

CORSO DI METODI MOLECOLARI E BIOINFORMATICA

LM Biologia Evoluzionistica, Università di Padova
Anno accademico 2021/22

***Dr. Andrea Binatti, Dr. Enrico Gaffo,
Prof.ssa Stefania Bortoluzzi***

Esercitazione 6 di Bioinformatica
Analisi dati NGS: Variant calling

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 “esercitazione6.zip” da
http://compgen.bio.unipd.it/~stefania/Didattica/AA2021-2022/MMOL_BIOINFO_BE/esercitazione6.zip
- Decomprimere il file “esercitazione6.zip”:

?

- Spostarsi nella cartella esercitazione6:

?

Preparazione dell'esercitazione

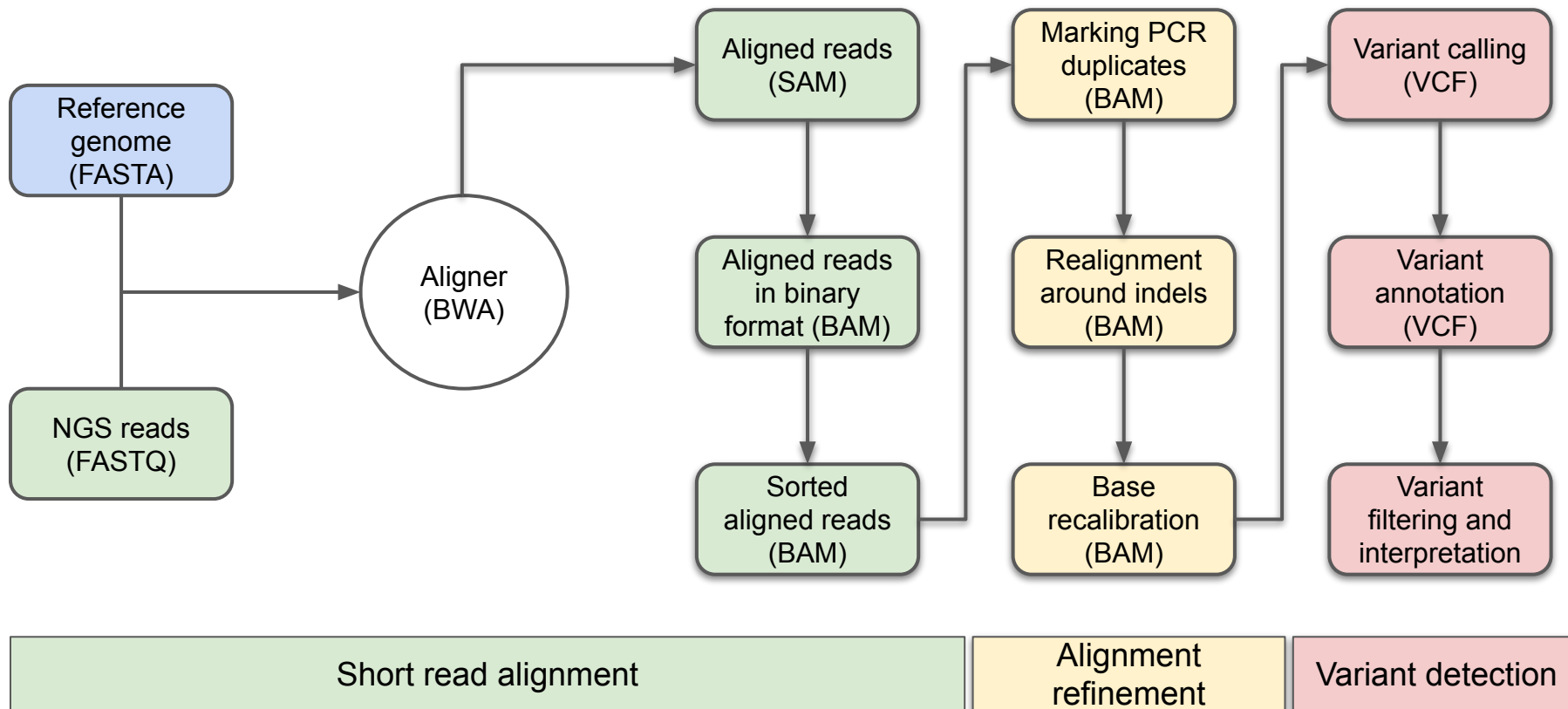
- Scaricare il file “esercitazione6.zip” da http://compgen.bio.unipd.it/~stefania/Didattica/AA2021-2022/MMOL_BIOINFO_BE/esercitazione6.zip
- Decomprimere il file “esercitazione6.zip” usando il comando **unzip**:

```
unzip esercitazione6.zip
```

- Spostarsi nella cartella esercitazione6:

```
cd esercitazione6
```

A typical workflow for variant calling

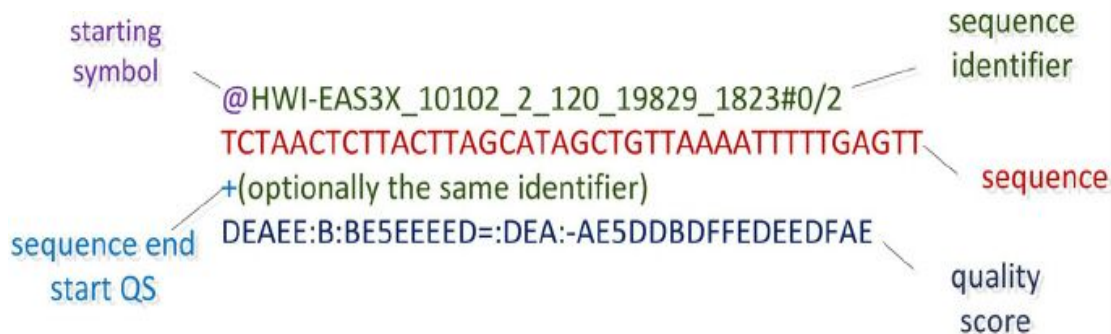


FASTQ format

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
TGCCTCCCTTTCCAATGGACTATTTTAGAAGAAATGGAGCTGTCACCCACATCAAGATTCAGAACACTG
+
?????ABA?DDDDDDDDFGGFGFFIIHHIIFHHHII@FHHIIIIIGFF>EHHHFFGHIFHFGHAFGH
@M70273:8:000000000-AJLMP:1:1101:16927:2095 1:N:0:1
CCTATCATATATGCCTTAGTTTGTGAAANATATTGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
+
??AA?BBBEDDEEEEGGGGGIHHIII#7AFHII#####5#####
@M70273:8:000000000-AJLMP:1:1101:18171:2095 1:N:0:1
TTGTGATCCACATTCTCTTCCATTGTAGNGCAAATNNNNNNNNNNNNNNNNNNNNNTNTNTNNNTN
+
?????BBBDDDDDDDDGGGGGIIHHI#7AEFHI#####7###55#55###5###
@M70273:8:000000000-AJLMP:1:1101:19337:2095 1:N:0:1
GCCCCCATGCGCGGCATGATGAACCTCCGCTGCTGNNNNNNNNNNNNNNNNNNNTNTNTNNNCAN
+
?????ABAADDDDDDDFFFFFIIHHIIIHHHHHHI#####5###55#55###44#
@M70273:8:000000000-AJLMP:1:1101:14484:2097 1:N:0:1
CTGGACTGATATGTGATTATTCTTTCAACAGCCACGCCAGATCCAGTGAAAAACAAGCTCTCATGTC
+
?????A?BB?DDDDDDBGGGGGGIIHHIIIHHHHHHHFGFHHHHHHHHHHHHHHHHHHHHHHHHHH
@M70273:8:000000000-AJLMP:1:1101:16321:2100 1:N:0:1
TAGATGCTTTTAACTAAGTTACCTGACTTNCCTTATNNNNNNNNNNNNNNNNNNNTNNNGCNGCNNCNC
+
?????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/

Unire i cromosomi in unico file FASTA:

?

Controllare che i cromosomi siano nell'ordine giusto:

?

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/

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
```

Controllare che i cromosomi siano nell'ordine giusto:

?

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/

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
```

Controllare che i cromosomi siano nell'ordine giusto:

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

Preparazione del genoma di riferimento

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
```

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

Header section

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
```

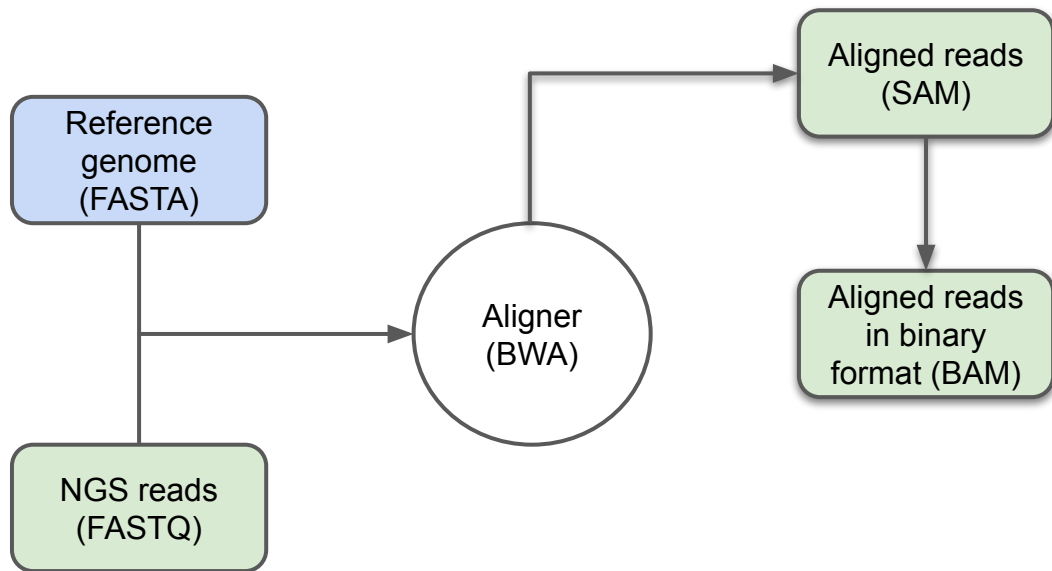
Alignment section

r001	99	ref	7	30	8M2I4M1D3M	=	37	39	TTAGATAAAGGATACTG	*	
r002	0	ref	9	30	3S6M1P1I4M	*	0	0	AAAAGATAAGGATA	*	
r003	0	ref	9	30	5S6M	*	0	0	GCCTAAGCTAA	*	SA:Z:ref,29,-,6H5M,17,0;
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

Annotations:

- QNAME** (query template name, aka. read ID)
- FLAG** (indicates alignment information about the read, e.g. paired, aligned, etc.)
- RNAME** (reference sequence name, e.g. chromosome /transcript id)
- POS** (1-based position)
- MAPQ** (mapping quality)
- CIGAR** (summary of alignment, e.g. insertion, deletion)
- RNEXT** (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)
- PNEXT** (Position of the primary alignment of the NEXT read in the template; corresponding to the POS column)
- TLEN** (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)
- SEQ** (read sequence)
- Optional fields in the format of TAG:TYPE:VALUE

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 binary sam. 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.

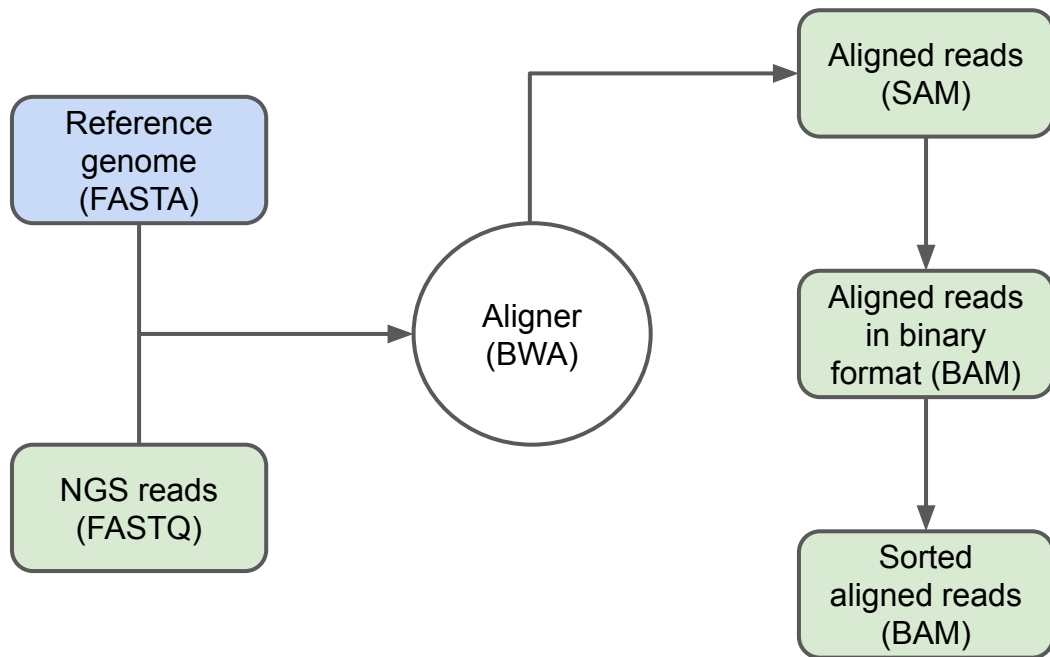
```
/opt/samtools/bin/samtools view -b mapping.sam > mapping.bam
```

...

-b output BAM

...

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 riferimento

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
```

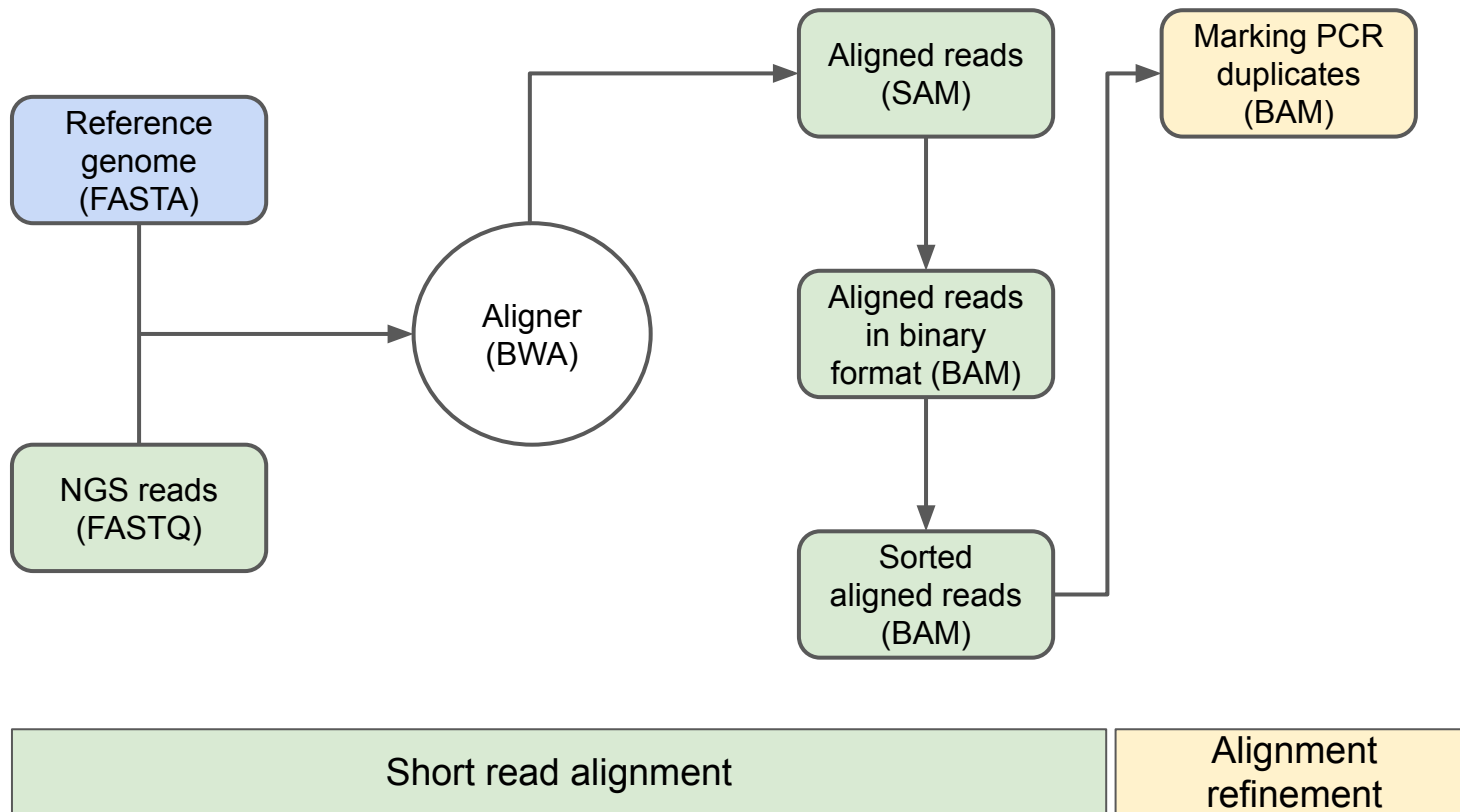
Visualizzare reads allineate al genoma di riferimento

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:

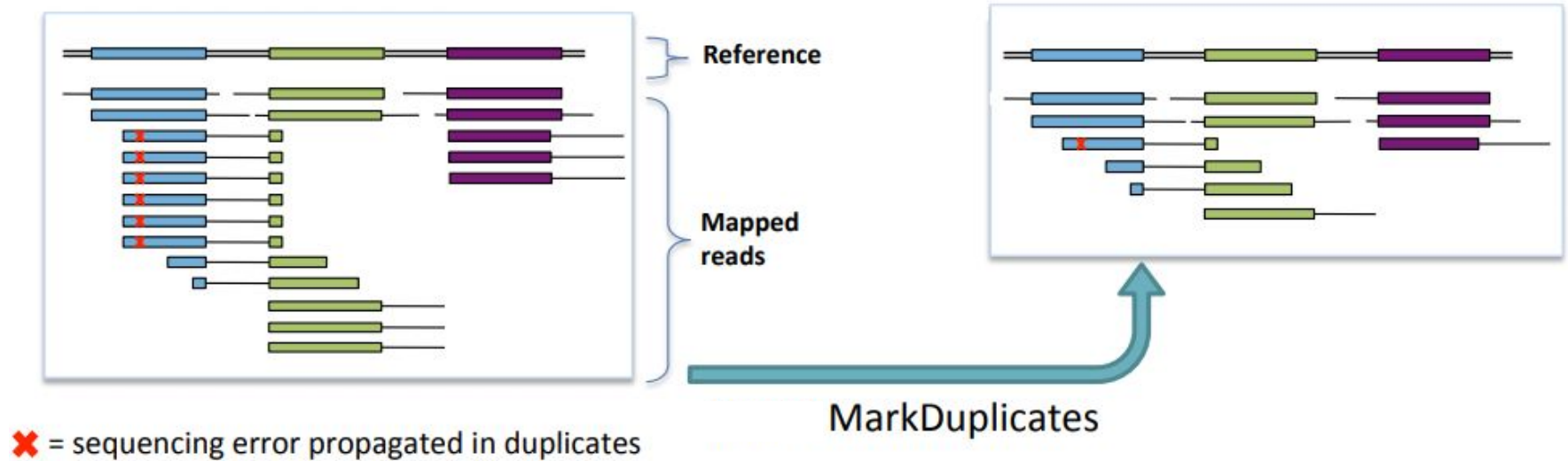
<http://www.broadinstitute.org/igv/> (/opt/IGV/igv.sh) (caricare sorted.bam)



A typical workflow for variant calling



Rimozione dei duplicati di PCR



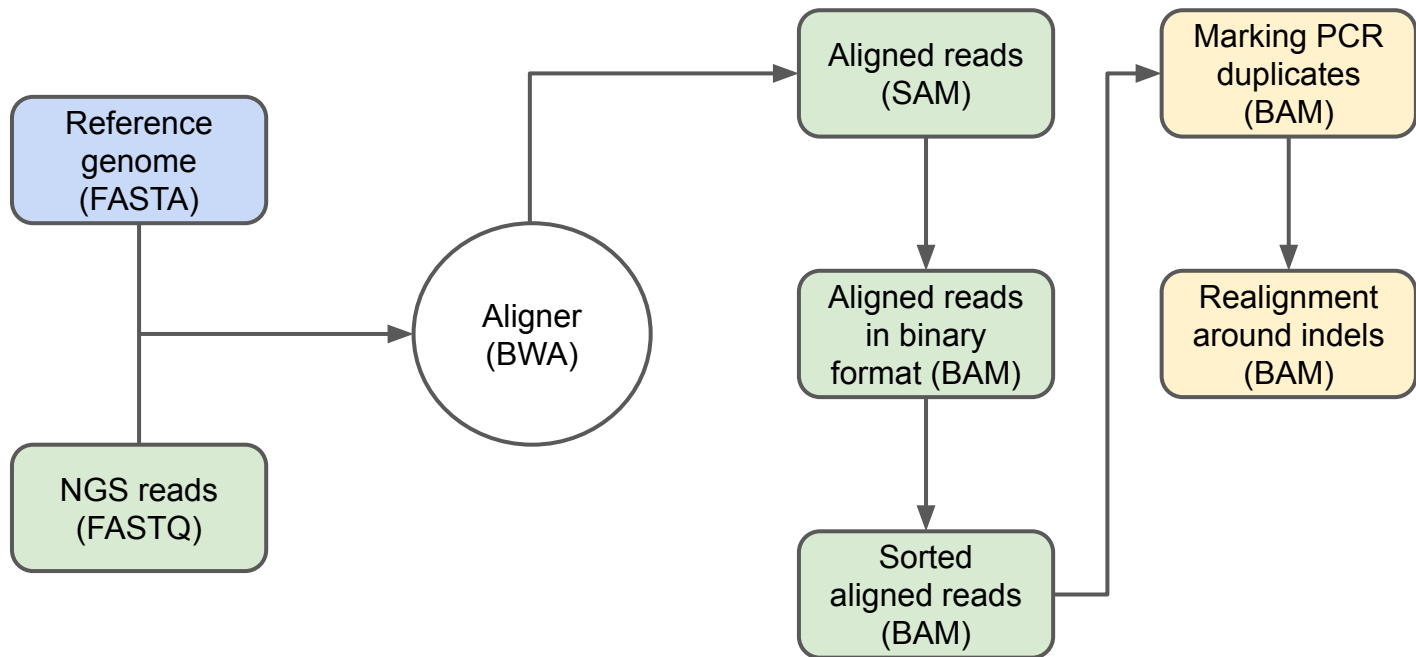
I duplicati possono falsificare un alto coverage portando a false “chiamate”.

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

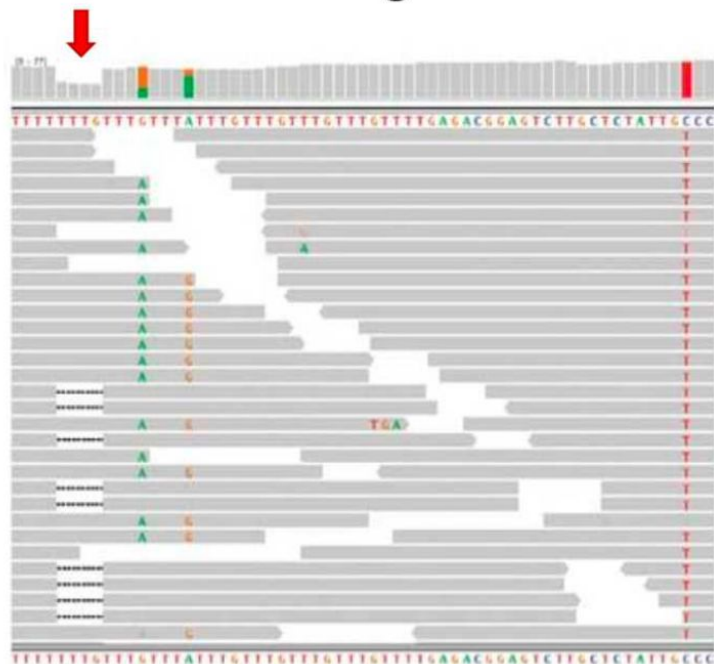


Short read alignment

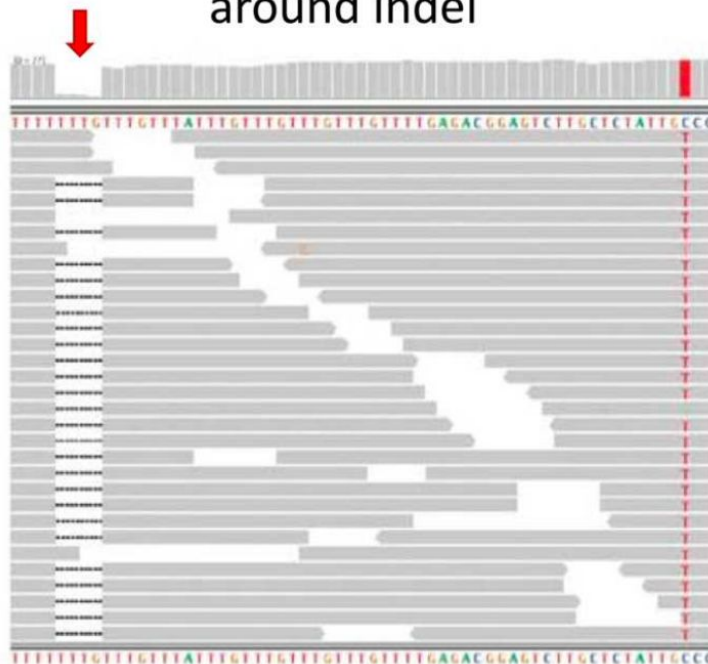
Alignment
refinement

Reallineamento locale attorno alle indels

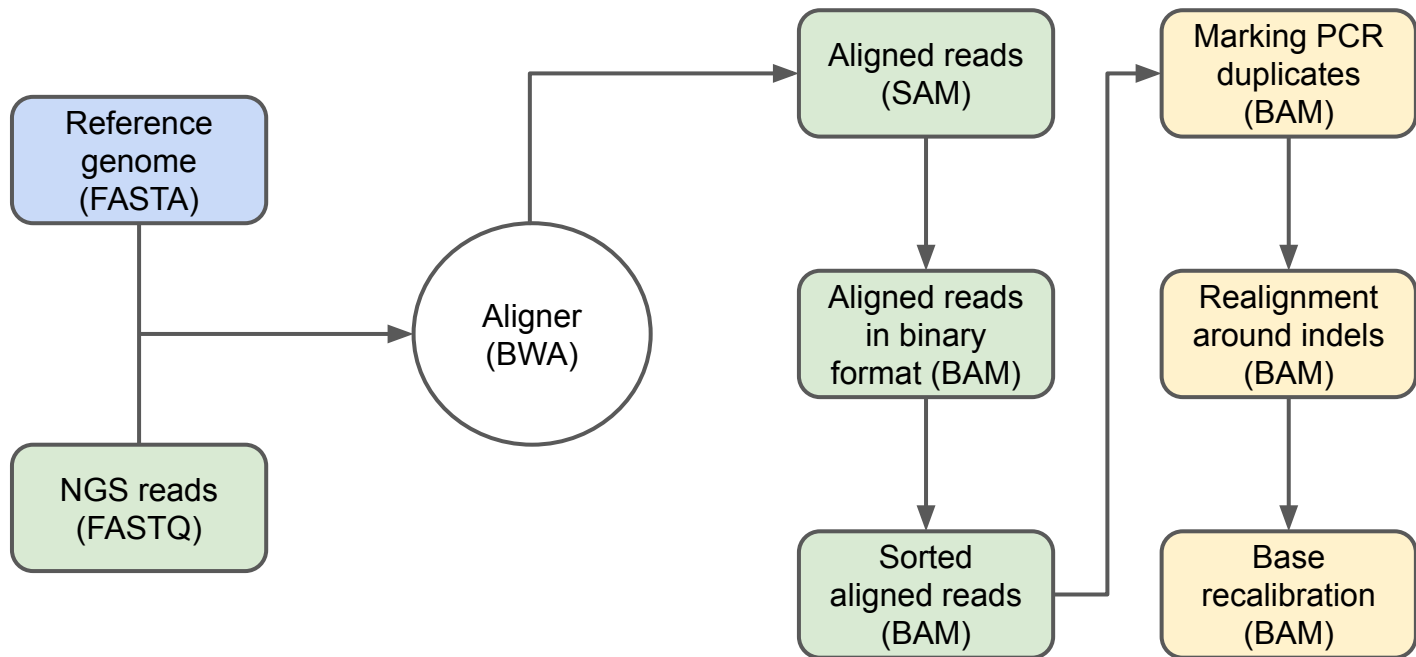
Raw BWA alignment



After local realignment around indel



A typical workflow for variant calling



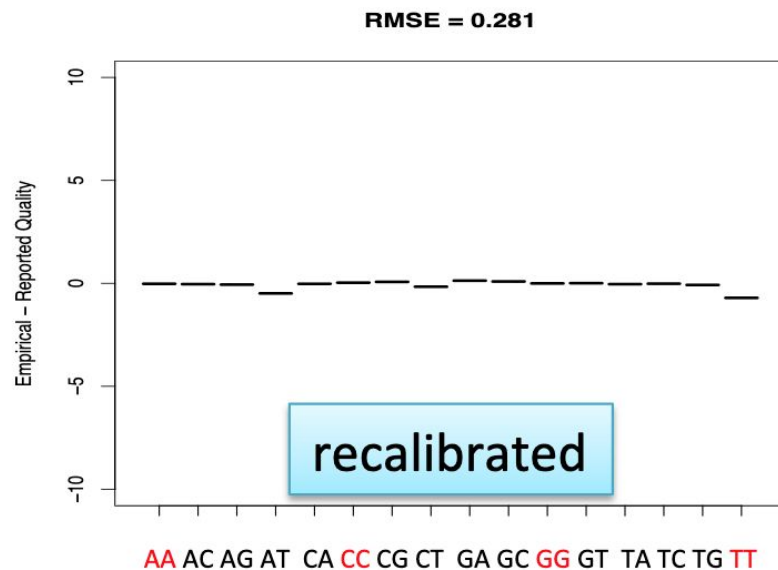
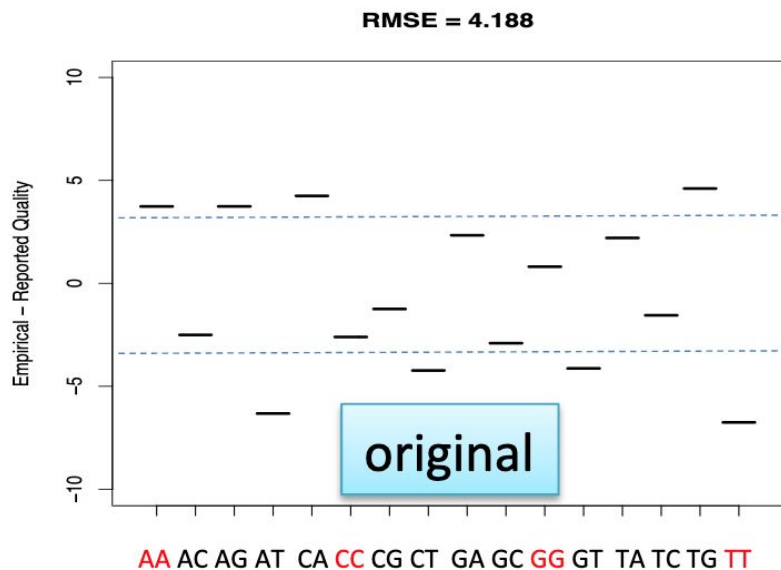
Short read alignment

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

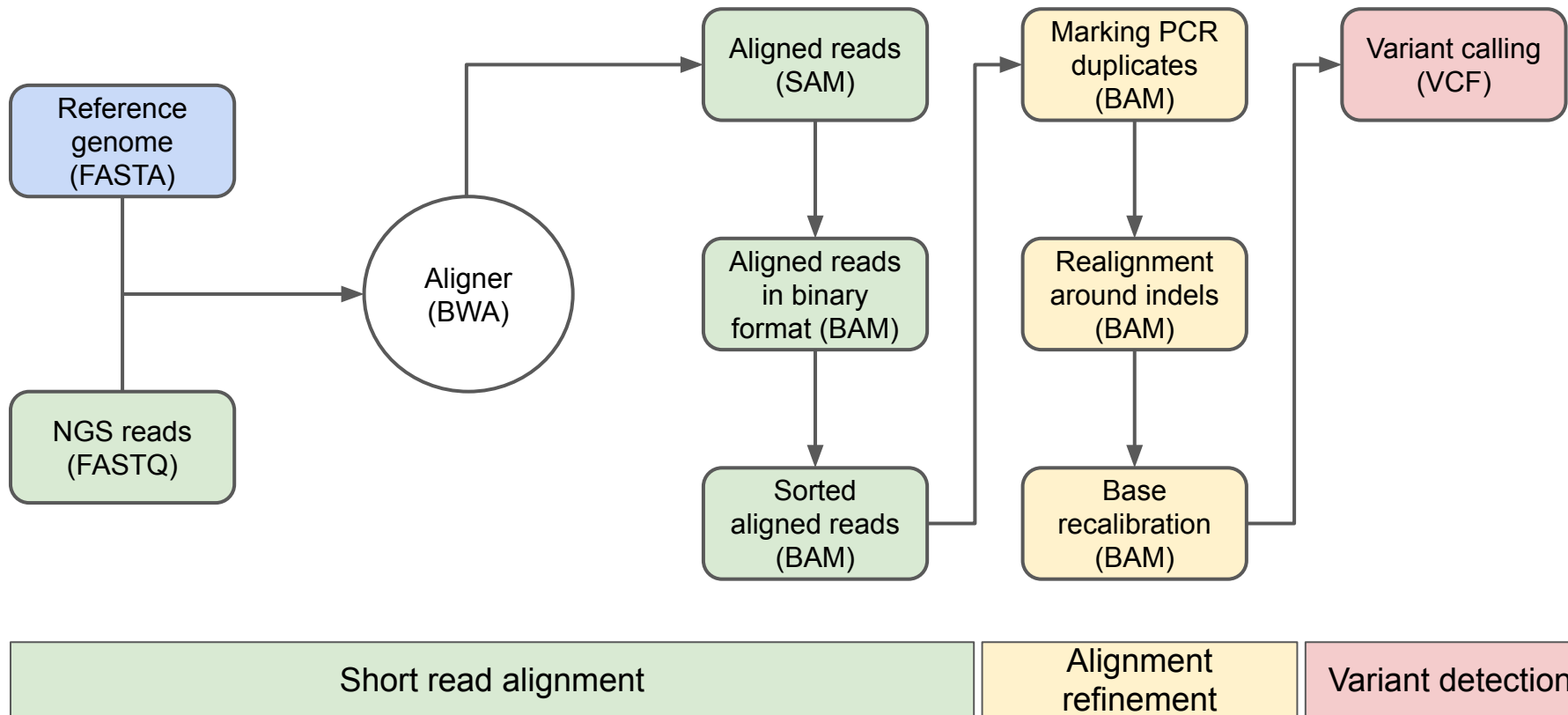
Per costruire il modello su cui ricalibrare la qualità usiamo la funzione BaseRecalibrator di GATK4:

```
/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:

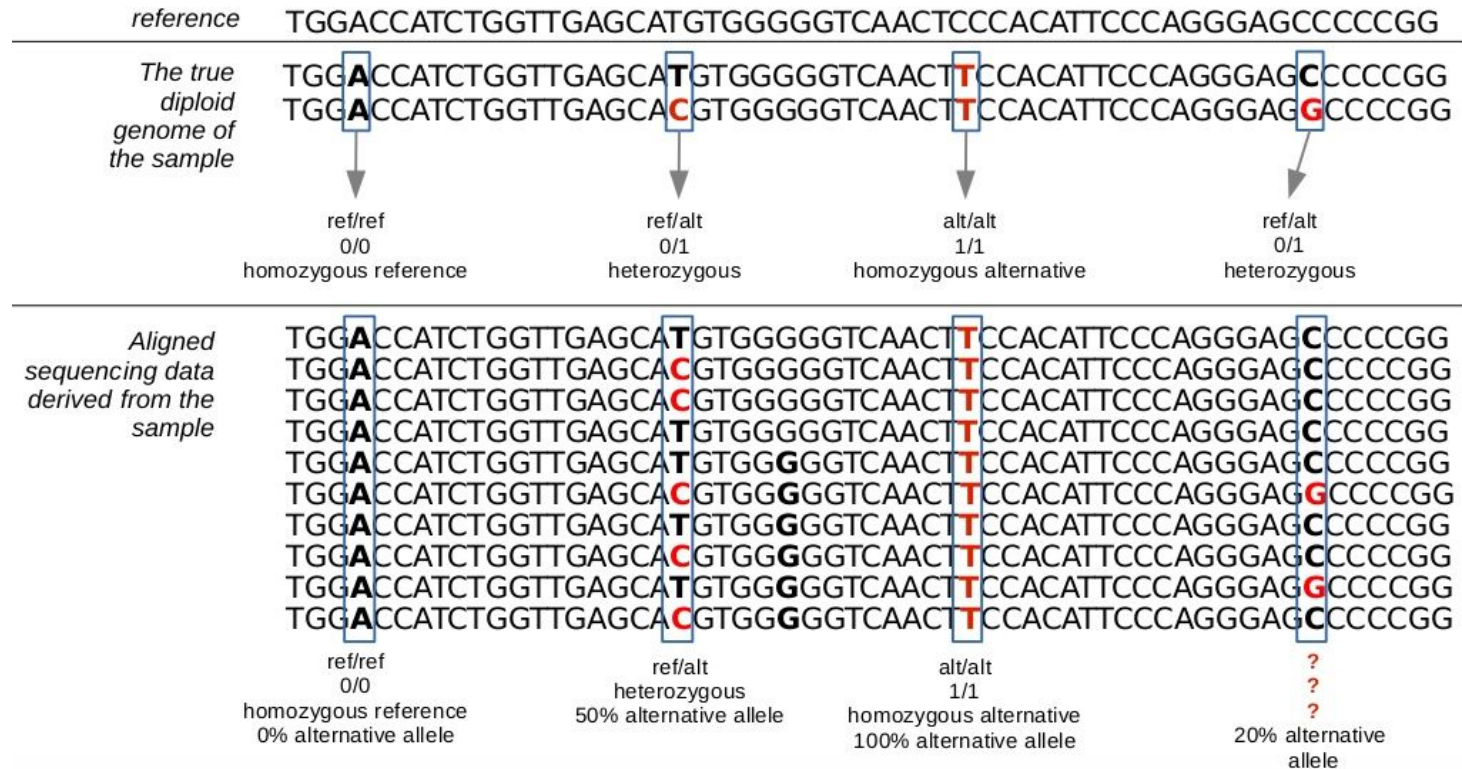
```
/opt/gatk/gatk ApplyBQSR -R ref.fa -I nodup.bam -bqsr model.grp -O  
recalibrated.bam
```

A typical workflow for variant calling



Cosa significa “Variant calling”?

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



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

```
##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 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

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
```

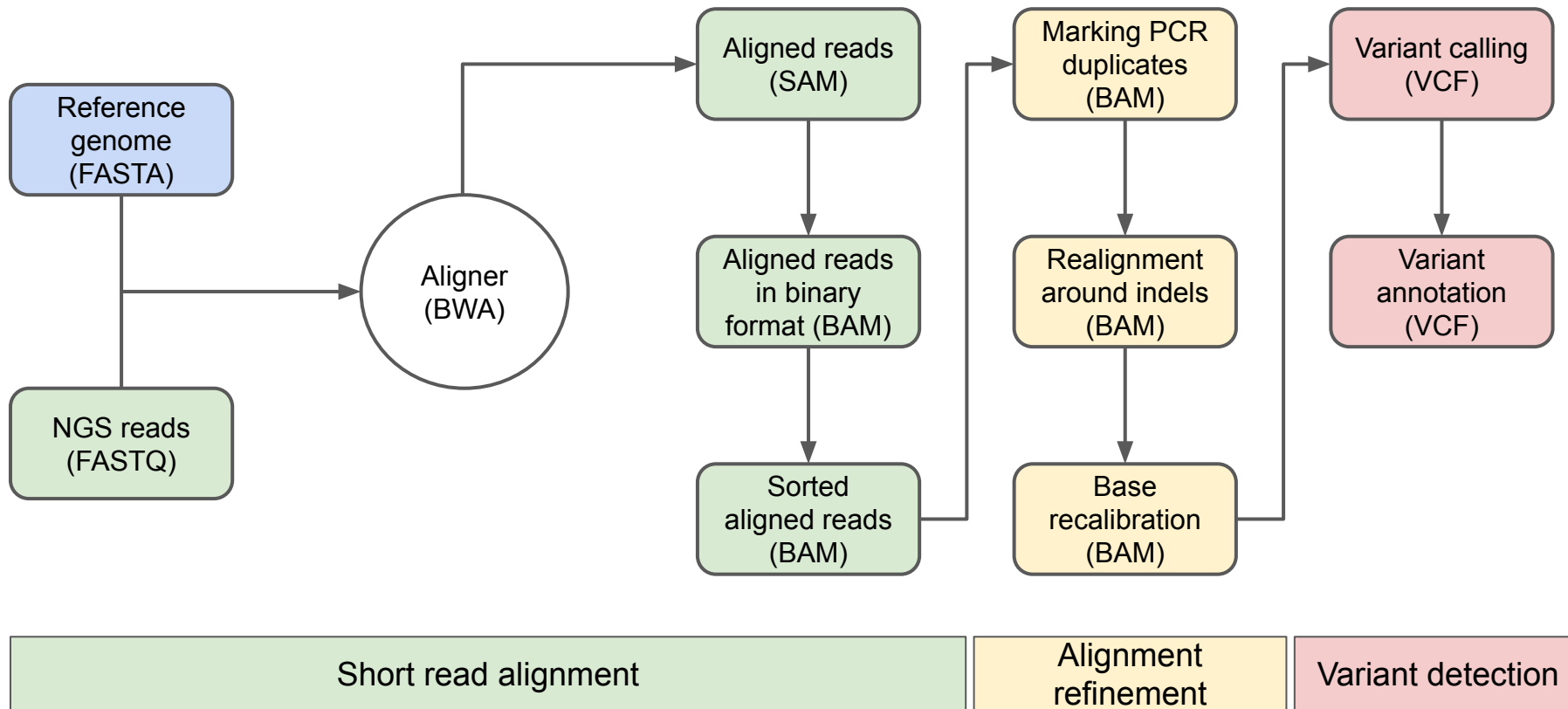
Variant technical filtering

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



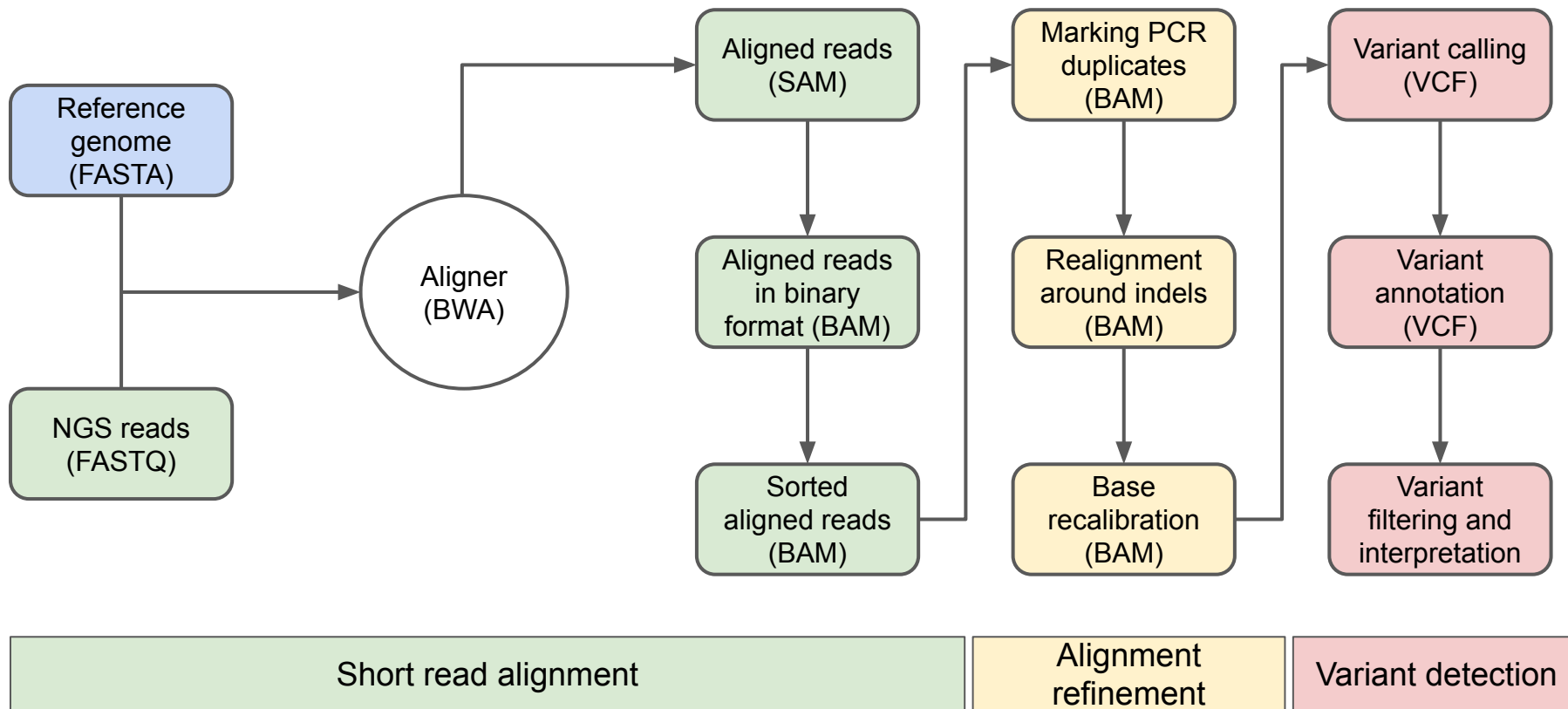
VCF annotation

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?.....

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

- 
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

[Switch to classic site](#)


[Download](#)
[f](#)
[t](#)
[s](#)

rs768520625

Organism

Homo sapiens

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](#)

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 [TGC] > * [TAG]	Coding Sequence Variant

Vedere annotazioni in dbSNP e COSMIC

COSMIC è un database per varianti trovate in studi oncologici

- Andare al sito di COSMIC: <https://cancer.sanger.ac.uk/cosmic/mutation/overview?id=24778681>
- Cercare l'ID "COSM5369532"

Mutation
COSV61070134

- Overview
- Tissue distribution
- Samples
- Pathways affected
- References

[Reset page](#)

Overview

This section shows a general overview of the selected mutation. It describes the source of the mutation i.e gene name/sample name/tissue name with unique ID, and also shows the mutation syntax at the amino acid and nucleotide sequence level. You can see more information on our [help pages](#).

Genomic Mutation ID	COSV61070134
Legacy Identifier	COSM5369532
Gene name	MAP2K1
AA mutation	p.K57E (Substitution - Missense, position 57, K→E)
CDS mutation	c.169A>G (Substitution, position 169, A→G)
SNP	No
Nucleotides inserted	n/a
Genomic coordinates	GRCh37, 15:66727453..66727453 , view Ensembl contig
CDD	NP_002746.1
HomoloGene	2063 , view the multiple sequence alignment
Ever confirmed somatic?	Yes
FATHMM prediction	Pathogenic (score 0.99)
Remark	n/a
Recurrent	n/a
Drug resistance	Resistance has been observed for the following drugs in samples curated with this mutation (or the DNA variant at the same genomic location on an alternative transcript, overlapping gene or fusion, which shares a COSM id) Note that the same resistance pattern may not apply to all samples. For more details, look at the Samples section. Dabrafenib
Alternative ids	n/a

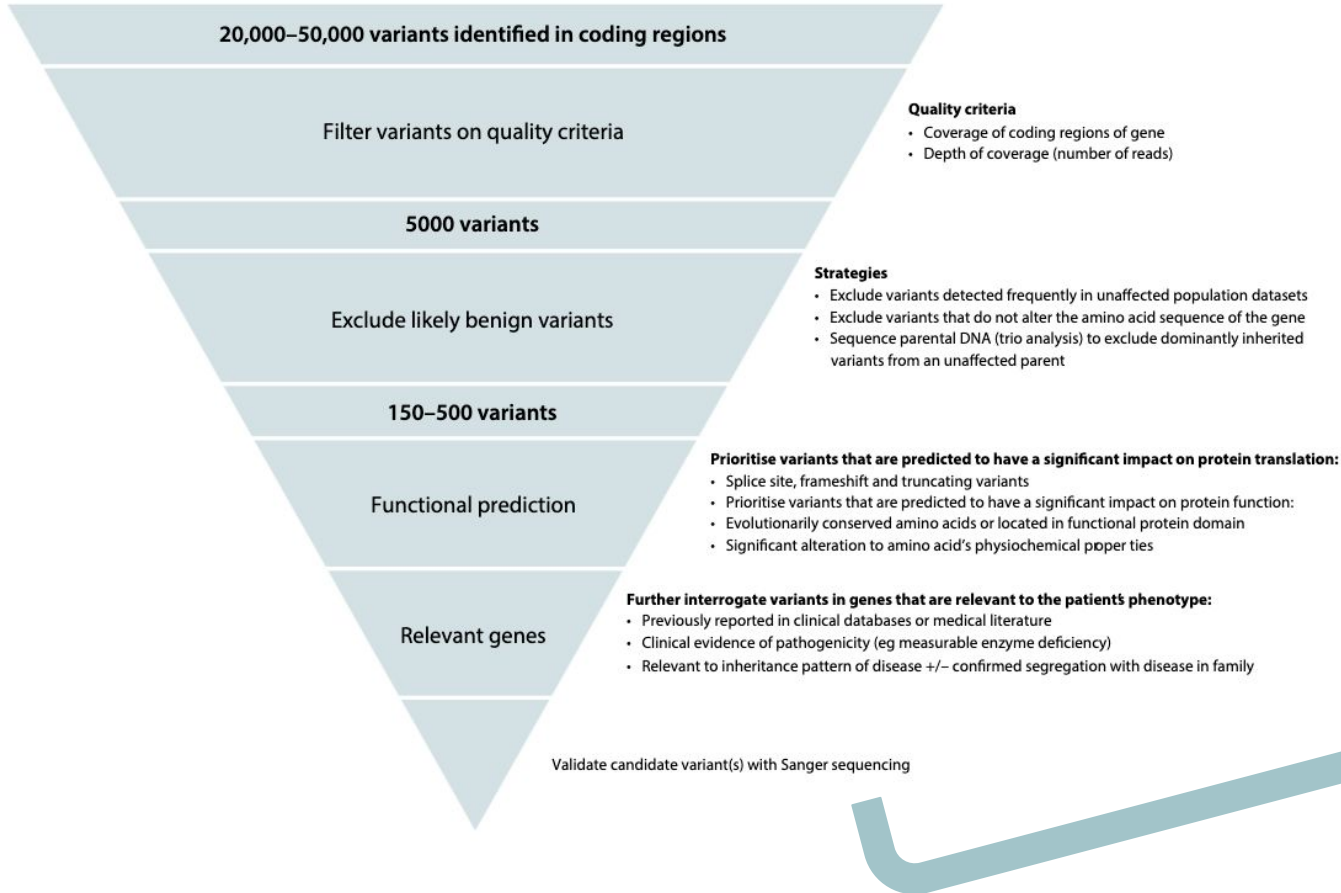
GRCh37 · COSMIC v92

Esempi di varianti annotate

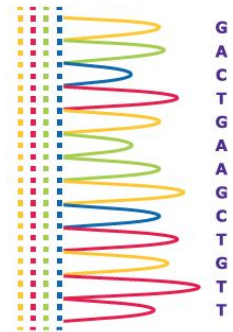
```
1      2488138 rs768520625      G      A      .      PASS      DB:SOMATIC;VT=SNP;ANN=A|stop_gained|HIGH|TNFRSF14|ENSG00000157873|transcript|ENST00000355716|protein_coding|1/8|c.35G>A|p.Trp12*|334/1707|35/852|12/283||A|stop_gained|HIGH|TNFRSF14|ENSG00000157873|transcript|ENST00000426449|protein_coding|2/6|c.35G>A|p.Trp12*|144/660|35/551|12/182||WARNING TRANSCRIPT I
NCOMPLETE,A|stop_gained|HIGH|TNFRSF14|ENSG00000157873|transcript|ENST00000434817|protein_coding|2/7|c.35G>A|p.Trp12*|175/834|35/694|12/230||WARNING TRANSCRIPT INCOMPLETE,A|stop_gained|
HIGH|TNFRSF14|ENSG00000157873|transcript|ENST00000435221|protein_coding|2/7|c.35G>A|p.Trp12*|566/1225|35/694|12/230||WARNING TRANSCRIPT INCOMPLETE,A|stop_gained|HIGH|TNFRSF14|ENSG00000
157873|transcript|ENST00000451778|protein_coding|2/7|c.35G>A|p.Trp12*|168/827|35/694|12/230||WARNING TRANSCRIPT INCOMPLETE,A|stop_gained|HIGH|TNFRSF14|ENSG00000157873|transcript|ENST00
000409119|protein_coding|2/7|c.35G>A|p.Trp12*|155/896|35/582|12/193||A|splice_region_variant&intron_variant|LOW|RP3-395M20.8|ENSG00000238164|transcript|ENST00000452793|antisense|1/3|n.56-3C>T||||A|splice_region_variant&intron_variant|LOW|RP3-395M20.8|ENSG00000238164|transcript|ENST00000432521|lincRNA|n.-2549C>T||||2549|A|upstream_gene_variant|MODIFIER|RP3-395
M20.8|ENSG00000238164|transcript|ENST00000432521|lincRNA|n.-2549C>T||||2549|A|upstream_gene_variant|MODIFIER|RP3-395M20.8|ENSG00000238164|transcript|ENST00000449660|lincRNA|n.-4593
C>T||||4593|A|upstream_gene_variant|MODIFIER|RP3-395M20.8|ENSG00000238164|transcript|ENST00000443892|lincRNA|n.-2454C>T||||2454|A|upstream_gene_variant|MODIFIER|TNFRSF14|ENSG00000
157873|transcript|ENST00000466750|retained_intron|n.-1114G>A||||1114|A|upstream_gene_variant|MODIFIER|TNFRSF14|ENSG00000157873|transcript|ENST00000471768|retained_intron|n.-2881G>A
||||2881|A|upstream_gene_variant|MODIFIER|TNFRSF14|ENSG00000157873|transcript|ENST00000463835|retained_intron|n.-3855G>A||||3855|A|upstream_gene_variant|MODIFIER|TNFRSF14|ENSG0000
0157873|transcript|ENST00000482602|processed_transcript|n.-3983G>A||||3983|A|intron_variant|MODIFIER|TNFRSF14|ENSG00000157873|transcript|ENST00000475523|retained_intron|1/5|n.70+244
G>A|||||A|non_coding_exon_variant|MODIFIER|TNFRSF14|ENSG00000157873|transcript|ENST00000442392|processed_transcript|1/4|n.295G>A|||||A|non_coding_exon_variant|MODIFIER|TNFRSF14|EN
SG00000157873|transcript|ENST00000482074|retained_intron|1/3|n.294G>A|||||A|non_coding_exon_variant|MODIFIER|TNFRSF14|ENSG00000157873|transcript|ENST00000496064|retained_intron|1/5|n.
143G>A|||||A|non_coding_exon_variant|MODIFIER|TNFRSF14|ENSG00000157873|transcript|ENST00000463471|retained_intron|1/5|n.129G>A|||||LOF=(TNFRSF14|ENSG00000157873|16|0.38);NMD=(TNFRS
F14|ENSG00000157873|16|0.38)      GT:AD:BQ:DP:FA:SS      0:31,0:..:31:0.00:0
```

```
15      66727453      COSM5369532;rs397516790      A      G      .      PASS      SOMATIC;VT=SNP;AA=p.K57E;CDS=c.169A>G;CNT=2;GENE=MAP2K1;STRAND=+;ASP;CLNACC=RCV000037591.3|RCV000158014.
1.RCV000443354.1;CLNALLE=1.2;CLNDBN=Cardio-facio-cutaneous syndrome|Rasopathy,Malignant melanoma;CLNDSDB=MedGen:Orphanet;SNOMED_CT|MedGen:Orphanet,MeSH:MedGen:SNOMED_CT;CLNDSDBID=C1275
081:ORPHA1340:403770008|CN166718:ORPHA98733,D008545:C0025202:2092003;CLNHGVS=NC 000015.9:g.66727453A>C,NC 000015.9:g.66727453A>G;CLNORIGIN=1,2;CLNREVSTAT=single|single,no_criteria;CLNS
IG=4|0,4;GENEINFO=MAP2K1:5604;LSD;NSM;PM;PMC;REF;RS=397516790;RSP05=66727453;SAO=3;SSR=0;VC=SNV;VP=0x05006000a05000002100100;WGT=1;dbSNPBuildID=138;dbNSFP_MetalR_pred=0;dbNSFP_MetaSV
M_pred=0;ANN=G|structural_interaction_variant|HIGH|MAP2K1|ENSG00000169032|interaction|3EQC:A 57-A 129:ENST00000307102|protein_coding|2/11|c.169A>G|||||G|structural_interaction_variant
|HIGH|MAP2K1|ENSG00000169032|interaction|3EQC:A 57-A 129:ENST00000307102|protein_coding|2/11|c.169A>G|||||G|structural_interaction_variant|HIGH|MAP2K1|ENSG00000169032|interaction|3EQ
D:A 57-A 119:ENST00000307102|protein_coding|2/11|c.169A>G|||||G|structural_interaction_variant|HIGH|MAP2K1|ENSG00000169032|interaction|3EQF:A 57-A 119:ENST00000307102|protein_coding|
2/11|c.169A>G|||||G|structural_interaction_variant|HIGH|MAP2K1|ENSG00000169032|interaction|3EQG:A 57-A 129:ENST00000307102|protein_coding|2/11|c.169A>G|||||G|structural_interaction
_variant|HIGH|MAP2K1|ENSG00000169032|interaction|3EQG:A 57-A 129:ENST00000307102|protein_coding|2/11|c.169A>G|||||G|structural_interaction_variant|HIGH|MAP2K1|ENSG00000169032|interac
tion|3EQG:A 57-A 129:ENST00000307102|protein_coding|2/11|c.169A>G|||||G|structural_interaction_variant|HIGH|MAP2K1|ENSG00000169032|interaction|3EQH:A 57-A 119:ENST00000307102|protein
_coding|2/11|c.169A>G|||||G|structural_interaction_variant|HIGH|MAP2K1|ENSG00000169032|interaction|3EQH:A 57-A 129:ENST00000307102|protein_coding|2/11|c.169A>G|||||G|structural int
eraction_variant|HIGH|MAP2K1|ENSG00000169032|interaction|3EQI:A 57-A 119:ENST00000307102|protein_coding|2/11|c.169A>G|||||G|structural_interaction_variant|HIGH|MAP2K1|ENSG00000169032
|interaction|3EQI:A 57-A 129:ENST00000307102|protein_coding|2/11|c.169A>G|||||G|structural_interaction_variant|HIGH|MAP2K1|ENSG00000169032|interaction|3W8Q:A 57-A 120:ENST00000307102
|protein_coding|2/11|c.169A>G|||||G|structural_interaction_variant|HIGH|MAP2K1|ENSG00000169032|interaction|3W8Q:A 57-A 129:ENST00000307102|protein_coding|2/11|c.169A>G|||||G|struct
ural_interaction_variant|HIGH|MAP2K1|ENSG00000169032|interaction|3ZLS:A 57-A 119:ENST00000307102|protein_coding|2/11|c.169A>G|||||G|structural_interaction_variant|HIGH|MAP2K1|ENSG000
00169032|interaction|3ZLW:A 57-A 119:ENST00000307102|protein_coding|2/11|c.169A>G|||||G|structural_interaction_variant|HIGH|MAP2K1|ENSG00000169032|interaction|3ZLX:A 57-A 129:ENST000
00307102|protein_coding|2/11|c.169A>G|||||G|structural_interaction_variant|HIGH|MAP2K1|ENSG00000169032|interaction|3ZLY:A 57-A 119:ENST00000307102|protein_coding|2/11|c.169A>G|||||G|struct
ural_interaction_variant|HIGH|MAP2K1|ENSG00000169032|interaction|3ZLY:A 57-A 129:ENST00000307102|protein_coding|2/11|c.169A>G|||||G|structural_interaction_variant|HIGH|MAP2K1|
ENSG00000169032|interaction|3ZM4:A 57-A 119:ENST00000307102|protein_coding|2/11|c.169A>G|||||G|structural_interaction_variant|HIGH|MAP2K1|ENSG00000169032|interaction|3ZM4:A 57-A 129
:ENST00000307102|protein_coding|2/11|c.169A>G|||||G|missense_variant|MODERATE|MAP2K1|ENSG00000169032|transcript|ENST00000307102|protein_coding|2/11|c.169A>G|p.Lys576Glu|700/3410|169/1
182|57/393||G|non_coding_transcript_exon_variant|MODIFIER|MAP2K1|ENSG00000169032|transcript|ENST00000425818|retained_intron|2/5|n.680A>G|||||GT:AD:BQ:DP:FA:SS      0:47,0:..:47:0.00
:0      0/1:106,12:27:119:0.102:2
```


Example of variant filtering

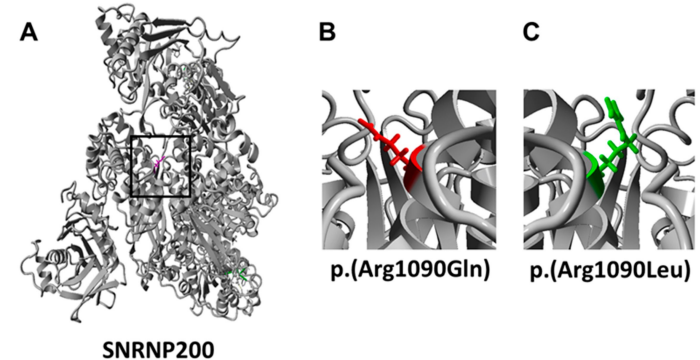
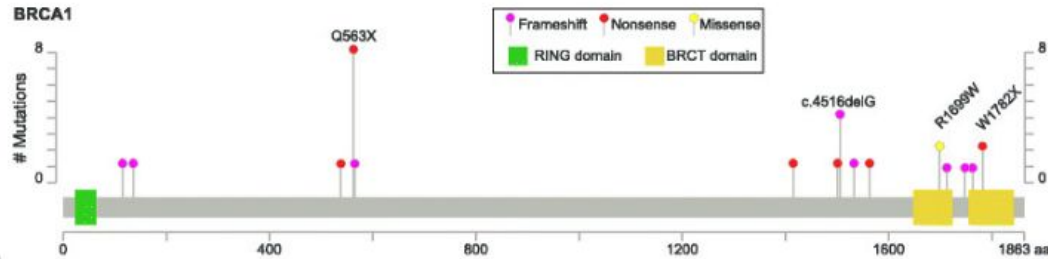
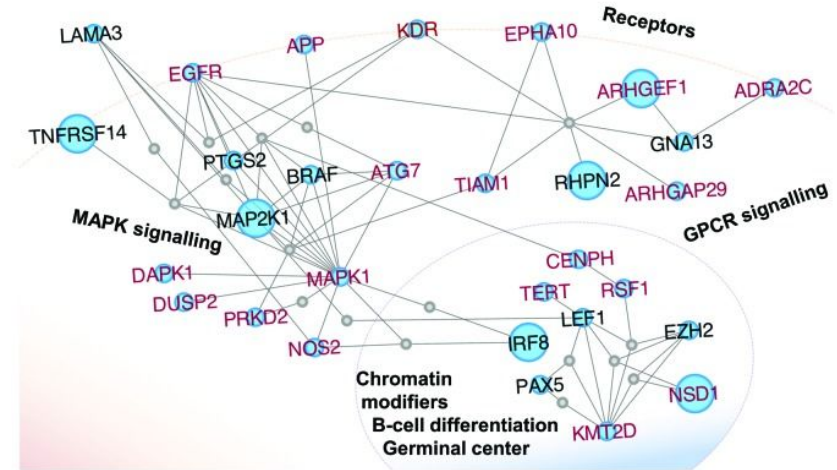
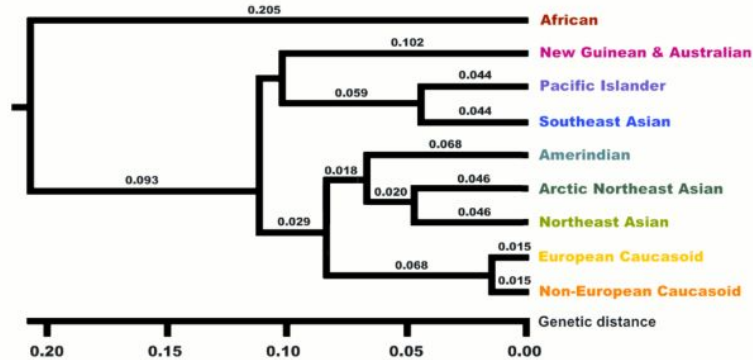


Sanger sequencing



Output chromatogram

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 **Esercitazione6_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 **Esercitazione6_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**