Introduction to high-throughput sequencing and its application in high-throughput sequencing and single-nucleotide variant analysis in cancer

Cancer genomics and transcriptomics course, 2023

Francesc Muyas Remolar, PhD
EMBL-EBI
10 – 07 - 2023

EMBL-EBI

Outline

1. Introduction
2. Somatic variant calling
3. Workflow for cancer genome analysis
4. Library preparation and sequencing
5. Fastq files
6. Alignment
7. Variant detection
8. Variant annotation

What do you know about these concepts?

- Next Generation Sequencing (NGS)
- Paired-read sequencing
- Fastq file
- BAM/SAM file
- VCF file
- Variant Allele Frequency (VAF)
- Depth of coverage
- Variant annotation

Importance of understanding the somatic landscape in cancer

Comprehensive Mutational Landscape

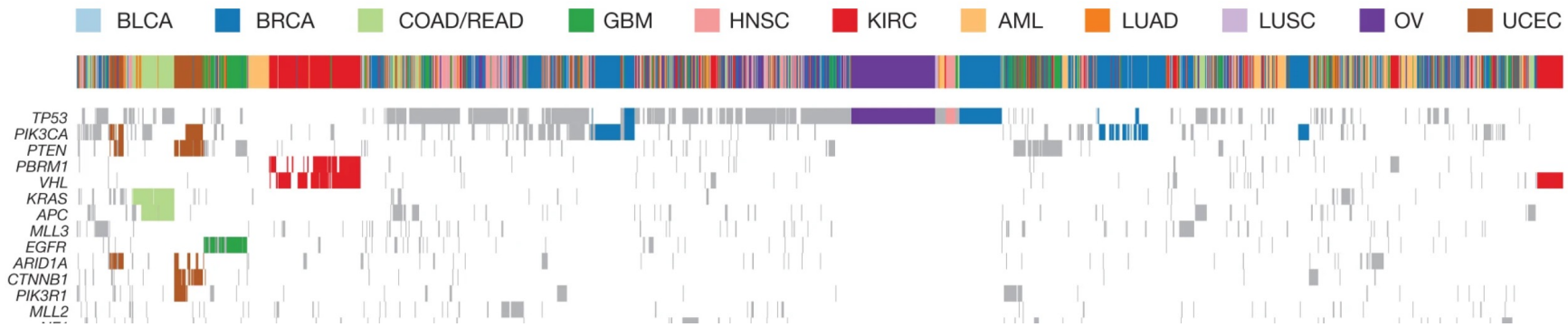
High-throughput sequencing technologies enable the analysis of cancer genomes, revealing a comprehensive view of mutational events (SNVs, SVs, CNAs…)

Research and Clinical Applications

Exome sequencing (WES) and whole-genome sequencing (WGS) are widely used in research to characterize genomic alterations

Precise and Informed Interventions

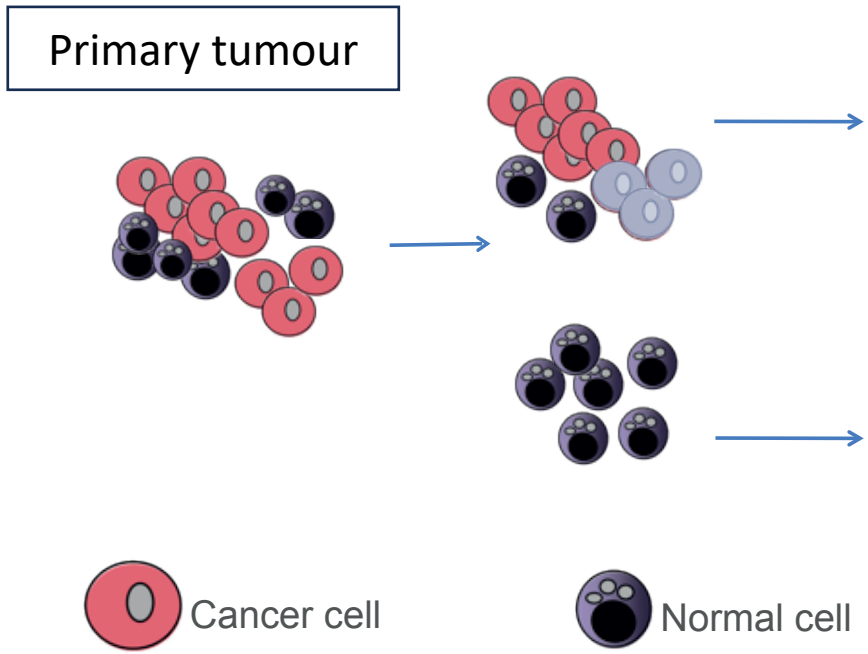
The information derived from sequencing cancer genomes allows for more precise interventions, including personalized treatment strategies and the identification of potential therapeutic targets, ultimately improving patient outcomes.

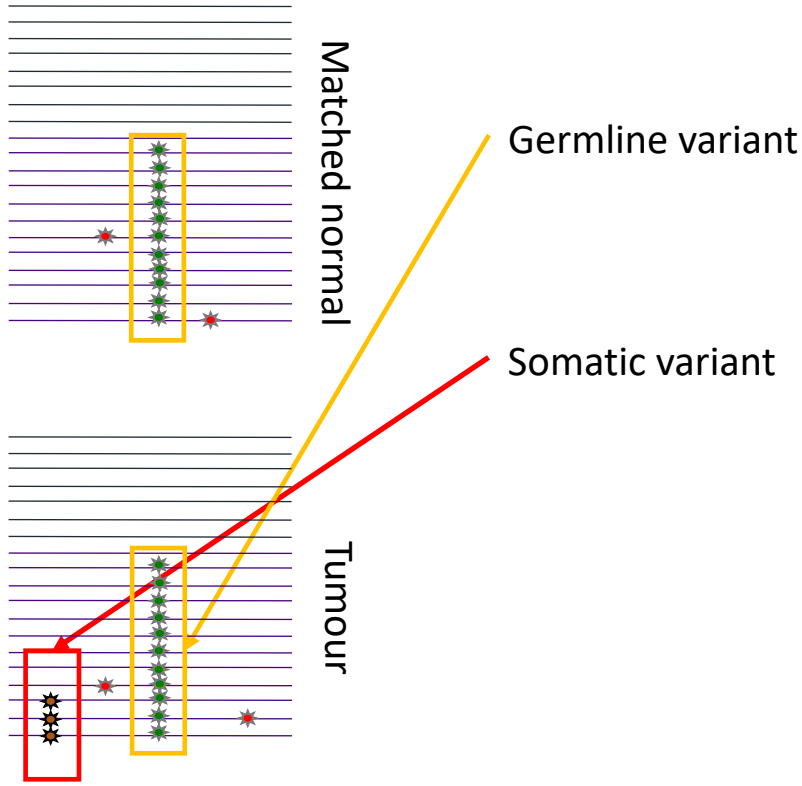


Kandoth *et al* 2013

Somatic variant calling

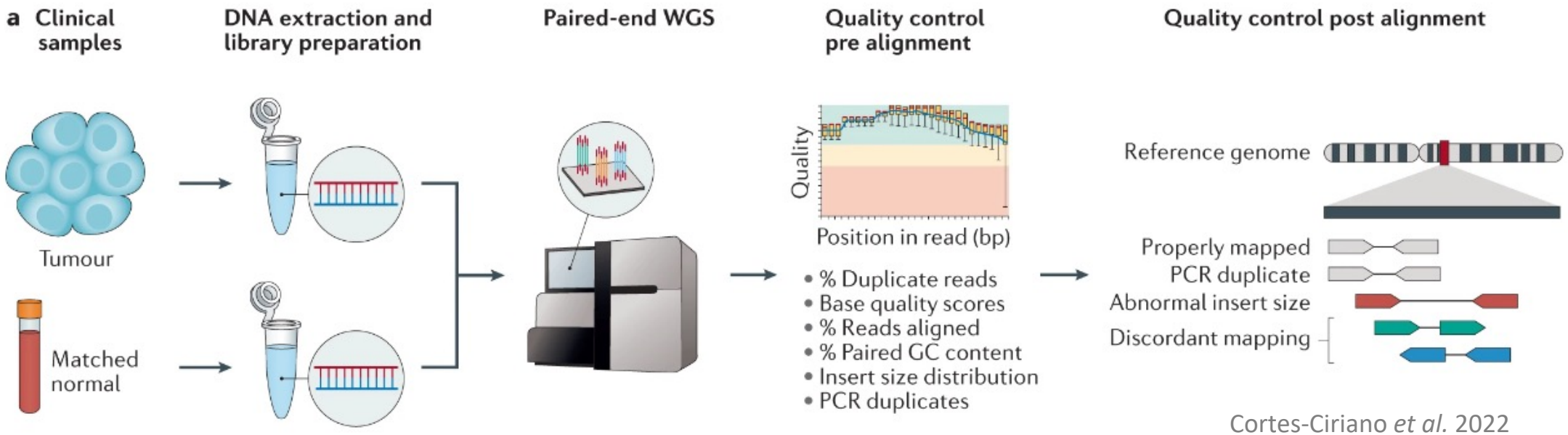
- Somatic variant calling typically involves sequencing both tumour and matched normal samples from an individual.
- The normal sample serves as a control, representing the non-mutated or germline cells, while the tumour sample contains genetic alterations specific to the tumour cells.





Workflow for cancer genome analysis

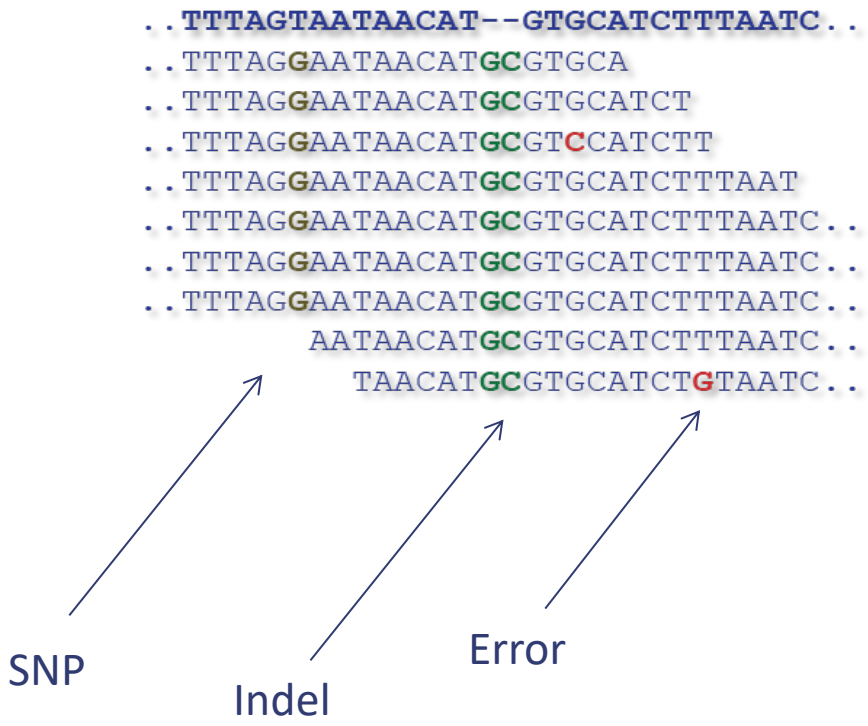
- Detection of somatic mutations in human cancers involves the extraction of DNA from a tumour and, ideally, a matched normal sample.
- Sequencing is usually performed in the paired-end mode with read lengths of 100–150 bp.
- Quality control is performed on the sequencing reads by assessing several metrics before and after alignment to the reference genome.



Cortes-Ciriano *et al.* 2022

Workflow for cancer genome analysis

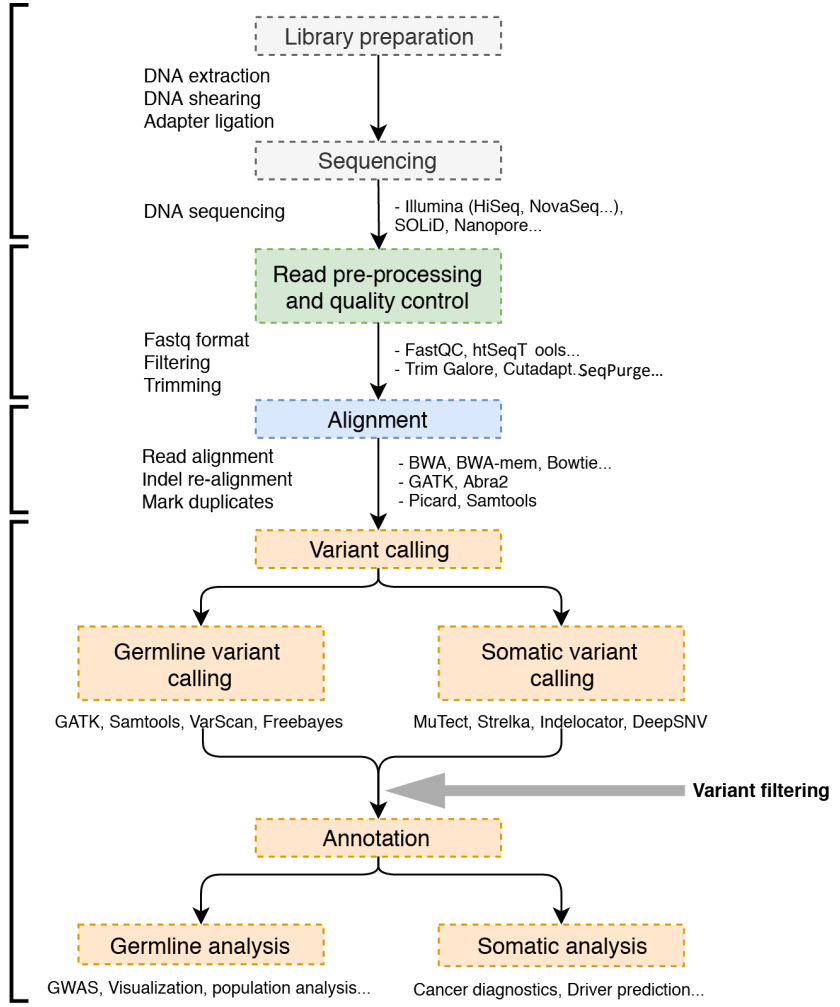
Detecting SNVs and Indels



Library preparation and sequencing

Read pre-processing and alignment

Somatic/Germline variant calling SNPs, Indels



Library preparation

Library preparation is a critical step in Next-Generation Sequencing (NGS) that involves the conversion of DNA or RNA samples into a format suitable for sequencing.

1. **Fragmentation**: The DNA or RNA sample is fragmented into smaller pieces to create a more manageable size range for sequencing (e.g., sonication, enzymathic methods)
2. **End Repair and A-Tailing**: The fragmented DNA ends are repaired to generate blunt ends, or in some cases, overhangs with specific adapters
3. **Adapter Ligation**: Short, synthetic DNA adapters containing sequences complementary to the sequencing platform are ligated to the repaired DNA fragments
4. **Amplification**: The ligated DNA fragments, now with adapters attached, are amplified using polymerase chain reaction (PCR)
5. **Quality Control**: The resulting DNA library is evaluated to assess its quality and quantity

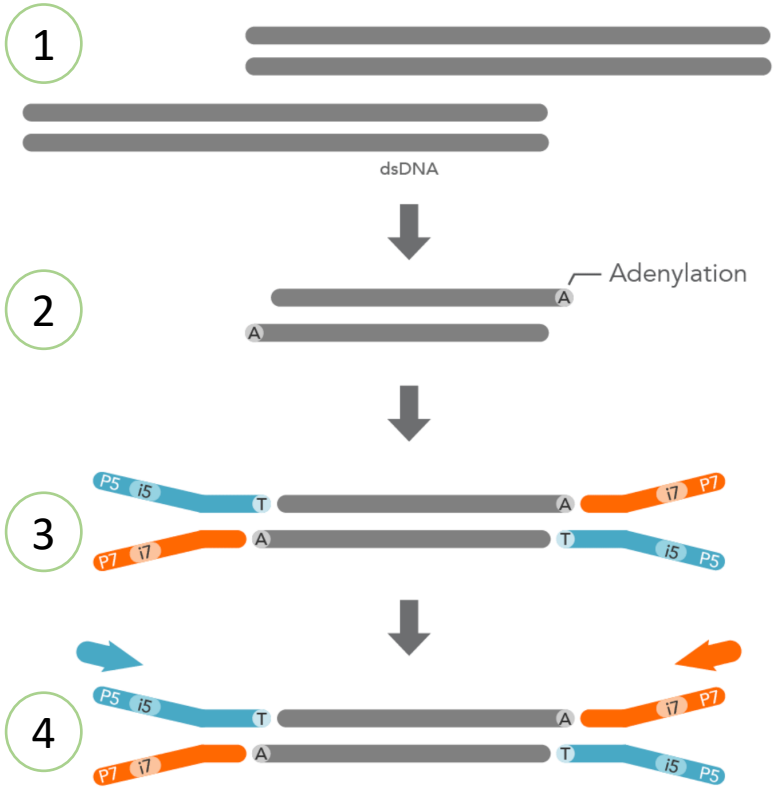
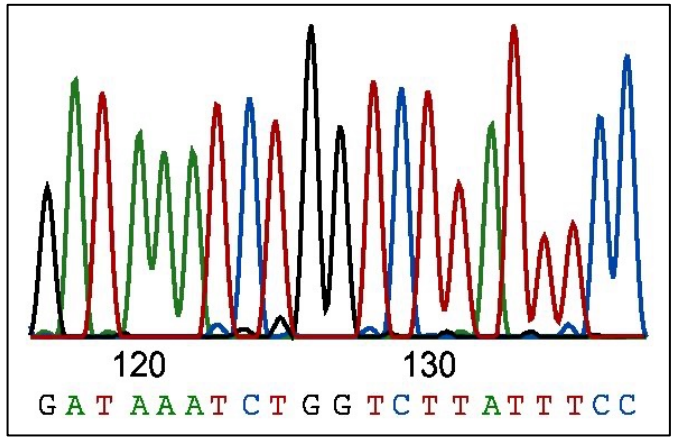


Figure: eu.idtdna.com

Sequencing technologies

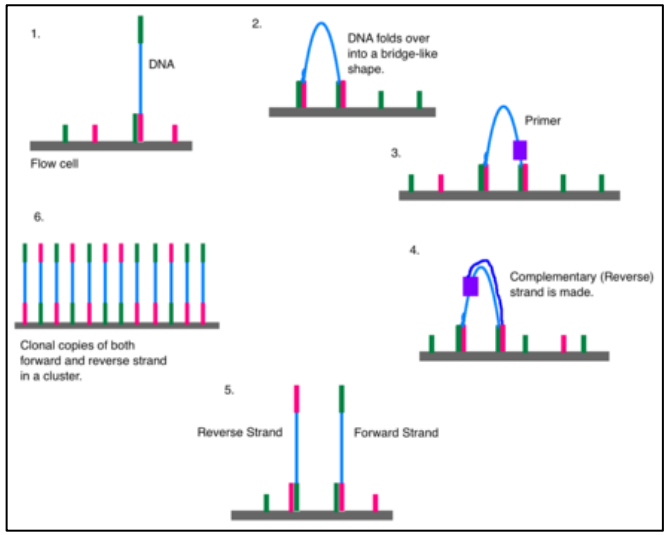
Sanger Sequencing

- Known as the chain termination method (first widely adopted sequencing technology)
- It relies on enzymatic replication of DNA and incorporates chain-terminating dideoxynucleotides labelled with fluorescent dyes.
- The resulting fragments are separated by capillary electrophoresis, allowing the determination of the DNA sequence.
- Often used for targeted sequencing and validation of genetic variants.



Next-Generation Sequencing (NGS)

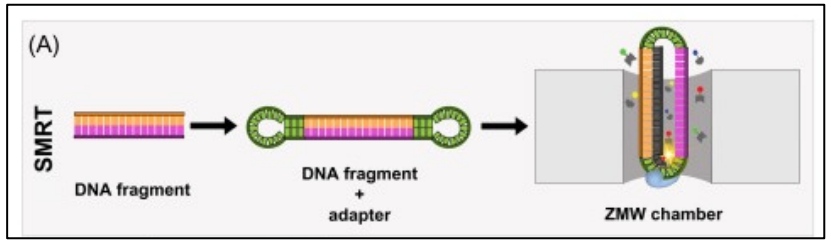
- Encompasses a group of high-throughput sequencing technologies that revolutionized genomic research.
- Illumina sequencing (most used) → massively parallel sequencing by synthesis
- They generate millions to billions of short DNA fragments simultaneously (low error rate)
- NGS offers high-throughput and cost-effective sequencing



Sequencing technologies

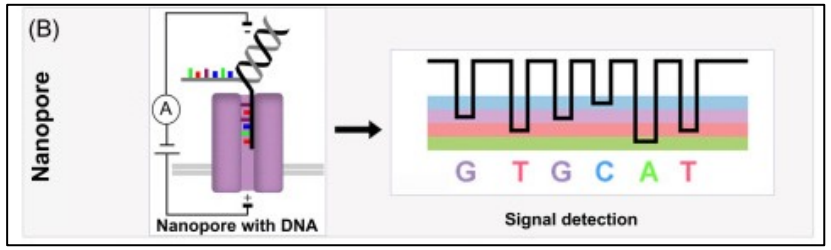
PacBio Single Molecule Real-Time Sequencing (SMRT)

- Employs a different technology known as single-molecule real-time sequencing.
- Uses specialized DNA polymerases and detects the incorporation of fluorescently labelled nucleotides in real-time as the DNA is being synthesized.
- Enables long-read sequencing → Useful for resolving complex genomic regions and structural variants



Oxford Nanopore Sequencing

- Portable and real-time sequencing technology.
- It involves passing DNA strands through nanopores, and the changes in electrical current as the DNA molecule passes through the pore are recorded.
- The current signal is then decoded to infer the DNA sequence.
- Nanopore sequencing offers long-read capabilities, allowing the sequencing of long DNA fragments without the need for extensive library preparation.
- It has applications in genomics & metagenomics



Figures: Pavlovic et al 2020

Sequencing technologies

Summary of the sequencing technologies

	Next generation						Third generation*	
	454	Illumina				Ion Torrent	PacBio	Oxford nanopore
Platform	GS FLX+	HiSeq 2500	MiSeq	NextSeq 500	NovaSeq S4	PGM 318	Sequel II System	MinION R9.4
Run time	~ 24 h	~ 6 days	2-3 days	12-30 h	~ 44 h	4-7 h	~ 30 h	24 h**
Output / run	700 Mb	1 Tb	15 Gb	120 Gb	6 Tb	2 Gb	300 Gb	50 Gb
Read length	1 kb	2 x 125 bp	2 x 300 bp	2 x 150 bp	2 x 150 bp	400 bp	10 -16 kb (average)	18 kb**
Error rate	~ 1 %	~ 0.1 %				~ 1 %	~ 14 %	~ 10 % **
Primary errors	Indels	SNVs				Indels	Indels	Indels
Advantages	- Long reads - Relative fast run time	- Highest throughput of all platforms and lowest per-base cost. - Low per-base cost - Low error rate				- Unmodified nucleotides - No optical scanning necessary - Fast run time	- Very long reads - Does not require PCR amplification before sequencing - No bias based on GC content	- Very long reads - Does not require PCR amplification before sequencing - No bias based on GC content -Portable and easy use
Limitations	- High error rate in homopolymers. - Low throughput - High cost - Cumbersome emPCR	-Short reads - Overloading results in overlapping clusters and poor sequence quality - Requirement for sequence complexity. Problems in low-complex regions - Bias in coverage correlated with GC content.				- High error rate in homopolymers - Cumbersome emPCR	- High cost - High error rates	- High error rates, specially in homopolymers and low-complexity regions

*Many updates in the last years

Adapted from Pfeifer et al 2017 and Ardui et al 2018

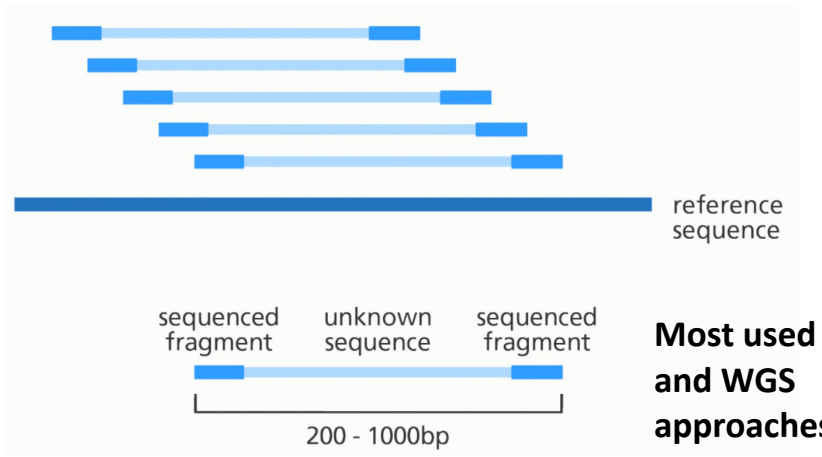
Single-end and paired-end sequencing

During short read sequencing using NGS platforms such as Illumina, there are two sequencing strategies available for implementation.

Single-end reads

- Single-read sequencing involves sequencing DNA from only one end → simplest way to utilize Illumina sequencing
- **Cost-effective uses:** This solution delivers large volumes of high-quality data, rapidly and economically
- **Specific applications:** good choice for certain methods like small RNA-Seq or ChIP-Seq

Paired-end reads



Most used in WES and WGS approaches

- Paired-end sequencing allows to sequence both ends of a fragment and generate high-quality, alignable sequence data.
- Paired-end DNA sequencing also detects common DNA rearrangements such as insertions, deletions, and inversions

Figures: https://www.yourgenome.org/

Limitations at different steps of the variant sequencing workflow

Each step of this workflow has particular limitations and issues, which might affect the downstream analysis and interpretation of the calls

Library preparation and sequencing

- **Oxidative damage during acoustic shearing** → C to A, G to T
- **PCR error rate** *Taq Polimerase* → $1\text{-}20 \times 10^{-5}$
- **Sequencing errors**
  - Crosstalk: A to C | G to T
  - Dephasing: Indels
  - Low coverage: GC content

Read pre-processing and alignment

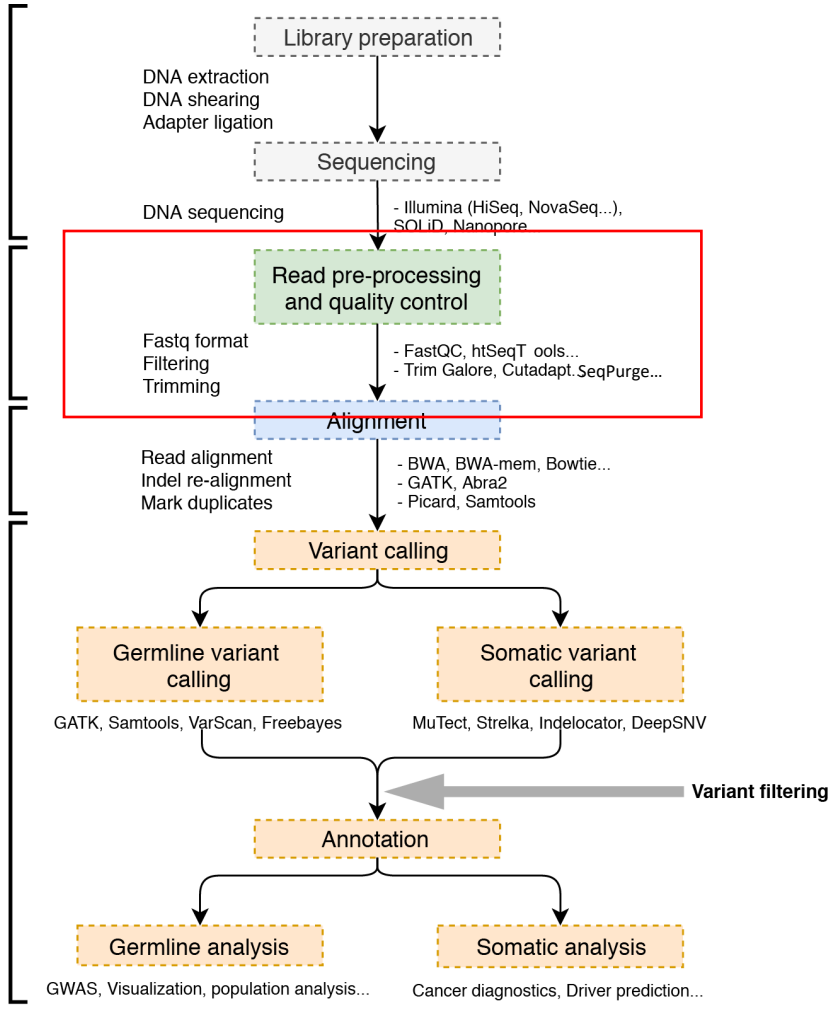
Somatic/Germline variant calling SNPs, Indels, SVs, CNVs

Read processing

Library preparation and sequencing

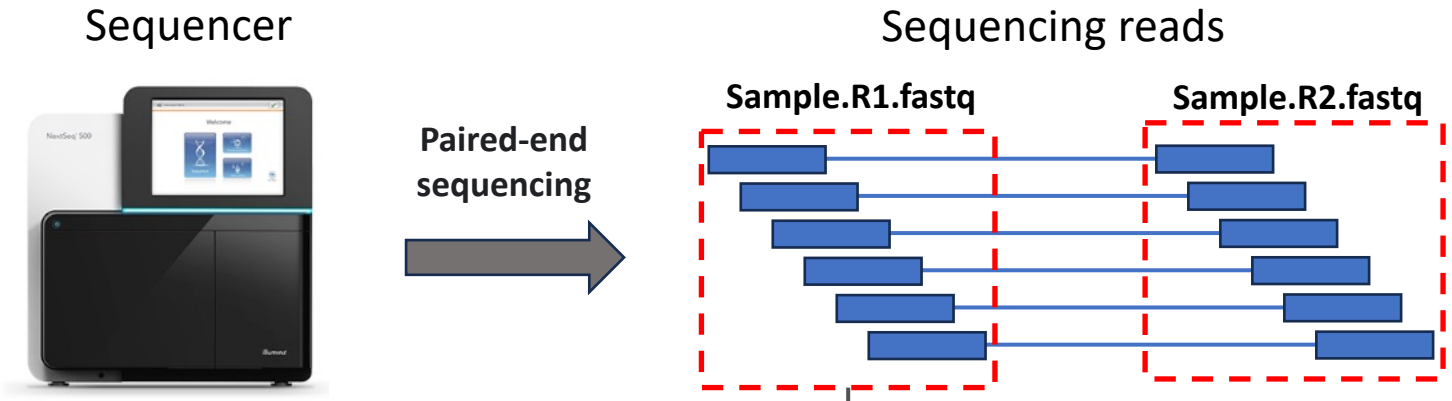
Read pre-processing and alignment

Somatic/Germline variant calling SNPs, Indels



Fastq files

- A FASTQ file is a commonly used file format in bioinformatics for storing DNA or RNA sequencing data
- It contains both the sequence information and corresponding quality scores for each nucleotide in the sequence.



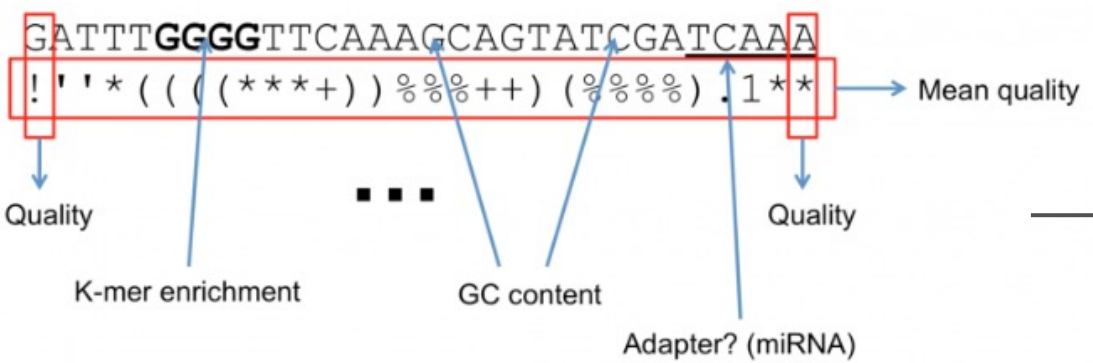
How does a fastq file look like?

Read name	@ST-E00287:121:HCT3FCCXY:3:2114:22972:45083
Read sequence	AAATCAGCGTTGTTCCTACATGATAATTCGGTCCACAAACTCACTTCAGTTTTA
Sequencing direction	+
Base qualities	?BBA@???8@@>?@>?A?>??>A>@@?A>7?@?@?>@@A>A?@>AA?@?>AAA@

Each sample will generate (depending on the read length) ~ 5Gb (100X WES) to ~ 300 Gb (100X WGS) of data

Fastq files

Base quality (Phread Quality Score): The quality value (*Q*) is a representation of the probability (*p*) that a base calls is incorrect



$$Q = -10 \log_{10} P \quad \Rightarrow \quad P = 10^{\frac{-Q}{10}}$$

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%

ascii encoded

```
>>> ord("A")-33
32
```

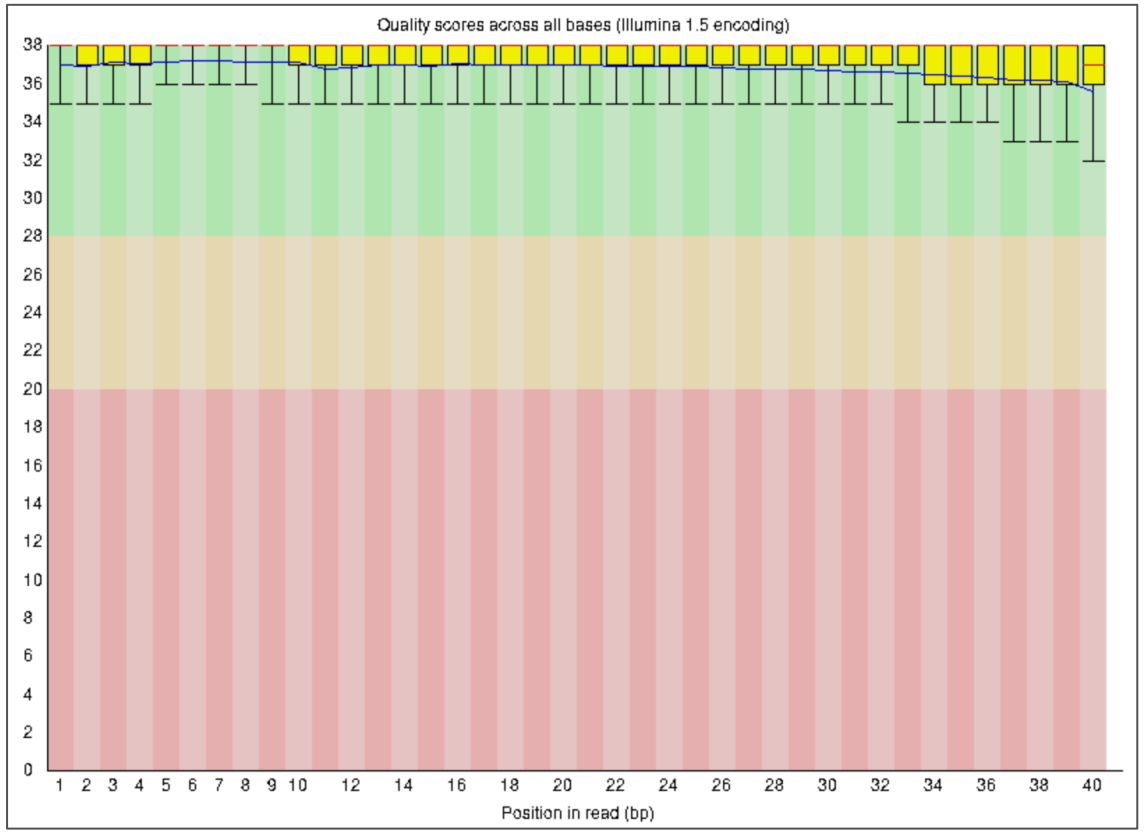
Example in python to transform letter quality to numeric quality value

Figures: https://learn.gencore.bio.nyu.edu/

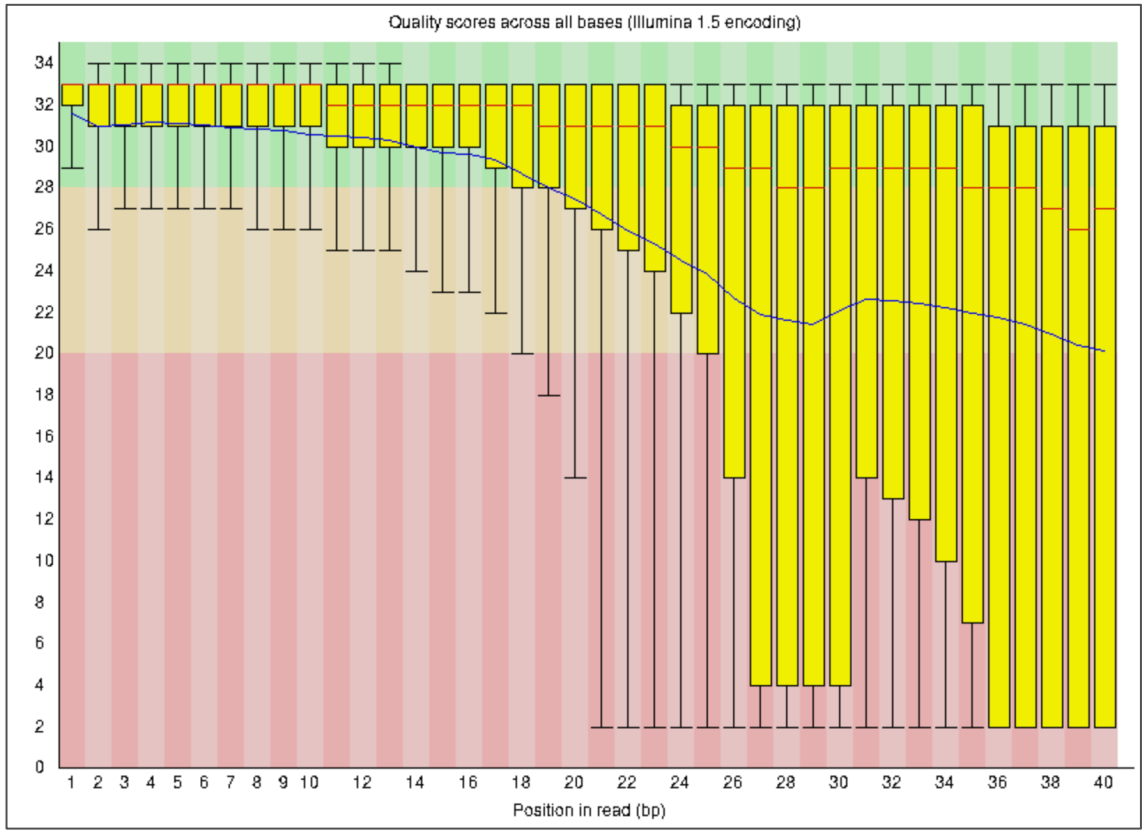
QC of raw sequences

- Low quality bases or low-quality reads can bias all the downstream analysis (alignment, SNV and indel calling…)
- This step can be performed with the **FastQC software**

Good quality illumina data



Bad quality illumina data

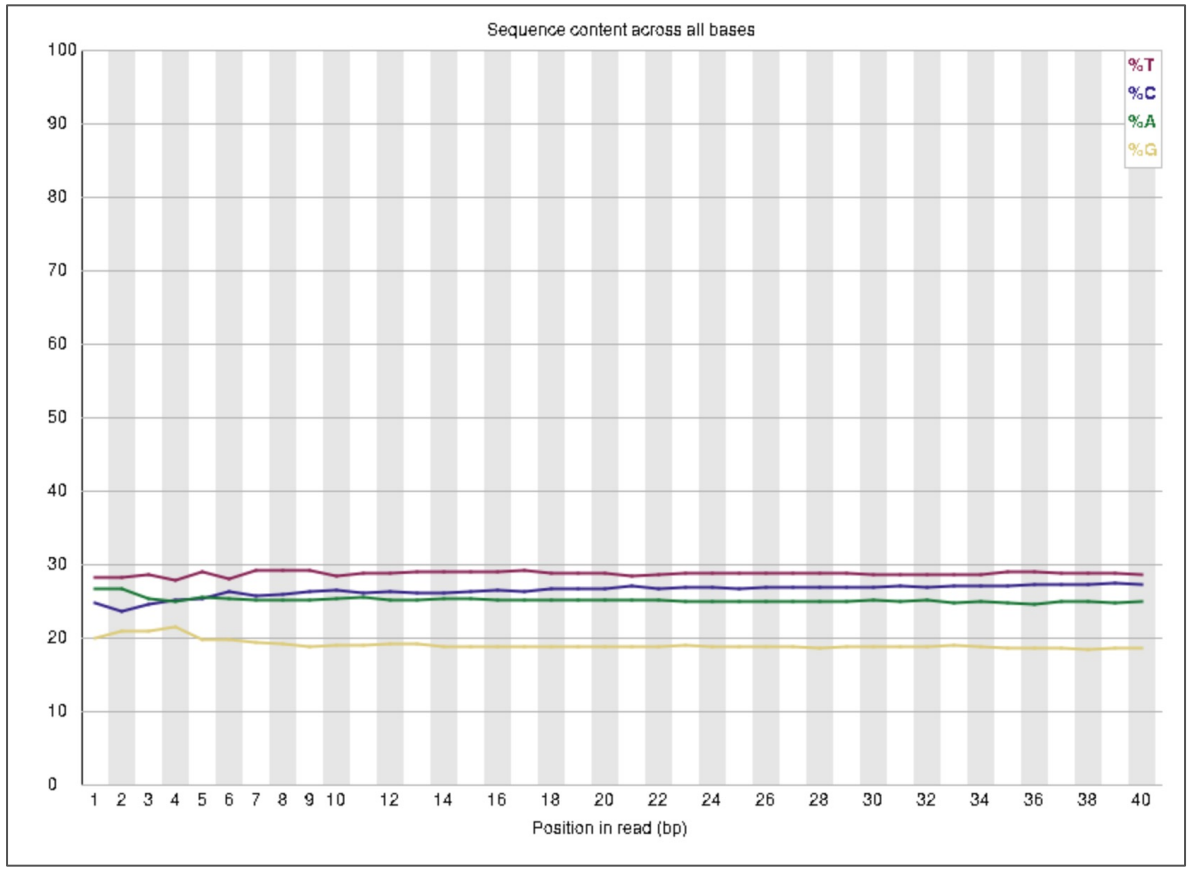


Figures: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

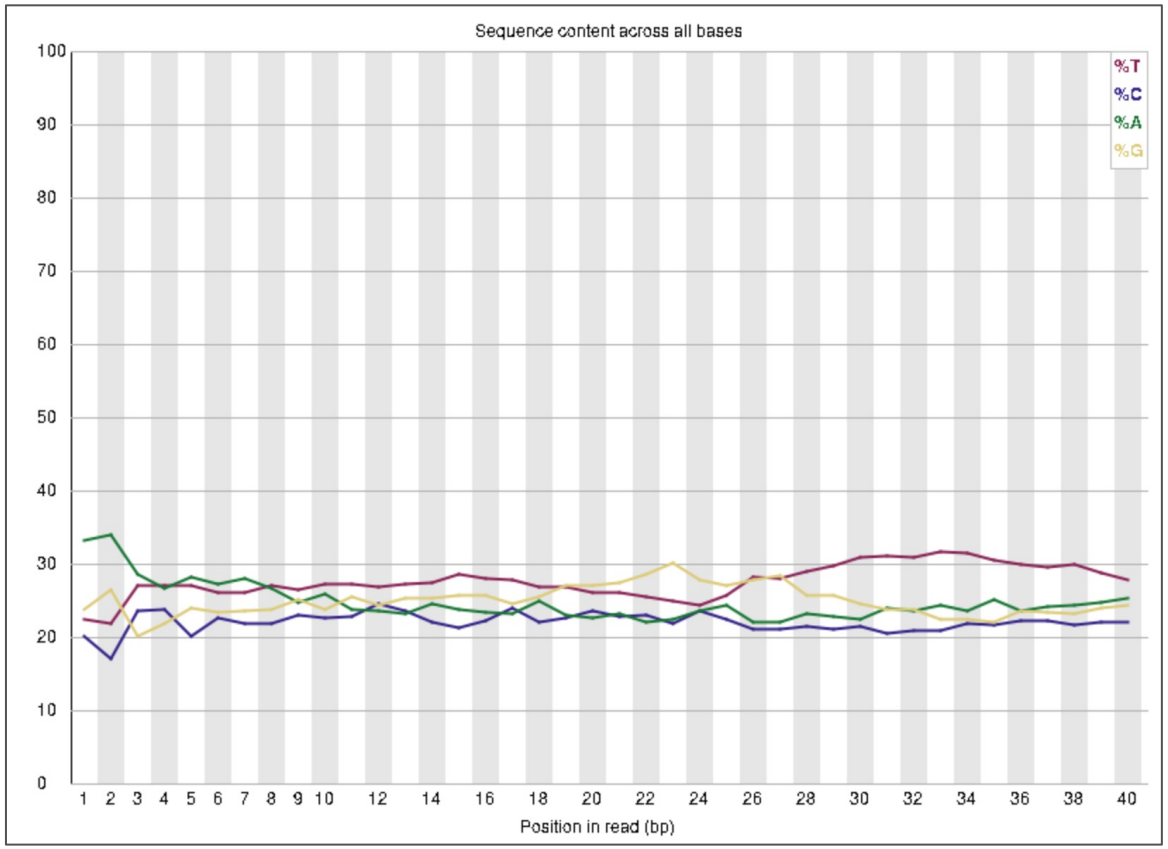
QC of raw sequences

Position base-content is usually consistent and uniform across the read

Good quality illumina data



Bad quality illumina data



Figures: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

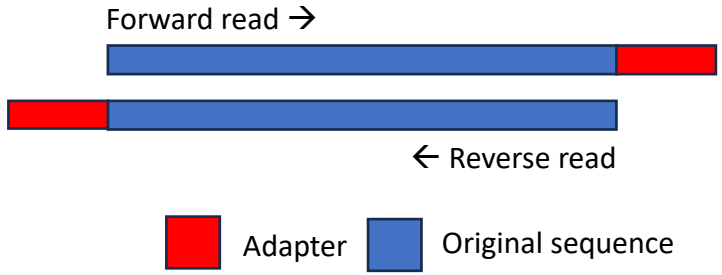
QC of raw sequences

Some samples have Illumina adapter contamination

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCT	8122	8.122	Illumina Paired End PCR Primer 2 (100% over 40bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGATCGGAAG	5086	5.086	Illumina Paired End PCR Primer 2 (97% over 36bp)
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTAC	1085	1.085	Illumina Single End PCR Primer 1 (100% over 40bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGGAAG	508	0.508	Illumina Paired End Sequencing Primer 2 (100% over 36bp)
AATTATACGGCGACCACCGAGATCTACACTCTTTCCCTAC	242	0.242	Illumina Single End PCR Primer 1 (97% over 40bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAAGATCGGAA	235	0.23500000000000001	Illumina Paired End Adapter 2 (96% over 31bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCGAGATCGGAAGA	228	0.22799999999999998	Illumina Paired End Adapter 2 (96% over 28bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGGACG	205	0.20500000000000002	Illumina Paired End Sequencing Primer 2 (100% over 36bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGGTCGGAAG	183	0.183	Illumina Paired End Adapter 2 (100% over 32bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGGATCGGAA	183	0.183	Illumina Paired End Adapter 2 (100% over 32bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGAACT	164	0.164	Illumina Paired End PCR Primer 2 (97% over 40bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGGTCT	129	0.129	Illumina Paired End PCR Primer 2 (97% over 40bp)
AATTATACTTCTACCACCTATATCTACACTCTTTCCCTAC	123	0.123	No Hit
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGGACT	122	0.122	Illumina Paired End Sequencing Primer 2 (100% over 36bp)
CGGTTCAGCAGGAATGCCGAGATCGGAAGAGCGGTTCAGC	113	0.11299999999999999	Illumina Paired End PCR Primer 2 (96% over 25bp)

These regions need to be trimmed before the alignment



Many tools available that can do this trimming:

- Cutadapt
- SeqPurge
- Trimmomatic
- …

Figures: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Alignment

Read mapping, also known as alignment or mapping, is a process in computational genomics that involves aligning or matching short DNA/RNA sequences (reads) to a reference genome (in our case, the human reference genome **Hg38**)

Mapping problems

- Mapping millions of short reads to a genome
- The genome is very large
- The human genome is highly repetitive (~45%) → many possible mapping locations
- Highly affected by sequencing errors

Many algorithms (tools)

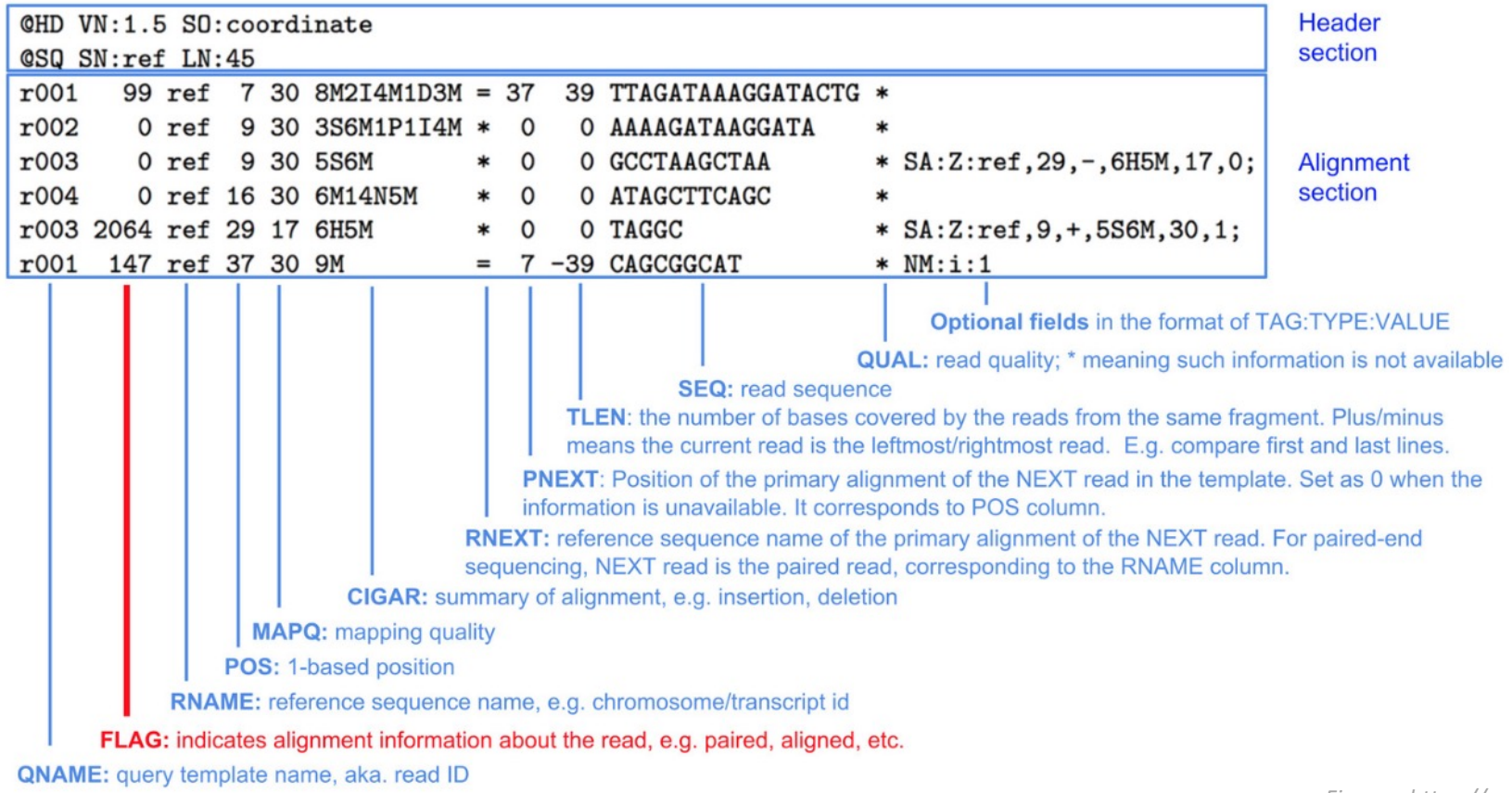
- **BWA***
- Novoalign
- Bowtie2
- GEM

Why do we use BWA(-mem)?

- It uses indexing technique called the Burrows-Wheeler transform to create a compressed version of the reference genome
- When aligning reads to the reference genome using BWA, it breaks each read into smaller fragments called seeds.
- Faster, less memory usage and same performance

Alignment

- **SAM** (Sequence Alignment/Map) and **BAM*** (Binary Alignment/Map) files are commonly used file formats in bioinformatics for storing DNA/RNA sequencing alignment data (between 10Gb and 500Gb per file)
- We obtain (in general) one for the tumour and another for the matched normal sample



*It also exists CRAM: compressed version of BAM (0.2-0.8 of bam size)

Figures: https://www.samformat.info/sam-format-flag

Limitations at different steps of the variant sequencing workflow

Each step of this workflow has particular limitations and issues, which might affect the downstream analysis and interpretation of the calls

Library preparation and sequencing

Read pre-processing and alignment

Somatic/Germline variant calling SNPs, Indels, SVs, CNVs

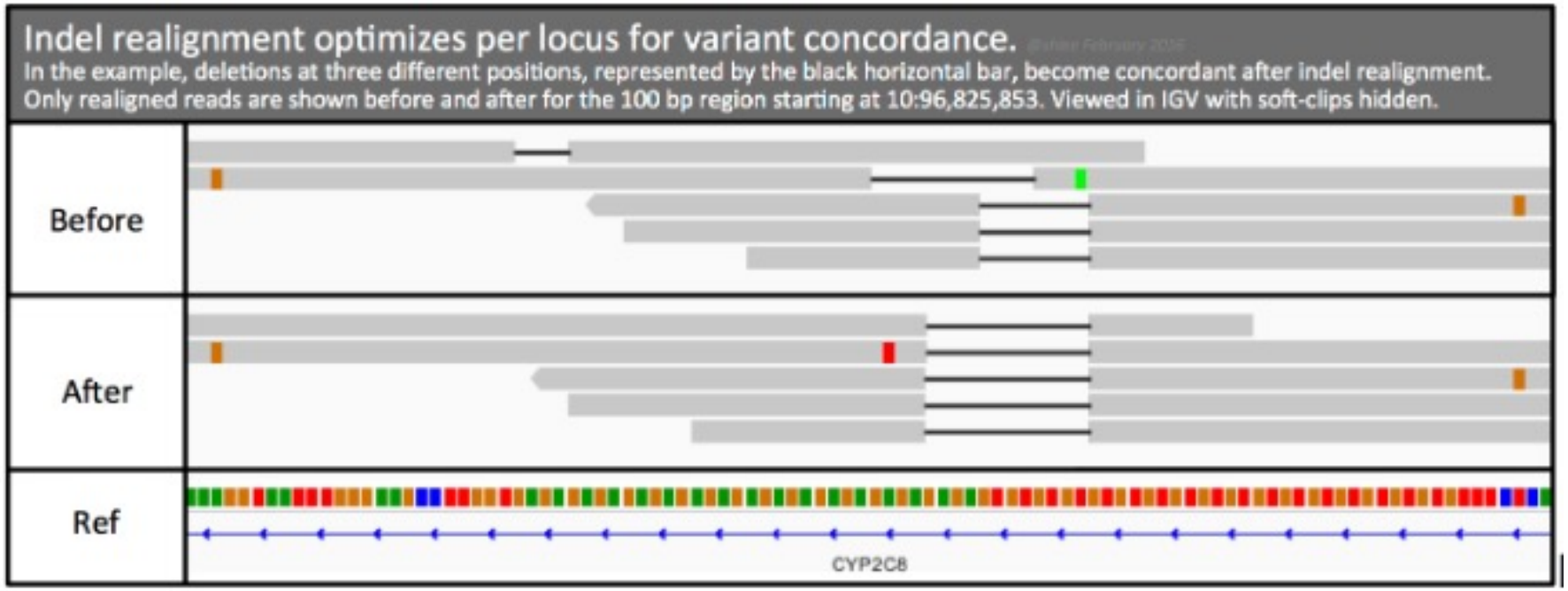
Low complexity regions	2 %
Repetitive regions	45 %
Reads spanning indels might introduce errors	Require indel realignment
Over-amplified reads: duplicates	Require dup identification

Many of them related with systematic errors

Processing bam files for variant calling

Indel realignment

- During the sequencing process, errors can occur, leading to misalignments around indel regions.
- These misalignments can result in false-positive or false-negative variant calls.
- Indel realignment aims to correct these misalignments by identifying and repositioning reads that span indel sites.
- In the past, it was an individual step in GATK3. In the new versions of GATK4, it is already implemented in the variant calling algorithm

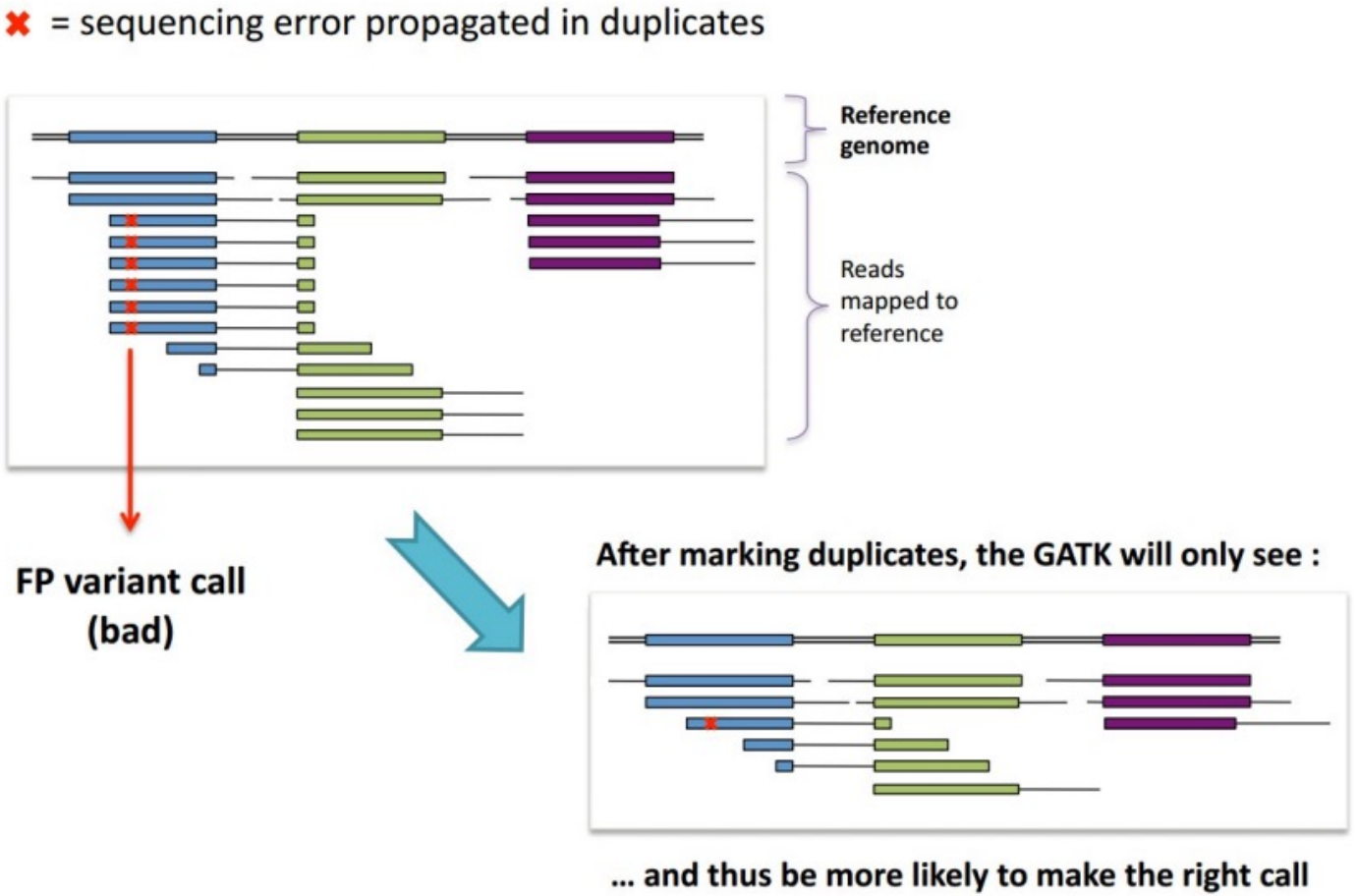


Figures: https://github.com/broadinstitute/gatk-docs

Processing bam files for variant calling

Marking duplicates

- Read duplicates refer to multiple identical or nearly identical DNA sequencing reads that arise from the same DNA fragment rather than representing distinct biological molecules.
- These duplicates can occur due to biases in library preparation and amplification steps (PCR).
- Understanding and identifying read duplicates is important in order to accurately interpret sequencing data and avoid potential biases in downstream analyses.
- Available tools: Picard, GATK4, SAMtools…

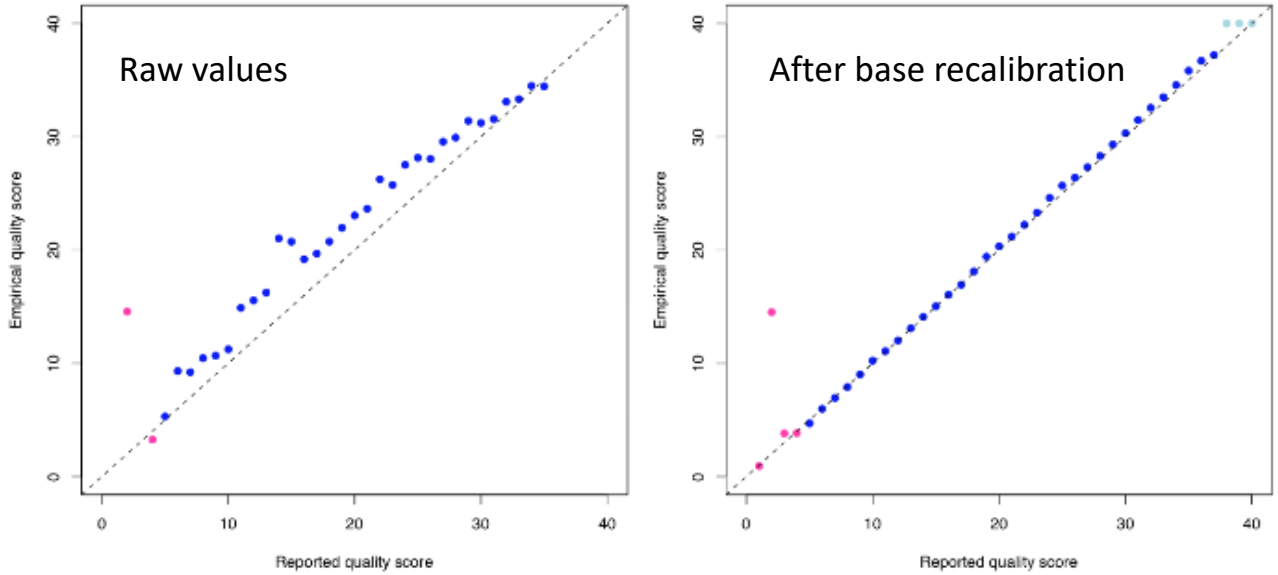


Figures: www.broadinstitute.org

Processing bam files for variant calling

Base recalibration

- The scores produced by the machines are subject to various sources of systematic (non-random) technical error, leading to over- or under-estimated base quality scores in the data.
- Base recalibration is a data pre-processing step that detects systematic errors made by the sequencing machine when it estimates the accuracy of each base call
- Base quality score recalibration (BQSR, **GATK**) is a process in which we apply machine learning to model these errors empirically and adjust the quality scores accordingly. (Recommended for GATK variant calling pipelines)



Source: https://gatk.broadinstitute.org/

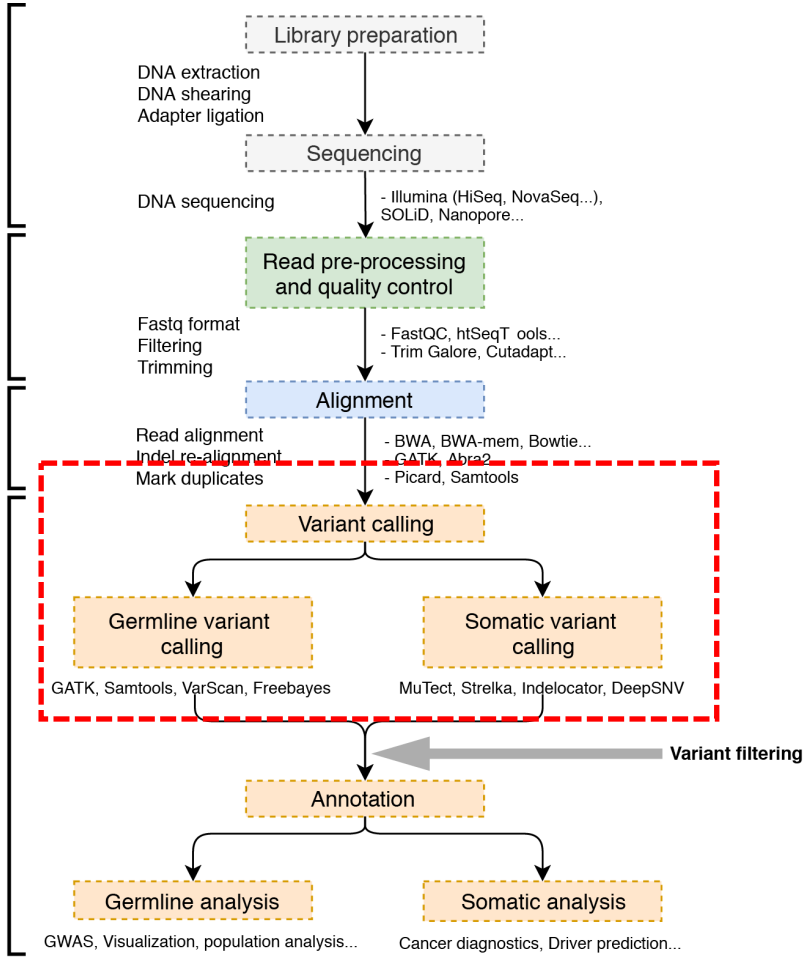
Variant calling (VC)

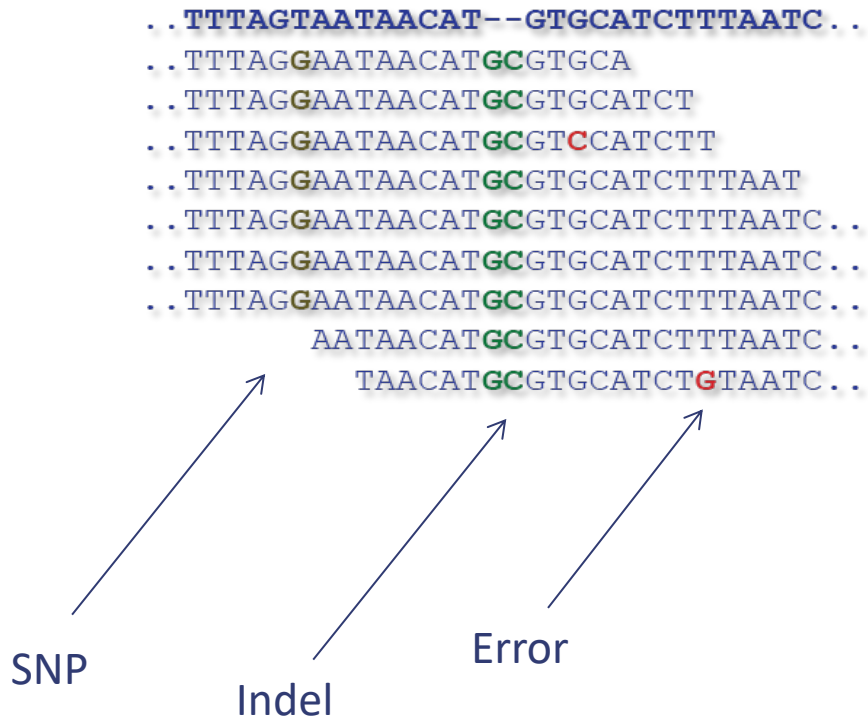
- VC refers to the process of identifying genetic variations or variants present in an individual's DNA sequence.
- VC compares the sequenced DNA reads to a reference genome and determines where the individual's sequence differs from the reference.

Library preparation and sequencing

Read pre-processing and alignment

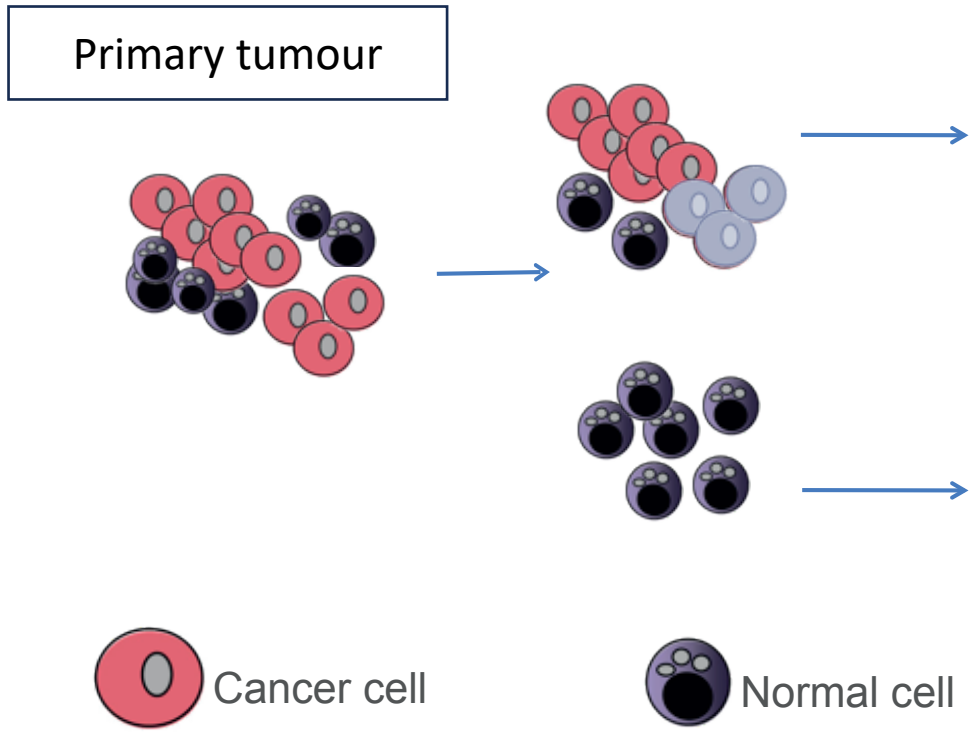
Somatic/Germline variant calling
SNPs, Indels, SVs, CNVs

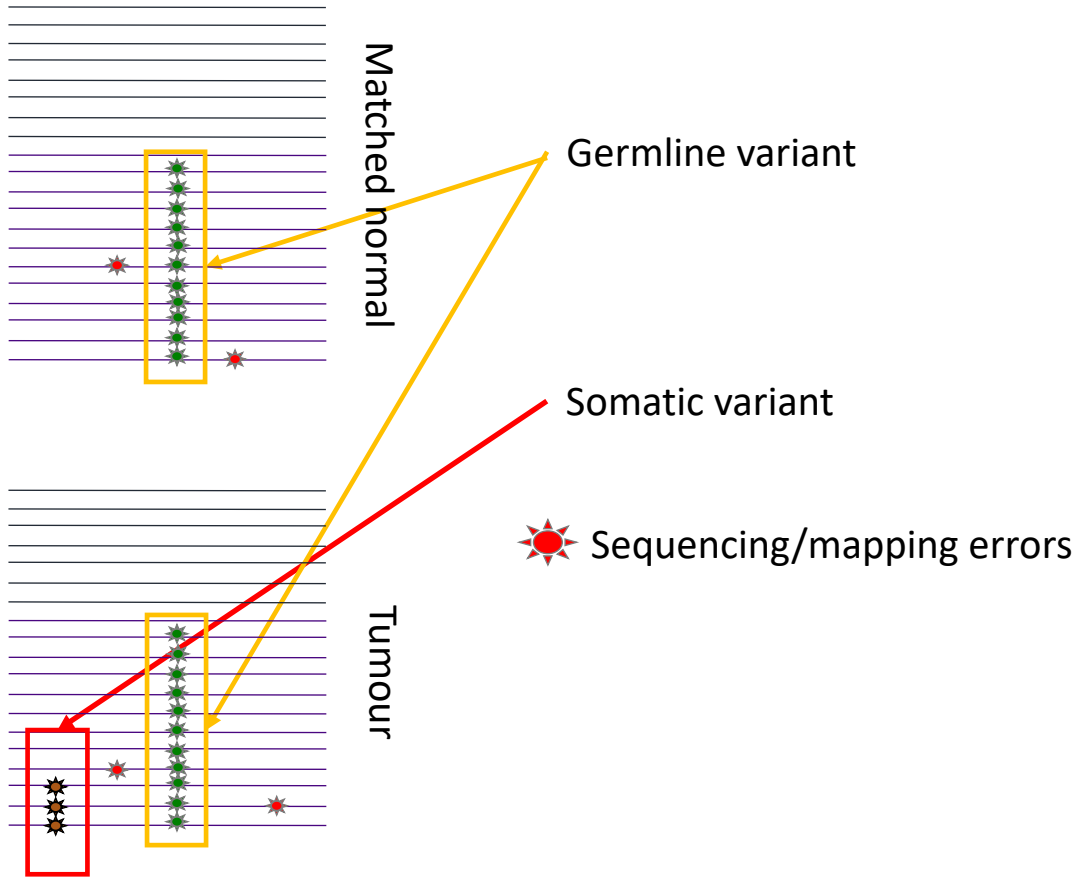




Somatic variant calling

- Somatic variant calling typically involves sequencing both tumour and matched normal samples from an individual.
- The normal sample serves as a control, representing the non-mutated or germline cells, while the tumour sample contains genetic alterations specific to the tumour cells.





Somatic variant calling

Limitations

Library preparation and sequencing

Read pre-processing and alignment

Somatic/Germline variant calling SNPs, Indels, SVs, CNVs

Sensible to all previous errors

High false positive rate

Germline calling

High performance

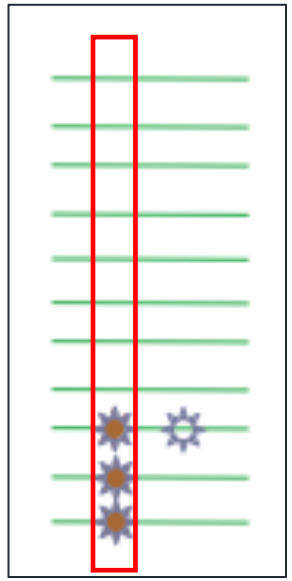
Somatic variant calling limitations

Low variant allele frequencies (**VAFs**)*

Purity and clonality

Require complex filtration

*VAF (Variant allele frequency)
Fraction of reads supporting the variant



VAF = 3 / 11 = 0.27

Huge importance of **read coverage**

Somatic variant calling

Available tools for somatic variant calling

- **MuTect2** (part of GATK)
- **Strelka2**

Tools we will learn to run today

- MuSe2
- VarScan2
- SAMtools
- VarDict
- Somatic Sniper
- Scalpel (indels in WES data)
- …

Please! Don't use germline callers for somatic variant detection

Somatic variant calling

Variant filtering

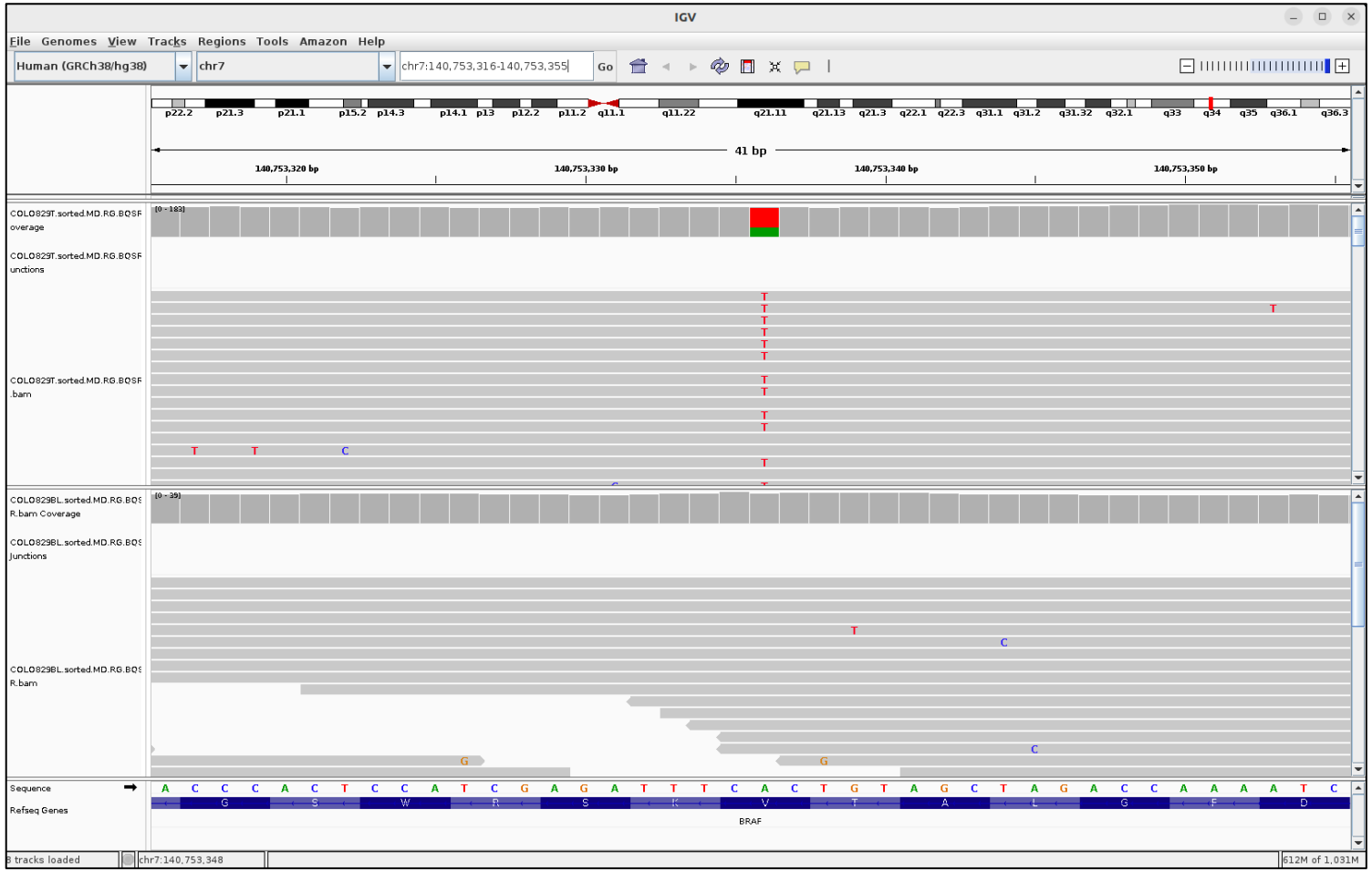
- Raw variant calls usually have lots of false positive calls
- Many algorithms have their own set of internal filters, but some post-processing is sometimes required
- Filters like minimum **VAF**, **coverage**, blacklist regions are sometimes useful
- Some tools like GATK (*GATK VariantRecalibrator*) can rank the mutations based on the quality

Some of these filters are usually applicable if sufficient number of samples can be analysed in the same batch (It will not be performed today)

Somatic variant calling

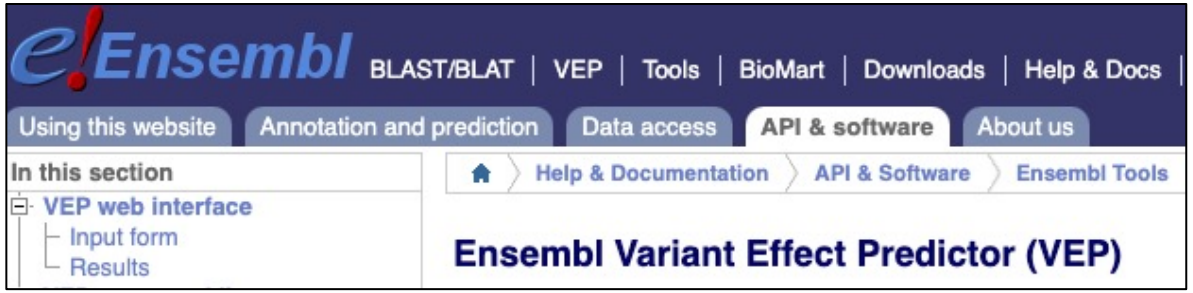
Variant visualization

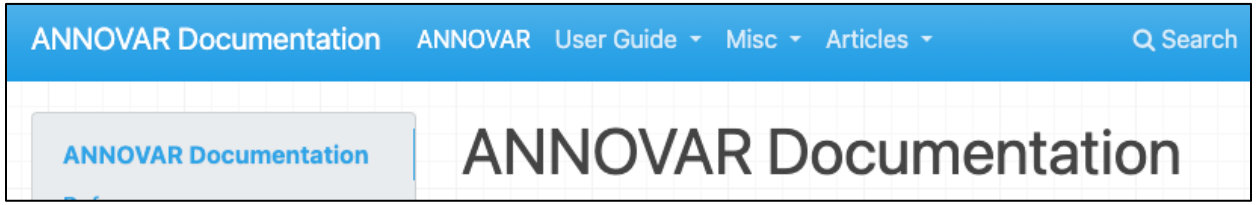
IGV (Integrative Genomics Viewer) software is very useful to check how the detected variants look in the alignment file (bam)



Variant annotation

- Detected variants can be annotated with additional information, such as functional consequences (e.g., synonymous or nonsynonymous mutations), predicted effects on proteins, and known associations with diseases.
- This annotation provides insights into the potential impact and significance of the variants.
- Some available tools:
  - **ANNOVAR** (ANNOtate VARiation)
  - VEP (Variant Effect Predictor)
  - SnpEff (SNP Effects)





Variant annotation

ANNOVAR output

- It provides insane amount of information.
- Very flexible with the type of datasets you can use annotate variants
- Gene affected, location, type of mutation, amino acid change, band location, variant population frequency (ExAC, GnomAD…), impact …

This is a small part of the columns of the file we will generate today…

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Chr	Start	End	Ref	Alt	Func.refGene	Gene.refGene	ExonicFunc.refGene	AAChange.refGene	cytoBand	ExAC_ALL	ExAC_AFR	ExAC_AMR	ExAC_EAS	ExAC_FIN
519	chr7	34658443	34658443	G	C	exonic	NPSR1	nonsynonymous SNV	NPSR1:NM_001300933:exon1:c.G31C:p.D11H,NPSR1:NM_00	7p14.3	.	.	.	.	.
559	chr7	36420245	36420245	A	G	exonic	ANLN	nonsynonymous SNV	ANLN:NM_001284301:exon10:c.A1835G:p.Q612R,ANLN:NM_	7p14.2	.	.	.	.	.
738	chr7	48643293	48643293	G	A	exonic	ABCA13	nonsynonymous SNV	ABCA13:NM_152701:exon60:c.G14843A:p.G4948D	7p12.3	.	.	.	.	.
1374	chr7	77052963	77052963	C	T	exonic	SPDYE18	synonymous SNV	SPDYE18:NM_001351348:exon5:c.G765A:p.E255E	7q11.23	.	.	.	.	.
1376	chr7	77196981	77196981	T	C	exonic	FGL2	synonymous SNV	FGL2:NM_006682:exon2:c.A618G:p.Q206Q	7q11.23	.	.	.	.	.
1448	chr7	82758692	82758692	C	T	exonic	PCLO	synonymous SNV	PCLO:NM_033026:exon25:c.G15312A:p.G5104G	7q21.11	.	.	.	.	.
1703	chr7	100819817	100819817	G	A	exonic	EPHB4	nonsynonymous SNV	EPHB4:NM_004444:exon6:c.C1037T:p.P346L	7q22.1	.	.	.	.	.
1705	chr7	100893810	100893810	G	A	exonic	ACHE	synonymous SNV	ACHE:NM_000665:exon2:c.C423T:p.S141S,ACHE:NM_001282	7q22.1	.	.	.	.	.
1706	chr7	100893949	100893949	G	A	exonic	ACHE	nonsynonymous SNV	ACHE:NM_000665:exon2:c.C284T:p.T95I,ACHE:NM_00128244	7q22.1	.	.	.	.	.
1712	chr7	101210439	101210439	G	A	exonic	PLOD3	synonymous SNV	PLOD3:NM_001084:exon14:c.C1506T:p.I502I	7q22.1	.	.	.	.	.
1736	chr7	103698005	103698005	C	T	exonic	RELN	nonsynonymous SNV	RELN:NM_005045:exon10:c.G991A:p.E331K,RELN:NM_17305	7q22.1	.	.	.	.	.
1873	chr7	113878248	113878248	C	T	exonic	PPP1R3A	synonymous SNV	PPP1R3A:NM_002711:exon4:c.G2844A:p.T948T	7q31.1	2.49E-05	0	0	0.0001	0
2355	chr7	140753336	140753336	A	T	exonic	BRAF	nonsynonymous SNV	BRAF:NM_004333:exon15:c.T1799A:p.V600E	7q34	1.65E-05	0	8.72E-05	0	0
2559	chr7	149071266	149071266	C	T	exonic	ZNF786	synonymous SNV	ZNF786:NM_152411:exon4:c.G1506A:p.R502R	7q36.1	.	.	.	.	.
2571	chr7	149860907	149860907	C	A	exonic	ZNF862	nonsynonymous SNV	ZNF862:NM_001099220:exon7:c.C1747A:p.Q583K	7q36.1	.	.	.	.	.
2592	chr7	150952580	150952580	G	A	exonic	KCNH2	nonsynonymous SNV	KCNH2:NM_001204798:exon2:c.C382T:p.L128F,KCNH2:NM_1	7q36.1	.	.	.	.	.
2730															

Variant annotation

ANNOVAR output (mutation impact)

- It also provides the **mutation impact** obtained by LJB dbNSFP (on non-synonymous variants annotation) included in the ANNOVAR output
- The LJB databases include several impact scores*
- It also can include *ClinVar**: check if mutation is known to have a clinical impact (***https://www.ncbi.nlm.nih.gov/clinvar/intro/)*** **(**Not used today)
- Check ***Intogen*** **(*https://www.intogen.org/)*** to see how frequent the genes are affected in each cancer type
- Check Cosmic (***https://cancer.sanger.ac.uk/census***) to see the list of Cancer Gene Census (CGC).

Score (dbtype)	# variants in LJB23 build hg19	Categorical Prediction
SIFT (sift)	77593284	D: Deleterious (sift<=0.05); T: tolerated (sift>0.05)
PolyPhen 2 HDIV (pp2_hdiv)	72533732	D: Probably damaging (>=0.957), P: possibly damaging (0.453<=pp2_hdiv<=0.956); B: benign (pp2_hdiv<=0.452)
PolyPhen 2 HVar (pp2_hvar)	72533732	D: Probably damaging (>=0.909), P: possibly damaging (0.447<=pp2_hdiv<=0.909); B: benign (pp2_hdiv<=0.446)
LRT (lrt)	68069321	D: Deleterious; N: Neutral; U: Unknown
MutationTaster (mt)	88473874	A" ("disease_causing_automatic"); "D" ("disease_causing"); "N" ("polymorphism"); "P" ("polymorphism_automatic"
MutationAssessor (ma)	74631375	H: high; M: medium; L: low; N: neutral. H/M means functional and L/N means non-functional
FATHMM (fathmm)	70274896	D: Deleterious; T: Tolerated
MetaSVM (metasvm)	82098217	D: Deleterious; T: Tolerated
MetaLR (metalr)	82098217	D: Deleterious; T: Tolerated
GERP++ (gerp++)	89076718	higher scores are more deleterious
PhyloP (phylop)	89553090	higher scores are more deleterious
SiPhy (siphy)	88269630	higher scores are more deleterious

*Check here for a more detailed information of each score:
https://annovar.openbioinformatics.org/en/latest/user-guide/filter/

Conclusions

1. We learnt the different steps of a somatic variant calling workflow:
    - Quality control of raw reads
    - Alignment
    - How to prepare the BAM files for the variant calling
    - Somatic variant calling
    - Variant annotation
2. We explained the different source of artefacts in the somatic variant calling workflow
3. We learnt the challenges of SNV detection and how to solve them (sequencing/mapping errors, VAF limitations, coverage…)

What is next?

Day one – Monday 10 July 2023		
09:00 – 09:30	Arrival and registration	
09:30 – 09:45	Welcome	Ajay Mishra
09:45 – 10:45	Course introduction	Isidro Cortes-Ciriano
10:45 – 11:00	Break	
11:00 – 12:30	Introduction to cancer genomics and its caveats	Tobias Rausch
12:30 – 13:30	Lunch	
13:30 – 14:30	High-throughput sequencing and single-nucleotide variant analysis in cancer	Francesc Muyas Remolar
14:30 – 15:00	Break	
15:00 – 17:30	Practical: Alignment and SNV analysis	Francesc Muyas Remolar
17:30 – 18:30	Accommodation check-in and free time	
18:30	Dinner at Hinxton Hall Conference Centre	

Acknowledgements

Cortes-Ciriano group (EMBL-EBI)

Course organisers

Isidro Cortes Ciriano, EMBL-EBI

Ajay Mishra, EMBL-EBI

Tobias Rausch, EMBL

Sophie Spencer, EMBL-EBI

All trainers