

MOLECULAR METHODS AND BIOINFORMATICS

LM Evolutionary Biology, University of Padua

Dr. Enrico Gaffo, Dr. Silvia Orsi,

Prof. Stefania Bortoluzzi

Practical lesson 5 “NGS data analysis: Variant calling”

Padua, November 19, 2025

Aim of the practical session

- Become familiar with the main file formats used for the storage of DNA-seq data
- Carry out the essential steps in the analysis of DNA-seq data obtained by next-generation sequencing for the detection of genomic variants ("variant calling").

Cases to analyze

- Analysis of a human exome affected by pediatric follicular lymphoma to identify germline or somatic variants that could be associated with the disease
- Exome analysis obtained from ancient DNA of a 14th century Venetian nobleman who died under mysterious circumstances.

Cases to analyze

- Analysis of a human exome affected by pediatric follicular lymphoma to identify germline or somatic variants that could be associated with the disease
- Exome analysis obtained from ancient DNA of a 14th century Venetian nobleman who died under mysterious circumstances.

Preparing the exercise

- Download the file “Practical_session_5.zip” from http://compgen.bio.unipd.it/~stefania/Didattica/AA2025-2026/MMOL_BIOINFO_EB/Practical_session_5.zip

- Unzip the file “Practical_session_5.zip”:

?

- Move to the Practical_session_5 folder:

?

Preparing the exercise

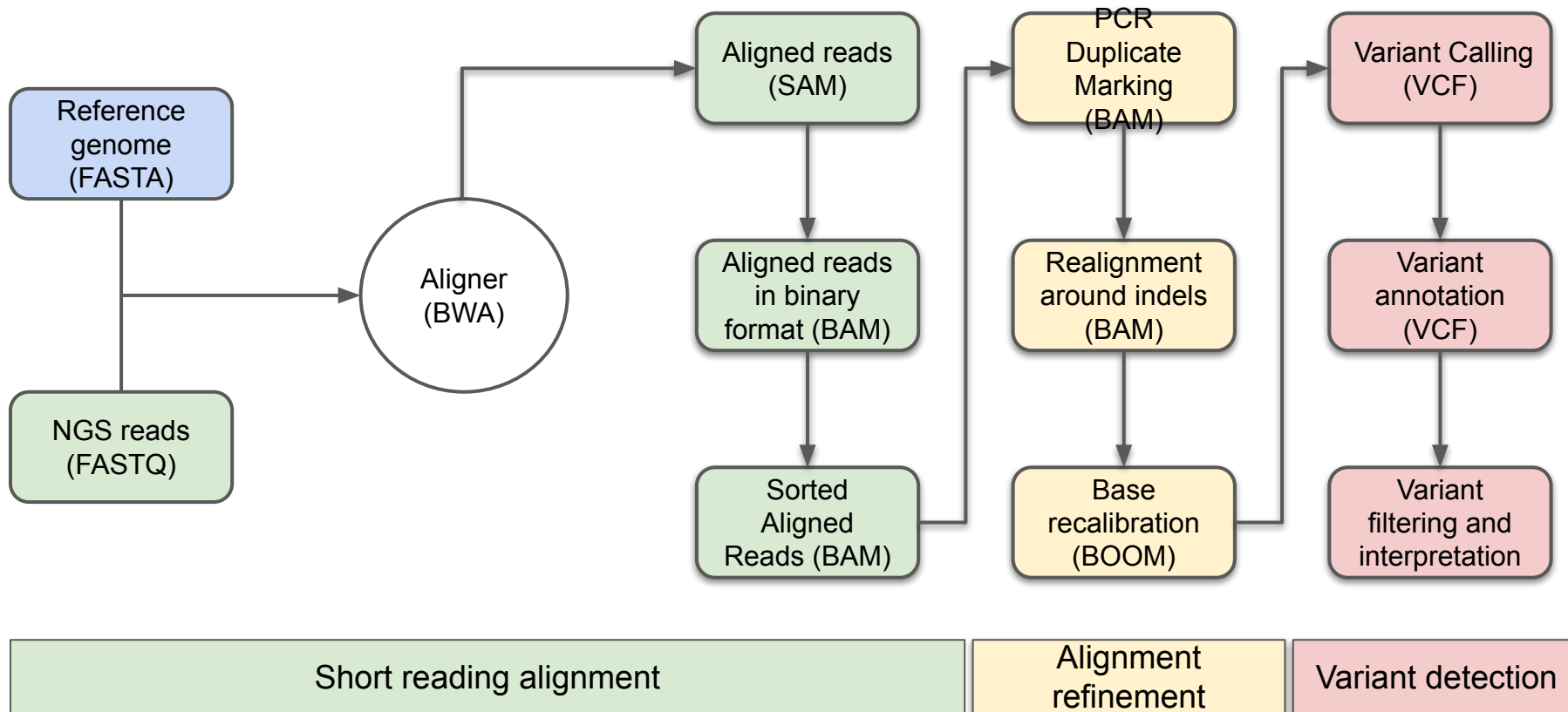
- Download the file “Practical_session_5.zip” from http://compgen.bio.unipd.it/~stefania/Didattica/AA2025-2026/MMOL_BIOINFO_EB/Practical_session_5.zip
- Unzip the file “Practical_session_5.zip”:

```
unzip Practical_session_5.zip
```

- Move to the Practical_session_5 folder:

```
cd Practical_session_5.zip
```

A typical workflow for variant calling

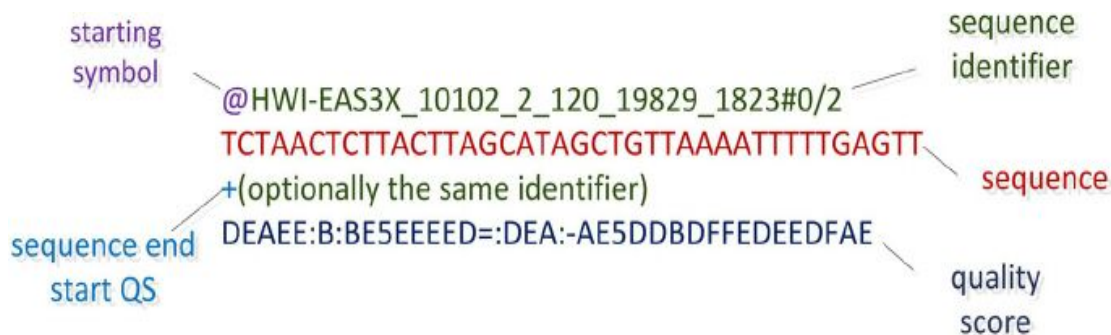


FASTQ format

To view a FastQ:

```
less 1.fastq
```

```
less 2.fastq
```



```
@M70273:8:000000000-AJLMP:1:1101:14452:1861 1:N:0:1
TAACTACTTTGGGGAATGTTAGCCTGGACAAACAACATTTGATGAATGTCTGTTTCTTTCTGAATT
+
5,,5</5<@A--++,+6-AC/.88A,+6-,-7,+7+8AC..9..9..9.-88CAEFFFECE---5A
@M70273:8:000000000-AJLMP:1:1101:14458:1948 1:N:0:1
CAGTGAAACGATATACTCCAGCCCGATTGCCCTGGGCTGCCAGGGTGCCAAACCAAGGAACCTCTT
+
=====99/@@@@@AAE8C;-8C>CC7EE-9.977+++7++A--++555@A-55>A+,+,-,AFFFE
@M70273:8:000000000-AJLMP:1:1101:14505:2082 1:N:0:1
GTGCTGTTTCATCACTGTGCCATTGCAGGTTTATTTGAAATACAACATGTCCAAGAGGAAAGCACTG
+
?????B??B?BBBBBBBFBFFHHHFFHHHHFH09EFFHDFEFEG@FHHFGFD?D-CEFFHDFE
@M70273:8:000000000-AJLMP:1:1101:14399:2091 1:N:0:1
TGCCTCCCTTTCCAATGGACTATTTTAGAAGAAATGGAGCTGTACCCACATCAAGATTGAGAACACTG
+
?????ABA?DDDDDDDDFGGFGFFIIHHIIFHHHII@FHHIIIIIGFF>EHHHFFGHHIFHFGHAFGH
@M70273:8:000000000-AJLMP:1:1101:16927:2095 1:N:0:1
CCTATCATATATGCCTTAGTTTGTATGAAANATATTGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
+
??AA?BBBEDDEEEEGGGGGIHHIIII#7AFHII#####5#####
@M70273:8:000000000-AJLMP:1:1101:18171:2095 1:N:0:1
TTGTGATTCCACATTCTCTTCCATTGTAGNGCAAATNNNNNNNNNNNNNNNNNNNNNNNTCNTTNNNTN
+
?????BBBDDDDDDDDGGGGGIIHHI#7AEFHI#####7###55#55###5###
@M70273:8:000000000-AJLMP:1:1101:19337:2095 1:N:0:1
GCCGCCCATGCGCGGCATGATGAAGTCCGCTGCTGTNNNNNNNNNNNNNNNNNNNNNTNTTNNNCAN
+
?????ABAADDDDDDDFFFFFIHHIIHHHHHHI#####5###55#55###4###
@M70273:8:000000000-AJLMP:1:1101:14484:2097 1:N:0:1
CTGGACTGATATGTGATTATTCTTTCAACAGCCACGCCAGATCCAGTGAAAAACAAGCTCTCATGTC
+
?????A?BB?DDDDDDDBGGGGGIIHHIIHHHHHHHFGFHHHHHHHHHHHHHHHHHHHHHHHHHH
@M70273:8:000000000-AJLMP:1:1101:16321:2100 1:N:0:1
TAGATGCTTTTAACTAAGTTACCTGACTTNCCTTATNNNNNNNNNNNNNNNNNNNNNNNGCNGCNNCN
+
?????BBBDDDDDDDDGFGGGGIIHHI#7AFHFG#####7###55#55###5###
```


Phred Quality Score

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%

$$Q = -10 \log_{10} P$$

$$P = 10^{-Q/10}$$

Preparation of the reference genome

Download the fasta files of human chromosomes 1, 15 and 17 from the Ensembl link http://ftp.ensembl.org/pub/grch37/release-104/fasta/homo_sapiens/dna/

Move to the “Downloads” folder:

?

Move the chromosome files into the “Practical_session_5” folder:

?

Merge chromosomes into a single FASTA file:

?

Preparation of the reference genome

Move to the “Downloads” folder:

```
cd ..
```

Move the chromosome files into the “Practical_session_5” folder:

```
mv Homo_sapiens.GRCh37.dna.chromosome.1* ./Practical_session_5/
```

Merge chromosomes into a single FASTA file:

```
cd Practical_session_5/
```

```
zcat Homo_sapiens.GRCh37.dna.chromosome.1.fa.gz
```

```
Homo_sapiens.GRCh37.dna.chromosome.15.fa.gz
```

```
Homo_sapiens.GRCh37.dna.chromosome.17.fa.gz > ref.fa
```

Preparation of the reference genome

Check that the chromosomes are in the correct order:

?

Preparation of the reference genome

Check that the chromosomes are in the correct order:

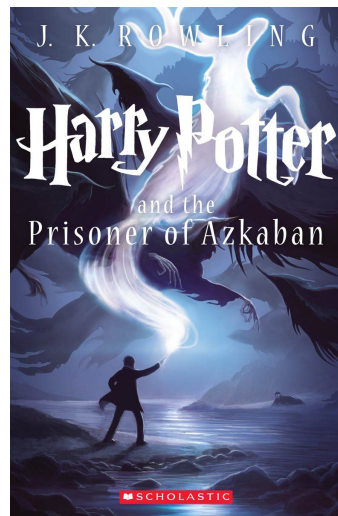
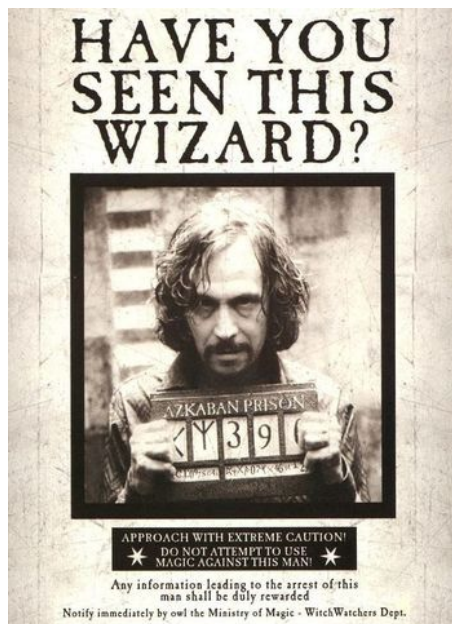
```
grep ">" ref.fa
```

Create reference sequence indices with 3 commands:

```
bwa index ref.fa
```

```
samtools faidx ref.fa
```

```
gatk CreateSequenceDictionary -R ref.fa -O ref.dict
```



Looking for Sirius Black

“Nonsense!” said Percy, looking startled. “You had too much to eat, Ron — had a nightmare —”

“I’m telling you —”

“Now, really, enough’s enough!”

Professor McGonagall was back. She slammed the portrait behind her as she entered the common room and stared furiously around.

“I am delighted that Gryffindor won the match, but this is getting ridiculous! Percy, I expected better of you!”

“I certainly didn’t authorize this, Professor!” said Percy, puffing himself up indignantly. “I was just telling them all to get back to bed! My brother Ron here had a nightmare —”

“IT WASN’T A NIGHTMARE!” Ron yelled. “PROFESSOR, I WOKE UP, AND SIRIUS BLACK WAS STANDING OVER ME, HOLDING A KNIFE!”

Professor McGonagall stared at him.

“Don’t be ridiculous, Weasley, how could he possibly have gotten through the portrait hole?”

“Ask him!” said Ron, pointing a shaking finger at the back of Sir Cadogan’s picture. “Ask him if he saw —”

Glaring suspiciously at Ron, Professor McGonagall pushed the portrait back open and went outside. The whole common room listened with bated breath.

“Sir Cadogan, did you just let a man enter Gryffindor Tower?”

REFERENCE INDEX: data structures that allow to narrow down the potential origin of a query sequence within the genome, saving both time and memory

Alignment of reads to the reference genome

We can view the list of produced files with the command:

```
ls -l
```

bwa has several subcommands which can be listed by simply running the command:

```
bwa
```

Alignment of reads to the reference genome

In this case we will use paired end reads produced with Illumina technology, so the reads will be in two different files.

To map the reads we use the following bwa command using the “mem” algorithm:

```
bwa mem -R "@RG\tID:sample\tLB:exome\tSM:sample\tPL:ILLUMINA" ref.fa  
1.fastq 2.fastq > mapping.sam
```

To see what the mapping.sam file contains:

```
less mapping.sam
```

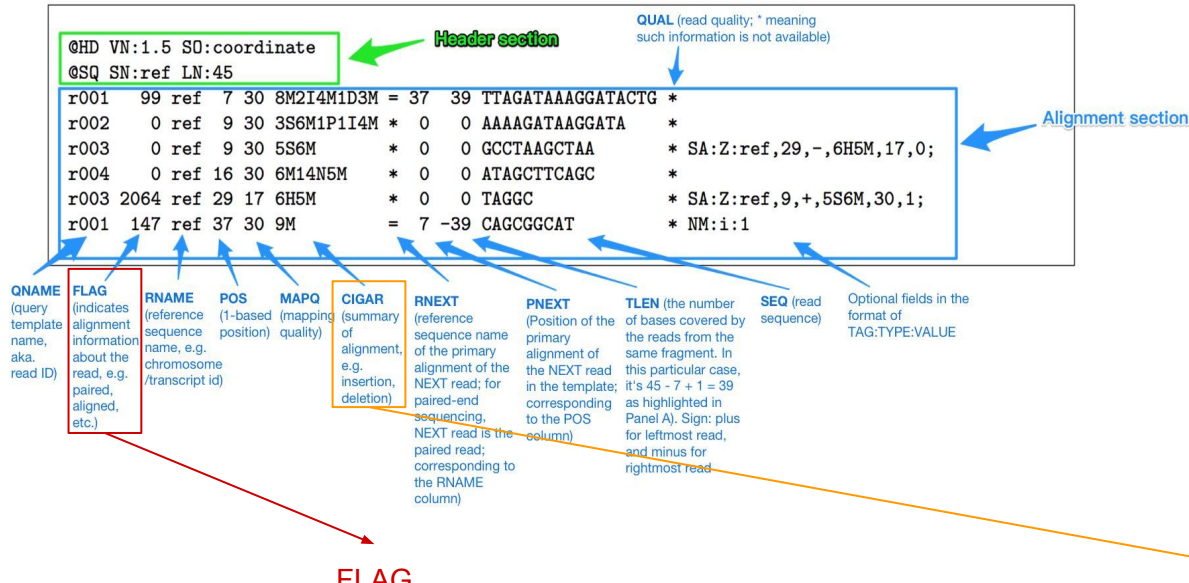

SAM Format

<https://samtools.github.io/hts-specs/SAMv1.pdf>

- TAB-delimited text
- header section (optional): lines start with '@'
- alignment section with 11 mandatory fields
- The BAM (=binary alignment map) is the compressed version of a SAM

QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	RNEXT	PNEXT	TLEN	SEQ	Optional fields in the format of TAG:TYPE:VALUE
(query template name, aka. read ID)	(indicates alignment information about the read, e.g. paired, aligned, etc.)	(reference sequence name, e.g. chromosome /transcript id)	(1-based position)	(mapping quality)	(summary of alignment, e.g. insertion, deletion)	(reference sequence name of the primary alignment of the NEXT read; for paired-end sequencing, NEXT read is the paired read; corresponding to the RNAME column)	(Position of the primary alignment of the NEXT read in the template; corresponding to the POS column)	(the number of bases covered by the reads from the same fragment. In this particular case, it's 45 - 7 + 1 = 39 as highlighted in Panel A). Sign: plus for leftmost read, and minus for rightmost read	(read sequence)	
Header section										
@HD VN:1.5 SO:coordinate										
@SQ SN:ref LN:45										
r001	99	ref	7	30	8M2I4M1D3M	= 37	39		TTAGATAAAGGATACTG	* SA:Z:ref,29,-,6H5M,17,0;
r002	0	ref	9	30	3S6M1P1I4M	* 0	0		AAAAGATAAGGATA	*
r003	0	ref	9	30	5S6M	* 0	0		GCCTAAGCTAA	*
r004	0	ref	16	30	6M14N5M	* 0	0		ATAGCTTCAGC	*
r003	2064	ref	29	17	6H5M	* 0	0		TAGGC	* SA:Z:ref,9,+,5S6M,30,1;
r001	147	ref	37	30	9M	= 7	-39		CAGCGGCAT	* NM:i:1

SAM Format



FLAG

there are 12 bits and each bit represents some information about the read as shown in the **Description** column

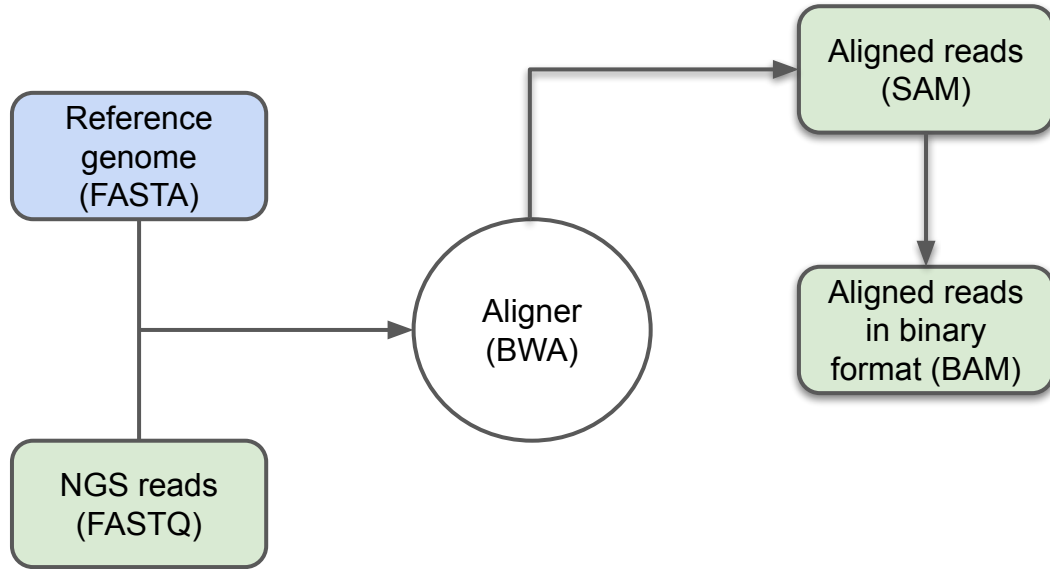
Bit	Description
1	0x1 template having multiple segments in sequencing
2	0x2 each segment properly aligned according to the aligner
4	0x4 segment unmapped
8	0x8 next segment in the template unmapped
16	0x10 SEQ being reverse complemented
32	0x20 SEQ of the next segment in the template being reverse complemented
64	0x40 the first segment in the template
128	0x80 the last segment in the template
256	0x100 secondary alignment
512	0x200 not passing filters, such as platform/vendor quality controls
1024	0x400 PCR or optical duplicate
2048	0x800 supplementary alignment

CIGAR STRING

It is a compressed representation of an alignment that is used in the [SAM file format](#).

Eg: 6M14N5M
 6 MATCH bases
 14 INTERVAL bases
 5 MATCH bases

A typical workflow for variant calling



Short reading alignment

Alignment of reads to the reference genome

SAM files are usually compressed into a binary format (not text and therefore only understandable by computers) called **BAM** and which stands for **binarySAM** . To perform this conversion, the program **samtools** is used , which is a software package for manipulating and extracting information from sam/bam files.

```
samtools view -b mapping.sam > mapping.bam
```

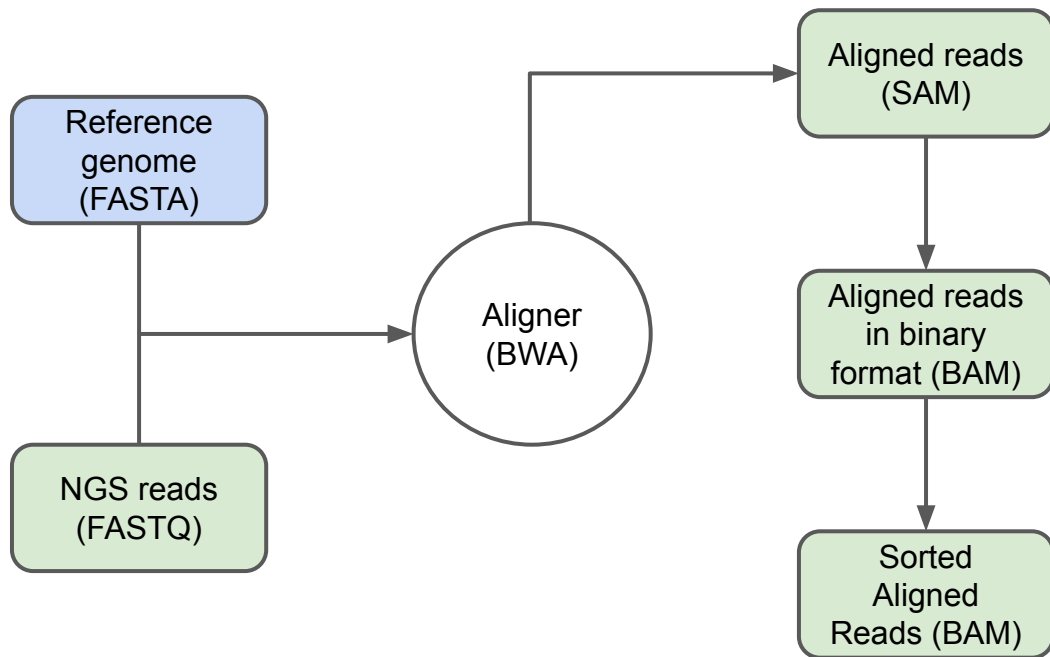
...

-b output BAM

...

Samtools is a set of utilities that manipulate alignments in the SAM (Sequence Alignment/Map), BAM, and CRAM formats. It converts between the formats, does sorting, merging and indexing, and can retrieve reads in any regions swiftly.

A typical workflow for variant calling



Short reading alignment

Alignment of reads to the reference genome

To sort the reads of the BAM file according to their position in the genome we use the samtools sort command:

```
samtools sort mapping.bam > sorted.bam
```

Now let's create the index for our sorted BAM file:

```
samtools index sorted.bam
```

View reads aligned to the reference genome

To view the reads aligned to the genome we use the samtools tview command:

```
samtools tview -p 1:2488138  
sorted.bam ref.fa
```

```
Program: samtools (Tools for alignments in the SAM format)  
Version: 1.14 (using htslib 1.14)  
  
Usage:  samtools <command> [options]  
  
Commands:  
  -- Indexing  
    dict          create a sequence dictionary file  
    faidx         index/extract FASTA  
    fqidx         index/extract FASTQ  
    index         index alignment  
  
  -- Editing  
    calmd         recalculate MD/NM tags and '=' bases  
    fixmate       fix mate information  
    reheader      replace BAM header  
    targetcut     cut fosmid regions (for fosmid pool only)  
    addreplacerg  adds or replaces RG tags  
    markdup       mark duplicates  
    ampliconclip  clip oligos from the end of reads  
  
  -- File operations  
    collate       shuffle and group alignments by name  
    cat           concatenate BAMs  
    merge         merge sorted alignments  
    mpileup       multi-way pileup  
    sort          sort alignment file  
    split         splits a file by read group  
    quickcheck    quickly check if SAM/BAM/CRAM file appears intact  
    fastq         converts a BAM to a FASTQ  
    fasta         converts a BAM to a FASTA  
    import        Converts FASTA or FASTQ files to SAM/BAM/CRAM  
  
  -- Statistics  
    bedcov        read depth per BED region  
    coverage      alignment depth and percent coverage  
    depth         compute the depth  
    flagstat      simple stats  
    idxstats     BAM index stats  
    phase         phase heterozygotes  
    stats         generate stats (former bamcheck)  
    ampliconstats generate amplicon specific stats  
  
  -- Viewing  
    flags         explain BAM flags  
    tview         text alignment viewer  
    view          SAM<->BAM<->CRAM conversion  
    depad         convert padded BAM to unpadded BAM  
    samples       list the samples in a set of SAM/BAM/CRAM files  
  
  -- Misc  
    help [cmd]    display this help message or help for [cmd]  
    version        detailed version information
```

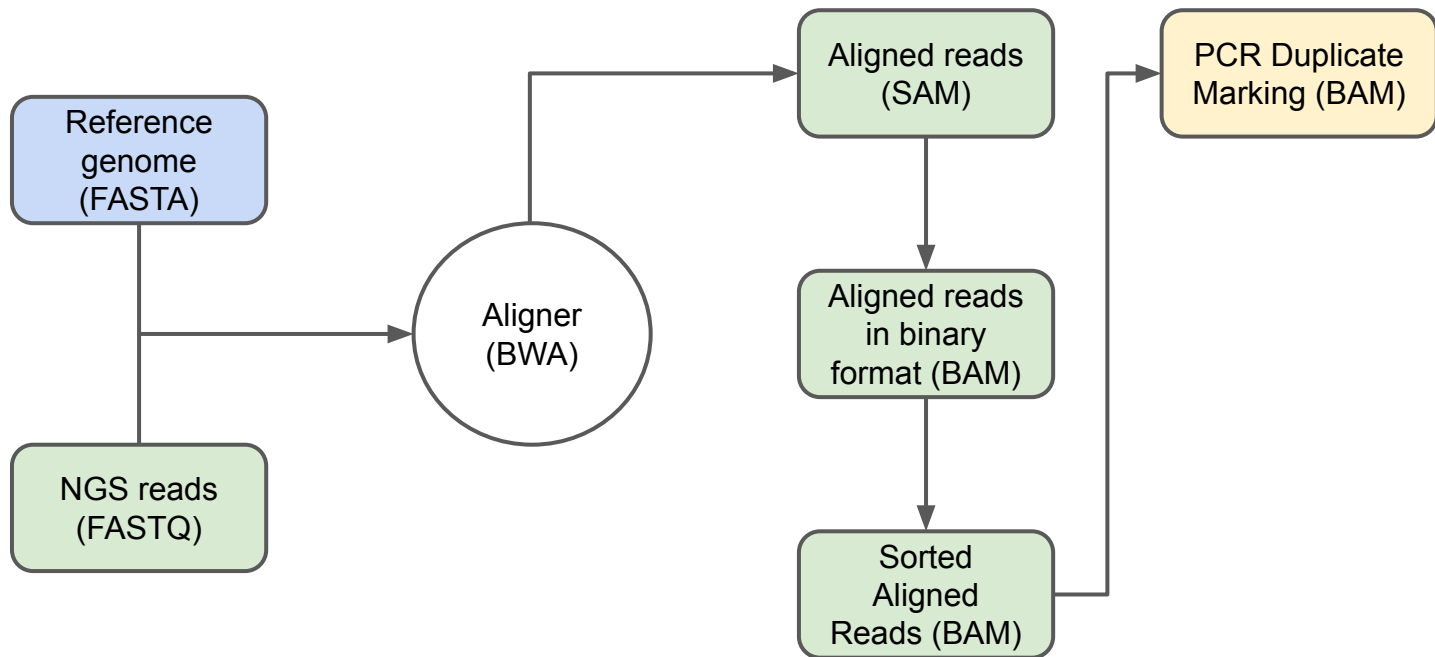
-p chr:pos

View reads aligned to the reference genome

A widely used program for viewing alignments in various formats, including sam/bam, is IGV (Integrative genomics viewer) which can be downloaded from the following address: <http://www.broadinstitute.org/igv/> (load sorted.bam)



A typical workflow for variant calling



Short reading alignment

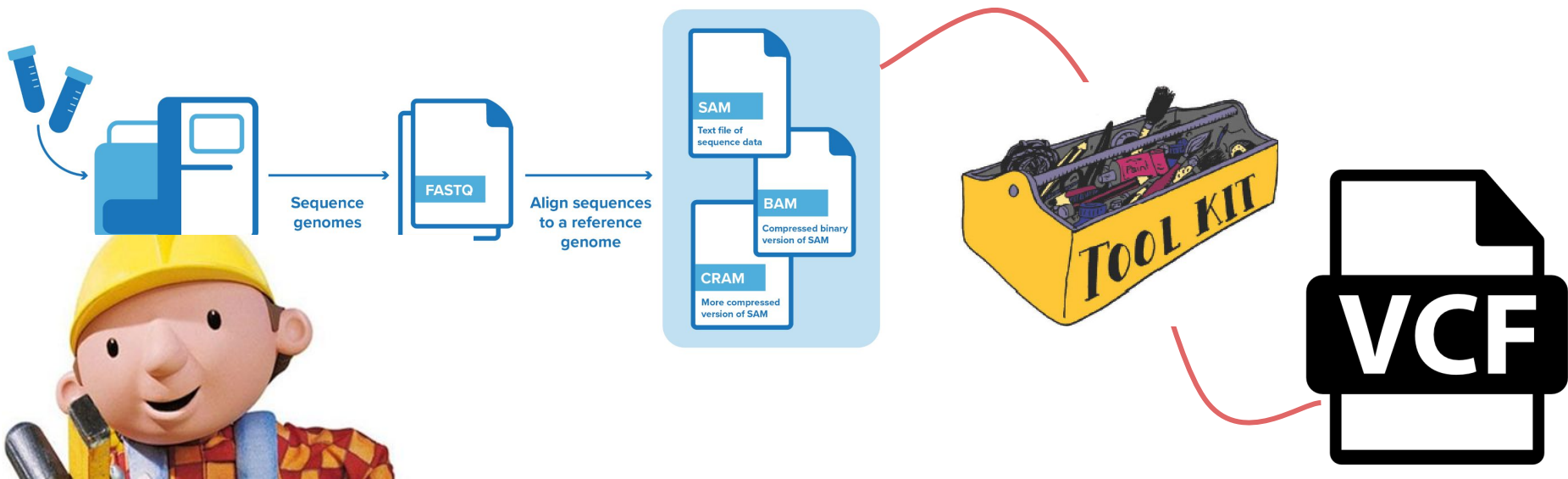
Alignment
refinement

Genome Analysis Tool Kit

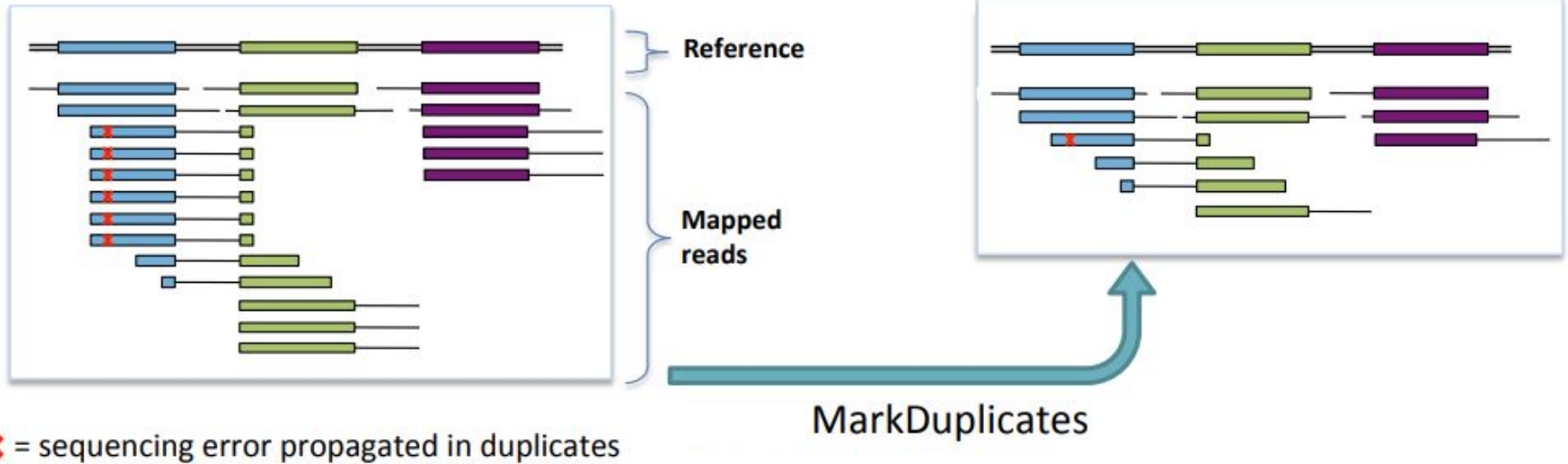


"Gee-ay-tee-kay" (/dʒi•er•ti•keɪ/) and **not** "Gat-kay" (/gæt•keɪ/)

It is a collection of **command-line tools** for analyzing high-throughput sequencing data with a primary focus on variant discovery. The tools can be used individually or chained together into complete workflows.



Removal of PCR duplicates



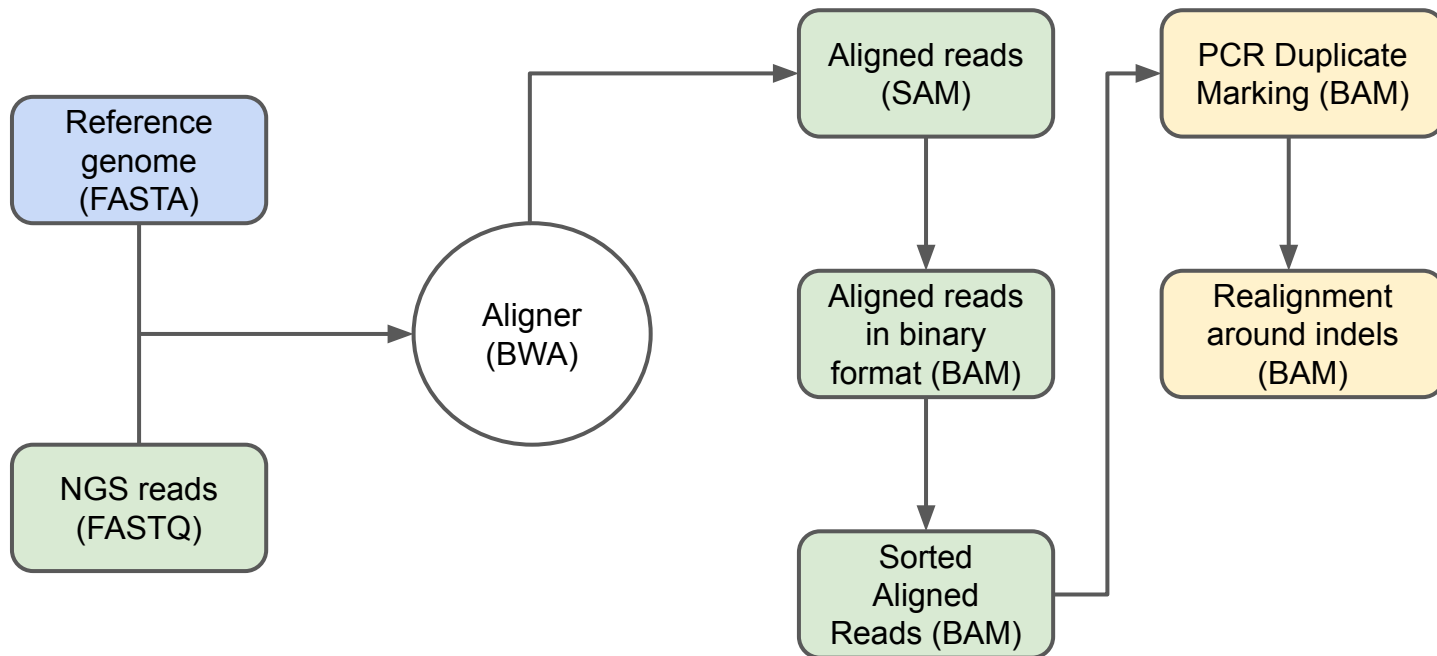
Duplicates can falsify high coverage leading to false “calls”.

Removal of PCR duplicates

To remove PCR duplicates we use a function of the GATK4 program:

```
gatk MarkDuplicates -I sorted.bam -O nodup.bam -M metrics.txt  
-REMOVE_DUPLICATES true -CREATE_INDEX true
```

A typical workflow for variant calling

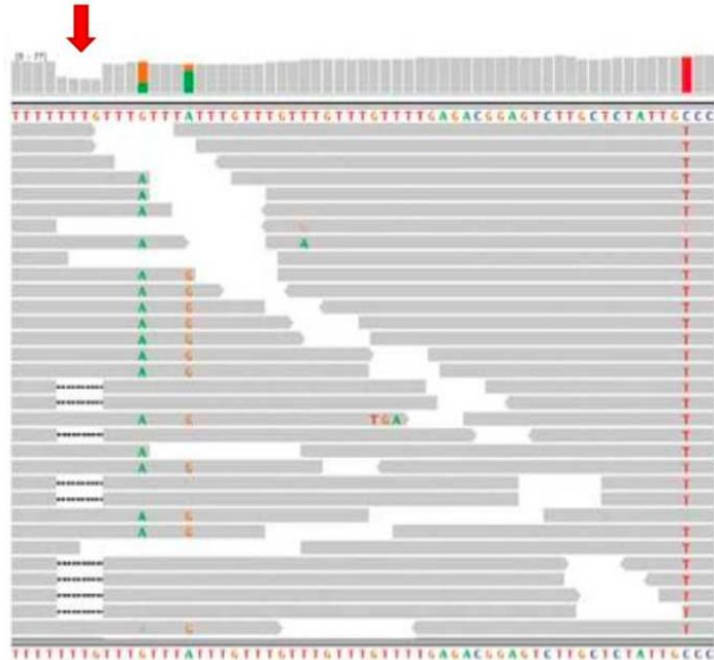


Short reading alignment

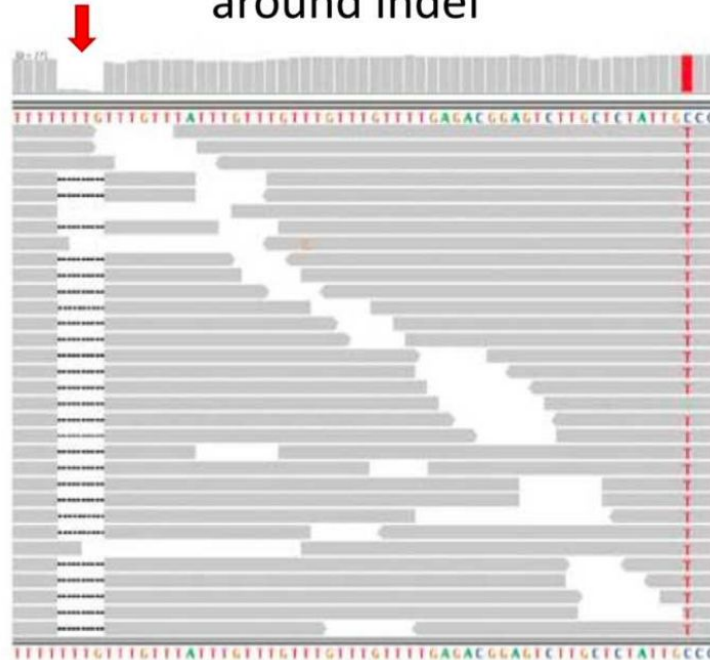
Alignment
refinement

Local realignment around indels

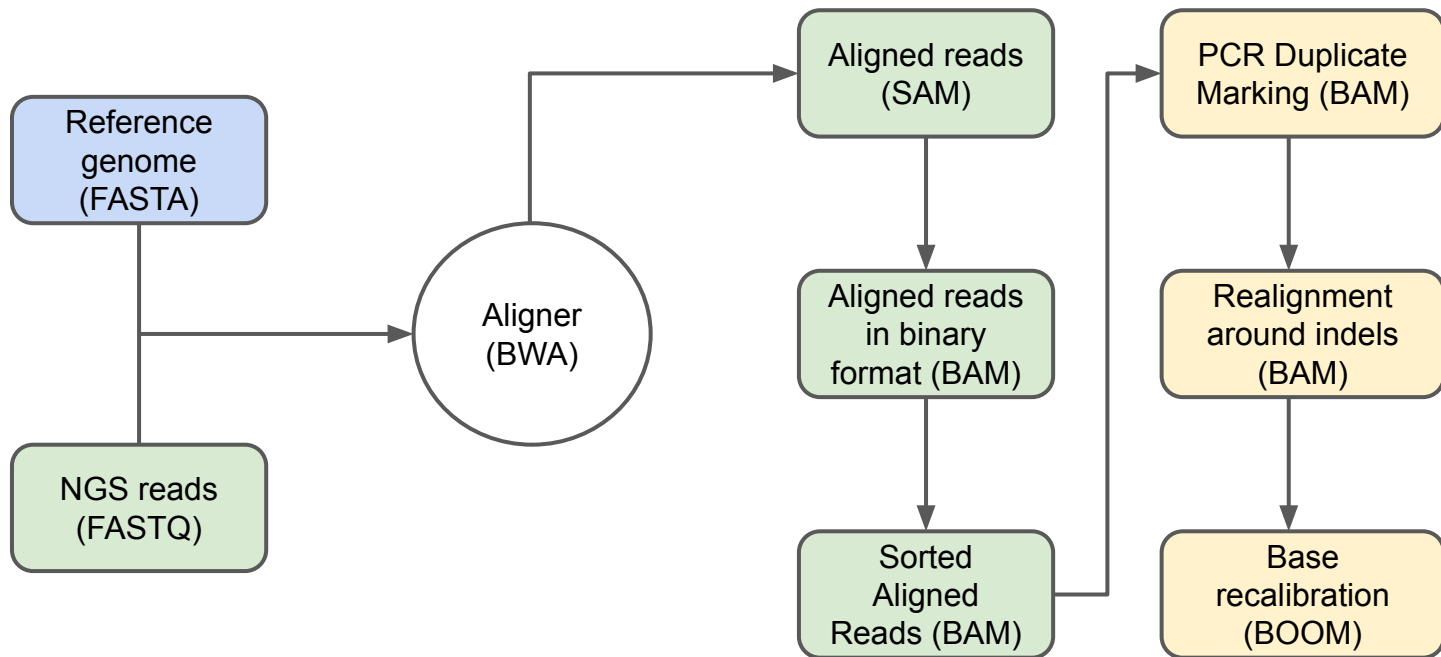
Raw BWA alignment



After local realignment around indel



A typical workflow for variant calling



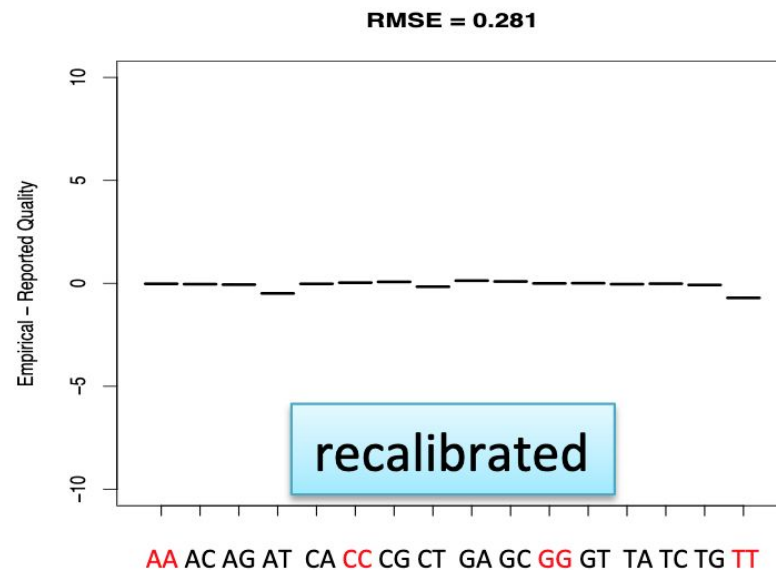
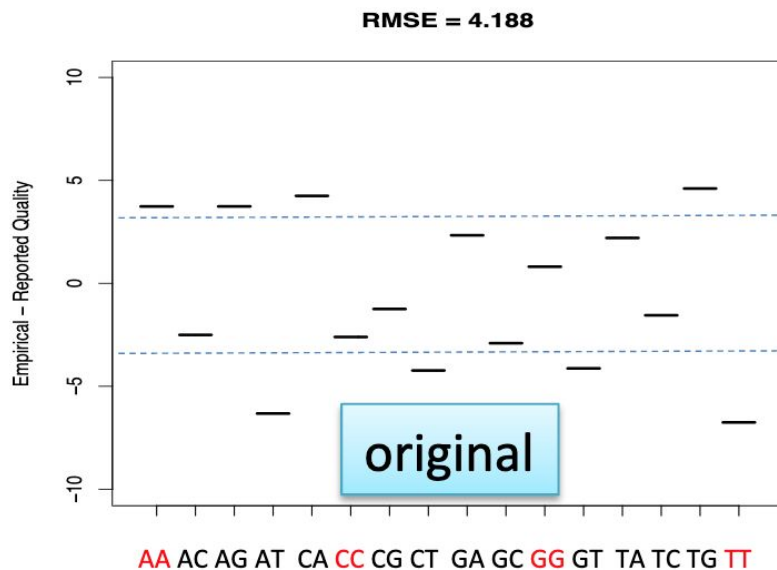
Short reading alignment

Alignment
refinement

Recalibration of the quality of the bases

Base quality scores are critical for variant calling but there are systematic biases that influence it

Example of bias: qualities reported depending on nucleotide context



Recalibration of the quality of the bases

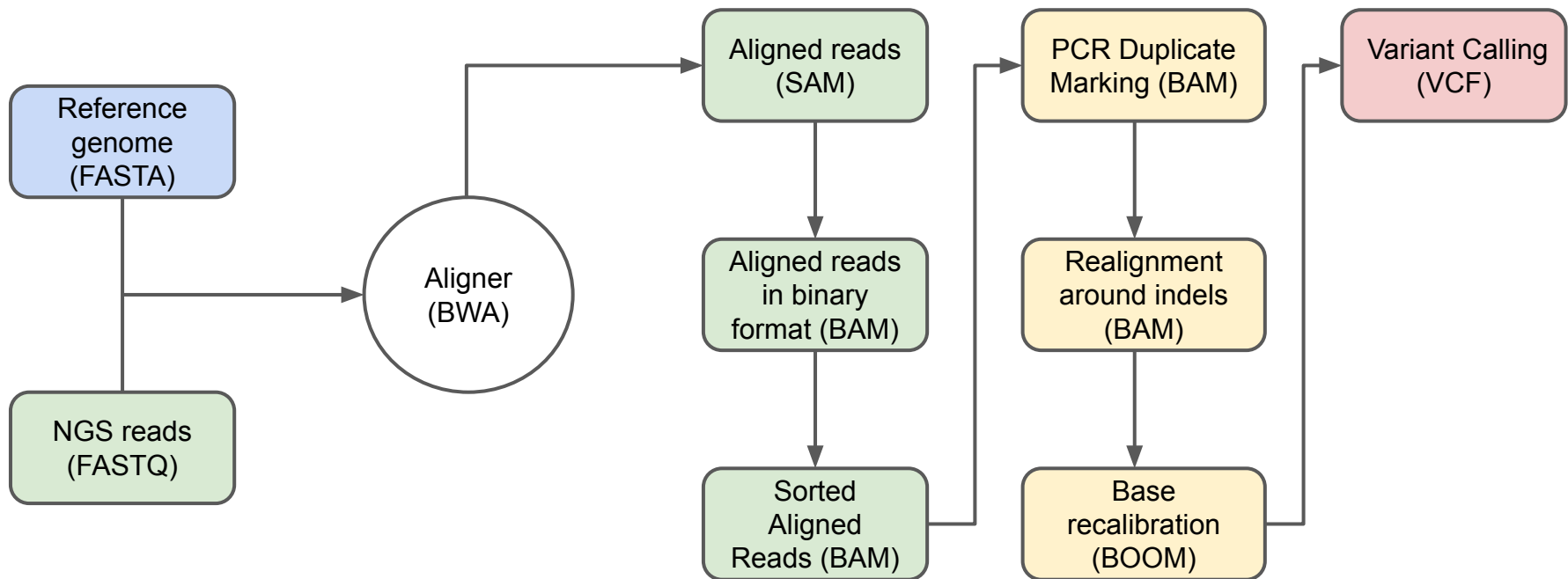
To build the model on which to recalibrate the quality we use the BaseRecalibrator function of GATK4:

```
gatk BaseRecalibrator -I nodup.bam -R ref.fa --known-sites  
dbSNP.vcf.gz -O model.grp
```

To recalibrate the quality we use the ApplyBQSR function of GATK4:

```
gatk ApplyBQSR -R ref.fa -I nodup.bam -bqsr model.grp -O  
recalibrated.bam
```

A typical workflow for variant calling



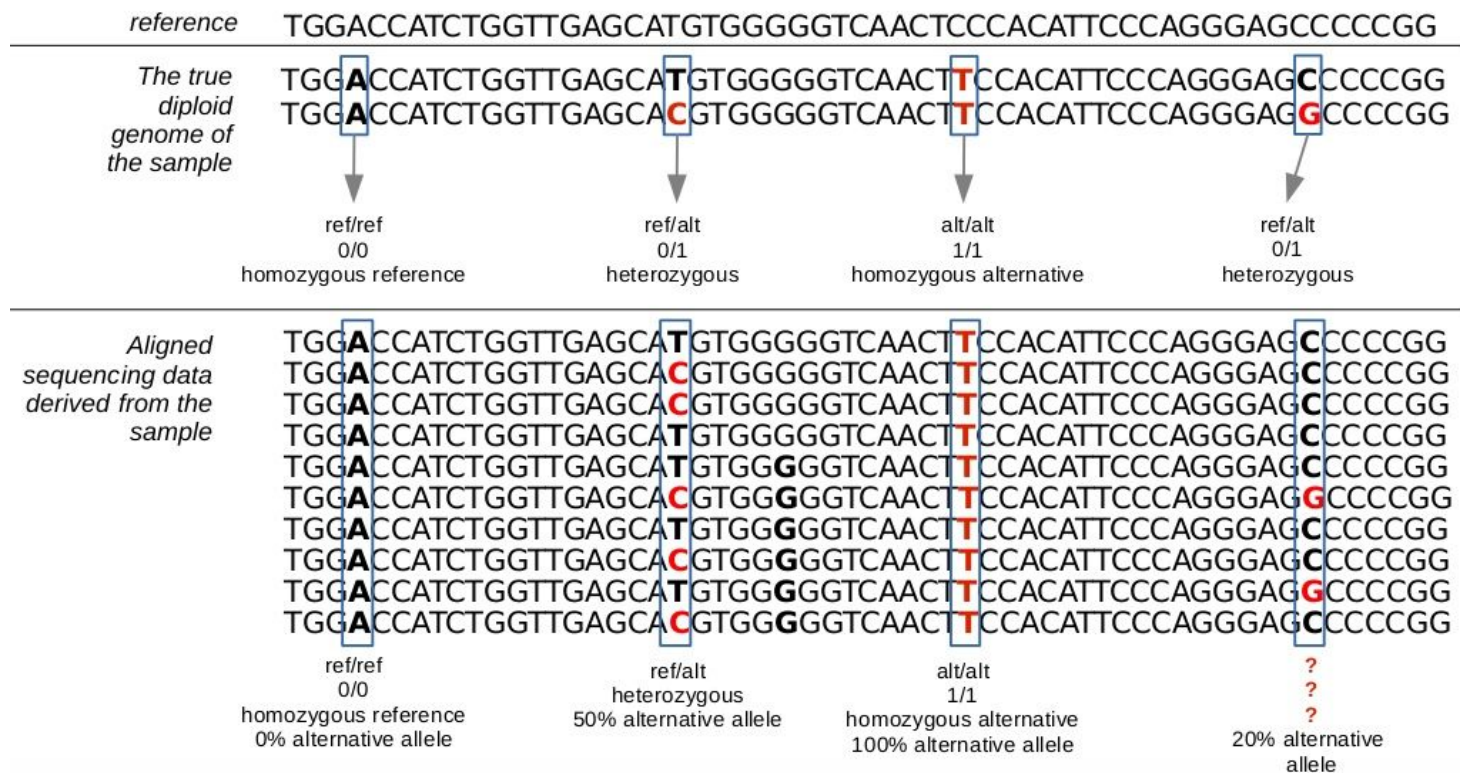
Short reading alignment

Alignment
refinement

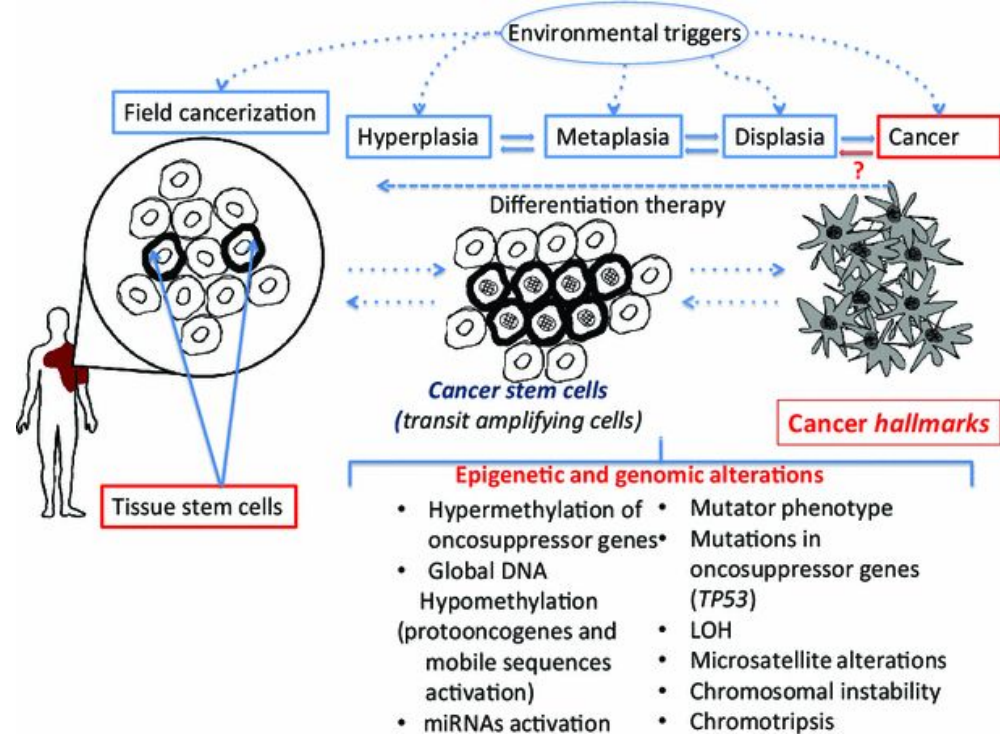
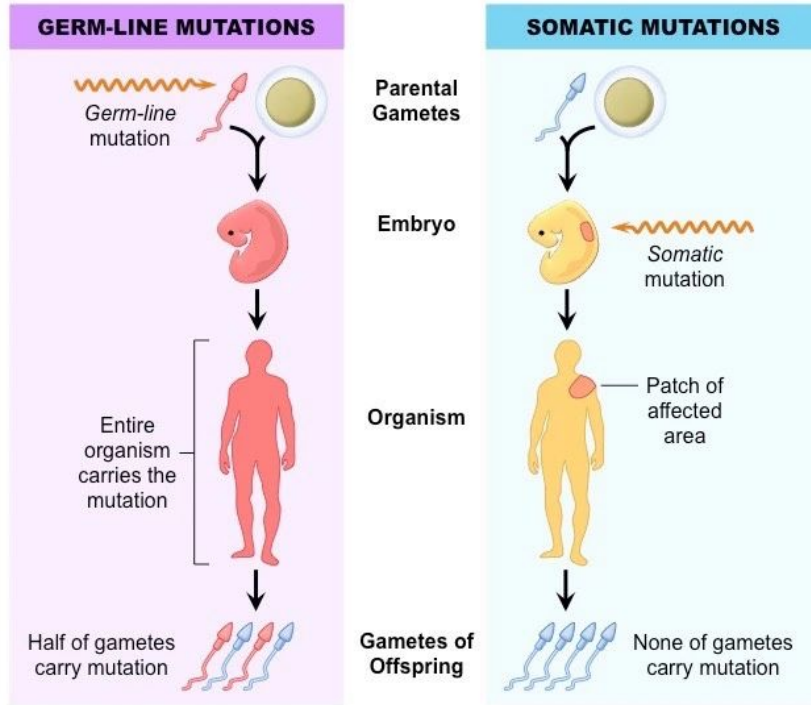
Variant detection

What does “Variant calling” mean?

Identify genetic differences by comparing sequenced reads to a reference genome



SOMATIC VS GERMLINE MUTATIONS



Variant Calling

Calling variants can be done using the GATK4 program HaplotypeCaller:

```
gatk HaplotypeCaller -R ref.fa -I recalibrated.bam -O germline.vcf
```

For calling somatic variants, the Mutect2 program from GATK4 can be used:

```
gatk Mutect2 -R ref.fa -I recalibrated.bam -O somatic.vcf
```

VCF file format

VCF header

```
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
```

Mandatory header lines

Optional header lines (meta-data about the annotations in the VCF body)

Body

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE1	SAMPLE2
1	1	.	ACG	A,AT	.	PASS	.	GT:DP	1/2:13	0/0:29
1	2	rs1	C	T,CT	.	PASS	H2;AA=T	GT:GQ	0 1:100	2/2:70
1	5	.	A	G	.	PASS	.	GT:GQ	1 0:77	1/1:95
1	100	.	T		.	PASS	SVTYPE=DEL;END=300	GT:GQ:DP	1/1:12:3	0/0:20

Deletion

SNP

Large SV

Insertion

Other event

Reference alleles (GT=0)

Alternate alleles (GT>0 is an index to the ALT column)

Phased data (G and C above are on the same chromosome)

VCF file format

- Lines starting with `##`: arbitrary number of meta-information lines
- Line starting with `#`: column definition (8 mandatory):
 - CHROM = chromosome
 - POS = start position of the variant
 - ID = unique identifier of the variant (eg Number for SNPs)
 - REF = reference allele
 - ALT = comma separated list of alternate alleles
 - QUAL = phred-scaled quality score
 - FILTER = site filtering information
 - INFO = user extensible annotation (eg snpEff, Annovar)
 - • FORMAT = an (optional) extensible list of fields for describing the SAMPLE column
 - • SAMPLE COLUMN = free

GERMLINE variant technical filtering

Filter germline variants with low coverage, low mapping quality, low variant call quality:

```
gatk VariantFiltration -V germline.vcf -filter "QUAL < 30.0"  
--filter-name "QUAL30" -filter "MQ < 40.0" --filter-name "MQ40"  
-filter "DP < 30" --filter-name "DP30" -O germline_filtered.vcf
```

```
gatk SelectVariants -R ref.fa -V germline_filtered.vcf  
--exclude-filtered -O germline_selected.vcf
```

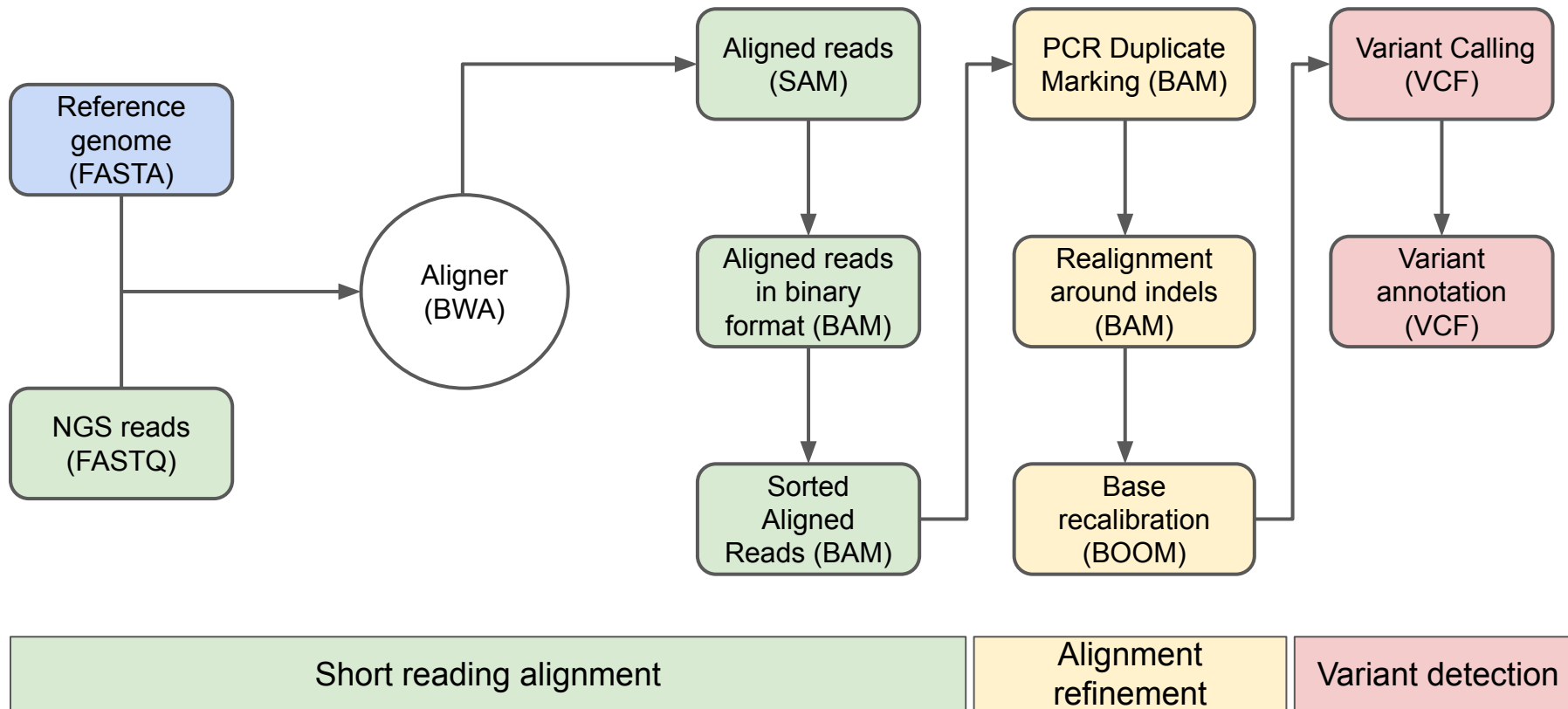

SOMATIC variant technical filtering

There is a function (FilterMutectCalls) in GATK4 which is used to eliminate false positives by considering various technical parameters:

```
gatk FilterMutectCalls -R ref.fa -V somatic.vcf -O somatic_filtered.vcf
```

```
gatk SelectVariants -R ref.fa -V somatic_filtered.vcf --exclude-filtered  
-O somatic_selected.vcf
```

A typical workflow for variant calling



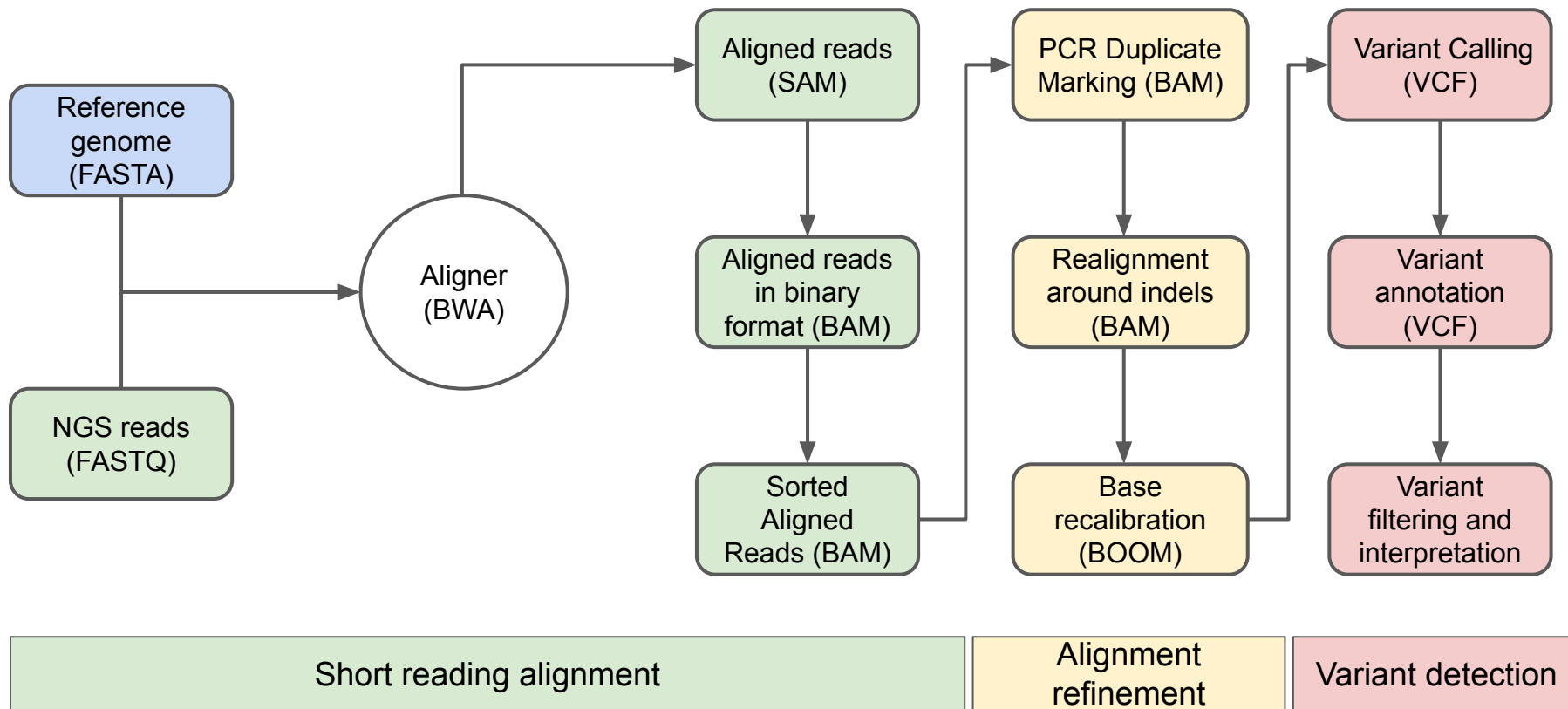
VCF annotation

Annotation of variants for dbSNP can be done with the VariantAnnotator function of GATK4:

```
gatk VariantAnnotator -R ref.fa -V germline_selected.vcf -O  
germ_annotated.vcf --dbsnp dbSNP.vcf.gz
```

```
gatk VariantAnnotator -R ref.fa -V somatic_selected.vcf -O  
som_annotated.vcf --dbsnp dbSNP.vcf.gz
```

A typical workflow for variant calling



VCF annotation

Look at the alignment of these two variants:

```
samtools tview -p 1:2488138 recalibrated.bam ref.fa
```

```
samtools tview -p 15:66727453 recalibrated.bam ref.fa
```

In the VCF file containing the somatic variants search for these two variants by their genomic position and retrieve the dbSNP ID in the “ID” column:

- 1:2488138 G>A (rs*****?; Het/Hom? Variant Allele Frequency?)
- 15:66727453 A>G (rs*****?; Het/Hom?; Variant Allele Frequency?)

VCF annotation

In the VCF file containing the somatic variants search for these two variants by their genomic position and retrieve the dbSNP ID in the “ID” column:


- 1:2488138 G>A (rs768520625, Het, 0.107)
- 15:66727453 A>G (rs397516790, Het, 0.112)

In which gene are they found? Do they impact the function of the protein? What is their allelic frequency in the population?.....





See annotations in dbSNP

dbSNP is a variant database and we search for the above two variants by their ID:

- Go to the dbSNP website: <https://www.ncbi.nlm.nih.gov/snp/>
- Search for the ID “rs768520625” and then “rs397516790”


Welcome to the Reference SNP (rs) Report
 All alleles are reported in the [Forward orientation](#). Click on the [Variant Details tab](#) for details on Genomic Placement, Gene, and Amino Acid changes. HGVS names are in the [HGVS tab](#).

Reference SNP (rs) Report

[Download](#)





[Switch to classic site](#)

rs768520625


Organism

Homosapiens

Clinical Significance

Not Reported in ClinVar

Position

chr1:2556699 (GRCh38.p12) 

Alleles

G>A

Variation Type

SNV Single Nucleotide Variation

Frequency

A=0.000004 (1/239996, GnomAD_exome)
 A=0.00001 (1/91792, ExAC)

Gene : Consequence

TNFRSF14 : Stop Gained
TNFRSF14-AS1 : Intron Variant

Publications

0 citations

Genomic View

[See rs on genome](#)

Current Build 154

Released April 21, 2020

Variant Details

Genomic Placements

Clinical Significance

Frequency

HGVS

Submissions

History

Publications

Flanks

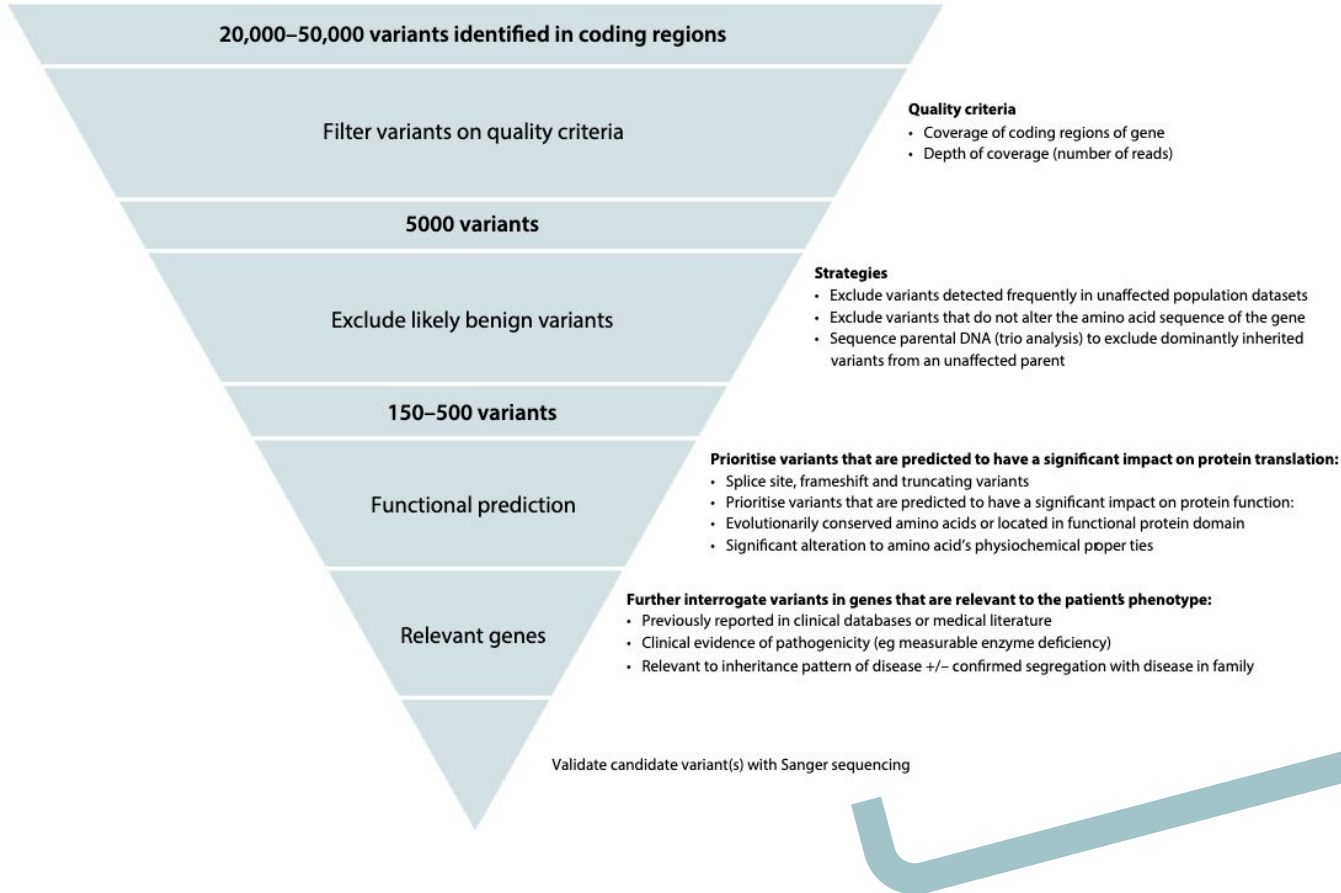
Sequence name	Change
GRCh37.p13 chr1	NC_000001.10:g.2488138G>A
GRCh38.p12 chr1	NC_000001.11:g.2556699G>A
GRCh38.p12 chr1 alt locus HSCHR1_1_CTG3	NT_187515.1:g.107889G>A
TNFRSF14 RefSeqGene	NG_047096.1:g.5335G>A

Gene: [TNFRSF14](#), TNF receptor superfamily member 14 (plus strand)

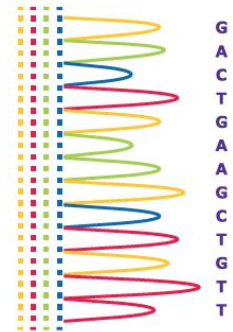
Molecule type	Change	Amino acid[Codon]	SO Term
TNFRSF14 transcript variant 1	NM_003820.3:c.35G>A	W [TGG] > * [TAG]	Coding Sequence Variant

FEEDBACK

Example of variant filtering

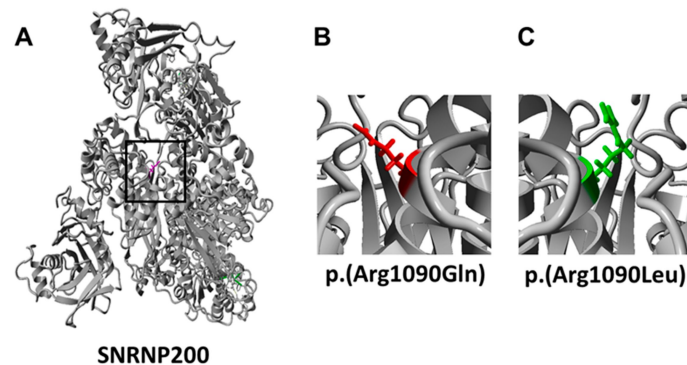
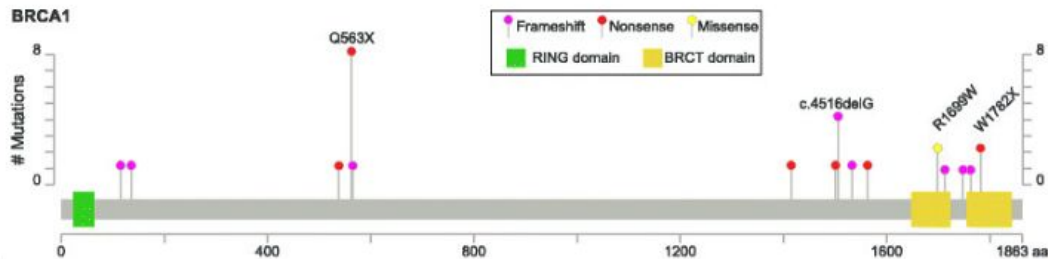
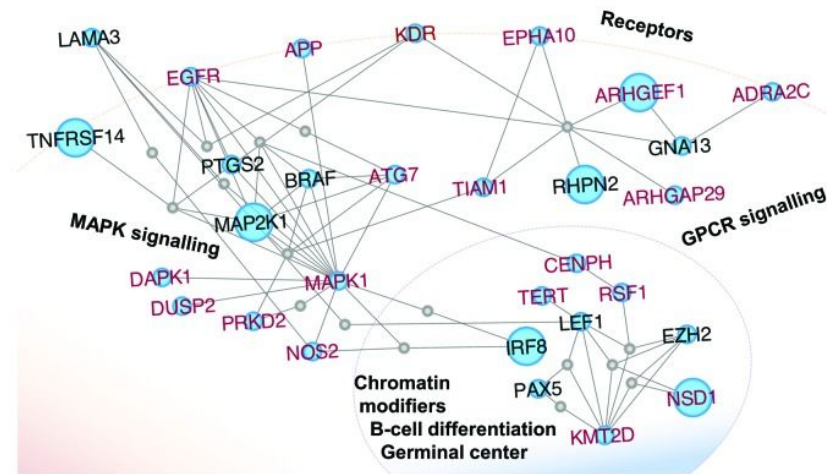
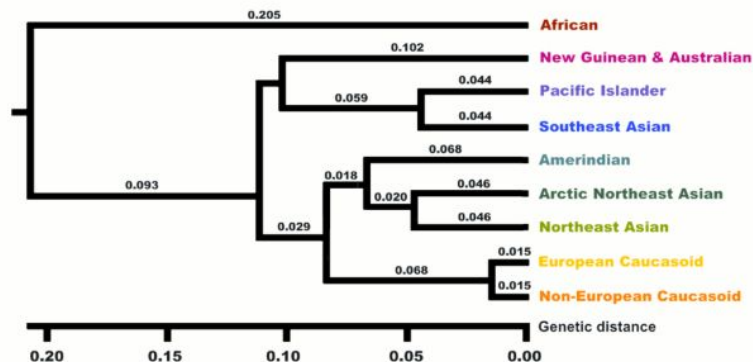


Sanger sequencing



Output chromatogram

Examples of subsequent analysis



Cases to analyze

- Analysis of a human exome affected by pediatric follicular lymphoma to identify germline or somatic variants that could be associated with the disease
- Exome analysis obtained from ancient DNA of a 14th century Venetian nobleman who died under mysterious circumstances.

Case Description

- Noble leader who died around 1300 and was buried in a marble tomb which favoured the mummification and preservation of the body.



- In 2004 the remains were studied and the historical documents reconsidered, discovering several things.

Historical data

- At the age of 23 he began to suffer from fatigue, fever, muscle cramps, breathing and heart problems which, after intense efforts, led him to abandon the battlefield several times despite being an experienced warrior.
- At the age of 34 he was ill for a long time, so much so that he was considered dead.
- He died in 1300 at the age of 38 after 3 days of fever and generic “flow” (hemorrhage? nausea? intestinal disease?)
- Some contemporary documents attribute death to poisoning, others to "congestion" after a hunting trip.

Data on remains (2004)

- He had no injuries
- Toxicological tests found traces of *Digitalis purpurea* .
- Digoxin is a potent poison, a cardiac glycoside that increases the force of contraction of the heart .
- Used for therapeutic purposes for arrhythmias and heart failure



Objectives of the study

- Exome sequencing data analysis from ancient DNA
 - **Find out the most likely cause of death**

Objectives of the study

- Check the exome for genetic variants that can be linked to a disease that the nobleman could have suffered from by following the guide **Practical_session_5_guide.pdf**
 - How many germline variants are found?
 - Do we find the variants in position 17:78078656 and 17:78084553 in the VCF?
 - If so, what is the alternative allele? What is their allele frequency? Are they heterozygous or homozygous? Are they in cis or in trans?
 - Do they have a dbSNP ID? If so, what are they?
 - Which gene do they affect? What is the function of the protein produced by this gene? The functional consequence (missense/synonymous)?
 - What is their frequency in the population? Are they rare or polymorphisms?
 - Do they have clinical significance noted in ClinVar? (Benign, Pathogenic, or Unknown) In which disease?
 - In ClinVar for the variant 17:78078656 is there a link for OMIM?
 - Is there a gene/phenotype relationship? Which one? Recessive or dominant syndrome?
 - Final thoughts? Diagnosis?

Solution

Check the exome for genetic variants that can be linked to a disease that the nobleman could have suffered from by following the guide **Exercise5_guide.pdf**

- How many germline variants are found? **35**
- Do we find the variants in position 17:78078656 and 17:78084553 in the VCF? **Yes**
- If so, what is the alternative allele? What is their allele frequency? Are they heterozygous or homozygous? Are they in cis or in trans? **A, 51.4% & 46.2%, trans**
- Do they have a dbSNP ID? If so, what are they? **rs1800299 ; rs398123169**
- Which gene do they affect? What is the function of the protein produced by this gene? The functional consequence (missense/synonym)? **GAA, glycogen breakdown, missense**
 - What is their frequency in the population? Are they rare or polymorphisms? **3% and rare**
 - Do they have a clinical significance noted in ClinVar? (Benign, Pathogenic?) In which disease? **Pathogenic and benign, Glycogen storage disease, type II**
 - In ClinVar for the variant 17:78078656 is there a link for OMIM? **Yes, 606800.0001**
 - Is there a gene/phenotype relationship? Which one? Recessive or dominant inheritance? **Glycogen storage disease II (GSD2) is an autosomal recessive**