

Bioinformatics · Practical notes

Genomic Variant Analysis

From Sequencing Reads to SNP and Indel Discovery,
Annotation, Visualization and Interpretation

This note explains reference-based small-variant analysis across organisms, from study design and ploidy to read evidence, variant discovery, filtering, annotation and interpretation. A shared foundation connects the biological question to the assumptions of each computational stage. Separate routes address human germline data, haploid isolates, non-model organisms, polyploid and pooled samples, long reads and somatic variation. Worked records and companion phased scripts illustrate decisions, execution contracts and failure modes; they do not constitute a completed real-data reproduction or a validated whole-genome pipeline. The English and Turkish editions preserve the same examples, scientific qualifications and bibliography.

Companion scripts:
main_scripts/genomic_variants/

Tuna Birgün
github.com/tunabirgun

Original note · 22 August 2026

Contents

1 From a biological question to a variant study **2**
1.1 Define the observation before selecting software . . . 2
1.2 Design information that analysis cannot reconstruct . . . 2
1.3 Further reading . . . 3
2 Alleles, genotypes and the reference genome **3**
2.1 Ploidy changes the question asked of the reads . . . 3
2.2 Four kinds of confidence . . . 3
3 Read the files as a chain of evidence **4**
3.1 A synthetic alignment . . . 4
3.2 VCF describes alleles and their sample states . . . 4
3.3 Coordinates require explicit conversion . . . 5
3.4 What a reference-confidence file adds . . . 5
4 Prepare a compatible reference and sample record **5**
4.1 Keep sample, library and run identities distinct . . . 6
4.2 Inspect before running expensive stages . . . 6
4.3 Reference choice is an analytical decision . . . 6
5 Make short-read evidence interpretable **7**
5.1 Alignment proposes genomic origins . . . 7
5.2 Duplicates describe dependence between observations . . . 7
5.3 Base-quality recalibration is conditional . . . 8
6 Discover alleles and genotype a cohort **8**
6.1 A bounded GATK example . . . 9
6.2 Why joint genotyping is a separate operation . . . 9
6.3 Inspect calls before assigning biological meaning . . . 9
7 Filtering: decide which evidence can support a result **10**
7.1 Three levels of filtering . . . 10
7.2 Choose rules from the experiment . . . 10
7.3 Callable regions define the denominator . . . 11
8 Normalization: make variant representations comparable **11**
8.1 Reference-aware normalization . . . 11
8.2 Multiallelic sites carry relationships . . . 12
8.3 Haplotypes and consequence interpretation . . . 12
9 Annotation: connect a sequence difference to a biological model **12**
9.1 Choose a compatible annotation bundle . . . 12
9.2 Illustrative annotation commands . . . 13
9.3 One allele can have several consequences . . . 13
9.4 Add evidence without losing its context . . . 14
10 Validation: measure the claim within an explicit territory **14**
10.1 Truth includes regions as well as variants . . . 14
10.2 Compare haplotypes and report the denominator . . . 15
10.3 Investigate discordances and selected candidates . . . 15
11 Special routes: change the model when the biology changes **15**
11.1 Human germline samples . . . 16
11.2 Other diploid organisms: rebuild the resource contract . . . 16
11.3 Haploid isolates, microbes, and fungi . . . 16
11.4 Polyploids and pooled samples . . . 17
11.5 Long reads and somatic variation . . . 17
12 Visualize evidence at callset and locus scales **18**
12.1 Choose a diagnostic display . . . 18
12.2 Review an allele in local IGV . . . 18
12.3 Reproduce the displayed summary . . . 18
13 Interpretation and reporting: preserve the chain of evidence **19**
13.1 A worked synthetic candidate . . . 19
13.2 Reproducibility . . . 20
13.3 State what was actually verified . . . 20
14 A staged workflow from reads to an evaluated callset **20**
14.1 Shared preparation and explicit biological branches . . . 21
14.2 Compare aligners and somatic callers fairly . . . 21
14.3 Reproducibility as an execution record . . . 22
14.4 Map the same protocol to a local Galaxy installation . . . 22
References **23**

1 From a biological question to a variant study

A genomic variant study asks how the DNA sequence carried by a sample differs from a defined reference and what those differences may mean. Sequencing does not observe a complete chromosome directly. It produces reads from DNA molecules; alignment proposes their genomic origins, and a caller evaluates the evidence for alternative sequence states. Annotation then relates a reported allele to genes, transcripts and other features. These operations answer different questions, so uncertainty introduced early in the analysis remains relevant when interpreting the final table.

This note follows reference-based discovery of single-nucleotide variants and short insertions and deletions in genomic DNA. The teaching spine is paired-end short-read germline sequencing. The same concepts recur in other organisms and technologies, but a human diploid workflow is not a species-independent recipe. Later sections separate haploid organisms, polyploids, pooled samples, long reads, organelles and somatic samples. Structural variants, copy-number changes, repeat expansions and assembly-based discovery need additional methods; a small-variant VCF cannot establish their absence.

1.1 Define the observation before selecting software

Begin with the biological unit. An individual organism, a tissue specimen, an isolate and a pool of individuals represent different sampling designs. A DNA extraction from a nominally single organism can also contain organellar DNA, symbionts or contaminants. Record what the sample is intended to represent and which sources of variation the experiment can distinguish. A low-frequency allele in a pooled sample, for example, does not identify the individual carrying it. In a mixed infection it need not represent heterozygosity within one genome.

The question also determines the region under study. Whole-genome sequencing samples both coding and noncoding sequence, whereas capture and amplicon assays deliberately enrich selected regions. A capture target is an intended design region, not a guarantee that every base has usable evidence. Assess the sequenced, callable portion of the target before reporting how many variants were absent. Comparisons between samples require compatible definitions of both the biological unit and the regions examined.

Germline discovery usually seeks alleles present in the inherited genome of an individual. Somatic discovery seeks changes acquired in non-germline cells after conception. Such changes may occur in some or all cells of a sampled clone; interpretation must account for purity, copy number and potentially a matched normal. The two tasks use different evidence models. The official GATK germline workflow provides one coherent route from prepared alignments through per-sample reference-confidence files to cohort genotyping; its existence does not validate that route for every assay or species [6].

Key parameters

Before downloading data, write a short analysis contract: biological unit; assay and read technology; reference assembly; expected ploidy by region; germline, somatic or mixture model; discovery regions; intended annotation; and available independent checks. Record which choices remain assumptions. Resolve choices that change the genotype model before interpreting calls.

1.2 Design information that analysis cannot reconstruct

Biological replication, sequencing depth and technical replication address different uncertainties. More reads from the same library can improve evidence at covered loci, but do not create more independent individuals. Resequencing a library may expose a run-specific problem while preserving errors introduced during extraction or library preparation. An independently prepared library or an orthogonal assay tests a different part of the evidence chain. Choose controls according to the claim they must support rather than treating all repeated measurements as interchangeable.

Balance batches across biological groups whenever the experiment permits it. If all members of one population were sequenced with one library protocol and all members of another with a different protocol, a downstream variant-frequency contrast may also reflect technical differences. Keeping batch labels permits investigation; it cannot always recover information missing from the design. Family relationships, population structure and related isolates similarly affect later inference and should be preserved in study metadata.

A useful study deliverable is therefore more than a list of differences. It includes the reference and sample definitions, the effective analysis regions, evidence and filter states, software and resource versions, and interpretation limits. The following chapters explain how each intermediate file contributes to that deliverable. Examples use invented contigs and sample names. They illustrate formats and decisions; no real dataset, complete pipeline run or performance result is claimed.

1.3 Further reading

For broader conceptual study, Brown's *Genomes 5* provides a genetics and genomics textbook, while Van der Auwera and O'Connor's *Genomics in the Cloud* offers a computational companion focused on genomic data analysis [1, 2]. These books provide conceptual background rather than specifications for current software behavior. Version-sensitive commands here are checked against official tool documentation.

2 Alleles, genotypes and the reference genome

A reference genome supplies coordinates and a comparison sequence. It is not a statement that its alleles are ancestral, common or biologically preferable. A sample can match the reference at one position and carry an alternative allele at another. Changing the assembly can change coordinates, local sequence context and the representation of a variant. A variant therefore needs an assembly and a contig, not merely a position and a pair of letters.

A single-nucleotide variant (SNV) substitutes one nucleotide for another. The term SNP is often used operationally in calling software, although a polymorphism historically implies variation within a population. An insertion adds sequence relative to the reference; a deletion removes it. These are relative descriptions: the allele labelled an insertion against one reference may not be the derived evolutionary event. Here, short indels mean the small insertion/deletion alleles handled by the chosen small-variant method, without imposing a universal length boundary.

2.1 Ploidy changes the question asked of the reads

Ploidy is the number of chromosome copies represented by the genotype model at a locus. In a diploid biallelic model, the unphased possibilities are two reference alleles, one reference and one alternative allele, and two alternative alleles. In a haploid model there is one allele copy; in a tetraploid model the alternative allele can have several dosages. A caller that can encode a tetraploid genotype still needs adequate evidence and an appropriate model to distinguish those dosages. Setting a ploidy argument does not resolve mapping between similar homeologous chromosomes.

Ploidy may vary within a sample. Sex chromosomes, organelles, aneuploid regions and copy-number changes can violate a single genome-wide value. Determine the biological and reference-specific regions to which each model applies. For a polyploid with a reference that collapses multiple homeologs, unusually deep coverage and apparent heterozygosity may reflect several genomic copies mapping onto one location. These observations motivate investigation of the assembly and alignment, not an automatic biological interpretation.

VCF genotype indices refer to allele order. Index 0 is the reference allele; 1 is the first alternative allele and 2 the second. A diploid `0/1` and a haploid `1` therefore express different models. A slash denotes unphased allele separation; a vertical bar denotes phased separation. Phase states that alleles have been assigned to chromosome copies within the specified phasing context; it should not be inferred merely because variants are nearby [3].

2.2 Four kinds of confidence

The Phred transformation expresses a probability on a logarithmic scale:

$$Q = -10 \log_{10}(p).$$

For an error probability of 10^{-3}, the corresponding score is 30. This is a mathematical illustration, not a filtering recommendation. The event represented by p matters. A base quality concerns a nucleotide observation; mapping quality concerns read placement; genotype quality concerns a genotype assignment

under a model; site quality concerns evidence associated with the variant record. Scores from different levels cannot be substituted merely because they use the same numerical scale [4, 3].

Consider a site with convincing evidence that an alternative allele exists somewhere in a cohort. One low-depth sample may nevertheless have an uncertain genotype at that site. Conversely, a sample can be confidently reference at a position where other samples carry an alternative allele. Site acceptance and genotype acceptance must therefore be separate decisions. A cohort-wide quality value does not make every sample's genotype reliable.

Watch out for

A missing genotype is not a reference genotype. In diploid notation, `./.` means that the genotype is not supplied; `0/0` asserts a reference/reference call. Replacing missing values with reference calls changes allele counts, missingness and downstream biological conclusions. An absent record in a variants-only VCF likewise does not establish adequate reference evidence at that position.

Read support is evidence rather than a vote count. Ten observations from distinct molecules with credible placement differ from ten duplicate reads from one amplified fragment. Strand, read position, local sequence context and nearby mismatches may explain an apparent allele. The caller combines some of this information according to its model; quality control and later review examine whether that model describes the sample well enough for the intended inference.

3 Read the files as a chain of evidence

FASTQ holds read sequences and base qualities. In a conventional four-line record, the identifier starts with `@`, the second line contains sequence, the third begins with `+`, and the fourth encodes per-base qualities. Sequence and quality strings must have equal lengths. Paired-end files describe the two sequenced ends of fragments; their pairing must survive filtering. A read name identifies an observation, not a biological replicate.

SAM is a text alignment format; BAM stores the corresponding alignment information in binary form. CRAM is another alignment format with its own reference and decoding considerations. An alignment describes where and how a read was placed, including flags, mapping quality and an alignment-operation string called CIGAR. A sorted, indexed BAM supports regional access, but sorting and indexing do not establish correct mapping or intact sample identity [4, 5].

3.1 A synthetic alignment

The following record is schematic; spaces are shown for readability where a real SAM uses tabs. Assume that bases 101–106 of contig `toy1` are `ACGTAC`.

Example records

```
r001  0  toy1  101  60  6M  *  0  0  ACGTTC  IIIIII
```

The read starts at the one-based reference position 101. Its six aligned bases span positions 101–106; the read's fifth base, T, differs from the assumed reference A at position 105. CIGAR `6M` means six alignment matches or mismatches, not six verified nucleotide identities. In this toy record, a mismatch can therefore occur inside `M`. The letters `I` in the quality string are encoded qualities, not insertion operations. A single such observation is insufficient to establish a sample genotype.

3.2 VCF describes alleles and their sample states

A VCF header declares the file version, contigs and the meanings of annotations. Each data record contains a chromosome, position, identifier, reference and alternative allele, quality, filter state and information fields. Sample columns are interpreted through the FORMAT field. BCF is a binary representation of variant data. The header is part of the scientific contract: identical-looking tags need compatible definitions before files can be combined [3].

Example records

```
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT sample_A
 toy1 105 .  A   T   60   PASS   .    GT:DP:GQ 0/1:18:35
 toy1 110 .  AT  A   45   PASS   .    GT:DP:GQ ./.:3:.
```

These invented rows are a readable excerpt, not a complete valid VCF: declarations and tab delimiters are omitted. The first proposes A-to-T variation at position 105 and supplies a diploid heterozygous genotype for `sample_A`. The second represents deletion of the T following an anchor A at position 110, but provides no genotype for this sample. Its `PASS` value concerns the record's declared site-filter scheme. It cannot override the missing genotype. The illustrative depth and quality numbers demonstrate fields rather than recommended cutoffs.

3.3 Coordinates require explicit conversion

SAM text positions and VCF positions use one-based coordinates. BED intervals conventionally use a zero-based start and an excluded end. Thus, reference bases 101 through 106 correspond to BED `toy1 100 106`. Their length is $106 - 100 = 6$. The single base at VCF position 105 corresponds to BED `toy1 104 105`. Command-line region syntax and internal binary coordinates need not use the same convention, so check the particular tool interface before transferring an interval [4].

For the anchored deletion above, the VCF reference allele covers positions 110–111, corresponding to BED `toy1 109 111`; the deleted T alone occupies `toy1 110 111`. These intervals answer different questions. Intersecting variants with features may consider an anchor, the altered bases or a tool-defined variant span. State which definition is used, particularly for insertions and deletions at exon boundaries. Never fix a suspected coordinate problem by shifting every position without testing a known synthetic case.

3.4 What a reference-confidence file adds

A variants-only VCF primarily records reported variant sites. A gVCF additionally carries information about evidence for reference and unobserved alternative states across examined regions. In the GATK route, blocks compress stretches with similar reference confidence, and symbolic alleles such as `<NON_REF>` participate in the model. They are not ordinary observed insertions or literal DNA sequences. A gVCF is an intermediate for compatible downstream genotyping, not simply a larger final VCF [7].

Reference-confidence information helps distinguish a sample with evidence for the reference from one lacking usable evidence when another sample contributes an alternative allele. Its interpretation still depends on coverage, caller conventions and the processed intervals. The filename suffix alone cannot establish interoperability between callers. Preserve the generating method and version, and verify that the selected joint-genotyping tool accepts that particular gVCF model.

4 Prepare a compatible reference and sample record

A reference bundle joins several resources that must describe the same coordinate system. The FASTA provides sequence; its index enables retrieval; a sequence dictionary describes contigs for software that requires it. Gene annotations, target intervals and known-variant resources add different information. Files bearing the same species name can still differ in assembly release, contig naming, organellar sequence, alternate scaffolds or annotation release. Compatibility must be demonstrated from their contents and provenance.

Start by selecting the assembly that fits the study and available resources. Record its provider, accession or release, included sequence set and any modifications. A chromosome-level assembly with alternate sequences is not interchangeable with a primary-only subset. Adding or removing sequences changes the alternatives available to the aligner. Preserve a local integrity record for the exact FASTA and derive indexes from that file. Reusing an index whose filename happens to match is unsafe when its underlying sequence has changed.

Resource	Compatibility question
Reference FASTA	Which assembly and sequence subset define the coordinates?
Gene annotation	Do contig names, coordinates and release provenance match that assembly?
Target intervals	Which coordinate convention and assay design generated them?
Known sites	Are alleles and positions valid against this exact reference?
Sample sheet	Do identifiers distinguish individuals, libraries and sequencing units?

Contig renaming is a limited operation. Changing `1` to `chr1` can reconcile names only when the sequences and coordinate systems already agree. It cannot convert one assembly to another. Coordinate conversion between assemblies requires a suitable mapping and verification of transformed alleles; some loci fail to map or change representation. For a new analysis, obtaining resources natively prepared for the chosen assembly often avoids preventable ambiguity.

4.1 Keep sample, library and run identities distinct

The sample sheet should preserve the relationship between biological sample, extraction, library and sequencing unit. Multiple lanes from one library contribute reads to the same biological sample. Different libraries from the same individual may need distinct library labels even when their reads are later analyzed together. Read groups carry identifiers within alignment files, including sample and library information. These fields affect grouping in downstream tools; they cannot be treated as decorative labels [4].

Before merging alignments, confirm that all intended inputs belong to the expected sample, use the same reference and retain the correct read-group relationships. A technically successful merge can still combine two individuals accidentally. Compare recorded identity with available experimental or genetic checks appropriate to the organism. Where metadata cannot be publicly shared, retain a controlled local mapping rather than placing sensitive information in read names or exported filenames.

The workflow record should connect each final result to its inputs, reference resources, software environment and parameter choices. Record changes that affect evidence, such as read trimming, duplicate policy, interval restriction, ploidy and filtering. Keep the original received data separate from derived files. This makes it possible to explain whether a changed callset resulted from new reads, a different model or a different reference rather than treating every file as a fresh independent result.

4.2 Inspect before running expensive stages

Small structural checks can catch incompatible inputs before alignment or calling consumes substantial resources. Confirm expected filenames and file types, inspect headers, verify sample names and compare contig dictionaries. Check whether compressed variant files use the compression and index accepted by the next tool. A nonempty file and a successful indexing operation show only limited properties; neither establishes biological validity.

Shell

```
# Illustrative checks on an existing prepared alignment.
samtools quickcheck sample.ready.bam
samtools view -H sample.ready.bam
samtools flagstat sample.ready.bam
```

These commands follow the official SAMtools interface [5]. `quickcheck` checks header and end-of-file features rather than reading every alignment for corruption. Header inspection exposes reference and read-group declarations; `flagstat` summarizes alignment flags. Interpret their outputs together with expected samples and assay characteristics. The commands have been checked against documentation, not executed on a study dataset in this note.

4.3 Reference choice is an analytical decision

A complete or more diverse reference can expose sequence that an older linear assembly represents poorly. The draft human pangenome study demonstrated changes in variant discovery under its evaluated graph-based workflows; this does not mean that replacing the FASTA in a linear pipeline reproduces that method [37]. A new assembly changes coordinates, alignments and resource compatibility. Rebuild indexes and

use matching truth and annotation resources; record a coordinate-conversion audit if comparing assemblies. The review by Olson and colleagues places these choices in the context of complete human genome sequences and benchmarking [36]. Neither review guidance nor human performance measurements establish accuracy in a non-model organism.

5 Make short-read evidence interpretable

Preprocessing prepares reads for a particular calling model. Its purpose is to identify technical problems and represent the evidence consistently, not to make every quality plot resemble an idealized example. Examine read quality profiles, adapter sequence, base composition, read lengths and overrepresented sequences in the context of the assay. A biased base composition can arise from the organism or library design; a quality warning is a diagnostic observation rather than an instruction to discard data.

Trimming is useful when unwanted adapter sequence or poor read ends impair the analysis, but it changes the reads supplied to the aligner. Preserve pair relationships and record the retained read lengths and counts. Aggressive trimming can remove genuine informative sequence and reduce unique placement. Choose any trimming rule for the library and calling route, then inspect its consequences. Do not adopt a universal minimum length or quality cutoff without considering the available sequence and intended analysis.

The optional companion uses fastp 1.4.0. Its `--qualified_quality_phred 15` defines a qualified base for a read-level filter; it does not trim every base below Q15 or require a mean quality of 15. The default `--unqualified_percent_limit 40` allows at most 40% unqualified bases. In an invented 100-base read, 30 bases below Q15 do not fail this rule, whereas 41 do; other filters can still reject either read. `--length_required 15` rejects reads shorter than 15 bases after processing. Adapter trimming is enabled, poly-G tail trimming is automatically enabled for recognized NextSeq/NovaSeq data, and the sliding-window quality-cut options and poly-X trimming are not enabled by this wrapper [42].

Raising the qualified-base threshold can make more bases count as unqualified without shortening the read. Enabling a quality-cut operation instead removes sequence and can then trigger the length filter. Compare retained pairs, the rejection reasons and length distributions in fastp reports, followed by mapping and callable coverage; stricter settings need not improve the final calls.

5.1 Alignment proposes genomic origins

An aligner seeks plausible placements of each read against the chosen reference. Paired-end information constrains placements through the relationship between fragment ends. Repeats and highly similar genes can support multiple origins; divergence from the reference and sequencing errors introduce further ambiguity. A mapping-quality score summarizes the aligner's placement confidence under its own model. It does not prove that the selected reference contains the correct genomic copy.

After alignment, inspect mapped fractions, coverage distribution, fragment-size behavior, duplicate metrics and discordant or supplementary alignments as appropriate to the assay. Compare samples and sequencing units rather than relying on one aggregate value. A library with adequate mean coverage may still leave GC-rich regions, repeated loci or capture targets poorly represented. Coverage across the intended analysis region therefore matters more than a single genome-wide average when interpreting missing calls.

Coordinate sorting arranges alignments for region-based processing; indexing enables efficient retrieval. Keep read groups intact and check expected contigs. Avoid reusing an RNA-seq alignment recipe for genomic DNA merely because both produce BAM files. Splice-aware alignment and transcript-derived coverage address different assumptions from genomic variant discovery. Likewise, a long-read caller may require a different aligner and platform-specific model from the short-read route described here.

5.2 Duplicates describe dependence between observations

Library amplification can produce several reads from copies of one original fragment. Treating these copies as independent molecules can exaggerate support for an allele. Duplicate marking assigns a flag so that downstream software can apply its documented policy. It differs from deleting all reads with identical

sequence. Naturally identical fragments, optical duplicates and amplified copies require interpretation in the context of the library and duplicate method [4, 7].

Do not apply an ordinary coordinate-based duplicate policy blindly to an amplicon assay: its design deliberately creates many molecules with the same endpoints. Libraries with unique molecular identifiers need a method that interprets those identifiers and their error model. Duplicate metrics can also differ with input amount, sequencing depth and library complexity. A high duplicate fraction motivates investigation; it does not by itself identify the cause or define a universal rejection threshold.

The order of operations matters. A duplicate tool may require mate information, a particular sort order or library labels before coordinate-sorted output is produced. Follow the chosen implementation's complete preparation contract. A short command copied out of that sequence may run yet mark a different set of reads than intended. The companion preparation scripts make this sequence explicit, including the intermediate sort orders.

5.3 Base-quality recalibration is conditional

Base quality score recalibration (BQSR) estimates systematic differences between reported base qualities and empirical mismatch patterns. It adjusts confidence values rather than repairing the nucleotide sequence. The GATK procedure models covariates such as sequence context and read cycle [10]. Known variant sites are masked so that genuine biological differences are not automatically treated as sequencing errors [8].

That requirement creates a practical distinction between well-resourced and poorly resourced organisms. A suitable known-sites set must match the reference and be reliable enough for its masking purpose; being downloadable is insufficient. A bootstrap route can use a preliminary callset to construct candidate masking sites, but dependence on the same reads and model remains. Iteration is not independent confirmation that the final calls are correct. Record how bootstrap sites were obtained and assess the procedure with separate evidence where possible.

BQSR is not a required ritual for every sequencing platform or caller. Determine whether the chosen workflow expects it, whether the data's quality model is appropriate and whether necessary resources exist. Compare diagnostic behavior rather than assuming that the presence of an extra preprocessing stage guarantees improvement. BQSR also differs from variant quality score recalibration, which models annotations on called variants later in the workflow. The similar names refer to different objects and decisions.

Watch out for

The output of preprocessing is an analysis-ready alignment only relative to a stated caller contract. Sorting, duplicate marking and recalibration do not establish that a sample is uncontaminated, that ploidy is correct or that every locus is callable. Preserve unresolved sample and reference problems before proceeding to genotype interpretation.

6 Discover alleles and genotype a cohort

Variant calling evaluates whether read evidence is better explained by one sequence state than by alternatives. In a simple conceptual model, the caller evaluates the likelihood of the observed reads D under candidate genotype G, written $P(D \mid G)$. A posterior genotype probability also depends on prior assumptions. This distinction explains why identical read counts need not produce identical calls under different models, ploidies or parameter settings. The model and its assumptions belong in the interpretation of a genotype.

A caller must also decide which candidate alleles to evaluate. Near an indel, several read alignments may describe the same local haplotype differently. Assembly-based methods can reconstruct candidate local sequences and compare read evidence with them rather than considering each mismatch independently. HaplotypeCaller is an example of a local-assembly caller; FreeBayes provides another haplotype-based framework. Their supported assumptions and outputs must be evaluated separately rather than treating all VCF-producing tools as interchangeable [7, 14].

6.1 A bounded GATK example

The following syntax illustrates one per-sample reference-confidence call. It assumes a compatible indexed reference with its dictionary, a prepared coordinate-sorted and indexed BAM with valid sample/read groups, and a genuinely diploid analysis region. The filenames are invented. They do not identify supplied data or constitute a validated execution environment.

Shell

```
gatk HaplotypeCaller \
  -R reference.fa \
  -I sample.ready.bam \
  -O sample.g.vcf.gz \
  -ERC GVCF \
  --sample-ploidy 2
```

The output is a per-sample gVCF for a compatible genotyping stage. The value 2 explicitly specifies the teaching example's diploid model; it is not a cross-species recommendation. HaplotypeCaller supports non-diploid use cases, but the number of possible genotypes and computational burden can grow with ploidy and allele count. Current documentation also provides region-dependent ploidy handling, whose active-region boundary behavior requires attention. Verify the exact installed version before adopting such options [7].

This compact example omits study-specific intervals, resource limits and options that depend on library preparation. It is deliberately not an executable workflow from raw FASTQ. For a documented PCR-free library, GATK 4.6.2.0 specifies `--pcr-indel-model NONE`: this disables the specialized PCR indel-error model rather than trimming reads or removing duplicates. The default is `CONSERVATIVE`; do not change it merely to obtain more indels. Review indel evidence after a justified change, and determine whether the chosen caller expects recalibrated qualities. Inspect the installed help and preserve the effective arguments. Successful syntax alone is not evidence that the biological model fits the sample.

6.2 Why joint genotyping is a separate operation

Independent variants-only VCFs do not contain equivalent evidence at all sites. If sample A reports an allele and sample B omits that position, B may be reference, poorly covered or outside the analyzed region. Simply merging rows cannot distinguish these cases. In a reference-confidence workflow, per-sample files retain information needed to revisit shared candidate sites, and joint genotyping produces a cohort representation with sample-specific genotypes [6].

The GATK route can import compatible single-sample gVCFs into GenomicsDB before genotyping. Import is a storage and data-access step; it is not itself the statistical decision that emits the final cohort genotypes. The import requires consistent references and compatible headers, unique and correct sample identities, and specified analysis intervals. The documented sample-map optimization relaxes some header-merging checks, making prior compatibility checks especially important [9].

Cohort membership becomes part of the result's provenance. Adding samples can change the set of discovered alleles and downstream annotations or filtering summaries; retain the cohort definition and version. Relatedness and population structure also matter to later analyses and any priors used. Joint genotyping does not remove contamination, repair a collapsed reference or turn low-confidence evidence into an independent observation.

6.3 Inspect calls before assigning biological meaning

An initial callset is a collection of model-based candidates. Inspect sample-level missingness, depth and quality distributions, variant classes and genomic distribution in relation to the assay. Look for batch-linked changes and unexpectedly abundant heterozygous calls in nominally haploid regions. Such signals can arise from contamination, mixed samples, duplicated regions, incorrect ploidy or genuine biology. Use independent evidence to distinguish explanations rather than naming a cause from one summary plot.

Review representative loci in an alignment viewer when local evidence is important. Examine placement ambiguity, read starts, strand and read-position patterns, nearby indels and the actual reference context. Visual review can uncover a mismatch between a reported record and its supporting reads; it is also

selective and observer-dependent. It cannot replace systematic benchmarking or establish the sensitivity of regions that were never called.

Practical note

Keep three outputs conceptually distinct: a candidate site, a sample genotype and a callable region. Later filtering may accept the site while masking one sample's genotype; a region may remain uncallable even when nearby variants pass. Carry these distinctions into annotation and downstream allele-frequency or association analyses.

The next stages normalize allele representation, apply justified site and genotype filters, assess performance and attach functional annotations. Each stage changes what can safely be concluded from the callset. Preserve the original candidate calls and a record of transformations so that a biologically interesting allele can be traced back to its evidence and model.

7 Filtering: decide which evidence can support a result

A caller assigns genotypes under a statistical model; filtering decides which of those assignments can support the intended analysis. These are different decisions. A site may have strong evidence that some alternate allele exists while one sample has too little evidence for a reliable genotype. Conversely, a convincing genotype in one sample does not rescue a site whose apparent variation is driven by misalignment across the cohort. Begin by separating site, genotype, sample, and region decisions rather than applying one quality cutoff to the whole VCF.

7.1 Three levels of filtering

A *site filter* evaluates a record or alternate allele using evidence such as mapping quality, strand imbalance, depth, and read-position bias. A *genotype filter* evaluates one sample's assignment using its genotype quality, informative depth, and allele support. A *sample filter* addresses failures such as contamination, unexpected identity, or inadequate genome-wide coverage. Record the level of every rule. A site-level `PASS` value is not a statement that every genotype in that record is usable.

The VCF header defines the meaning and scope of annotations. `INFO/DP` and `FORMAT/DP` need not describe the same quantity, and caller-specific depth may exclude reads rejected by internal filters. `QUAL` concerns a site-level hypothesis; `GQ` concerns a genotype conditional on a model. Neither should be treated as a universal calibrated probability across callers. Preserve filter labels and the original callset so that later decisions can be revisited. Missing filter status, represented by a dot, is distinct from an explicitly evaluated passing status [3].

An allele-balance plot is useful because it connects a genotype to its supporting reads. For an ordinary diploid heterozygote, comparable support for the two alleles is a useful expectation, not an exact requirement. Sampling variation, reference bias, local copy number, and mapping artifacts can change the ratio. The expectation changes with ploidy and biological context. Do not impose a diploid balance rule on haploid isolates, polyploids, pooled samples, or low-frequency somatic variants.

Consider an invented biallelic diploid site with three samples: `0/0`, `0/1` and `1/1`, with FORMAT/DP values 20, 5 and 30 respectively. Initially `AN=6`, `AC=3` and `AF=3/6=0.5`, where AN counts called alleles and AC counts copies of the alternate allele. If an illustrative `FORMAT/DP<10` rule masks the second sample, the genotypes become `0/0`, `./.` and `1/1`; recalculation gives `AN=4`, `AC=2` and `AF=2/4=0.5`. The unchanged frequency conceals a smaller denominator. The threshold is a teaching choice, not a recommended filter. Masking does not turn that genotype into `0/0`, and the companion masking phase does not automatically refresh INFO counts. Before frequency analysis, recompute the counts from the retained genotypes, for example with `bcftools +fill-tags` and `-t AN,AC,AF` [17].

7.2 Choose rules from the experiment

Inspect annotation distributions separately for SNPs and indels, relevant coverage strata, and sequencing batches. Compare plausible true variants with independently supported errors where such evidence exists. A very high depth can indicate a collapsed repeat, whereas low depth can make the genotype uncertain;

the same numerical cutoff cannot represent both situations across capture experiments, genomes, and organisms. Review variants near proposed boundaries, and record how many sites and genotypes each rule changes.

GATK distinguishes hard filtering from variant quality score recalibration (VQSR). Hard filtering applies explicit annotation thresholds. VQSR learns an annotation model using suitable resources and enough informative variants; it is not an automatic improvement for every dataset. Published human examples explain annotation behavior but do not establish thresholds for a different species, small target panel, or altered laboratory protocol. A known-site resource used during processing is also not automatically independent truth [11].

Population checks need their own assumptions. Unexpected heterozygosity, missingness, relatedness, or allele-frequency patterns may reveal a problem, but can also reflect real biology. A Hardy–Weinberg test assumes an appropriate population model; selection, structure, inbreeding, clonality, sex linkage, and ascertainment can make a simple exclusion rule misleading. Examine the experimental design before using a population statistic as an error detector. In a case-control analysis, different missingness between groups deserves particular attention because technical absence can imitate a biological difference [26].

7.3 Callable regions define the denominator

An absent variant record does not prove a reference genotype. A position may lack coverage, fail mapping criteria, lie outside capture targets, or be excluded by the caller. Define a callable region using the evidence required by the particular question, retain a region mask, and report its relationship to the intended analysis territory. A sample-specific mask answers where that sample was evaluable; a cohort mask answers a different question and must specify how samples were combined.

For example, comparing variant burden between two samples is difficult to interpret if one has much less callable sequence. Counting retained variants alone hides that difference. Likewise, a negative finding in a gene should distinguish confidently assessed coding bases from uncovered exons and difficult paralogous regions. A BED file usually uses zero-based, half-open intervals, whereas VCF positions are one-based; verify the conventions before intersecting them. Boundaries around indels require particular care because an anchored record can overlap a region differently from its biological event.

Watch out for

Filtering changes the set of claims the data can support. Report separately the number of records retained, genotypes set to missing, samples excluded, and bases or intervals outside the callable territory. A clean-looking VCF is not evidence that excluded sequence was successfully examined.

8 Normalization: make variant representations comparable

A VCF record is a representation of a sequence difference, not the difference itself. Within a repeated sequence, deleting one repeated base can be written at several positions while producing the same alternate haplotype. Adjacent substitutions can appear as one multinucleotide variant or as several SNPs. Before matching databases, merging records, or comparing callers, decide what equivalence means for the task.

8.1 Reference-aware normalization

For small variants, normalization commonly removes redundant shared sequence while retaining valid VCF alleles and left-aligns indels against the chosen reference. The reference FASTA is therefore part of the operation. Its contig names and sequences must agree with the VCF; changing an assembly label without transforming coordinates does not repair a mismatch. A normalization tool should reveal inconsistent reference alleles rather than silently make a different biological claim [16, 3].

Consider a synthetic reference segment `CAAAAG`. If the C is position 1, the VCF representation `POS=4, REF=AA, ALT=A` becomes `POS=1, REF=CA, ALT=C` after reference-aware left alignment. Both produce `CAAAG`; the retained anchor makes the deletion a valid VCF record. Left alignment chooses a consistent representation relative to that exact sequence. It does not establish which physical A was deleted, improve read evidence, or prove that two nearby calls arose on the same chromosome. A variant written on the reverse-strand transcript still needs to be reconciled with the genomic reference convention.

The following command illustrates a documented `bcftools norm` operation. It assumes a matching indexed reference and an input VCF whose records have been reviewed. The filenames are placeholders; the command is not evidence that any dataset in this note was processed.

Shell

```
bcftools norm -f reference.fa -m -any \
  -Oz -o normalized.vcf.gz calls.vcf.gz
bcftools index normalized.vcf.gz
```

Here `-f` supplies the sequence used for reference checks and normalization, `-m -any` splits multiallelic records, and `-Oz` writes compressed VCF. Splitting is a deliberate downstream choice, not a necessary definition of normalization. Inspect the output header and representative genotypes after transformation; an apparently tidy file can still have unsuitable semantics for the next analysis [16, 17].

8.2 Multiallelic sites carry relationships

At a diploid site with two alternate alleles, `GT=1/2` means that the sample carries the first and second alternate alleles. It does not mean two independent heterozygotes relative to reference. When a tool splits this record, allele-indexed annotations and genotype likelihoods require appropriate remapping. Header declarations such as `Number=A`, `Number=R`, and `Number=G` describe values per alternate allele, per allele including reference, and per genotype. Check that the chosen transformation handles the fields actually present [3].

Retain the original record or a traceable correspondence when allele relationships matter. Allele counts and frequencies may need recalculation after genotype masking or sample subsetting. A site-level annotation copied into each split record is not necessarily allele-specific evidence. A downstream program that accepts only biallelic records imposes an input restriction; that restriction should not be mistaken for a biological model of the locus.

8.3 Haplotypes and consequence interpretation

Two substitutions in the same codon can produce a different amino-acid consequence when interpreted together than when annotated separately. Whether the substitutions are in *cis* or *trans* therefore matters. An unphased genotype does not answer that question. Similarly, a complex replacement represented by several indels may match a truth haplotype without matching its individual VCF lines. Record-level equality is too strict for some comparisons and too permissive for others if genotype agreement is ignored.

Use a haplotype-aware comparison method for benchmarking, and state whether the comparison evaluates alleles, genotypes, or phase. Preserve a representation appropriate to annotation rather than decomposing every event solely to make rows look uniform. Normalization, decomposition, phasing, and assembly conversion solve different problems; performing one does not certify the others. After any representation change, compare counts by variant class and inspect changed complex loci before continuing [23].

9 Annotation: connect a sequence difference to a biological model

Variant annotation adds statements about a variant using external models and evidence. A consequence predictor can ask whether an allele changes a coding sequence, overlaps a splice region, or lies within an annotated regulatory interval. A population database can supply an observed frequency. An experimental study may support a functional effect. These statements answer different questions and should remain distinguishable in the annotated output.

9.1 Choose a compatible annotation bundle

A valid annotation bundle combines the correct assembly sequence, gene models, transcript identifiers, and resource versions. Confirm contig naming, coordinate conventions, genetic code where relevant, and the provenance of every database. A species name is insufficient: two assemblies of the same species can have different coordinates and gene models. A transcript accession without a version may also conceal

a changed coding sequence. Record both the genomic allele and the transcript used to describe its consequence.

Ensembl VEP can use a species cache, supported GFF/GTF transcript models, and custom annotation sources. Its cache release should match the software's documented compatibility requirements, and its assembly must match the input. Offline operation needs the local resources required by the selected mode; a cache directory is not a substitute for checking the reference and configuration. SnpEff uses its own genome databases and can build a database from compatible reference sequences and gene annotations. These are alternative annotation routes, not independent experimental confirmations of a predicted effect [18, 19, 21, 22].

For a non-model organism, first examine the gene annotation itself. Missing exons, incomplete coding sequences, or an incorrect translation table can produce misleading consequences even when the variant is called correctly. A custom SnpEff database should be checked against available coding and protein sequences, with discrepancies investigated. For VEP with custom gene models, verify the supported format and required indexing from the versioned manual. If no suitable gene model exists, report overlap and sequence evidence at the level actually supported; do not assign a confident protein consequence from an unrelated species' coordinates.

9.2 Illustrative annotation commands

The following VEP example assumes the software, a compatible local cache, and a matching indexed FASTA are already available. Set the shell variables to the verified species name, assembly label, and cache location for the experiment; these placeholders deliberately do not impose a human reference. The local cache version must satisfy the installed VEP version's compatibility rules.

Shell

```
vep --input_file normalized.vcf.gz \
  --output_file annotated.vep.vcf --vcf \
  --offline --cache --dir_cache "$cache_dir" \
  --species "$species" --assembly "$assembly" \
  --fasta reference.fa
```

The command requests VCF output and local annotation resources. Before parsing VEP fields, inspect the VCF header declaration for `INFO/CSQ`: its `Format:` description specifies the field order, which should not be guessed from an old example. Review warnings and representative annotated alleles. Successful completion does not prove that every transcript model agrees with the reference [19, 27].

An alternative SnpEff example assumes a previously checked, locally available database with an assembly and annotation release matching the input. Set `database_id` to that exact database identifier; a species name alone is insufficient.

Shell

```
java -jar snpEff.jar "$database_id" \
  normalized.vcf.gz > annotated.snpeff.vcf
```

Inspect the `ANN` header and any allele-level warnings rather than treating every emitted consequence as equally reliable. If a suitable prebuilt database is unavailable, construct and check a custom database from the matching FASTA and gene models before this annotation step. The database-build route requires its own sequence and translation checks. These commands are alternatives, not an instruction to concatenate two annotation outputs without reconciling their schemas [28, 22].

9.3 One allele can have several consequences

A variant can be coding in one transcript, intronic in another, and outside a third. Preserve transcript-level results before selecting a summary. The most severe predicted consequence is useful for triage but can privilege a poorly supported transcript. A principal or canonical transcript provides a reproducible selection rule, not a guarantee that it is the biologically relevant transcript for the tissue or condition. State the rule and retain the alternatives needed to review important candidates [18].

For an invented plus-strand example under the standard nuclear genetic code, a genomic `G>A` substitution changes transcript T1's codon from `GAA` to `AAA`, predicting glutamate to lysine, a missense consequence. Suppose the same genomic base lies in an intron of transcript T2: its annotation is intronic rather than the T1 coding substitution. These are two transcript-dependent descriptions of one allele, not two variants or measurements of transcript expression. Retaining transcript IDs and versions makes the difference reviewable.

A synonymous variant preserves the encoded amino acid in a specified transcript; the term does not establish that every molecular process is unaffected. A missense variant changes an amino acid but does not itself demonstrate a functional defect. A predicted frameshift or stop-gained consequence depends on transcript structure, reading frame, and sequence context. The consequence is a model-based description of the allele. Its biological importance requires evidence appropriate to the question.

Watch out for

Labels such as `HIGH` or `MODERATE` summarize predicted consequence categories. They are not probabilities of pathogenicity, measurements of fitness, or proof that a phenotype is caused by the variant. Keep sequence consequence, functional evidence, population evidence, and the final interpretation in separate fields [20].

9.4 Add evidence without losing its context

Custom databases can add population frequencies, conservation scores, curated phenotypes, or project-specific intervals. Before matching, reconcile assembly and allele representation. An interval overlap supports a different statement from an exact normalized allele match. A missing database hit may mean absence from the database, unavailable coverage, an unmatched representation, or a resource outside the organism's scope; it does not automatically mean that an allele is rare.

For a frequency, retain the population definition and denominator when available. A frequency estimate from one population cannot be assigned unchanged to another, and an allele observed in a database is not thereby harmless. For predicted functional scores, retain the method and version, and recognize shared training data or features when combining predictors. Counting correlated predictions as separate lines of evidence can exaggerate support.

Finish annotation with a small audit of known synthetic or well-characterized examples: one coding substitution, one noncoding allele, one indel, and a transcript with a relevant boundary. Check allele orientation, transcript identity, and the expected category. This audit tests configuration and interpretation; it does not measure sensitivity or specificity on the full experiment.

10 Validation: measure the claim within an explicit territory

A plausible workflow needs evidence that it performs the intended task. Benchmarking against a characterized sample measures agreement with a reference callset under defined comparison rules. Reviewing read alignments investigates individual candidates. An independent laboratory assay can test a selected allele. None of these alone answers every validation question. Choose the evidence to match the claim, and distinguish workflow evaluation from confirmation of a particular result.

10.1 Truth includes regions as well as variants

A small-variant benchmark normally includes a truth VCF and a confident-region BED file. The BED identifies territory in which the reference callset supports a comparison; sequence outside it should not be treated as confidently reference. Use the exact corresponding assembly and benchmark release. Human Genome in a Bottle (GIAB) resources provide characterized human samples and benchmark regions. Their usefulness for human workflow development does not establish performance in a fungus, plant, pooled population, or tumor [24, 23].

For a capture experiment, comparison may additionally be restricted to the intended target territory. State the resulting intersection and how boundary-spanning events are handled. A truth variant outside the evaluated territory is not an ordinary false negative. Likewise, an apparent extra call outside confident

truth territory cannot be classified as a false positive merely because it is absent from the truth VCF. Keep excluded and unassessed results visible.

The truth evidence should be as independent as the claim requires. Tuning filters until they match a single benchmark and then reporting the same sample as an unbiased performance estimate introduces evaluation leakage. Separate development decisions from final evaluation where resources permit. Agreement between two callers on the same reads is useful diagnostic evidence, but shared mapping errors and correlated models prevent it from serving as complete independent truth.

10.2 Compare haplotypes and report the denominator

Haplotype-aware tools such as `hap.py` address representation differences that make direct line matching misleading. The comparison must also define genotype matching, filtering, reference sequence, regions, and variant categories. Inspect the tool's output conventions before interpreting its counts, because truth-side and query-side event accounting can differ after representation-aware matching [23].

For a simple event-counting example with an agreed matching definition, precision and recall are

$$\text{precision} = \frac{TP}{TP + FP}, \qquad \text{recall} = \frac{TP}{TP + FN}.$$

Precision asks what fraction of assessed calls are supported by truth; recall asks what fraction of assessed truth variants were recovered. If a denominator is zero, the corresponding metric is undefined rather than automatically perfect. Report the underlying counts and evaluated territory with the metrics. Quantify uncertainty with an interval appropriate to the sampling and dependence structure when estimating performance; a small number of evaluated variants cannot establish precise performance across all contexts [23]. In a synthetic example with 90 matched calls, 10 unsupported calls, and 20 missed truth variants, precision is $90/100 = 0.90$ and recall is $90/110 \approx 0.818$. These are arithmetic illustrations, not observed workflow results.

A genome-wide average can hide errors relevant to the experiment. Stratify SNPs and indels, indel lengths, coverage, and difficult sequence contexts when the benchmark supports those categories. Repeats and low-complexity regions may behave differently from easier sequence. Report the benchmark's coverage limits, including variant classes it does not assess. A small-variant benchmark does not certify structural variants, copy-number changes, or every medically relevant locus [23].

10.3 Investigate discordances and selected candidates

Read visualization can reveal strand restriction, soft clipping, nearby indels, low mapping quality, or concentration of support at read ends. Compare alternate and reference reads and inspect surrounding sequence rather than only the marked base. These observations explain why a call deserves investigation; visual confidence alone is not a calibrated accuracy metric. A reviewer should be able to distinguish an observed read pattern from the proposed explanation for it.

Independent assays require their own design and detection limits. An assay chosen for a high-frequency germline allele may be unsuitable for a low-frequency mosaic allele or a repetitive locus. Report which candidates were tested, why they were selected, and what the assay could detect. Confirming a few strong candidates does not estimate whole-callset precision if selection excluded uncertain calls. When no suitable benchmark or orthogonal evidence exists, describe the checks completed and the remaining uncertainty instead of inventing a universal performance figure.

11 Special routes: change the model when the biology changes

The common workflow establishes reference compatibility, read quality, variant evidence, representation, and annotation. The model used within that workflow must still fit the sample. Before choosing a caller, write down the organism, expected ploidy by region, whether the material is a single individual or a mixture, sequencing technology, and the intended allele-frequency range. Tool support for an input type is only the beginning; performance needs evaluation in the relevant setting.

11.1 Human germline samples

Human germline analysis benefits from well-developed reference assemblies, population resources, and benchmark samples. These resources are assembly-specific and have different coverage across populations and genomic regions. Choose sex-chromosome and other region-specific ploidy deliberately; one sample-wide diploid assumption is not correct for every genomic compartment. The mitochondrial genome is also a distinct analysis problem when heteroplasmy is of interest.

11.2 Other diploid organisms: rebuild the resource contract

For a diploid model organism, the overall sequence of operations may remain useful, but human known sites, frequency cutoffs, transcript priorities, and truth resources do not transfer automatically. Check assembly quality and strain divergence. An allele called "alternate" differs from the selected reference; it is not necessarily newly arisen, rare, deleterious, or absent from the species. Reference choice affects both mapping and interpretation.

DeepVariant documentation distinguishes supported models and use cases, including restrictions relevant to ploidy and non-human applications. A model trained in one setting should not be described as validated for another merely because the software runs. Select a model appropriate to technology and assay, and evaluate the result using relevant evidence rather than transferring a published accuracy value [12, 13].

For this route, prepare the organism's reference and gene models first, confirm diploidy for the material and analysis intervals, then perform read QC, alignment and library-aware duplicate preparation. Use the supplied diploid calling route only after reviewing caller priors and assay compatibility. Omit optional BQSR if compatible known sites and a defensible protocol are unavailable; do not substitute a human resource. Choose and evaluate filters on the organism's evidence. Normalize against the same FASTA, annotate with matching models and review coverage, genotype and locus displays. Without an appropriate benchmark, report diagnostic checks and selected independent confirmations; do not label agreement between callers as measured accuracy.

11.3 Haploid isolates, microbes, and fungi

For a clonal haploid isolate, a haploid genotype model can be appropriate. Apparent mixed alleles require investigation: possible explanations include contamination, mixed infection or culture, within-sample diversity, duplicated sequence, and mapping artifacts. Diploid-style heterozygosity is not a neutral default explanation. Bacterial genomes may include plasmids with different copy numbers; coverage and allele fractions therefore need interpretation at the correct genomic compartment.

Fungi cannot be assigned one universal ploidy. Haploid, diploid, polyploid, and aneuploid states can occur depending on species and material. A survey of *Saccharomyces cerevisiae* isolates illustrates ploidy and chromosome-copy variation within one yeast species [29]. Establish the relevant biological state before choosing genotype and allele-balance expectations. When a close reference is unavailable, divergent or absent sequence may fail to map; a reference-based small-variant callset then describes the alignable portion of the sample rather than all genome differences. Gene presence/absence, rearrangements, and accessory sequence require additional analysis routes.

Annotation also needs organism-specific choices. Confirm gene models and genetic code, including organellar or lineage-specific exceptions where applicable. Do not use a human consequence database as a substitute for a compatible microbial or fungal annotation. A catalog of resistance-associated variants, when relevant to a research question, needs its own organism, allele, and evidence scope; a generic protein-impact label does not establish a resistance phenotype.

The supplied haploid route starts after shared preparation and uses BCFtools with explicit ploidy one. Inspect contamination and unexpectedly mixed allele evidence before interpretation. Normalize calls, choose organism-specific filters, then annotate against matching genes and inspect both depth and candidate loci. This is a single-isolate workflow; a metagenomic community, pooled culture or mixed infection needs a mixture model and cannot inherit its genotype interpretation.

In BCFtools 1.25, `mpileup` defaults include a raw depth cap of 250 per input file (`--max-depth`), minimum mapping quality 0 (`--min-MQ`) and minimum base/BAQ quality 1 (`--min-BQ`). Raising `--max-depth` can retain more read evidence and increase memory or work; lowering it can discard evidence before genotyping. For this single-sample route, `--max-idepth 250` instead excludes indel calling at pileup depths of 250

or more. It does not subsample those loci to 250 reads. Raising this threshold permits indel evaluation at deeper positions; lowering it excludes more high-depth positions. Raising quality minima excludes more alignments or bases. None guarantees better inference, and BAM depth can exceed the evidence actually used. The companion preserves these defaults and explicitly requests `FORMAT/AD,FORMAT/DP` with `-a`. AD is already a default annotation in this version; DP is optional. AD lists reference then alternate allele depths, while FORMAT/DP counts high-quality bases under the caller's rules. Inspect header definitions and BAM evidence before choosing a denominator; missing fields in other callsets remain missing rather than zero [17].

11.4 Polyploids and pooled samples

In a polyploid individual, genotype inference must account for allele dosage and the number of genomic copies. Coverage adequate to distinguish two diploid genotypes may not separate neighboring dosage states. In allopolyploids, related homeologous sequences can also create mapping ambiguity. Increasing a ploidy parameter does not repair a reference that collapses distinct copies. Interpret dosage uncertainty and reference structure together.

FreeBayes documents explicit ploidy and pooling modes, which make it an example of a tool whose model can be configured for these settings. Choosing such a mode still requires suitable coverage, parameter review, and validation. A pooled experiment measures evidence contributed by multiple genomes; read depth is not the number of independently sampled organisms. Unequal DNA contribution and sampling at both the organism and read levels affect allele-frequency estimates. Do not report pooled frequencies as individually resolved genotypes [14, 15].

The supplied polyploid route uses FreeBayes for a declared fixed integer ploidy above two. Establish whether a suitable collapsed, haplotype-resolved or subgenome-resolved reference supports the intended copy model before aligning. Inspect depth, allele dosage uncertainty and homeologous mapping; use compatible custom annotations and dosage-aware summaries after normalization and filtering. An aneuploid sample or changing interval copy number falls outside this fixed-ploidy script. The package does not implement pooled-frequency inference.

11.5 Long reads and somatic variation

Long reads can change alignment, phasing, and difficult-region evidence, while their error profiles depend on platform, chemistry, and processing. Use an aligner and small-variant model intended for the actual read type, and retain the relevant model version. A short-read configuration does not become appropriate merely by accepting a long-read BAM. Small variants and structural variants remain different output classes with different evaluation requirements [13].

Somatic analysis seeks acquired non-germline changes, which can be clonal or subclonal in the sampled tissue. Their observed allele fractions also depend on germline variation, purity, copy number and contamination. A diploid germline genotype model does not answer this question. A matched normal can help distinguish germline alleles and some artifacts, but normal contamination, tumor contamination of the normal, and detection limits still matter. Tumor-only analysis has different uncertainty and resource requirements. Mutect2 is an example of a dedicated somatic route; its calls require the corresponding filtering workflow rather than a germline `PASS` rule [25].

For the human matched somatic route, prepare tumor and normal consistently, verify identity and intended pairing, then select a dedicated caller. Mutect2 outputs feed its documented filtering for contamination and orientation bias as appropriate; the companion README lists the required resources and filtering phases. Retain filtered calls and allele-level evidence, normalize the selected output, then annotate with human assembly-matched resources and review tumor/normal displays. Germline population frequency is supporting context, not a direct substitute for a matched normal or proof of somatic origin. Tumor-only inference is not implemented by the supplied matched-pair wrappers.

These routes define where the teaching workflow must branch. None supplies a universal list of thresholds. When the biological state or technology falls outside the validated model, document the mismatch and obtain appropriate evidence before interpreting apparent absence or low-frequency variation.

The 2026 review by Zheng and colleagues surveys long-read small-variant methods, including deep-learning, assembly and pangenome approaches. Its abstract emphasizes distinct error profiles and benchmarking newly accessible regions. This provides a current reading direction, not an executed long-read workflow or an accuracy guarantee for these short-read scripts [41].

12 Visualize evidence at callset and locus scales

Visualization answers two complementary questions: whether the callset behaves as expected across samples and whether the reads support a particular allele. Begin with a declared question and retain the plotted data, selection rule and denominator. An attractive genome-wide track is not an accuracy measurement; a convincing alignment screenshot is not independent confirmation.

12.1 Choose a diagnostic display

Compare depth, missingness and genotype quality by sample and batch before summarizing biological groups. Use distributions rather than only means; show missing values separately from zero. For SNP/indel counts or consequence categories, specify whether the unit is a VCF record, normalized allele, sample genotype or transcript consequence. Splitting a multiallelic record or annotating several transcripts changes these counts without creating new biological observations. Compare variant density only within compatible callable territories; dividing by a whole reference when coverage differs can manufacture a difference.

Plot allele depth against total allele depth, or allele fraction against depth, when the caller supplies interpretable allele counts. Define the denominator and preserve multiallelic ordering from the header. For an example diploid heterozygote, a fraction near one half may be compatible with the model, but sampling, mapping bias and copy number affect its distribution. In fixed tetraploid regions, dosage states suggest different expectations; in somatic samples, purity and copy number prevent a simple conversion of fraction into cell prevalence. Do not impose a diploid balance window on all organisms.

For a benchmark comparison, show precision and recall separately for SNPs and indels, with the same truth set, region mask and matching policy. Stratify difficult regions where the benchmark supports it. The assembly-based X/Y benchmark illustrates why apparently strong aggregate performance does not establish performance in every genomic compartment [38]. A consequence bar chart describes annotation composition; it does not establish functional enrichment without a justified background and statistical model.

12.2 Review an allele in local IGV

Load the exact reference FASTA, indexed BAM/CRAM, indexed VCF and compatible gene models in IGV. Zoom to the locus; an empty track at a broad viewing scale may reflect the visibility window. Inspect indels and read details, group reads by strand when useful, and review mapping quality. Preserve the viewer version and display settings; coverage coloring and base-quality shading affect what is visible [39, 40].

Review a prespecified set of passing, filtered and discordant loci. For somatic pairs, display tumor and normal together on comparable coverage scales. Record the genomic allele, selection reason, observed evidence, reviewer and unresolved alternatives. Inspect nearby repeats and clipping rather than cropping away contradictory reads. Keep screenshots and sessions local when they contain protected information. Viewer sessions may retain file locations. Visual review supports investigation, not a sensitivity estimate for unobserved alleles.

12.3 Reproduce the displayed summary

The companion phase `09_prepare_review.sh` saves BCFtools statistics and a long allele/sample table from the explicitly selected VCF. It does not silently select only `PASS`, mask genotypes, infer missing tags or manufacture plots. The next phase, `10_plot_record_counts.py`, creates a stacked class/FILTER count display as PNG and PDF, with its source TSV. It counts each VCF record once, keeping multiallelic and symbolic records separate; a record count is not an allele count or accuracy estimate. Declare

the comparison and preserve the script. For local review, load the matching reference and alignments separately; the summary VCF alone cannot reconstruct read evidence.

Before sharing a display, reconcile plotted totals with the source table and independently count a small known example. Check axes, units, labels, missing values and the exact sample/region selection. Deliberately changing an allele or filter state should change the relevant summary. Exported scientific figures need both raster and vector inspection; the teaching package supplies executable plotting commands without claiming a real-data result.

13 Interpretation and reporting: preserve the chain of evidence

A useful report lets another reader trace a conclusion from the experimental question through the reads, genotype, annotation, and external evidence. The final candidate table should therefore contain more than a gene symbol and an impact label. Retain the assembly, normalized genomic allele, sample genotype or allele-frequency estimate, filter status, relevant depth and support, transcript identifier, consequence, and the evidence that motivated follow-up. Distinguish a candidate selected for investigation from a demonstrated biological explanation.

13.1 A worked synthetic candidate

Assume an educational experiment on a diploid organism with a compatible reference and gene annotation. The synthetic candidate is a biallelic coding substitution, and the caller assigns an unphased heterozygous genotype. Suppose the retained observations are 24 reads counted as reference support and 18 as alternate support. The observed alternate fraction among these counted reads is

$$\frac{18}{24+18} = \frac{18}{42} \approx 0.429.$$

The numbers are invented to illustrate interpretation. They are not a recommended depth requirement, an error model, or results from a real sample. Other reads may have been excluded by caller-specific rules, so this fraction should not silently replace the caller's own annotation.

The first question is whether the genotype is technically supported. Inspect mapping quality, read placement, strand distribution, nearby variation, and the region's callability. A moderate alternate fraction alone neither confirms nor rejects a heterozygote. If the locus overlaps a duplicated region, the same counts could arise from reads assigned to the wrong copy. If its ploidy is uncertain, the assumed genotype model needs revision before biological interpretation.

The second question is what the allele changes under the selected gene model. Suppose annotation reports a missense consequence in one transcript and an intronic consequence in another. Preserve both, and state why the coding transcript is relevant to the question. An amino-acid change is a predicted sequence consequence; it does not measure protein activity. If the variant has no hit in a compatible population resource, report that absence with the resource's scope rather than calling the allele globally rare.

The third question is what would justify a biological conclusion. Co-segregation, association, an independent functional assay, or other evidence may be relevant depending on the experiment. Each requires an appropriate design and controls. Agreement between VEP and SnpEff about a codon change mainly supports consistency of annotation under their inputs; it is not two independent demonstrations of phenotype causation. A sensible result at this stage can be "technically plausible candidate requiring targeted follow-up."

Separate observations from interpretations

Layer	Supported statement in the synthetic example	Statement not established
Read evidence	Counted reads support both alleles.	Every supporting read originates from the intended locus.
Genotype model	The caller assigns a diploid heterozygote.	Ploidy and copy number are known from the genotype label alone.
Annotation	One transcript predicts an amino-acid substitution.	Protein function or organismal phenotype is altered.
Database lookup	No matching allele was found in a specified resource.	The allele is absent from the species or population.
Follow-up	The candidate meets a declared investigation rule.	The candidate is the causal explanation.

13.2 Reproducibility

Preserve the exact reference and annotation releases, software versions, model identifiers, command arguments, interval definitions, and filtering rules. Retain logs and machine-readable intermediate summaries locally under the project's data-governance rules. A workflow name alone is insufficient because its versions, defaults, and resource bundles can change. For stochastic methods, retain the relevant seed and runtime assumptions where the implementation exposes them. Record deviations from the planned analysis and their reasons.

A practical report separates attempted analysis territory, callable territory, and benchmark territory. Include sample exclusions, genotype masking, missingness, and variant counts by class; explain which counts refer to records and which to alleles. If a database changed after annotation, preserve the original result alongside any explicitly versioned reannotation. A negative report should state the regions and variant classes assessed and the important unresolved gaps.

13.3 State what was actually verified

For this note, embedded commands and companion scripts illustrate documented interfaces and the surrounding reasoning. They do not constitute a completed real-data workflow, and the synthetic examples do not estimate performance. In an executed project, report separately which steps ran, which outputs were inspected, what benchmark or independent validation was used, and what remains untested. Successful file creation, indexing, and compilation demonstrate technical operations; they do not establish biological correctness.

The deliverable from a research analysis is an auditable statement with appropriate limits. A small-variant callset can support selected questions about the evaluated reference-aligned sequence. It cannot by itself exclude every genomic cause of a phenotype, certify unassessed structural variation, or establish clinical readiness. Match the final wording to the experiment, the validated territory, and the evidence available for each candidate.

14 A staged workflow from reads to an evaluated callset

The companion directory `main_scripts/genomic_variants` supplies phased Bash scripts, a sample-sheet example and separate software environments. It connects the concepts in this note to inspectable commands. Run one phase at a time and inspect its outputs before continuing; the package does not claim a completed real-data reproduction or a validated clinical pipeline. Its bilingual README gives the current commands, arguments and execution limits.

The chapter order separates filtering decisions from representation; the companion execution order normalizes first and then applies the reviewed filtering policy (Sections 7 and 8).

14.1 Shared preparation and explicit biological branches

Keep reads and derived data in an external analysis directory selected by `GENOMICS_PROJECT`. The sample sheet describes one paired sequencing unit per row and distinguishes sample, physical library and platform unit. Repeated lanes of one library share its library identifier; each unit has a unique read-group identifier. For Illumina reads, `PL:ILLUMINA` identifies the platform, `SM` the sample, `LB` the library and `PU` the platform unit. Assigning every read group the same identifier loses these distinctions [4].

Example records

```
main_scripts/genomic_variants/
  scripts/00_setup/       # validate manifest and inputs
  scripts/01_reference/   # index the selected reference
  scripts/02_reads/       # optional fastp
  scripts/03_alignment/   # BWA or Bowtie2, per sequencing unit
  scripts/04_preparation/ # library duplicates; optional BQSR
  routes/                # biological calling models
```

Build mapper indexes from the actual selected FASTA. Choose BWA-MEM or Bowtie2 explicitly, preserving their arguments and mapping diagnostics; their placement and quality models need not yield identical evidence [30, 31]. Optional trimming needs a stated reason and review of retained reads. Merge units from the same library for duplicate marking, then combine libraries for the sample. The supplied SAMtools route uses name sorting for `fixmate -m`, followed by coordinate sorting and duplicate marking. Its final calling input is coordinate-sorted and indexed; the temporary name-sorted BAM is not the final indexed input [5]. Optional BQSR requires compatible known sites and a justified caller protocol. Preserve the choice to use marked or recalibrated BAMs across comparisons.

Before running a fixed-ploidy wrapper, review a BED file containing only territory compatible with that model. The full FASTA and mapping reference remain unchanged. Each of the diploid GATK, haploid BCFtools and polyploid FreeBayes wrappers accepts this reviewed file as its optional second argument; it validates coordinates against the prepared reference and saves a copy with the calls. For example:

Shell

```
bash routes/diploid_germline/05_haplotypecaller.sh \
  sample1 reviewed-diploid.bed
```

The BED uses zero-based, half-open coordinates: `toy1 100 200` denotes one-based bases 101–200, with tab separators in the actual file. The wrappers pass GATK `-L`, BCFtools mpileup `--regions-file` or FreeBayes `--targets`. Without BED, the model must be justified throughout the whole reference; a manifest ploidy value does not check every contig. Removing reference contigs can alter mapping competition and is not an equivalent restriction. Calling territory and later evaluation masks answer different questions; an evaluation mask cannot repair an inappropriate calling model. Assembly context, anchored indels and reference-confidence blocks require boundary review. No extra interval padding is requested; if padding is introduced, its territory must also fit the model.

The calling branch follows the biological question. The diploid germline route produces per-sample gVCFs and then joint genotypes from compatible gVCFs. The haploid route uses an explicit haploid model; the fixed-ploidy polyploid route requires the declared copy number. These examples do not infer ploidy or repair contradictory input metadata. Mixed ploidy, aneuploidy, organelles, pools and long reads require additional decisions described in the route boundaries. Human paired somatic analysis instead uses matched normal and tumor samples; its mutation model must address the uncertainty discussed in Section 11.

14.2 Compare aligners and somatic callers fairly

A bounded human somatic comparison can evaluate BWA versus Bowtie2 and Mutect2, Strelka2 or VarScan2, but a matrix of filenames is not an experimental design. Hold the input read pairs, sample pairing, reference, trimming decision, duplicate policy, BQSR condition and evaluated territory constant where the comparison requires them. Preserve each caller's documented input and filtering contract.

Mutect2 needs its somatic filtering stages; VarScan's status and high-confidence labels are caller classifications rather than independent truth. Strelka2 is retained as a legacy comparison: its repository is archived, and its Python 2 runtime belongs in an isolated compatible environment, separate from modern tools [25, 32, 33].

If a preprocessing choice is itself under study, declare it as a factor and interpret the resulting comparison accordingly. Capture samples require the actual intersection of the benchmark's high-confidence regions and the exome target mask, with a stated boundary policy. Applying two masks in prose without constructing and using their intersection does not restrict the evaluation. A germline truth set characterizes germline alleles; it does not automatically label low-frequency somatic events. Use evidence suited to the event class and allele-frequency range, and keep unassessed calls separate [23, 24].

Report TP, FP and FN under explicit matching rules, with precision and recall as defined in Section 10. Accuracy also requires a defined true-negative universe and denominator; ordinary variant-only VCF counts do not supply them. Do not rank callers by raw call count or an unsupported accuracy percentage. Select summaries that answer the question, such as SNP/indel precision–recall, coverage or discordance categories; neither PCA nor a fixed number of VCF files is a requirement for a meaningful comparison.

14.3 Reproducibility as an execution record

For each phase, retain the input source and release, exact command, effective arguments, software and environment versions, reference and resource identity, output locations and exit status. Record wall-clock time, allocated threads and memory with the hardware context if reporting runtime. Preserve random seeds where an operation exposes stochastic behavior; record non-applicability rather than inventing a seed for every command. A version-pinned environment file helps restore tools, while a solved environment export records the packages actually used on one platform.

Define output gates before running. Examples include valid paired reads, compatible reference contigs, expected BAM sample/read groups, coordinate order and index, expected VCF samples and ploidy, reference-consistent alleles, and a nonempty evaluated region. Test gates with deliberately malformed synthetic inputs and require a failure. File integrity checks establish consistency; independent benchmarks or assays address biological validity. Report syntax checks, synthetic smoke execution and real-data reproduction separately, including stages that remain unexecuted. Keep raw calls and a record of normalization, filtering, annotation and comparison so that a changed result can be traced to its cause.

14.4 Map the same protocol to a local Galaxy installation

In an authorized local Galaxy installation, map the same logical stages to available tools and retain the dataset relationships in the history. Capture the installed tool identifier and version, parameters, reference selection and job outcome; export the history and reusable workflow locally as complementary records. The history describes what ran, while the workflow describes reusable connections. Check tool availability and versions on the actual installation, including the required recalibration tools, rather than assuming that every Galaxy has the same capabilities [34, 35]. This conceptual mapping does not establish equivalence with command-line results; compare effective parameters and outputs under the same biological contract.

References

[1] Brown TA. *Genomes 5*. 5th ed. CRC Press; 2023. ISBN: 9780367674076.

[2] Van der Auwera GA, O'Connor BD. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. O'Reilly Media; 2020. ISBN: 9781491975190.

[3] Global Alliance for Genomics and Health File Formats Team. *The Variant Call Format Specification: VCFv4.5 and BCFv2.2*. 25 February 2026. `https://samtools.github.io/hts-specs/VCFv4.5.pdf`

[4] SAM/BAM Format Specification Working Group. *Sequence Alignment/Map Format Specification*. Version 1.6; 12 August 2025. `https://samtools.github.io/hts-specs/SAMv1.pdf`

[5] SAMtools developers. *samtools(1) manual page*. Version 1.25; 27 October 2026. `https://www.htslib.org/doc/samtools.html`

[6] GATK Team. *Germline short variant discovery (SNPs + Indels)*. Updated 24 July 2024. `https://gatk.broadinstitute.org/hc/en-us/articles/360035535932-Germline-short-variant-discovery-SNPs-Indels`

[7] GATK Team. *HaplotypeCaller*. Updated 20 August 2026. `https://gatk.broadinstitute.org/hc/en-us/articles/54698694431131-HaplotypeCaller`

[8] GATK Team. *Known variants—Training resources—Truth sets*. Updated 25 June 2024. `https://gatk.broadinstitute.org/hc/en-us/articles/360035890831-Known-variants-Training-resources-Truth-sets`

[9] GATK Team. *GenomicsDBImport*. Updated 6 October 2022. `https://gatk.broadinstitute.org/hc/en-us/articles/360057439331-GenomicsDBImport`

[10] GATK Team. *Base Quality Score Recalibration (BQSR)*. Updated 29 August 2024. `https://gatk.broadinstitute.org/hc/en-us/articles/360035890531-Base-Quality-Score-Recalibration-BQSR`

[11] GATK Team. *Hard-filtering germline short variants*. Updated 6 November 2024. `https://gatk.broadinstitute.org/hc/en-us/articles/360035890471-Hard-filtering-germline-short-variants`

[12] Poplin R, Chang PC, Alexander D, et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol*. 2018;36(10):983–987. `https://doi.org/10.1038/nbt.4235`

[13] DeepVariant developers. *DeepVariant FAQ*. Release r1.10. `https://github.com/google/deepvariant/blob/r1.10/docs/FAQ.md`

[14] Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. *arXiv* [preprint]. 2012;1207.3907. `https://doi.org/10.48550/arXiv.1207.3907`

[15] FreeBayes developers. *FreeBayes: a haplotype-based variant detector*. Software documentation. `https://github.com/freebayes/freebayes`

[16] Danecek P, Bonfield JK, Liddle J, et al. Twelve years of SAMtools and BCFtools. *GigaScience*. 2021;10(2):giab008. `https://doi.org/10.1093/gigascience/giab008`

[17] BCFtools developers. *bcftools(1) manual page*. `https://samtools.github.io/bcftools/bcftools`

[18] McLaren W, Gil L, Hunt SE, et al. The Ensembl Variant Effect Predictor. *Genome Biol*. 2016;17:122. `https://doi.org/10.1186/s13059-016-0974-4`

[19] Ensembl. *Variant Effect Predictor: cache documentation and software guide*. `https://www.ensembl.org/info/docs/tools/vep/script/vep_cache.html`; `https://github.com/Ensembl/ensembl-vep/tree/release/115`

[20] Ensembl. *Calculated variant consequences*. `https://static.ensembl.org/info/genome/variation/prediction/predicted_data.html`

[21] Cingolani P, Platts A, Wang LL, 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*. 2012;6(2):80–92. https://doi.org/10.4161/fly.19695

[22] SnpEff developers. *Building databases*. https://pcingola.github.io/SnpEff/snpeff/build_db/

[23] Krusche P, Trigg L, Boutros PC, et al. Best practices for benchmarking germline small-variant calls in human genomes. *Nat Biotechnol*. 2019;37(5):555–560. https://doi.org/10.1038/s41587-019-0054-x

[24] National Institute of Standards and Technology. *Genome in a Bottle*. https://www.nist.gov/programs-projects/genome-bottle

[25] GATK Team. *Mutect2*. https://gatk.broadinstitute.org/hc/en-us/articles/54698777535387-Mutect2

[26] PLINK developers. *PLINK 1.9: Input filtering*. https://www.cog-genomics.org/plink/1.9/filter#hwe

[27] Ensembl. *VEP command-line configuration*. Release 115. https://github.com/Ensembl/ensembl-vep/blob/release/115/modules/Bio/EnsEMBL/VEP/Config.pm

[28] SnpEff developers. *Running SnpEff*. https://pcingola.github.io/SnpEff/snpeff/running/

[29] Peter J, De Chiara M, Friedrich A, et al. Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates. *Nature*. 2018;556:339–344. https://doi.org/10.1038/s41586-018-0030-5

[30] BWA developers. *BWA manual*. https://bio-bwa.sourceforge.net/bwa.shtml

[31] Bowtie 2 developers. *Bowtie 2 manual*. https://bowtie-bio.sourceforge.net/bowtie2/manual.shtml

[32] Illumina. *Strelka2 germline and somatic small variant caller*. Archived software repository and user guide. https://github.com/Illumina/strelka

[33] VarScan developers. *Somatic Mutation Calling*. https://varscan.sourceforge.net/somatic-calling.html

[34] Galaxy Training Network. *Understanding Galaxy history system*. https://training.galaxyproject.org/training-material/topics/galaxy-interface/tutorials/history/tutorial.html

[35] Galaxy Training Network. *Creating, Editing and Importing Galaxy Workflows*. https://training.galaxyproject.org/training-material/topics/galaxy-interface/tutorials/workflow-editor/tutorial.html

[36] Olson ND, Wagner J, Dwarshuis N, et al. Variant calling and benchmarking in an era of complete human genome sequences. *Nat Rev Genet*. 2023;24:464–483. https://doi.org/10.1038/s41576-023-00590-0

[37] Liao WW, Asri M, Ebler J, et al. A draft human pangenome reference. *Nature*. 2023;617:312–324. https://doi.org/10.1038/s41586-023-05896-x

[38] Wagner J, Olson ND, McDaniel J, et al. Small variant benchmark from a complete assembly of X and Y chromosomes. *Nat Commun*. 2025;16:497. https://doi.org/10.1038/s41467-024-55710-z

[39] IGV developers. *IGV Desktop: Alignments basics*. https://igv.org/doc/desktop/UserGuide/tracks/alignments/viewing_alignments_basics/

[40] IGV developers. *IGV Desktop: Variants (VCF)*. https://igv.org/doc/desktop/UserGuide/tracks/vcf/

[41] Zheng Z, Sedlazeck FJ, Luo R. Small variant calling in the era of long-read sequencing. *Nat Rev Genet*. Published online 20 October 2026. https://doi.org/10.1038/s41576-026-01020-7

[42] OpenGene. *fastp documentation*. https://github.com/OpenGene/fastp