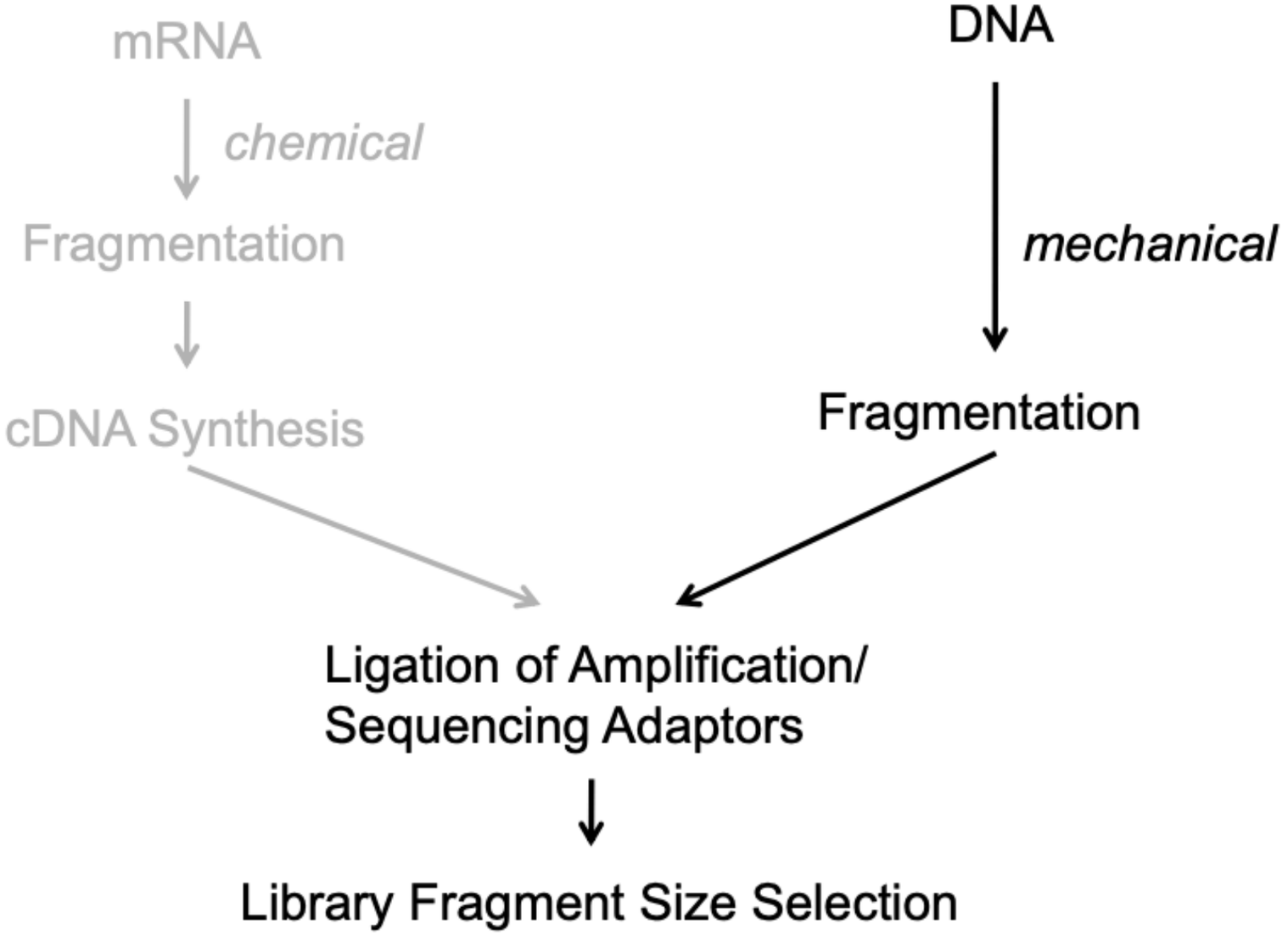
DNA sequencing

2025-03-24

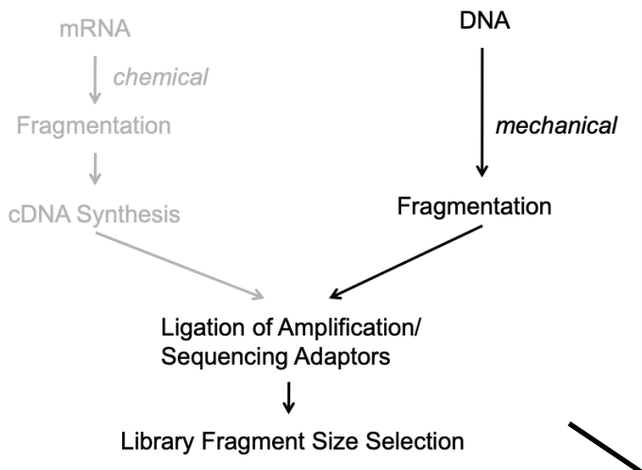
C.G.B.

UNIVERSITÀ DEGLI STUDI DI MILANO

Key steps in sequencing



Main types of DNA sequencing

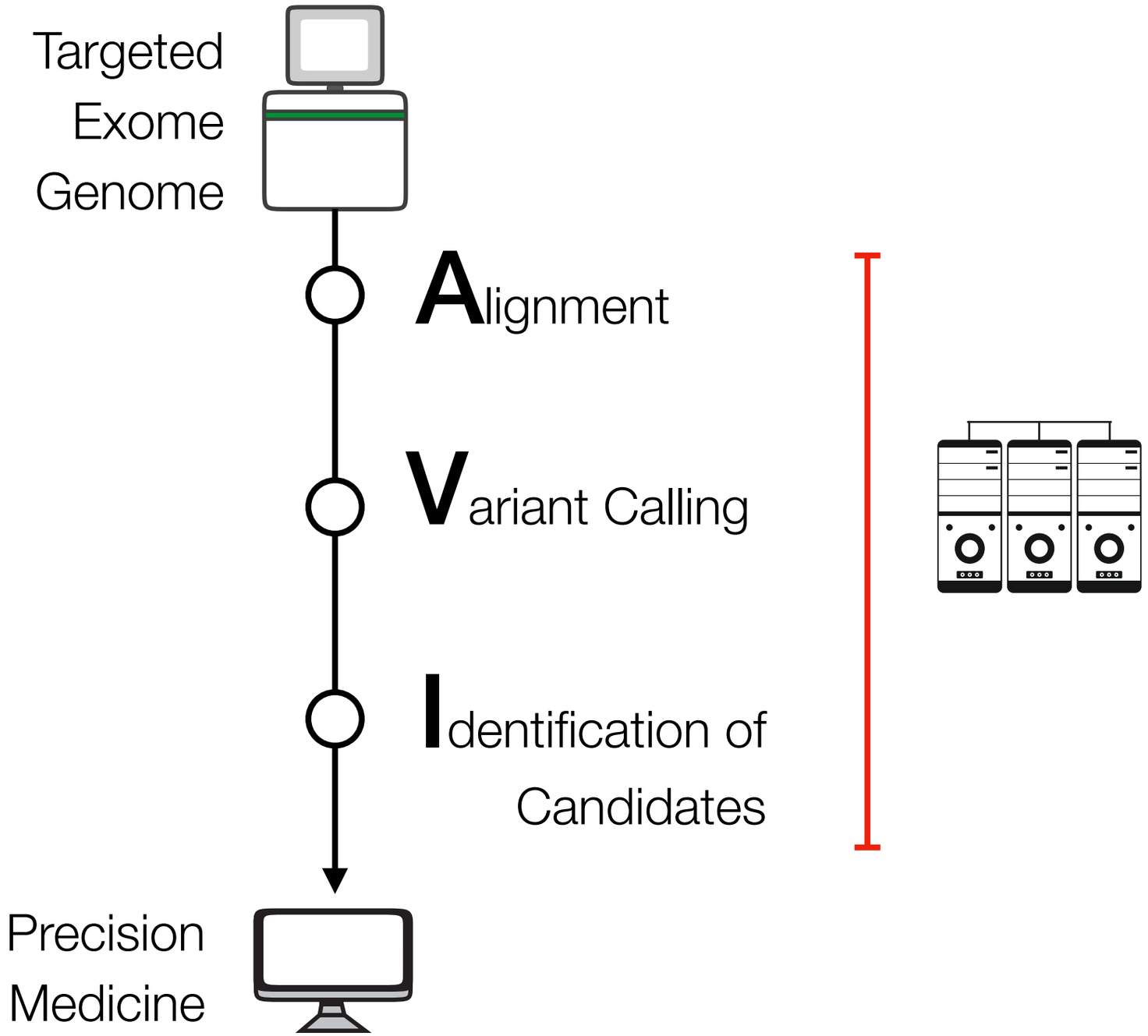


Targeted sequencing:
genes or regions of interest

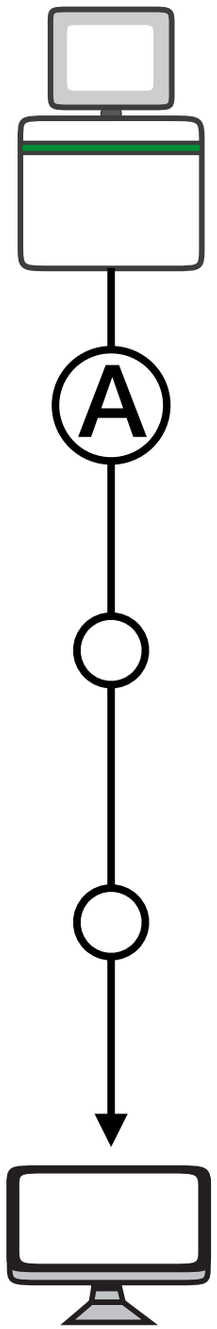
Whole EXome Sequencing
(WES/WXS): exome (coding sequences)

Whole Genome Sequencing
(WGS) : genome

Deciphering DNA-seq data



Alignment

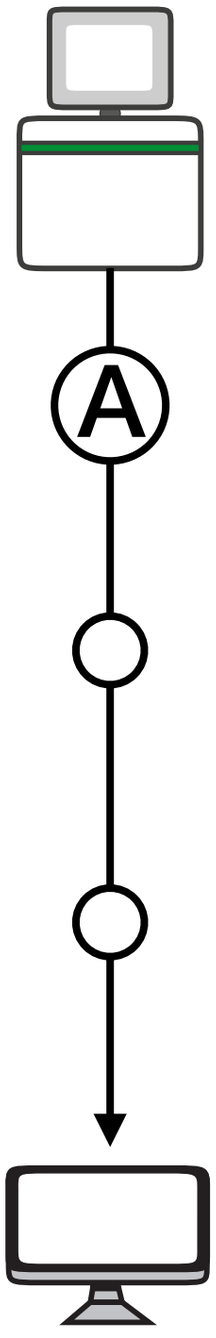


The first step in data analysis is alignment i.e. we need to understand where reads map on the genome.

Raw reads are usually found in **FASTQ format,** while the final output of the alignment is a **SAM/BAM** file.

The alignment requires a genome reference. The most recent release is GRCh38 (2013).

Base quality



The most used quality measure for sequencing data is the **Phred score**.

$$Q_{\text{PHRED}} = -10 \times \log_{10}(P_e)$$

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

In **fastq** format base quality is encoded in **ASCII**.

FastQC to do quality control on reads

FastQC Report

Summary

- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

Base quality distribution along the length of the read

Mean base quality distribution across reads

Percentage of each base along the sequence

Distribution of the percentage of GC along the sequence

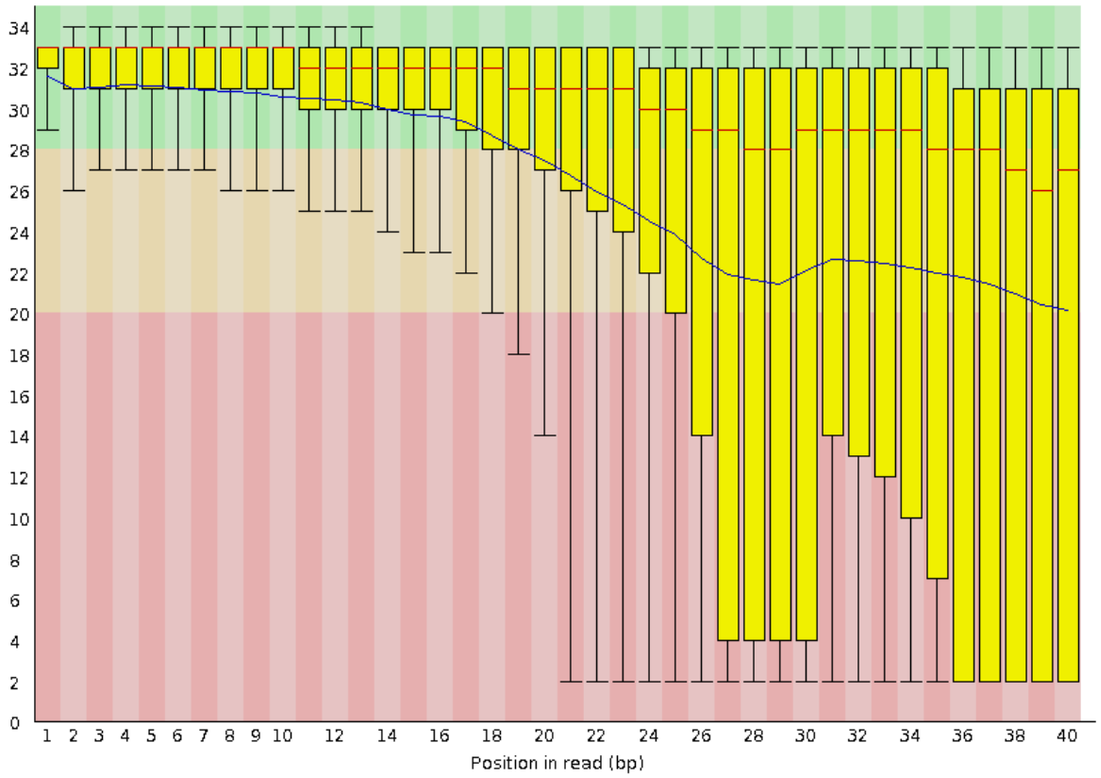
Distribution of the percentage of N bases (not properly called) along the sequence

Fragment length distribution

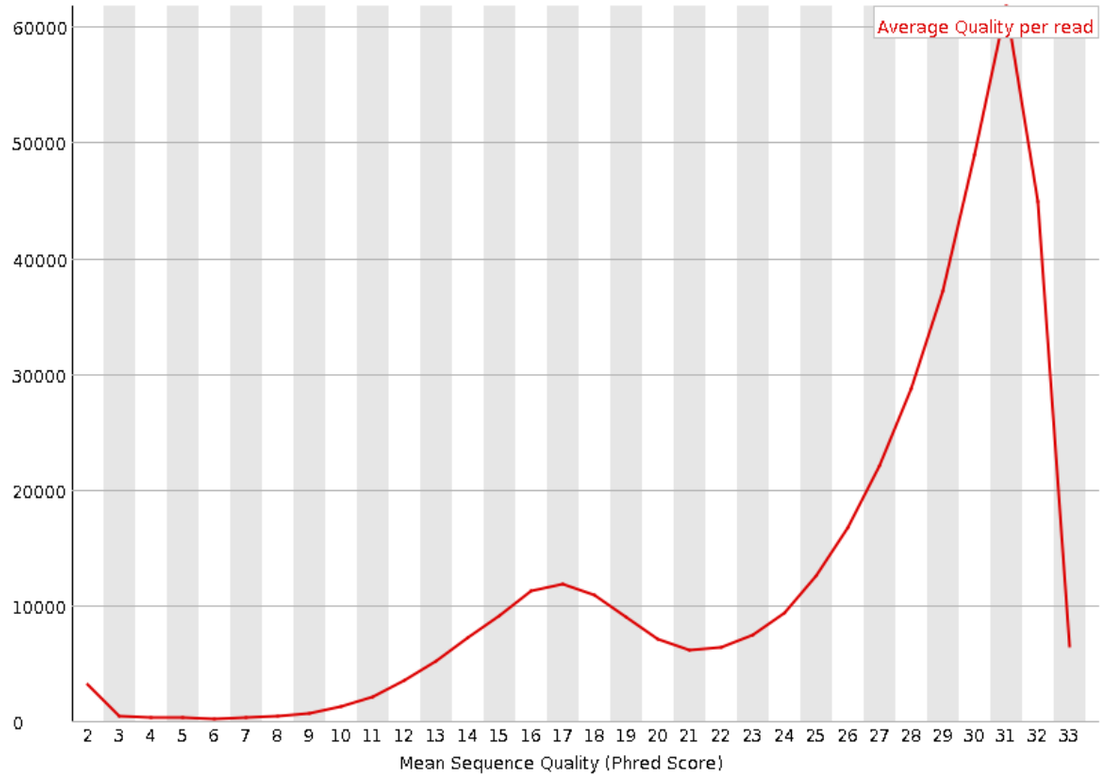
Duplication rate

If there are recurrent identical sequences

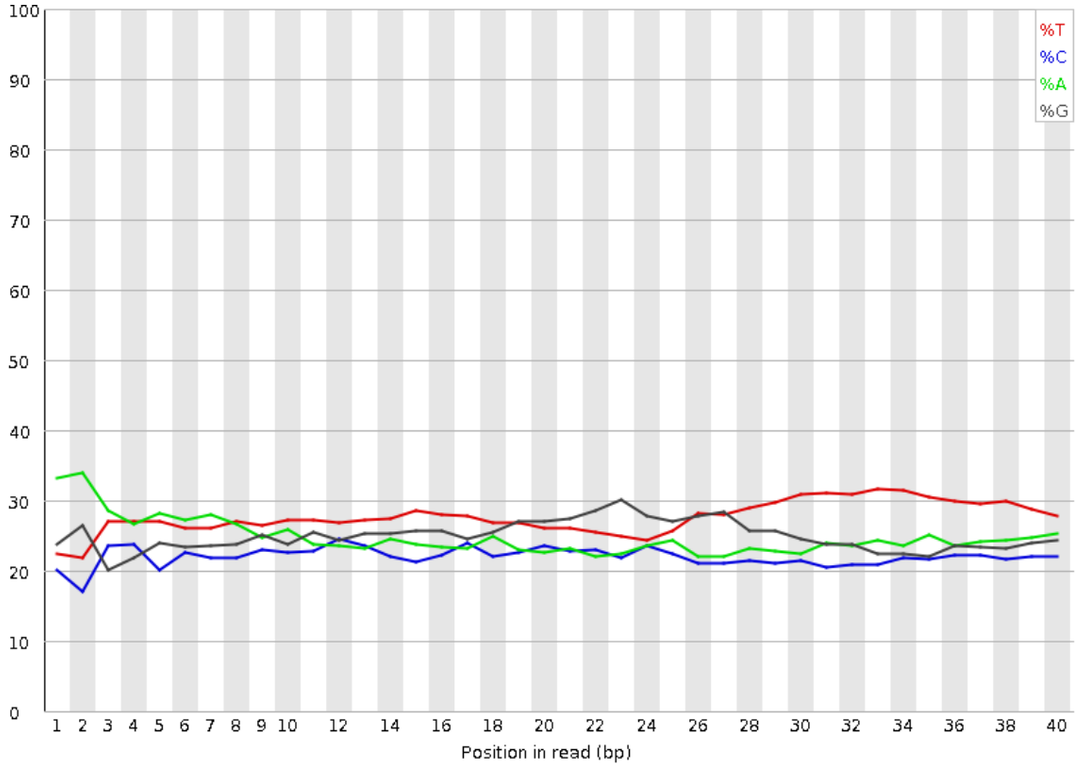
Base quality distribution along the read



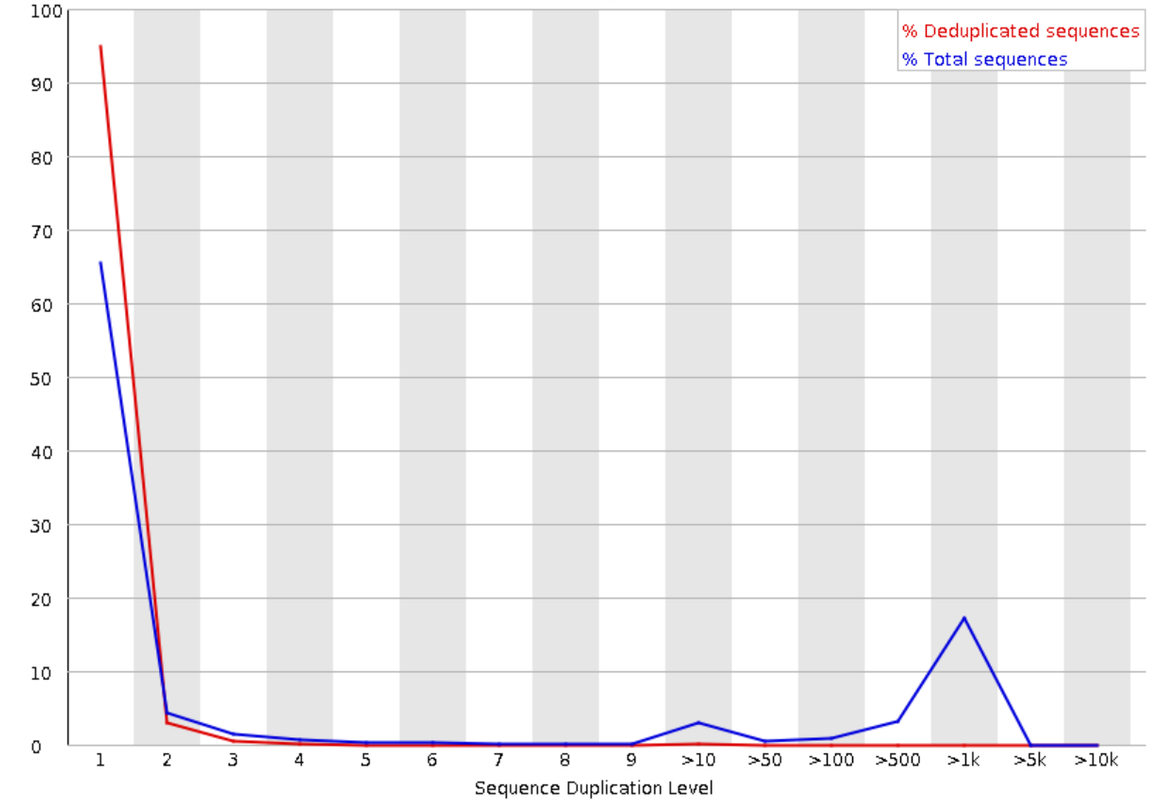
Mean base quality distribution



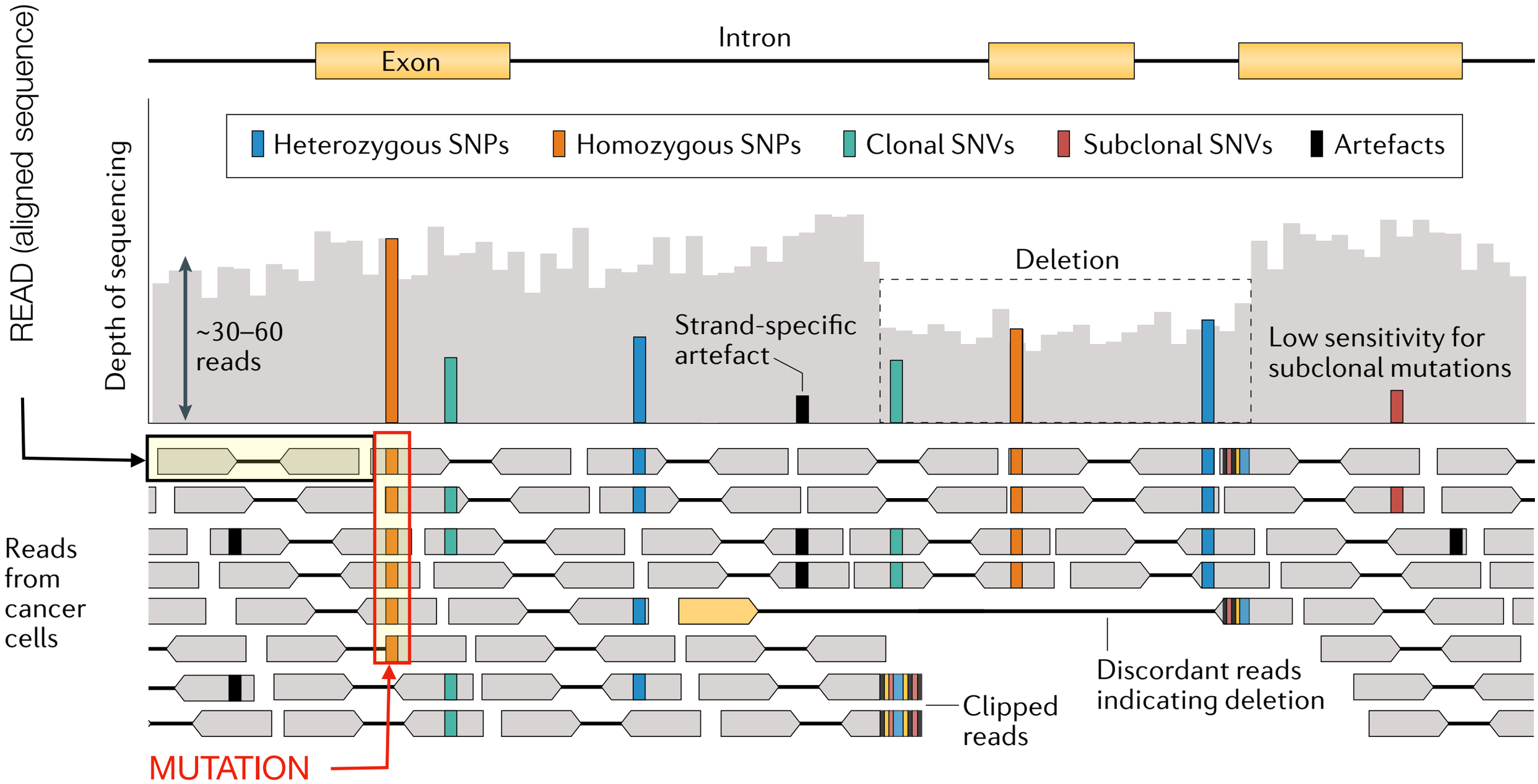
Percentage of each base along the sequence



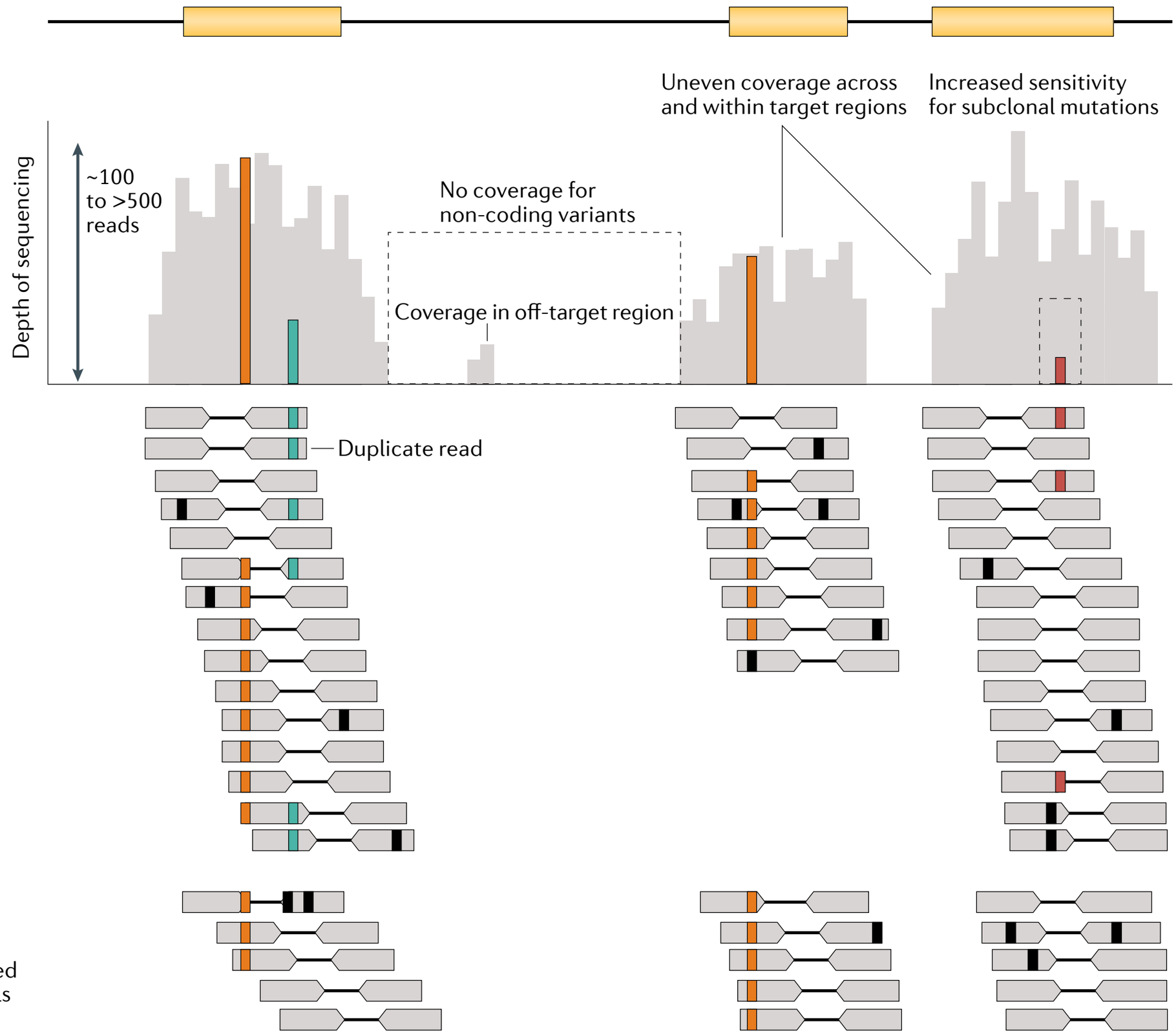
Duplication rate



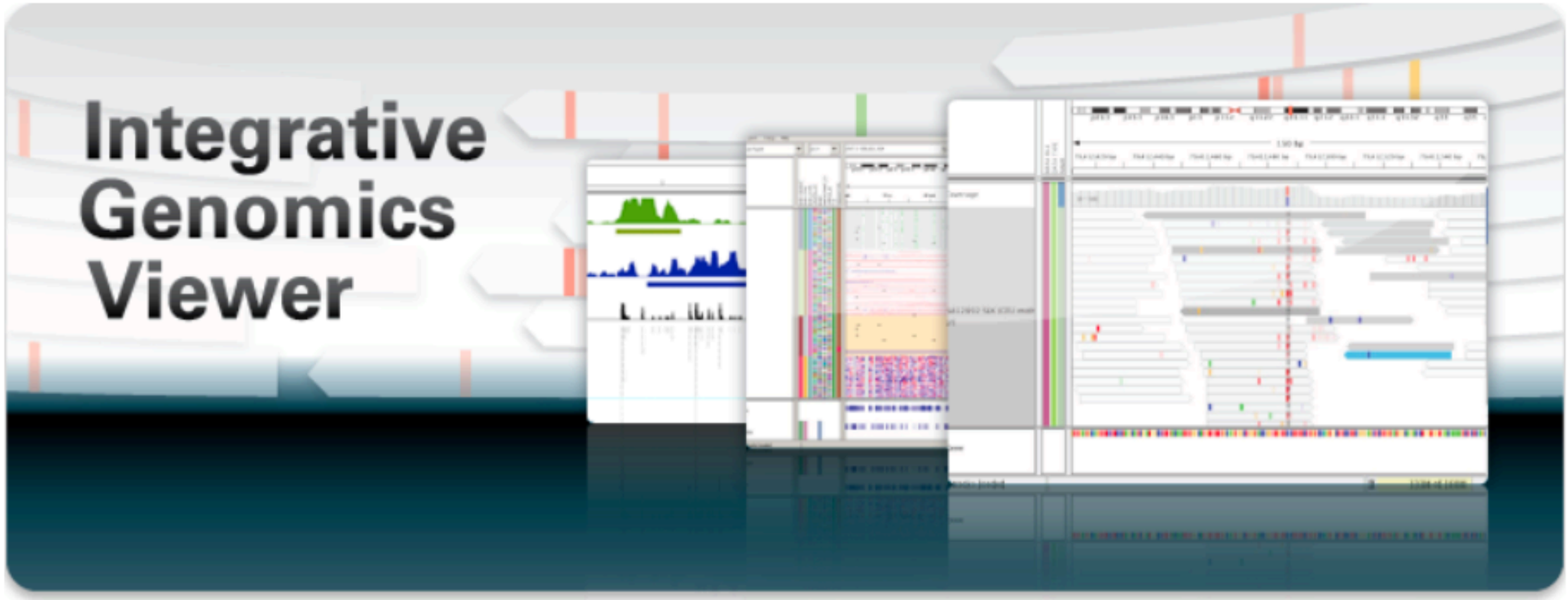
Whole-genome sequencing data



Whole-exome sequencing data



How can we see aligned data? Using IGV



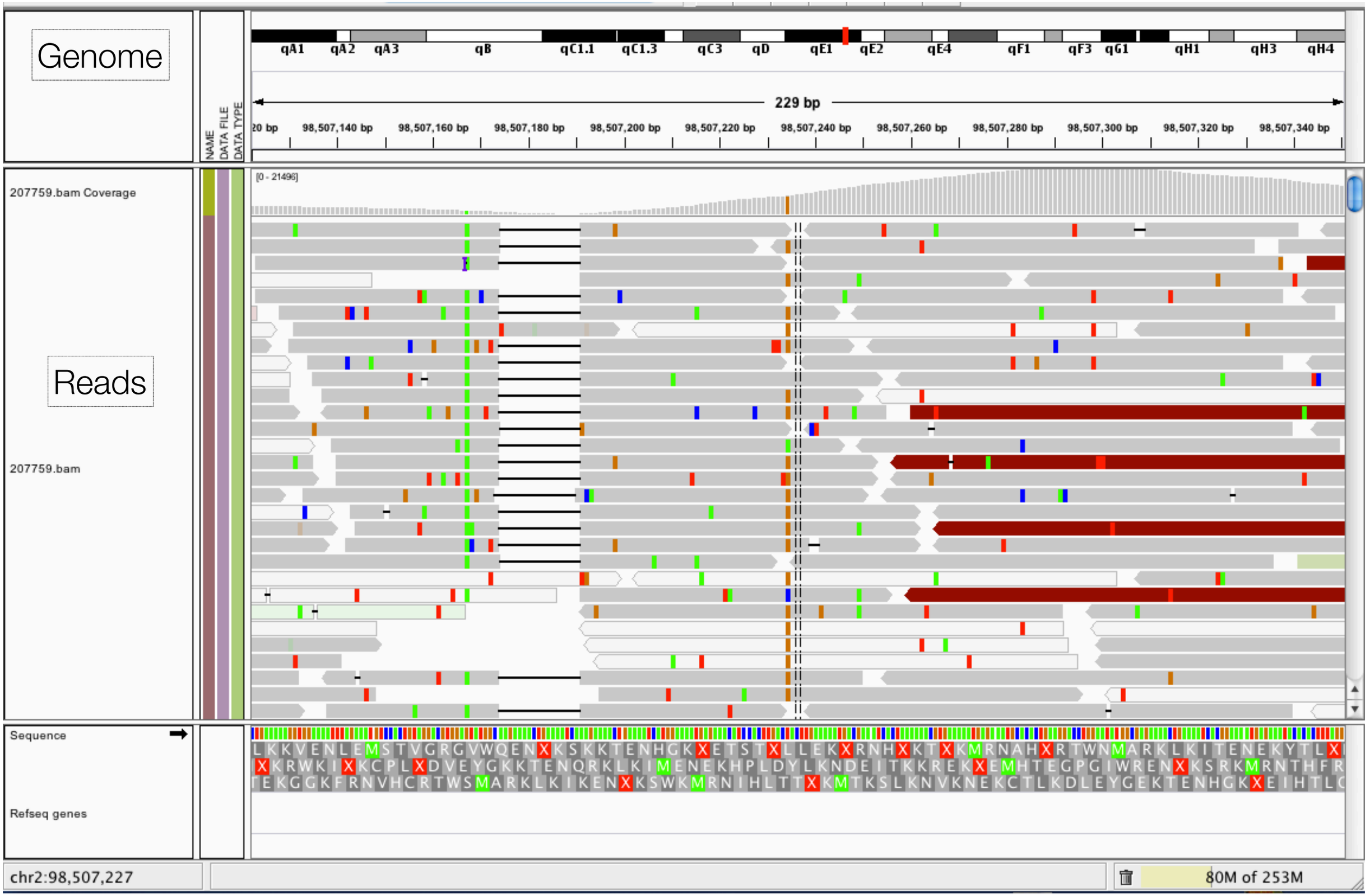
“The **Integrative Genomics Viewer (IGV)** is a high-performance, easy-to-use, interactive tool for the visual exploration of genomic data. It supports flexible integration of all the common types of genomic data and metadata, investigator-generated or publicly available, loaded from local or cloud sources.

IGV is available in multiple forms, including:

- the original **IGV** - a Java desktop application,
- **IGV-Web** - a web application,
- **igv.js** - a JavaScript component that can be embedded in web pages *(for developers)”*

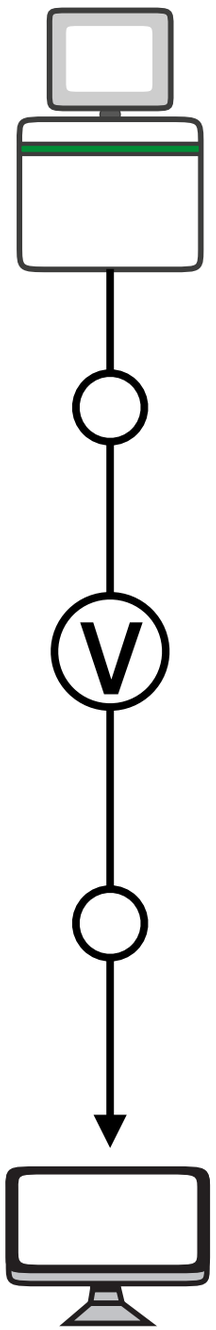
https://igv.org/doc/desktop/

The Integrative Genomics Viewer (IGV)



Mutations

Mutations identification

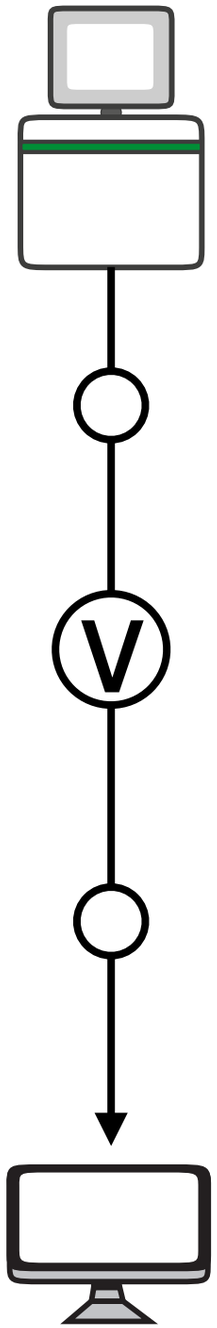


The main goal of variant identification is to evaluate if the alternative alleles supported by sequencing reads are true mutations or artefacts.

Vocabulary:

- **Single Nucleotide Polymorphisms** (SNPs): is a germline substitution of a single nucleotide at a specific position in the genome
- **Single Nucleotide Variants** (SNVs): a DNA sequence variation that occurs when a single nucleotide (adenine, thymine, cytosine, or guanine) in the genome sequence is altered. It is usually used for somatic mutations, nevertheless usually we can find the term “variant” for both somatic and germline mutations (Be aware of the context!!)
- **Small Insertions or Deletions** (InDels) (<50 bp)

Mutations identification



A lot of different tools are available for variant identification:

SOAP Short Oligonucleotide Analysis Package

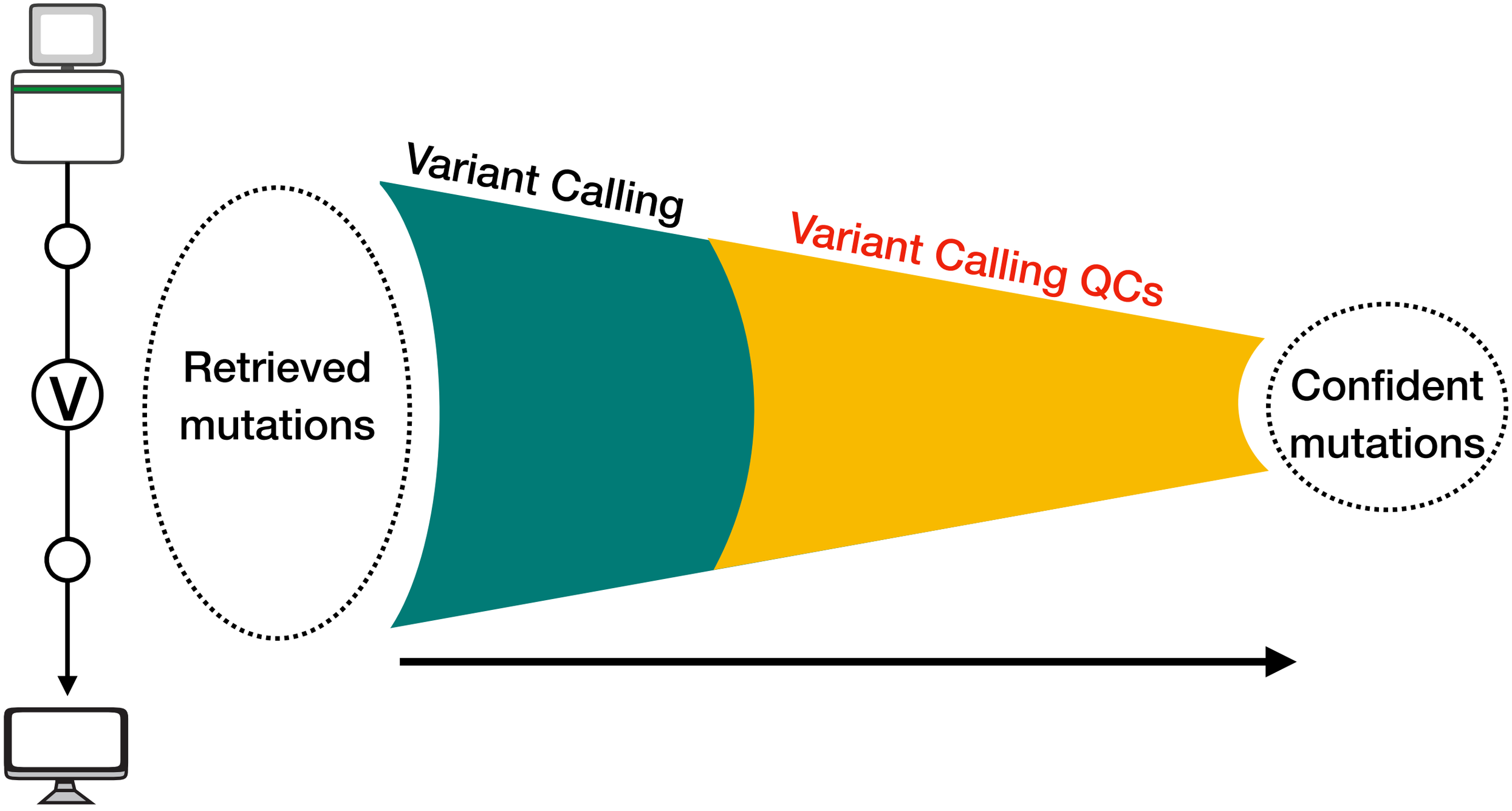
Strelka

SomaticSniper

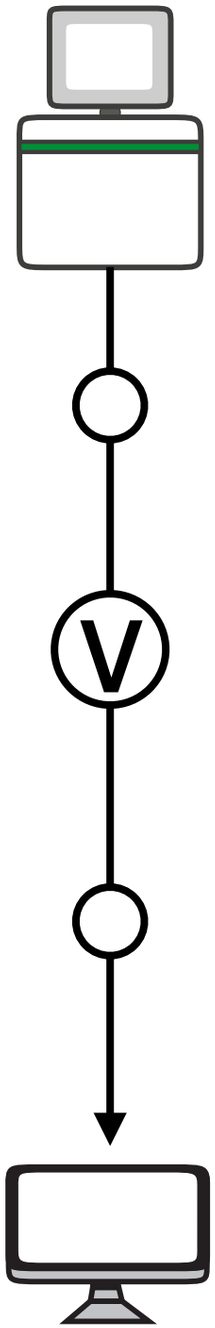
gatk

VarScan

Error remotion



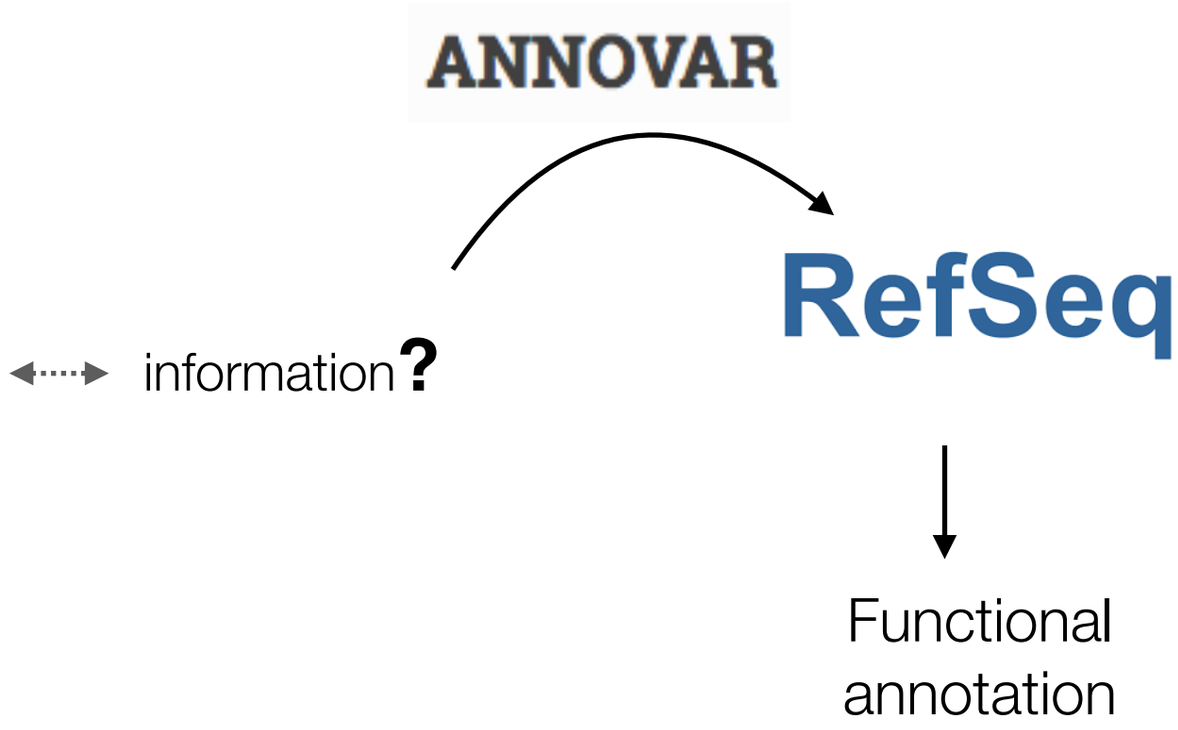
Functional annotation



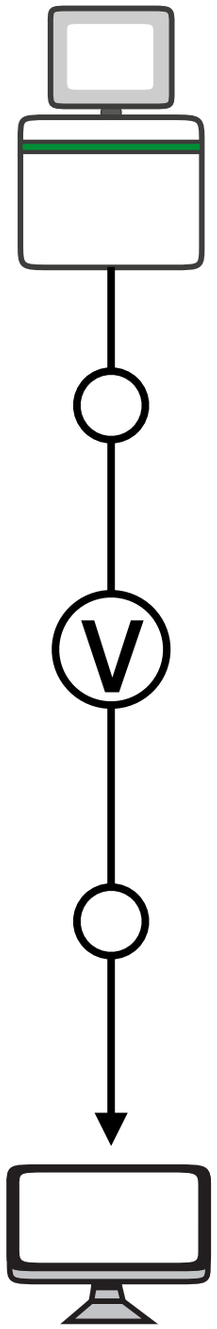
Mutated genomic positions have to be **annotated** to understand their **biological meaning**.

Can the identified SNP/SNV/InDel cause changes in protein coding and interested amino acids?

chrom	position	ref	var
chr17	17697102	G	A
chr17	21319977	A	G
chr17	26679861	G	A
chr17	28890301	G	A
chr17	36716758	G	A
chr17	40369149	G	A



Functional annotation

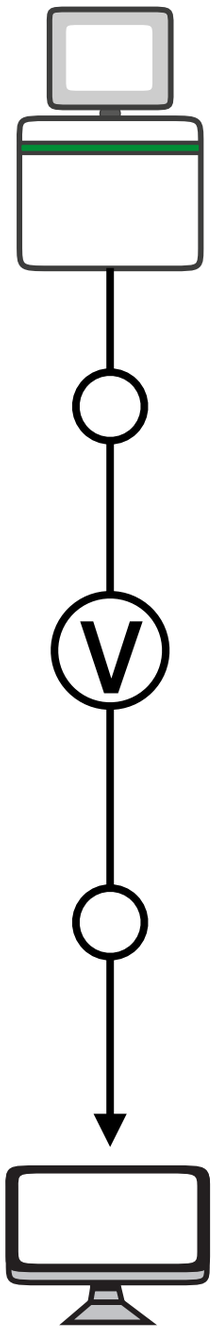


Functional annotation

chrom	position	ref	var	func	exonic.func	gene
chr17	17697102	G	A	exonic	synonymous SNV	RAI1
chr17	21319977	A	G	UTR3	-	KCNJ18
chr17	26679861	G	A	intronic	-	POLDIP2
chr17	28890301	G	A	exonic	nonsynonymous SNV	TBC1D29
chr17	36716758	G	A	intronic	-	SRCIN1
chr17	40369149	G	A	intronic	-	STAT5B

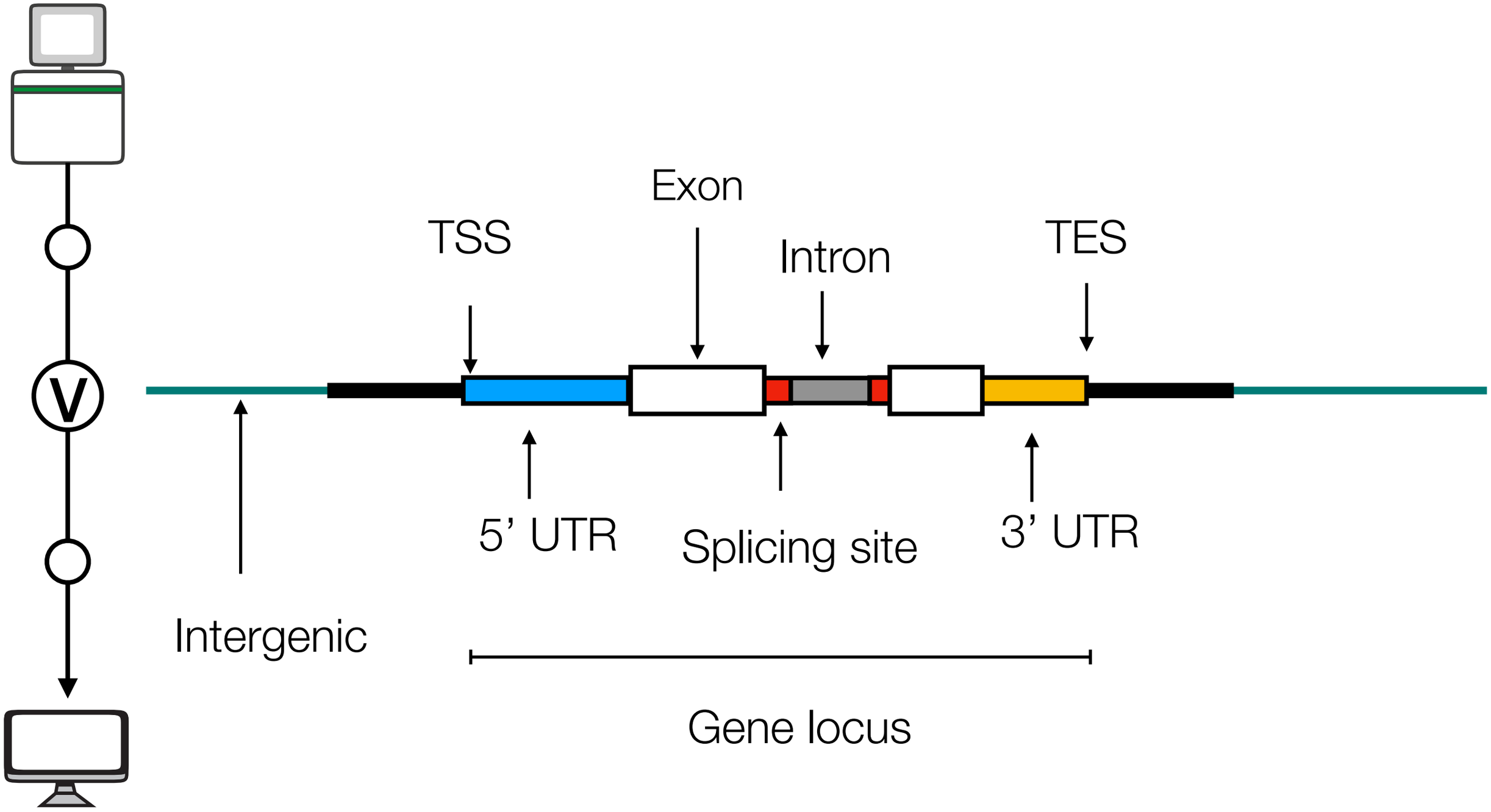
Exonic functional annotation

Functional annotation

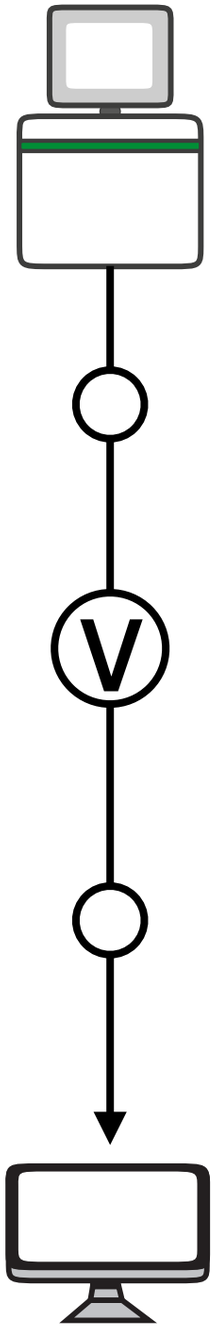


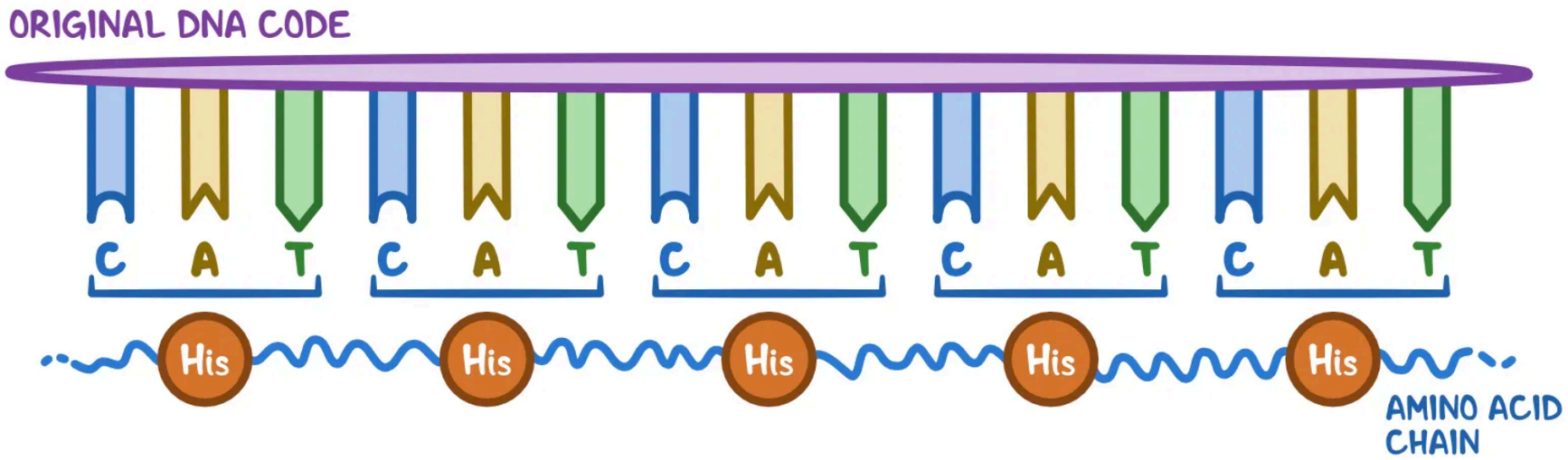
Value	Rank	Explanation
exonic	1	variant overlaps a coding
splicing	1	variant is within 2-bp of a splicing junction
ncRNA	2	variant overlaps a transcript without coding annotation in the gene definition
UTR5	3	variant overlaps a 5' untranslated region
UTR3	3	variant overlaps a 3' untranslated region
intronic	4	variant overlaps an intron
upstream	5	variant overlaps 1-kb region upstream of transcription start site
downstream	5	variant overlaps 1-kb region downtream of transcription end site
intergenic	6	variant is in intergenic region

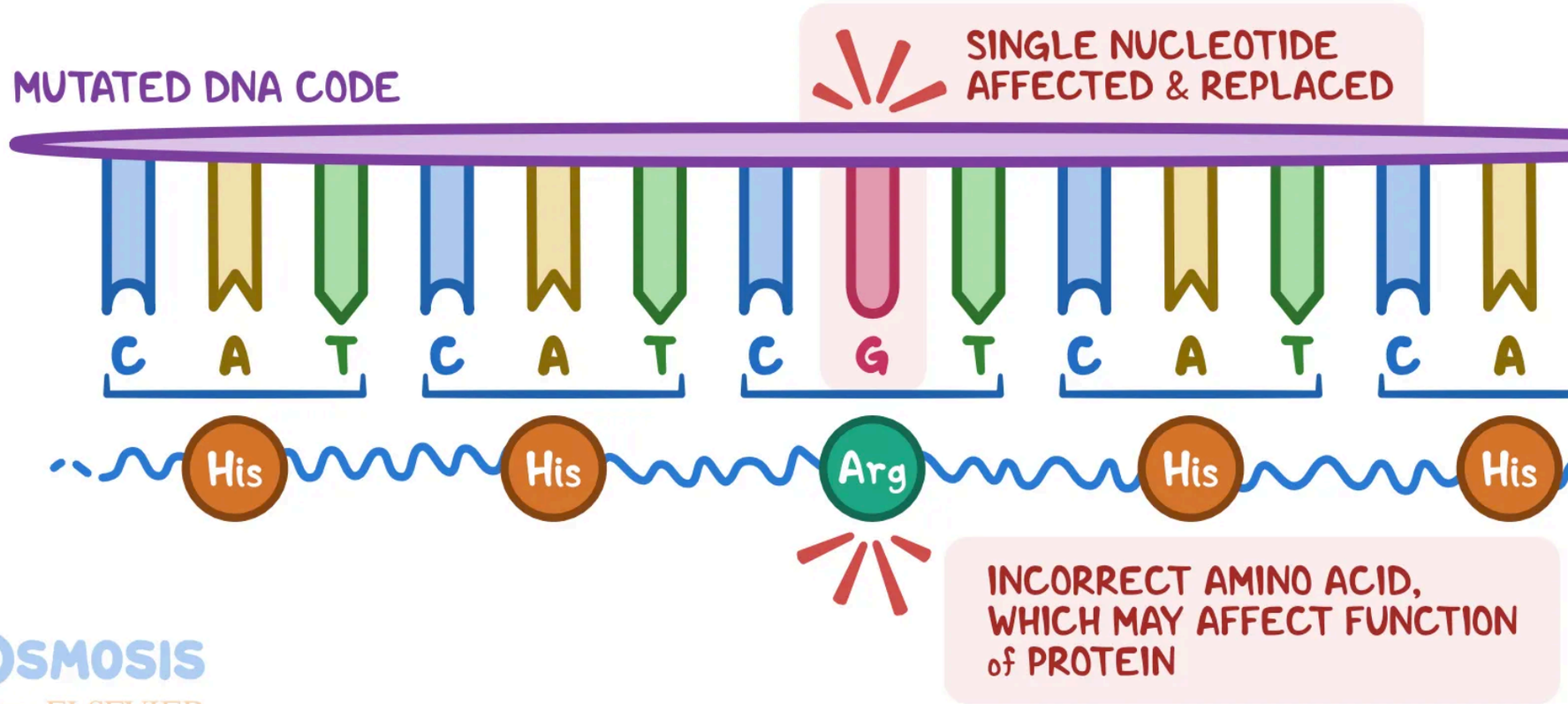
Functional annotation



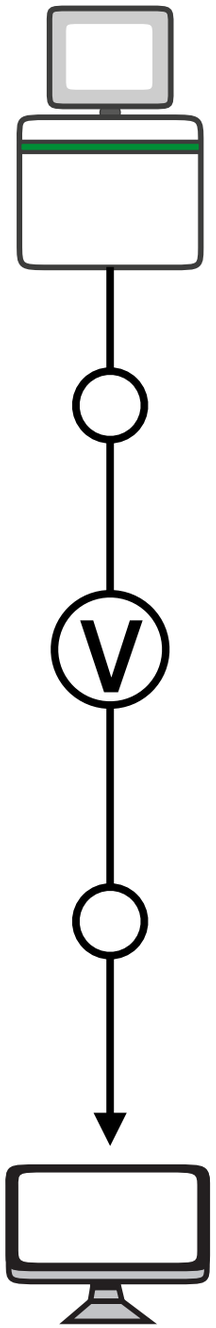
The effect of a mutation on protein





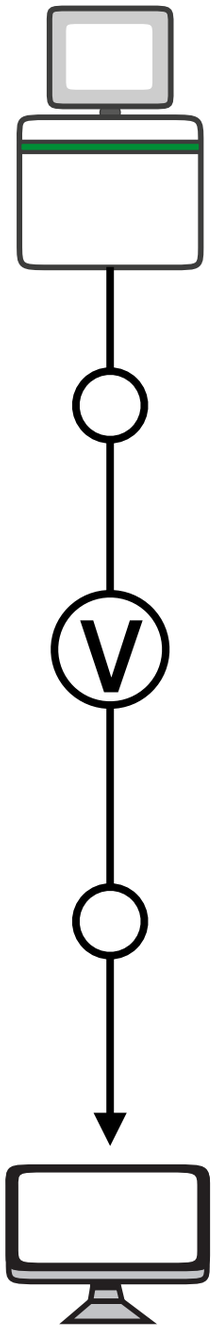


Exonic functional annotation



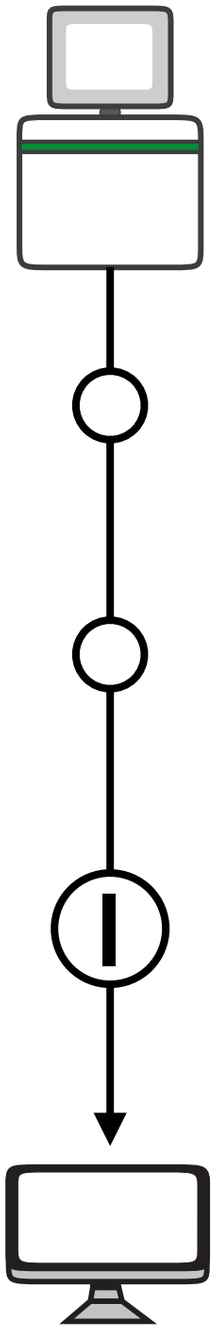
Annotation	Rank	Explanation
frameshift insertion	1	an insertion of one or more nucleotides that cause frameshift changes in protein coding sequence
frameshift deletion	2	a deletion of one or more nucleotides that cause frameshift changes in protein coding sequence
frameshift block substitution	3	a block substitution of one or more nucleotides that cause frameshift changes in protein coding sequence
stopgain	4	a variant lead to the immediate creation of stop codon at the variant site.
stoploss	5	a variant that lead to the immediate elimination of stop codon at the variant site
nonframeshift insertion	6	an insertion of 3 or multiples of 3 nucleotides that do not cause frameshift changes in protein coding sequence
nonframeshift deletion	7	a deletion of 3 or mutliples of 3 nucleotides that do not cause frameshift changes in protein coding sequence
nonframeshift block substitution	8	a block substitution of one or more nucleotides that do not cause frameshift changes in protein coding sequence
nonsynonymous SNV	9	a single nucleotide change that cause an amino acid change
synonymous SNV	10	a single nucleotide change that does not cause an amino acid change
unknown	11	unknown function (due to various errors in the gene structure definition in the database file)

Exonic functional annotation (nonsilent mutations)



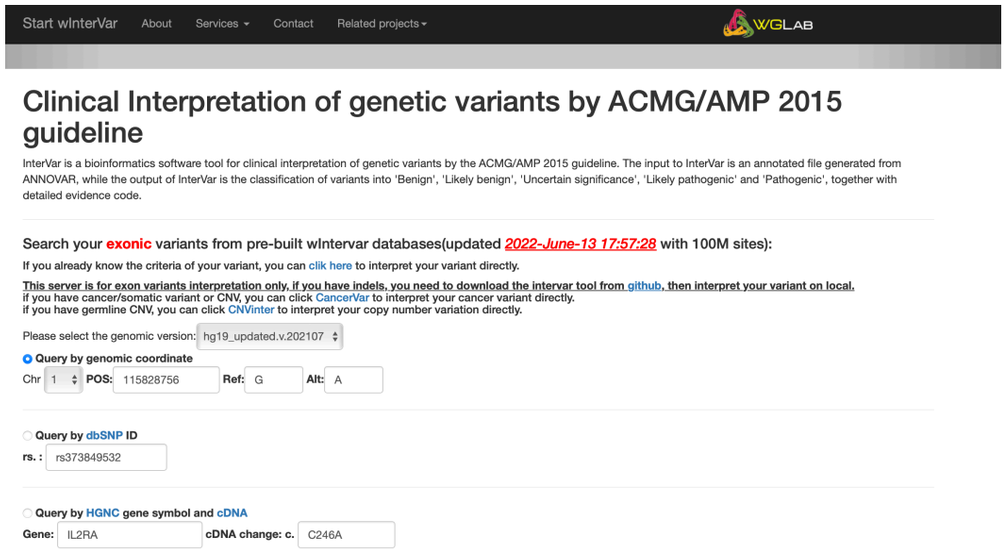
Annotation	Rank	Explanation
frameshift insertion	1	an insertion of one or more nucleotides that cause frameshift changes in protein coding sequence
frameshift deletion	2	a deletion of one or more nucleotides that cause frameshift changes in protein coding sequence
frameshift block substitution	3	a block substitution of one or more nucleotides that cause frameshift changes in protein coding sequence
stopgain	4	a variant lead to the immediate creation of stop codon at the variant site.
stoploss	5	a variant that lead to the immediate elimination of stop codon at the variant site
nonframeshift insertion	6	an insertion of 3 or multiples of 3 nucleotides that do not cause frameshift changes in protein coding sequence
nonframeshift deletion	7	a deletion of 3 or mutliples of 3 nucleotides that do not cause frameshift changes in protein coding sequence
nonframeshift block substitution	8	a block substitution of one or more nucleotides that do not cause frameshift changes in protein coding sequence
nonsynonymous SNV	9	a single nucleotide change that cause an amino acid change
synonymous SNV	10	a single nucleotide change that does not cause an amino acid change
unknown	11	unknown function (due to various errors in the gene structure definition in the database file)

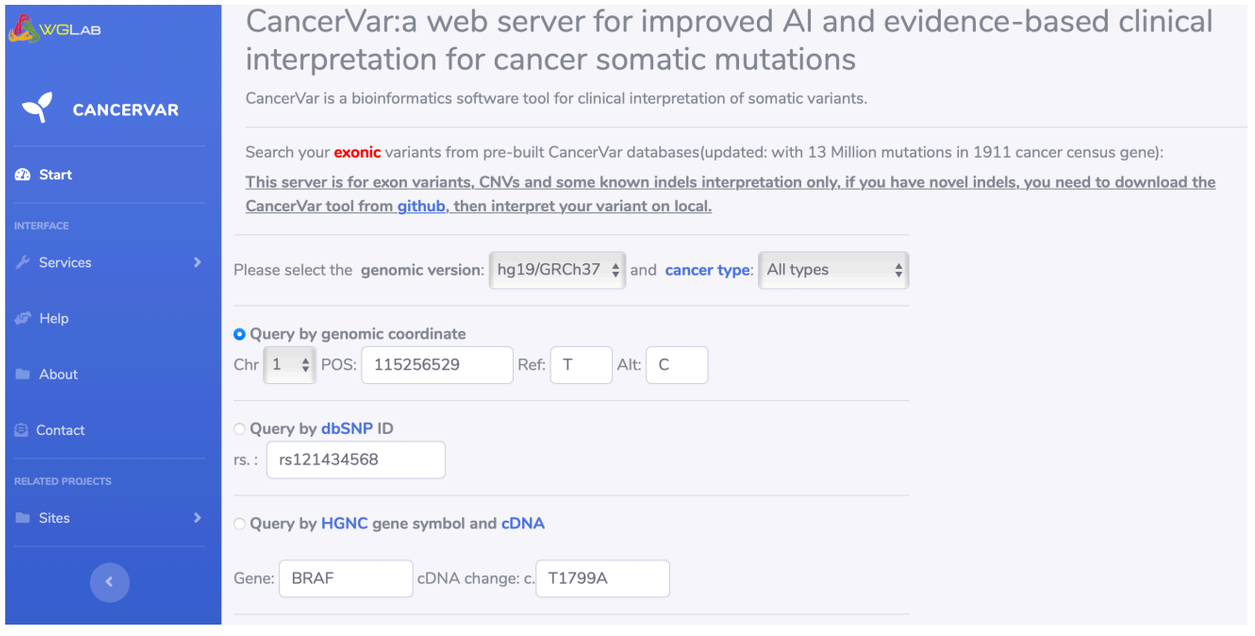
Clinical interpretation of variants



InterVar e **CancerVar** are curated databases for clinical interpretation. They contain catalogues of mutations previously pointed out to be pathogenetic or probably related to a disease.

https://wintervar.wglab.org/



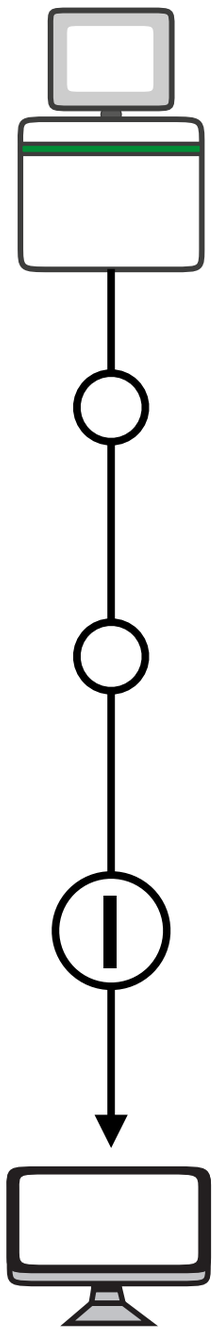


https://cancervar.wglab.org/

C.G.B.

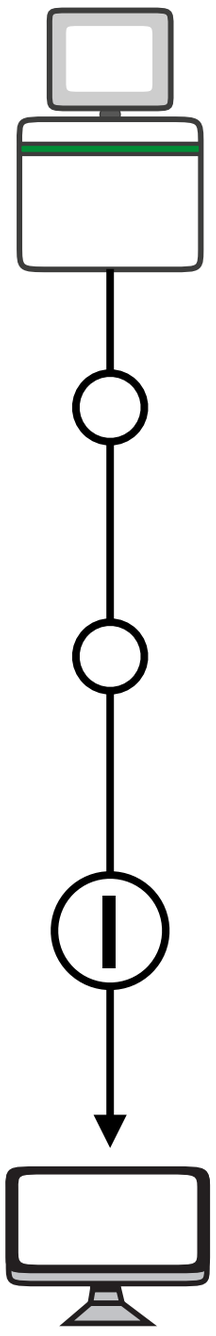
UNIVERSITÀ DEGLI STUDI DI MILANO

Clinical interpretation of germline mutations



In 2015, the ’American College of Medical Genetics and Genomics (ACMG) e l'Association for Molecular Pathology (AMP) published standard criterions and updated guidelines for the clinical interpretation of sequence variants (relations between variants and human diseases).

InterVar generates an automatic interpretation based on 28 criterions, classifying mutations as '**Benign**', '**Likely benign**', '**Uncertain significance**', '**Likely pathogenic**' and '**Pathogenic**'



InterVar:Guideline

The American College of Medical Genetics and Genomics (ACMG)and the Association for Molecular Pathology(AMP) published in 2015 the updated standards and guidelines for the clinical interpretation of sequence variants,based on 28 criteria. However, variability between individual interpreters may be extensive due to lack of standard algorithms that implement these guidelines. This ACMG/AMP2015 guideline is at here

Clinical Interpretation of genetic variants by ACMG/AMP 2015 guideline

InterVar is a bioinformatics software tool for clinical interpretation of genetic variants by the ACMG/AMP 2015 guideline. The input to InterVar is an annotated file generated from ANNOVAR, while the output of InterVar is the classification of variants into 'Benign', 'Likely benign', 'Uncertain significance', 'Likely pathogenic' and 'Pathogenic', together with detailed evidence code.

Search your exonic variants from pre-built wIntervar databases(updated *2022-June-13 17:57:28* with 100M sites):

If you already know the criteria of your variant, you can clik here to interpret your variant directly.

This server is for exon variants interpretation only, if you have indels, you need to download the intervar tool from github, then interpret your variant on local.
if you have cancer/somatic variant or CNV, you can click CancerVar to interpret your cancer variant directly.
if you have germline CNV, you can click CNVinter to interpret your copy number variation directly.

Please select the genomic version: hg19_updated.v.202107

○ **Query by genomic coordinate**

Chr 1 **POS:** 115828756 **Ref:** G **Alt:** A

○ **Query by dbSNP ID**

rs. : rs373849532

○ **Query by HGNC gene symbol and cDNA**

Gene: IL2RA **cDNA change: c.** C246A

◉ **Query by HGNC gene symbol and Protein Change**

Gene: KRAS **Protein change: p.** G12C

C.G.B.

Clinical Interpretation of genetic variants by ACMG/AMP 2015 guideline

InterVar is a bioinformatics software tool for clinical interpretation of genetic variants by the ACMG/AMP 2015 guideline. The input to InterVar is an annotated file generated from ANNOVAR, while the output of InterVar is the classification of variants into 'Benign', 'Likely benign', 'Uncertain significance', 'Likely pathogenic' and 'Pathogenic', together with detailed evidence code.

Warning: All listed results were from the automated interpretation on default parameters!
Users are advised to examine detailed evidence and use prior knowledge on ethnicity/disease to perform manual adjustments.

Database version:hg19_update
You searched by HGNC gene symbol with name as **KRAS**, and Protein change: **p.G12C**

Show/hide columns | Restore columns | Copy to clipboard | Download result as CSV

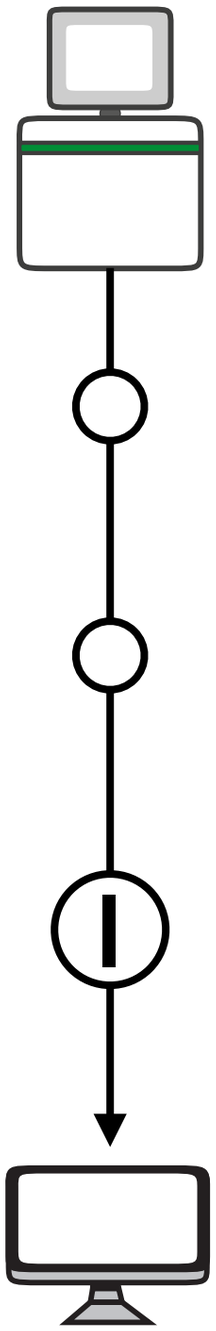
Search:

Chr	Position	Ref	Alt	Gene (refGene)	Intervar	ExonicFunc (refGene)	SNP	Transcripts (Ref)	MAF in gnomAD_ALL(genome)	Disease in OrphaNet
12	25398285	C	A	KRAS	Pathogenic (Details&Adjust)	nonsynonymous SNV	rs121913530(details of MAF)	NM_004985 p.G12C NM_033360 p.G12C	. (show in 7 POPs)	1340 268114 648 519 1333 2612 26106 227535

Showing 1 to 1 of 1 entries

Previous 1 Next

Clinical interpretation of somatic mutations



CancerVar is a bioinformatic tool for the clinical interpretation of somatic variants based on guidelines of the AMP/ASCO/CAP/ CGC 2017-2019

CancerVar classifies somatic variants as:

- Tier I/**Pathogenic**
- Tier II/**Likely pathogenic**
- Tier III/**Variants of Unknown Clinical Significance (VUS)**
- Tier IV/**Benign or Likely Benign Variants**

Clinical interpretation of somatic mutations

Tier I: Variants of Strong Clinical Significance

Therapeutic, prognostic & diagnostic

Level A Evidence

FDA-approved therapy

Included in professional guidelines

Level B Evidence

Well-powered studies with consensus from experts in the field

Tier II: Variants of Potential Clinical Significance

Therapeutic, prognostic & diagnostic

Level C Evidence

FDA-approved therapies for different tumor types or investigational therapies

Multiple small published studies with some consensus

Level D Evidence

Preclinical trials or a few case reports without consensus

Tier III: Variants of Unknown Clinical Significance

Not observed at a significant allele frequency in the general or specific subpopulation databases, or pan-cancer or tumor-specific variant databases

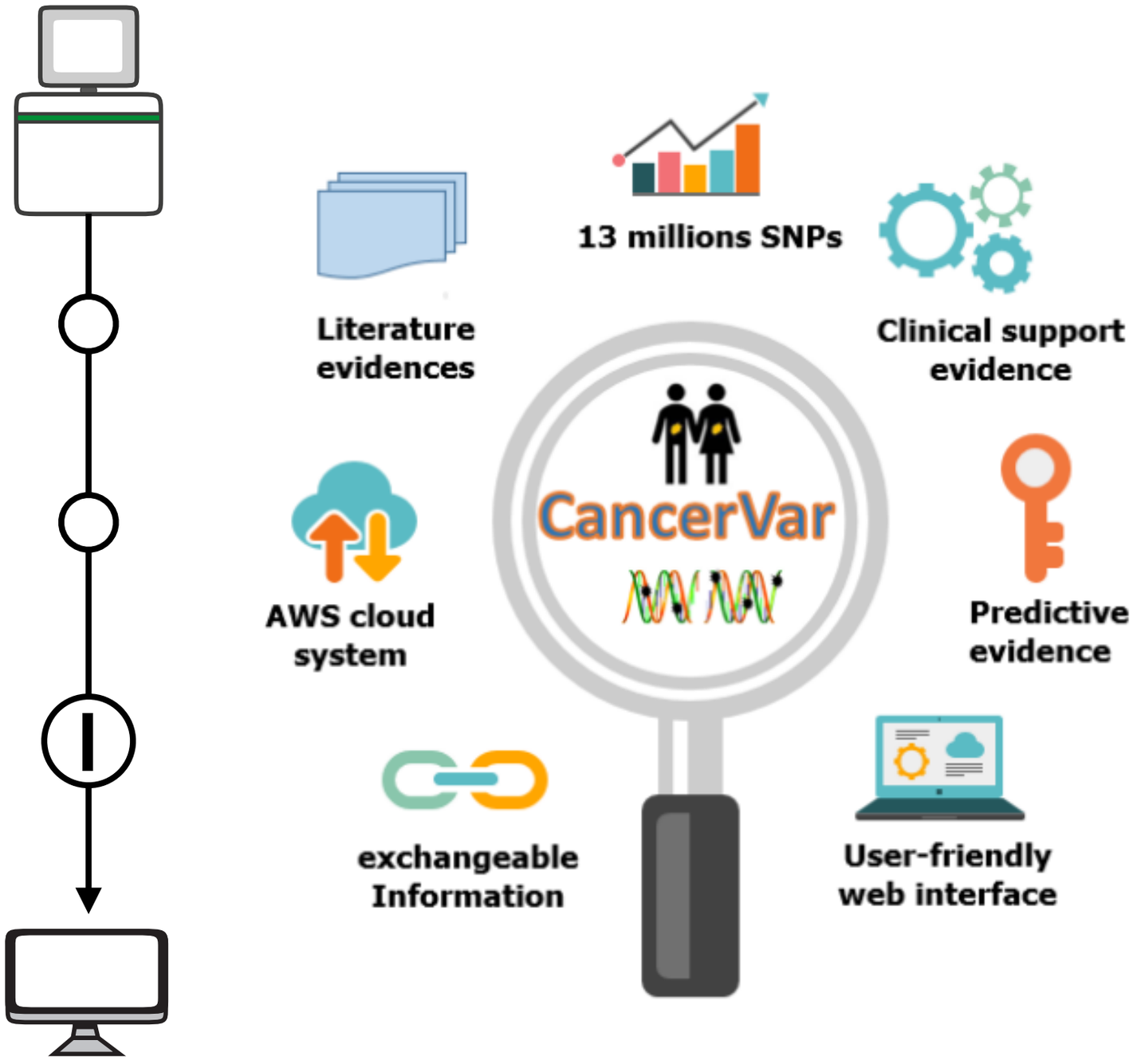
No convincing published evidence of cancer association

Tier IV: Benign or Likely Benign Variants

Observed at significant allele frequency in the general or specific subpopulation databases

No existing published evidence of cancer association

Clinical interpretation of somatic mutations



clinical-based prediction

Therapeutic
FDA approved or investigational with strong evidence

Diagnostic
In Professional Guideline or reported evidence with consensus

Prognostic
In Professional Guideline or reported evidence with consensus

Mutation type
Activating, LOF (missense, nonsense, indel, splicing)

Population data
Absent or extremely low MAF in population databases

Somatic data
Most likely present in COSMIC and ICGC

Predictive Softwares
Most of Prediction showed pathogenic in

Pathway involvement
Involve disease-associated pathways or pathogenic pathways

Germline database
Present in Clinvar/HGMD as Pathogenic

Publication database
Publication reported pathogenic evidence with consensus

CancerVar:a web server for improved AI and evidence-based clinical interpretation for cancer somatic mutations

CancerVar is a bioinformatics software tool for clinical interpretation of somatic variants.

Search your **exonic** variants from pre-built CancerVar databases(updated: with 13 Million mutations in 1911 cancer census gene):

This server is for exon variants, CNVs and some known indels interpretation only, if you have novel indels, you need to download the CancerVar tool from github, then interpret your variant on local.

Please select the **genomic version**: hg19/GRCh37 and **cancer type**: All types

○ **Query by genomic coordinate**

Chr 1 POS: 115256529 Ref: T Alt: C

○ **Query by dbSNP ID**

rs. : rs121434568

○ **Query by HGNC gene symbol and cDNA**

Gene: BRAF cDNA change: c. T1799A

◉ **Query by HGNC gene symbol and Protein Change**

Gene: KRAS Protein change: p. G12C

○ **Query by HGNC gene symbol or Alternations**

Gene: ERBB2 Copy Number

Submit Reset

CancerVar Results

You searched by HGNC gene symbol with name as **KRAS**, cancer types: **All_types** and Protein change: **P.G12C**

GENOMIC VERSION
hg19

CANCER TYPE
All_types

ALTERNATIONS
Mutation

INTERPRETATION SUMMARY

Based on Evidence(CBPs), CancerVar assign the clinical significance of your somatic mutation in KRAS as:

Tier_I_strong and Score: **11(with sub-scores)** **Deep learning Score from OPAI: 0.99(Oncogenic)**

This variant is nonsynonymous SNV in gene KRAS, located in chromosome 12:25398285. There are some clinical and/or experimental evidence showed strong/potential clinical significance in Therapeutic,Diagnosis,Prognosis, From the population databases(gnomAD,1000 genome etc), this variant is absent or extremely low minor allele frequency(MAF<0.1%). In Germline database of Clinvar/HGMD, this variant's clinical significane is ***Pathogenic***. In most of the pathogenic or deleterious prediction softwares/algorithms, this variant was predicted as ***Pathogenic***. When check the occurrence of somatic database in COSMIC or ICGC, this variant shows in ***both*** of them. From the KEGG pathway database, the gene of this variant **does** involve in disease-associated pathways or pathogenic pathways. Currently, searching the pubmed website, there are **some** publications from functional study, population study or other study as supporting evidence for clinical/biological significance.

Please review the cards below to get the detail of the interpretation.

Evidence Overview

CBP1: Therapeutic: FDA approved or investigational with strong evidence.In total of 78 records (you specified cancer types as: ***All_types***),most of the cancers are located in:**Lung 35%,Colorectal 33%,Cancer 8%,**and they have been mostly treated with:**Melphalan 5%,Cetuximab 4%,EGFR mAb inhibitor 4%,**the treatment of the drugs are:**Responsive 35%,Sensitivity/Response 28%,Resistance 21%,**if you need more information, please click Detail...

CBP2: Diagnostic: In Professional guideline or reported evidence with consensus.In total of 1 records (you specified cancer types as: ***All_types***),most of the cancers are located in:**Lung 100%,** the diagnostic are:**Positive 100%,**if you need more information, please click Detail...

CBP3: Prognostic: In Professional guideline or reported evidence with consensus.In total of 4 records (you specified cancer types as: ***All_types***),most of the cancers are located in:**Lung 75%,Other 25%,** the prognostic are:**Poor Outcome 100%,**if you need more information, please click Detail...

CBP4: Mutation type: Activating, LOF (missense, nonsense, indel, splicing), CNAs, fusions.

CBP5: Variant frequencies:Mostly mosaic. Need user's knowledge.

CBP6: Potential germline: Mostly nonmosaic. Need user's knowledge.

CBP7: Population databases: Absent or extremely low MAF. MAF In GnomAD_genome . (show in 7 POPs)

CBP8: Germline databases: may be present in HGMD/ClinVar **Pathogenic**.

CBP9: Somatic databases: Most present in COSMIC, ICGC, My Cancer Genome, TCGA.

CBP10: Predictive from: SIFT, PolyPhen2,MutationAssessor,MetaSVM,MetaLR,FATHMM,GERP++_RS, and mostly as **Pathogenic**.

CBP11: Pathway: involve in Disease-associated pathways or pathogenic pathways. KEGG Pathway

CBP12: Publications: Convincing evidence from Functional study, population study, other.

Evidence

Mutation Information

Chromosome: 12
Position: 25398285
Reference Allele: C
Alternative Allele: A
Function in refGene: nonsynonymous SNV
Minor allele frequency in(. means absent):
ESP6500:.
1000 genome:.
gnomAD genome_ALL:. More
ExAC:1.976E-5
Transcript in refGene: NM_033360
Exon location:exon2
Nucleotide change:c.34G>T
Residue change :p.G12C
Clinvar: Pathogenic/Likely_pathogenic
While the KRAS G12 region is a widely studied recurrent region in cancer, its impact on clinical action is still debated. Often associated with tumors that are wild-type for other drivers (EGFR and ALK specifically), the prognosis for patients with this mutation seems to be worse than the KRAS wild-type cohort in patients with colorectal and pancreatic cancer, however this hypothesis is in need of further validation. This mutation, along with the mutations affecting the neighboring G13 position, may result in a less responsive tumor when treated with first-generation TKI's like gefitinib. However, cetuximab treatment was shown to extend survival in a cohort of colorectal patients.
Deep learning Score: 0.99(OPAI,Oncogenic)
CancerVar: Tier_I_strong with score: 11
Need to review scores or adjust and manually re-interpret? Please Click Adjust!

Clinical significance Unknown Benign

Mutation

Gene Information

KRAS

Name: Kirsten rat sarcoma viral oncogene homolog
Location: chr12:25357723-25403870(Grch37)
Cytoband: 12p12.1

GeneCards Summary:

KRAS (KRAS Proto-Oncogene, GTPase) is a Protein Coding gene. Diseases associated with KRAS include Oculoectodermal Syndrome and Noonan Syndrome 3. Among its related pathways are Oocyte meiosis and Oxytocin signaling pathway. Gene Ontology (GO) annotations related to this gene include GTP binding. An important paralog of this gene is NRAS. .... More on GeneCards

CIViC Summary:

Mutations in the RAS family of proteins are frequently observed across cancer types. The amino acid positions that account for the overwhelming majority of these mutations are G12, G13 and Q61. The different protein isoforms, despite their raw similarity, also behave very differently when expressed in non-native tissue types, likely due to differences in the C-terminal hyper-variable regions. Mis-regulation of isoform expression has been shown to be a driving event in cancer, as well as missense mutations at the three hotspots previously mentioned. While highly recurrent in cancer, attempts to target these RAS mutants with inhibitors have not been successful, and has not yet become common practice in the clinic. The prognostic implications for KRAS mutations vary between cancer types, but have been shown to be associated with poor outcome in colorectal cancer, non-small cell lung cancer, and others.

Gene

Clinical publications

Pathway

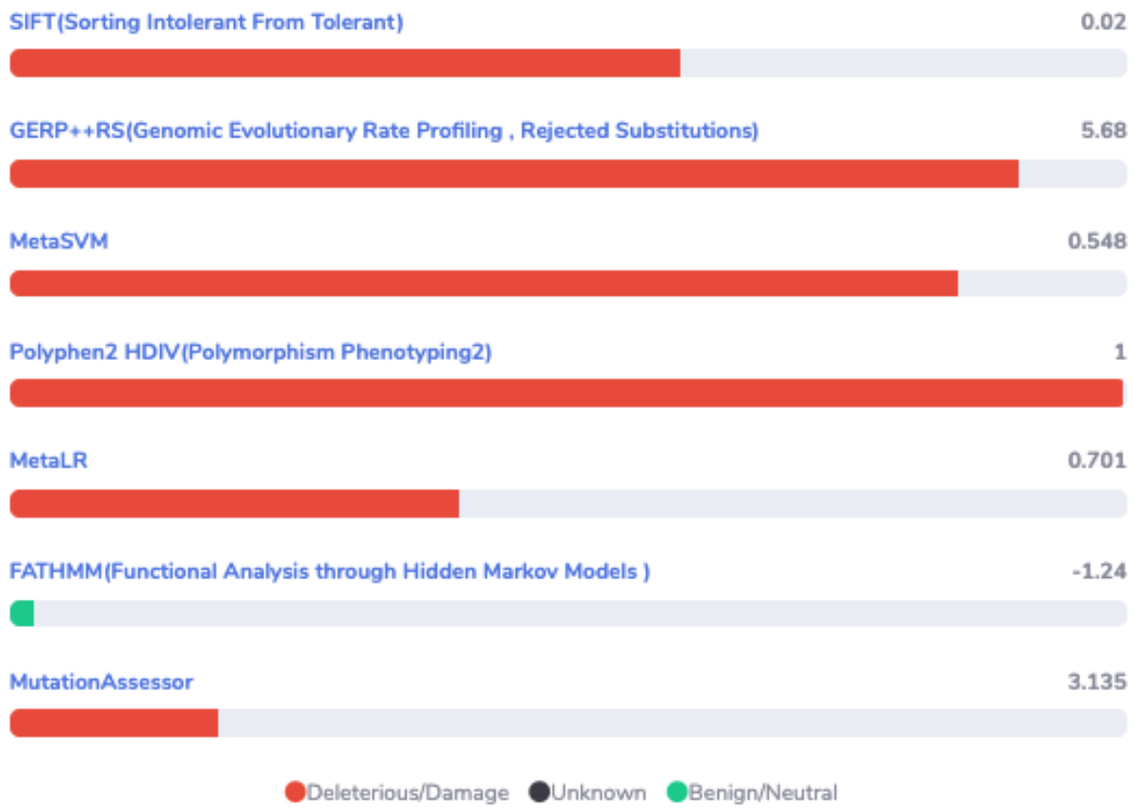
Domain

P-loop containing nucleoside triphosphate hydrolase;Small GTP-binding protein domain

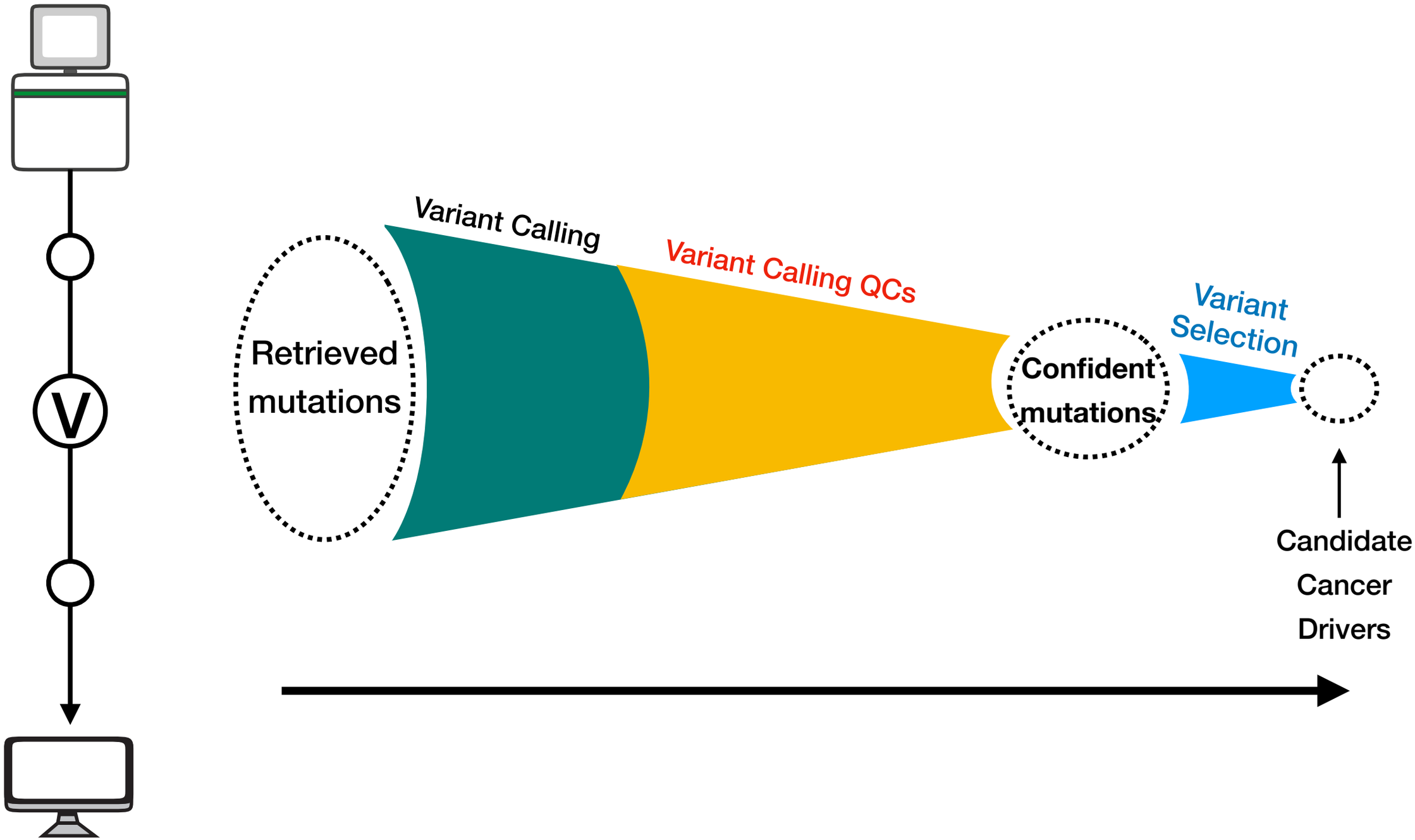
More resources

Gene on Oncokb
Variant on Oncokb
Clinvar
Cosmic

Deleterious Predictions from other softwares(. means no prediction)

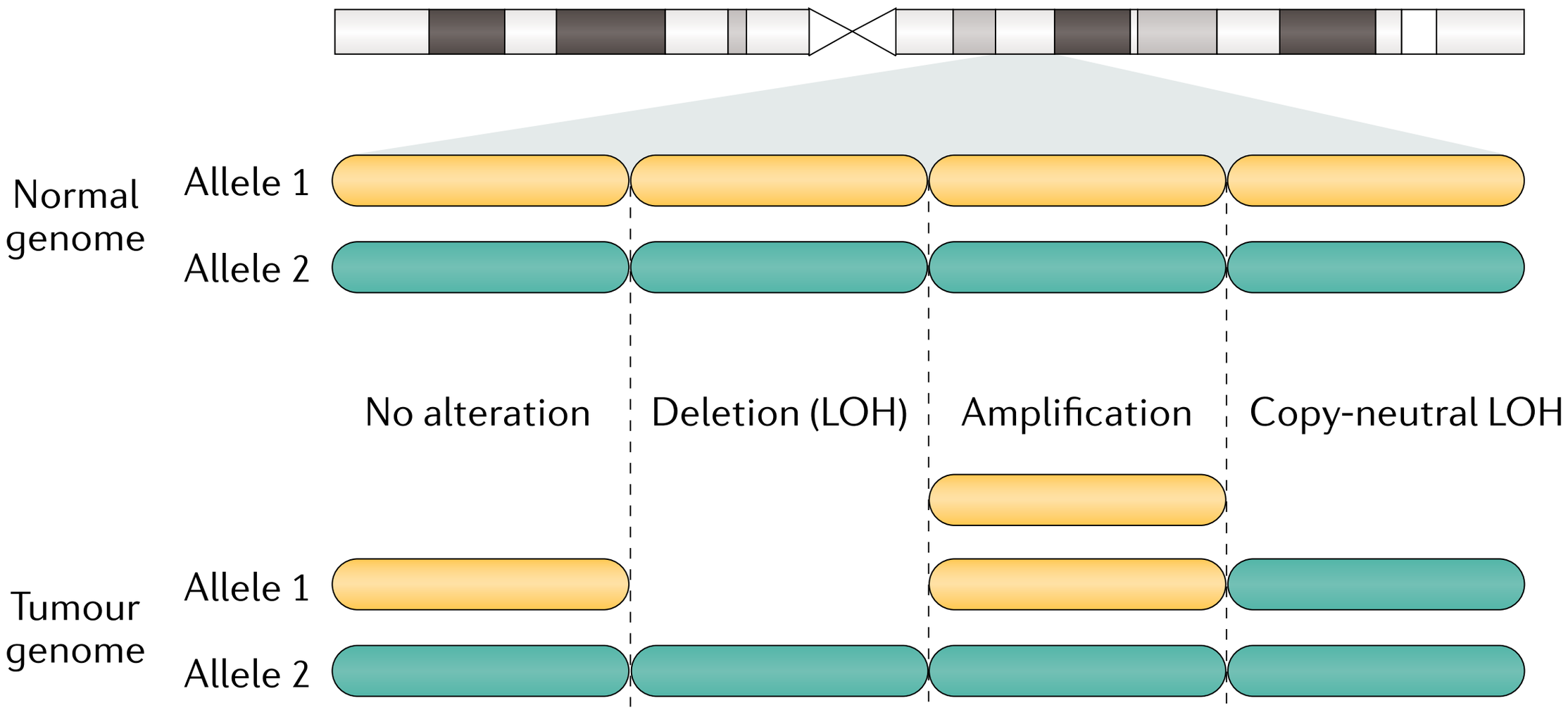


Candidate cancer driver

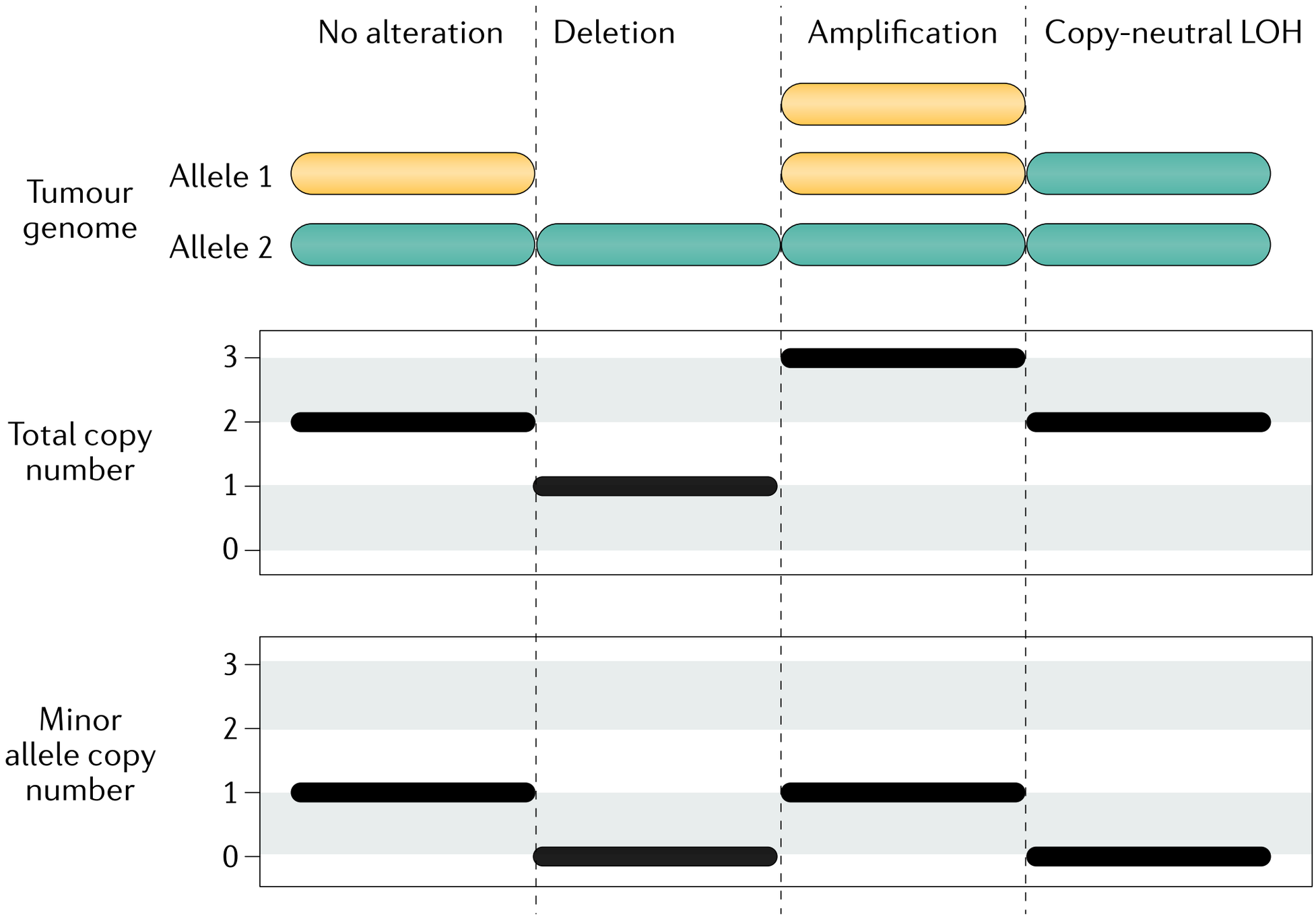


Copy Number Alterations

Allelic copies variation



Allelic copies variation



Genome CNV profile