcf)
#     sites.txt ->list of positions considered in the comparison
# NOTE: -n=2 ensures only variants present in both files are included in 0000.vcf.
bcftools isec -n=2 c1.vcf c2.vcf -p consensus_output/
```

CLI Example Code 4. LiftOver with CrossMap

```
# Convert VCF coordinates from GRCh37 to GRCh38 using CrossMap
# INPUT: chain file (GRCh37_to_GRCh38.chain), VCF file, reference genome
# OUTPUT: somatic_lifted.vcf
CrossMap.py vcf GRCh37_to_GRCh38.chain somatic_filtered.vcf.gz \
  GRCh38.fa somatic_lifted.vcf
```

Annotate data comprehensively

Annotation adds biological context, enables prioritization of meaningful variants, and reduces the need for manual review. Variant annotations can be associated with a specific variant (e.g., *KRAS* c.34G>T, p.G12C), groups of related variants at

the same codon (e.g., *KRAS* codon 12 mutations: G12C, G12D, G12V), gene (e.g., all pathogenic variants in the *KRAS* gene), or broader regions of the genome (e.g., SVs or amplifications spanning the *KRAS* locus on 12p12.1) [52].

Annotations are obtained through **variant annotators**, commonly via the CLI or alternatively, web-based platforms. Ensembl's **Variant Effect Predictor** (VEP) [53] is widely used, offering annotations like population frequencies, clinical significance, and predicted pathogenicity. ANNOVAR [54] and SnpEff [55] are popular alternatives.

Several annotation sources are particularly relevant for somatic cancer analysis. Population allele frequency databases (e.g., **gnomAD** [56] and TOPmed [57]) help exclude common germline polymorphisms. **Pathogenicity predictors** estimate the impact of variants on protein function or gene regulation (Table 5). Curated databases, including **ClinVar** [58], VarSome [59], Franklin by GenoOx [60], OncoKB [61,62], Genomenon Cancer Knowledgebase (Formerly JaxKB) [63], and the **Catalogue of Somatic Mutations in Cancer** (COSMIC) [64] consolidate expert-reviewed literature, functional data, and clinical annotations. COSMIC data are freely accessible for academic use following registration. In the demo dataset provided with this tutorial, the Genome Screens Mutant dataset was used. Beyond these resources, specialized annotations from the literature or pathway databases can offer insight into drug response, regulatory impact, or broader pathways.

Comprehensive annotation is important; however, excessive annotations can inflate file sizes and complicate variant filtering or interpretation. Select complementary resources that align with your research objectives [52] using recent literature and tool or database documentation [52,65].

Table 5. Commonly used pathogenicity predictors.

Tool Name	Variant Type	Tool Purpose	Common Threshold for Pathogenicity
Combined Annotation Dependent Depletion (CADD) [66]	SNV Indel Coding Non-coding	• Scores deleteriousness of variants. • Integrates multiple annotations into one metric. • C-scores increase with increasing variant deleteriousness.	C-score = >15 or >20
ClinPred [67]	Missense SNV	• Focus on disease-relevant non-synonymous variants. • Integrates multiple annotations, including ClinVar. • ClinPred score increases with increasing variant pathogenicity.	ClinPred score ≥0.5 or ≥0.7
BayesDel [68]	SNV Indel Coding Non-coding	• Integrates multiple annotations into one metric. • BayesDel score increases with increasing predicted pathogenicity. • Scores range from −1.29334 to 0.75731.	BayesDel score > -0.0570105
Rare Exome Variant Ensemble Learner (REVEL) [69]	Missense SNV	• Integrates multiple annotations into one metric. • REVEL score ranges from 0 to 1 where higher scores indicate greater likelihood that the variant is deleterious.	REVEL score = ≥0.5 or ≥0.75
MetaLR [70]	Missense SNV	• Integrates nine independent variant deleteriousness scores with variant allele frequency information to predict missense variant pathogenicity. • MetaLR scores range from 0 to 1 where variants with higher scores are more likely to be deleterious. • Variants are categorized as damaging or tolerated.	D = damaging T = tolerated / = N/A
AlphaMissense [71]	Missense SNV	• Combines structural context and evolutionary conservation to predict variant pathogenicity. • AlphaMissense score used to categorize variants into three groups: likely benign, ambiguous, likely pathogenic.	3 variant groups: likely benign (<0.34) ambiguous (0.34-0.564) likely pathogenic (>0.564)
SpliceAI [72]	Splice site variants	• Robust tool for predicting splicing defects caused by DNA variations. • Delta score increases with the likelihood of splicing defects.	Delta score = >0.5 or >0.8

Summary of computational tools used for pathogenicity prediction. Description of tools used for pathogenicity predictions, the variant types for which they are appropriate, and the thresholds recommended by each tool to indicate pathogenicity.

https://doi.org/10.1371/journal.pcbi.1013924.t005

```
Example CLI Code 5. Annotation with VEP
# Annotate variants using Ensembl VEP
# INPUT: somatic.vcf.gz
# OUTPUT: somatic_annotated.vcf with annotations
vep \
  --input_file somatic.vcf.gz \
  --output_file somatic_annotated.vcf \
  --cache \
  --assembly GRCh38 \
  --vcf
```

Phase 3: Variant filtering and validation

Filter and prioritize candidate variants

After variant annotation, reduce the variant list by **quality filtering** (remove unreliable variants [73,74]), and **functional prioritization** (elevate those most likely to be biologically or clinically relevant [75]).

Quality filtering. Quality filtering uses caller metrics and sequencing parameters to remove artifacts. Here we discuss filtering considerations; however, thresholds will vary by dataset. During variant calling, variants receive a "PASS" FILTER **flag** if they meet all the caller's quality requirements. Alternative FILTER field flags are defined by individual variant callers, described in the output VCF header or in the software documentation. Retaining only PASS flags may exclude true variants, but including non-PASS variants risks admitting artifacts. Publicly available VCFs are often pre-filtered.

VAF often informs PASS criteria. Depending on the assay's detection limit, additional VAF-specific filtering may be necessary. Typical somatic cancer minimum VAF thresholds are 5%–10% of total reads [76–78]; however, dynamic thresholds can be used to account for variability in depth of coverage. Variants observed at highly similar VAFs across many samples may indicate run-specific artifacts [79]. Control samples with known VAFs can help empirically define the lower limit of detection for the sequencing run.

Variant callers aggregate base quality scores (Phred-scaled, 30=99.9% confidence [80]) and other signals to estimate confidence in the variant as a **variant quality score**. Minimum scores of 30 are commonly used to balance sensitivity and specificity [81,82].

Functional prioritization. Functional prioritization ranks variants by biological or clinical relevance using annotations, either within annotation tools (e.g., Ensembl's VEP) [53] or *post hoc*. Functional prioritization follows either clinical-grade binning or research-focused prioritization [2,3,75].

Clinical frameworks from organizations like the **American Society of Clinical Oncology** (ASCO) [2] and the **American College of Medical Genetics and Genomics** (ACMG) [3] classify variants into tiers or pathogenicity categories. These strategies are aimed at clinical decision-making, as their high stringency may omit variants that could be of interest in research.

Research prioritization strategies weigh features like predicted functional impact, evolutionary conservation, presence in known cancer gene lists or curated databases, and occurrence within your cohort or in public datasets [75]. Fig 3 illustrates an example prioritization scheme.

During prioritization, common germline polymorphisms are excluded using population databases (e.g., gnomAD [56] and TopMED [57] with typical cutoffs of 0.01%–1% [83], while curated tumor lists can be used to help identify expected versus novel variants [75]. Most prioritization strategies emphasize protein-coding variants; however, adjust prioritization of non-coding or regulatory variants if of interest.

Employ a robust variant interpretation strategy

Interpretation integrates annotations, literature, databases, and prior knowledge to generate biologically meaningful hypotheses. Passing filters does not make a variant meaningful; a variant must relate to pathology by impacting gene expression, protein structure or function, regulatory mechanisms, or downstream molecular pathways [75].

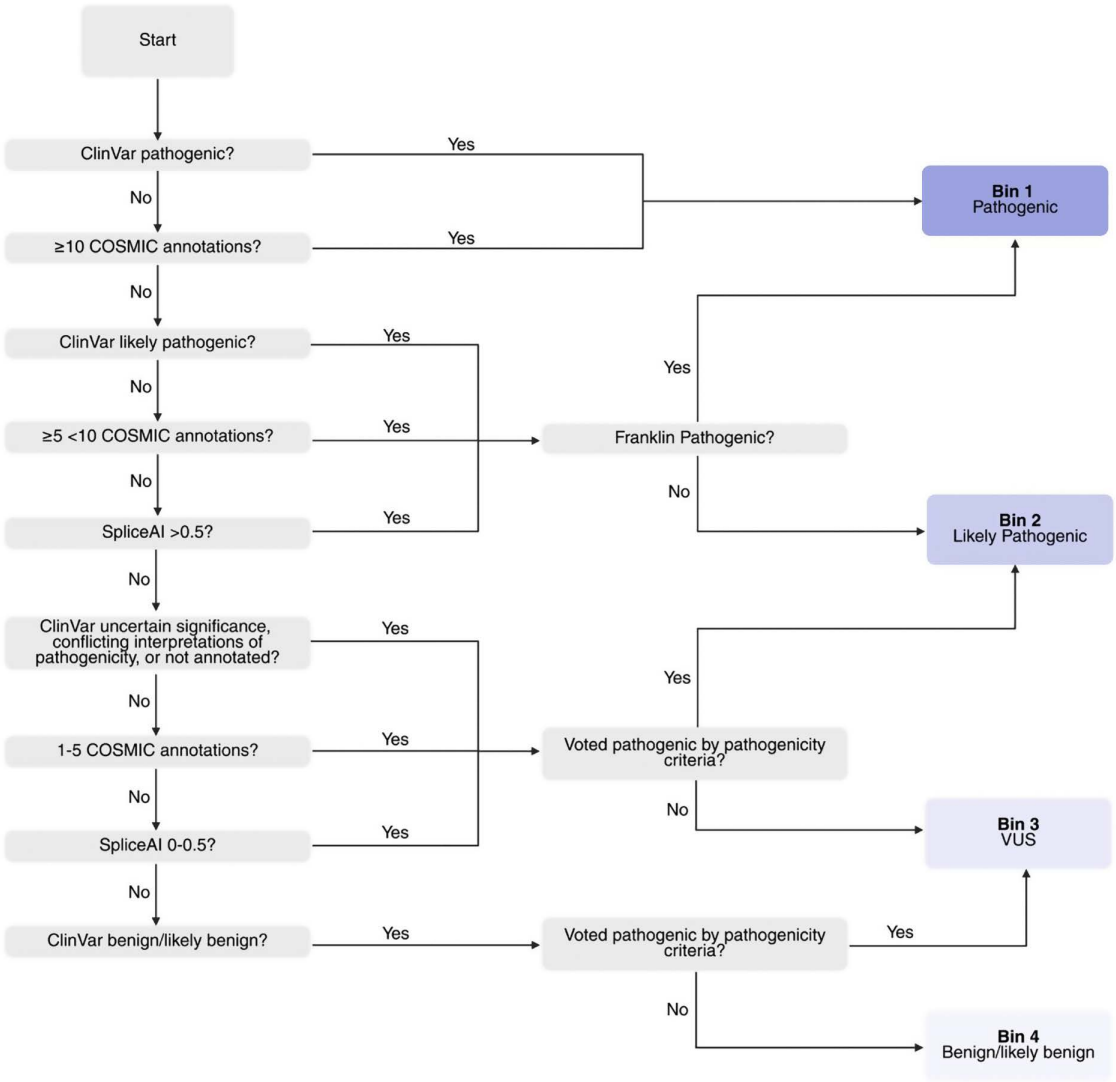

Fig 3. Example variant prioritization scheme. The flowchart illustrates a strategy for prioritizing variants into four bins based on predicted pathogenicity. Variants are initially assigned to a bin using criteria including ClinVar annotations, COSMIC frequency, and SpliceAI scores. Variants may then be reclassified to a different bin based on additional pathogenicity criteria, such as Franklin classifications (for promotion from Bin 2 to Bin 1), or other computational predictors including CADD, REVEL, and phastCons conservation scores (for promotion from Bin 4 to Bin 3, or Bin 3 to Bin 2). Bin 1 contains pathogenic variants, Bin 2 contains likely pathogenic variants, Bin 3 contains variants of uncertain significance (VUS), and Bin 4 represents likely benign or benign variants. Created in BioRender. Arseneau, R. (2026) https://BioRender.com/o4pbmzu.

https://doi.org/10.1371/journal.pcbi.1013924.g003

Cancer interpretation often focuses on oncogenes with activating mutations or amplifications [84], or tumor suppressor genes, which exhibit deletions, truncations, or inactivating mutations [85]. ClinVar [58] and COSMIC [64] remain central repositories for variant-level information. Online databases (e.g., OncoKB [62], VarSome [59], Franklin [60], and the Clinical Interpretation of Variants in Cancer [86]) provide additional information, including clinical significance, relevant publications, ACMG/ASCO classifications, pharmacogenomic associations, and community-submitted interpretation. Pathway analysis and Gene Set Enrichment Analysis [87,88] can reveal broader relevance for variants that may appear marginal in isolation. The list of molecular changes relevant to cancer continues to expand; thus, a comprehensive literature review is essential [89]. Ultimately, your scientific judgement is essential for variant interpretation.

Manually validate variants where appropriate

Pipeline quality controls may miss false positives, so manual review is essential for confirming variants. Tools like **Integrative Genomics Viewer** (IGV) [90] allow inspection of read alignment, the variants position within reads, and local CNV [91]. IGV is essential for novel or unexpected findings, and guidelines are available elsewhere [91].

For paired normal-tumor sequencing, both sequencing alignments (tumor and normal) should be evaluated to confirm somatic status [22,91]. Similarly, when a control sample has been sequenced, it should be compared with the sample of interest.

Phase 4: Disseminating and storage

Use standardized nomenclature

When disseminating results, standardized nomenclature ensures variants are universally understandable, traceable to reference data, and correctly interpreted [2,92] (Table 6).

Gene and protein nomenclature. Use gene and protein symbols from the **Human Genome Organization (HUGO) Gene Nomenclature Committee** (HGNC) [92], maintaining one consistent name and introducing aliases at first mention (e.g., *CD274*, a.k.a. *PDL1* or *B7H1*) [92]. Specify the reference transcript used for annotation, preferably the Matched Annotation from NCBI and EMBL-EBI (MANE) [93].

Variant reporting. Report variants at the DNA level following **Human Genome Variation Society** (HGVS) guidelines [94]. Present designations in both the manuscript text and a table that includes DNA, RNA, and protein nomenclature where applicable [95]. Verify variant descriptions with tools like Mutalyzer [96] for HGVS compliance and formatting [97].

CLI Example Code 6. Verifying variant descriptions with Mutalyzer

```
# Normalize an HGVS description (use canonical form for reporting)
# INPUT: Genomic or transcript HGVS (e.g., "GRCh38 (chr<CHR>):g.<POS><REF>><ALT>" or
"NM_<TRANSCRIPT>:c.<...>")
# OUTPUT: JSON with normalized_description (report this)
curl -sS "https://v3.mutalyzer.nl/api/normalize/<YOUR_HGVS_DESCRIPTION>"
# Map a normalized description to a specific transcript (e.g., MANE Select)
# INPUT: description=<YOUR_NORMALIZED_HGVS>; target_selector=<TRANSCRIPT_ACCESSION> (e.g.,
NM_########.#)
# OUTPUT: JSON with c.-notation for the requested selector
curl -sS --data-urlencode "description=<YOUR_NORMALIZED_HGVS>" \
        --data-urlencode "target_selector=<TRANSCRIPT_ACCESSION>" \
        "https://v3.mutalyzer.nl/api/map/"
# List available selectors (transcripts) for a reference sequence
# INPUT: reference_id=<REFERENCE_ACCESSION> (e.g., NC_########.##)
# OUTPUT: JSON array of selector IDs (choose MANE when available)
curl -sS https://v3.mutalyzer.nl/api/get_selectors/<REFERENCE_ACCESSION>
# Convert reference positions to selector-oriented coordinates (genome to c.-notation)
# INPUT: Genomic HGVS (e.g., "GRCh38 (chr<CHR>):g.<POS><REF>><ALT>" or "NC_########.##:g.<...>")
# OUTPUT: JSON with c.-level coordinates aligned to the chosen selector
curl -sS --data-urlencode "description=<YOUR_GENOMIC_HGVS>" \
         "https://v3.mutalyzer.nl/api/position_convert/"
```

Genomic data storage and compliance

Genomic data management should follow Findable, Accessible, Interoperable, and Reusable principles (FAIR) [98].

Working directory storage. During active analysis, use hierarchical directory system for raw data, intermediate files, results, and metadata [99]. Apply a version control system to track changes in scripts and metadata [98].

Table 6. Variant reporting checklist.

Checklist items	Complete?
HGNC-approved DNA, RNA, and protein symbols are used consistently and correctly formatted.	☐
Transcript(s) used for variant annotation are clearly specified (preferably MANE). If MANE transcripts are unavailable, document the transcript and rationale.	☐
Variant descriptions have been externally validated.	☐
A table with DNA (including genomic coordinates), RNA, and protein information for reported variants is included in the text or supplementary data.	☐

https://doi.org/10.1371/journal.pcbi.1013924.t006

Long-term storage. Retain files essential for future auditing, reanalysis, or validation [100], including raw data, analysis scripts, auxiliary files, and selected results. Genomic files are large, making compression essential for storage. Two primary types of compression are available: lossless and lossy. **Lossless** formats like FASTQ.gz and BAM are preferred for permanent storage. When **lossy** compression is used, its impact on downstream analyses should be considered [101]. All transformations should be logged with details on software, versions, and parameters [98].

Storage scalability, redundancy, and security. Combine institutional servers, cloud storage, and external drives for redundancy [99,100,102] and ensure compliance with ethical, legal, and institutional standards [103], including encryption and secure transfer protocols [100,104]. Deposit data in secure external repositories when possible [98].

Structure and share code for reproducibility

Genomic analyses rely on complex workflows that must be documented for reproducibility, validation, and reuse [105] (Table 7). Share code when possible [98,105] via repositories such as GitHub [106] or GitLab [107]. Use workflow managers (e.g., Snakemake [108], Nextflow [109]) and package/container managers (e.g., Conda [110], Docker [111]) to standardize environments and automate pipelines [112]. Include README files detailing file structures, workflows, and expected outputs [100,113]. Code availability statements can be referenced from *The American Journal of Human Genetics* [114] or *Oxford Academic* [115].

Table 7. Code reproducibility and sharing checklist.

Checklist items	Complete?
Archive the finalized scripts, workflows, software versions, and the date of dataset download alongside all datasets used in the analysis.	☐
Write well-annotated code that can be easily adapted to the environments of other users. Prioritize modular structures and functions and avoid hard-coded path names.	☐
Use version control systems, such as Git [116], along with external repositories to maintain code integrity.	☐
A include README and configuration files with details about the computational environment to ensure replicability.	☐

https://doi.org/10.1371/journal.pcbi.1013924.t007

Conclusions

This framework provides practical guidance for tumor sequencing analysis, covering the full workflow—from study design to data interpretation and dissemination—while emphasizing code sharing to foster reproducibility and collaborative science. By promoting a structured, reproducible approach, these guidelines support consistency in variant interpretation and reporting, contributing to greater clarity, transparency, and comparability across studies in cancer genomics.

Supporting information

S1 Fig. Variant interrogation in practice: step-by-step questions using demo data. Guiding questions for demo data used to demonstrate variant filtering and prioritization in tumor and control samples. Box 1 focuses on interrogating the dataset in question. Box 2 focuses on variant annotations and their use cases. Box 3 and Box 4 encompass quality filtering and functional prioritization of variants. Box 5 guides the identification and reporting of final variants of interest. Box colors correspond to processing phase: green (Planning and pre-processing), yellow (Variant calling and annotation), and blue (Filtering and validation). Created in BioRender. Arseneau, R. (2026) https://BioRender.com/gexi4i4.
(TIF)

S1 Data. Example data and associated code for working through variant annotation and interrogation following the guidelines laid out in this manuscript.
(ZIP)

Acknowledgments

The authors acknowledge the support of the Terry Fox Research Institute Marathon of Hope, and the Beatrice Hunter Cancer Research Institute.

RA is a trainee in the Cancer Research Training Program of the Beatrice Hunter Cancer Research Institute; funds for RA are provided by GIVETOLIVE, by the Kilpatrick Trust through the Dalhousie Faculty of Medicine 2023 Graduate Studentship program, and by the Terry Fox MOHCCN Health Informatics and Data Scientist Award.LM is a trainee in the Cancer Research Training Program of the Beatrice Hunter Cancer Research Institute; funds for LM are provided by the Canadian Institute for Health Research Doctoral Research Reward (FRN 183293), and by the Killam Predoctoral Scholarship held at Dalhousie University.

Author contributions

Conceptualization: Riley J. Arseneau, Leah K. MacLean, Jeanette E. Boudreau, Daniel Gaston.

Data curation: Riley J. Arseneau, Leah K. MacLean.

Funding acquisition: Jeanette E. Boudreau.

Methodology: Riley J. Arseneau, Leah K. MacLean.

Supervision: Jeanette E. Boudreau, Daniel Gaston.

Writing – original draft: Riley J. Arseneau, Leah K. MacLean.

Writing – review & editing: Riley J. Arseneau, Leah K. MacLean, Jeanette E. Boudreau, Daniel Gaston.

References

1. Satam H, Joshi K, Mangrolia U, Waghoo S, Zaidi G, Rawool S, et al. Next-generation sequencing technology: current trends and advancements. Biology (Basel). 2023;12(7):997. https://doi.org/10.3390/biology12070997 PMID: 37508427
2. Li MM, Datto M, Duncavage EJ, Kulkarni S, Lindeman NI, Roy S, et al. Standards and guidelines for the interpretation and reporting of sequence variants in cancer. J Mol Diagn. 2017;19(1):4–23.

3. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, 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. Genet Med. 2015;17(5):405–24. https://doi.org/10.1038/gim.2015.30 PMID: 25741868
4. The Biostar Handbook. 2nd ed [Internet]. [cited 2025 May 25]. Available from: https://www.biostarhandbook.com/index.html
5. Abbasi A, Alexandrov LB. Significance and limitations of the use of next-generation sequencing technologies for detecting mutational signatures. DNA Repair (Amst). 2021;107:103200. https://doi.org/10.1016/j.dnarep.2021.103200 PMID: 34411908
6. Williams MJ, Sottoriva A, Graham TA. Measuring clonal evolution in cancer with genomics. Annu Rev Genom Hum Genet. 2019;20(1):309–29.
7. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet. 2014;15(2):121–32. https://doi.org/10.1038/nrg3642 PMID: 24434847
8. Zare F, Dow M, Monteleone N, Hosny A, Nabavi S. An evaluation of copy number variation detection tools for cancer using whole exome sequencing data. BMC Bioinformatics. 2017;18(1):286. https://doi.org/10.1186/s12859-017-1705-x PMID: 28569140
9. Zanardo ÉA, Monteiro FP, Chehimi SN, Oliveira YG, Dias AT, Costa LA, et al. Application of whole-exome sequencing in detecting copy number variants in patients with developmental delay and/or multiple congenital malformations. J Mol Diagn. 2020;22(8):1041–9. https://doi.org/10.1016/j.jmoldx.2020.05.007 PMID: 32497716
10. Logsdon GA, Vollger MR, Eichler EE. Long-read human genome sequencing and its applications. Nat Rev Genet. 2020;21(10):597–614. https://doi.org/10.1038/s41576-020-0236-x PMID: 32504078
11. Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. Opportunities and challenges in long-read sequencing data analysis. Genome Biol. 2020;21(1):30. https://doi.org/10.1186/s13059-020-1935-5 PMID: 32033565
12. Illumina [Internet]. [cited 2025 May 25]. Sequencing Coverage for NGS Experiments. Available from: https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/coverage.html
13. Yu H, Yu H, Zhang R, Peng D, Yan D, Gu Y, et al. Targeted gene panel provides advantages over whole-exome sequencing for diagnosing obesity and diabetes mellitus. J Mol Cell Biol. 2023;15(6):mjad040. https://doi.org/10.1093/jmcb/mjad040 PMID: 37327085
14. Naito Y, Aburatani H, Amano T, Baba E, Furukawa T, Hayashida T, et al. Clinical practice guidance for next-generation sequencing in cancer diagnosis and treatment (edition 2.1). Int J Clin Oncol. 2021;26(2):233–83.
15. Munchel S, Hoang Y, Zhao Y, Cottrell J, Klotzle B, Godwin AK, et al. Targeted or whole genome sequencing of formalin fixed tissue samples: potential applications in cancer genomics. Oncotarget. 2015;6(28):25943–61.
16. Horizon Discovery [Internet]. [cited 2025 May 25]. Mimix™ Structural Multiplex (gDNA) Reference Standard. Available from: https://horizondiscovery.com/en/reference-standards/products/structural-multiplex-reference-standard-gdna
17. UMCCR Genomics Platform Group [Internet]. Panel of normals. 2019 [cited 2025 Nov 21]. Available from: https://umccr.org/blog/panel-of-normals/
18. Matched tumor-normal sequencing: The preferred method for identifying somatic mutations driving tumorigenesis [Internet]. SOPHiA GENETICS [cited 2025 Nov 21]. Available from: https://www.sophiagenetics.com/resource/matched-tumor-normal-sequencing-preferred-method-identifying-somatic-mutations-driving-tumorigenesis/
19. Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. Cell. 2018;173(2):291-304.e6. https://doi.org/10.1016/j.cell.2018.01.018
20. Miura T, Yasuda S, Sato Y. A simple method to estimate the in-house limit of detection for genetic mutations with low allele frequencies in whole-exome sequencing analysis by next-generation sequencing. BMC Genom Data. 2021;22(1):8. https://doi.org/10.1186/s12863-020-00956-x PMID: 33602132
21. National Cancer Institute. The Cancer Genome Atlas Program (TCGA) - NCI [Internet]. 2022 [cited 2025 May 8]. Available from: https://www.cancer.gov/ccg/research/genome-sequencing/tcga
22. Koboldt DC. Best practices for variant calling in clinical sequencing. Genome Med. 2020;12(1):91. https://doi.org/10.1186/s13073-020-00791-w PMID: 33106175
23. He B, Zhu R, Yang H, Lu Q, Wang W, Song L, et al. Assessing the impact of data preprocessing on analyzing next generation sequencing data. Front Bioeng Biotechnol. 2020;8:817. https://doi.org/10.3389/fbioe.2020.00817 PMID: 32850708
24. MAQ. FASTQ Format [Internet]. [cited 2025 May 25]. Available from: https://maq.sourceforge.net/fastq.shtml
25. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078–9. https://doi.org/10.1093/bioinformatics/btp352
26. vcftools-spec List Signup and Options [Internet]. [cited 2025 Aug 18]. Available from: https://sourceforge.net/projects/vcftools/lists/vcftools-spec
27. Kappelmann-Fenzl M. Computer setup. In: Kappelmann-Fenzl M, editor. Next Generation Sequencing and Data Analysis [Internet]. Cham: Springer International Publishing; 2021 [cited 2025 Jun 10]. p. 59–69. Available from: https://doi.org/10.1007/978-3-030-62490-3_5
28. Illumina [Internet]. DRAGEN secondary analysis. Available from: https://www.illumina.com/content/dam/illumina/gcs/assembled-assets/marketing-literature/dragen-bio-it-data-sheet-m-gl-00680/dragen-bio-it-data-sheet-m-gl-00680.pdf
29. Benjamin D, Sato T, Cibulskis K, Getz G, Stewart C, Lichtenstein L. Calling Somatic SNVs and Indels with Mutect2 [Internet]. bioRxiv; 2019 [cited 2025 May 8]. p. 861054. Available from: https://www.biorxiv.org/content/10.1101/861054v1

30. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Auwera GAVd, et al. Scaling accurate genetic variant discovery to tens of thousands of samples [Internet]. bioRxiv; 2018 [cited 2025 May 8]. p. 201178. Available from: https://www.biorxiv.org/content/10.1101/201178v3
31. Wong M, Liew B, Hum M, Lee NY, Lee ASG. Benchmarking of variant calling software for whole-exome sequencing using gold standard datasets. Sci Rep. 2025;15(1):13697. https://doi.org/10.1038/s41598-025-97047-7 PMID: 40258889
32. Garcia-Prieto CA, Martínez-Jiménez F, Valencia A, Porta-Pardo E. Detection of oncogenic and clinically actionable mutations in cancer genomes critically depends on variant calling tools. Bioinformatics. 2022;38(12):3181–91. https://doi.org/10.1093/bioinformatics/btac306 PMID: 35512388
33. Karimnezhad A, Perkins TJ. Empirical Bayes single nucleotide variant-calling for next-generation sequencing data. Sci Rep. 2024;14(1):1550. https://doi.org/10.1038/s41598-024-51958-z PMID: 38233494
34. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012;22(3):568–76. https://doi.org/10.1101/gr.129684.111 PMID: 22300766
35. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing [Internet]. arXiv; 2012 [cited 2025 May 25]. Available from: http://arxiv.org/abs/1207.3907
36. 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. https://doi.org/10.1371/journal.pcbi.1004873 PMID: 27100738
37. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. 2011;21(6):974–84. https://doi.org/10.1101/gr.114876.110 PMID: 21324876
38. Roller E, Ivakhno S, Lee S, Royce T, Tanner S. Canvas: versatile and scalable detection of copy number variants. Bioinformatics. 2016;32(15):2375–7.
39. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. 2014;15(6):R84. https://doi.org/10.1186/gb-2014-15-6-r84 PMID: 24970577
40. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELly: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 2012;28(18):i333-9.
41. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics. 2009;25(21):2865–71. https://doi.org/10.1093/bioinformatics/btp394
42. Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics. 2016;32(8):1220–2.
43. Eisfeldt J, Vezzi F, Olason P, Nilsson D, Lindstrand A. TIDDIT, an efficient and comprehensive structural variant caller for massive parallel sequencing data. F1000Res. 2017;6:664. https://doi.org/10.12688/f1000research.11168.2 PMID: 28781756
44. Guille A, Adélaïde J, Finetti P, Andre F, Birnbaum D, Mamessier E, et al. A benchmarking study of individual somatic variant callers and voting-based ensembles for whole-exome sequencing. Brief Bioinform. 2024;26(1):bbae697. https://doi.org/10.1093/bib/bbae697 PMID: 39828270
45. Bian X, Zhu B, Wang M, Hu Y, Chen Q, Nguyen C, et al. Comparing the performance of selected variant callers using synthetic data and genome segmentation. BMC Bioinformatics. 2018;19(1):429. https://doi.org/10.1186/s12859-018-2440-7 PMID: 30453880
46. Zverinova S, Guryev V. Variant calling: considerations, practices, and developments. Hum Mutat. 2022;43(8):976–85. https://doi.org/10.1002/humu.24311 PMID: 34882898
47. National Center for Biotechnology Information (NCBI). Genome. NCBI. [cited 2025 May 8]. Available from: https://www.ncbi.nlm.nih.gov/datasets/genome/#/GCF_000001405.40/?utm_source=gquery&utm_medium=referral&utm_campaign=KnownItemSensor:acc
48. Dyer SC, Austine-Orimoloye O, Azov AG, Barba M, Barnes I, Barrera-Enriquez VP, et al. Ensembl 2025. Nucleic Acid Res. 2024;53(D1):D948–57.
49. Guo Y, Dai Y, Yu H, Zhao S, Samuels DC, Shyr Y. Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis. Genomics. 2017;109(2):83–90. https://doi.org/10.1016/j.ygeno.2017.01.005 PMID: 28131802
50. Genovese G, Rockweiler NB, Gorman BR, Bigdeli TB, Pato MT, Pato CN, et al. BCFtools/liftover: an accurate and comprehensive tool to convert genetic variants across genome assemblies. Bioinformatics. 2024;40(2):btae038. https://doi.org/10.1093/bioinformatics/btae038 PMID: 38261650
51. Zhao H, Sun Z, Wang J, Huang H, Kocher JP, Wang L. CrossMap: a versatile tool for coordinate conversion between genome assemblies. Bioinformatics. 2014;30(7):1006–7.
52. Hebbar P, Sowmya SK. Genomic variant annotation: A comprehensive review of tools and techniques. In: Abraham A, Gandhi N, Hanne T, Hong TP, Nogueira Rios T, Ding W, editors. Intelligent systems design and applications. Cham: Springer International Publishing; 2022. p. 1057–67.
53. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016;17(1):122.
54. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):e164. https://doi.org/10.1093/nar/gkq603 PMID: 20601685
55. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012;6(2):80–92. https://doi.org/10.4161/fly.19695 PMID: 22728672
56. Chen S, Francioli LC, Goodrich JK, Collins RL, Kanai M, Wang Q, et al. A genomic mutational constraint map using variation in 76,156 human genomes. Nature. 2024;625(7993):92–100. https://doi.org/10.1038/s41586-023-06045-0 PMID: 38057664

57. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature. 2021;590(7845):290–9. https://doi.org/10.1038/s41586-021-03205-y PMID: 33568819
58. Landrum MJ, Chitipiralla S, Kaur K, Brown G, Chen C, Hart J, et al. ClinVar: updates to support classifications of both germline and somatic variants. Nucleic Acids Res. 2025;53(D1):D1313–21. https://doi.org/10.1093/nar/gkae1090 PMID: 39578691
59. Kopanos C, Tsiolkas V, Kouris A, Chapple CE, Albarca Aguilera M, Meyer R, et al. VarSome: the human genomic variant search engine. Bioinformatics. 2019;35(11):1978–80.
60. Franklin by Genoox [Internet]. [cited 2025 May 8]. Available from: https://franklin.genoox.com/clinical-db/home
61. Suehnholz SP, Nissan MH, Zhang H, Kundra R, Nandakumar S, Lu C, et al. Quantifying the expanding landscape of clinical actionability for patients with cancer. Cancer Discov. 2024;14(1):49–65. https://doi.org/10.1158/2159-8290.CD-23-0467 PMID: 37849038
62. Chakravarty D, Gao J, Phillips SM, Kundra R, Zhang H, Wang J, et al. OncoKB: a precision oncology knowledge base. JCO Precis Oncol. 2017;2017:PO.17.00011. https://doi.org/10.1200/PO.17.00011 PMID: 28890946
63. Patterson SE, Liu R, Statz CM, Durkin D, Lakshminarayana A, Mockus SM. The clinical trial landscape in oncology and connectivity of somatic mutational profiles to targeted therapies. Hum Genomics. 2016;10:4. https://doi.org/10.1186/s40246-016-0061-7 PMID: 26772741
64. Sondka Z, Dhir NB, Carvalho-Silva D, Jupe S, Madhumita, McLaren K, et al. COSMIC: a curated database of somatic variants and clinical data for cancer. Nucleic Acid Res. 2024;52(D1):D1210–7.
65. Tuteja S, Kadri S, Yap KL. A performance evaluation study: Variant annotation tools - the enigma of clinical next generation sequencing (NGS) based genetic testing. J Pathol Inform. 2022;13:100130.
66. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acid Res. 2019;47(Database issue):D886–94.
67. Alirezaie N, Kernohan KD, Hartley T, Majewski J, Hocking TD. ClinPred: prediction tool to identify disease-relevant nonsynonymous single-nucleotide variants. Am J Hum Genet. 2018;103(4):474–83.
68. Tian Y, Pesaran T, Chamberlin A, Fenwick RB, Li S, Gau C-L, et al. REVEL and BayesDel outperform other in silico meta-predictors for clinical variant classification. Sci Rep. 2019;9(1):12752. https://doi.org/10.1038/s41598-019-49224-8 PMID: 31484976
69. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. Am J Hum Genet. 2016;99(4):877–85. https://doi.org/10.1016/j.ajhg.2016.08.016 PMID: 27666373
70. Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum Mol Genet. 2015;24(8):2125–37. https://doi.org/10.1093/hmg/ddu733 PMID: 25552646
71. Cheng J, Novati G, Pan J, Bycroft C, Žemgulytė A, Applebaum T, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science. 2023;381(6664):eadg7492. https://doi.org/10.1126/science.adg7492 PMID: 37733863
72. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li YI. Predicting splicing from primary sequence with deep learning. Cell. 2019;176(3):535-548.e24.
73. Bao R, Huang L, Andrade J, Tan W, Kibbe WA, Jiang H, et al. Review of current methods, applications, and data management for the bioinformatics analysis of whole exome sequencing. Cancer Inform. 2014;13(Suppl 2):67–82.
74. Carson AR, Smith EN, Matsui H, Brækkan SK, Jepsen K, Hansen J-B, et al. Effective filtering strategies to improve data quality from population-based whole exome sequencing studies. BMC Bioinformatics. 2014;15:125. https://doi.org/10.1186/1471-2105-15-125 PMID: 24884706
75. Sefid Dashti MJ, Gamieldien J. A practical guide to filtering and prioritizing genetic variants. BioTechniques. 2017;62(1):18–30.
76. Malcikova J, Tausch E, Rossi D, Sutton LA, Soussi T, Zenz T, et al. ERIC recommendations for TP53 mutation analysis in chronic lymphocytic leukemia-update on methodological approaches and results interpretation. Leukemia. 2018;32(5):1070–80. https://doi.org/10.1038/s41375-017-0007-7 PMID: 29467486
77. Cheng Y-W, Stefaniuk C, Jakubowski MA. Real-time PCR and targeted next-generation sequencing in the detection of low level EGFR mutations: Instructive case analyses. Respir Med Case Rep. 2019;28:100901. https://doi.org/10.1016/j.rmcr.2019.100901 PMID: 31367517
78. Pandzic T, Ladenvall C, Engvall M, Mattsson M, Hermanson M, Cavelier L, et al. Five percent variant allele frequency is a reliable reporting threshold for TP53 variants detected by next generation sequencing in chronic lymphocytic leukemia in the clinical setting. Hemasphere. 2022;6(8):e761. https://doi.org/10.1097/HS9.0000000000000761 PMID: 35935605
79. Chen H, Zhang Y, Wang B, Liao R, Duan X, Yang C, et al. Characterization and mitigation of artifacts derived from NGS library preparation due to structure-specific sequences in the human genome. BMC Genomics. 2024;25(1):227. https://doi.org/10.1186/s12864-024-10157-w PMID: 38429743
80. Broad Institute. (How to) Filter variants either with VQSR or by hard-filtering. GATK; 2025 [cited 2025 May 8]. Available from: https://gatk.broadinstitute.org/hc/en-us/articles/360035531112--How-to-Filter-variants-either-with-VQSR-or-by-hard-filtering
81. Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 1998;8(3):186–94.
82. Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. 1998;8(3):175–85.

83. McNulty SN, Parikh BA, Duncavage EJ, Heusel JW, Pfeifer JD. Optimization of population frequency cutoffs for filtering common germline polymorphisms from tumor-only next-generation sequencing data. J Mol Diagn. 2019;21(5):903–12. https://doi.org/10.1016/j.jmoldx.2019.05.005 PMID: 31251990
84. Cooper GM. Oncogenes. In: The cell: a molecular approach [Internet]. 2nd ed. Sunderland (MA): Sinauer Associates; 2000 [cited 2025 May 25]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK9840/
85. Knudson AG. Two genetic hits (more or less) to cancer. Nat Rev Cancer. 2001;1(2):157–62. https://doi.org/10.1038/35101031 PMID: 11905807
86. Griffith M, Spies NC, Krysiak K, McMichael JF, Coffman AC, Danos AM, et al. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. Nat Genet. 2017;49(2):170–4. https://doi.org/10.1038/ng.3774 PMID: 28138153
87. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(43):15545–50. https://doi.org/10.1073/pnas.0506580102 PMID: 16199517
88. Mootha VK, Lindgren CM, Eriksson K-F, Subramanian A, Sihag S, Lehar J, et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet. 2003;34(3):267–73. https://doi.org/10.1038/ng1180 PMID: 12808457
89. Hanahan D. Hallmarks of cancer: new dimensions. Cancer Discov. 2022;12(1):31–46. https://doi.org/10.1158/2159-8290.CD-21-1059 PMID: 35022204
90. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29(1):24–6. https://doi.org/10.1038/nbt.1754 PMID: 21221095
91. Robinson JT, Thorvaldsdóttir H, Wenger AM, Zehir A, Mesirov JP. Variant review with the integrative genomics viewer (IGV). Cancer Res. 2017;77(21):e31–4.
92. Bruford EA, Braschi B, Denny P, Jones TEM, Seal RL, Tweedie S. Guidelines for human gene nomenclature. Nat Genet. 2020;52(8):754–8. https://doi.org/10.1038/s41588-020-0669-3 PMID: 32747822
93. Morales J, Pujar S, Loveland JE, Astashyn A, Bennett R, Berry A, et al. A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. Nature. 2022;604(7905):310–5. https://doi.org/10.1038/s41586-022-04558-8 PMID: 35388217
94. Den Dunnen JT, Dalgleish R, Maglott DR, Hart RK, Greenblatt MS, McGowan-Jordan J. HGVS recommendations for the description of sequence variants: 2016 Update. Human Mutation. 2016;37(6):564–9.
95. MacArthur DG, Manolio TA, Dimmock DP, Rehm HL, Shendure J, Abecasis GR, et al. Guidelines for investigating causality of sequence variants in human disease. Nature. 2014;508(7497):469–76.
96. Lefter M, Vis JK, Vermaat M, den Dunnen JT, Taschner PEM, Laros JFJ. Mutalyzer 2: next generation HGVS nomenclature checker. Bioinformatics. 2021;37(18):2811–7. https://doi.org/10.1093/bioinformatics/btab305
97. Zhang J, Yao Y, He H, Shen J. Clinical interpretation of sequence variants. Curr Protoc Hum Genet. 2020;106(1):e98. https://doi.org/10.1002/cphg.98 PMID: 32176464
98. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:160018. https://doi.org/10.1038/sdata.2016.18 PMID: 26978244
99. Teperek M. Data Management Guide [Internet]. 2015 [cited 2025 May 8]. Available from: https://www.data.cam.ac.uk/data-management-guide
100. Tanjo T, Kawai Y, Tokunaga K, Ogasawara O, Nagasaki M. Practical guide for managing large-scale human genome data in research. J Hum Genet. 2021;66(1):39–52. https://doi.org/10.1038/s10038-020-00862-1 PMID: 33097812
101. Ochoa I, Hernaez M, Goldfeder R, Weissman T, Ashley E. Effect of lossy compression of quality scores on variant calling. Brief Bioinform. 2017;18(2):183–94. https://doi.org/10.1093/bib/bbw011 PMID: 26966283
102. Government of Canada I. Tri-Agency Research Data Management Policy - Frequently Asked Questions [Internet]. Innovation, Science and Economic Development Canada; 2024 [cited 2025 May 8]. Available from: https://science.gc.ca/site/science/en/interagency-research-funding/policies-and-guidelines/research-data-management/tri-agency-research-data-management-policy-frequently-asked-questions
103. Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. Identifying personal genomes by surname inference. Science. 2013;339(6117):321–4.
104. Sorani MD, Yue JK, Sharma S, Manley GT, Ferguson AR, Cooper SR, et al. Genetic data sharing and privacy. Neuroinformatics. 2015;13(1):1–6.
105. Sandve GK, Nekrutenko A, Taylor J, Hovig E. Ten simple rules for reproducible computational research. PLoS Comput Biol. 2013;9(10):e1003285. https://doi.org/10.1371/journal.pcbi.1003285 PMID: 24204232
106. GitHub [Internet]. GitHub: Build and ship software on a single, collaborative platform. 2025 [cited 2025 May 8]. Available from: https://github.com/
107. The most-comprehensive AI-powered DevSecOps platform [Internet]. [cited 2025 May 8]. Available from: https://about.gitlab.com/
108. Köster J, Rahmann S. Snakemake-a scalable bioinformatics workflow engine. Bioinformatics. 2018;34(20):3600. https://doi.org/10.1093/bioinformatics/bty350 PMID: 29788404
109. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol. 2017;35(4):316–9. https://doi.org/10.1038/nbt.3820 PMID: 28398311
110. Grüning B, Dale R, Sjödin A, Chapman BA, Rowe J, Tomkins-Tinch CH, et al. Bioconda: sustainable and comprehensive software distribution for the life sciences. Nat Methods. 2018;15(7):475–6. https://doi.org/10.1038/s41592-018-0046-7 PMID: 29967506

111. Merkel D. Docker: lightweight Linux containers for consistent development and deployment. Linux J. 2014;2014(239):2:2.
112. Wratten L, Wilm A, Göke J. Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. Nat Methods. 2021;18(10):1161–8. https://doi.org/10.1038/s41592-021-01254-9 PMID: 34556866
113. GitHub Docs [Internet]. About READMEs. [cited 2025 May 26]. Available from: https://docs-internal.github.com/en/repositories/managing-your-repositorys-settings-and-features/customizing-your-repository/about-readmes
114. The American Journal of Human Genetics: Cell Press [Internet]. [cited 2025 May 8]. Available from: https://www.cell.com/ajhg/home
115. Oxford Academic [Internet]. Research data. [cited 2025 May 8]. Available from: https://academic.oup.com/pages/open-research/research-data
116. Chacon S, Straub B. Pro Git [Internet]. 2nd ed. Berkeley (CA): Apress; 2014 [cited 2025 Jun 13]. Available from: https://git-scm.com/book/en/v2