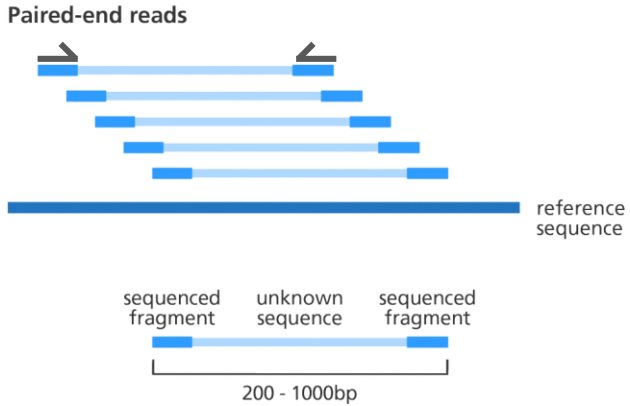
Basic concepts

The reads

Read = DNA fragment end

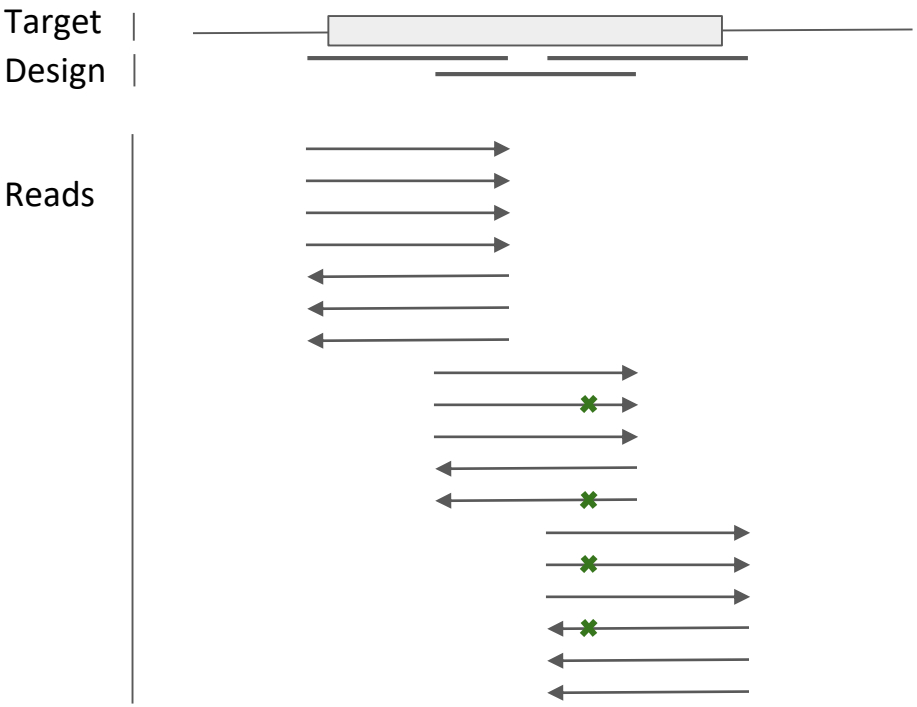
- Single-end
  Sequencing only 1 end

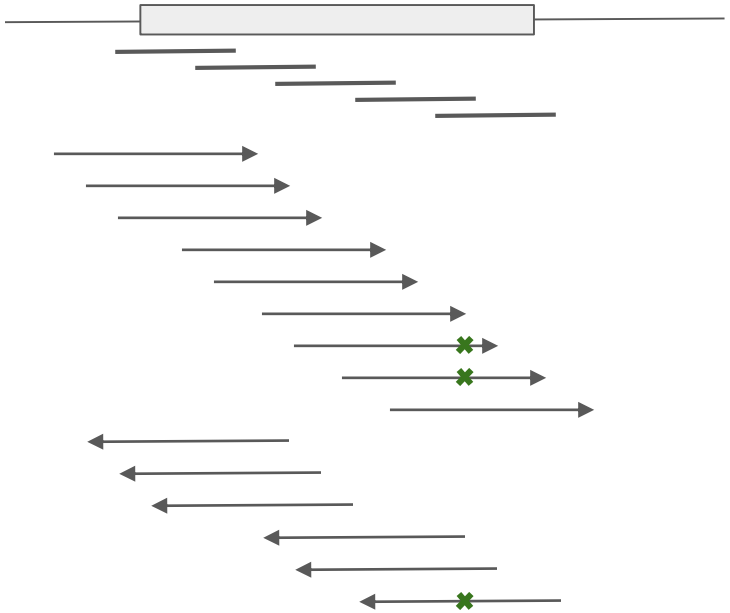
- Paired-end
  Sequencing both ends
  Reads orientation



Amplicon Library

Fragmented Library





Variants are misaligned bases relatively to a reference sequence !

There is not only one reference, the reference is still evolving

SPECIES	UCSC VERSION	RELEASE DATE	RELEASE NAME	STATUS
MAMMALS				
Human	hs1	Jan. 2022	T2T Consortium CHM13v2.0	Available
	hg38	Dec. 2013	Genome Reference Consortium GRCh38	Available
	hg19	Feb. 2009	Genome Reference Consortium GRCh37	Available
	hg18	Mar. 2006	NCBI Build 36.1	Available
	hg17	May 2004	NCBI Build 35	Available
	hg16	Jul. 2003	NCBI Build 34	Available
	hg15	Apr. 2003	NCBI Build 33	Archived
	hg13	Nov. 2002	NCBI Build 31	Archived
	hg12	Jun. 2002	NCBI Build 30	Archived
	hg11	Apr. 2002	NCBI Build 29	Archived (data only)
	hg10	Dec. 2001	NCBI Build 28	Archived (data only)
	hg8	Aug. 2001	UCSC-assembled	Archived (data only)
	hg7	Apr. 2001	UCSC-assembled	Archived (data only)
	hg6	Dec. 2000	UCSC-assembled	Archived (data only)
	hg5	Oct. 2000	UCSC-assembled	Archived (data only)
	hg4	Sep. 2000	UCSC-assembled	Archived (data only)
	hg3	Jul. 2000	UCSC-assembled	Archived (data only)
	hg2	Jun. 2000	UCSC-assembled	Archived (data only)
	hg1	May 2000	UCSC-assembled	Archived (data only)

Warning : all annotations are associated to only one given genome assembly : dbSNP, ClinVar, transcripts, UTR, variants from old data, ...

Human Genome Assembly from Ensembl

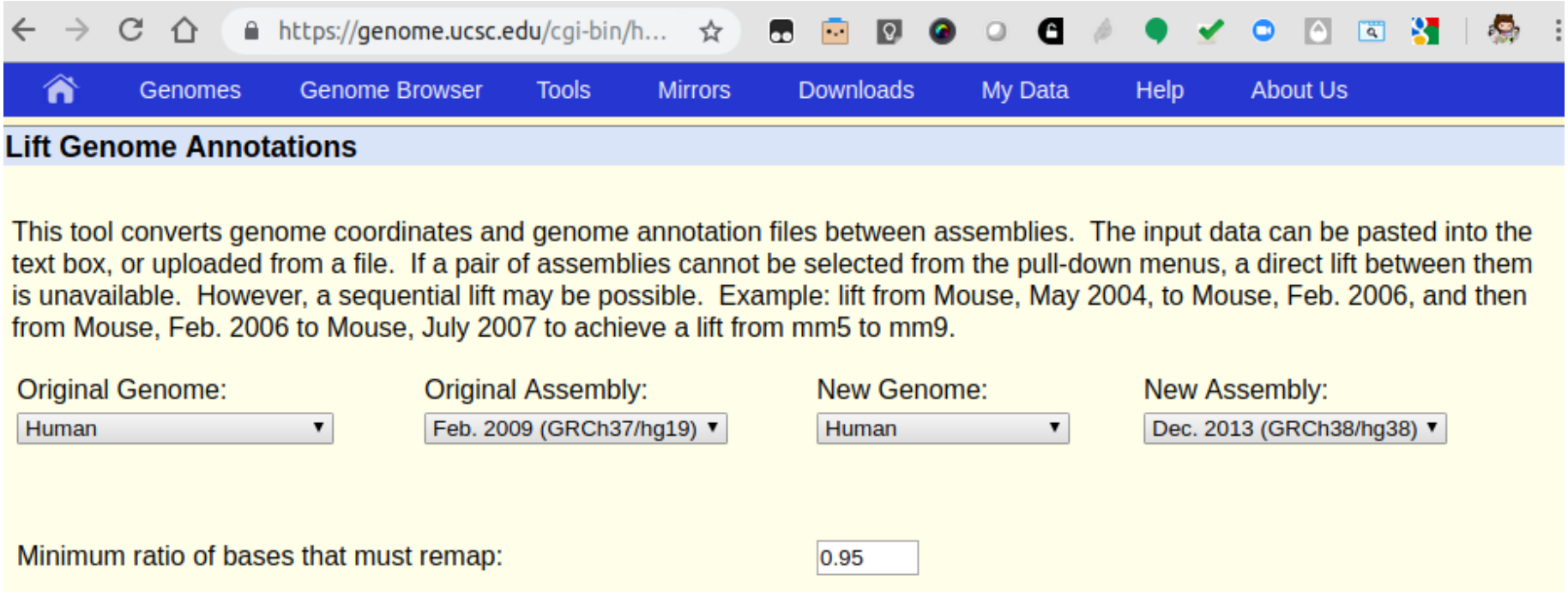
Assembly	GRCh38.p13 (Genome Reference Consortium Human Build 38), INSDC Assembly GCA_000001405.28, Dec 2013
Base Pairs	3,096,649,726
Golden Path Length	3,096,649,726
Annotation provider	Ensembl
Annotation method	Full genebuild
Genebuild started	Jan 2014
Genebuild released	Jul 2014
Genebuild last updated/patched	Aug 2020
Database version	103.38
Gencode version	GENCODE 37

13th patch

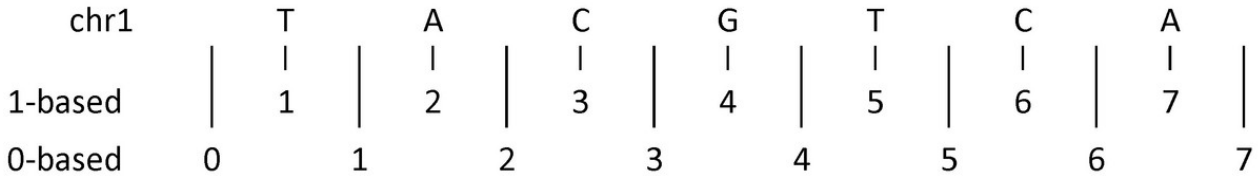
Annotation version (Release 103.38). Annotation are updated since the initial build of the genome (2013 for GRCh38) The annotation are (re)computed for each assembly and for each update in annotation (three months cycle)

Version of the GENCODE used for annotation (Ensembl annotation and HAVANA team manually corrected annotation)

You can convert old data to a new assembly and reciprocally



One-Based Vs Zero-Based Coordinate Systems

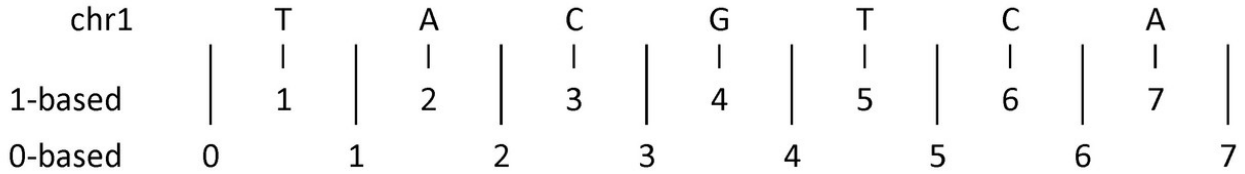


The example above shows (an imaginary) first seven nucleotides of sequence on chromosome 1:

- 1-based coordinate system
  - Numbers nucleotides directly
- 0-based coordinate system
  - Numbers between nucleotides

https://www.biostars.org/p/84686/

One-Based Vs Zero-Based Coordinate Systems



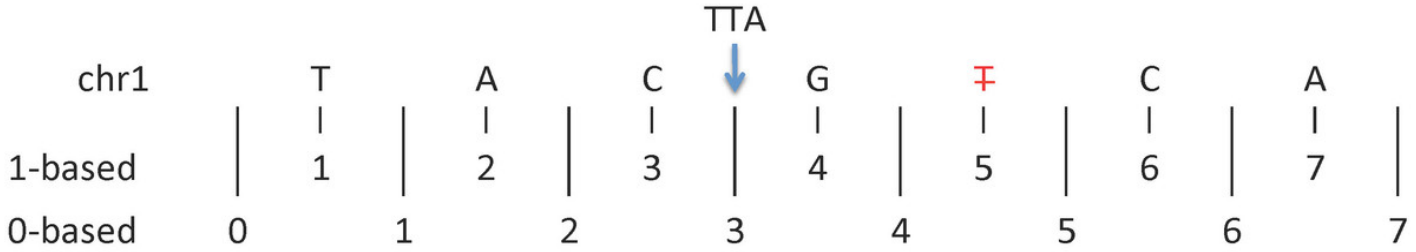
	1-based	0-based
Indicate a single nucleotide	chr1:4-4 G	chr1:3-4 G
Indicate a range of nucleotides	chr1:2-4 ACG	chr1:1-4 ACG
Indicate a single nucleotide variant	chr1:5-5 T/A	chr1:4-5 T/A

- 1-based coordinate system

  Single nucleotides, variant positions, or ranges are specified directly by their corresponding nucleotide numbers
- 0-based coordinate system

  Single nucleotides, variant positions, or ranges are specified by the coordinates that flank them

One-Based Vs Zero-Based Coordinate Systems



	1-based	0-based
Indicate a deletion	chr1:5-5 T/-	chr1:4-5 T/-
Indicate an insertion	chr1:3-4 -/TTA	chr1:3-3 -/TTA

	Deletions	Insertions
1-based coordinate	positions of the deleted bases	coordinates of the bases that flank the insertion
0-based coordinate	coordinates that flank the deleted bases	coordinate position where the insertion occurs

One-Based Vs Zero-Based Coordinate Systems

- Moving from UCSC browser/tools to Ensembl browser/tools or back
  - Ensembl uses 1-based coordinate system
  - UCSC uses 0-based coordinate system
- Some file formats are 1-based (GFF, SAM, VCF) and others are 0-based (BED, BAM)
- cheap length calculations : `m-n` (0-based) instead of `(m-n)+1` (1-based)

The BED Format

One line per feature, each containing 3-12 columns of data, plus optional track definition lines

Required fields

1. **chrom** - name of the chromosome or scaffold..
2. **chromStart** - Start position of the feature in standard chromosomal coordinates (i.e. first base is 0).
3. **chromEnd** - End position of the feature in standard chromosomal coordinates

```
chr1  213941196  213942363
chr1  213942363  213943530
chr1  213943530  213944697
chr2  158364697  158365864
chr2  158365864  158367031
chr3  127477031  127478198
```

Optional fields

4. **name** - Label to be displayed.
5. **score** - A score between 0 and 1000.
6. **strand** - defined as + (forward) or - (reverse).
7. **thickStart** - coordinate at which to start drawing the feature as a solid rectangle
8. **thickEnd** - coordinate at which to stop drawing the feature as a solid rectangle
9. **itemRgb** - an RGB colour value (e.g. 0,0,255).
10. **blockCount** - the number of sub-elements (e.g. exons) within the feature
11. **blockSizes** - the size of these sub-elements
12. **blockStarts** - the start coordinate of each sub-element

```
chr7  127471196  127472363  Pos1  0  +  127471196  127472363  255,0,0
chr7  127472363  127473530  Pos2  0  +  127472363  127473530  255,0,0
chr7  127473530  127474697  Pos3  0  +  127473530  127474697  255,0,0
chr7  127474697  127475864  Pos4  0  +  127474697  127475864  255,0,0
chr7  127475864  127477031  Neg1  0  -  127475864  127477031  0,0,255
chr7  127477031  127478198  Neg2  0  -  127477031  127478198  0,0,255
chr7  127478198  127479365  Neg3  0  -  127478198  127479365  0,0,255
chr7  127479365  127480532  Pos5  0  +  127479365  127480532  255,0,0
chr7  127480532  127481699  Neg4  0  -  127480532  127481699  0,0,255
```

http://genome.ucsc.edu/FAQ/FAQformat#format1

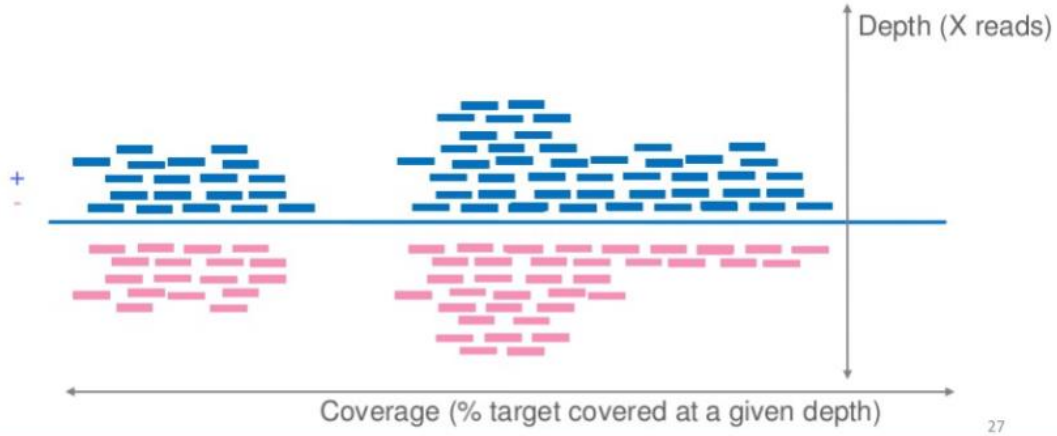
Coverage and Depth Of Coverage

- **Depth of coverage** = average number of reads covering a base (X)
  Example: 30X for normal sample, 100X for tumor sample

Profondeur

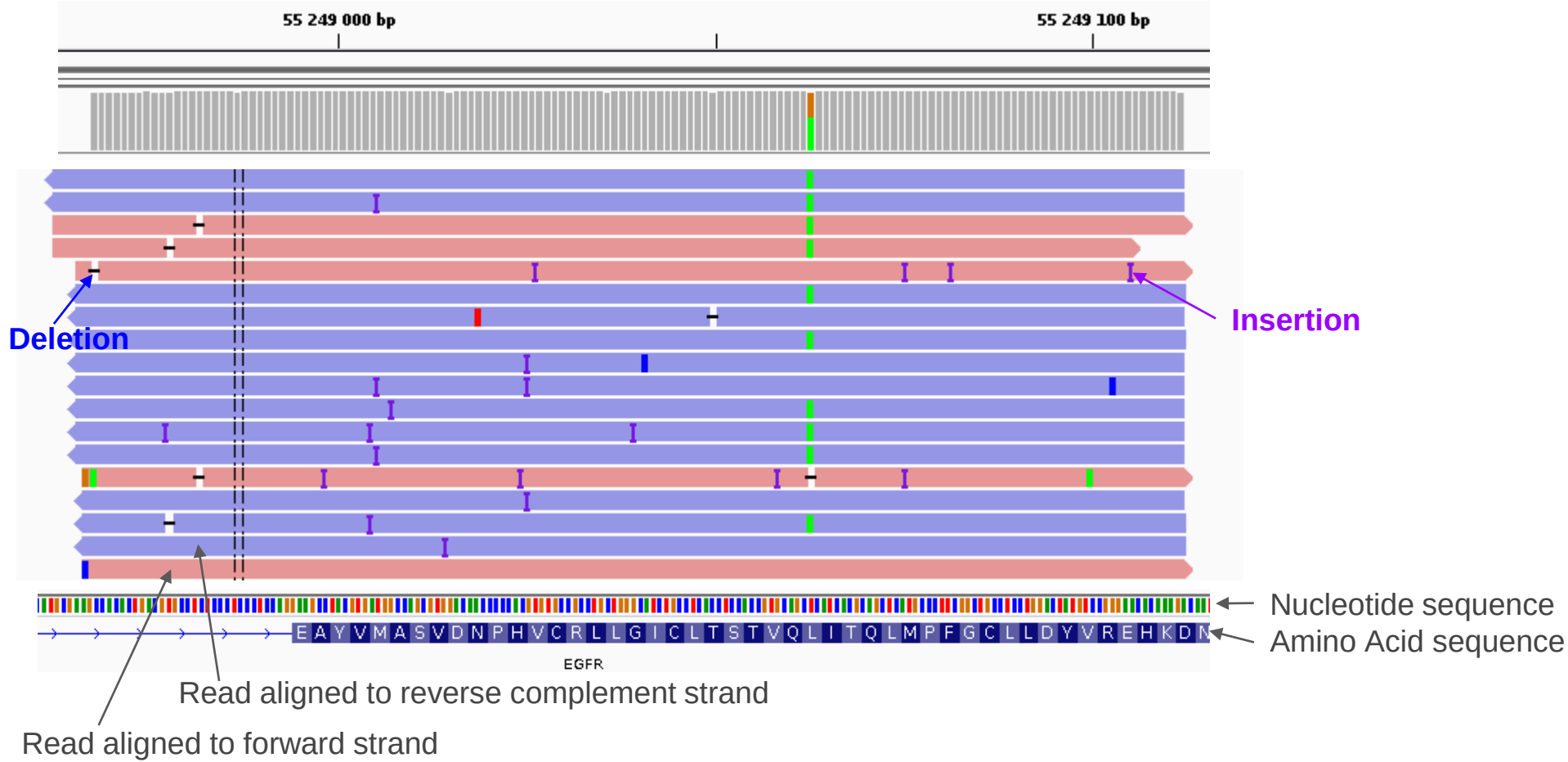
- **Coverage** = percentage of the targeted regions covered by at least X read
  Example: >=80% of your exome target is covered by 20X for normal sample

Couverture

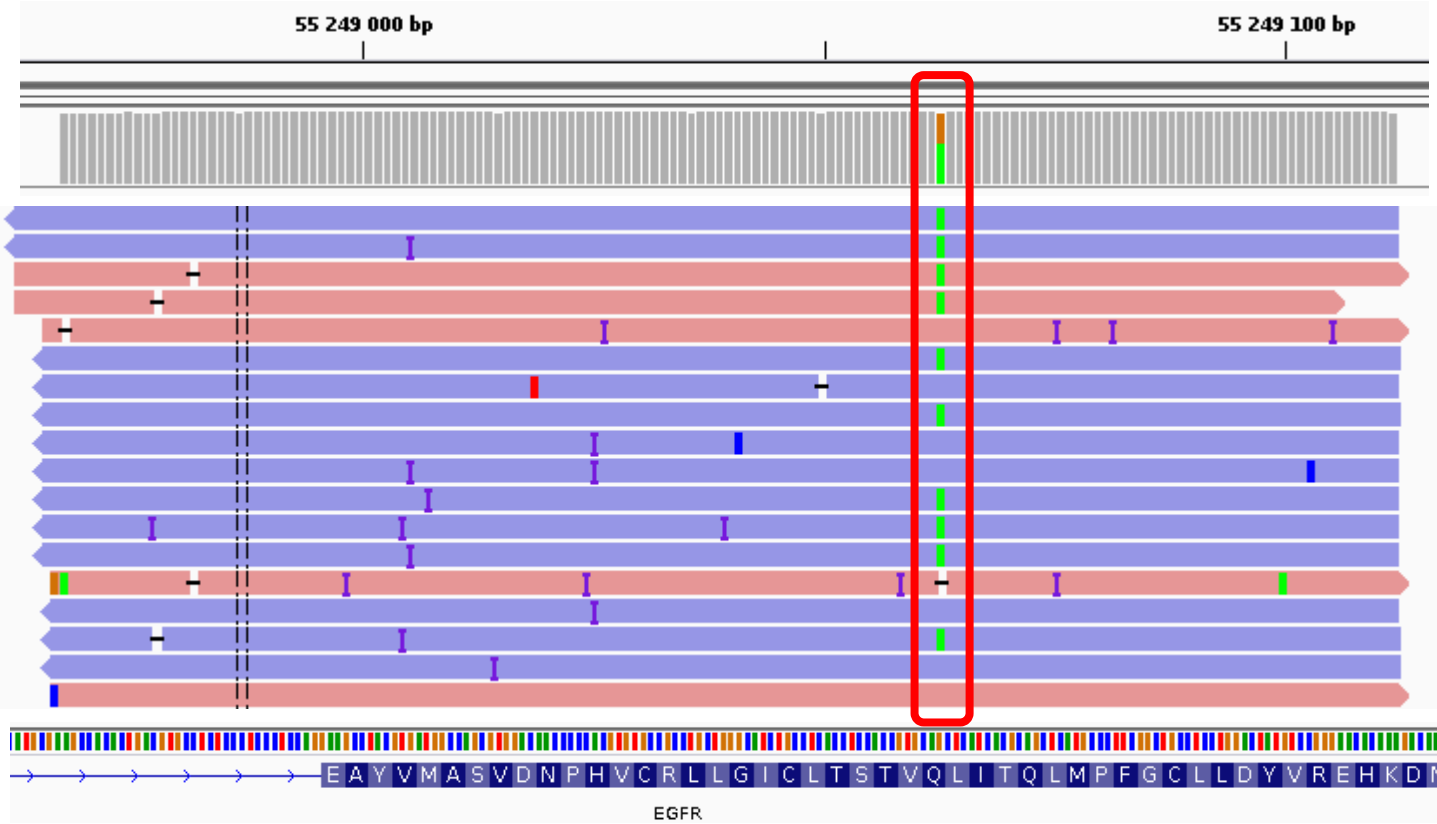


Source : Élodie Girard , 5ème Ecole de bioinformatique AVIESAN-IFB 2016
http://www.france-bioinformatique.fr/sites/default/files/V01_ITMO_2016_EG_from_fastq_to_mapping_1.pdf

Variant Calling Sketch on real data : Integrative Genomic Viewer (IGV)



Variant Calling Sketch on real data : Integrative Genomic Viewer (IGV)



Many sequencing or PCR errors.

One variant :
8/13 reverse reads
2/4 forward reads

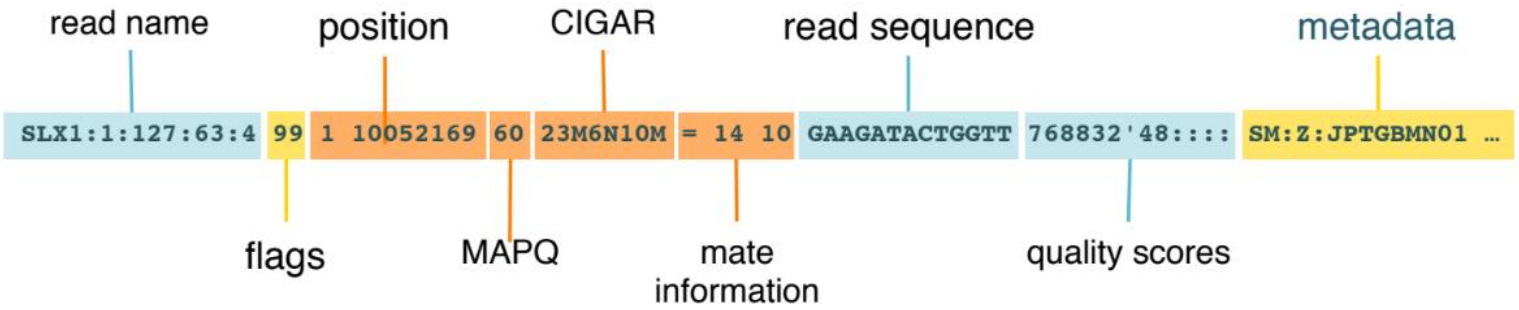
Variant Allele Frequency (VAF) = (8+2)/(13+4) = 0.58

Nucleotide sequence

Amino Acid sequence

Output format: Sequence/Binary Alignment Map (SAM/BAM)

HEADER containing metadata (sequence dictionary, read group definitions etc)
RECORDS containing structured read information (1 line per read record)



- Added mapping info summarizes **position**, **quality**, and **structure** for each **read**

http://samtools.github.io/hts-specs/SAMv1.pdf

From GATK Best Practices for Variant Discovery Presentation