Universidad Francisco de Vitoria UFV Madrid

NGS applications

Variant detection: SNPs, CNVs, structural variants, epigenetic variation

Elena Piñeiro - epineiro@cnio.es

BU CNIO BIOINFORMATICS UNIT

cnio stop cancer

Genomic Variants

What are genomic variants?

- Genomic variants are **permanent** changes in the DNA sequence of an organism.

- They can emerge by different mechanisms:

  - Recombination during gametes formation.
  - Errors during the DNA replication.
  - External factors like radiation, viruses, transposons, tobacco, UV light.

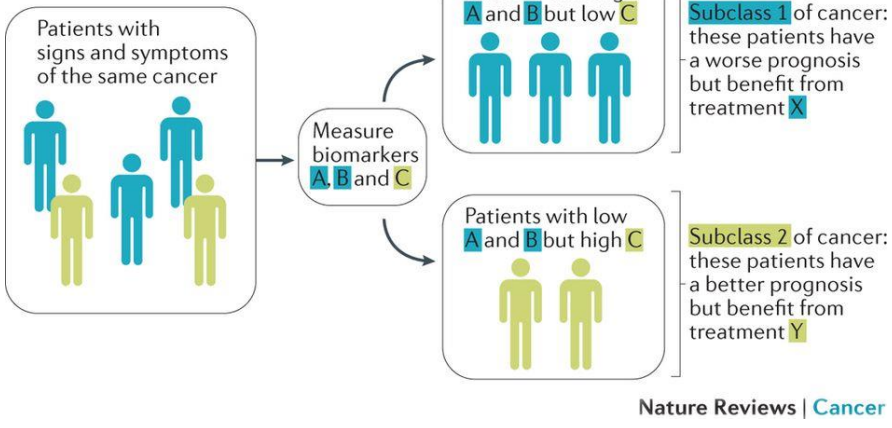
Relevance of genomic variants

- They are a source of genomic variability, laying the foundations for the evolutionary mechanisms.
- Allow phenotypic differences between individuals (hair color, skin color, ...).
- Involved in diseases and drug response.

Areas of application of variant

- Population and demographical studies.
- Phylogenetic studies.
- Epidemiological studies.
- Forensic applications.
- Clinical applications: disease susceptibility, drug response.

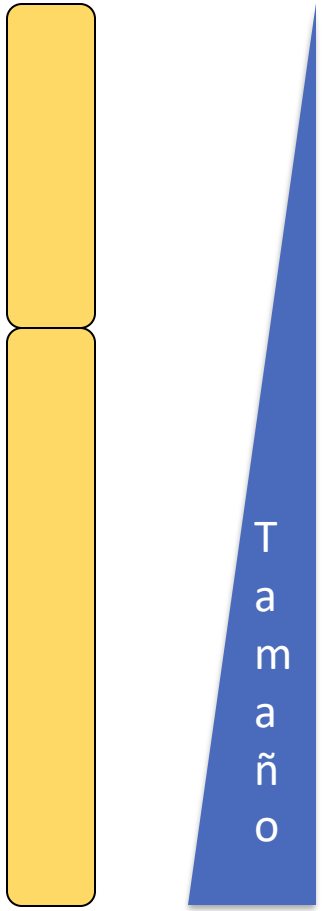
Personalized Medicine



Types of variants

- Different types of variants according to different criteria:
  - Variant size
  - Position of the variant in the DNA sequence
  - Consequence of the variant in transcription and translation
  - Clinical implication

Classification of variants according to size



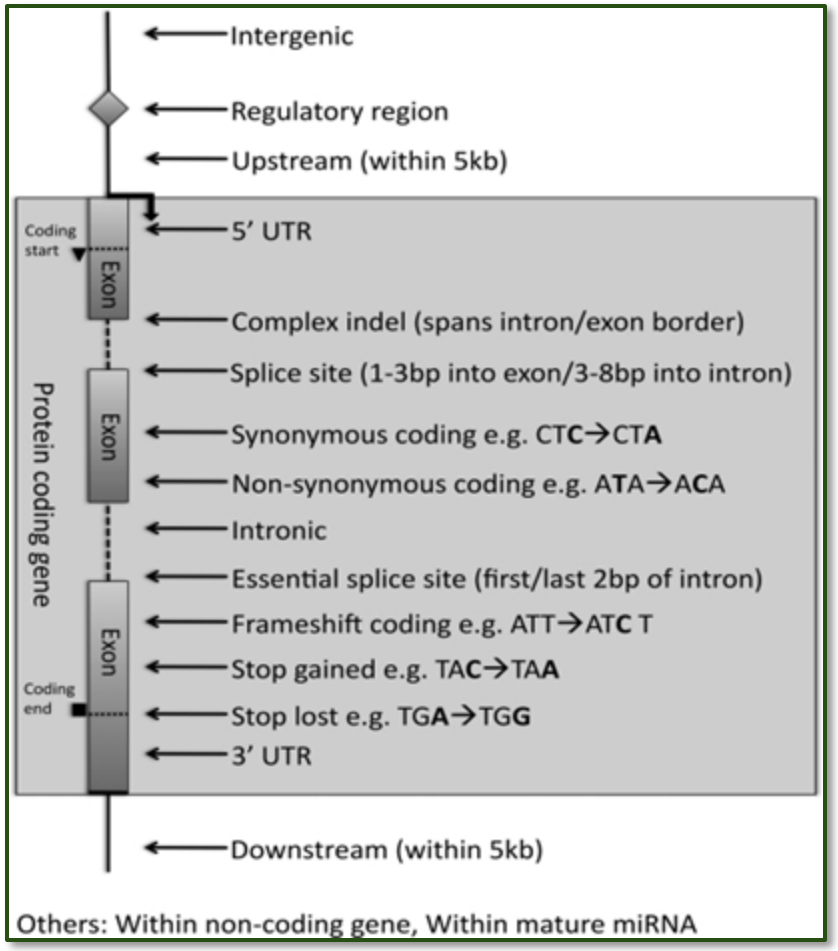
SNVs Single Nucleotide Variation
Indels Small insertions and deletions

VNTRs (Micro, Minisatellites)
Variable number of tandem repetitions

CNVs Copy number variation
Translocations, Inversions

Aneuploidies

Classification of variants according to the sequence position



McLaren et al. Bioinformatics. 2010.

- Intergenic
- In regulatory regions
- Upstream
- Downstream
- In genes
    - Untranslated regions: 5'UTR y 3'UTR
    - Exons
    - Introns
    - Splicing sites

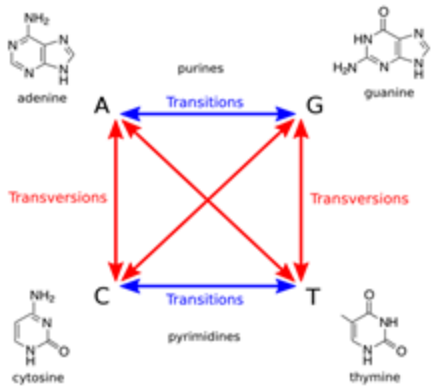
Consequences of SNV in protein coding regions

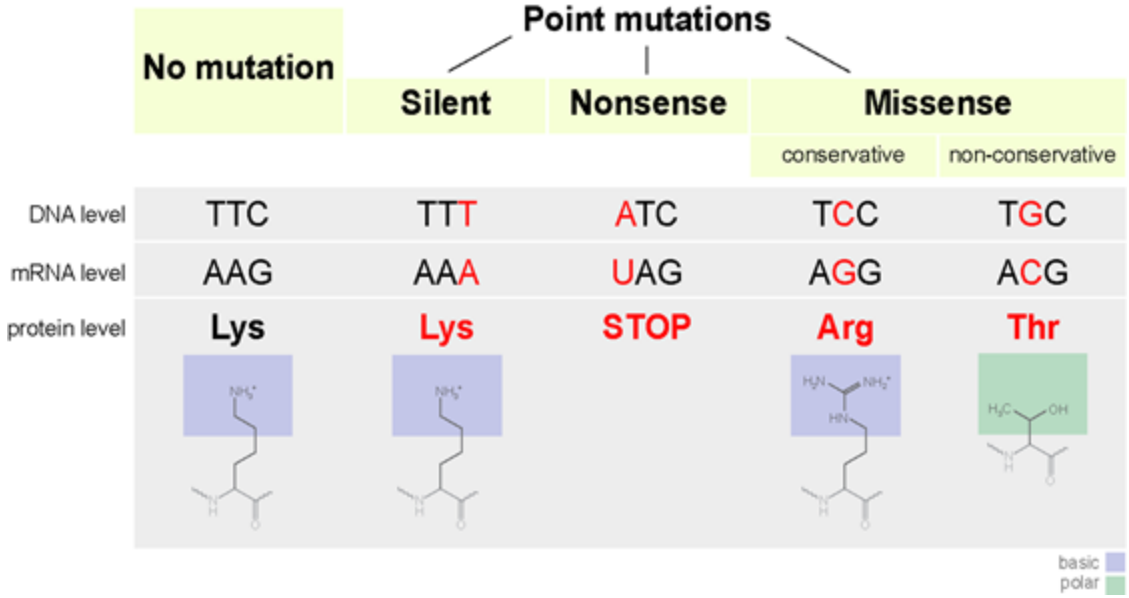
SNV: Single-Nucleotide Variant

Wild-type sequence (reference)
ATCTTCAGCCTAAAAGATGAAGTT

Transition (green)
ATCTTCAGCCTGAAAGATGAAGTT

Transversion (blue)
ATCTTCAGCCTCAAAGATGAAGTT





Stop loss, start gain, start loss

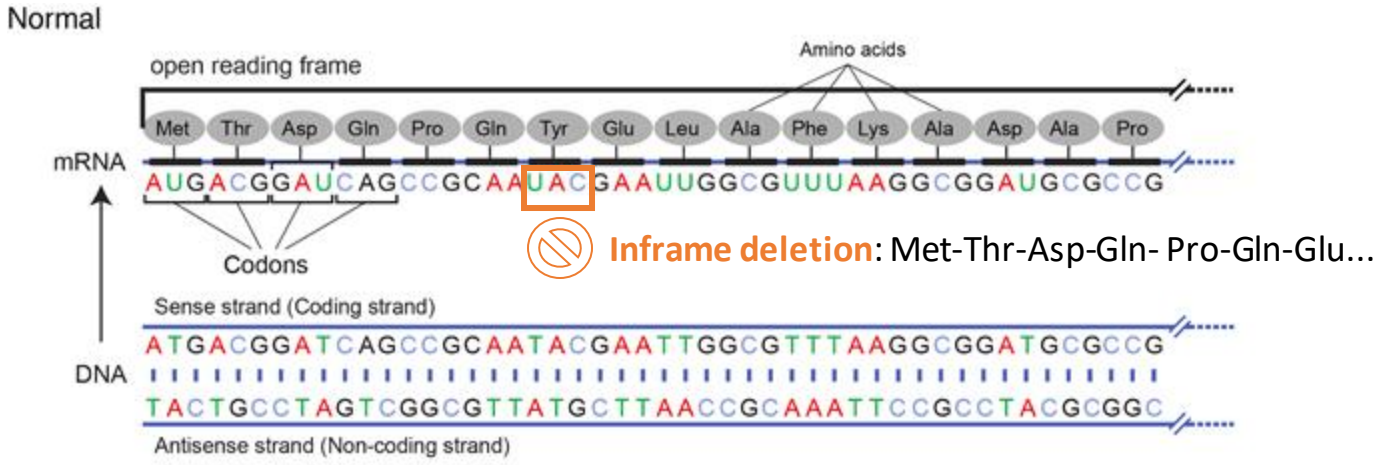
Consequence of INDEL in protein coding regions

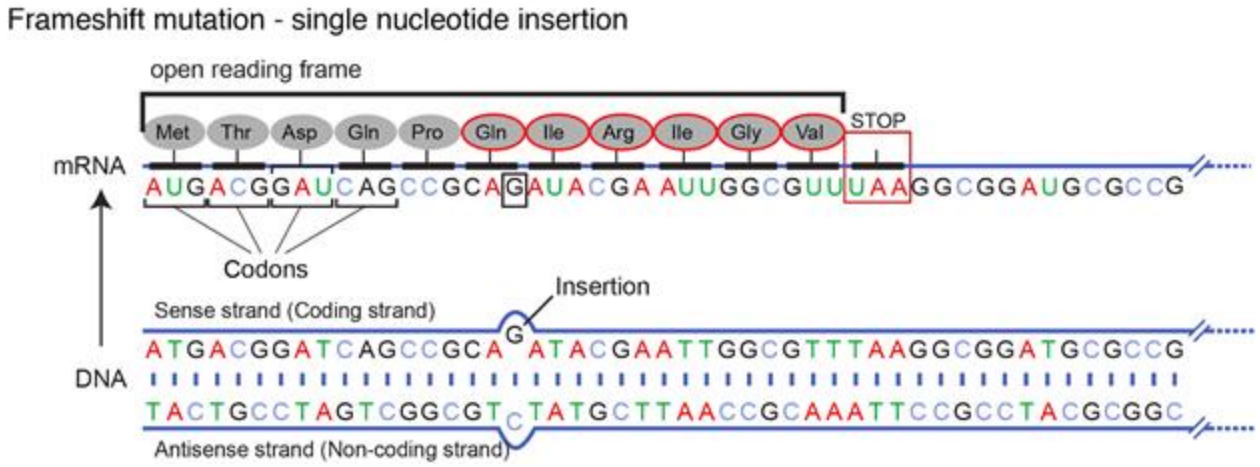
Indel: short Insertions and deletions

Wild-type sequence (reference)
ATCTTCAGCCTAAAAGATGAAGTT

3 bp deletion (light blue)
ATCTTCAGCCTA---GATGAAGTT

4bp insertion (orange)
ATCTTCAGCCTATGTGAAAGATGAAGTT

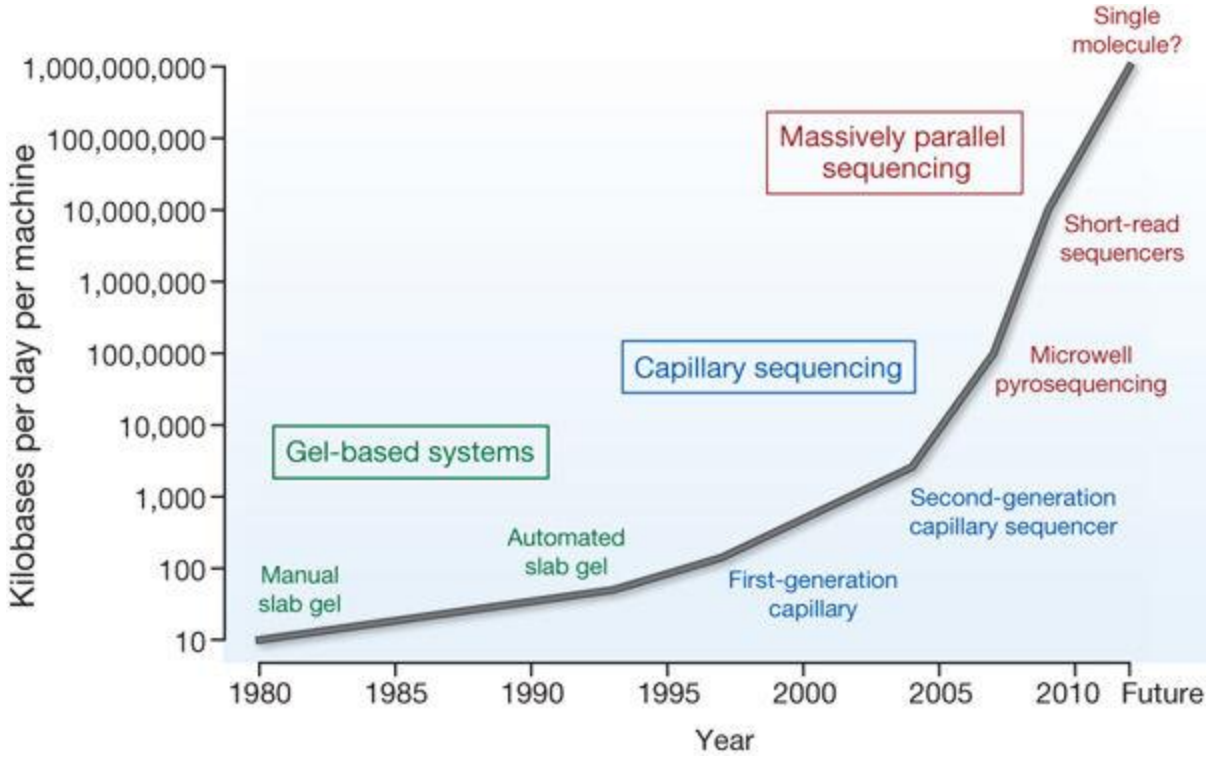




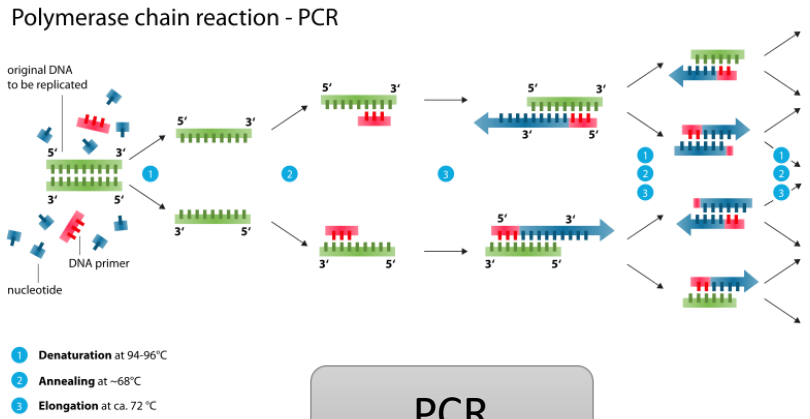
Source: https://www.genome.gov/dmd/

Variant Detection with NGS

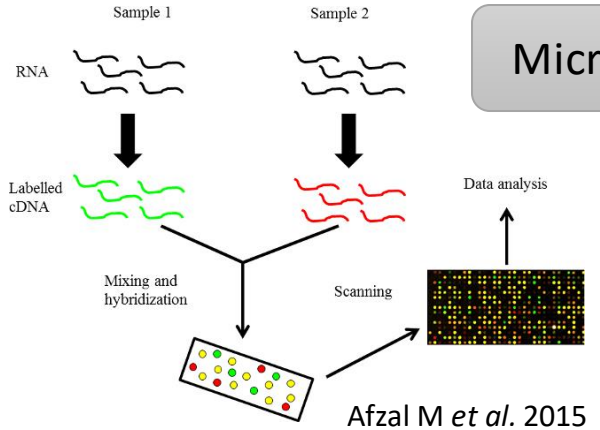
Variant detection



DNA sequencing



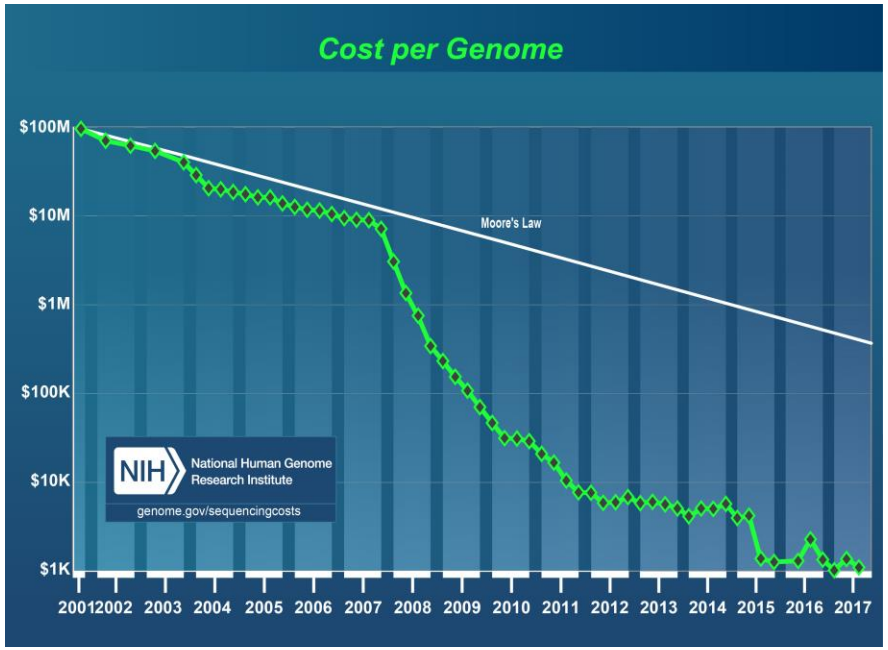
PCR

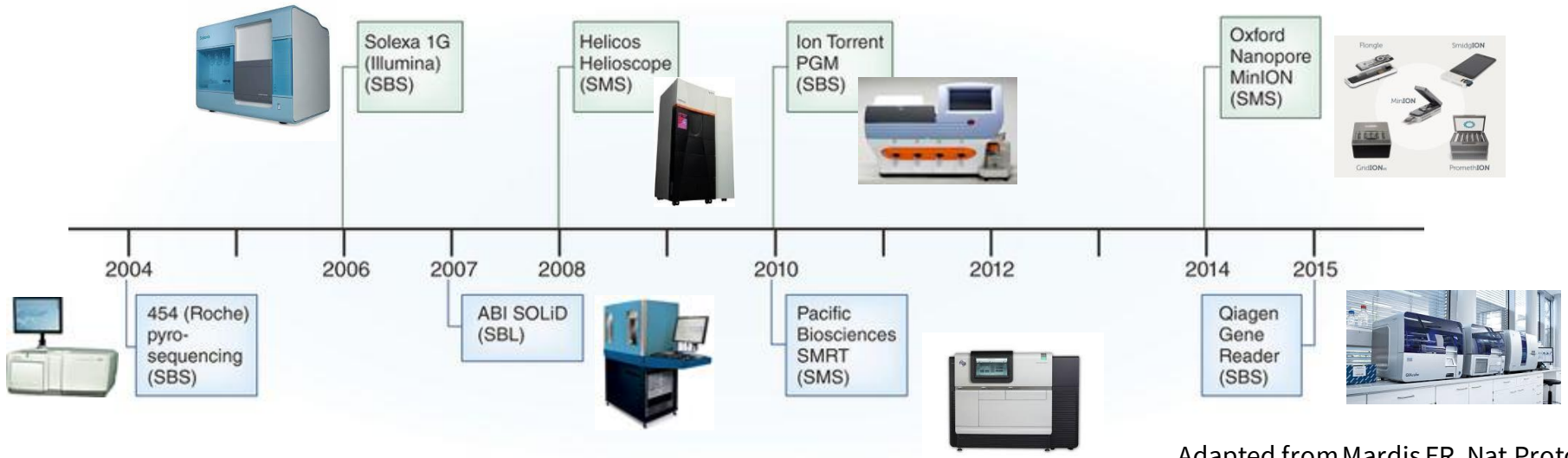


Microarrays

Afzal M *et al.* 2015

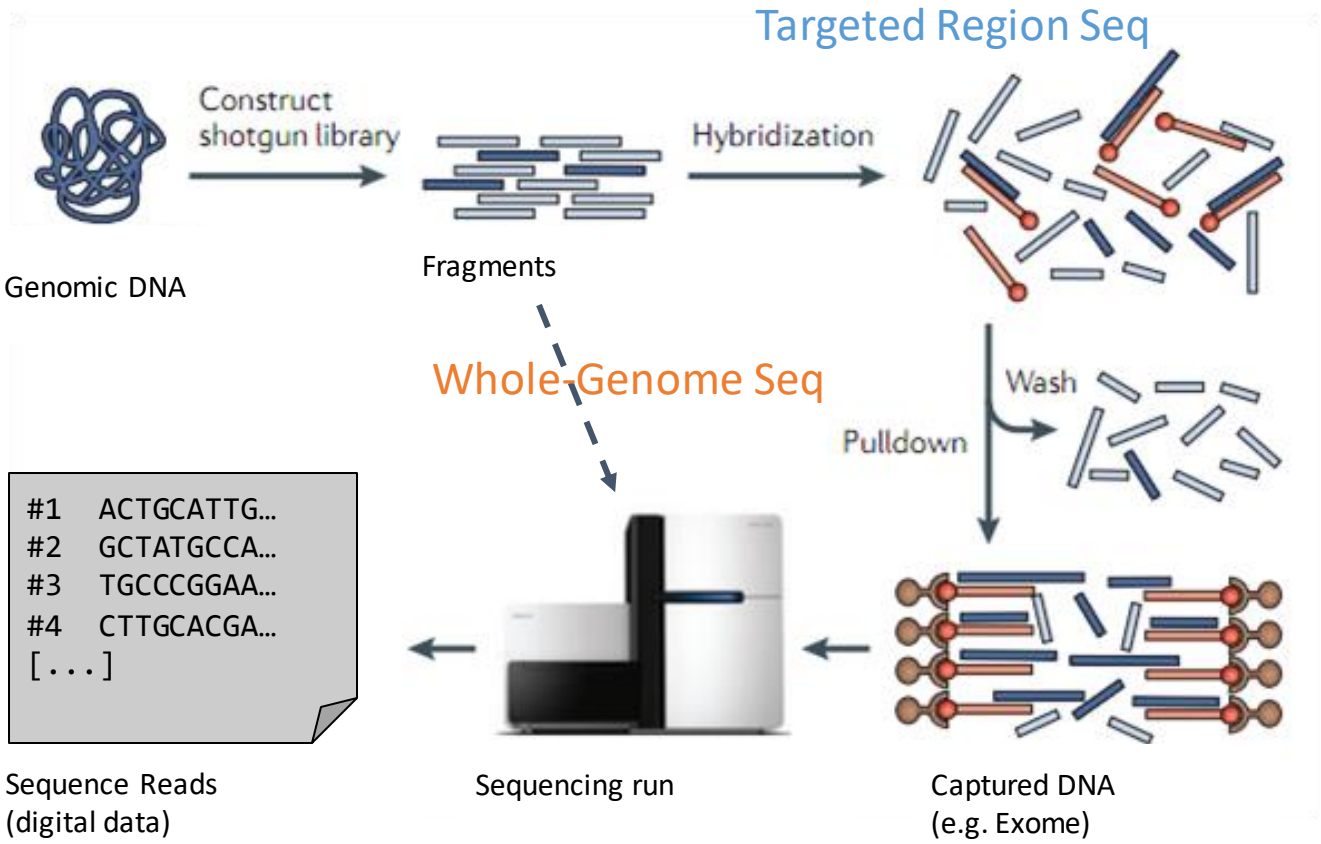
Sequencing cost has been coming down





Adapted from Mardis ER. Nat Protoc 2017

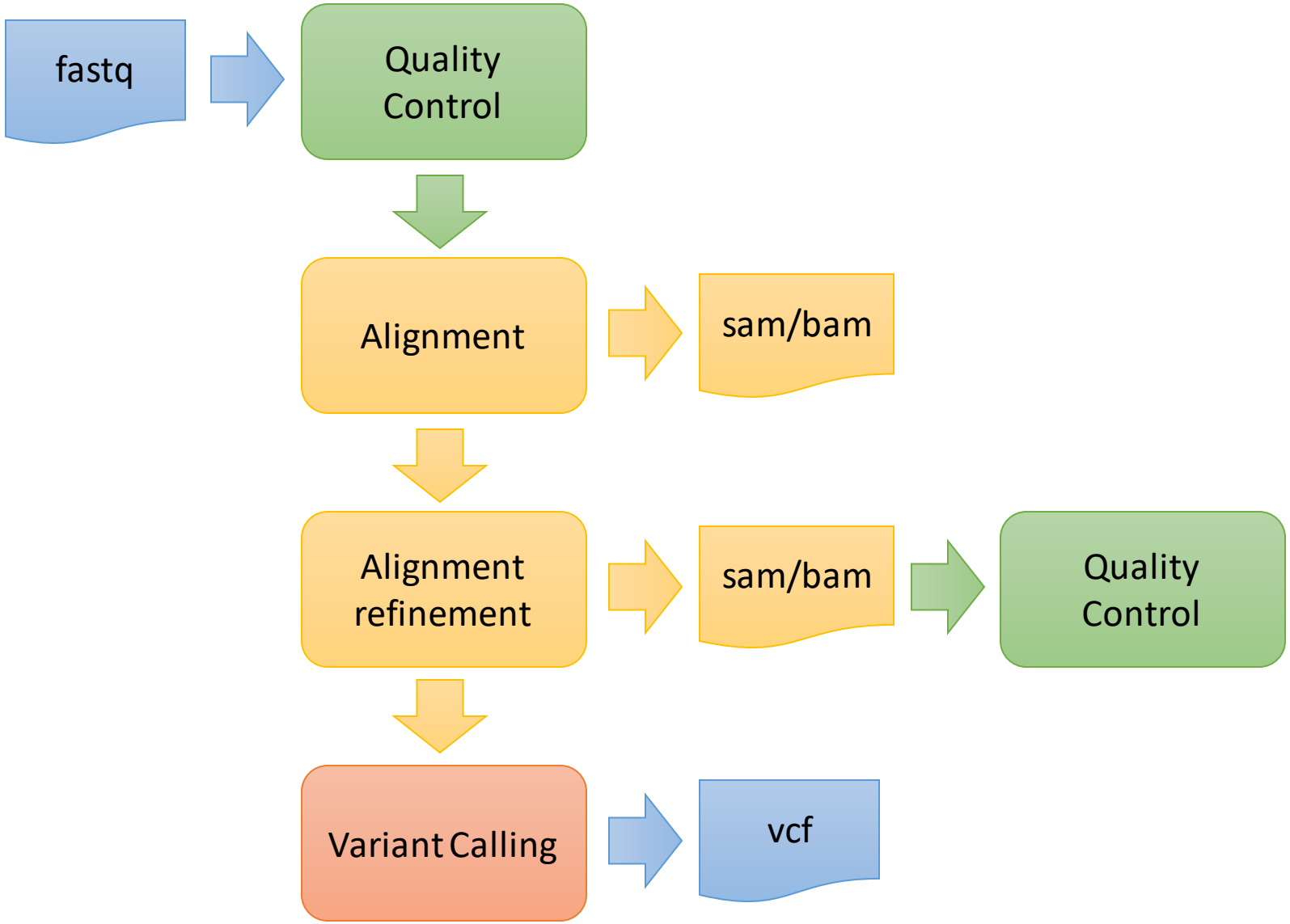
DNA Sequencing data generation



The sequence reads belong from the ends of the original fragment.

Steps in bioinformatic analysis

General steps



Fasta & Quality scores

FASTA

- Typical extensions: .fasta, .fas, .fa, .fna, .fsa
- Each sequence is composed by at least two consecutive lines:
  - ">" Sequence name and optional description (space separated)
  - Line(s) with the whole sequence

```
>DNA_SEQUENCE_1
NNNNNNCTCTGGGGGACAGAACCCTATGGTGGCCCCGGCTCCTCCCCAGTATCCAGTCCT
CCGTGAAGATGGAGCCATATTCC
```

60 chars

```
>DNA_SEQUENCE_1
NNNNNNCTCTGGGGGACAGAACCCTATGGTGGCCCCGGCTCCTCCCCAGTATCCAGTCCT
CCGTGAAGATGGAGCCATATTCC
>DNA_SEQUENCE_2
GGGGGACAGAACCCTATGGTGGCCCCGGCTCCTCCCCAGTATCCAGTCCT
>DNA_SEQUENCE_3
CTCCTCCCCAGTATCCAGTCCTGGGGGACAGAACCCTATGGTGGCCCCGCCCAGTATCCA
```

We can have multiple sequences in the same file (multifasta)

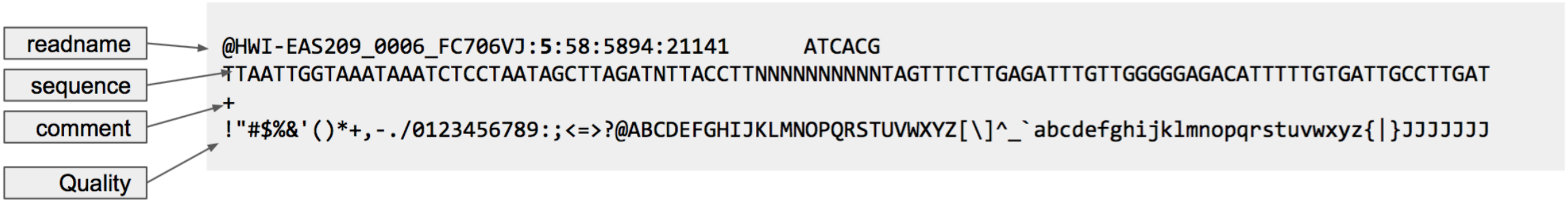
FASTA

IUPAC nucleotide code	Base
A	**Adenine**
C	**Cytosine**
G	**Guanine**
T (or U)	**Thymine (or Uracil)**
R	A or G
Y	C or T
S	G or C
W	A or T
K	G or T
M	A or C
B	C or G or T
D	A or G or T
H	A or C or T
V	A or C or G
N	any base
. or -	gap

Nucleotide codes (IUPAC)

FASTQ

- Typical extensions: .fq, .fastq
- Each read is composed by 4 lines:
  - "@" Read name and optional description (space separated)
  - Sequence
  - "+" (optionally: repeat the read name)
  - Base Quality Score

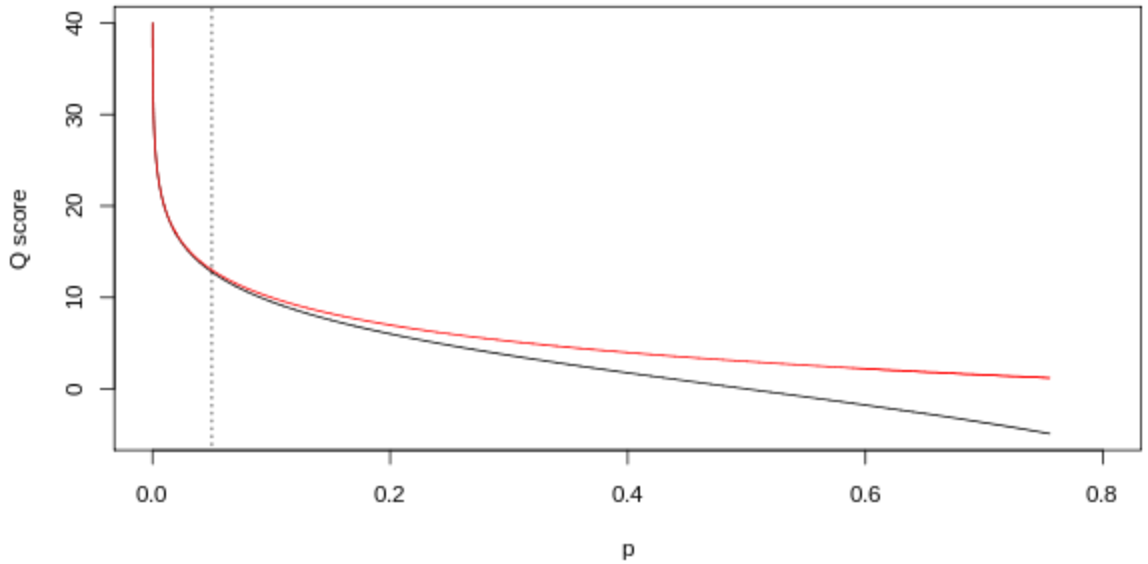


Quality Score

- Phred quality scores (**QPhred**) are defined as a property which is **logarithmically related to the** base-calling **error probabilities *p***

$$QPhred = -10 \log_{10}(p)$$

- The higher the *QPhred* , the lower the probability that the base calling is erroneous



- The score is written as the character whose ASCII code is
  - QPhred + number (usually 33)

ASCII code

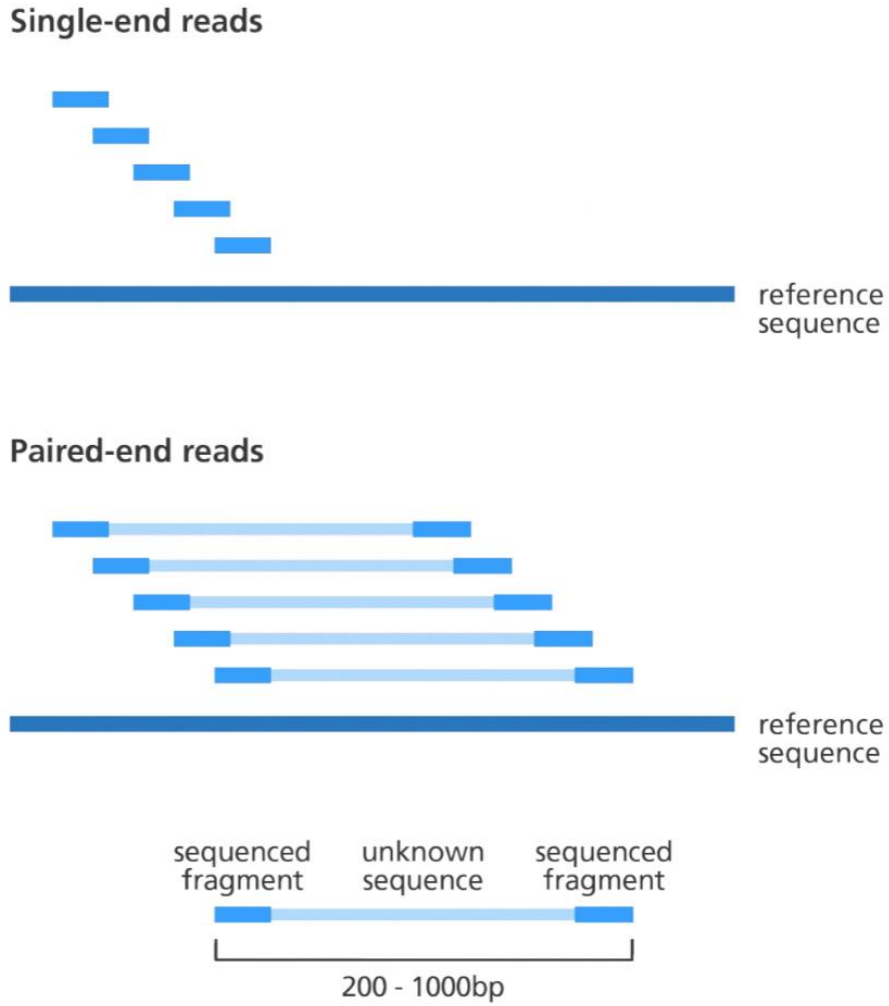
Dec	Hx	Oct	Char	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr
0	0	000	NUL (null)	32	20	040		Space	64	40	100	@	@	96	60	140	`	`
1	1	001	SOH (start of heading)	33	21	041	!	!	65	41	101	A	A	97	61	141	a	a
2	2	002	STX (start of text)	34	22	042	"	"	66	42	102	B	B	98	62	142	b	b
3	3	003	ETX (end of text)	35	23	043	#	#	67	43	103	C	C	99	63	143	c	c
4	4	004	EOT (end of transmission)	36	24	044	$	$	68	44	104	D	D	100	64	144	d	d
5	5	005	ENQ (enquiry)	37	25	045	%	%	69	45	105	E	E	101	65	145	e	e
6	6	006	ACK (acknowledge)	38	26	046	&	&	70	46	106	F	F	102	66	146	f	f
7	7	007	BEL (bell)	39	27	047	'	'	71	47	107	G	G	103	67	147	g	g
8	8	010	BS (backspace)	40	28	050	(	(	72	48	110	H	H	104	68	150	h	h
9	9	011	TAB (horizontal tab)	41	29	051	)	)	73	49	111	I	I	105	69	151	i	i
10	A	012	LF (NL line feed, new line)	42	2A	052	*	*	74	4A	112	J	J	106	6A	152	j	j
11	B	013	VT (vertical tab)	43	2B	053	+	+	75	4B	113	K	K	107	6B	153	k	k
12	C	014	FF (NP form feed, new page)	44	2C	054	,	,	76	4C	114	L	L	108	6C	154	l	l
13	D	015	CR (carriage return)	45	2D	055	-	-	77	4D	115	M	M	109	6D	155	m	m
14	E	016	SO (shift out)	46	2E	056	.	.	78	4E	116	N	N	110	6E	156	n	n
15	F	017	SI (shift in)	47	2F	057	/	/	79	4F	117	O	O	111	6F	157	o	o
16	10	020	DLE (data link escape)	48	30	060	0	0	80	50	120	P	P	112	70	160	p	p
17	11	021	DC1 (device control 1)	49	31	061	1	1	81	51	121	Q	Q	113	71	161	q	q
18	12	022	DC2 (device control 2)	50	32	062	2	2	82	52	122	R	R	114	72	162	r	r
19	13	023	DC3 (device control 3)	51	33	063	3	3	83	53	123	S	S	115	73	163	s	s
20	14	024	DC4 (device control 4)	52	34	064	4	4	84	54	124	T	T	116	74	164	t	t
21	15	025	NAK (negative acknowledge)	53	35	065	5	5	85	55	125	U	U	117	75	165	u	u
22	16	026	SYN (synchronous idle)	54	36	066	6	6	86	56	126	V	V	118	76	166	v	v
23	17	027	ETB (end of trans. block)	55	37	067	7	7	87	57	127	W	W	119	77	167	w	w
24	18	030	CAN (cancel)	56	38	070	8	8	88	58	130	X	X	120	78	170	x	x
25	19	031	EM (end of medium)	57	39	071	9	9	89	59	131	Y	Y	121	79	171	y	y
26	1A	032	SUB (substitute)	58	3A	072	:	:	90	5A	132	Z	Z	122	7A	172	z	z
27	1B	033	ESC (escape)	59	3B	073	;	;	91	5B	133	[	[	123	7B	173	{	{
28	1C	034	FS (file separator)	60	3C	074	<	<	92	5C	134	\	\	124	7C	174			|
29	1D	035	GS (group separator)	61	3D	075	=	=	93	5D	135	]	]	125	7D	175	}	}
30	1E	036	RS (record separator)	62	3E	076	>	>	94	5E	136	^	^	126	7E	176	~	~
31	1F	037	US (unit separator)	63	3F	077	?	?	95	5F	137	_	_	127	7F	177	&#127;	DEL

Source: www.LookupTables.com

FASTQ – Single-end/Paired-end

One unique sample can have 1 or 2 files:

- If **single-end** Seq -> 1 file (name ".fastq")
- If **paired-end** Seq -> 2 files (names "_R1.fastq" "_R2.fastq")



Different NGS technologies have different capabilities

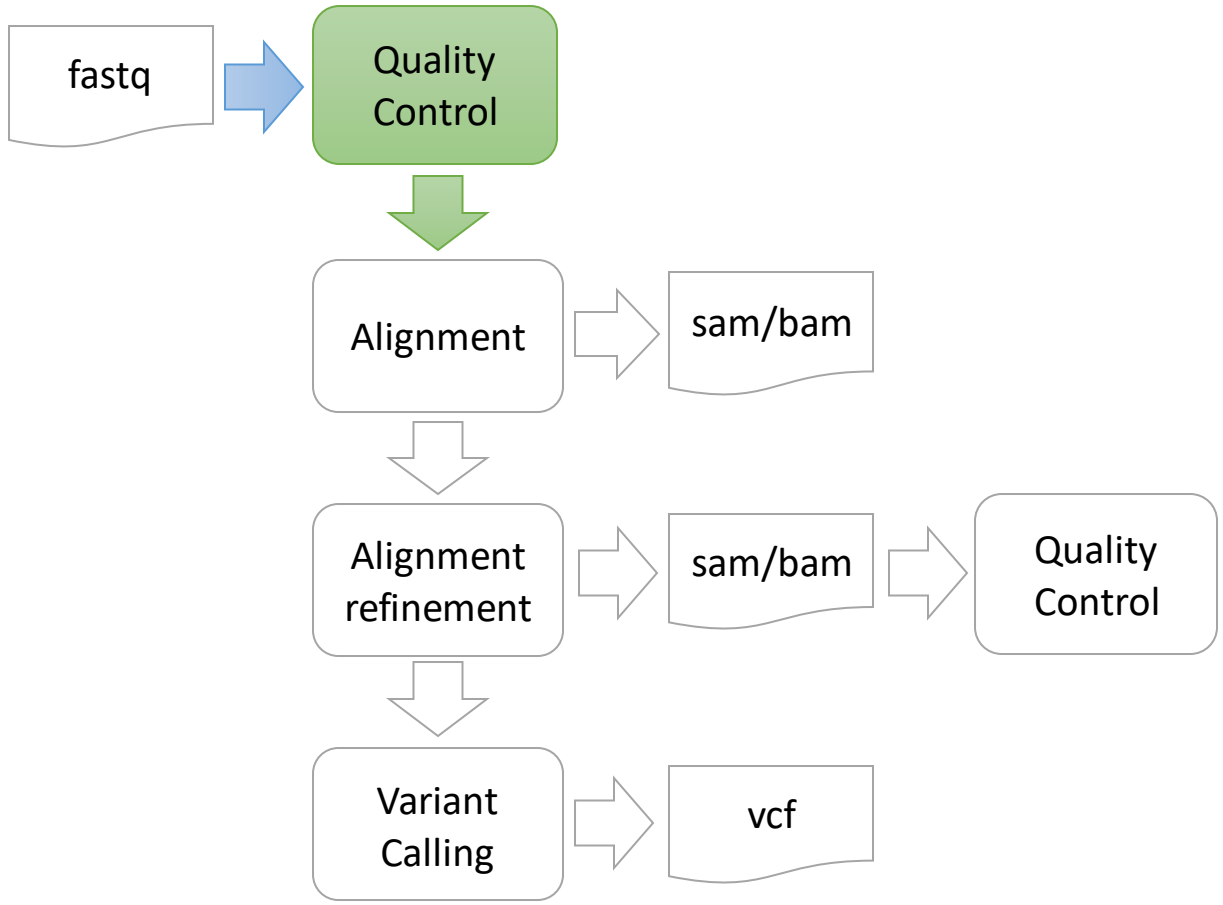
Table 1 | **Main characteristics of current NGS technologies**

Technology	Run type			Maximum read length	Quality scores	Error rates	Refs
	Single end	Paired end	Mate pair				
Illumina	Yes	Yes	Yes	300 bp	>30	0.0034–1%	59
SOLiD	Yes	Yes	Yes	75 bp	>30	0.01–1%	60
IonTorrent	Yes	Yes	No	400 bp	~20	1.78%	22
454	Yes	Yes	No	~700 bp (up to 1 kb)	>20	1.07–1.7%	53,61
Nanopore	Yes	No	No	5.4–10 kb	NA	10–40%	62–66
PacBio	Yes	No	No	~15 kb (up to 40 kb)	<10	5–10%	22,67–69

454, 454 pyrosequencing (Roche); NA, not applicable; Nanopore, Oxford Nanopore Technologies; NGS, next-generation sequencing; PacBio, Pacific Biosciences; SOLiD, sequencing by oligonucleotide ligation and detection (Thermo Fisher).

Nature Reviews Genetics 17,459–469(2016)doi:10.1038/nrg.2016.57

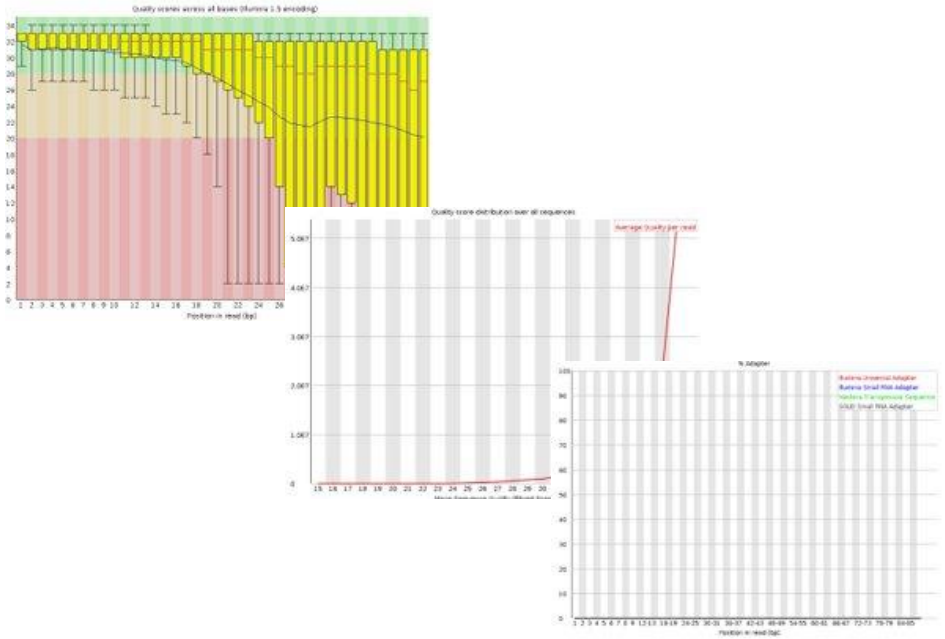
Quality control



Quality control

- Is the read number in the expected range?
- Have bases a good quality?
- Is there any contamination in the samples?

- Software:
  - FastQC
  - FastQ Screen



FastQC

- **Input**: FastQ files. **Output report**: html (web-like) with plots

- Raw files: https://fundacioncnio-my.sharepoint.com/:u:/g/personal/epineiro_cnio_es/EaWrloqPdb5GjXxak1O4TzkBUmlI6sqnupz8SLuJ2aT5-A?e=hv0Nss

Patient suffering ovarian cancer.

Whole-exome sequencing data from two samples from the patient:

- Tumour sample.
- Matched normal sample (healthy tissue) from epithelium.

Library protocol: Agilent SureSelect V5 Human All Exons.

Sequencing platform: HiSeq 2000 (Illumina)

NOTE: This data was simulated and reduced in order to perform the computational analysis in class time.

FastQC

```
> docker pull osvaldogc/ufv:2.0
```
1

```
> SAMPLES_LOCAL=/Path_directory_raw_data/
> SAMPLES_DOCKER=/SAMPLES
> RESULTS_LOCAL=/Path_directory_RESULTS
> RESULTS_DOCKER=/RESULTS
```
2

```
> chmod 777 $RESULTS_LOCAL
```
3

```
> docker run --rm  -v $SAMPLES_LOCAL:$SAMPLES_DOCKER -v $RESULTS_LOCAL:$RESULTS_DOCKER -it osvaldogc/ufv:2.0 /bin/bash
```
4

```
> fastqc /SAMPLES/fastq_file_name -o /RESULTS
```

FastQC execution

FastQC Report

Mon 7 Sep 2020
WEx_Normal_R1.fastq

Summary

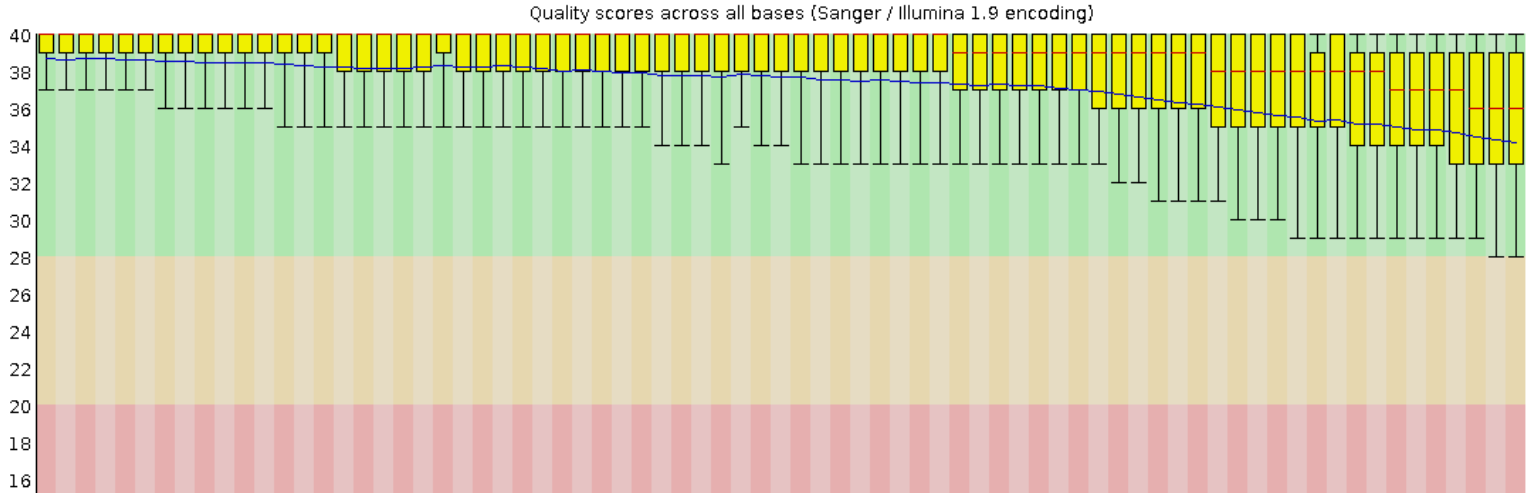
Basic Statistics

Measure	Value
Filename	WEx_Normal_R1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1400000
Sequences flagged as poor quality	0
Sequence length	75
%GC	53

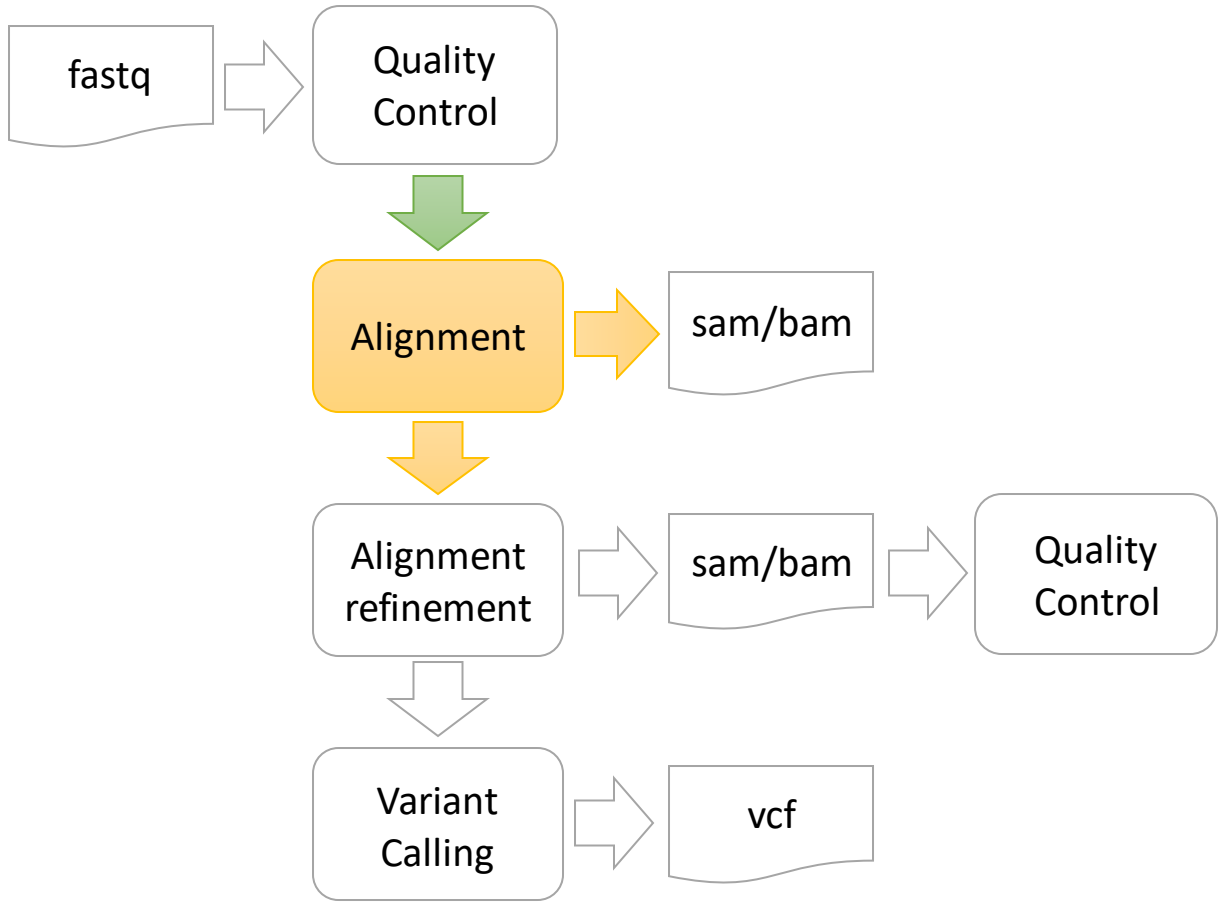
Phred+33 listed as Sanger/Illumina 1.9

Phred+64 listed as Illumina 1.5 (or lower)

Per base sequence quality



Alignment

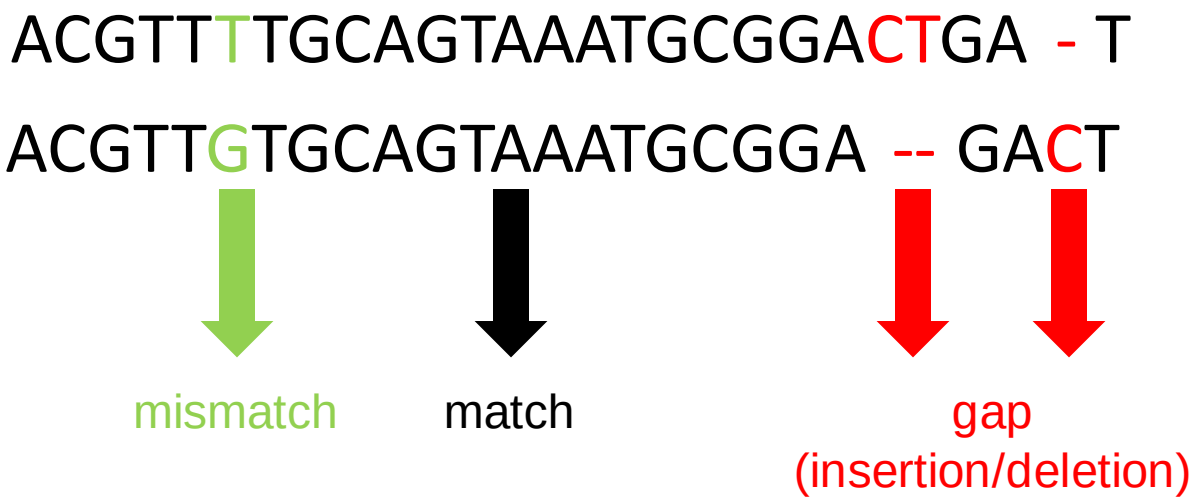

fastq
Quality Control
Alignment
sam/bam
Alignment refinement
sam/bam
Quality Control
Variant Calling
vcf

What is an alignment?

ACGTCTTGACTGG -TTAAAATAC
AC - TCTTGACTGGATTAACATAC

Sequence alignment is a way of arranging the sequences of DNA, RNA, or protein to **identify regions of similarity** that may be a consequence of functional, structural, or evolutionary relationships between the sequences.

Elements



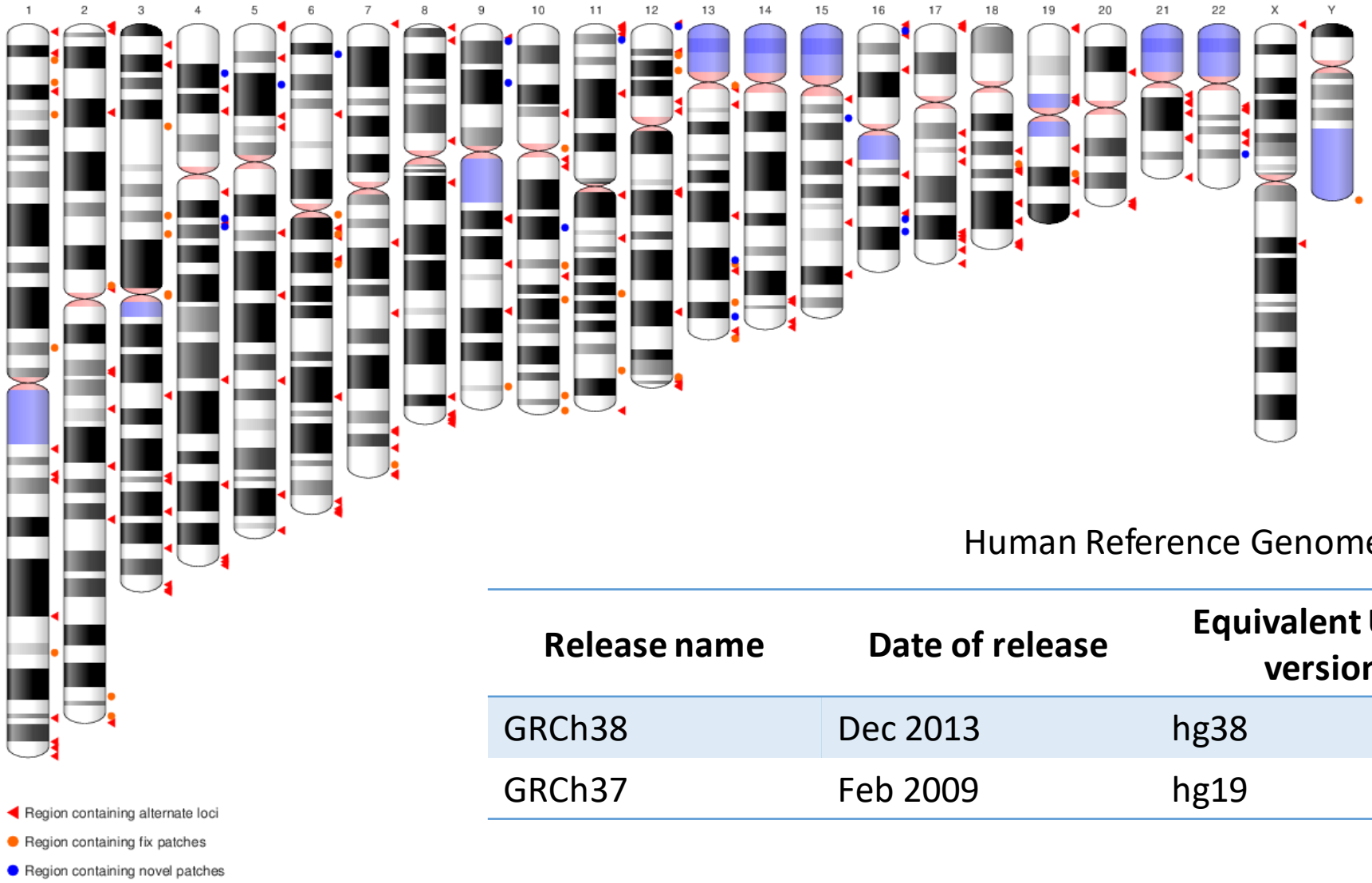
Alignment seeks to **reduce gaps and mismatches** and **maximize matches**.
In the construction, each of these components has a penalty value associated. For gaps there is a penalty value for opening the gap and another for extending it.

ACGTTTTGCAGTAAATGCGGACTGAT
ACGTTGTGCAGTAAATGCGGA-GACT
1 gap

ACGTTTTGCAGTAAATGCGGACTGAT
ACGTTGTGCAGTAAATGCGGA -- GACT
1 extended gap

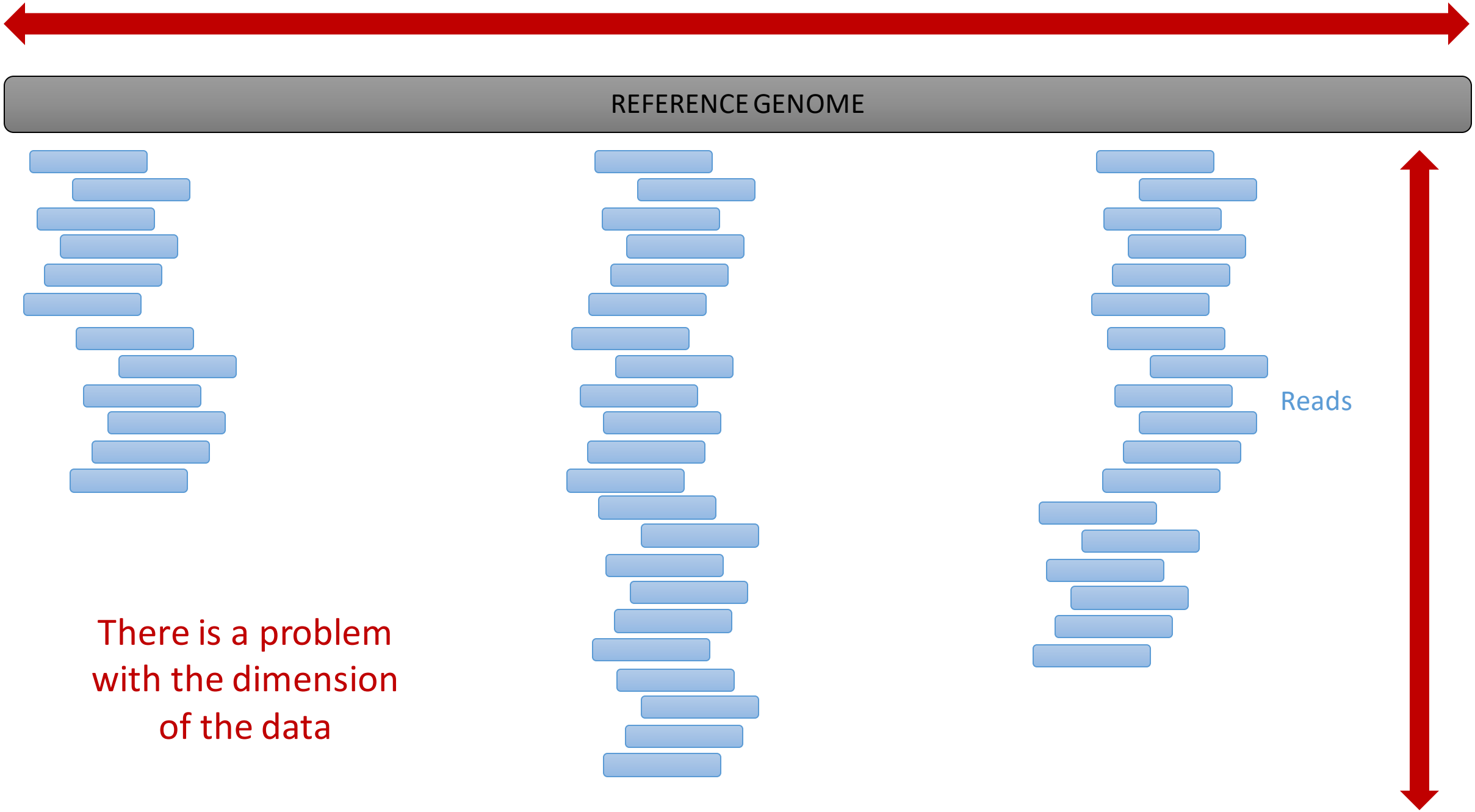
ACGTTTTGCAGTAAATGCGGACTGA -T
ACGTTGTGCAGTAAATGCGGA --GACT
2 gaps

Reference Genome



Human Reference Genome

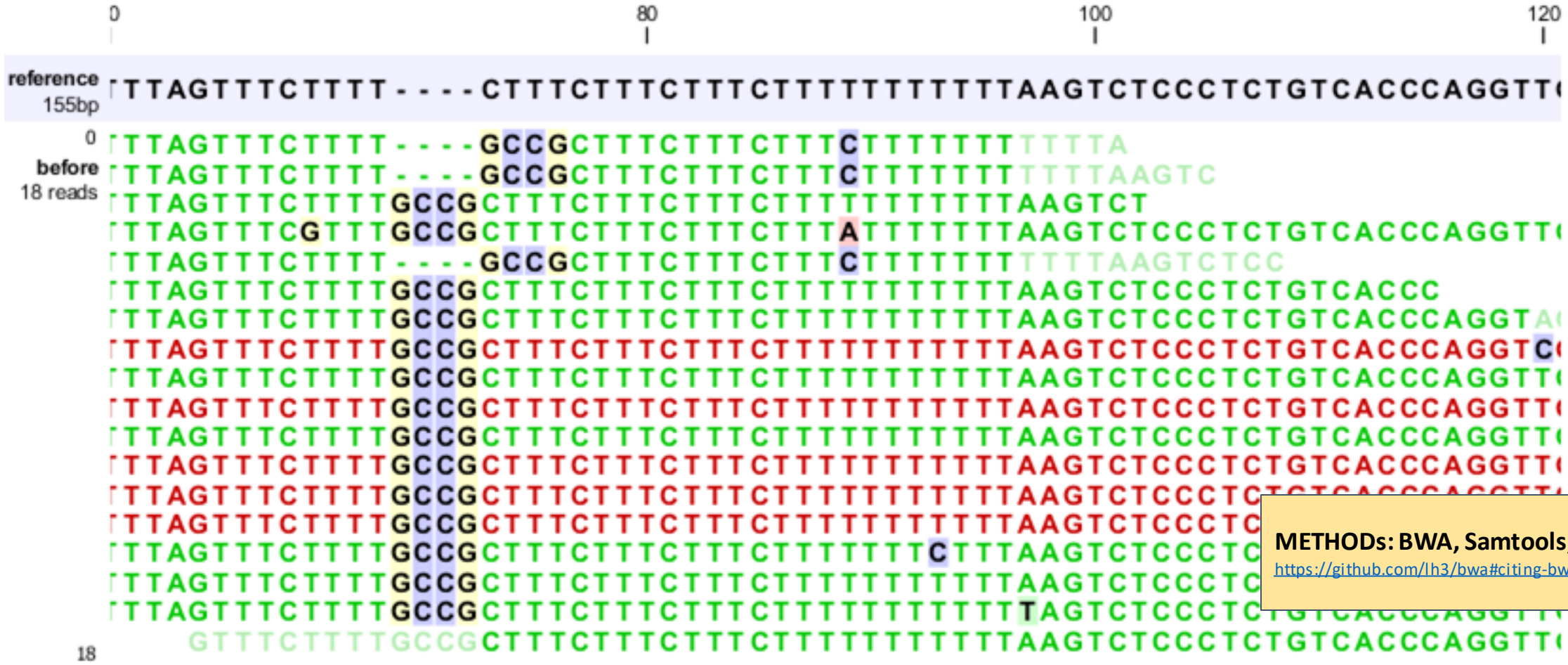
Release name	Date of release	Equivalent UCSC version	Base Pairs
GRCh38	Dec 2013	hg38	3,609,003,417
GRCh37	Feb 2009	hg19	3,326,743,047



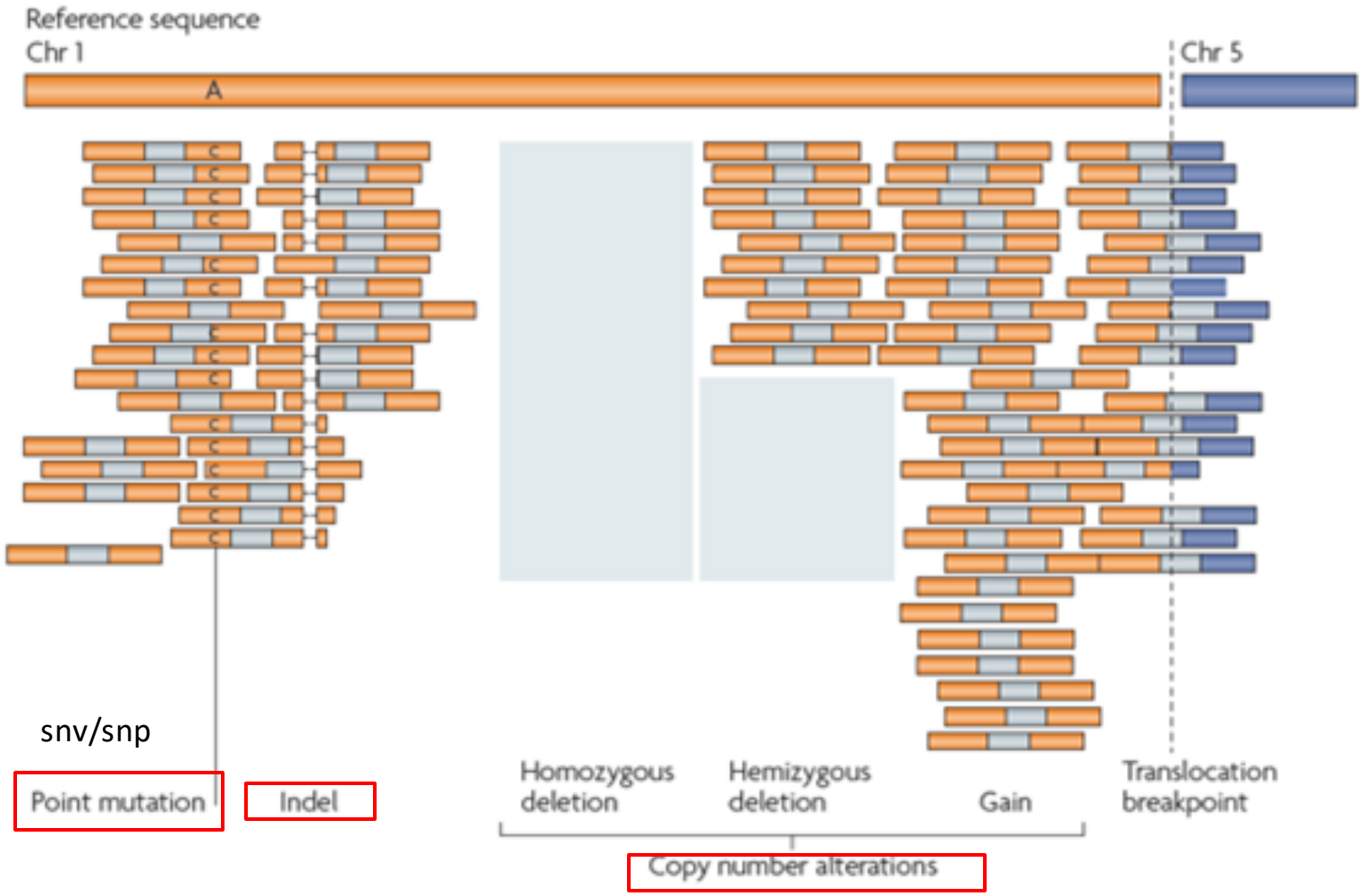
There is a problem with the dimension of the data

- Fast mapping on the reference genome by creating indexes. It is computationally intensive, but it is done only once.

- Search for candidate sites to align a given read by using seeds (fragments of a read).

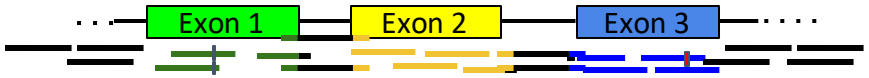


Different types of variants detected by mapping reads

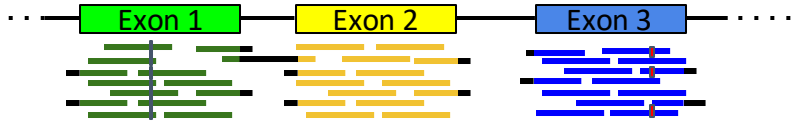


Strategies on DNA-seq

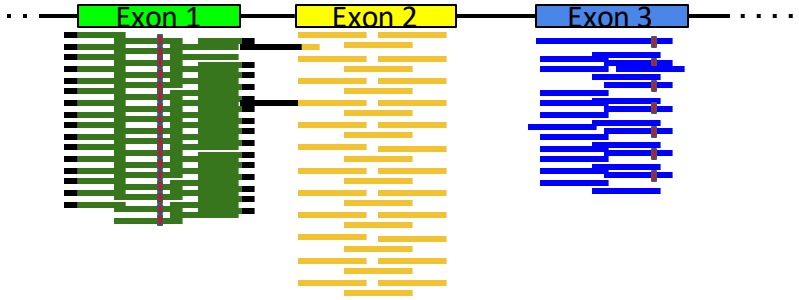
Whole-Genome Sequencing

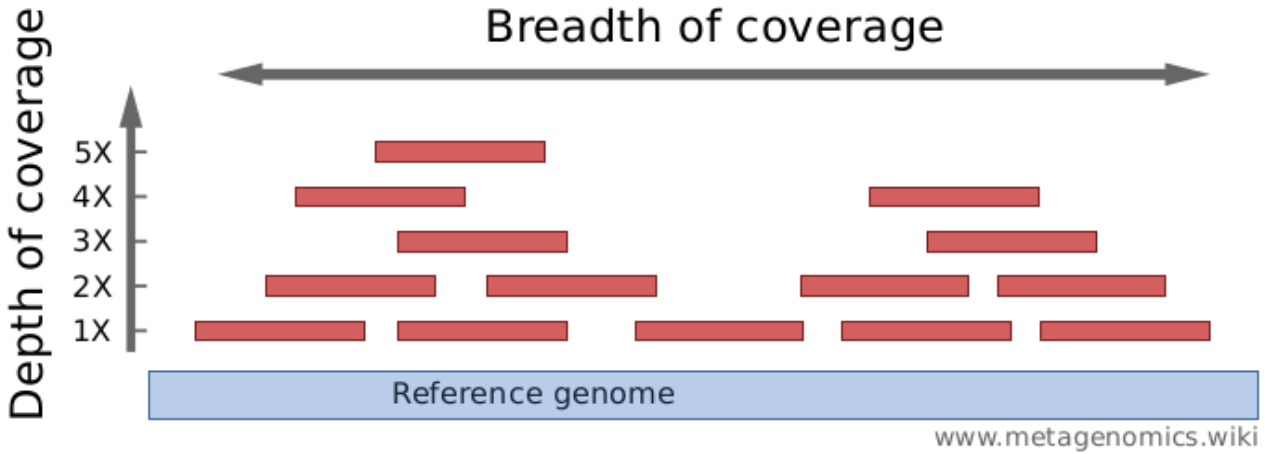


Whole-Exome Sequencing



Panel of genes by amplicon/PCR approach





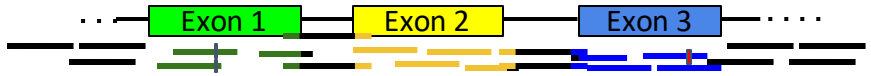
Depth: times that a base is sequenced (in average)

$$\frac{\text{Number of times a bases is covered by reads}}{\text{length of the sequenced genome}}$$

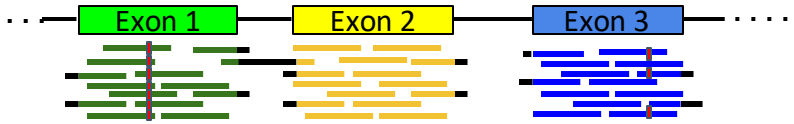
Breadth: percentage of the sequenced genome that is covered by the reads at a certain depth

Strategies on DNA-seq

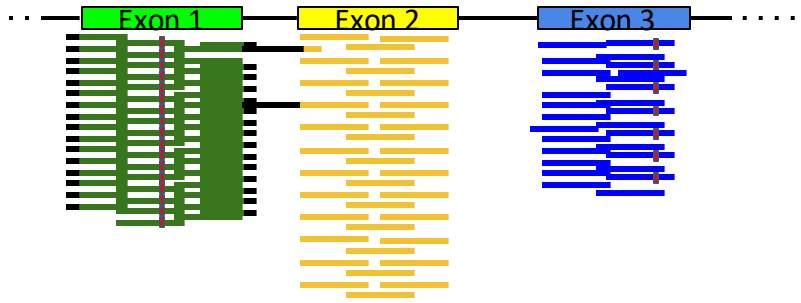
Whole-Genome Sequencing



Whole-Exome Sequencing



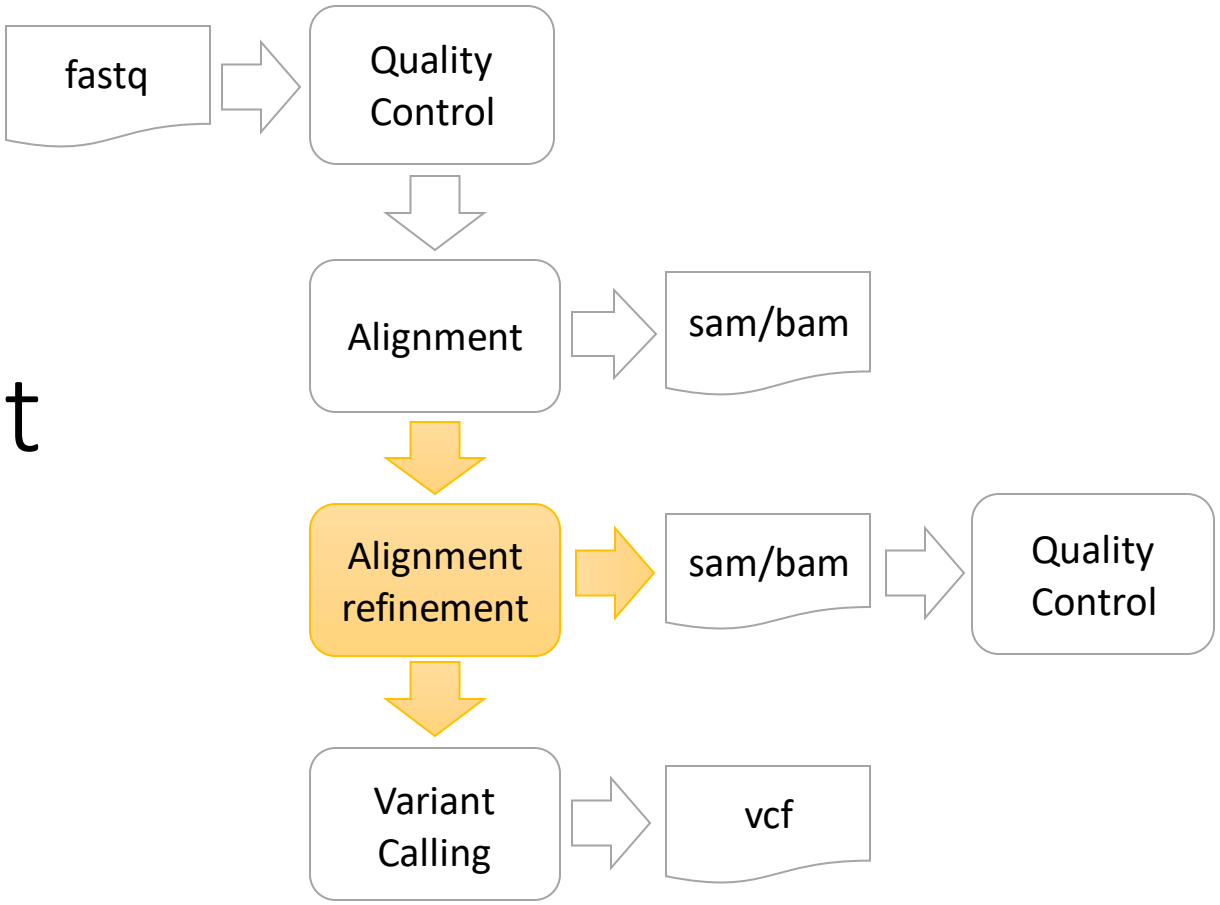
Panel of genes by amplicon/PCR approach



# bp pos seq	Type of variants discovered	Avg depth per pos	Cost
~100 Gb	- **coding variants***, intronic and regulatory sites. - **Structural variants** - **CNA** #Variants= 3M - 4M.	10x	High
~32Mb 50Mb	- **coding variants***. - Some intronic and regulatory sites. - **CNA** (challenging). **#**Variants= 20k - 60k.	20x - 80x	Low
ND	Depends on the design - Particular **coding variants*** - **CNA** (challenging) # variants = *ND*	1000x - 5000x	Low

***coding variants**: missense, stop gained, stop lost, frameshift, splice region...

Alignment refinement



Remove duplicates

- Duplicates derive from PCR amplification (library preparation): one fragment is sequenced multiple times.
- An error at the beginning of the PCR (first steps) is propagated.

- Therefore, duplicates are **worthless** for the analysis:
  Duplicates are source of False Positives calls while only provide redundancy.

Solution: retrieve the best one, discard the duplicates:

METHOD: by **Picard-tools**
http://broadinstitute.github.io/picard/
(alternatives : **samtools**)

Adapted from GATK

Duplicates share the **same alignment properties** : sequence, start and end positions

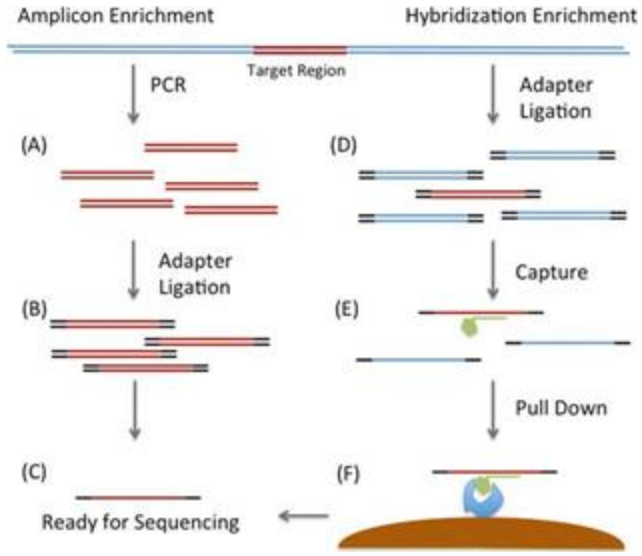
✖ = sequencing error propagated in duplicates

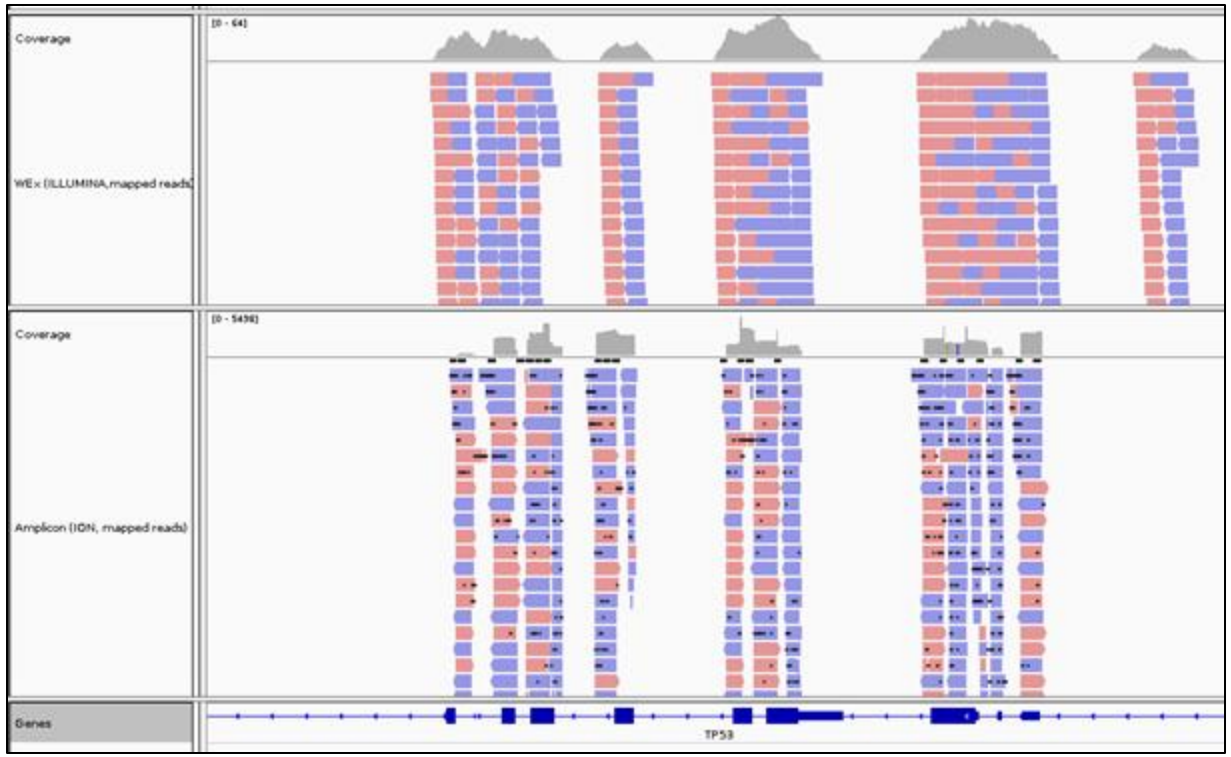
MARK DUPLICATES

After marking duplicates, the variant caller will only see :

… and thus be more likely to make the right call

Remove duplicates: Amplicon seq





WARNING: **Do NOT remove duplicates in data derived from** amplicon techniques (**Ion Torrent**).
More info.: http://gatkforums.broadinstitute.org/discussion/5847/remove-duplicates-from-targetted-sequencing-using-amplicon-approach

Indel realignment

- Algorithms align reads very fast with high accuracy, but not perfectly.

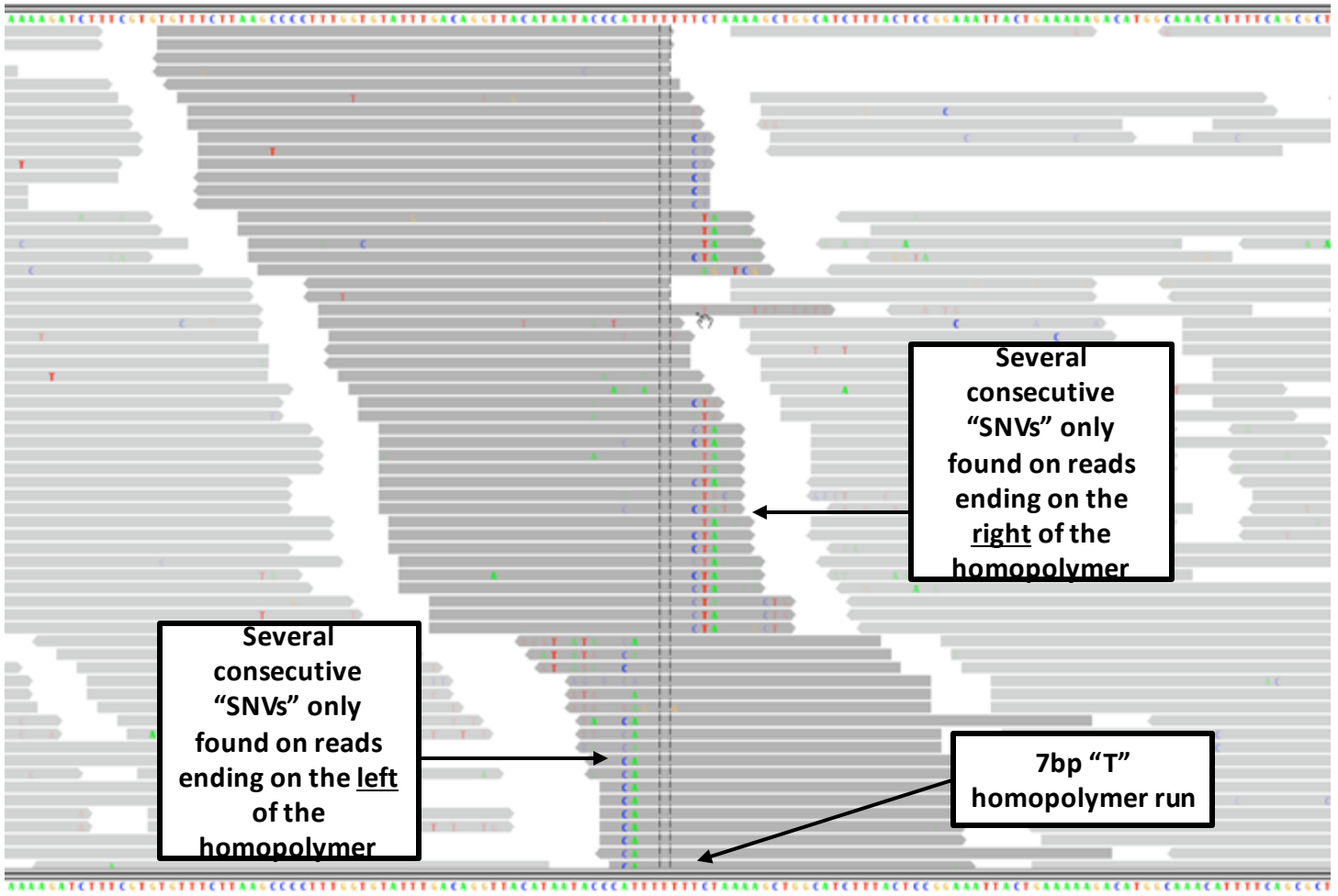
  During alignment, penalties on mismatches are much cheaper than gaps (indels).

- Also, there are sometimes multiple solutions (alignments) for a given read. Aligners choose one randomly.

- Variant calling requires the most perfect alignment as possible to avoid False Positives.

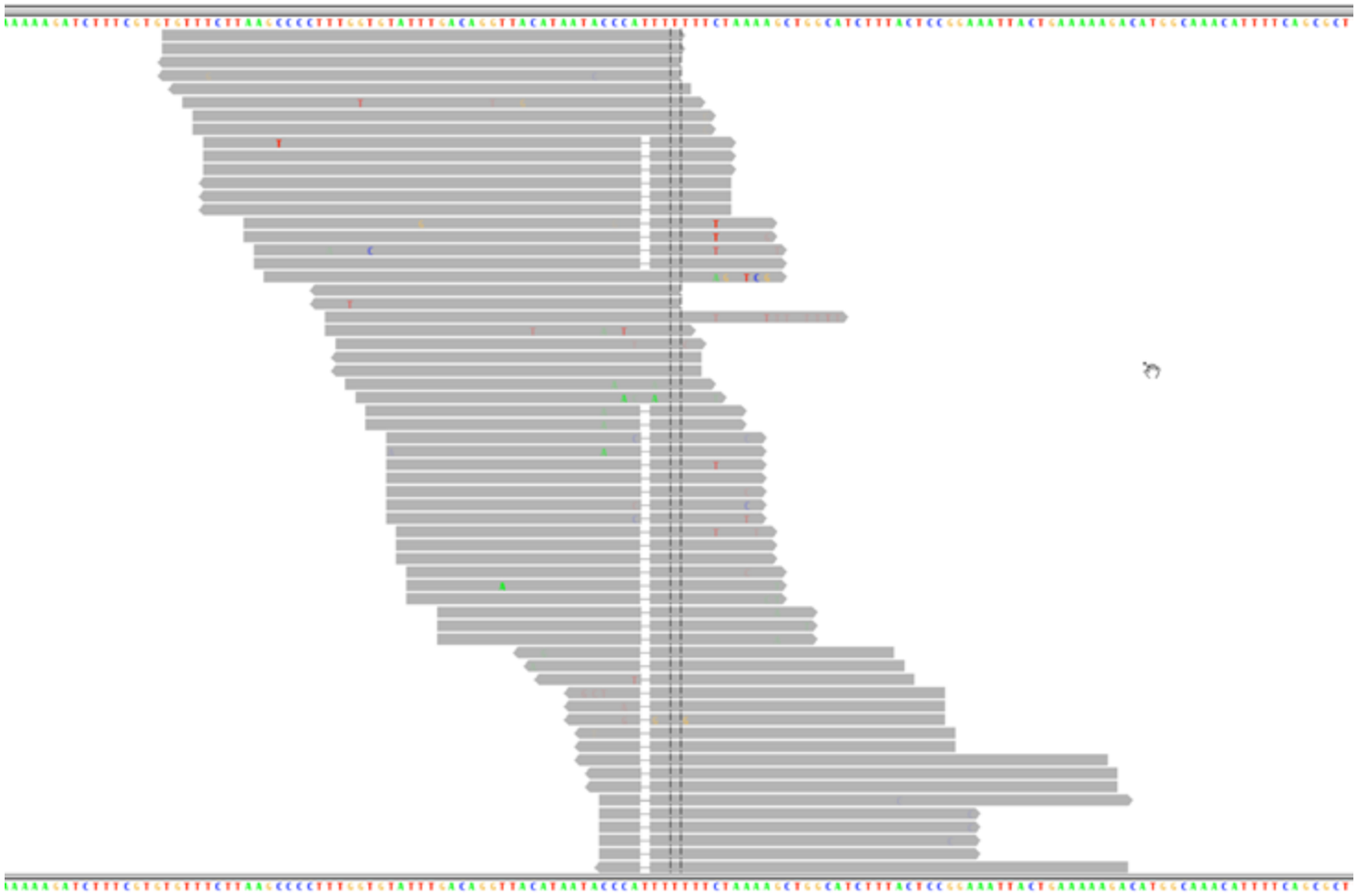
METHOD: by **GATK**
https://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_gatk_tools_walkers_indels_IndelRealigner.php

Indel realignment



Taken from GATK team

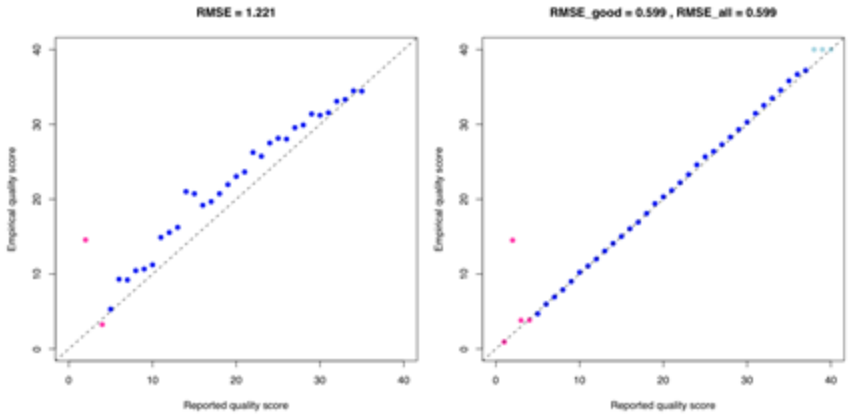
Indel realignment



Taken from GATK team

Base Quality Score Recalibration

- **Phred Quality score**: each position of the sequence has its particular **base Quality score**.
- The individual quality measures are crucial during Variant calling.
- Different NGS technologies have their particular **bias in Quality Score** depending on the context. Recalibration **correct empirically** these biases.



Original After BQSR recalibration

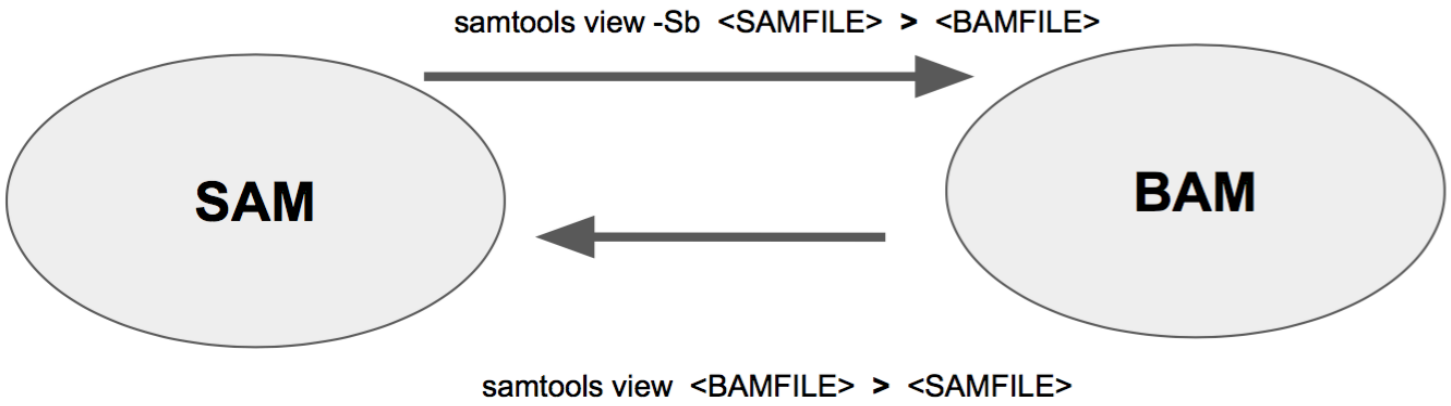
METHOD: by **GATK**
http://gatkforums.broadinstitute.org/discussion/44/base-quality-score-recalibration-bqsr

Alignment file

sam & bam

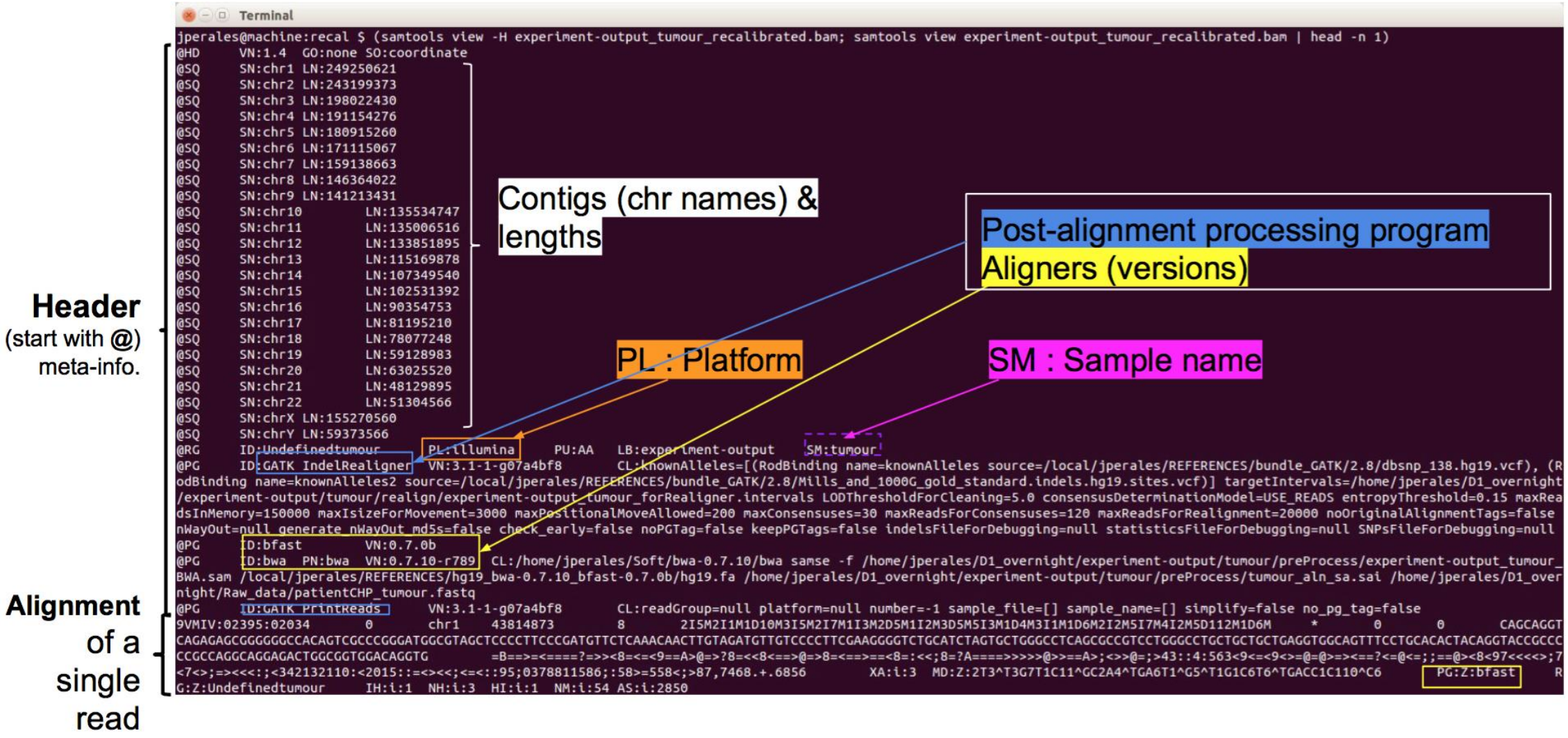
Alignment - SAM/BAM

- **SAM** is the human readable text format (.sam extension)
- **BAM** is the binary, machine efficient format (.bam extension)
- Both contains exactly the same information and are interconvertible (samtools)



File specifications: https://samtools.github.io/hts-specs/SAMv1.pdf

Alignment - SAM/BAM - Header



Alignment - SAM/BAM - Alignments

#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12
ReadName	99	chr10	2	30	3MD2M1I1M	=	14	20	CATCTG	jjjjjjj	z:Aligner

If single-end:
7. reference sequence name of the alignment of the next read in sequence
8. position in the alignment of the next read in sequence
9. number of bases covered by reads from the same fragment. Plus/minus means the current read is the leftmost/rightmost read

FIELDS

#Col	Field	Description
1.	QNAME	read name
2.	FLAG	bitwise FLAG* (unmapped, pair unmapped, properly mapped, …)
3.	RNAME	Reference sequence name (e.g. chr1).
4.	POS	1-based leftmost position.
5.	MAPQ	Mapping Quality (Phred-scaled). Scale 0 to 255.
6.	CIGAR	extended CIGAR string.
7.	MRNM	Paired-end: Mate Reference sequence Name (= if same as RNAME).
8.	MPOS	Paired-end: 1-based Mate position.
9.	TLEN	Paired-end: Insert size
10.	SEQ	Read sequence
11.	QUAL	Base Quality Score from the Read sequence.
12.	OPT	Optional Tags.

SAM FLAGS

Decimal	Description of read
1	Read paired
2	Read mapped in proper pair
4	Read unmapped
8	Mate unmapped
16	Read reverse strand
32	Mate reverse strand
64	First in pair
128	Second in pair
256	Not primary alignment
512	Read fails platform/vendor quality checks
1024	Read is PCR or optical duplicate
2048	Supplementary alignment

One of the reads is unmapped:
73, 133, 89, 121, 165, 181, 101, 117, 153, 185, 69, 137

Both reads are unmapped:
77, 141

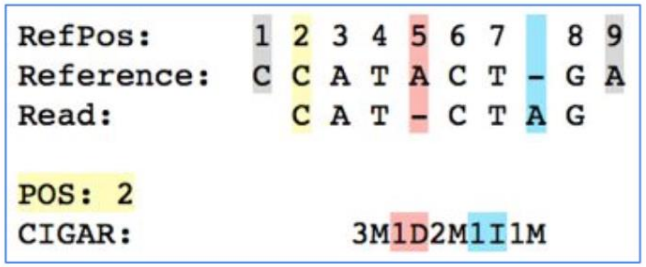
Mapped within the insert size and in correct orientation:
99, 147, 83, 163

Mapped within the insert size but in wrong orientation:
67, 131, 115, 179

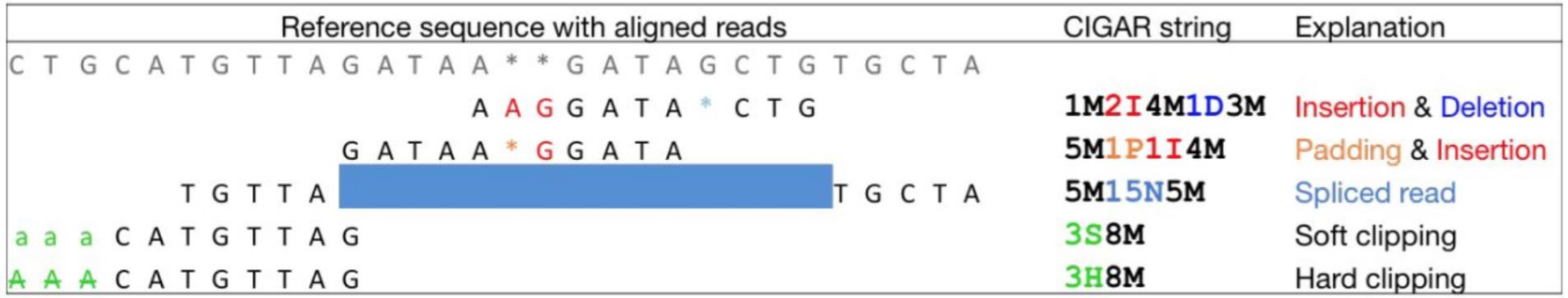
Mapped uniquely, but with wrong insert size:
81, 161, 97, 145, 65, 129, 113, 177

Alignment - SAM/BAM - CIGAR

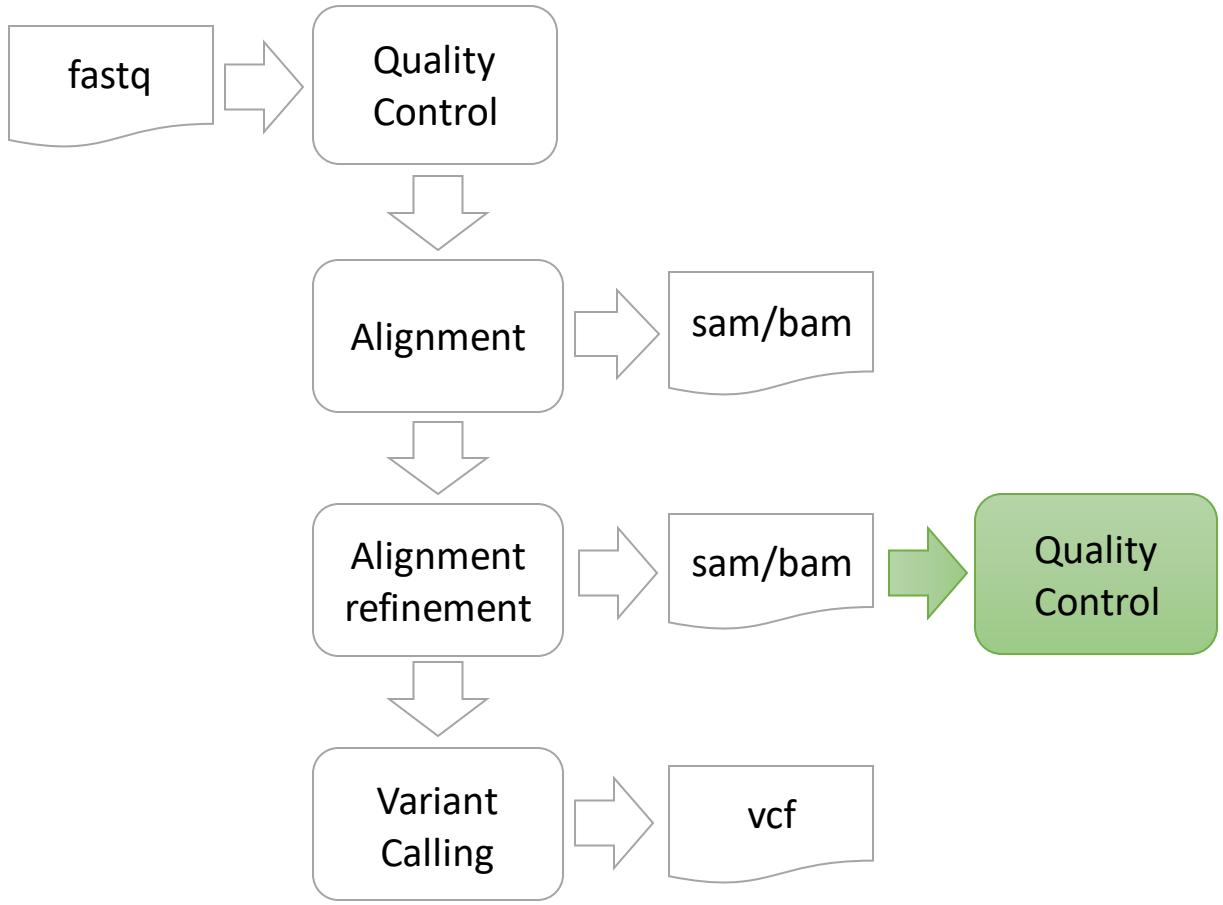
- *Concise Idiosyncratic Gapped Alignment Report*
- It is a compressed representation of an alignment
- **Format**: A CIGAR string is made up of <integer><op> pairs
- Here, "op" is an operation specified as a single character, usually an upper-case letter (see table)



Op	Description
M	alignment match (can be a sequence match or mismatch)
I	insertion to the reference
D	deletion from the reference
N	skipped region from the reference
S	soft clipping (clipped sequences present in SEQ)
H	hard clipping (clipped sequences NOT present in SEQ)
P	padding (silent deletion from padded reference)
=	sequence match
X	sequence mismatch



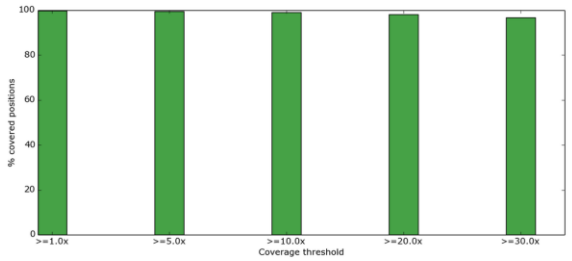
Alignment Quality Control

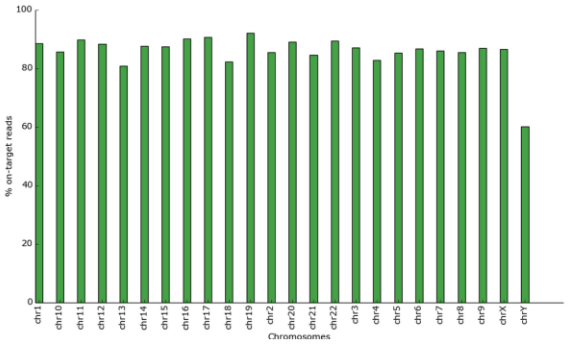


Alignment Quality Control

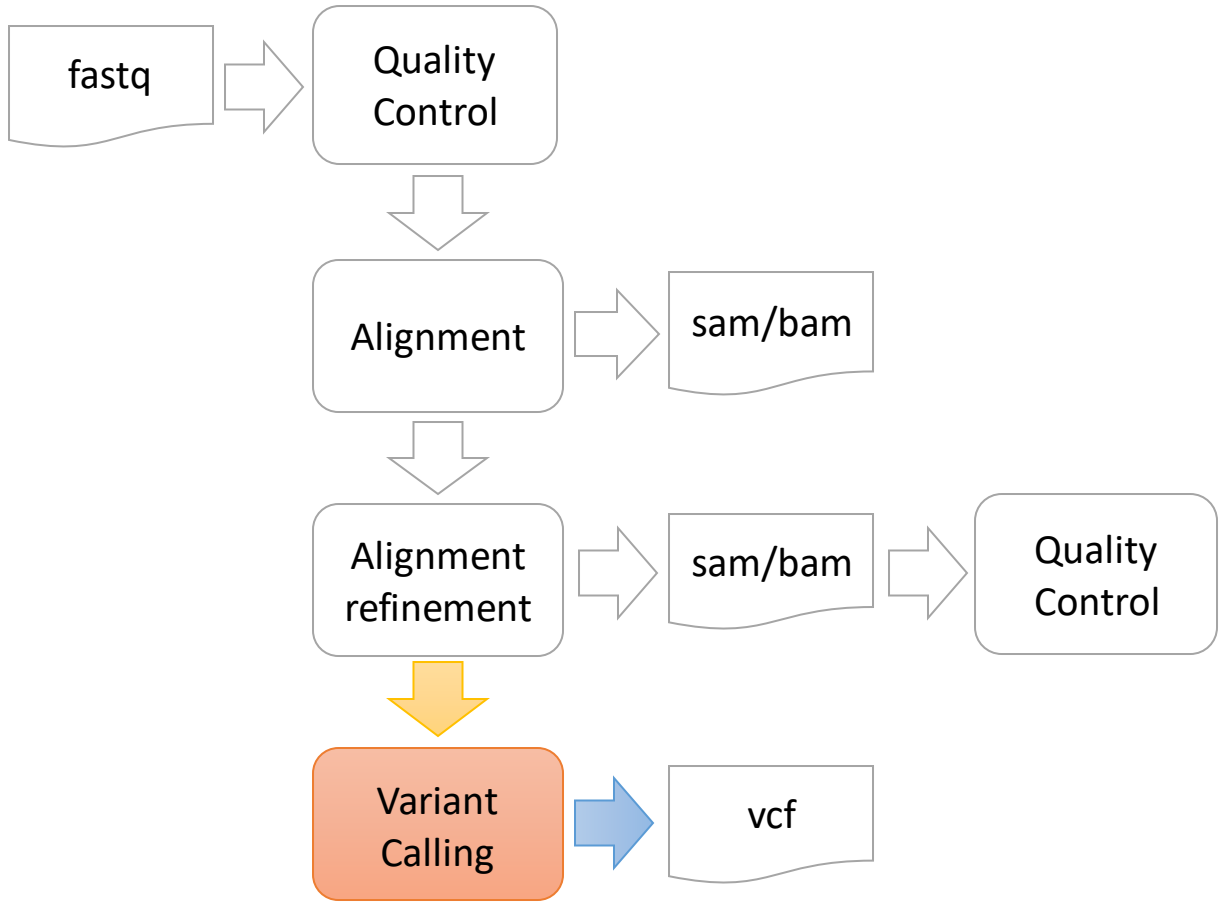
- Mean sequencing depth
- Is there enough coverage in regions of interest?
- Are the reads on-target?

- Software:
  - ngsCAT
  - QualiMap



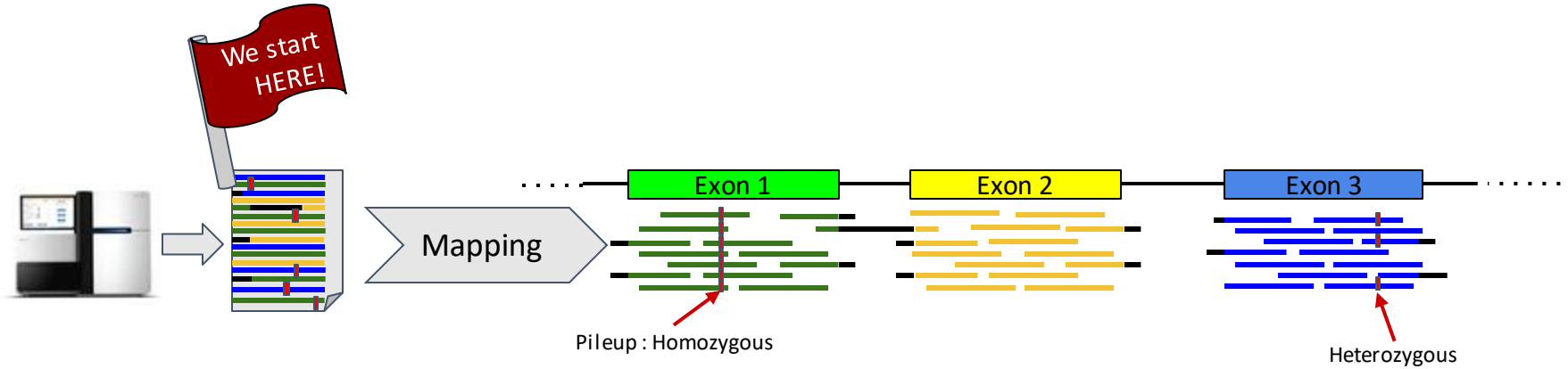


Variant calling



Fundamentals of Variant Calling

- How we detect genetic variants ?



1

Identify the most likely **genotype for each genomic position from the individual** using statistical methods.

2

Identify the differences by comparing with the reference genome.

What is Crucial in Variant calling

- For clinical practices, the use of **gold standard methods** and **reproducible analysis** are mandatory.
- The analysis is based on the comparison against the **reference genome**:

A single consensus sequence for the whole genome. It was built up from a high quality set of representative samples of the specie (from different populations).
It is the first-line comparison during analysis.

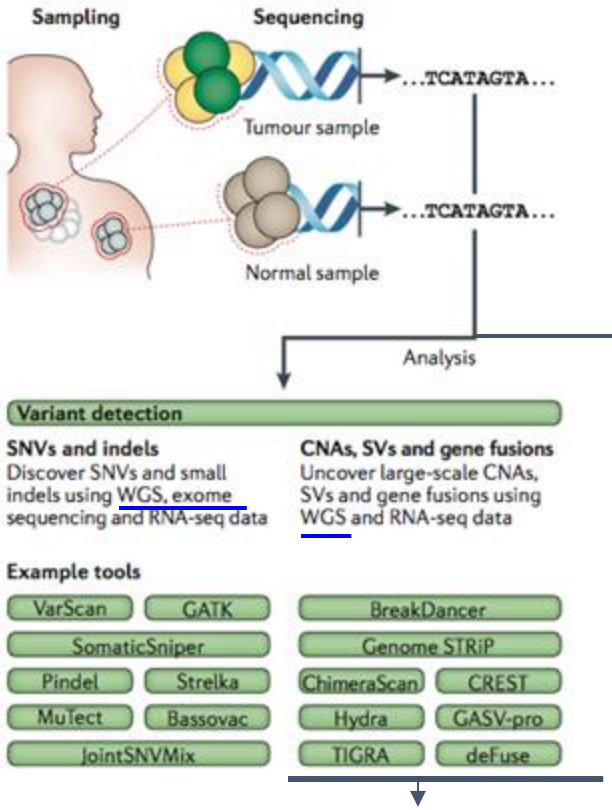
By **Genome Reference Consortium (GRC)** (http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/)

- **Human assemblies (Versions):**
  - **GRCh37/hg19** : former version. Released in 2012. It is still used for analysis.
  - **CRCh38/hg38** : current version (Sep. 2017). Released in 2014. More accurate, comprehensive (includes Haplotypes) and sophisticated.

We must keep consistency in the Genome Reference Version through the variant analysis.

- We must know what **regions along the genome were sequenced** in the experiment, that is, the sequencing library.

Algorithms for Variant Calling



Several Methods have been published.

More on CNAs

Tool	Year	Language	Paired or pooled data	Segmentation	Feature
ADTEx	2014	Python, R	Both	HMM	Noise reduction Ploidy estimation
CONTRA	2012	Python, R	Both	CBS	GC correction
Control-FREEC	2011	C++, R	Paired	LASSO	GC correction, mappability
EXCAVATOR	2013	Perl, R	Both	HSLM	GC correction, mappability, exon-size correction
ExomeCNV	2011	R	Paired	CBS	GC correction, mappability
Varscan2	2012	Java, Perl, R	Paired	CBS	GC correction

Only WGS

Appropriate methods for Whole-Exome seq

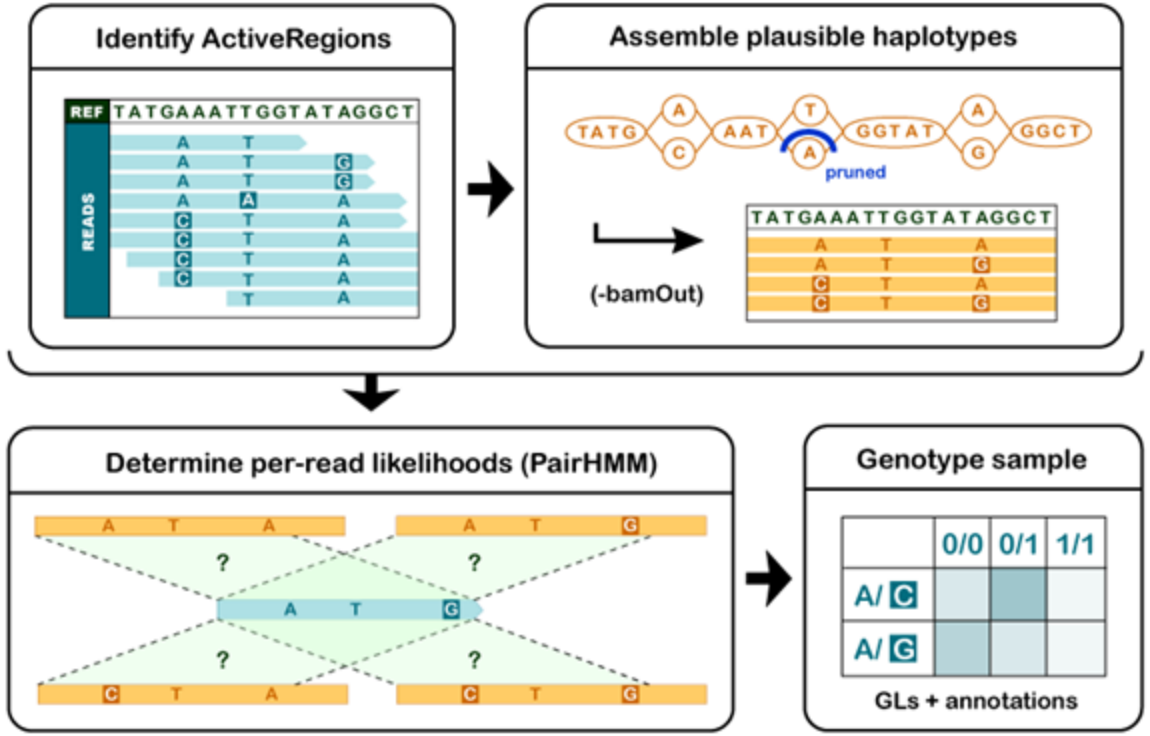
Further reading:

Ding, L., Wendl, M. C., McMichael, J. F. & Raphael, B. J. Nat Rev Genet – (2014). doi:10.1038/nrg3767

Nam J.N. et al. Evaluation of somatic copy number estimation tools for whole-exome sequencing data. Brief. Bioinformatics (2015)

Variant Calling for SNVs and Indels

Haplotype Caller : Variant calling based on the calculation of genotype likelihoods:



Assumptions: It bases the calling in the indicated ploidy (e.g. 2n)
Limited detection of low allele frequencies.

Further reading:
http://gatkforums.broadinstitute.org/discussion/4148/hc-overview-how-the-haplotypecaller-works
HC steps 1-4: https://software.broadinstitute.org/gatk/documentation/topic?name=methods

Variant Calling for somatic variants: MuTect

SNV and Indel caller**.**

Similar logic to Haplotype Caller but:

- It **allows variable allele frequencies**.
- It **includes logic to avoid germline variants**.

CGA
Cancer Genome Analysis

Login

Search this site:

CGA
- Home
- Help Forum
- ABSOLUTE
- BreakPointer
- ChainFinder
- ContEst
- Copynumber Pipeline
- D-TrackG
- dRanger
- Firehose
- GISTIC
- HAPSEG
- Indelocator
- IntVEx
- MuTect
  - How do I get MuTect?
  - How do I run MuTect?
  - How do I get help?
- MutSig
- Oncotator
- PathSeq
- RNA-SeQC
- SegSeq
- TranspoSeq

BROAD INSTITUTE
© 2013 Broad Institute

For site help, please contact cga-webmaster@broadinstitute.org

Home »

MuTect

What is MuTect?

MuTect is a method developed at the Broad Institute for the reliable and accurate identification of somatic point mutations in next generation sequencing data of cancer genomes.

For complete details, please see our publication in Nature Biotechnology:

Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnology (2013).doi:10.1038/nbt.2514

How does it work?

In brief, muTect consists of three steps.

1. Preprocessing the aligned reads in the tumor and normal sequencing data. In this step we ignore reads with too many mismatches or very low quality scores since these represent noisy reads that introduce more noise than signal.
2. A statistical analysis that identifies sites that are likely to carry somatic mutations with high confidence. The statistical analysis predicts a somatic mutation by using two Bayesian classifiers – the first aims to detect whether the tumor is non-reference at a given site and, for those sites that are found as non-reference, the second classifier makes sure the normal does not carry the variant allele. In practice the classification is performed by calculating a LOD score (log odds) and comparing it to a cutoff determined by the log ratio of prior probabilities of the considered events. For the tumors we calculate

$$LOD_T = \log_{10}\left(\frac{P(\text{observed data in tumor}\,|\,\text{site is mutated})}{P(\text{observed data in tumor}\,|\,\text{site is reference})}\right)$$, and for the normal $$LOD_N = \log_{10}\left(\frac{P(\text{observed data in normal}\,|\,\text{site is reference})}{P(\text{observed data in normal}\,|\,\text{site is mutated})}\right)$$

Since we expect somatic mutations to occur at a rate of ~1 in a Mb, we require $LOD_T > \log_{10}(0.5\times10^{-6}) = 6.3$ which guarantees that our false positive rate, due to noise in the tumor, is less than half of the somatic mutation rate. In the normal, not in dbSNP sites, we require $LOD_N = \log_{10}(0.5\times10^{-3}) = 2.3$ since non-dbSNP germline variants occur roughly at a rate of 100 in a Mb. This cutoff guarantees that the false positive somatic call rate, due to missing the variant in the normal, is also less than half the somatic mutation rate.

3. Post-processing of candidate somatic mutations to eliminate artifacts of next-generation sequencing, short read alignment and hybrid capture. For example, sequence context can cause hallucinated alternate alleles but often only in a single direction. Therefore, we test that the alternate alleles supporting the mutations are observed in both directions.

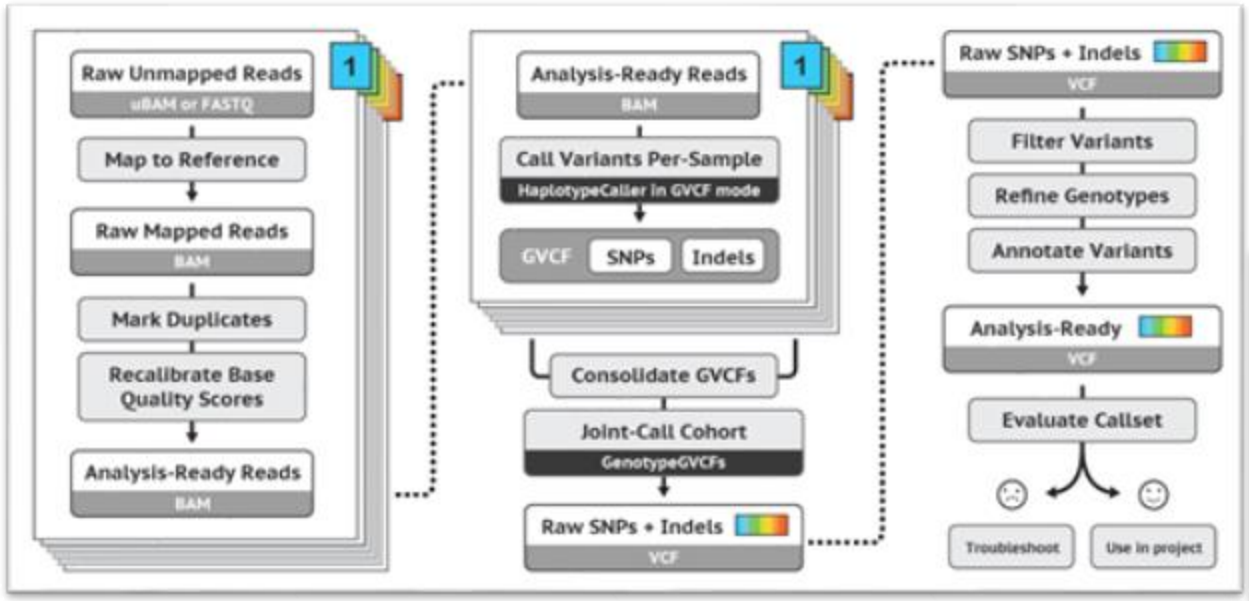
As muTect attempts to call mutations it also generates a coverage file (in a wiggle file format, which indicates for every base whether it is sufficiently covered in the tumor and normal to be sensitive enough to call mutations). We currently use cutoffs of at least 14 reads in the tumor and at least 8 in the normal (these cutoffs are applied after removing noisy reads in the preprocessing step). In addition, wiggle files can also be generated of the observed depth in the tumor and in the normal.

Most cancer genome studies at the Broad Institute have made use of MuTect and have validated the mutation calls as a part of their cancer biology papers, showing that MuTect has a very low false positive rate. A summary of validation rates from these papers are show below:

publication	technology	candidates	validated	no result	validation rate
Multiple Myeloma[1]	Sequenom	97	92	5	94.85%
Ovarian[2]	Sequenom/PCR/454	1655	1483	172	89.61%
Ovarian[3]	Capture/Illumina	6497	6232	265	95.92%
Head and Neck[4]	Sequenom	321	288	33	89.72%
Breast[5]	Sequenom/PCR/454	455	428	0	94.07%

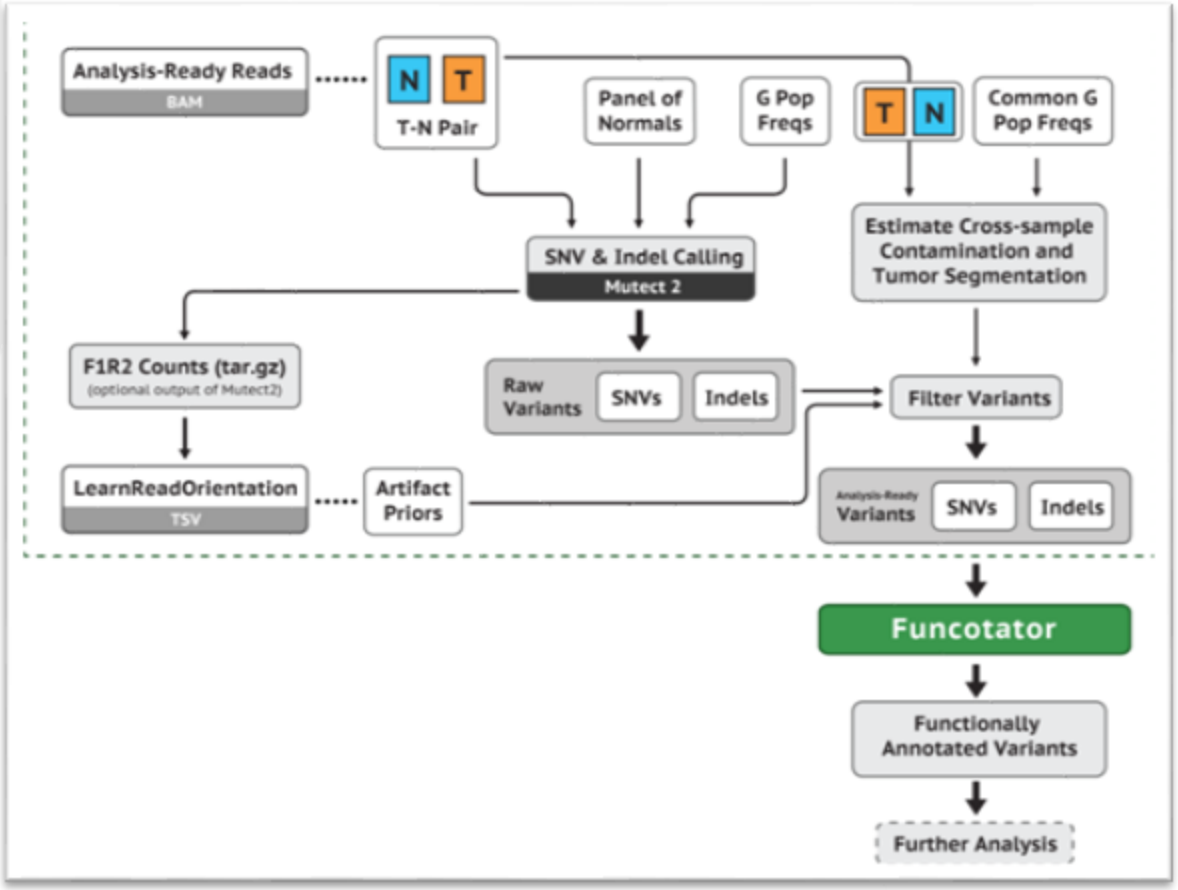
Cibulskis, K. et al.
Nat Biotechnology (2013).doi:10.1038/nbt.2514

GATK Best Practices



Germline small-scale variants

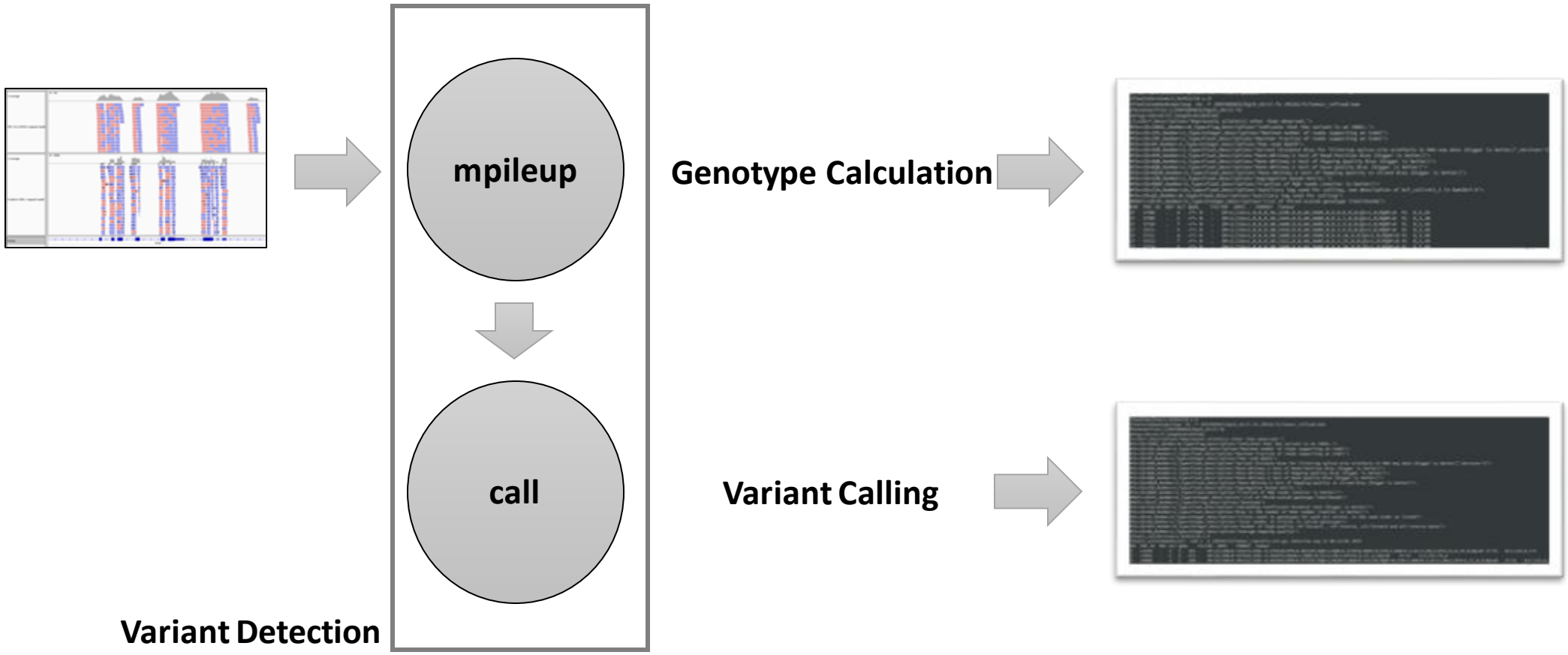
Somatic small-scale variants



bcftools

http://samtools.github.io/bcftools/

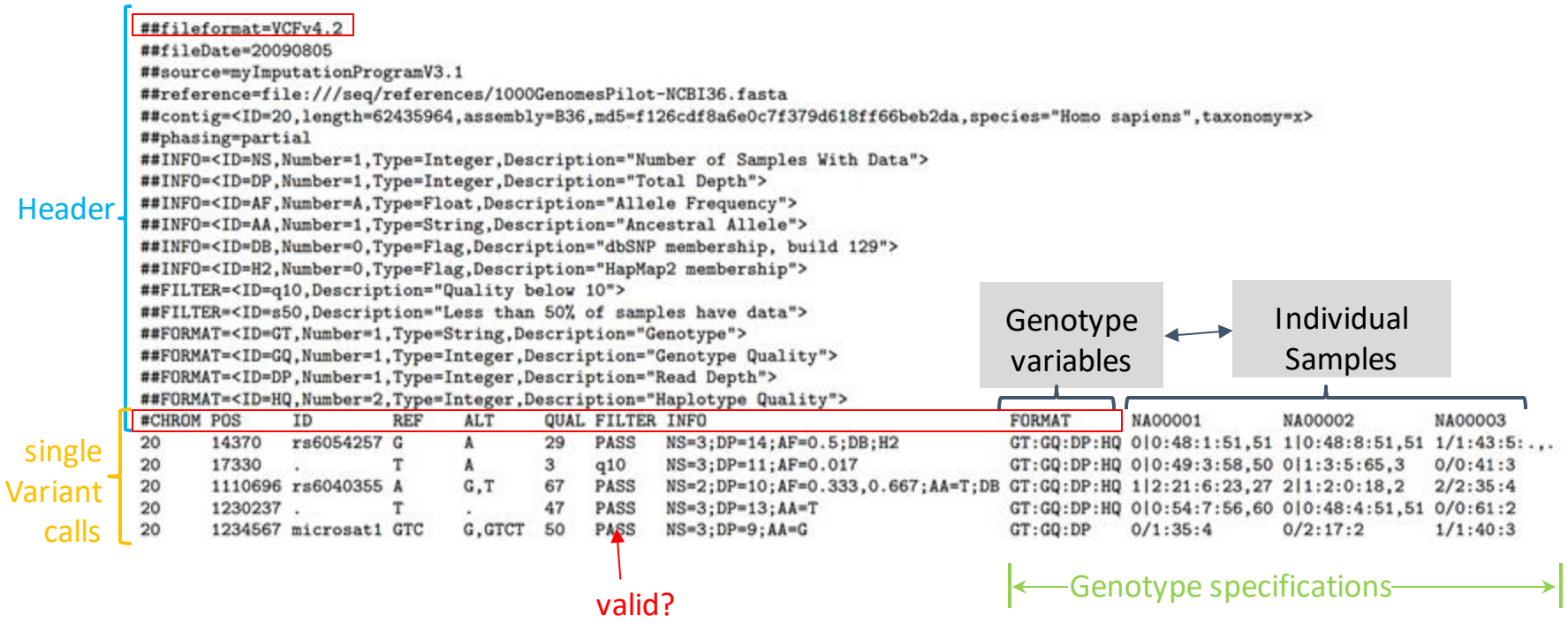
Set of tools to call variants and manage VCF files.



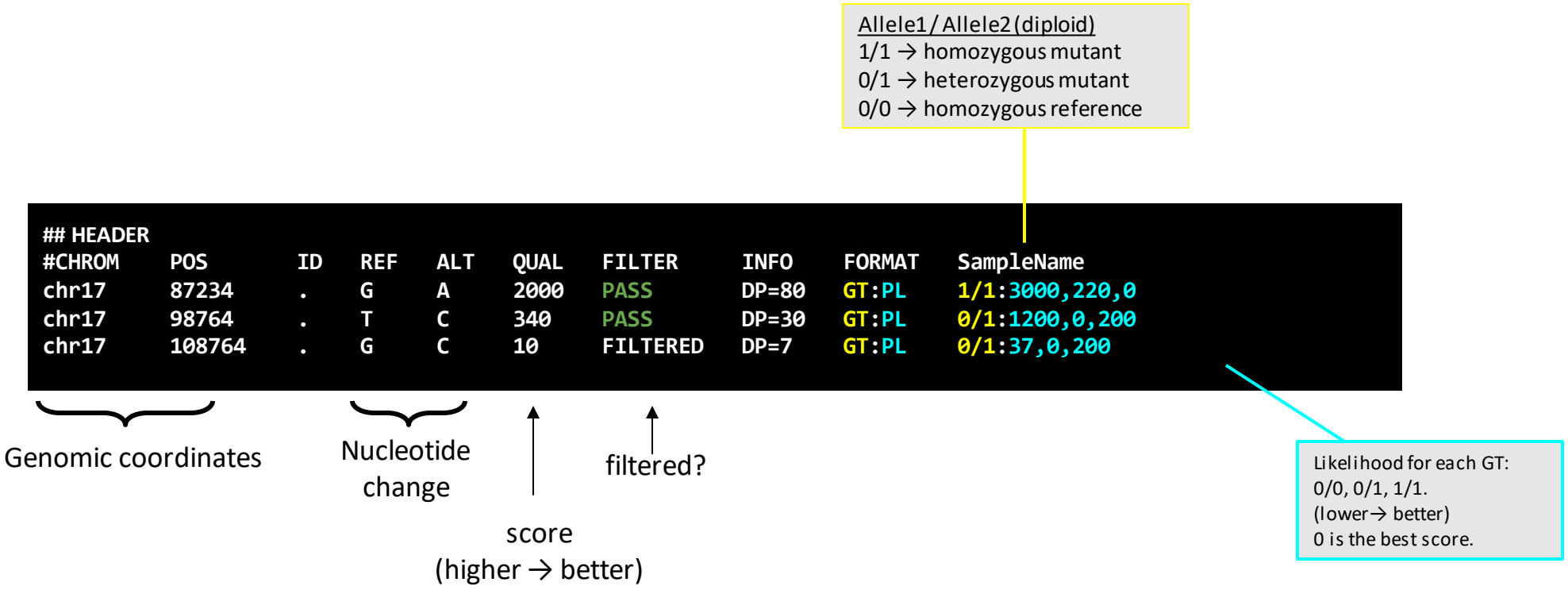
Variant's file

vcf

Variant Calling Format - VCF



- Typical extension: .vcf (.bcf binary counterpart)
- Not all records in a VCF are true calls, the FILTER column specifies those which passed the calling
- QUAL is the score assigned to a given call. The greater QUAL is, the more reliable is. It is in log-scale



More info.:
https://samtools.github.io/hts-specs/VCFv4.2.pdf
https://www.broadinstitute.org/gatk/guide/article?id=1268

Variant Detection

Hands on

Data

- **Input** : FastQ files.
- **Output vcf**: variants in tumor sample, in normal sample and in both samples
- Raw files: https://fundacioncnio-my.sharepoint.com/:u:/g/personal/epineiro_cnio_es/EaWrloqPdb5GjXxak1O4TzkBUmlI6sqnupz8SLuJ2aT5-A?e=hv0Nss

Patient suffering ovarian cancer.

Whole-exome sequencing data from two samples from the patient:

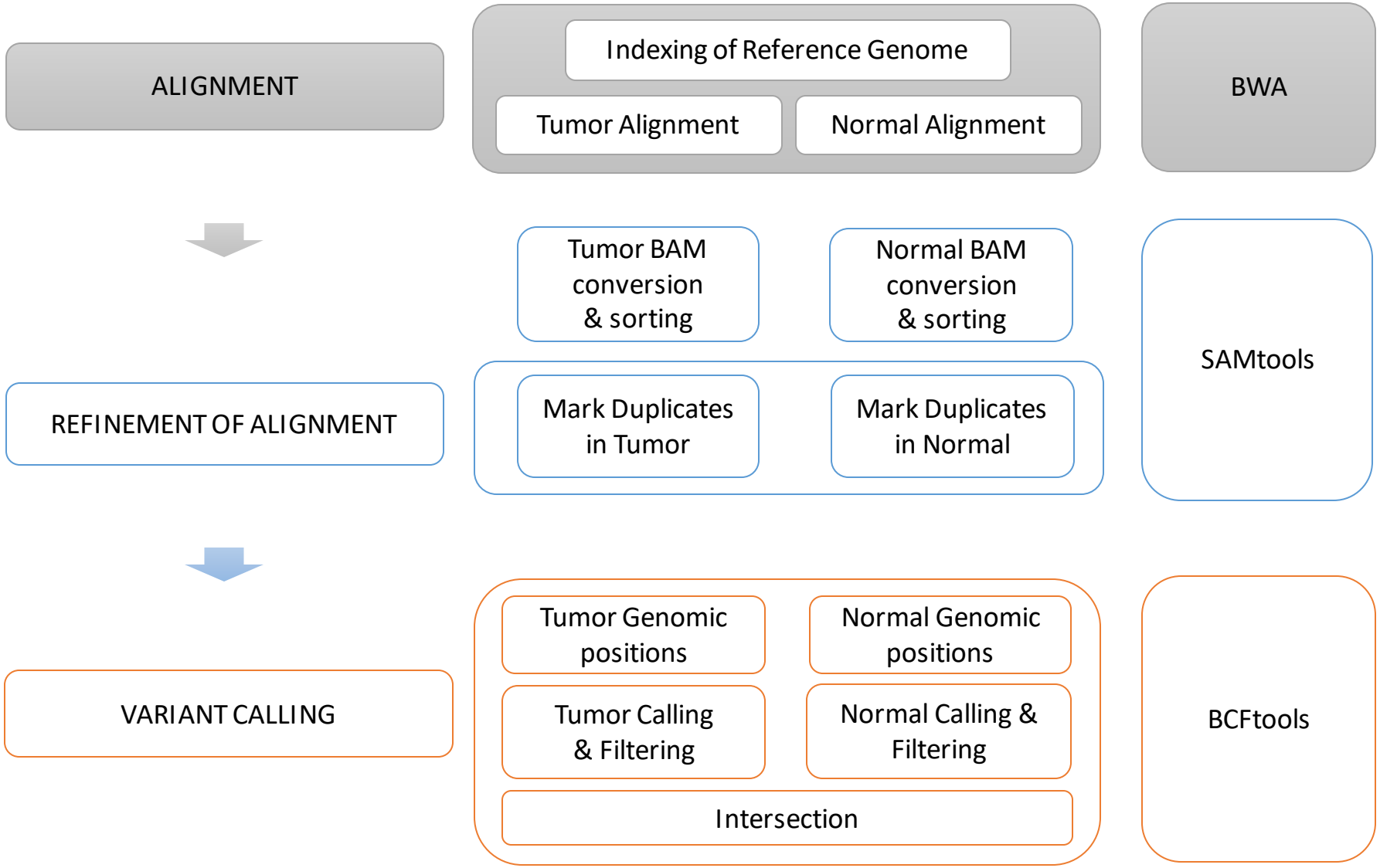
- Tumour sample.
- Matched normal sample (healthy tissue) from epithelium.

Library protocol: Agilent SureSelect V5 Human All Exons.

Sequencing platform: HiSeq 2000 (Illumina)

NOTE: This data was simulated and reduced in order to perform the computational analysis in class time.

ROAD MAP


ALIGNMENT
Indexing of Reference Genome
Tumor Alignment
Normal Alignment
BWA
REFINEMENT OF ALIGNMENT
Tumor BAM conversion & sorting
Normal BAM conversion & sorting
Mark Duplicates in Tumor
Mark Duplicates in Normal
SAMtools
VARIANT CALLING
Tumor Genomic positions
Normal Genomic positions
Tumor Calling & Filtering
Normal Calling & Filtering
Intersection
BCFtools

Step 1. Index the human reference genome

Reference genome: https://fundacioncnio-my.sharepoint.com/:u:/g/personal/epineiro_cnio_es/ETg0aiGbVzdChTX2VNv3PZ0BJMOmovsLtaDbMSbMoR6i8A?e=1ymXkA

```
> docker pull osvaldogc/ufv:2.0 1

> SAMPLES_LOCAL=/Path_directory_raw_data/
> SAMPLES_DOCKER=/SAMPLES
> REFERENCE_LOCAL=/Path_to_reference_fasta/ 2
> REFERENCE_DOCKER=/REFERENCE
> RESULTS_LOCAL=/Path_directory_RESULTS
> RESULTS_DOCKER=/RESULTS

> chmod 777 $REFERENCE_LOCAL 3
> chmod 777 $RESULTS_LOCAL

> docker run --rm  -v $SAMPLES_LOCAL:$SAMPLES_DOCKER -v $RESULTS_LOCAL:$RESULTS_DOCKER -v 4
$REFERENCE_LOCAL:$REFERENCE_DOCKER -it osvaldogc/ufv:2.0 /bin/bash
```

Step 1. Index the human reference genome

```
> bwa index /REFERENCE/hg19_chr17.fa
```

Indexing the human reference genome

Note:
- The indexing of the reference genome must be done only the first time your are going to perform the alignment of any sample on it.
- During the indexing, several files are generated with the same prefix as the fasta file (hg19_chr17.fa.*)

Check that those files were generated:

```
> ls -l /REFERENCE
```

Check the creation of the index files

hg19_chr17.fa.amb
hg19_chr17.fa.ann
hg19_chr17.fa.bwt
hg19_chr17.fa.pac
hg19_chr17.fa.sa

Step 2. Alignment

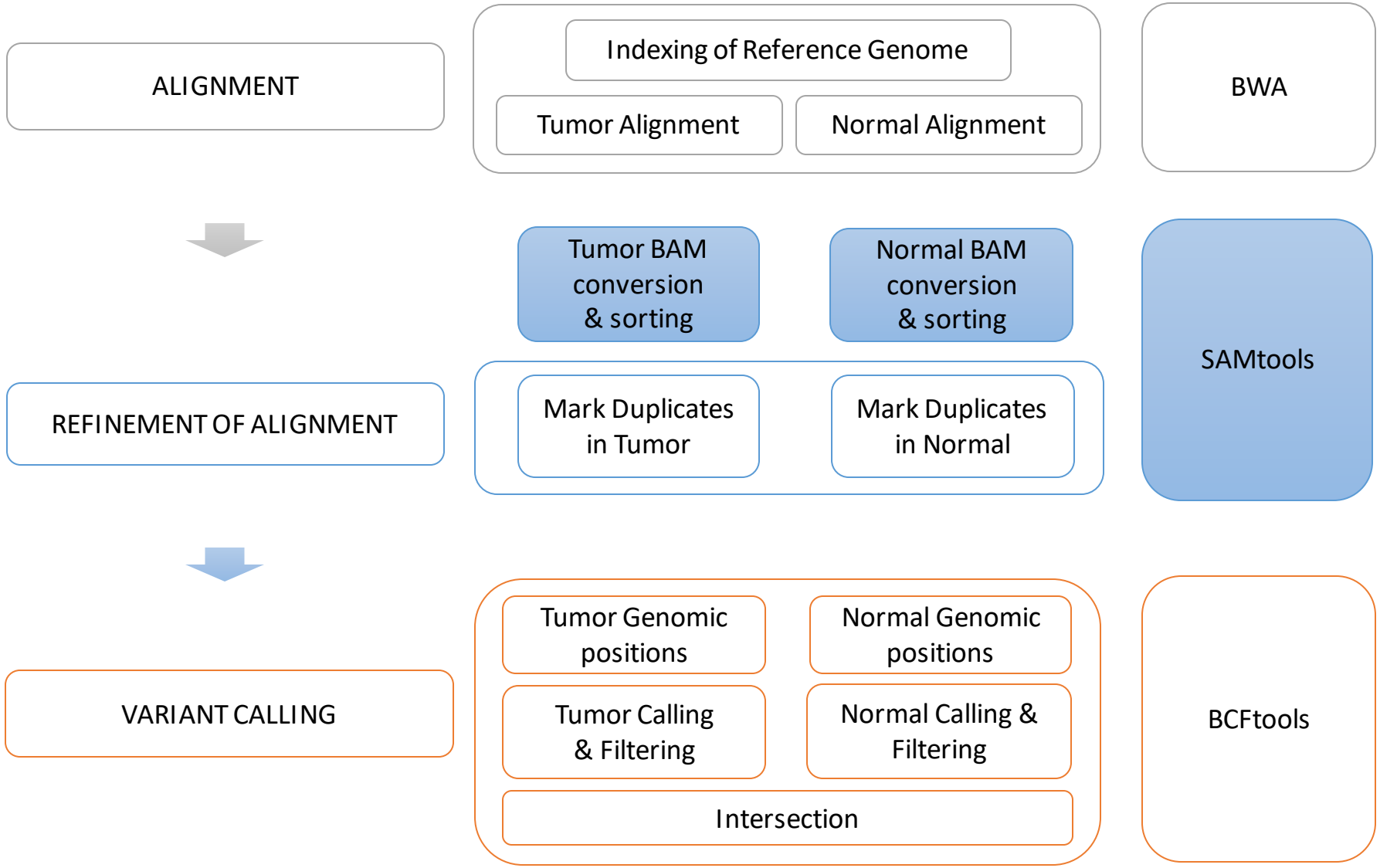
```
> bwa mem -R '@RG\tID:OVCA\tSM:sample'
/REFERENCE/hg19_chr17.fa
/SAMPLES/WEx_sample_R1.fastq
/SAMPLES/WEx_sample_R2.fastq > /RESULTS/sample.sam
```

Alignment of each of the 2 samples against the reference genome

[M::main_mem] read 133334 sequences (10000050 bp)...
[M::mem_pestat] # candidate unique pairs for (FF, FR, RF, RR): (0, 60917, 0, 0)
[M::mem_pestat] skip orientation FF as there are not enough pairs
[M::mem_pestat] analyzing insert size distribution for orientation FR...
[M::mem_pestat] (25, 50, 75) percentile: (333, 364, 400)
[M::mem_pestat] low and high boundaries for computing mean and std.dev: (199, 534)
[...]
[M::worker2@0] performed mate-SW for 10461 reads
[main] Version: 0.7.5a-r405
[main] CMD: bwa mem -R @RG\tID:OVCA\tSM:tumor /REFERENCE/hg19_chr17.fa /SAMPLES/WEx_Tumour_R1.fastq /SAMPLES/WEx_Tumour_R2.fastq
[main] Real time: 169.040 sec; CPU: 129.535 sec

WARNING: You must replace **`sample`** with the 'sample name' (i.e. Tumour or Normal)

ROAD MAP


ALIGNMENT
Indexing of Reference Genome
Tumor Alignment
Normal Alignment
BWA
REFINEMENT OF ALIGNMENT
Tumor BAM conversion & sorting
Normal BAM conversion & sorting
Mark Duplicates in Tumor
Mark Duplicates in Normal
SAMtools
VARIANT CALLING
Tumor Genomic positions
Normal Genomic positions
Tumor Calling & Filtering
Normal Calling & Filtering
Intersection
BCFtools

Step 3. Compression and sorting

```
> samtools view -S -b /RESULTS/sample.sam > /RESULTS/sample.bam
```

SAM to BAM conversion

BWA sometimes leaves unusual paired information on SAM records. We fix that using this function:

```
> samtools fixmate /RESULTS/sample.bam /RESULTS/sample_fixmate.bam
```

Fix paired information in files

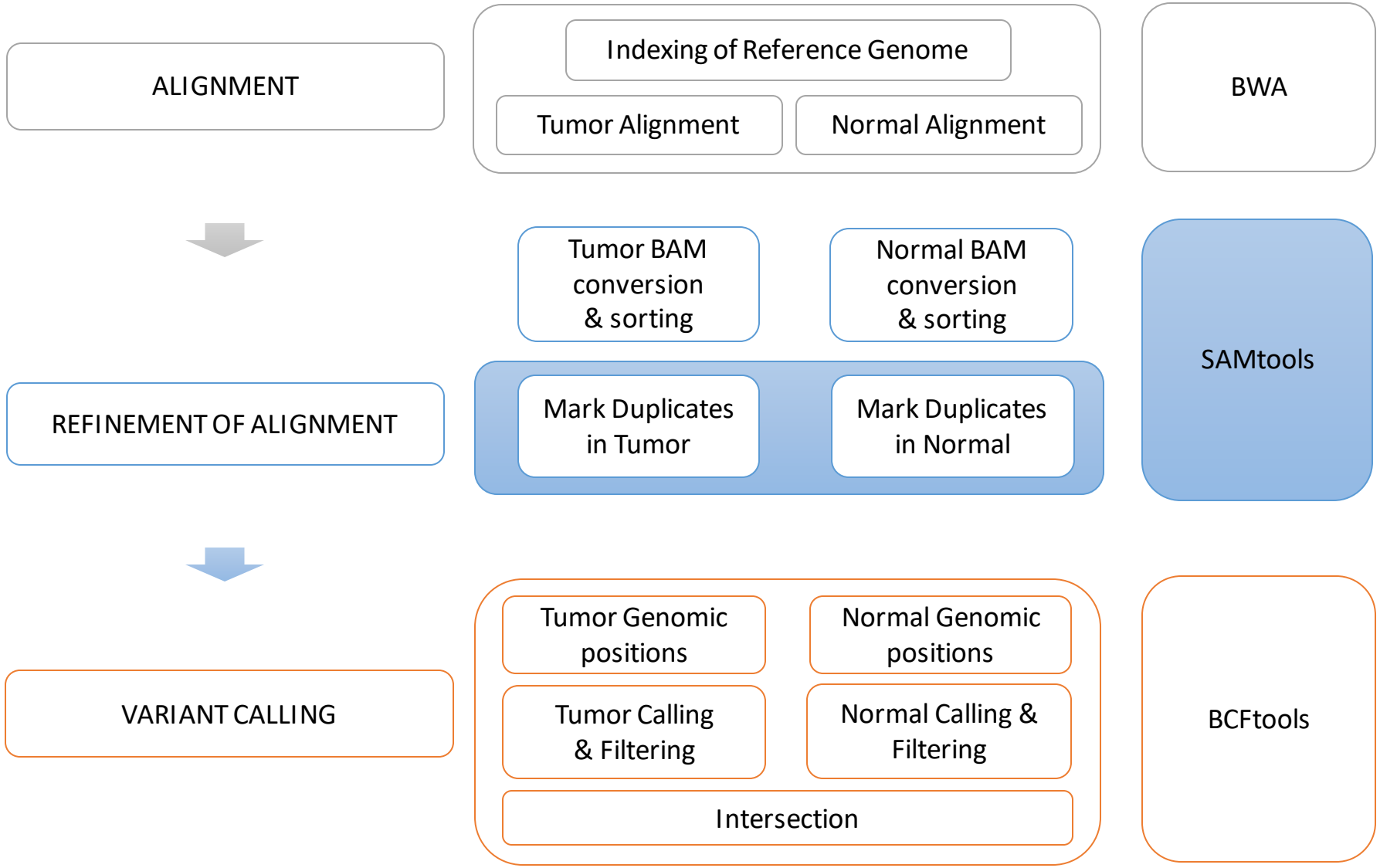
The variant caller requires that the alignment is sorted by genomic positions:

```
> samtools sort /RESULTS/sample_fixmate.bam /RESULTS/sample_sorted
```

Sort BAM files

WARNING: You must replace **`sample`** with the 'sample name' (i.e. Tumour or Normal)

ROAD MAP


ALIGNMENT
Indexing of Reference Genome
Tumor Alignment
Normal Alignment
BWA
REFINEMENT OF ALIGNMENT
Tumor BAM conversion & sorting
Normal BAM conversion & sorting
Mark Duplicates in Tumor
Mark Duplicates in Normal
SAMtools
VARIANT CALLING
Tumor Genomic positions
Normal Genomic positions
Tumor Calling & Filtering
Normal Calling & Filtering
Intersection
BCFtools

Step 4. Refinement of the alignment

Next, we mark duplicates (from the PCR). So the Variant Caller will ignore them:

```
> samtools rmdup -S /RESULTS/sample_sorted.bam /RESULTS/sample_refined.bam
```

Mark duplicates

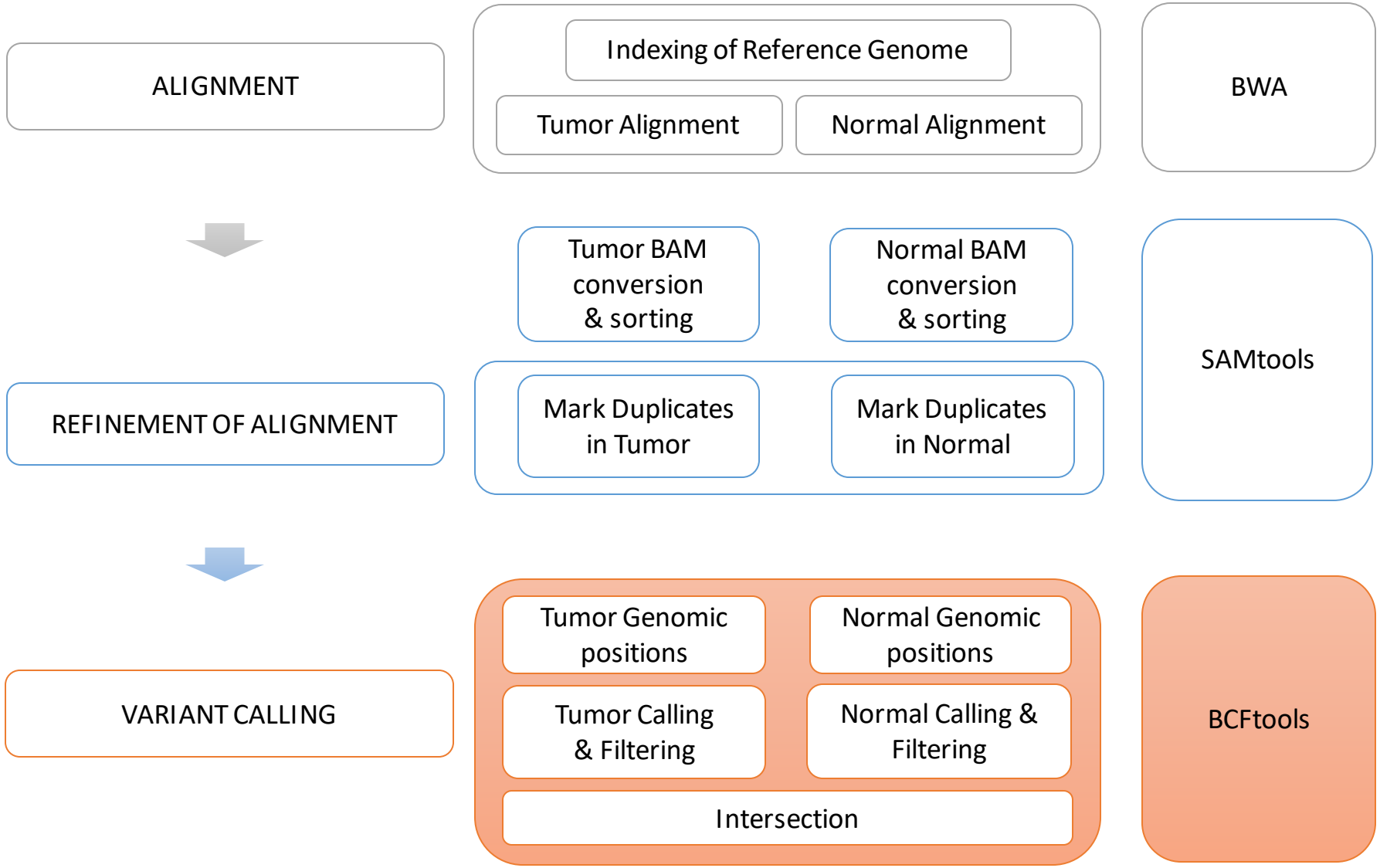
We index the BAM file:

```
> samtools index /RESULTS/sample_refined.bam
```

Index BAM files

WARNING: You must replace **sample** with the 'sample name' (i.e. Tumour or Normal)

ROAD MAP


ALIGNMENT
Indexing of Reference Genome
Tumor Alignment
Normal Alignment
BWA
Tumor BAM conversion & sorting
Normal BAM conversion & sorting
REFINEMENT OF ALIGNMENT
Mark Duplicates in Tumor
Mark Duplicates in Normal
SAMtools
VARIANT CALLING
Tumor Genomic positions
Normal Genomic positions
Tumor Calling & Filtering
Normal Calling & Filtering
Intersection
BCFtools

Step 5. Variant calling

Identify "Active regions" & calculate the most likely genotype :

```
> bcftools mpileup -Ou -f /REFERENCE/hg19_chr17.fa /RESULTS/sample_refined.bam | bcftools call -vmO z -o /RESULTS/sample_rawcalls.vcf.gz
```

Variant calling

Create the index for the vcf file:

```
> bcftools index /RESULTS/sample_rawcalls.vcf.gz
```

Indexing

WARNING: You must replace **sample** with the 'sample name' (i.e. Tumour or Normal)

Step 6. Variant calling filtering

Retrieve calls with high quality (include option):

```
> bcftools filter -i 'QUAL>10&&DP>10'  /RESULTS/sample_rawcalls.vcf.gz
```

Filtering Option 1

Retrieve calls with high quality (exclude option):

```
> bcftools filter -e 'DP<10'  /RESULTS/sample_rawcalls.vcf.gz
```

Filtering Option 2

WARNING: You must replace **sample** with the 'sample name' (i.e. Tumour or Normal)

Step 7. Intersection. Discern between germline/somatic variants

Intersection of detected variants in the two samples from the patient:

```
> bcftools isec -i 'DP>10'  /RESULTS/Tumour_rawcalls.vcf.gz /RESULTS/Normal_rawcalls.vcf.gz -p /RESULTS/
```

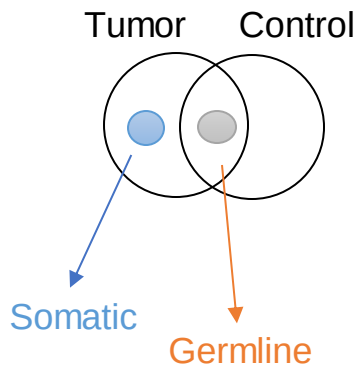
Intersection

What set of variants are stored in each vcf:

```
> cat /RESULTS/README.txt
```

Check distribution of variants in vcf

Using the following file names:
/RESULTS//0000.vcf for records private to /RESULTS/Tumor_rawcalls.vcf.gz
/RESULTS//0001.vcf for records private to /RESULTS/Normal_rawcalls.vcf.gz
/RESULTS//0002.vcf for records from /RESULTS/Tumor_rawcalls.vcf.gz shared by both /RESULTS/Tumor_rawcalls.vcf.gz /RESULTS/Normal_rawcalls.vcf.gz
/RESULTS//0003.vcf for records from /RESULTS/Normal_rawcalls.vcf.gz shared by both /RESULTS/Tumor_rawcalls.vcf.gz /RESULTS/Normal_rawcalls.vcf.gz
root@f280f8d0f046:/SOFTWARE/bcftools-1.9# cat /RESULTS/README.txt



*Case study :: **Mutation results***

- Germline variants

There were detected __ germline variants in total:
- __ Single Nucleotide Variants.
- __ Indels.

- Somatic variants

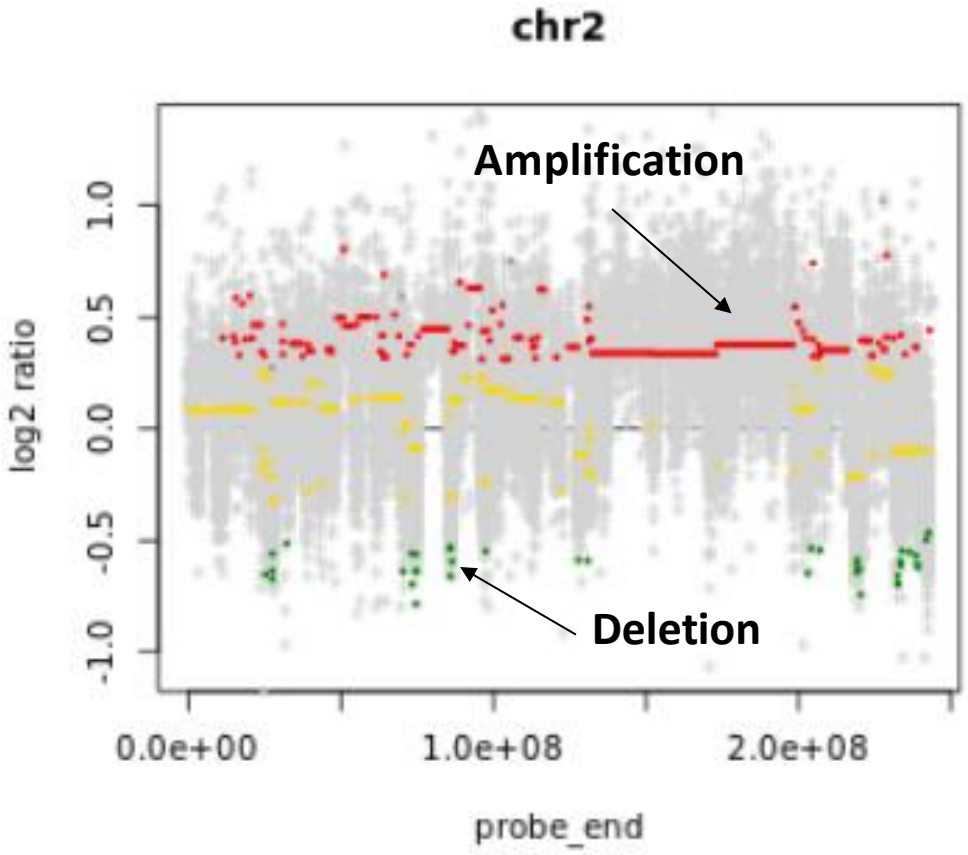
There were detected __ somatic variants in total:
- __ Single Nucleotide Variants.
- __ Indels.

Copy Number Variation (CNV)

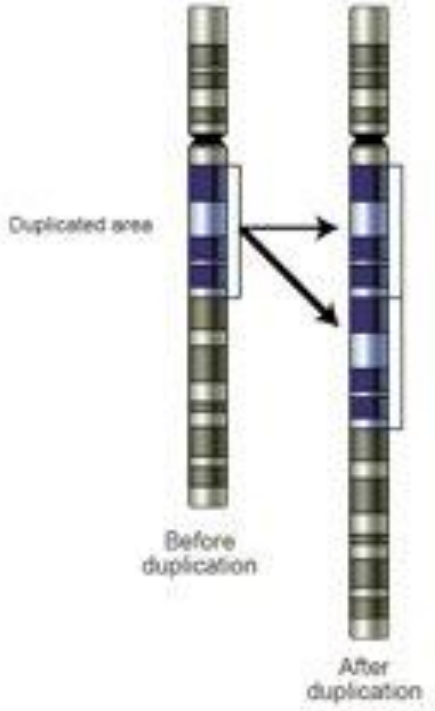
CNV

- Variations in the copy number of DNA regions.
- Longer than 1 kb (50 bp in some definitions)
- Can be deletions or amplifications
- Can change the expression of the genes

- Associated to phenotypes and diseases:
  - Cancer
  - Alzheimer
  - Autism
  - Diabetes



CNV detection



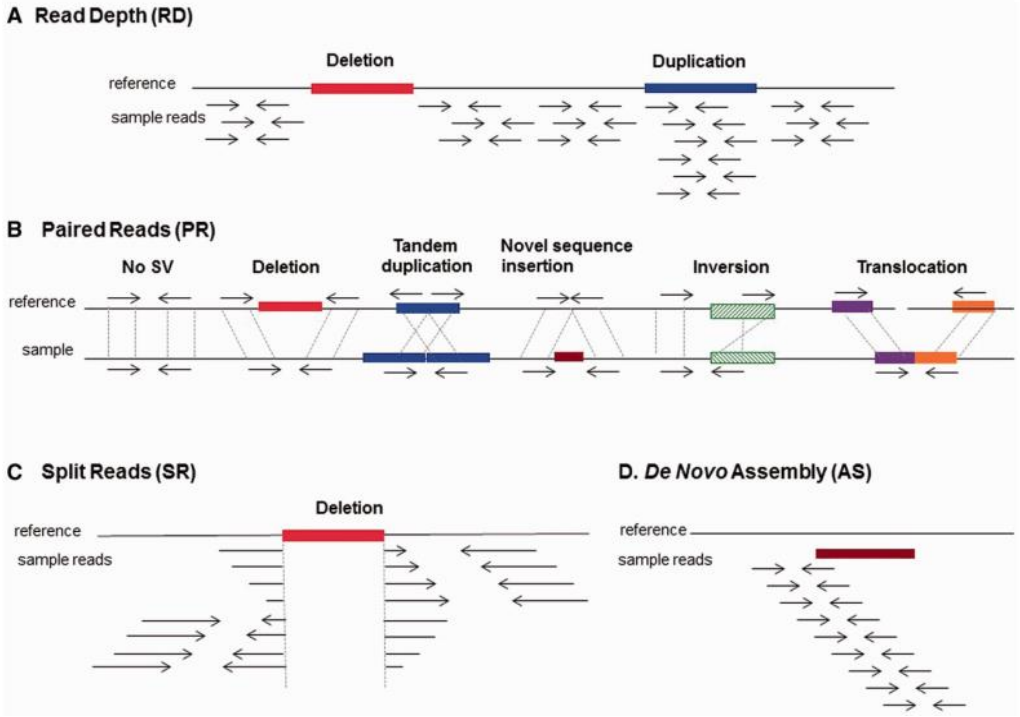
Classical detection:

- FISH (Fluorescence In Situ Hybridization)
- aCGH (Comparative Genomic Hybridization)

NGS:

- Read Count
- Read pair
- Split reads
- Assembly

All of them rely in depth information, therefore, enough and uniform coverage is required for optimal results.



Briefings in Functional Genomics, 2015

Read Depth methodology

- Two main steps:
  - **Pre-processing**: reduce noise and sequencing bias
  - **Segmentation**: join by statistical methods regions with a similar number of overlapping reads
- Examples of segmentation methods:
  - CBS (Circular Binary Segmentation)
  - HMM (Hidden Markov Models)

Algorithms for CNV in WGS

CNV detection methods on WGS data.

Software	Methods	Algorithm detail	Input data	Publish	Latest update	Accessibility	URL	Programing Language	#Citations
#Canvas	RD	Expectation-maximization (EM) clustering	BAM	2011	2018/3	Y	https://github.com/Illumina/canvas	C#	29
#cn.MOPS	RD	Mixture Poisson model	BAM	2012	2018/10	Y	http://www.bioinf.jku.at/software/cnmops/cnmops.html	R	226
CNVeM	RD	Expectation-maximization (EM) algorithm	CSV	2013	NA	Y	https://omictools.com/cnvem-tool	C	14
CNVer	RP	Maximum-likelihood, Graphic flow	BAM	2010	2011/5	N	NA	C	158
#CNVnator	RD	Mean shift algorithm	BAM	2011	2016/11	Y	https://github.com/abyzovlab/CNVnator	C++	640
CNVrd2	RD	Expectation-maximization (EM) algorithm	BAM/SAM	2014	2015/11	Y	https://bioconductor.org/packages/release/bioc/html/CNVrd2.html	R	13
#Control-FREEC	RD	LASSO regression	BAM/SAM	2011	2018/8	Y	http://boevalab.com/FREEC/	C++	190
#GROM-RD	RD	Quantile normalization	BAM	2015	2017/5	Y	http://grigoriev.rutgers.edu/software/	C	7
#iCopyDAV	RD	DoC approaches	BAM	2018	2018/3	Y	https://github.com/vogetihrsh/icopydav	R,C++	1
JointSLM	RD	Population-based approach	SAM/BAM	2011	NA	N	NA	R	49
#LUMPY	RD, PEM	A probabilistic framework	BAM/CRAM	2014	2016/3	Y	https://github.com/arq5x/lumpy-sv	C++	157
mrCaNaVAR	RD	mrFAST	SAM	2009	2013/9	Y	http://mrcanavar.sourceforge.net/	C	685
#RDXplorer	RD	Event-wise testing algorithm	BAM	2009	2013/4	Y	https://sourceforge.net/projects/rdxplorer/	Python	496
#ReadDepth	RD	Circular binary segmentation algorithm	Bed Files	2011	2014/8	Y	https://github.com/chrisamiller/readDepth	R	150
#RSICNV	RD	Negative binomial transformations	BAM	2017	2017/7	Y	https://github.com/yhwu/rsicnv	C++	2

PLOS Computational Biology. 2019.

Algorithms for CNV in WES

Table 1

Selected tools for the performance analysis of CNV detection tools using WES data

Tool name Chara- Cteristics	ADTEx	CONTRA	cn.MOPS	ExomeCNV	VarScan 2
Control set required	Yes	Yes	No	Yes	No
Prog. Language	Python, S/R	Python, R	R	R	Java
Input format	BAM, BED	BAM, SAM, BED	BAM, Read count matrices	BAM, Pileup, GTF	BAM, Pileup
Segmentation Algorithm	HMM	CBS	CBS	CBS	NA[a]
OS	GNU, Linux	Linux, Mac OS	Linux, Mac OS, windows	Linux, Mac OS, windows	Linux, Mac OS, windows
Methodology characteristic	DWT[c] for de-noising, use BAF[d]	Base-level log-ratio	Bayesian approach for de-noising	Statistical test for analyzing BAF data	CMDS[b] for generating read counts
Year	2014	2012	2012	2011	2012
URL	http://adtex.sourceforge.net	https://sourceforge.net/projects/contra-cnv/	http://www.bioinf.jku.at/software/cnmops/cnmops.html	https://secure.genome.ucla.edu/index.php/ExomeCNV_User_Guide	http://varscan.sourceforge.net/

[a]Segmentation is not imbedded in the tool. CBS is recommended for segmentation

[b]Correlation Matrix Diagonal Segmentation

[c]Discrete wavelet transform

[d]B allele frequencies

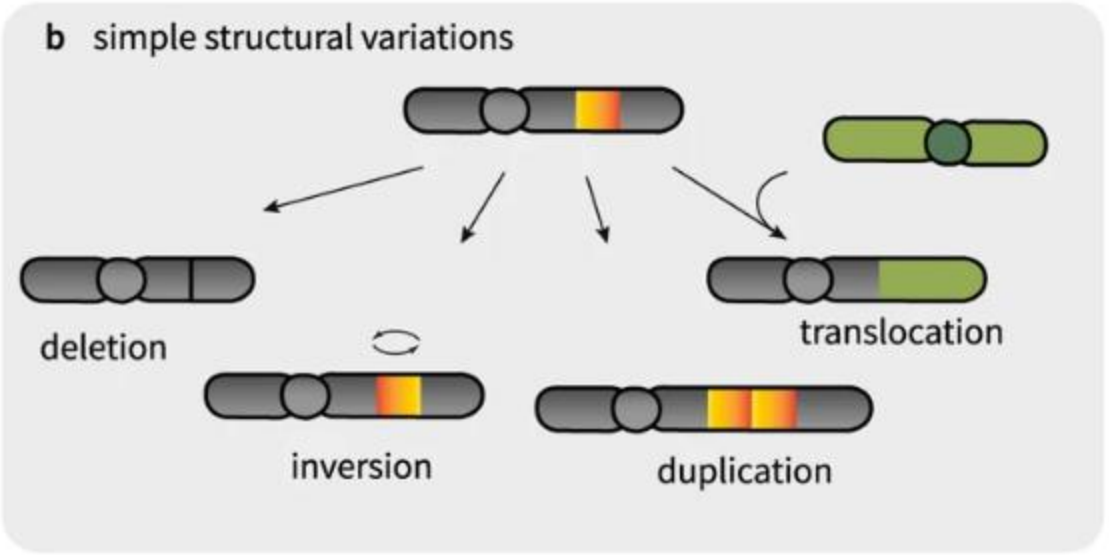
BMC Bioinformatics. 2017.

Structural Variants

Structural Variants

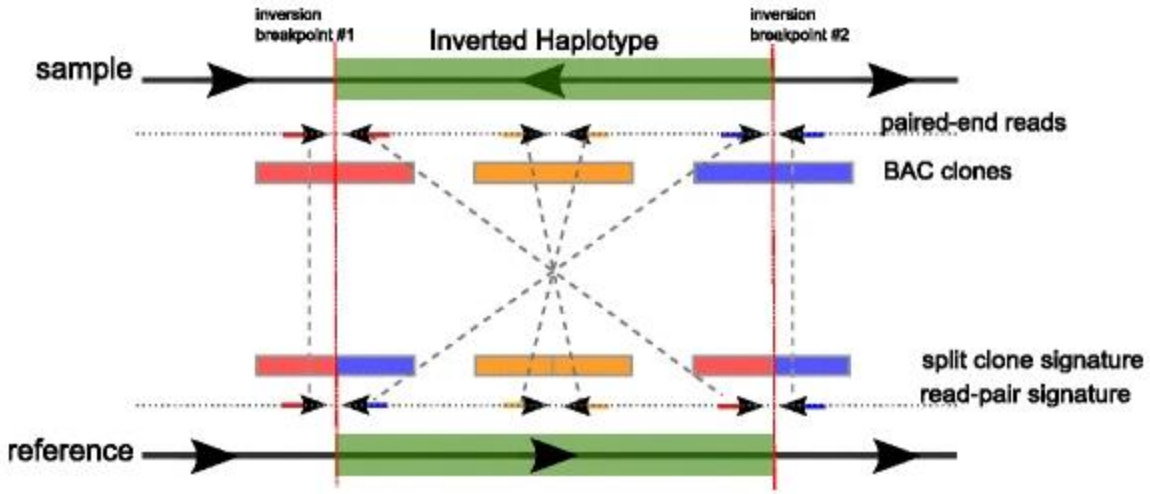
- Changes that affect to DNA structure
- Longer than 1 kb (> 50 bp in some definitions)
- CNV is a subclass of Structural Variants
- Another structural variants are inversions and translocations

- Associated to phenotypes and diseases

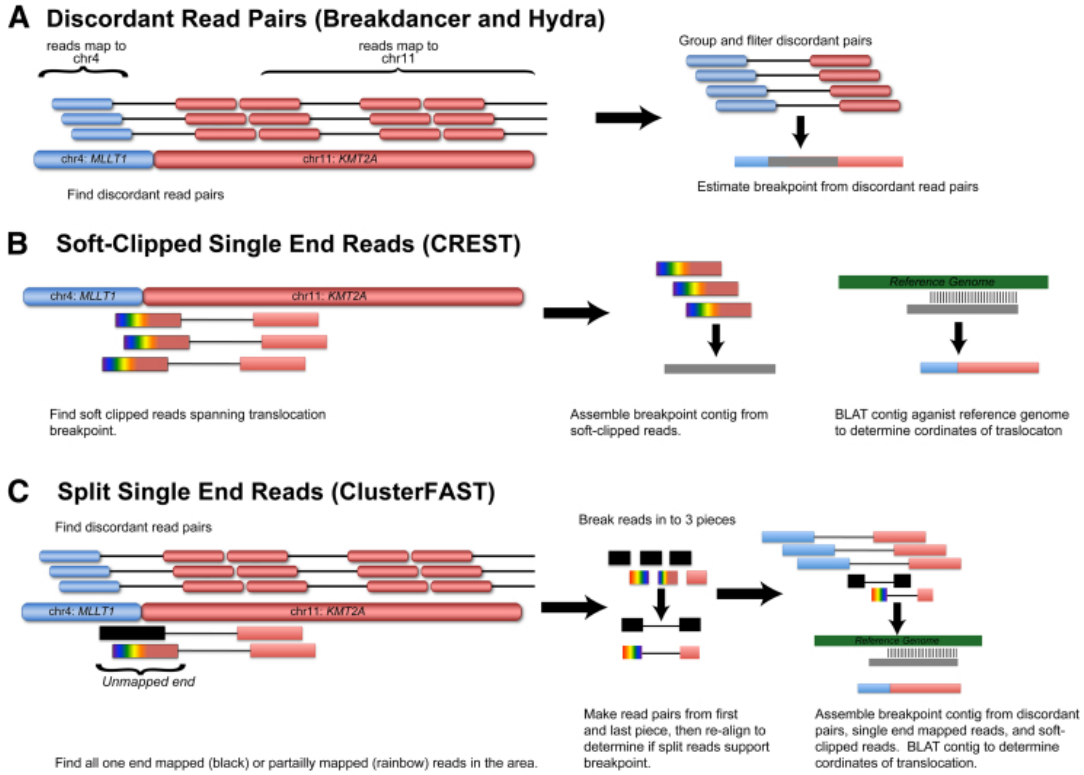


Experimental & Molecular Medicine. 2018

Detection of inversions and translocations



BMC Genomics, 2017



J Mol Diagn., 2014

Inversions and translocations are named balanced rearrangements

Algorithms for inversions and translocations

Algorithm	Inversions	Translocations	Discovery stage	Validation stage	Techniques
LUMPY	Yes	Yes	RD;SC;PR	RD;SC;PR	CL;SA
MetaSV	Yes	Yes	RD;SC;PR	RD;SC;PR	CL;SA;CA
Meerkat	Yes	Yes	SC;PR;OEA	SC;OEA	CL;SA
SVMerge	Yes	Yes	SC;PR;OEA	SC;PR;OEA	CL;SA;CA
SoftSV	Yes	Yes	SC;PR	SC;PR	CL;SA
BreaKmer	Yes	Yes	SC;OEA	PR	CA
ClipCrop	Yes	Yes	SC	SC	CL;SA
CREST	Yes	Yes	SC	SC	CA
Gustaf	Yes	Yes	SC	SC	SA
Socrates	Yes	Yes	SC	SC	CL;SA
Bellerophon		Yes	PR	SC;PR	CL;SA
BreakDancer	Yes	Yes	PR		CL;ST
DELLY	Yes	Yes	PR	SC;OEA	CL;SA
FACTERA		Yes	PR	SC	CL;SA
GASV	Yes	Yes	PR		CL
GASVPro	Yes	Yes	PR	RD;PR	CL;ST
HYDRA	Yes	Yes	PR	PR	CL
HYDRA-Multi	Yes	Yes	PR	PR	CL
inGAP-SV	Yes	Yes	PR	RD	

Algorithm	Inversions	Translocations	Discovery stage	Validation stage	Techniques
PEMer	Yes	Yes	PR		CL
PeSV-Fisher	Yes	Yes	PR	RD	CL
PRISM	Yes	Yes	PR	OEA	CL;SA
SVDetect	Yes	Yes	PR	PR	CL
SVMiner	Yes		PR	RD;PR	CL;ST
Ulysses	Yes	Yes	PR	RD;PR	CL;ST
SLOPE		Yes	OEA	OEA	CL;SA
SMUFIN	Yes	Yes			CL;CA

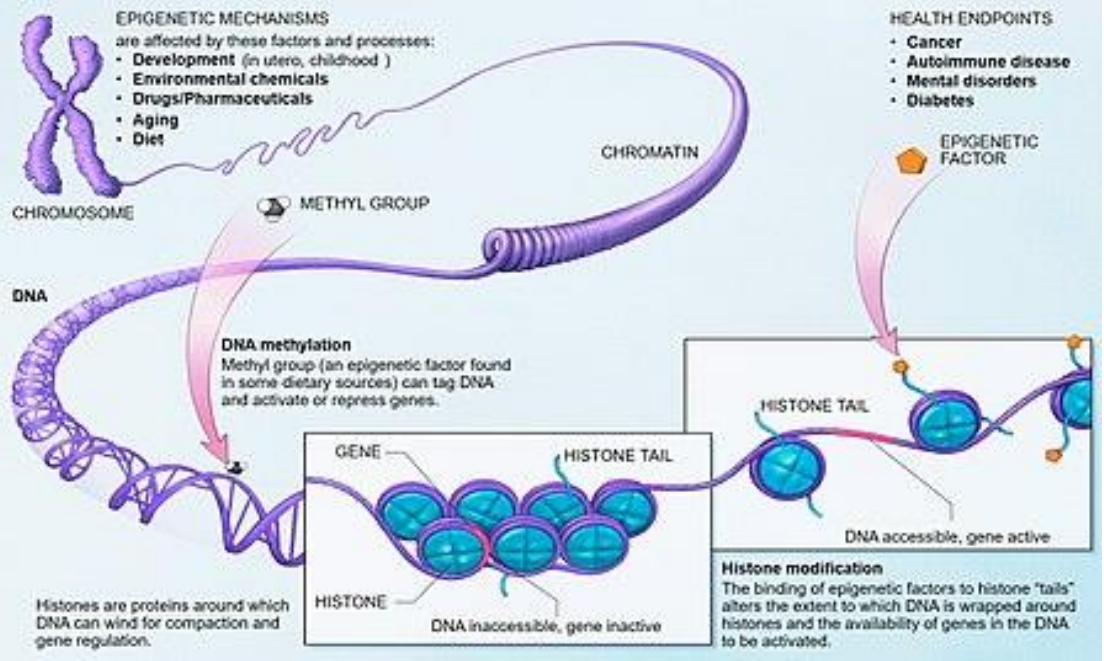
Adapted from Methods., 2016

RD: Read Depth
SC: Soft Clipped
PR: Paired Reads
OEA: One End Anchored

CL: Clustering
SA: Split-reads Alignment
CA: Contig Assembly
ST: Statistical Testing

Epigenetic variation

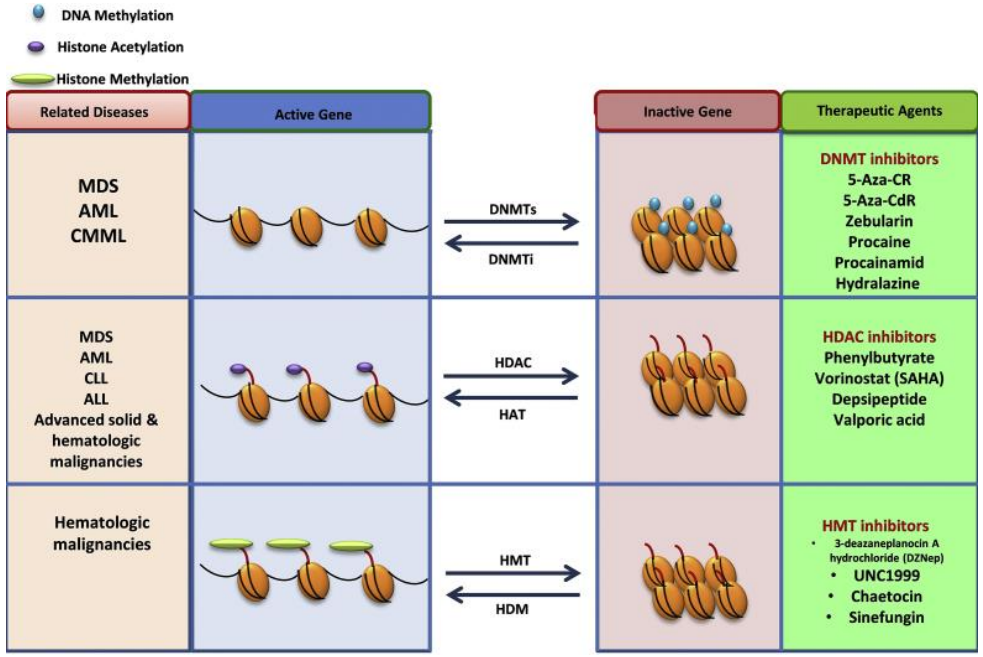
Epigenome variants



- Changes not concerning nucleotide sequence. They can be reversible.
- Can be caused by external, environmental or biological factors.

- Genetic and epigenetic factors are interconnected.

Epigenome variants



- Types

  - DNA methylation
  - Histone modification: covalent modification, methyl groups, acetyl groups, …
  - Binding of proteins to regulatory regions (silencers o enhancers)
  - Non coding RNAs

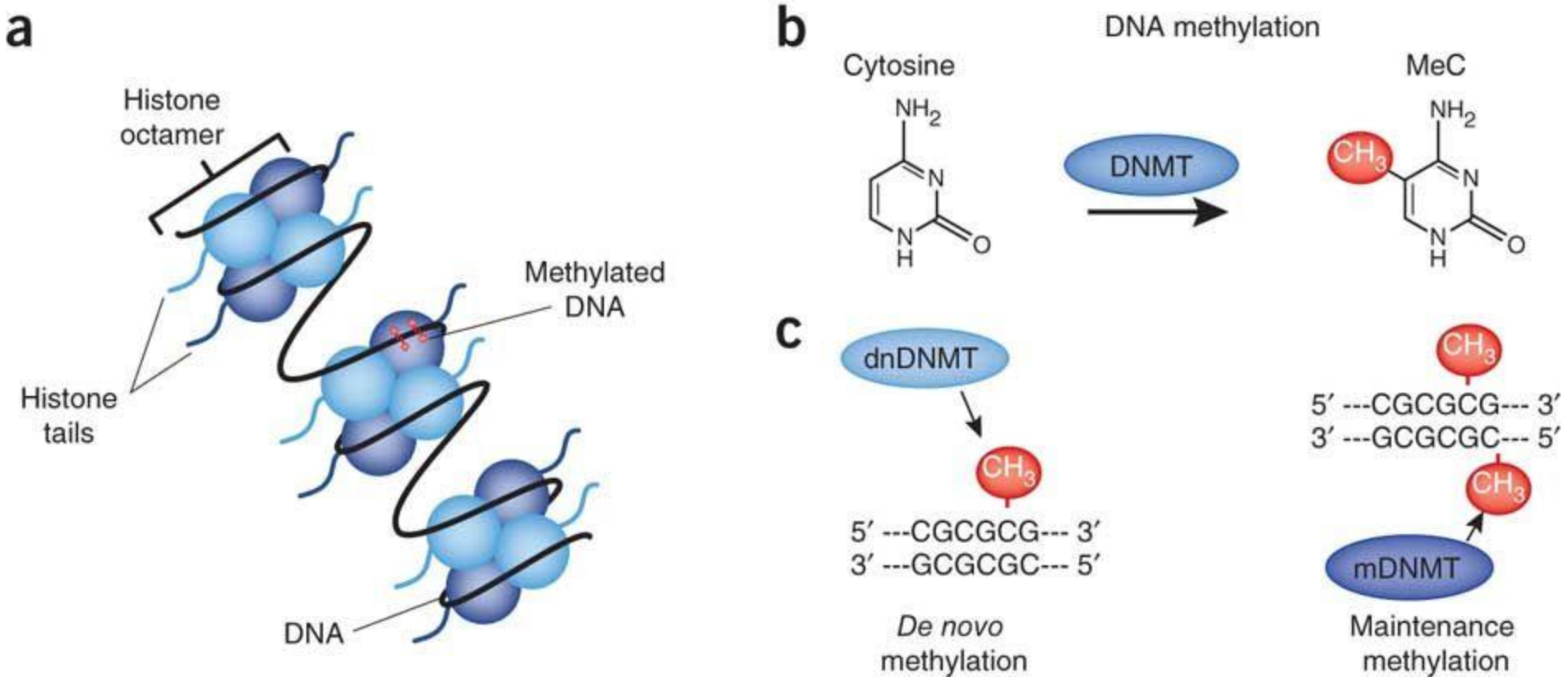
- Consequences

  - Development: X chromosome inactivation, celular lineage
  - Phenotypic differences. I.e.: monozygotic twins
  - Diseases: AS (UBE3A gene), cancer (IGF2)

NGS and epigenomics

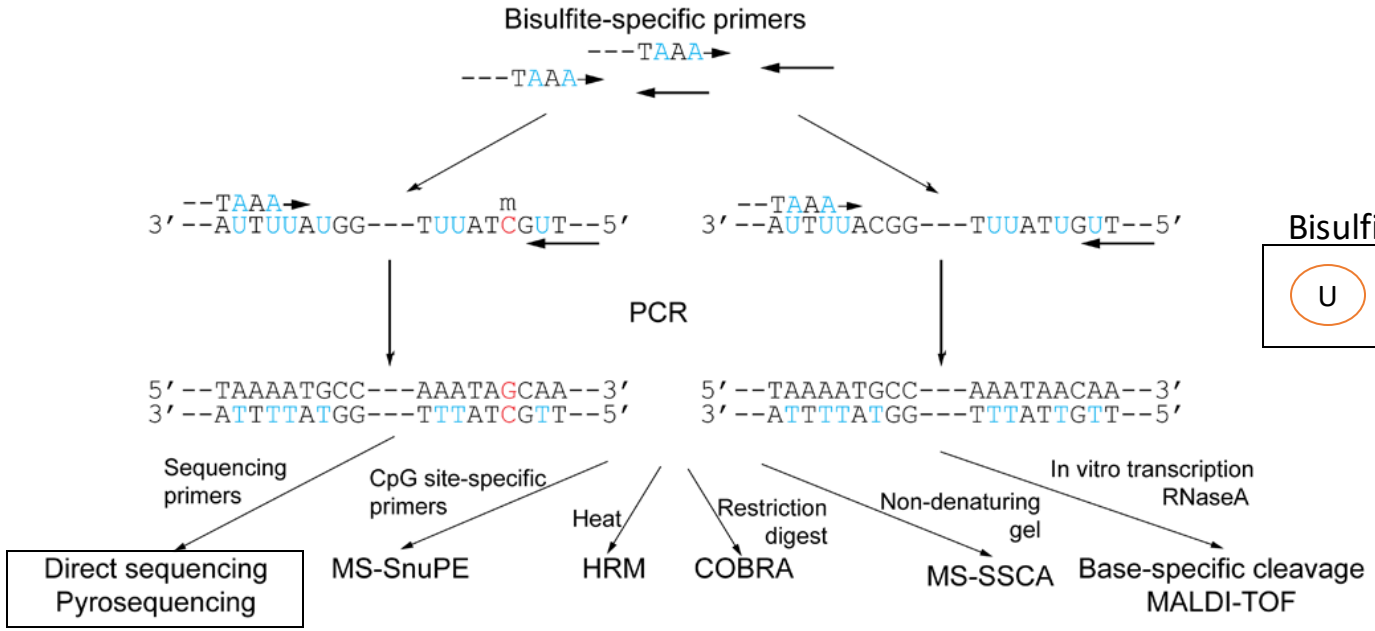
- **Bisulfite sequencing:** DNA methylation
- **Chip-seq:** protein-DNA interactions, histone modifications

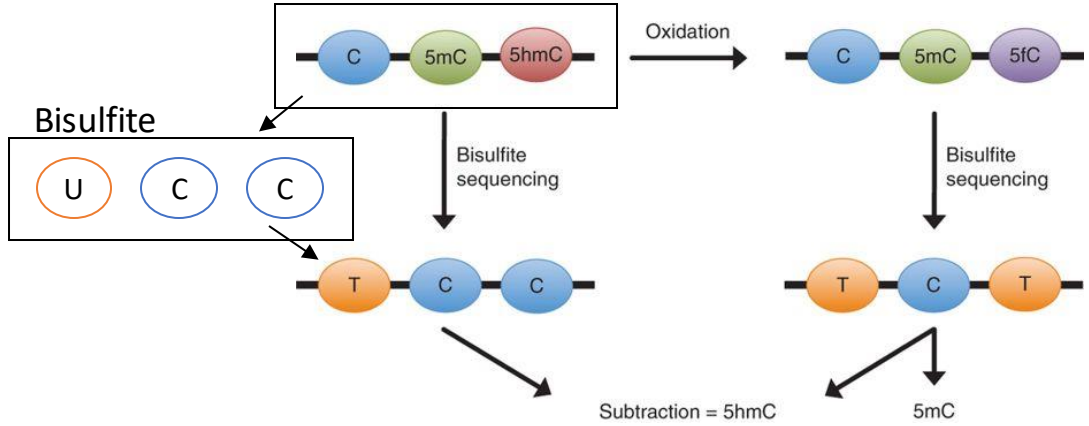
DNA methylation



Nature Neuroscience **volume13**, pages1319–1323 (2010)

Bisulfite sequencing





Nature Protocols **volume8**, pages1841–1851 (2013)

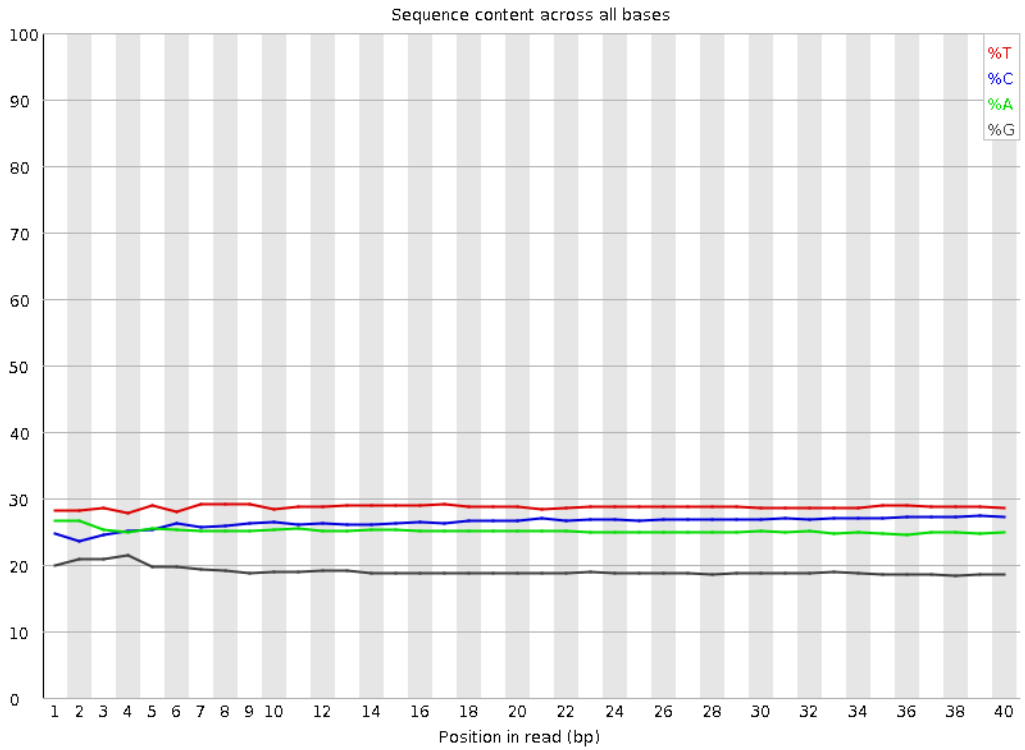
Bisulfite sequencing - bioinformatic tools

Step	Bismark workflow	bwa-meth workflow
Generate Reference Genome Index *(optional)*	Bismark	bwa-meth
Raw data QC	FastQC	FastQC
Adapter sequence trimming	Trim Galore!	Trim Galore!
Align Reads	Bismark	bwa-meth
Deduplicate Alignments	Bismark	Picard MarkDuplicates
Extract methylation calls	Bismark	MethylDackel
Sample report	Bismark	-
Summary Report	Bismark	-
Alignment QC	Qualimap	Qualimap
Sample complexity	Preseq	Preseq
Project Report	MultiQC	MultiQC

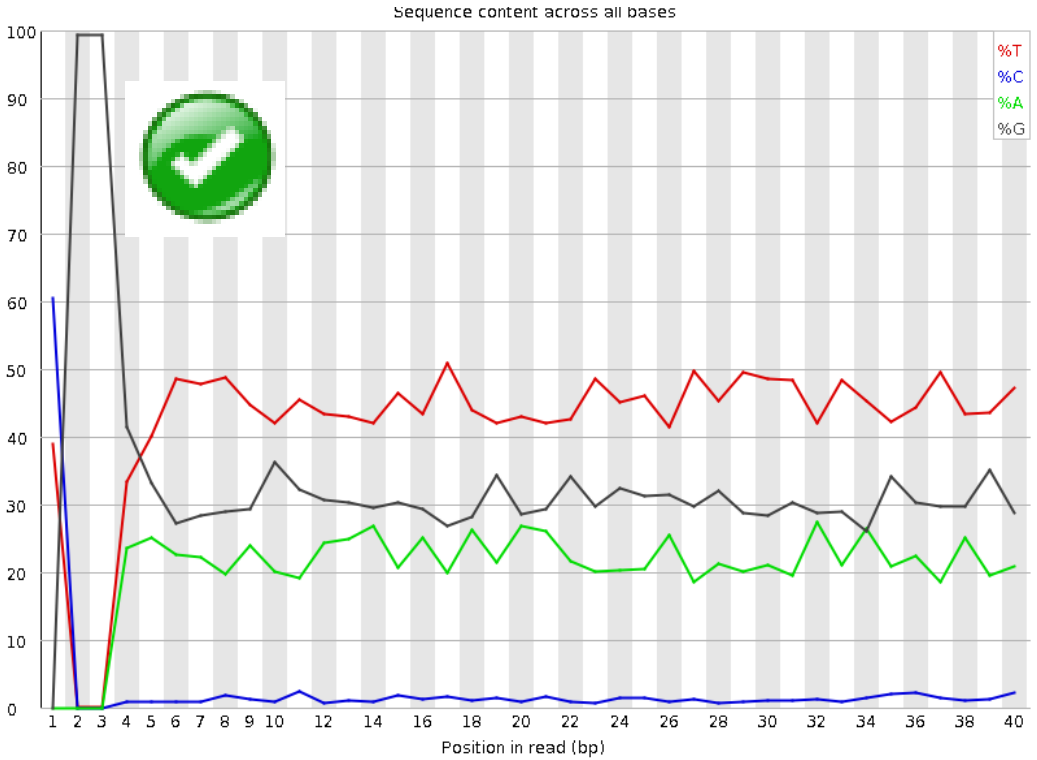
https://github.com/nf-core/methylseq

Bisulfite sequencing - FastQC

Per base sequence content



DNA sequencing

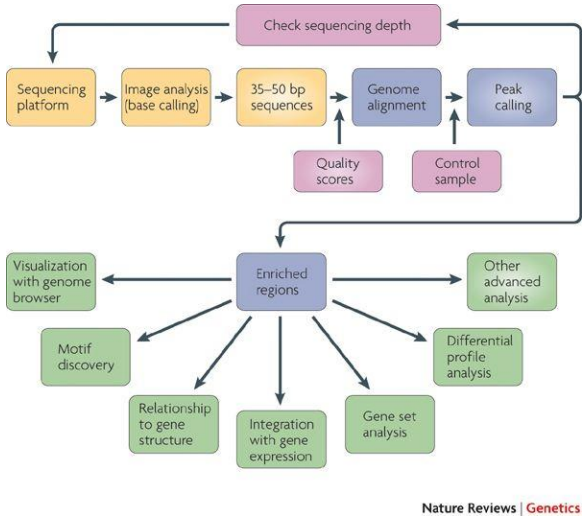


bisulfute sequencing

Bisulfite sequencing - Alignment

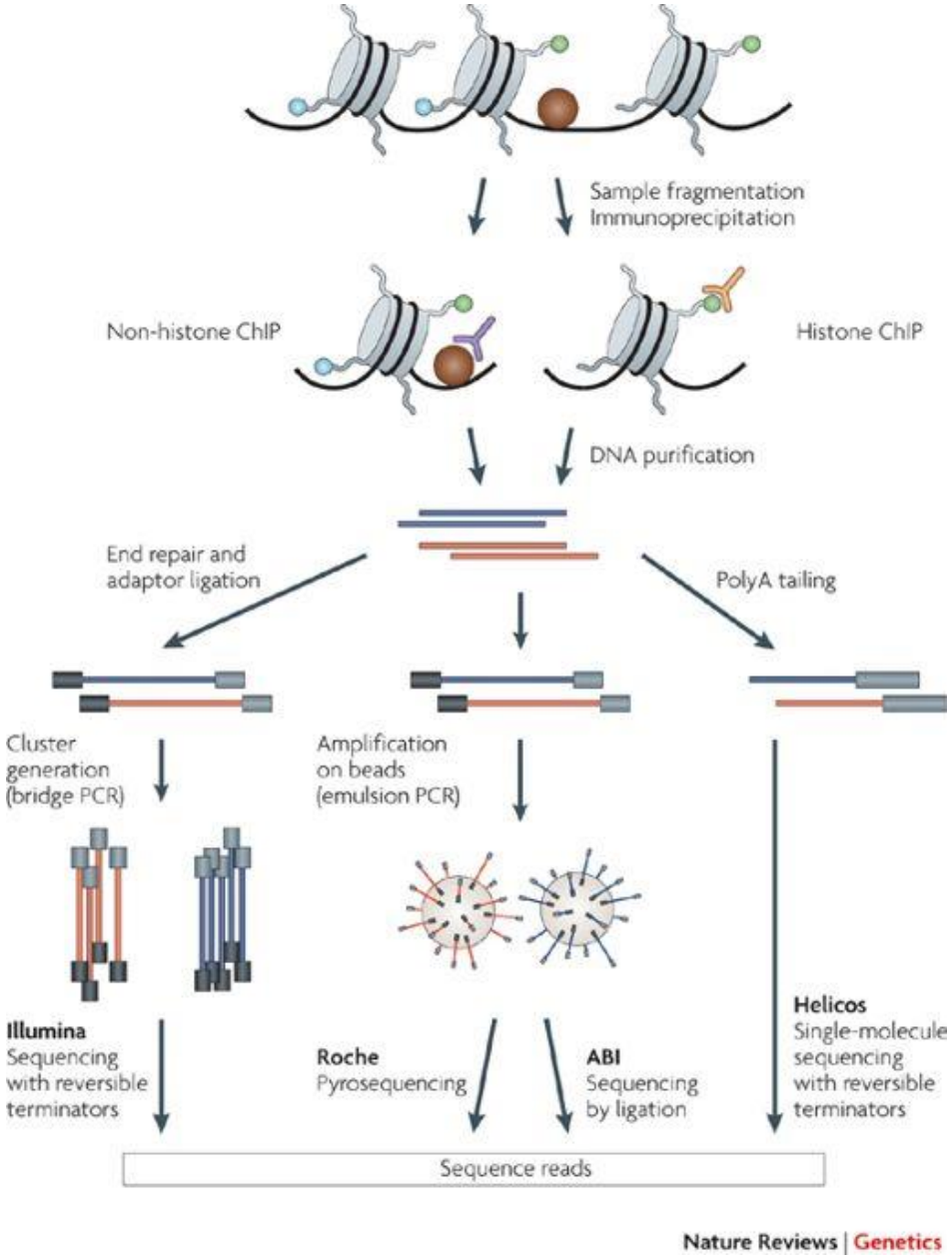
- More differences in bases, in relation to reference genome (C/T)
- Conventional aligners for DNA sequencing are not appropriate
- Examples of bisulfite sequencing specific aligners:
  - Bwa-meth
  - Bismark

Chip-seq

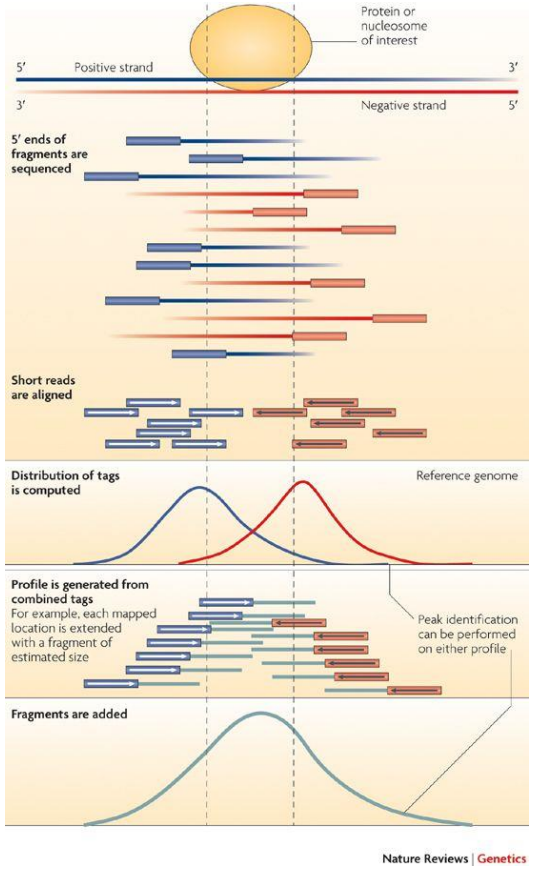


Downstream analysis

- Protein binding sites motifs
- Annotate peak location



Chip-seq



Types of peaks

- Sharp: protein-DNA binding sites or histone modifications in regulatory elements
- Broad: histone modifications marking domains (transcribed or repressed regions)
- Mixed: RNA polymerase