

CORSO DI METODI MOLECOLARI E BIOINFORMATICA

LM Biologia Evoluzionistica, Università di Padova

Dr. Enrico Gaffo, Dr. Silvia Orsi,

Prof. Stefania Bortoluzzi

Esercitazione 5 **“Analisi dati NGS: Variant calling”**

Padova, 23 Novembre 2023

Obiettivo dell'esercitazione

- Familiarizzare con i formati di file essenziali per immagazzinare i dati di DNA-seq
- Effettuare i passaggi essenziali dell'analisi di dati DNA-seq ottenuti tramite Next-Generation Sequencing per rilevare varianti genomiche (“Variant calling”)

Casi da analizzare

- Analisi di un esoma umano affetto da linfoma follicolare pediatrico per individuare varianti germline o somatiche che potrebbero essere associate alla malattia
- Analisi dell'esoma ottenuto da DNA antico di un nobile veneto del 1300 morto in circostanze misteriose.

Casi da analizzare

- Analisi di un esoma umano affetto da linfoma follicolare pediatrico per individuare varianti germline o somatiche che potrebbero essere associate alla malattia
- Analisi dell'esoma ottenuto da DNA antico di un nobile veneto del 1300 morto in circostanze misteriose.

Preparazione dell'esercitazione

- Scaricare il file “esercitazione5.zip” da
http://compgen.bio.unipd.it/~stefania/Didattica/AA2023-2024/MMOL_BIOINFO_BE/esercitazione5.zip
- Decomprimere il file “esercitazione5.zip”:

?

- Spostarsi nella cartella esercitazione5:

?

Preparazione dell'esercitazione

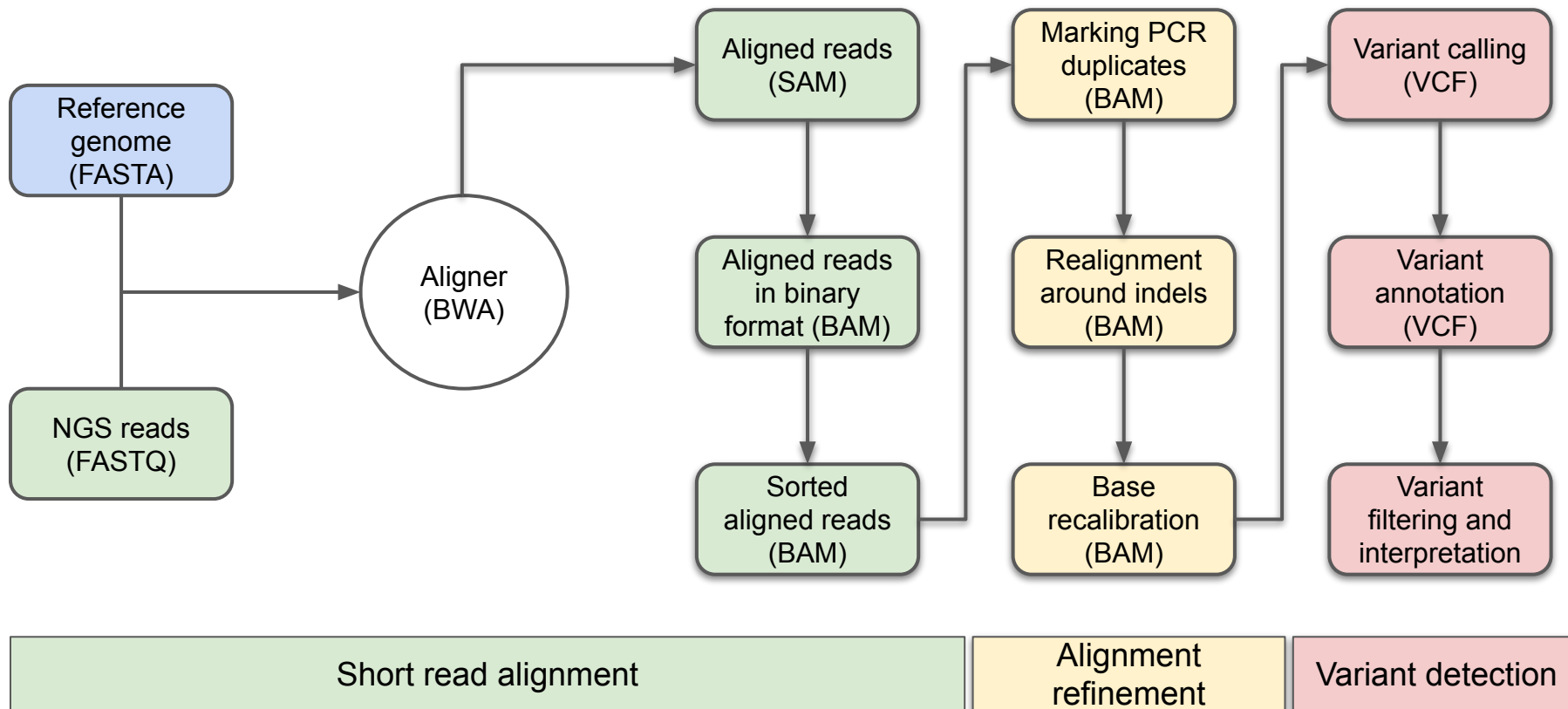
- Scaricare il file “esercitazione5.zip” da http://compgen.bio.unipd.it/~stefania/Didattica/AA2022-2023/MMOL_BIOINFO_BE/esercitazione5.zip
- Decomprimere il file “esercitazione5.zip” usando il comando **unzip**:

```
unzip esercitazione5.zip
```

- Spostarsi nella cartella esercitazione5:

```
cd esercitazione5
```

A typical workflow for variant calling

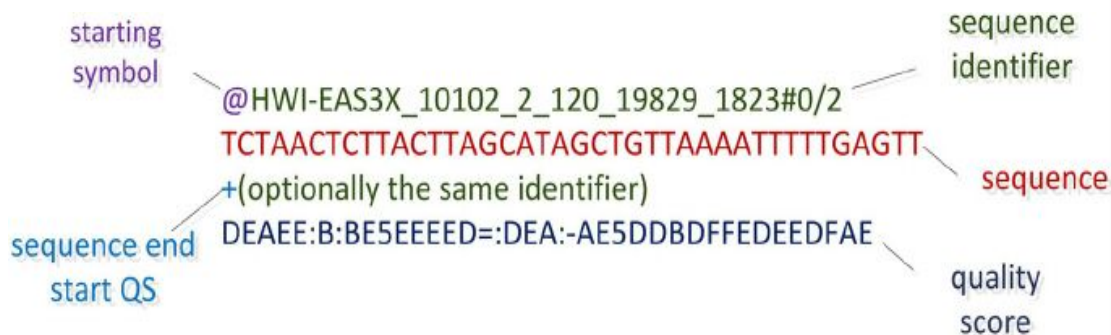


FASTQ format

Per vedere un 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?BBBBBBBFBFFHHHHFFHHHHFH09EFFHDFEFEG@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
CCTATCATATATGCCTTAGTTTGTGAAANATATTGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
+
??AA?BBBEDDEEEEGGGGGIHHIIII#7AFHII#####5#####
@M70273:8:000000000-AJLMP:1:1101:18171:2095 1:N:0:1
TTGTGATTCCACATTCTCTTCCATTGTAGNGCAAATNNNNNNNNNNNNNNNNNNNNNTNNTCNTTNNNTN
+
?????BBBDDDDDDDDGGGGGIIHHI#7AEFHI#####7###55#55###5###
@M70273:8:000000000-AJLMP:1:1101:19337:2095 1:N:0:1
GCCCCCATGCGCGGCATGATGAACCTCCGCTGCTGNNNNNNNNNNNNNNNNNNNTNNTTNTTNNNCAN
+
?????ABAADDDDDDDFFFFFIIHHIIHHHHHHI#####5###55#55###44#
@M70273:8:000000000-AJLMP:1:1101:14484:2097 1:N:0:1
CTGACTGATATGTGATTATTCTTTCAACAGCCACGCCAGATCCAGTGAAAAACAAGCTCTCATGTC
+
?????A?BB?DDDDDDBGGGGGGIIHHIIIIHHHHHHHFGFHHHHHHHHHHHHHHHHHHHHHHHH
@M70273:8:000000000-AJLMP:1:1101:16321:2100 1:N:0:1
TAGATGCTTTTAACTAAGTTACCTGACTTNCCTTATNNNNNNNNNNNNNNNNNNNTNNGCNGCNNCNCN
+
?????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% |

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

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

Preparazione del genoma di riferimento

Scaricare i file fasta dei cromosomi umani 1, 15 e 17 dal link di Ensembl

http://ftp.ensembl.org/pub/grch37/release-104/fasta/homo_sapiens/dna/

Spostarsi nella cartella “Scaricati”:

?

Spostare i file dei cromosomi dentro alla cartella “esercitazione5”:

?

Unire i cromosomi in unico file FASTA:

?

Preparazione del genoma di riferimento

Spostarsi nella cartella “Scaricati”:

```
cd ..
```

Spostare i file dei cromosomi dentro alla cartella “esercitazione5”:

```
mv Homo_sapiens.GRCh37.dna.chromosome.1* ./esercitazione5/
```

Unire i cromosomi in unico file FASTA:

```
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
```

Preparazione del genoma di riferimento

Controllare che i cromosomi siano nell'ordine giusto:

?

Preparazione del genoma di riferimento

Controllare che i cromosomi siano nell'ordine giusto:

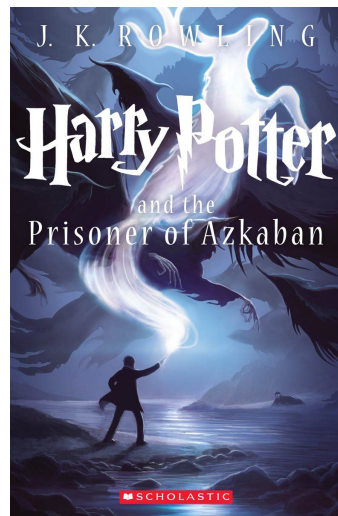
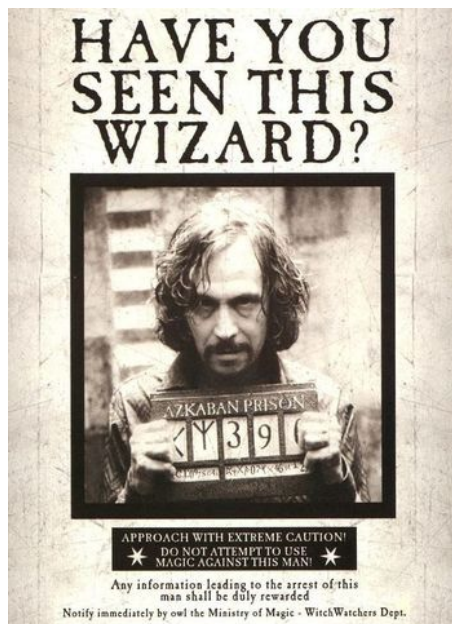
```
grep ">" ref.fa
```

Creare gli indici della sequenza di riferimento con 3 comandi:

```
bwa index ref.fa
```

```
/opt/samtools/bin/samtools faidx ref.fa
```

```
/opt/gatk/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

Allineamento delle reads al genoma di riferimento

Possiamo visualizzare l'elenco dei file prodotti con il comando:

```
ls -l
```

bwa ha vari sottocomandi che possono essere elencati lanciando semplicemente il comando:

```
bwa
```

Allineamento delle reads al genoma di riferimento

In questo caso useremo reads paired end prodotti con la tecnologia Illumina, quindi le reads si troveranno in due file diversi.

Per mappare le reads usiamo il seguente comando di bwa usando l'algoritmo "mem":

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

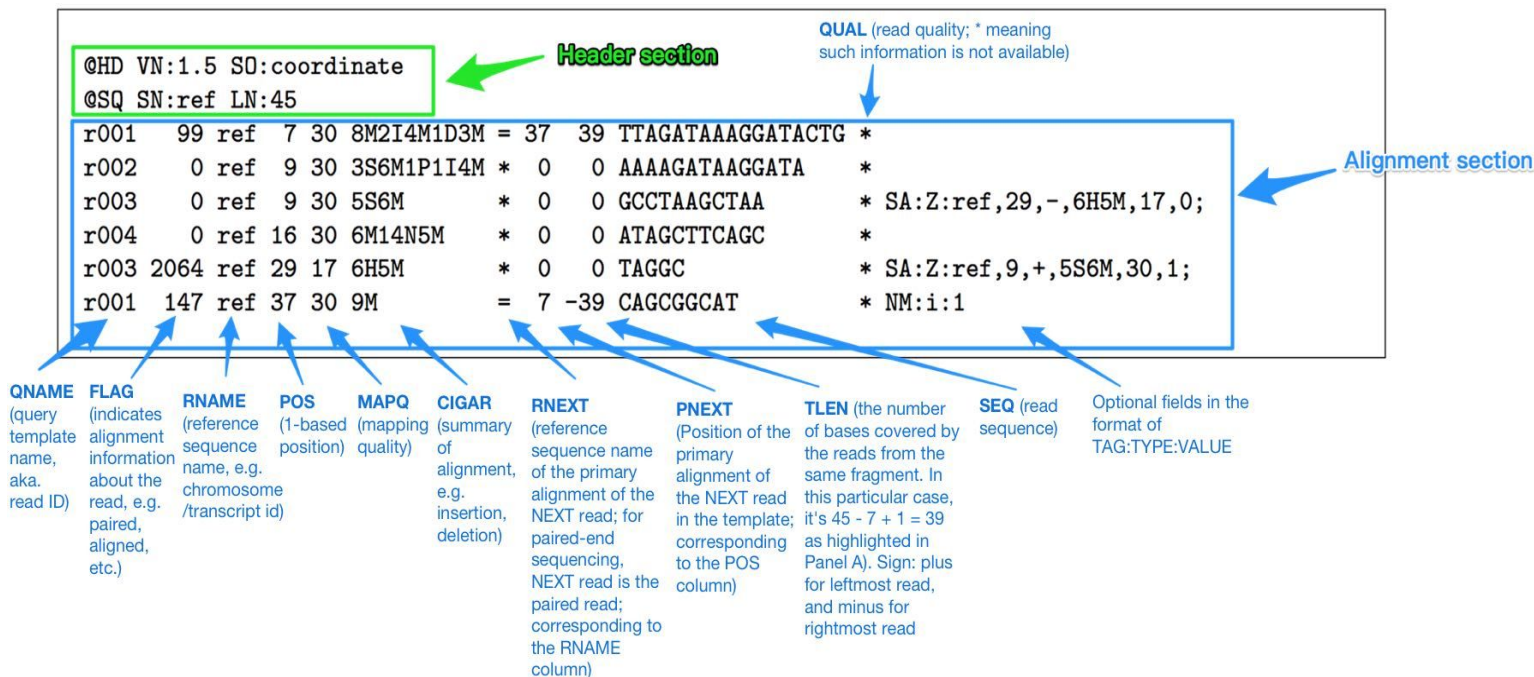
Per vedere cosa contiene il file mapping.sam:

```
less mapping.sam
```

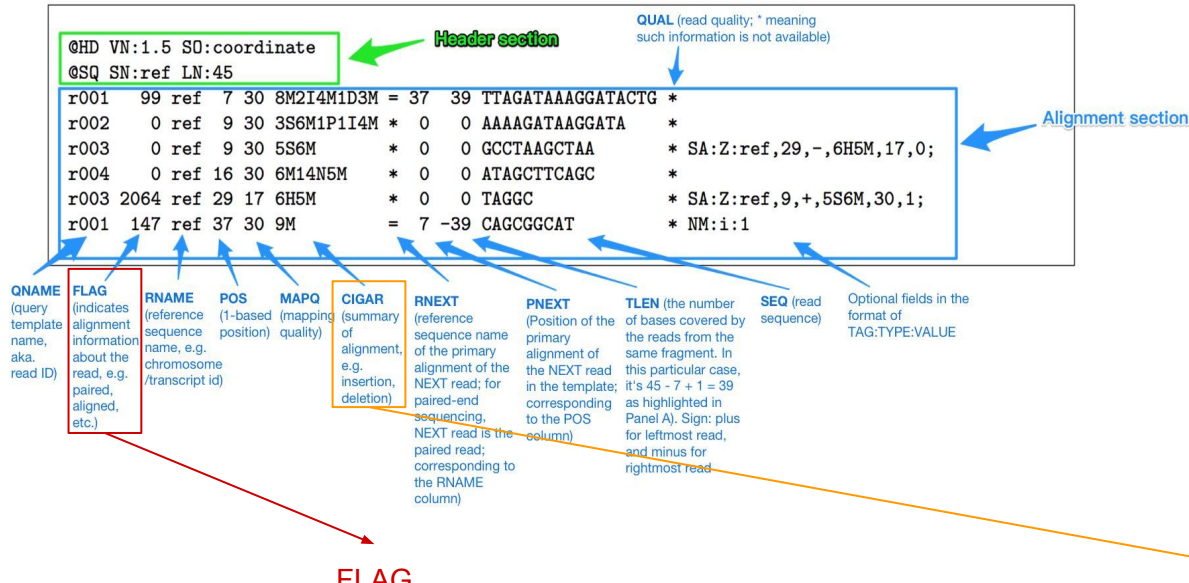

SAM Format

<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



SAM Format



FLAG

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

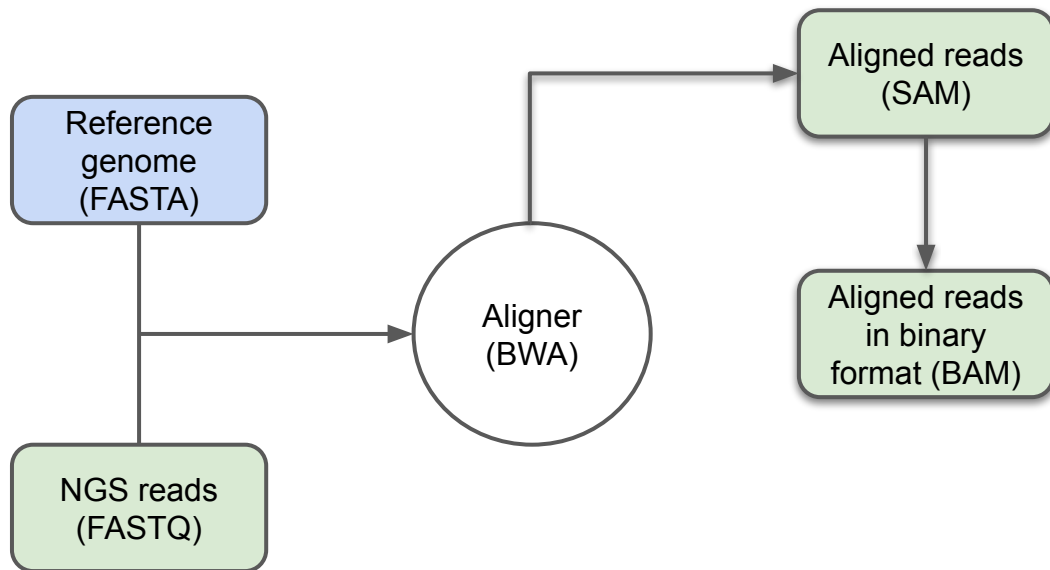
| Bit | Description |
|------|-------------------------------------------------------------------------|
| 1 | 0x1 template having multiple segments in sequencing |
| 2 | 0x2 each segment properly aligned according to the aligner |
| 4 | 0x4 segment unmapped |
| 8 | 0x8 next segment in the template unmapped |
| 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 basi MATCH
 14 basi INTERVAL
 5 basi MATCH

A typical workflow for variant calling



Short read alignment

Allineamento delle reads al genoma di riferimento

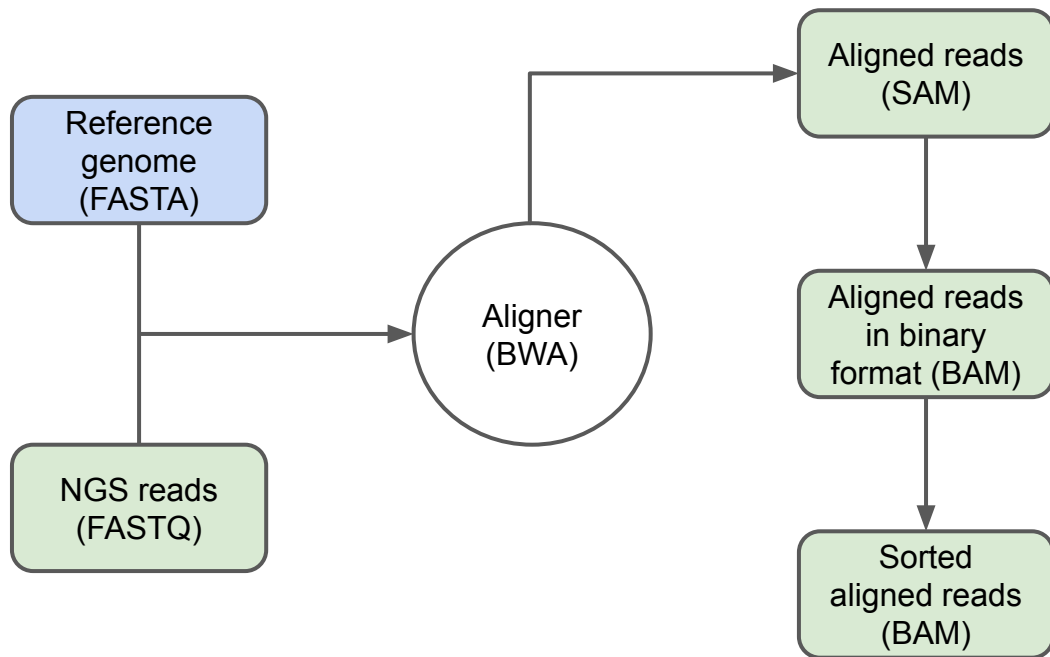
I file **SAM** di solito vengono compressi in un formato binario (non di testo e quindi comprensibile solo al computer) che si chiama **BAM** e che sta per **binarySAM**. Per portare a termine questa conversione si usa il programma **samtools** che è un pacchetto software per la manipolazione e l'estrazione di informazione dai file sam/bam.

```
/opt/samtools/bin/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 read alignment

Allineamento delle reads al genoma di riferimento

Per ordinare le reads del file BAM in base alla loro posizione nel genoma usiamo il comando sort di samtools:

```
/opt/samtools/bin/samtools sort mapping.bam > sorted.bam
```

Ora creiamo l'indice per il nostro file BAM ordinato:

```
/opt/samtools/bin/samtools index sorted.bam
```

Visualizzare reads allineate al genoma di referimento

Per visualizzare le reads allineate al genoma usiamo il comando `tvview` di `samtools`:

```
/opt/samtools/bin/samtools
```

```
tvview -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  
    tvview        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

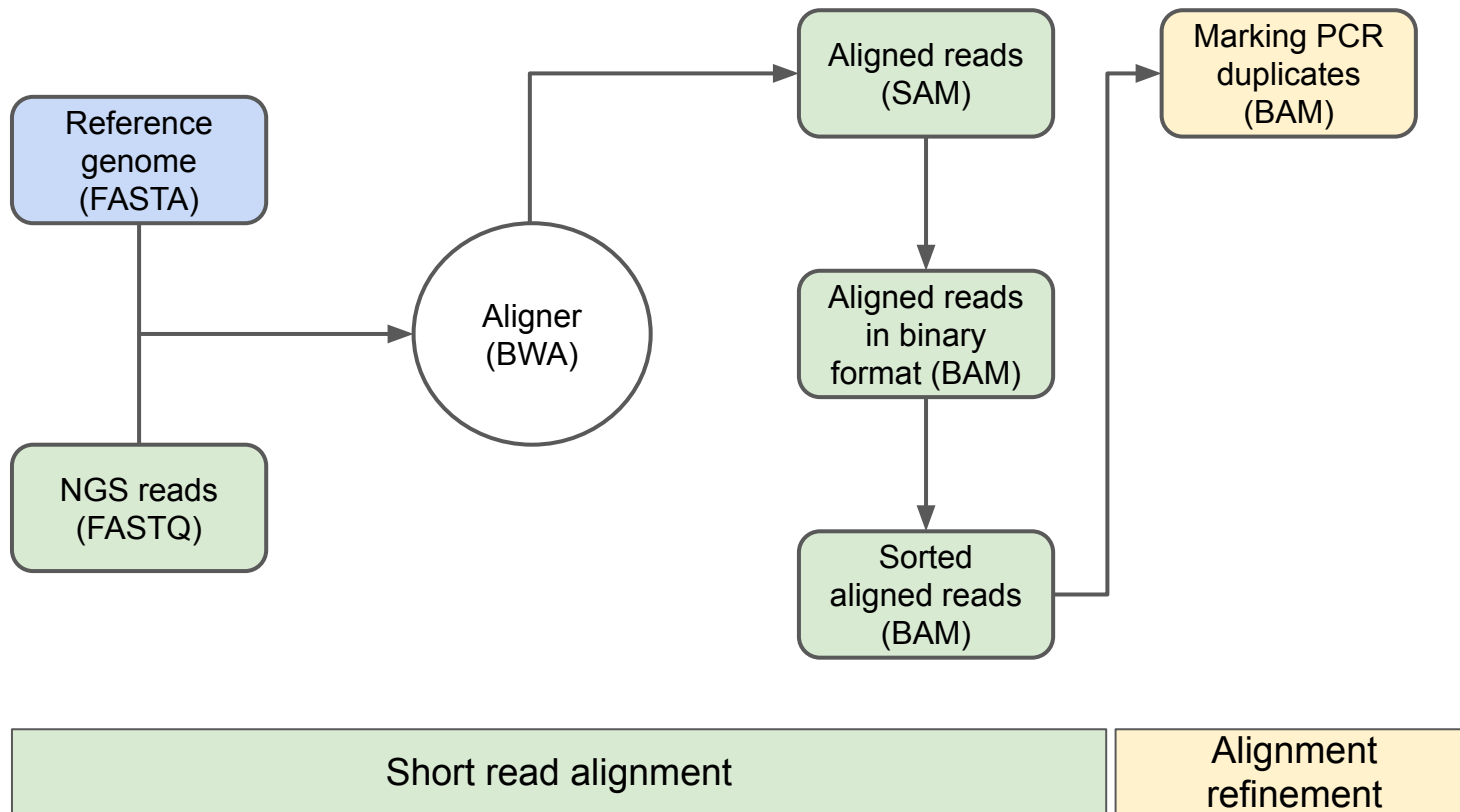
Visualizzare reads allineate al genoma di riferimento

Un programma molto utilizzato per visualizzare gli allineamenti in vari formati, tra cui i sam/bam è IGV (Integrative genomics viewer) che può essere scaricato dal seguente indirizzo:

<http://www.broadinstitute.org/igv/> (/opt/IGV/igv.sh) (caricare sorted.bam)



A typical workflow for variant calling

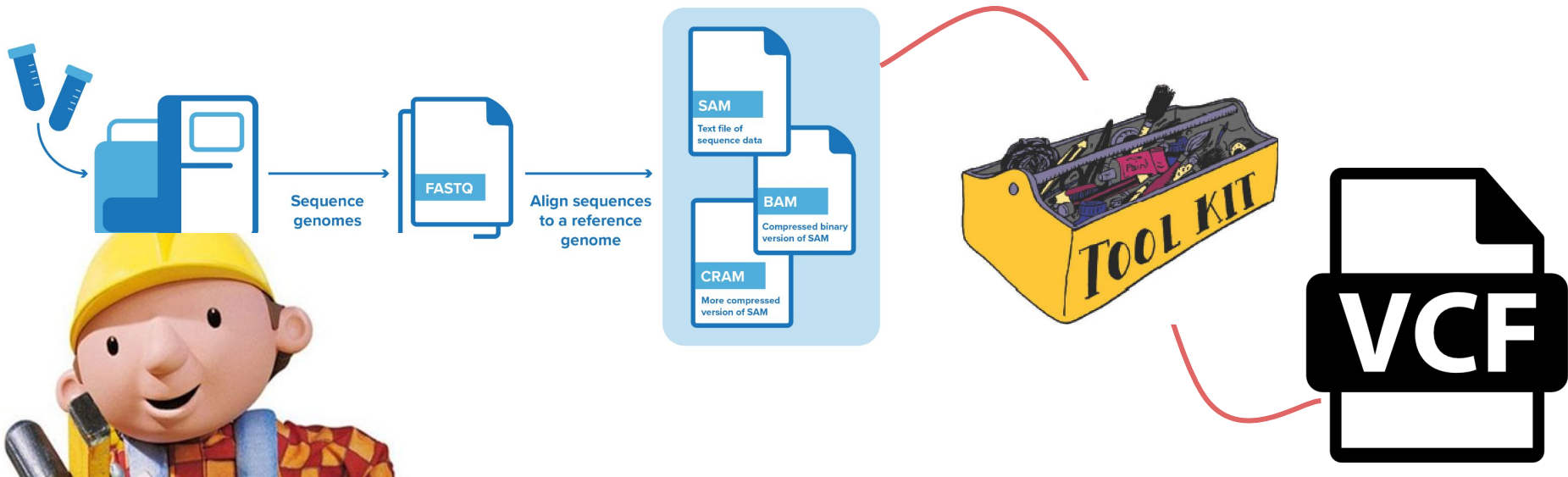


Genome Analysis ToolKit

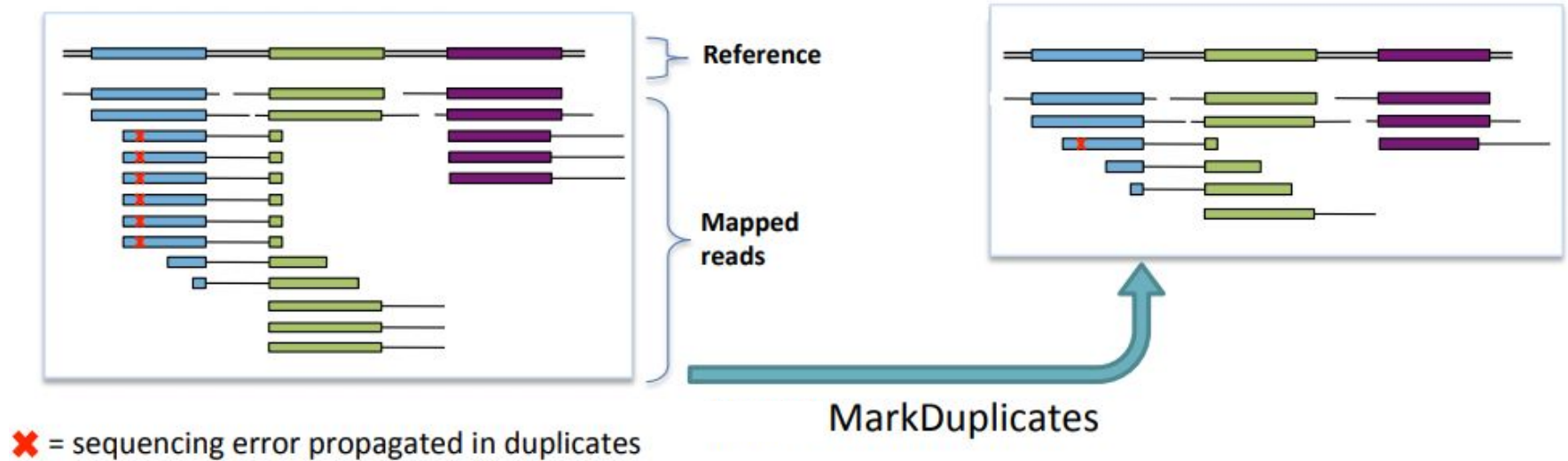


"Gee-ay-tee-kay" (/dʒi•eɪ•ti•keɪ/) and not "Gat-kay" (/gæt•keɪ/)

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.



Rimozione dei duplicati di PCR



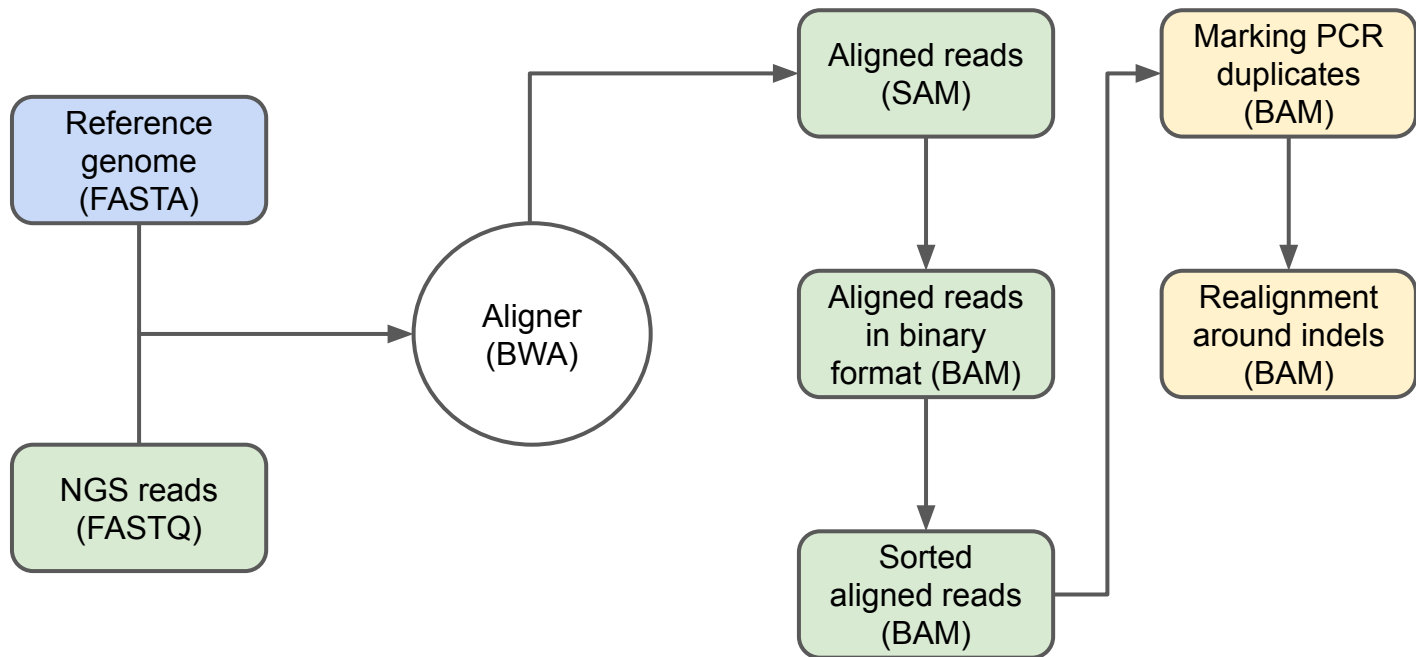
I duplicati possono falsificare un alto coverage portando a false “chiamate”.

Rimozione dei duplicati di PCR

Per rimuovere i duplicati di PCR utilizziamo una funzione del programma GATK4:

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

A typical workflow for variant calling

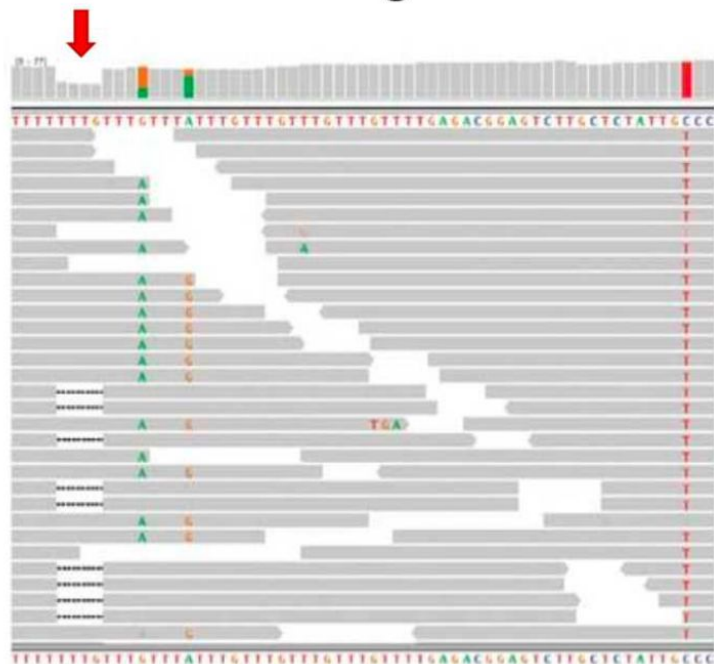


Short read alignment

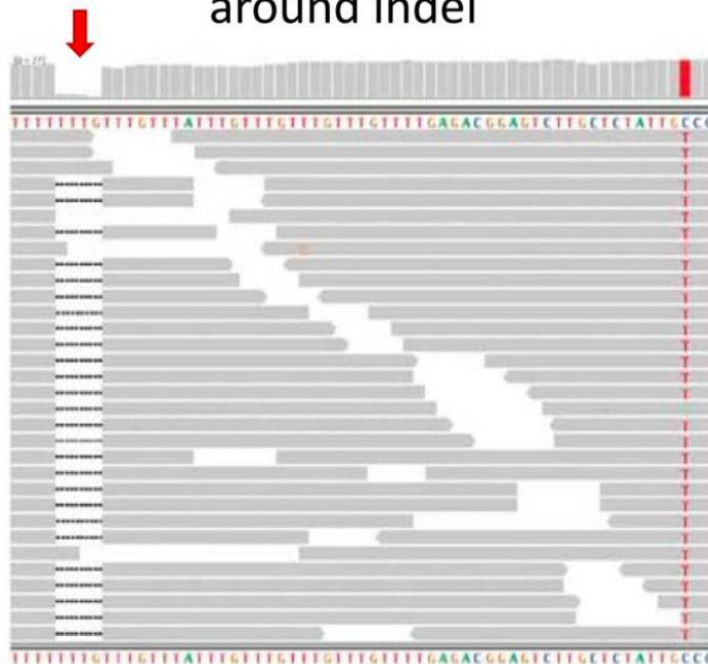
Alignment
refinement

Reallineamento locale attorno alle indels

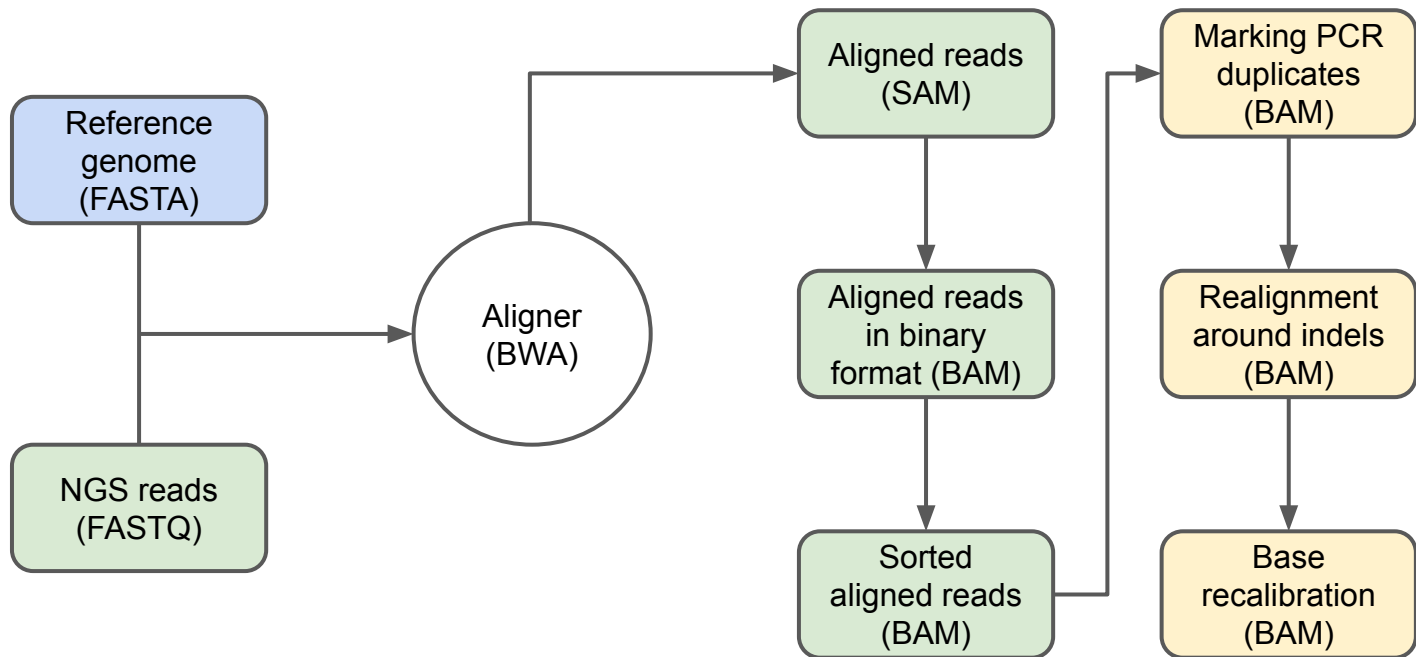
Raw BWA alignment



After local realignment around indel



A typical workflow for variant calling



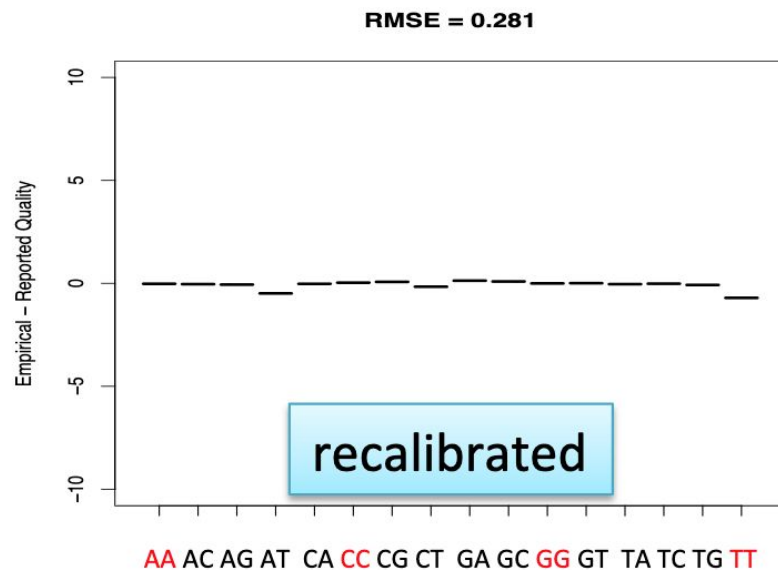
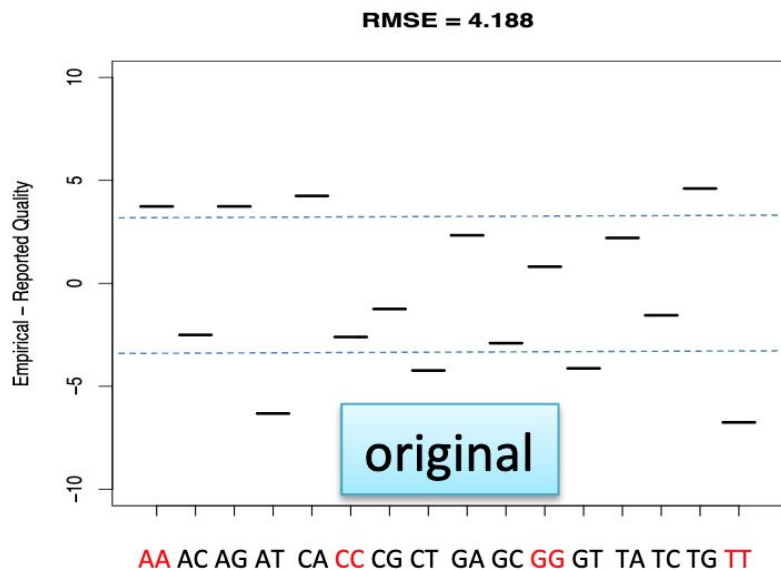
Short read alignment

Alignment
refinement

Ricalibrazione della qualità delle basi

I punteggi della qualità delle basi è fondamentale per la chiamata delle varianti ma ci sono bias sistematici che la influenzano

Example of bias: qualities reported depending on nucleotide context



Ricalibrazione della qualità delle basi

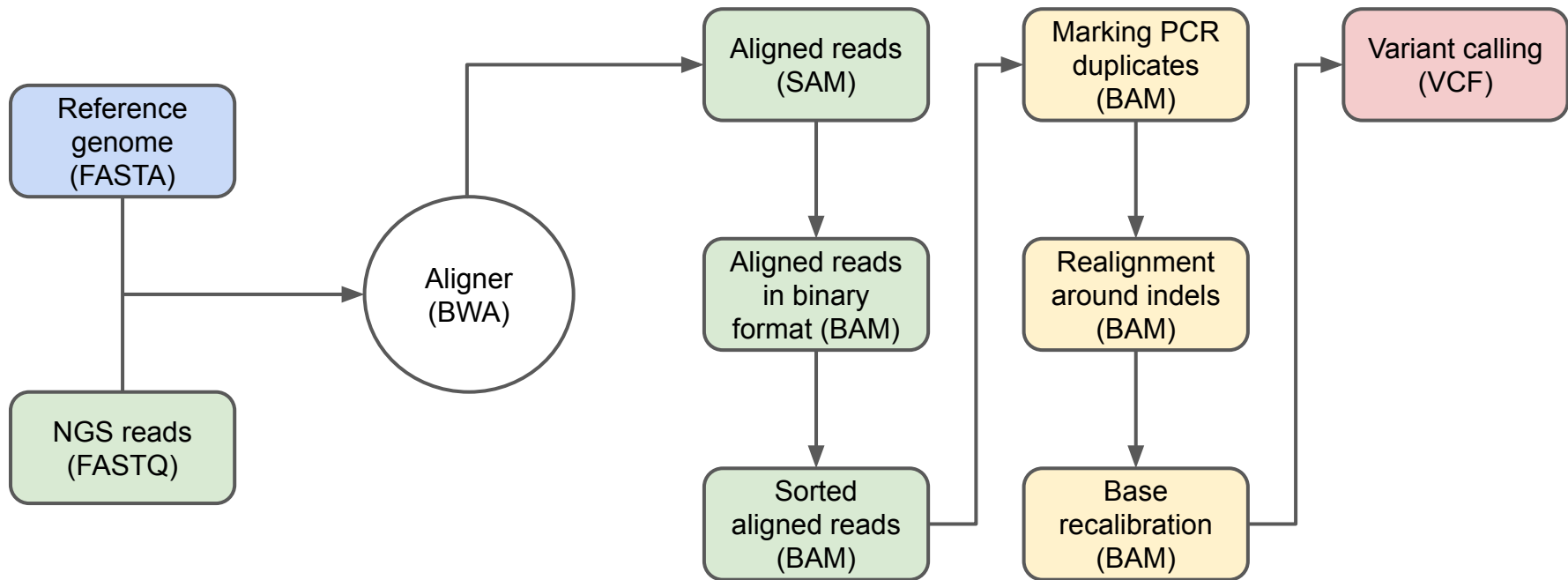
Per costruire il modello su cui ricalibrare la qualità usiamo la funzione BaseRecalibrator di GATK4:

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

Per ricalibrare la qualità usiamo la funzione ApplyBQSR di GATK4:

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

A typical workflow for variant calling



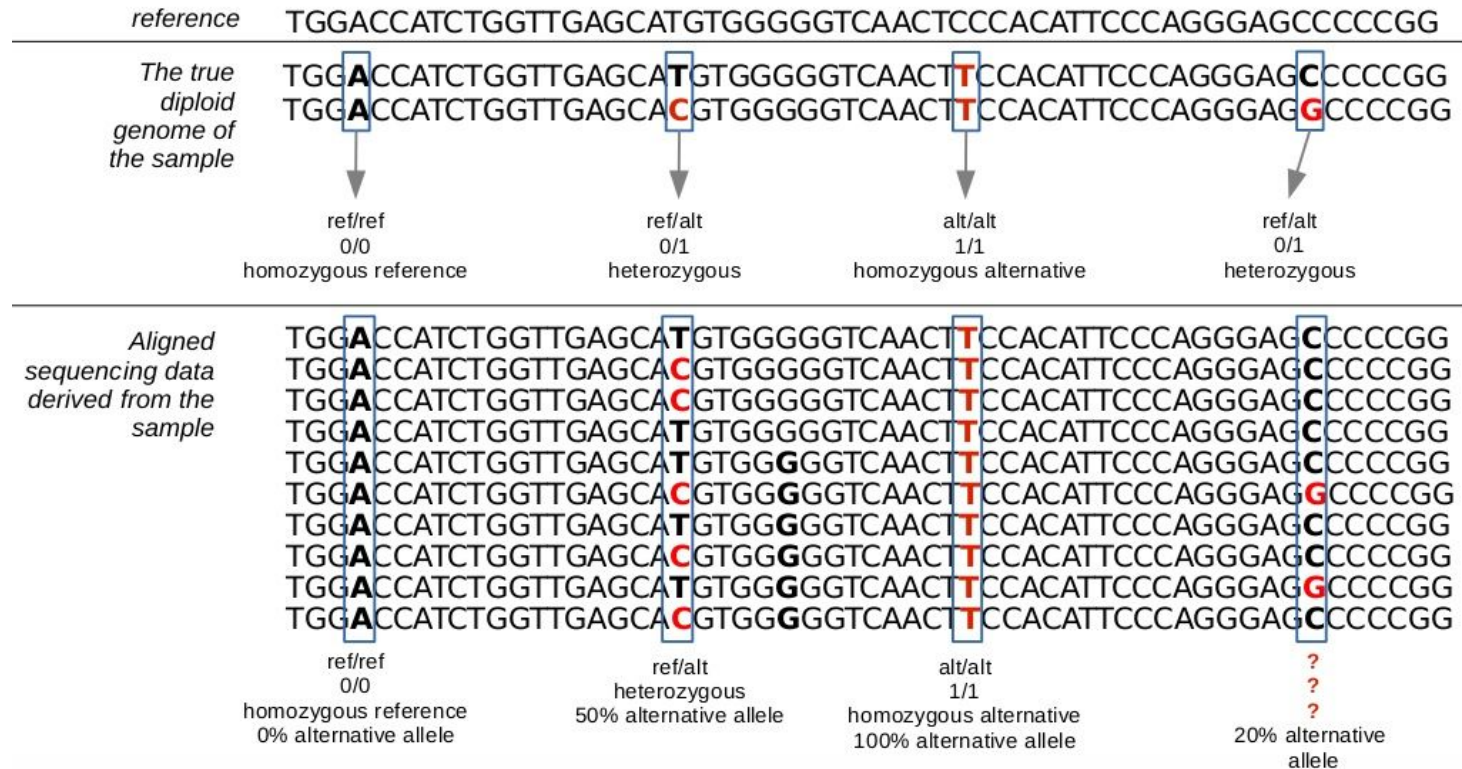
Short read alignment

Alignment
refinement

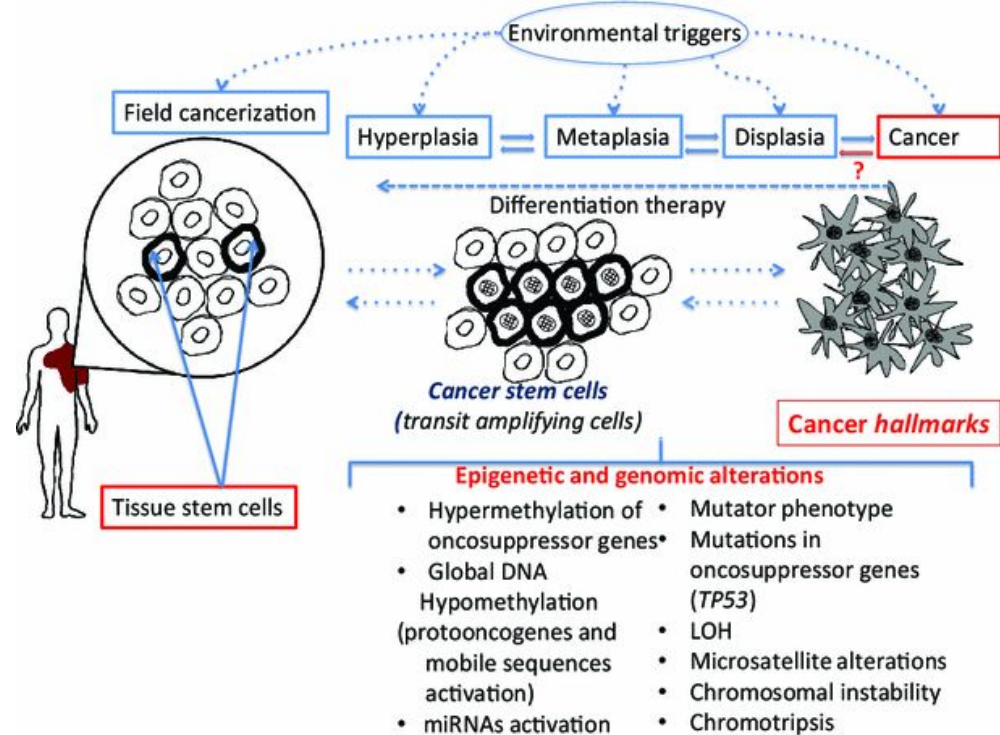
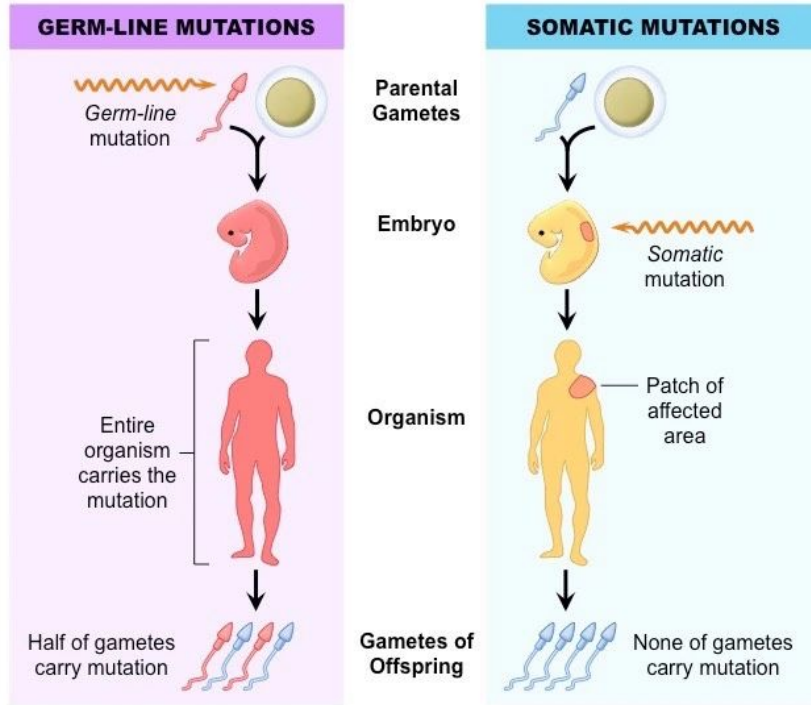
Variant detection

Cosa significa “Variant calling”?

Identificare le differenze genetiche paragonando le reads sequenziate ad un genoma di riferimento



SOMATIC VS GERMLINE MUTATIONS



Variant calling

La chiamata delle varianti può essere fatta utilizzando il programma HaplotypeCaller di GATK4:

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

Per la chiamata di varianti somatiche può essere utilizzato il programma Mutect2 di GATK4:

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

VCF format file

VCF header

```
##format=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">
```

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 |

Annotations:

- Mandatory header lines:** Lines starting with ## that define the VCF format and file information.
- Optional header lines (meta-data about the annotations in the VCF body):** Lines starting with ## that provide additional information about the annotations.
- Reference alleles (GT=0):** The first allele in the ALT column (e.g., A, T, G,).
- Alternate alleles (GT>0 is an index to the ALT column):** The other alleles in the ALT column (e.g., C, CT, G,).
- Deletion:** Indicated by the ALT value .
- SNP:** Indicated by the REF and ALT values (e.g., C to T).
- Large SV:** Indicated by the SVTYPE=DEL and END=300.
- Insertion:** Indicated by the ALT value G.
- Other event:** Indicated by the ALT value CT.
- Phased data (G and C above are on the same chromosome):** Indicated by the vertical bar (|) in the GQ field.

VCF format file

- 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 (e.g. 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 (e.g. snpEff, Annovar)
 - • FORMAT = an (optional) extensible list of fields for describing the SAMPLE column
 - • SAMPLE COLUMN = free

GERMLINE variant technical filtering

Filtrare le varianti germline con basso coverage, bassa qualità di mappaggio, bassa qualità della chiamata della variante:

```
/opt/gatk/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
```

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

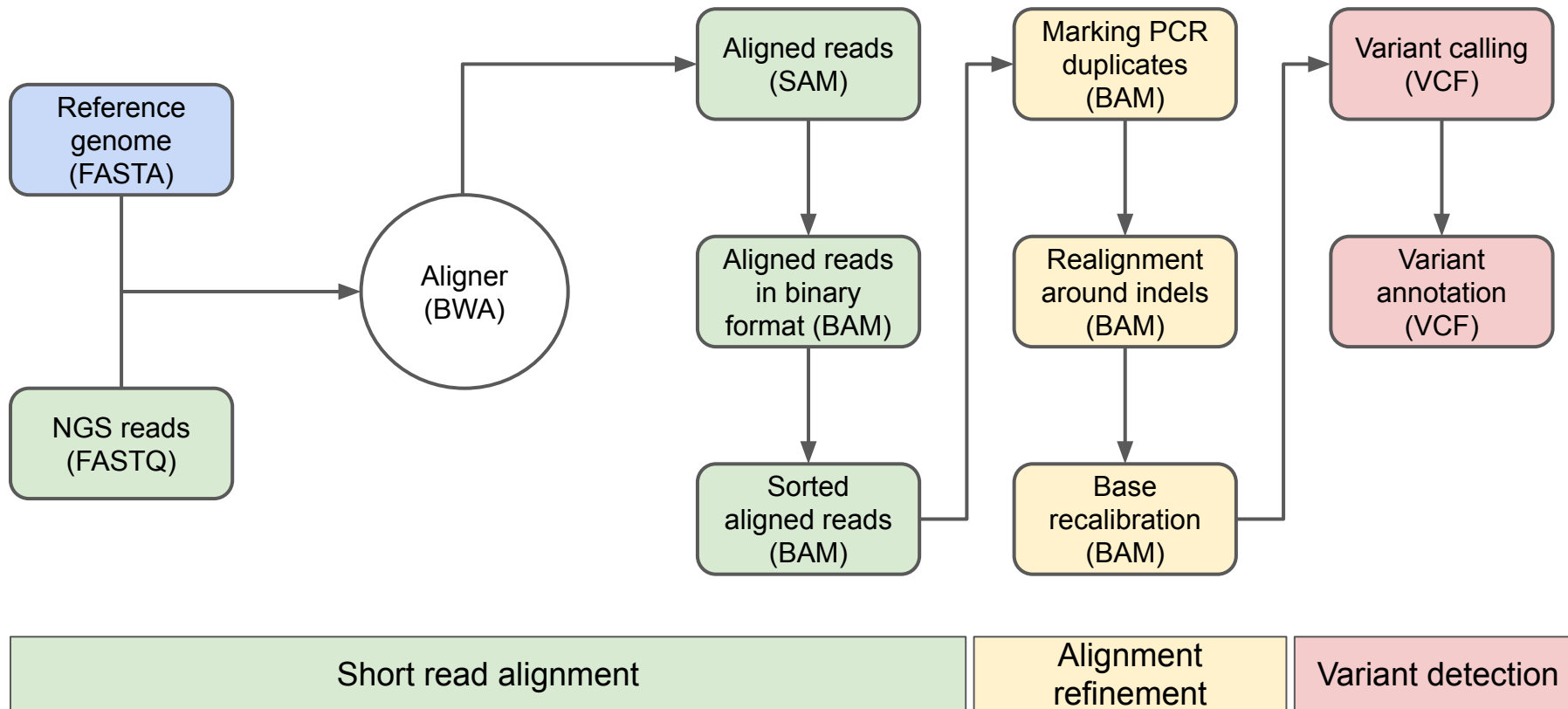

SOMATIC variant technical filtering

C'è una funzione (FilterMutectCalls) in GATK4 che serve per eliminare i falsi positivi considerando vari parametri tecnici:

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

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

A typical workflow for variant calling



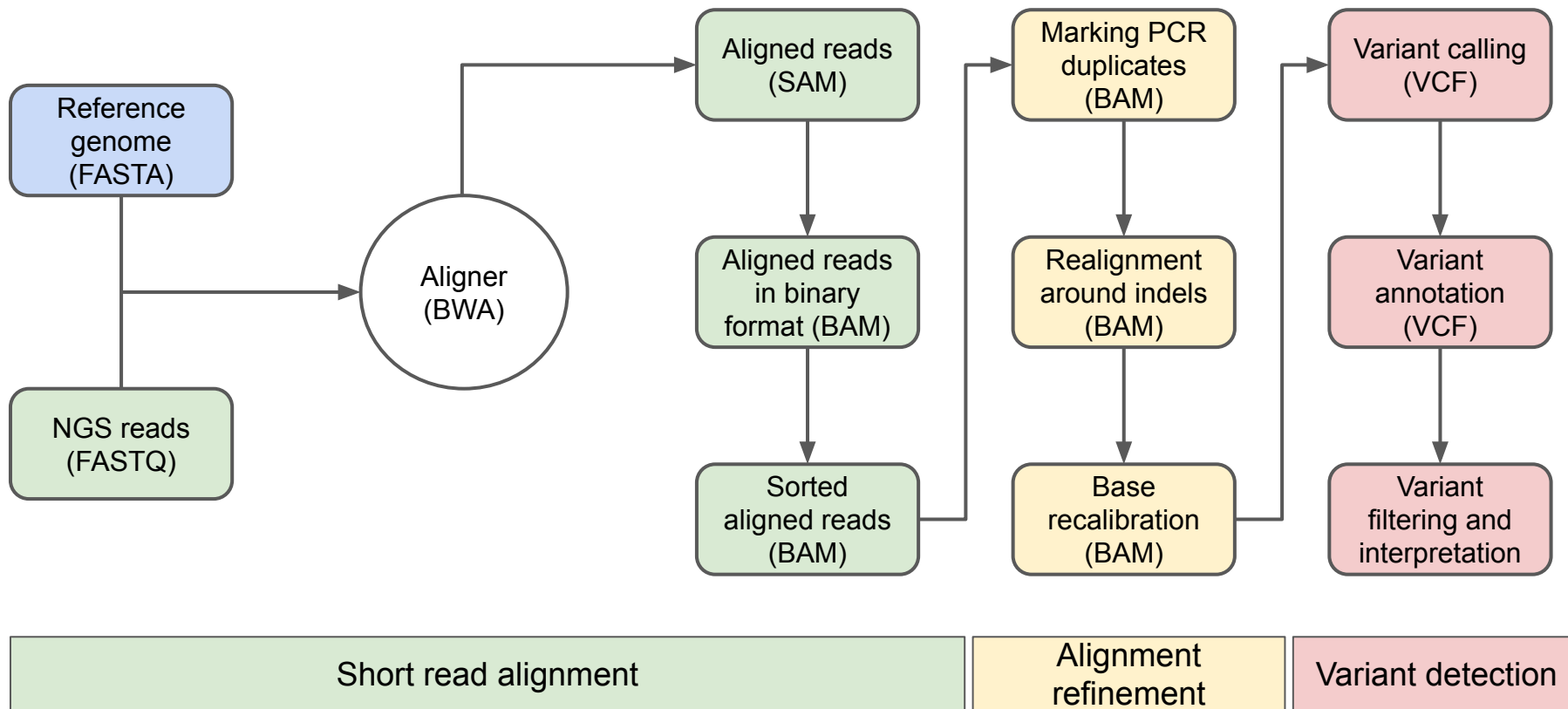
VCF annotation

L'annotazione delle varianti per dbSNP può essere fatta con la funzione VariantAnnotator di GATK4:

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

```
/opt/gatk/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

Guardare allineamento di queste due varianti:

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

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

Nel VCF file contenente le varianti somatiche cercare queste due varianti tramite la loro posizione genomica e recuperare il dbSNP ID nella colonna “ID”:

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

VCF annotation

Nel VCF file contenente le varianti somatiche cercare queste due varianti tramite la loro posizione genomica e recuperare il dbSNP ID nella colonna “ID”:

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

In quale gene si trovano? Hanno impatto sulla funzione della proteina? Qual è la loro frequenza allelica nella popolazione?.....

Vedere annotazioni in dbSNP

dbSNP è un database di varianti e cerchiamo le due varianti precedenti tramite il loro ID:

- Andare al sito di dbSNP: <https://www.ncbi.nlm.nih.gov/snp/>
- Cercare l'ID "rs768520625" e successivamente "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

Current Build 154
 Released April 21, 2020

| | |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Organism <i>Homo sapiens</i> 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) | Clinical Significance Not Reported in ClinVar Gene : Consequence TNFRSF14 : Stop Gained TNFRSF14-AS1 : Intron Variant Publications 0 citations Genomic View See rs on genome |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

Variant Details

Clinical Significance
 Frequency
 HGVS
 Submissions
 History
 Publications
 Flanks

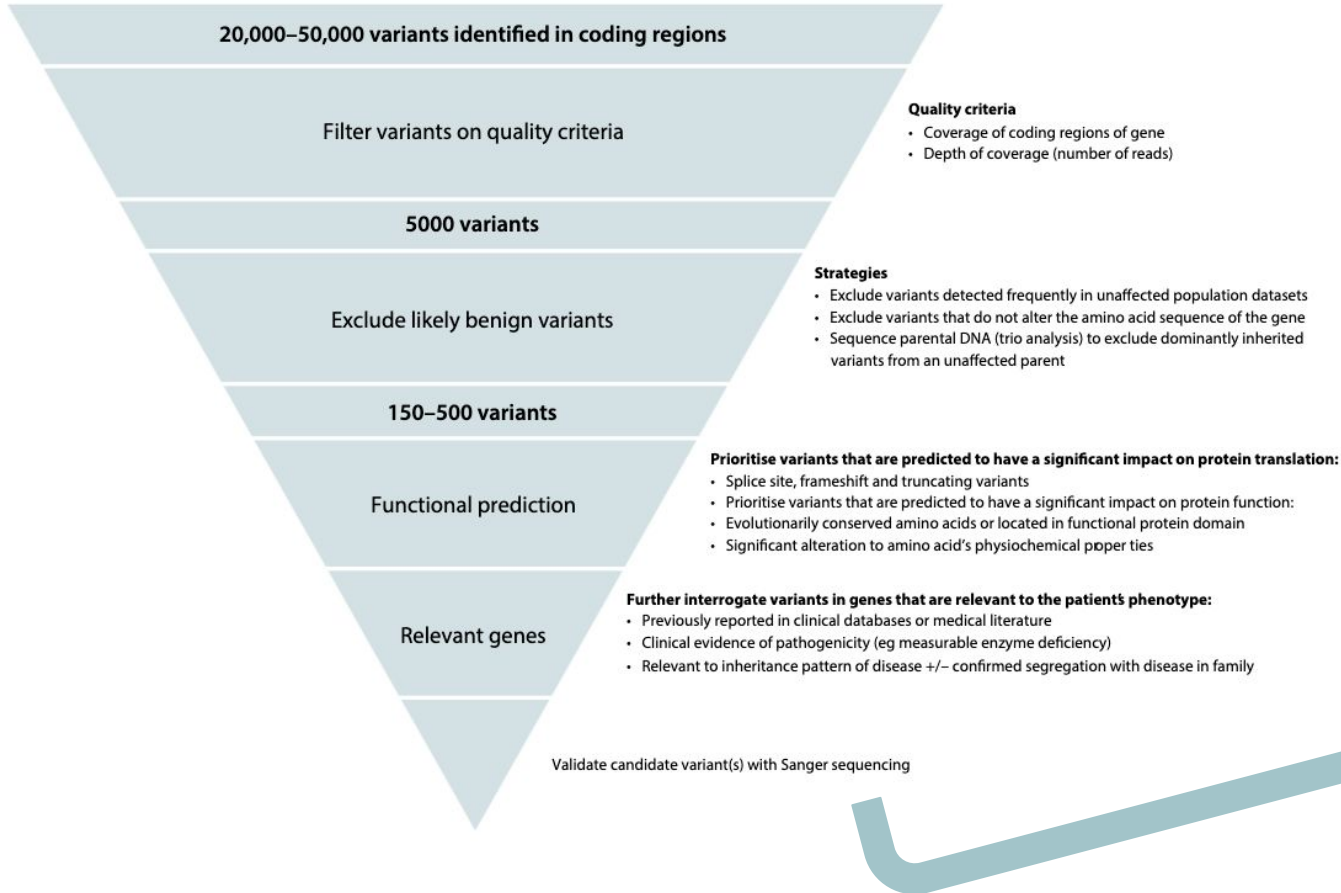
Genomic Placements

| 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 |

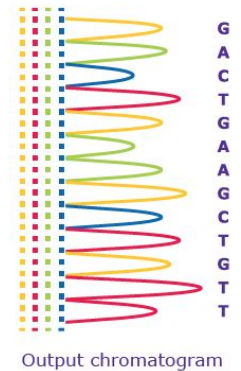
Gene: [TNFRSF14](#), TNF receptor superfamily member 14 (plus strand)

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

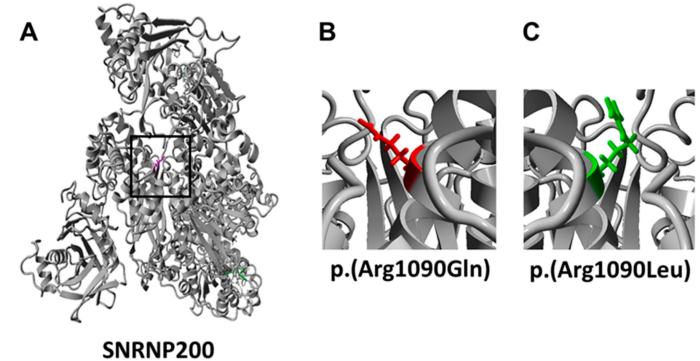
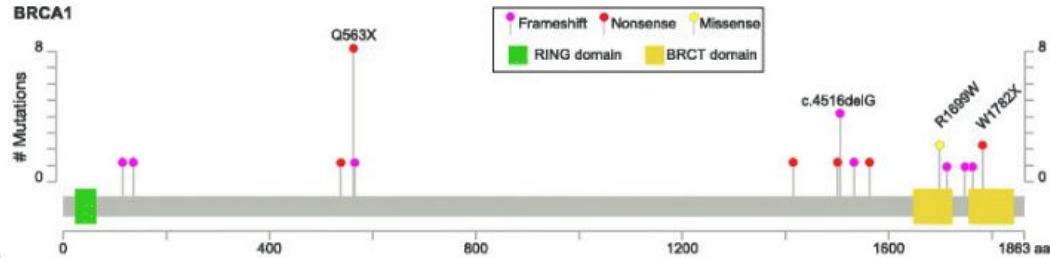
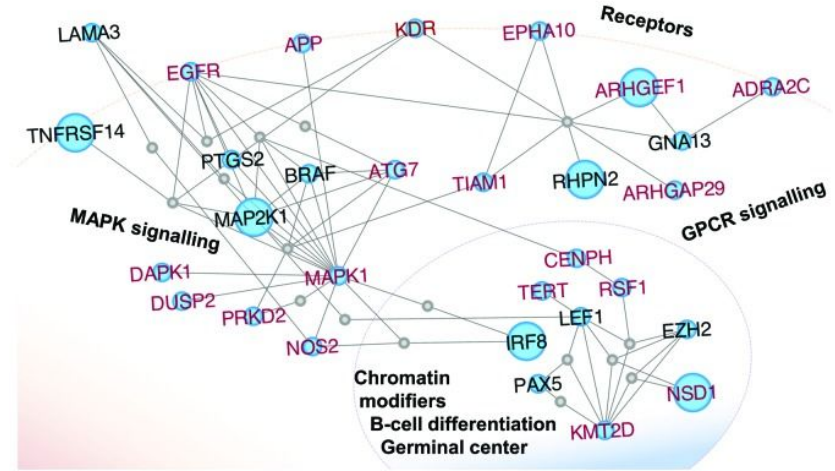
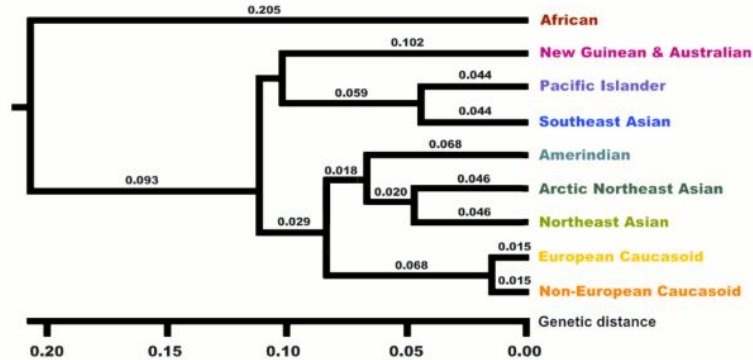
Example of variant filtering



Sanger sequencing



Esempi di analisi successive



Obiettivo dell'esercitazione

- Analisi di un esoma umano affetto da linfoma follicolare pediatrico per individuare varianti germline o somatiche che potrebbero essere causative della malattia
- Analisi dell'esoma ottenuto da DNA antico di un nobile veneto del 1300 morto in circostanze misteriose.

Descrizione del caso

- Nobile condottiero morto intorno al 1300 e sepolto in una tomba di marmo che ha favorito la mummificazione e la conservazione del corpo.



- Nel 2004 le spoglie sono state studiate e i documenti storici riconsiderati scoprendo diverse cose

Dati storici

- All'età di 23 anni aveva iniziato ad accusare stanchezza, febbre, crampi muscolari, difficoltà respiratorie e cardiache che, dopo sforzi intensi, l'hanno portato ad abbandonare più volte il campo di battaglia nonostante fosse un guerriero esperto
- All'età di 34 anni si è ammalato per lungo tempo, tanto che l'avevano considerato morto
- Morto nel 1300 all'età di 38 anni dopo 3 giorni di febbre e generico “flusso” (emorragia? nausea? malattia intestinale?)
- Alcuni documenti contemporanei attribuiscono la morte ad avvelenamento, altri alla “congestione” dopo una battuta di caccia

Dati sulle spoglie (2004)

- Non presentava ferite
- Esami tossicologici hanno trovato tracce di *Digitalis purpurea*.
- La *digossina* è un potente veleno, un glicoside cardiaco che aumenta la forza di contrazione del cuore
- Usata per fini terapeutici per le aritmie e l'insufficienza cardiaca



Obiettivi dello studio

- Analisi dei dati di sequenziamento dell'esoma a partire da DNA antico
- **Scoprire la più probabile causa della morte**

Obiettivi dello studio

- Verificare nell'esoma se ci sono varianti genetiche ricollegabili ad una malattia di cui poteva essere affetto il nobile seguendo la guida **Esercitazione5_guida.pdf**
 - Quante varianti germline si trovano?
 - Troviamo nel VCF le varianti in posizione 17:78078656 e 17:78084553?
 - Se sì, qual è l'allele alternativo? La loro frequenza allelica? Sono eterozigoti o omozigoti? Sono in cis o in trans?
 - Hanno un dbSNP ID? Se sì, quali sono?
 - Quale gene colpiscono? Qual è la funzione della proteina prodotta da questo gene? La conseguenza funzionale (missenso/sinonime)?
 - Qual è la loro frequenza nella popolazione? Sono rare o polimorfismi?
 - Hanno una significatività clinica annotata in ClinVar? (Benigne, Patogeniche o sconosciute) In quale malattia?
 - In ClinVar per la variante 17:78078656 c'è un link per OMIM?
 - C'è una relazione gene/fenotipo? Quale? Sindrome recessiva o dominante?
 - Considerazioni finali? Diagnosi?

Soluzione

- Verificare nell'esoma se ci sono varianti genetiche ricollegabili ad una malattia di cui poteva essere affetto il nobile seguendo la guida **Esercitazione5_guida.pdf**
 - Quante varianti germline si trovano? **35**
 - Troviamo nel VCF le varianti in posizione 17:78078656 e 17:78084553? **Sì**
 - Se sì, qual è l'allele alternativo? La loro frequenza allelica? Sono eterozigoti o omozigoti? Sono in cis o in trans? **A, 51.4% & 46.2%, trans**
 - Hanno un dbSNP ID? Se sì, quali sono? **rs1800299 ; rs398123169**
 - Quale gene colpiscono? Qual è la funzione della proteina prodotta da questo gene? La conseguenza funzionale (missenso/sinonime)? **GAA, demolizione glicogeno, missenso**
 - Qual è la loro frequenza nella popolazione? Sono rare o polimorfismi? **3% e rara**
 - Hanno una significatività clinica annotata in ClinVar? (Benigne, Patogeniche?) In quale malattia? **Patogenica e benigna, Glycogen storage disease, type II**
 - In ClinVar per la variante 17:78078656 c'è un link per OMIM? **Sì, 606800.0001**
 - C'è una relazione gene/fenotipo? Quale? Sindrome recessiva o dominante? **Glycogen storage disease II (GSD2), autosomica recessiva**
 - Considerazioni finali? Diagnosi? **Eterozigote composto per GAA, GSD2 tardivo**