Applied Computational Multi-Omics

Genetic Alterations and Functional Impact

Daniel Sobral, INSA

17/03/2025

Genetic Alterations and Functional Impact

Learning Objectives

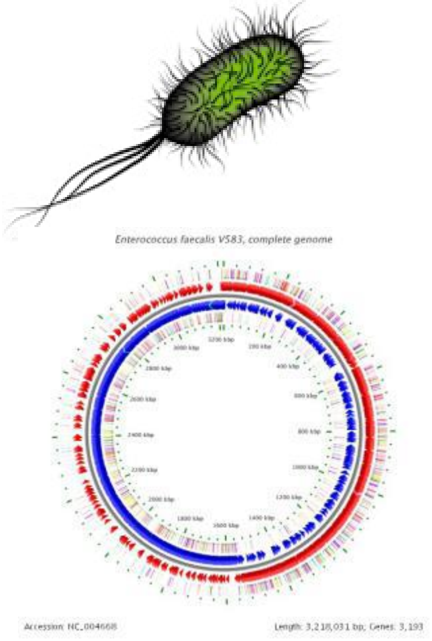
Introduction to variant calling

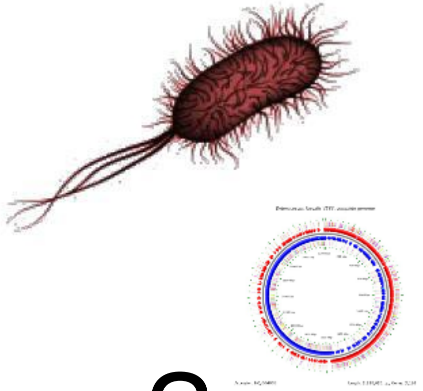
- 2 major classes: SNPs / indels and large structural variants
- Factors that influence the variant calling process
- Overview of the VCF file format
- Variant quality and Genotype quality

Introduction to Variant Annotation

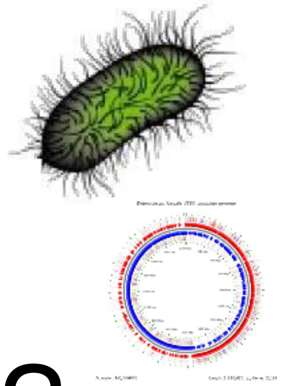
Variant Calling

Common question: find mutations underlying phenotypes


Enterococcus faecalis V583, complete genome
Accession: NC_004668
Length: 3,218,031 bp; Genes: 3,193



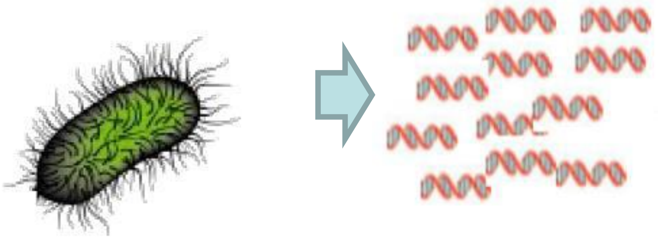
?



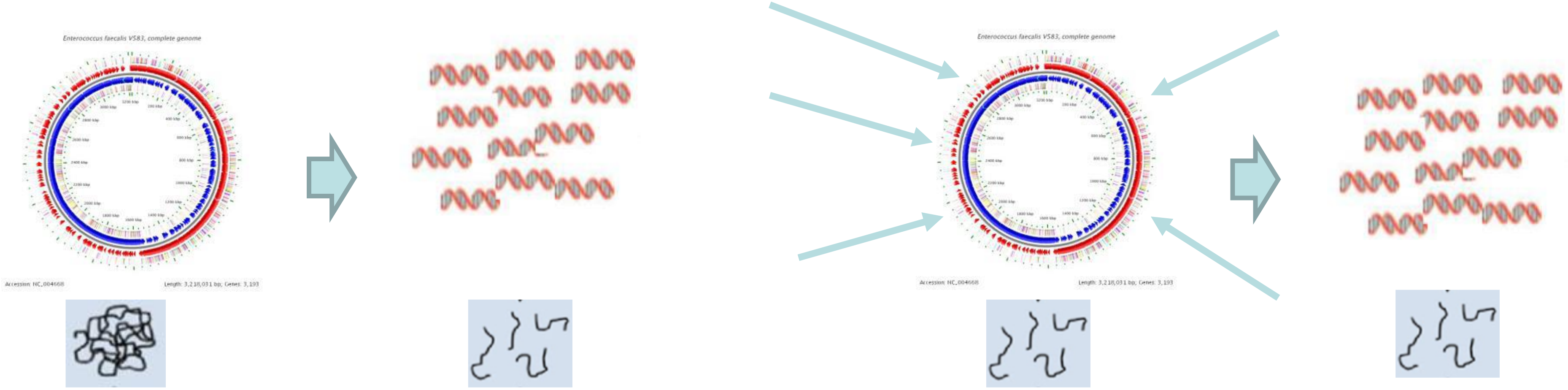
?

Data pre-processing for variant discovery

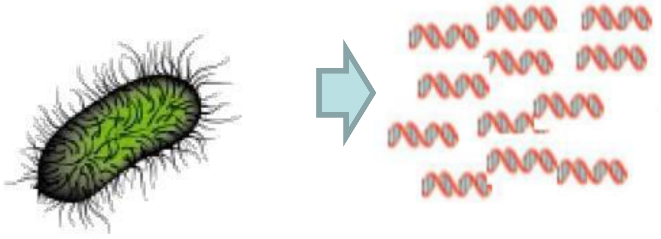
DNA Extraction



Whole Genome VS Targeted


Enterococcus faecalis V583, complete genome
Accession: NC_004668
Length: 3,218,031 bp; Genes: 3,193
Enterococcus faecalis V583, complete genome
Accession: NC_004668
Length: 3,218,031 bp; Genes: 3,193

DNA Extraction



Eg. TruSight One Enrichment Panel (Human)
Eg. SARS-CoV-2 Artic Amplicon panel

Note: Amplicon VS Enrichment

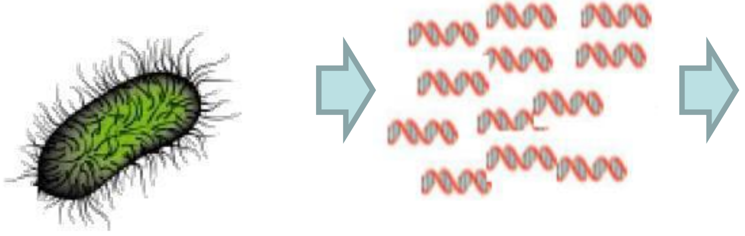
How to calculate coverage

Example for the Human Genome (~$3x10^9$ bp):

- WGS 30x coverage, 150bp read pairs
  - $30 \times 3x10^9 / (150x2) = 3x10^8$ = 300 million read pairs

- WES (~1-2% genome) 30x coverage, 150bp read pairs
  - 6 million read pairs (theoretical minimum, but usually more)

Data pre-processing for variant discovery

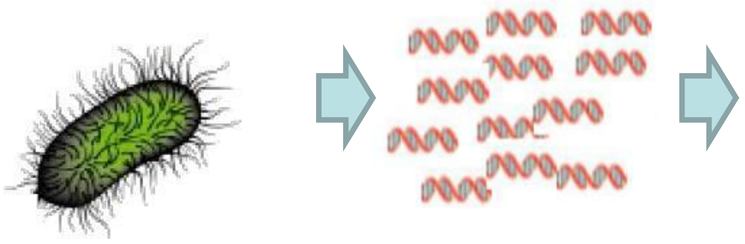


DNA Extraction

Sequencing

TGCTCAGTTA TGAC ATGGAGT TTT CGTTGT **(“raw” fastQ)**

Data pre-processing for variant discovery



DNA Extraction **Sequencing**

TGCTCAGTT TGAC ATGGAGT GTTGT **(“clean” fastQ)**

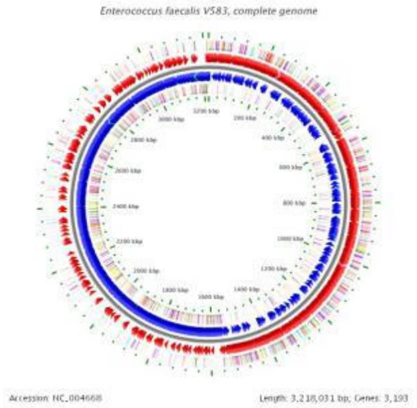
QC and pre-processing

TGCTCAGTTA TGAC ATGGAGT TTT CGTTGT **(“raw” fastQ)**

Data pre-processing for variant discovery

(fasta file)

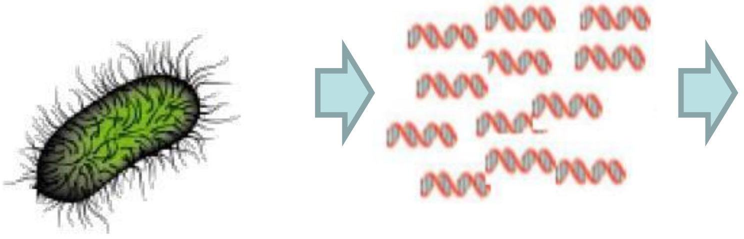
Reference Genome



geneA geneB geneC

AAGCGATGACTGCATGCACAGTTGTGTGTTTTCACGTGAC

DNA Extraction **Sequencing**



TGCTCAGTT TGAC ATGGAGT GTTGT **(“clean” fastQ)**

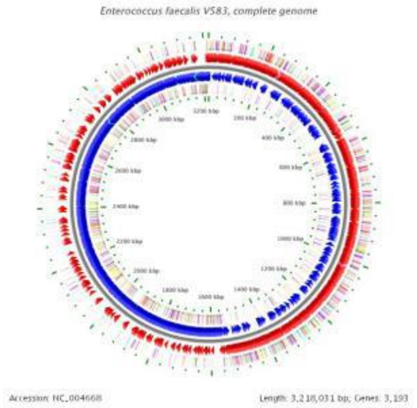
QC and pre-processing

TGCTCAGTTA TGAC ATGGAGT TTT CGTTGT **(“raw” fastQ)**

Data pre-processing for variant discovery

(fasta file)

Reference Genome

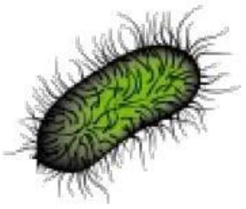

Enterococcus faecalis V583, complete genome
Accession: NC_004668
Length: 3,218,031 bp; Genes: 3,193

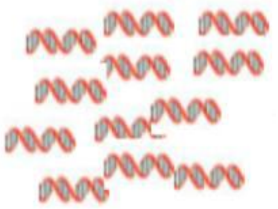
TGCTCAGTT

geneA **geneB** **geneC**

AAG**CGATGAC**TGCAT**GCACAG**TTGTGT**GTTTTCAC**GTGAC

DNA Extraction **Sequencing**





TGCTCAGTT TGAC ATGGAGT GTTGT **(“clean” fastQ)**

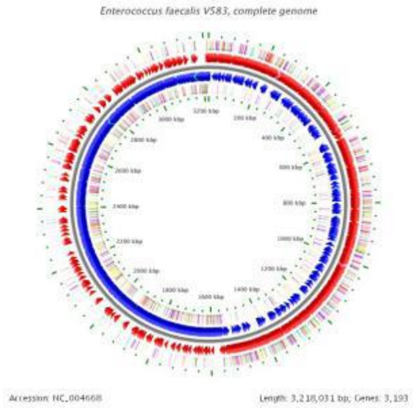
QC and pre-processing

TGCTCAGTTA TGAC ATGGAGT TTT CGTTGT **(“raw” fastQ)**

Data pre-processing for variant discovery

(fasta file)

Reference Genome

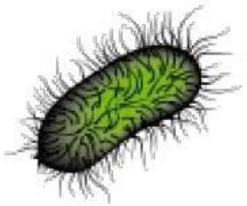

Enterococcus faecalis V583, complete genome
Accession: NC_004668
Length: 3,218,031 bp; Genes: 3,193

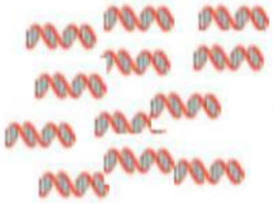
TGCTCAGTT

geneA geneB geneC

AAGCGATGACTGCATGCACAGTTGTGTGTTTTCACGTGAC

DNA Extraction **Sequencing** **Alignment**





TGCTCAGTT TGAC ATGGAGT GTTGT **(“clean” fastQ)**

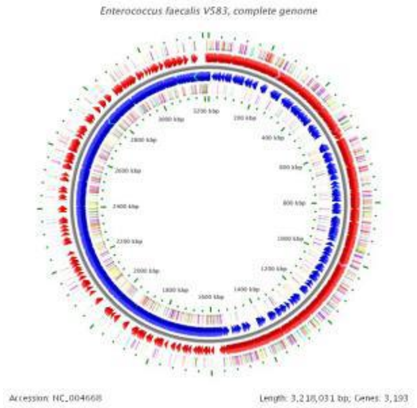
QC and pre-processing

TGCTCAGTTA TGAC ATGGAGT TTT CGTTGT **(“raw” fastQ)**

Data pre-processing for variant discovery

(fasta file)

Reference Genome


Enterococcus faecalis V583, complete genome
Accession: NC_004668
Length: 3,218,031 bp; Genes: 3,193

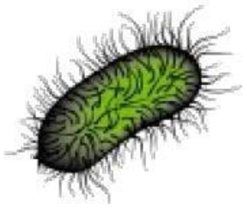
ACTCCAT GTTGT

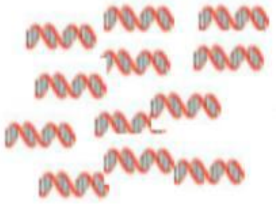
TGCTCAGTT

geneA geneB geneC

AAGCGATGACTGCATGCACAGTTGTGTGTTTTCACGTGAC

DNA Extraction **Sequencing** **Alignment**





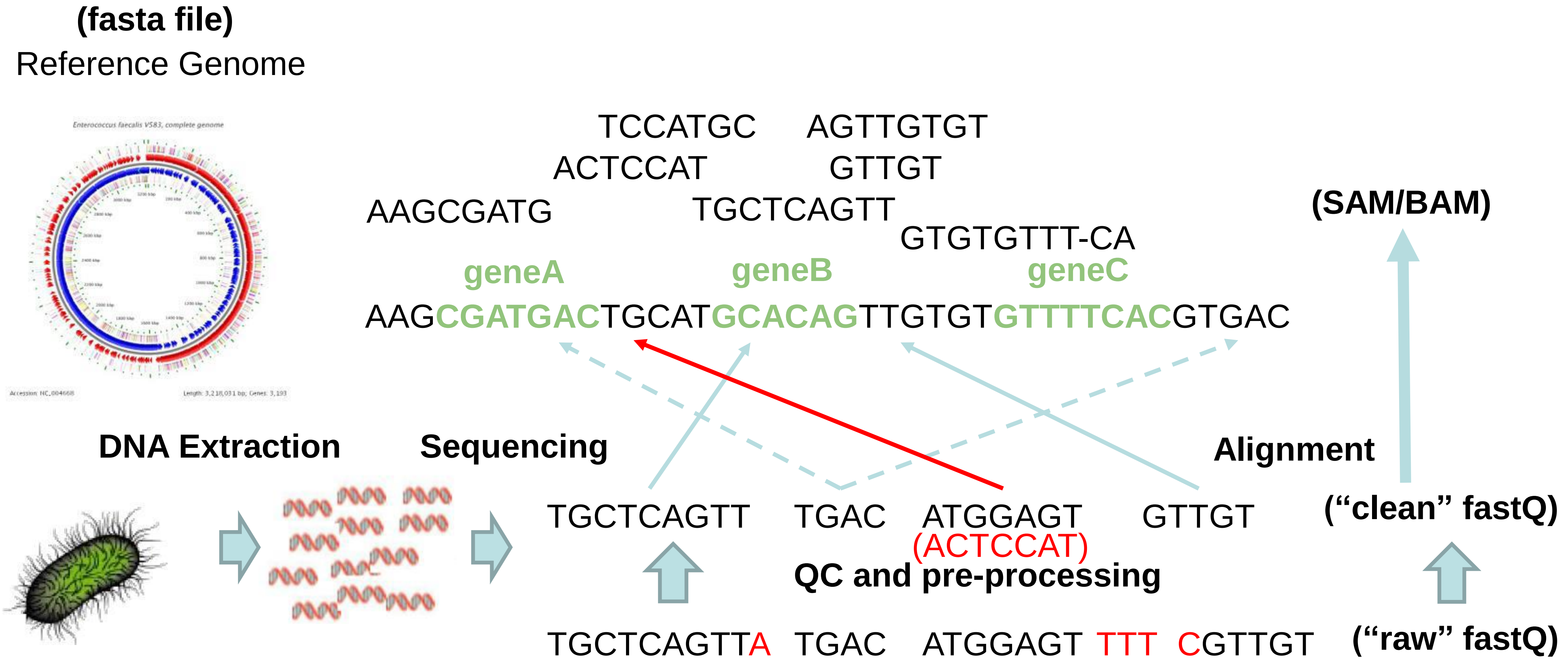
TGCTCAGTT TGAC ATGGAGT GTTGT **(“clean” fastQ)**

(ACTCCAT)

QC and pre-processing

TGCTCAGTTA TGAC ATGGAGT TTT CGTTGT **(“raw” fastQ)**

Data pre-processing for variant discovery



Enterococcus faecalis V583, complete genome
Accession: NC_004668
Length: 3,218,031 bp; Genes: 3,193

(fasta file)

Reference Genome

TCCATGC AGTTGTGT

ACTCCAT GTTGT

AAGCGATG TGCTCAGTT

GTGTGTTT-CA

geneA geneB geneC

AAGCGATGACTGCATGCACAGTTGTGTGTTTTCACGTGAC

(SAM/BAM)

DNA Extraction **Sequencing** **Alignment**

TGCTCAGTT TGAC ATGGAGT GTTGT **(“clean” fastQ)**

(ACTCCAT)

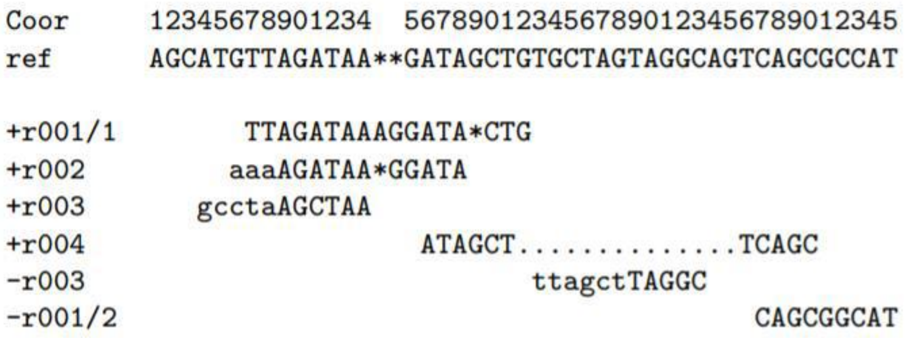
QC and pre-processing

TGCTCAGTTA TGAC ATGGAGT TTT CGTTGT **(“raw” fastQ)**

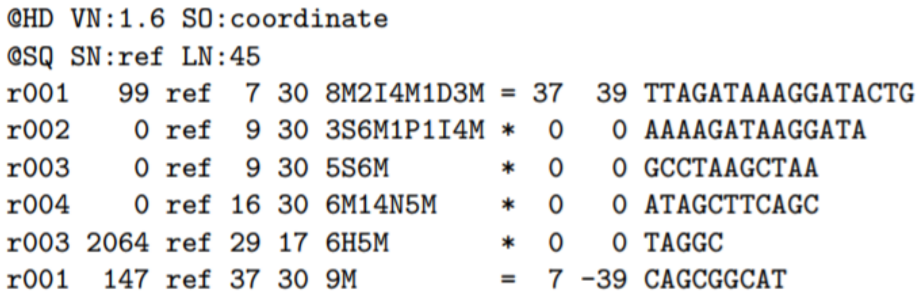
SAM/BAM format

A file format to represent alignments

BAM -> binary form of SAM



```
Coor     12345678901234  5678901234567890123456789012345
ref      AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1        TTAGATAAAGGATA*CTG
+r002         aaaAGATAA*GGATA
+r003       gcctaAGCTAA
+r004                     ATAGCT..............TCAGC
-r003                            ttagctTAGGC
-r001/2                                        CAGCGGCAT
```



```
@HD VN:1.6 SO:coordinate
@SQ SN:ref LN:45
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
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC
r001  147 ref 37 30 9M         =  7 -39 CAGCGGCAT
```

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

How alignment is made in practice

BWA (Burrows-Wheeler Aligner) is the most popular tool for WGS/WES

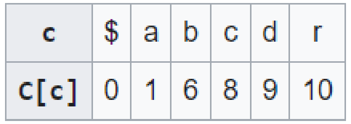
- Align millions of short reads to a human-sized genome in minutes

It is based on the FM-index of the Burrows-Wheeler Transform

"abracadabra$" $\xrightarrow{\text{BWT}}$ "ard$rcaaaabb" + C[c] table + Occ table

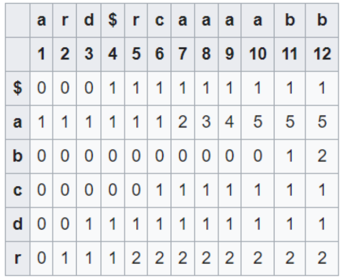
(genome)

C[c] of "ard$rcaaaabb"



c	$	a	b	c	d	r
C[c]	0	1	6	8	9	10

Occ(c, k) of "ard$rcaaaabb"



	a	r	d	$	r	c	a	a	a	a	b	b
	1	2	3	4	5	6	7	8	9	10	11	12
$	0	0	0	1	1	1	1	1	1	1	1	1
a	1	1	1	1	1	1	2	3	4	5	5	5
b	0	0	0	0	0	0	0	0	0	0	1	2
c	0	0	0	0	0	1	1	1	1	1	1	1
d	0	0	1	1	1	1	1	1	1	1	1	1
r	0	1	1	1	2	2	2	2	2	2	2	2

$\underbrace{\text{"ard\$rcaaaabb"} + C[c] + Occ(c, k)}_{\text{FM - index}}$

FM - index

https://en.wikipedia.org/wiki/FM-index

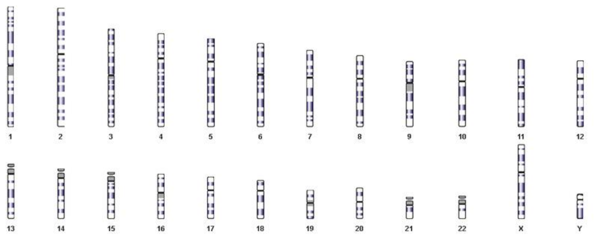
How alignment is made in practice

Genome BWT FM-índex needs to be created (only once)

$ bwa index genome.fasta

(genome.fasta file)

Reference Genome

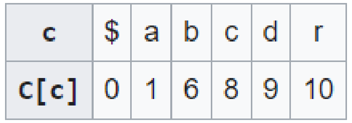


3x10^9 bases

AAGCGATGACTGCATGCACAGTTGTGTGTTTTCACGTGAC

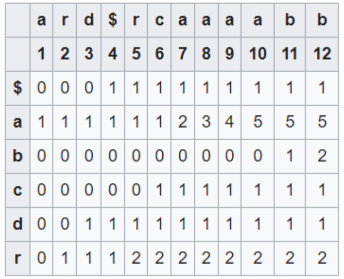
BWT

+



c	$	a	b	c	d	r
c[c]	0	1	6	8	9	10

+



	a	r	d	$	r	c	a	a	a	a	b	b
	1	2	3	4	5	6	7	8	9	10	11	12
$	0	0	0	1	1	1	1	1	1	1	1	1
a	1	1	1	1	1	1	2	3	4	5	5	5
b	0	0	0	0	0	0	0	0	0	0	1	2
c	0	0	0	0	0	1	1	1	1	1	1	1
d	0	0	1	1	1	1	1	1	1	1	1	1
r	0	1	1	1	2	2	2	2	2	2	2	2

FM - index

How alignment is made in practice

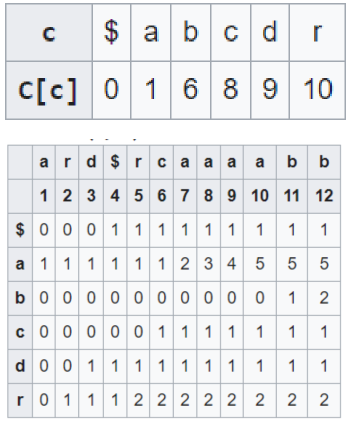
Basic command to generate alignments with BWA:

```
$ bwa mem genome.fasta reads_R1.fastq(.gz) reads_R2.fastq(.gz)>output.sam
```

reads_R1.fastq
TGCTCAGTT

reads_R2.fastq
ACGTCCGA

BWT



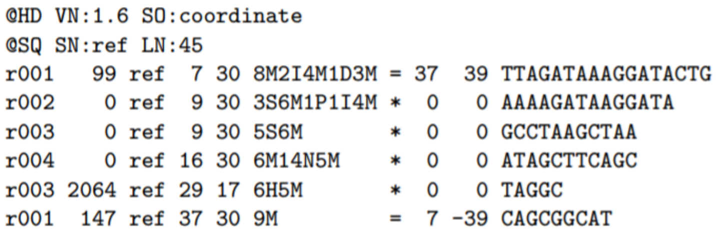
c	$	a	b	c	d	r
C[c]	0	1	6	8	9	10

	a	r	d	$	r	c	a	a	a	a	b	b
	1	2	3	4	5	6	7	8	9	10	11	12
$	0	0	0	1	1	1	1	1	1	1	1	1
a	1	1	1	1	1	1	2	3	4	5	5	5
b	0	0	0	0	0	0	0	0	0	0	1	2
c	0	0	0	0	0	1	1	1	1	1	1	1
d	0	0	1	1	1	1	1	1	1	1	1	1
r	0	1	1	1	2	2	2	2	2	2	2	2

Output in SAM format

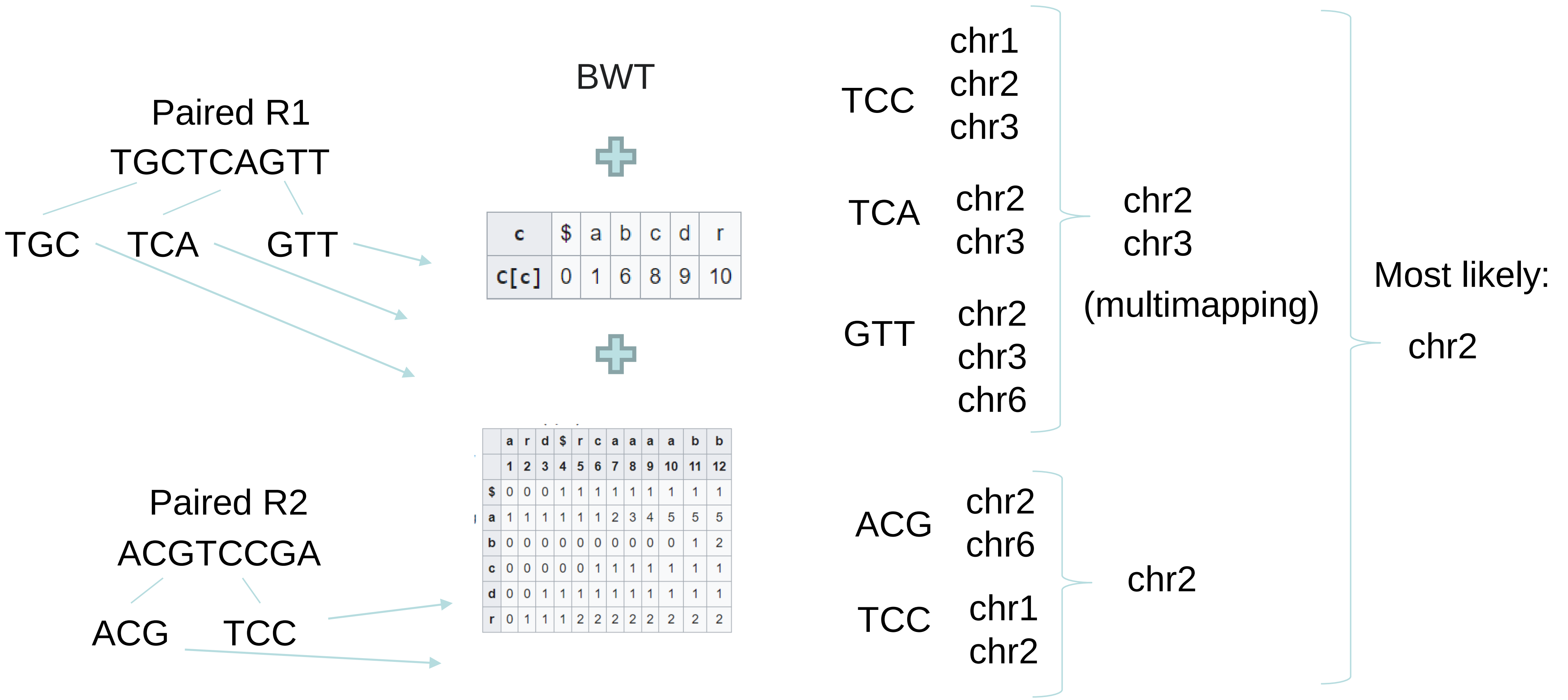
TGCTCAGTT Chr2 position xxx

ACGTCCGA Chr2 position yyy



```
@HD VN:1.6 SO:coordinate
@SQ SN:ref LN:45
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
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC
r001  147 ref 37 30 9M         =  7 -39 CAGCGGCAT
```

How alignment is made in practice



Paired R1

TGCTCAGTT

TGC TCA GTT

BWT

c	$	a	b	c	d	r
c[c]	0	1	6	8	9	10

	a	r	d	$	r	c	a	a	a	a	b	b
	1	2	3	4	5	6	7	8	9	10	11	12
$	0	0	0	1	1	1	1	1	1	1	1	1
a	1	1	1	1	1	1	2	3	4	5	5	5
b	0	0	0	0	0	0	0	0	0	0	1	2
c	0	0	0	0	0	1	1	1	1	1	1	1
d	0	0	1	1	1	1	1	1	1	1	1	1
r	0	1	1	1	2	2	2	2	2	2	2	2

Paired R2

ACGTCCGA

ACG TCC

How alignment is made in practice: Summary

- Special algorithms are used to have fast alignments
  - They are not guaranteed to be perfect but most of the time they are very good

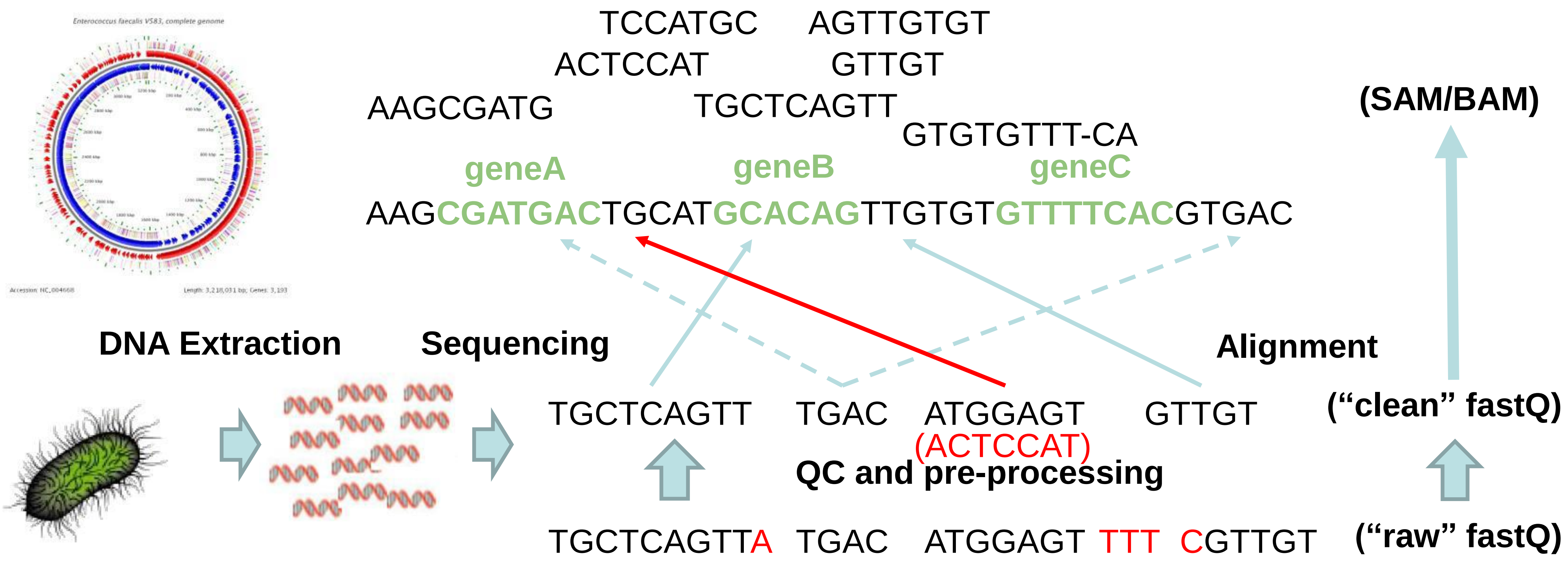
- A read can map equally well to multiple regions (multimappings)
  - BWA reports **one primary alignment (randomly chosen) with mapping quality of 0**
  - Depending on the software, it can generate **secondary alignments**
  - Information of paired reads are used to disambiguate multimappings if possible

- Alignments are made piece-wise (a read is split in segments)
  - A read alignment can be split in a primary and **supplementary alignment(s)**
    - Eg. splicing in RNA-Seq; large deletions
  - Sometimes, only a part of the read is aligned (the rest is “masked”/hidden)
    - Particularly in repetitive areas this can lead to false alignments

Variant Calling

(fasta file)

Reference Genome



Enterococcus faecalis V583, complete genome
Accession: NC_004668
Length: 3,218,031 bp; Genes: 3,193

TCCATGC AGTTGTGT

ACTCCAT GTTGT

AAGCGATG TGCTCAGTT

GTGTGTTT-CA

geneA geneB geneC

AAGCGATGACTGCATGCACAGTTGTGTGTTTTCACGTGAC

(SAM/BAM)

DNA Extraction

Sequencing

Alignment

TGCTCAGTT TGAC ATGGAGT GTTGT

(ACTCCAT)

(“clean” fastQ)

QC and pre-processing

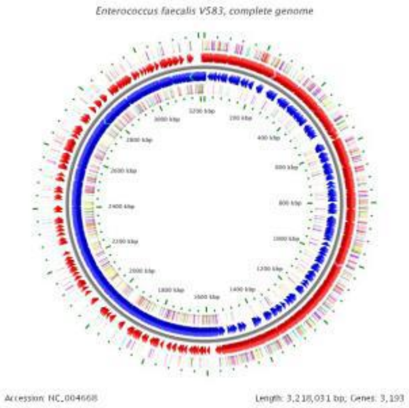
TGCTCAGTTA TGAC ATGGAGT TTT CGTTGT

(“raw” fastQ)

Variant Calling

(fasta file)

Reference Genome


Enterococcus faecalis V583, complete genome
Accession: NC_004668
Length: 3,218,031 bp; Genes: 3,193

SNV **SNV**

TCCATGC AGTTGTGT

ACTCCAT GTTGT

AAGCGATG TGCTCAGTT

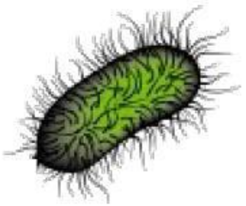
GTGTGTTT-CA

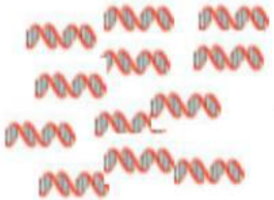
geneA geneB geneC

AAGCGATGACTGCATGCACAGTTGTGTGTTTTCACGTGAC

(SAM/BAM)

DNA Extraction **Sequencing** **Alignment**





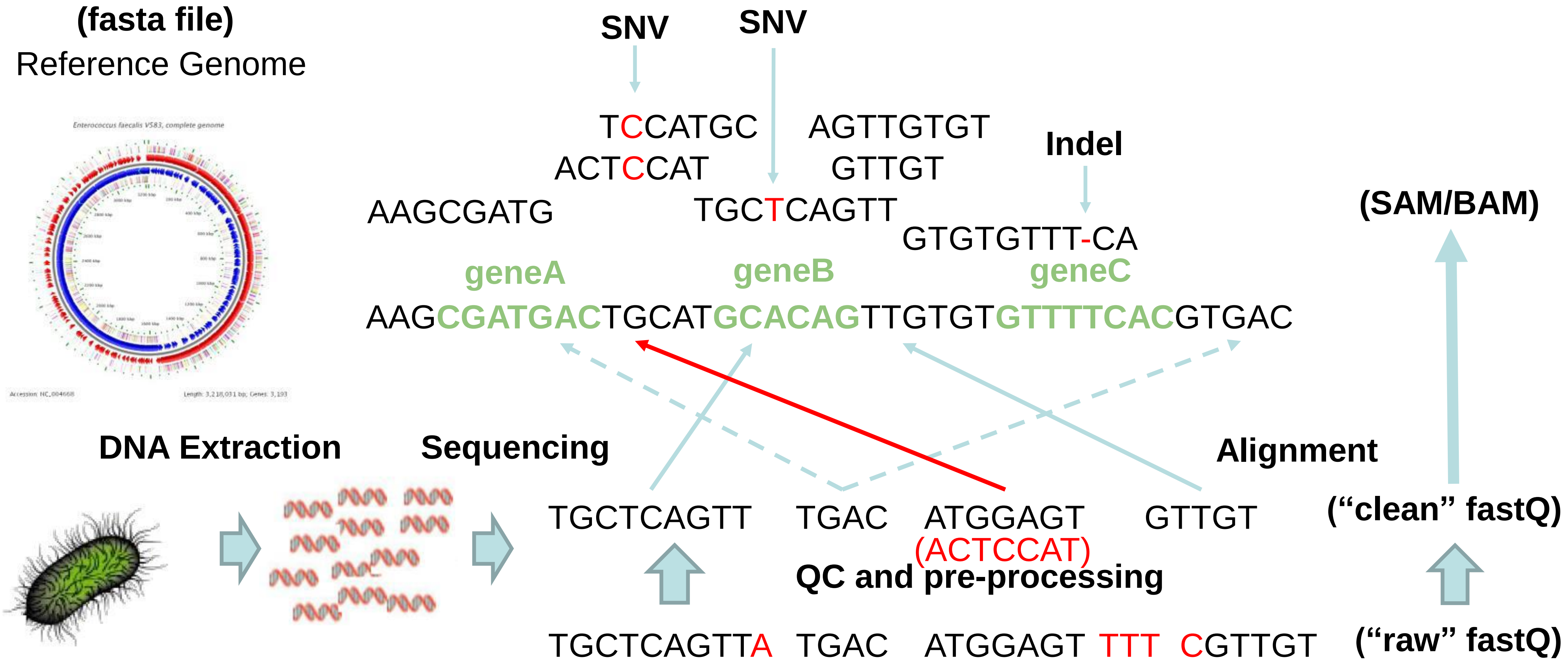
TGCTCAGTT TGAC ATGGAGT GTTGT **(“clean” fastQ)**

(ACTCCAT)

QC and pre-processing

TGCTCAGTTA TGAC ATGGAGT TTT CGTTGT **(“raw” fastQ)**

Variant Calling



(fasta file)

Reference Genome

Enterococcus faecalis V583, complete genome
Accession: NC_004668
Length: 3,218,031 bp; Genes: 3,193

SNV

SNV

Indel

TCCATGC AGTTGTGT

ACTCCAT GTTGT

AAGCGATG TGCTCAGTT

GTGTGTTT-CA

geneA geneB geneC

AAGCGATGACTGCATGCACAGTTGTGTGTTTTCACGTGAC

(SAM/BAM)

DNA Extraction

Sequencing

Alignment

TGCTCAGTT TGAC ATGGAGT GTTGT

(ACTCCAT)

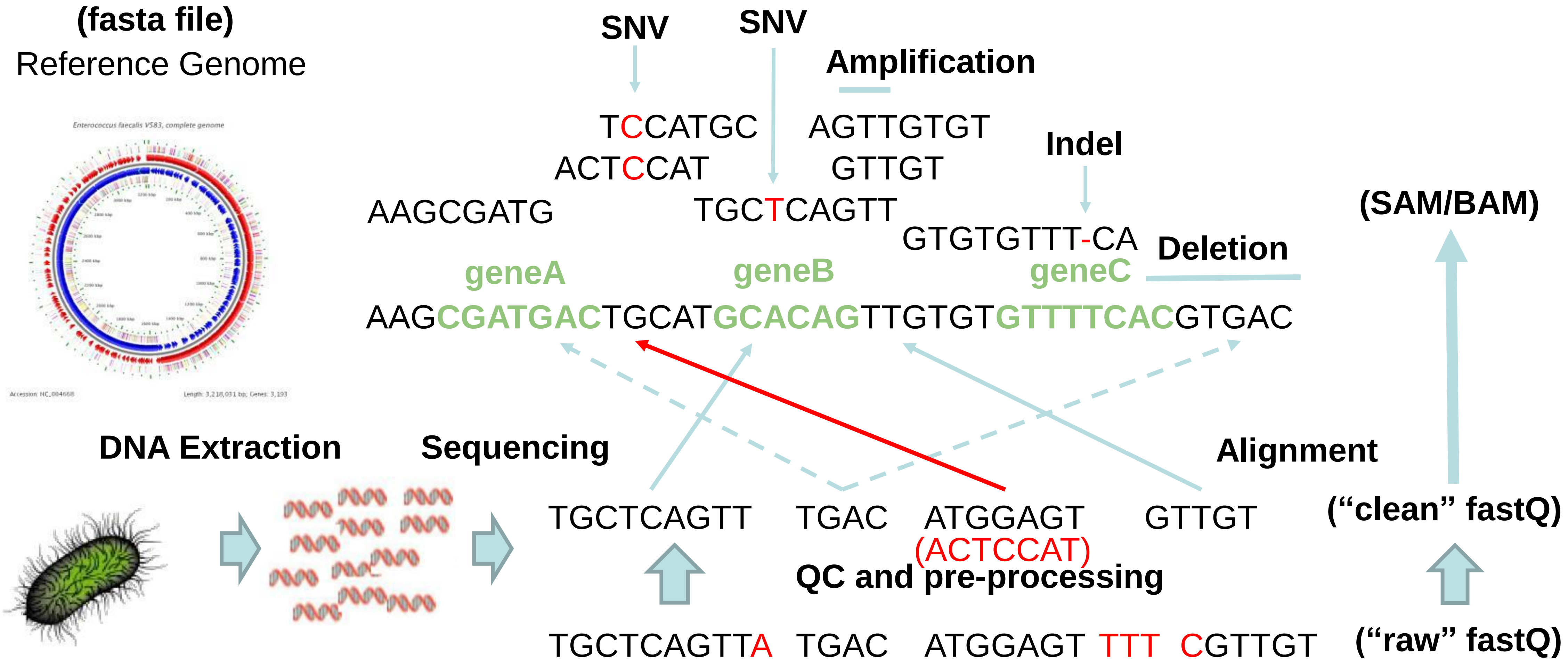
("clean" fastQ)

QC and pre-processing

TGCTCAGTTA TGAC ATGGAGT TTT CGTTGT

("raw" fastQ)

Variant Calling



Enterococcus faecalis V583, complete genome
Accession: NC_004668
Length: 3,218,031 bp; Genes: 3,193

(fasta file)

Reference Genome

SNV

SNV

Amplification

Indel

Deletion

(SAM/BAM)

TCCATGC AGTTGTGT

ACTCCAT GTTGT

AAGCGATG TGCTCAGTT

GTGTGTTT-CA

geneA geneB geneC

AAGCGATGACTGCATGCACAGTTGTGTGTTTTCACGTGAC

DNA Extraction

Sequencing

Alignment

TGCTCAGTT TGAC ATGGAGT (ACTCCAT) GTTGT

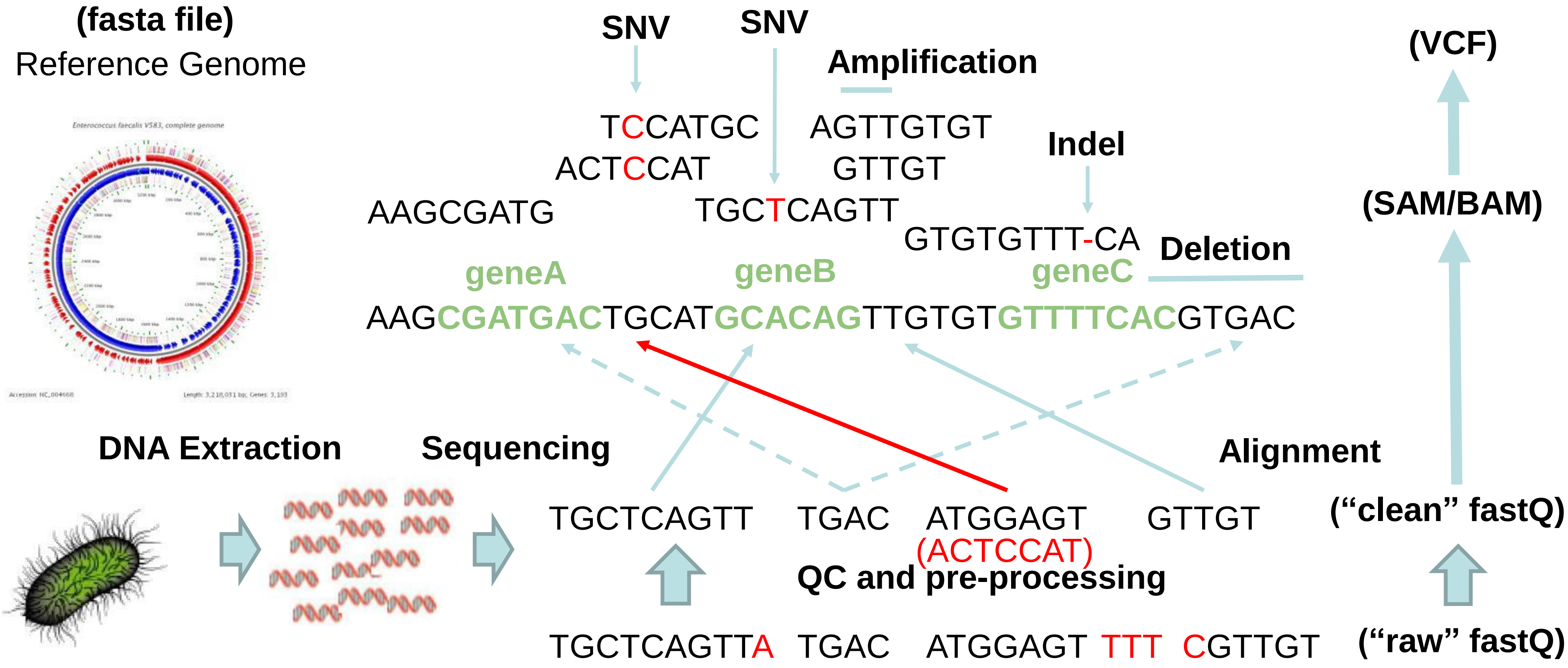
(“clean” fastQ)

QC and pre-processing

TGCTCAGTTA TGAC ATGGAGT TTT CGTTGT

(“raw” fastQ)

Variant Calling



Enterococcus faecalis V583, complete genome
Accession: NC_004668
Length: 3,218,031 bp; Genes: 3,193

(fasta file)
Reference Genome

SNV

SNV

Amplification

TCCATGC

AGTTGTGT

ACTCCAT

GTTGT

Indel

AAGCGATG

TGCTCAGTT

GTGTGTTT-CA

Deletion

geneA geneB geneC

AAGCGATGACTGCATGCACAGTTGTGTGTTTTCACGTGAC

(VCF)

(SAM/BAM)

DNA Extraction

Sequencing

Alignment

TGCTCAGTT TGAC ATGGAGT GTTGT

(ACTCCAT)

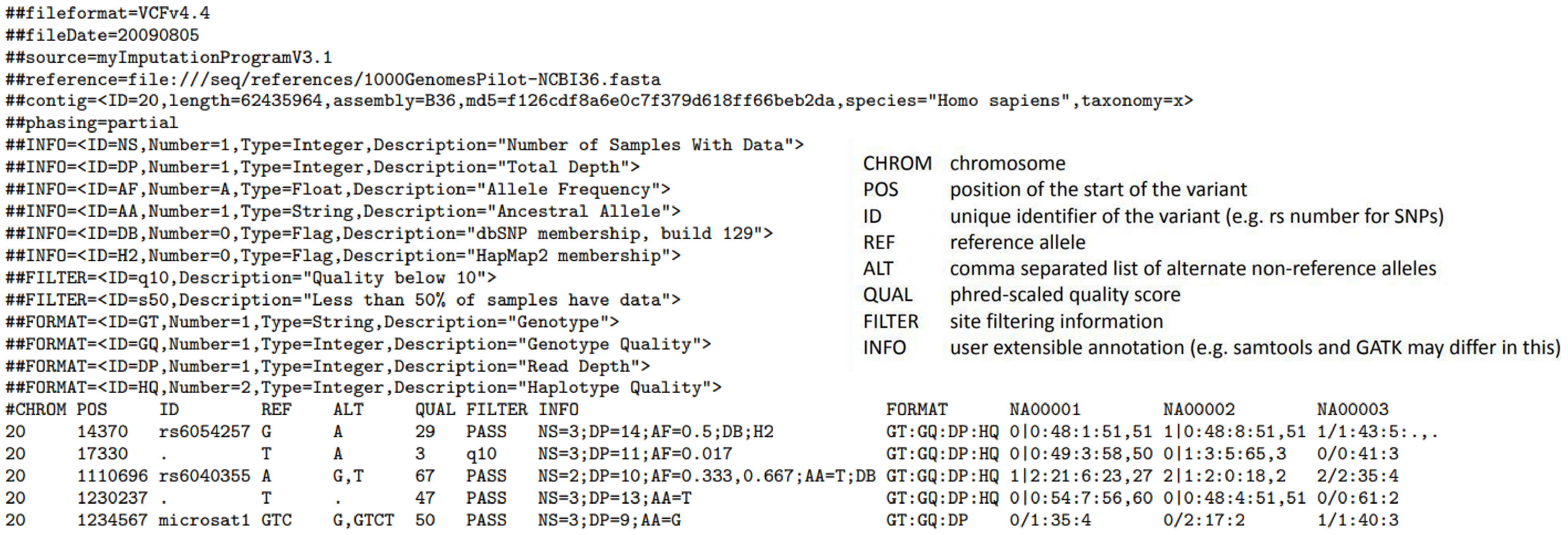
("clean" fastQ)

QC and pre-processing

TGCTCAGTTA TGAC ATGGAGT TTT CGTTGT

("raw" fastQ)

The VCF format

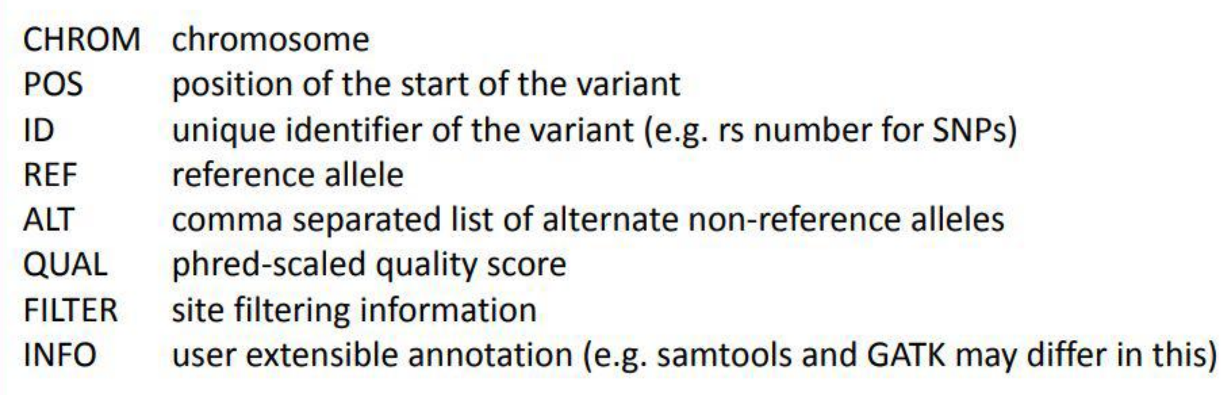


```
##fileformat=VCFv4.4
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

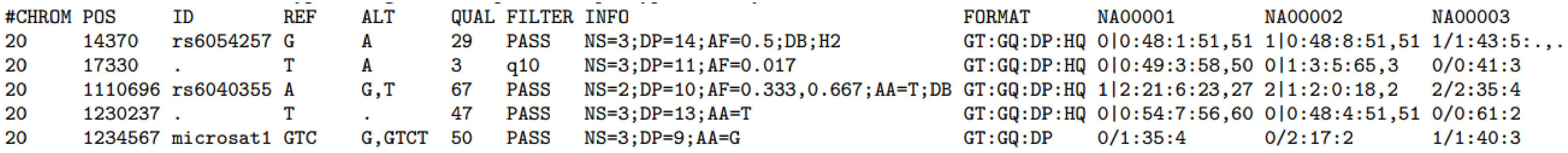
CHROM	chromosome
POS	position of the start of the variant
ID	unique identifier of the variant (e.g. rs number for SNPs)
REF	reference allele
ALT	comma separated list of alternate non-reference alleles
QUAL	phred-scaled quality score
FILTER	site filtering information
INFO	user extensible annotation (e.g. samtools and GATK may differ in this)

https://samtools.github.io/hts-specs/VCFv4.4.pdf

The VCF format



CHROM	chromosome
POS	position of the start of the variant
ID	unique identifier of the variant (e.g. rs number for SNPs)
REF	reference allele
ALT	comma separated list of alternate non-reference alleles
QUAL	phred-scaled quality score
FILTER	site filtering information
INFO	user extensible annotation (e.g. samtools and GATK may differ in this)

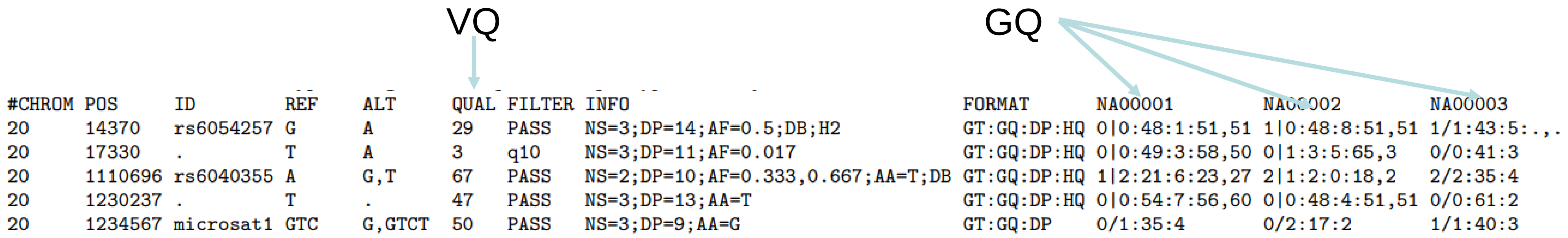


```
#CHROM POS     ID         REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257  G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .          T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355  A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .          T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1  GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

https://samtools.github.io/hts-specs/VCFv4.4.pdf

Variant Quality vs Genotype Quality

- Variant Quality
  - – Phred score estimating if variant is likely to be an artifact
- Genotype Quality
  - – Phred score estimating accuracy of estimated sample genotype



VQ

GQ

```
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

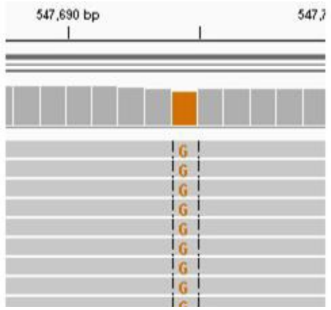
Variant Calling

- 2 Major types of Variants
  - – Single Nucleotide Variants (SNV) and small Indels
    - Smaller than the size of one read
  - – Large Structural Variants
    - Usually larger than the size of one read

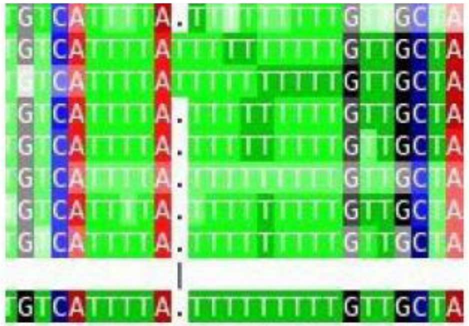
Single Nucleotide Variants (SNV) and small Indels

Variants detected within reads (smaller than size of read)

- SNVs:
  – Change of a single nucleotide



- Indels:
  – “Small” deletion or amplification

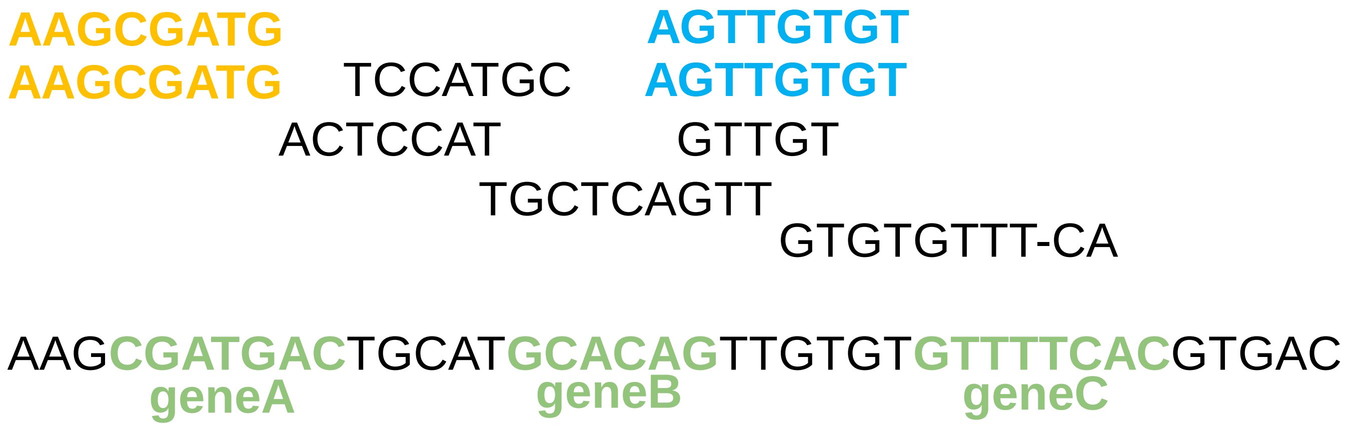


Single Nucleotide Variants (SNV) and small Indels

- Main factors affecting detection of SNVs and Indels
  - Number of reads (coverage supporting a variant)
  - Base quality (affects confidence in the SNVs)
  - PCR amplification bias (can generate duplicates and other biases)
  - Repetitive areas (mostly affects indels, but also affects SNVs)

Duplicated Reads

- Duplicate reads (same fragment) can appear
  - – In library preparation during amplification (eg. WES)
  - – In the amplification process while sequencing (optical duplicates)

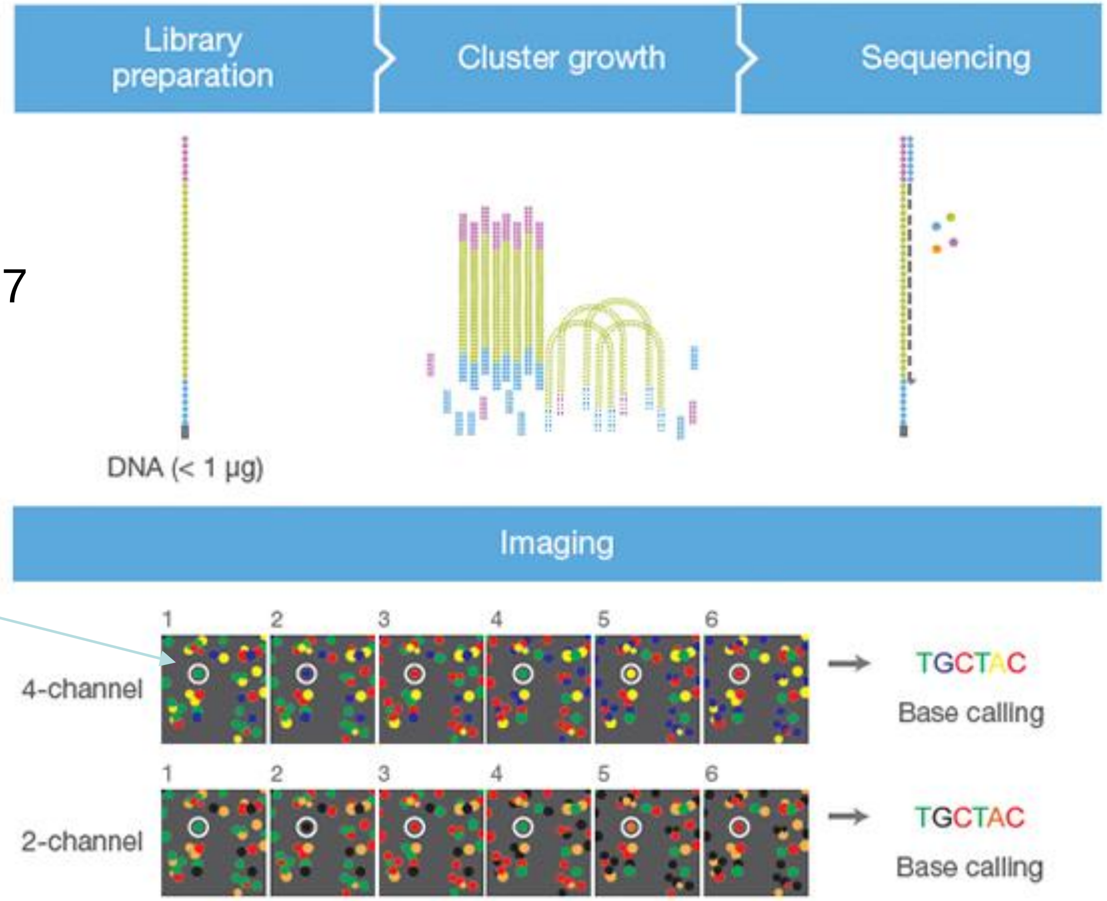


```
                   AGTTGTGT
AAGCGATG  TCCATGC  AGTTGTGT
        ACTCCAT     GTTGT
              TGCTCAGTT
                       GTGTGTTT-CA

AAGCGATGACTGCATGCACAGTTGTGTGTTTTCACGTGAC
    geneA       geneB       geneC
```

Duplicated Reads: optical duplicates

@MN00723:33:000H3MCVT:1:11102:7591:1087 1:N:0:47

Machine Flow Cell Lane Position



Optical duplicates are duplicate reads that are very close in the flow cell

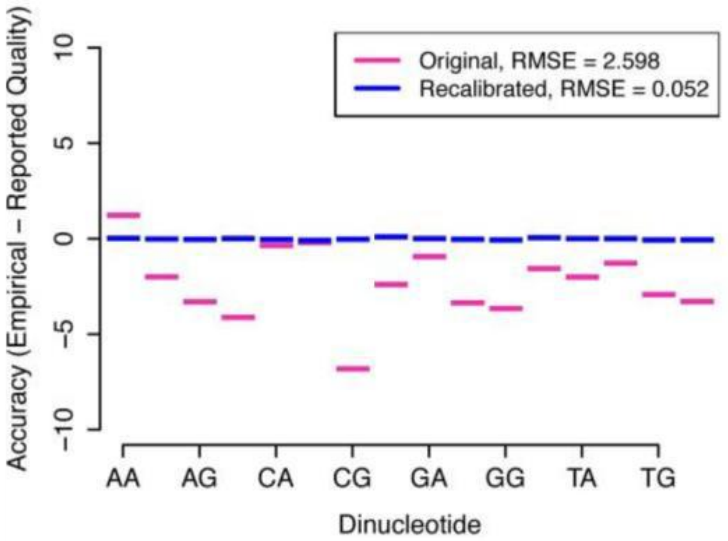
Duplicated Reads

- **The recommended practice is to ignore duplicates**
  - – Only consider one of the duplicates for variant calling
    - Usually the one with the best quality
  - – This may remove good information (eg. with high coverage, targeted)
  - – Duplicates are marked and later ignored (or not)
  - – Benefits of marking duplicates not always obvious
    - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4965708/
    - Eg. one can chose to only ignore reads marker as optical duplicates

Base Quality Recalibration

Base Quality Depends on several factors:

- Sample Quality (DNA)
- Nucleotide context
- Machine and cycle of sequencing
- Type of variant (SNP or Indel)

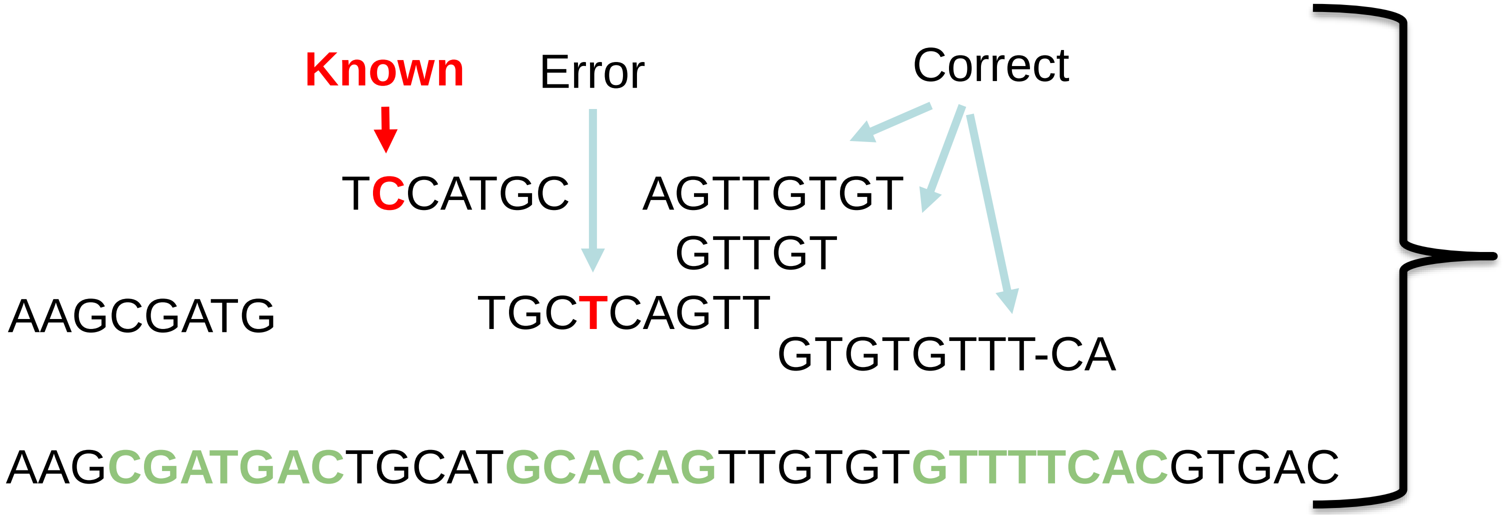


The machine may not estimate well the base quality score

Base Quality Recalibration

Use list of known variants to estimate correct quality values

- All bases different from reference **not in the provided list** of known variants are considered to be errors



Known

Error

Correct

TCCATGC

AGTTGTGT

GTTGT

AAGCGATG

TGCTCAGTT

GTGTGTTT-CA

AAGCGATGACTGCATGCACAGTTGTGTGTTTTCACGTGAC

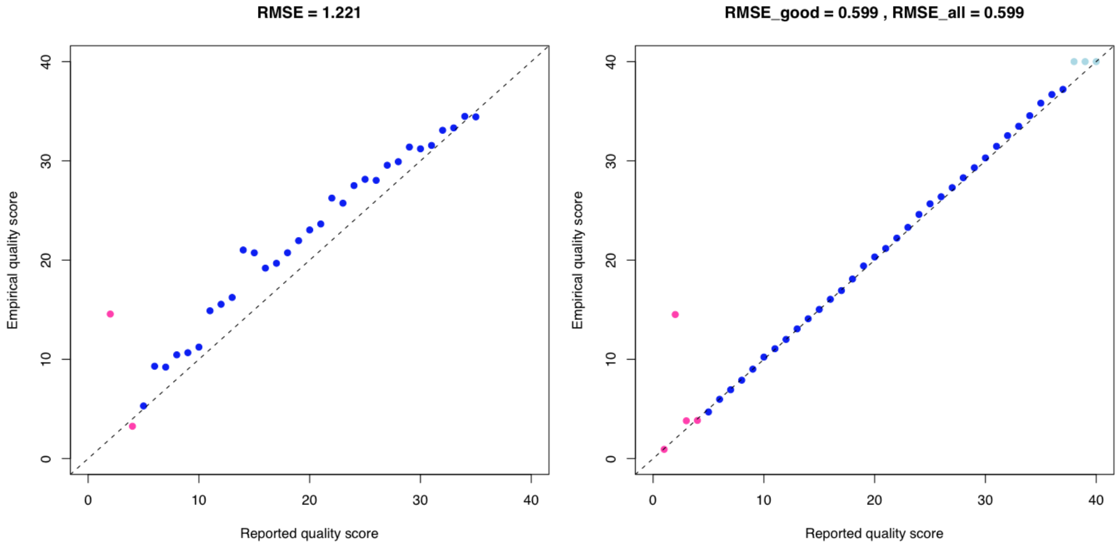
Count occurrences and compare Q value marked by the machine with the observed % errors

Base Quality Recalibration

Base Quality Recalibration:

The covariates being used here:

- ReadGroupCovariate
- QualityScoreCovariate
- ContextCovariate
- CycleCovariate



Original Data

After GATK Recalibration

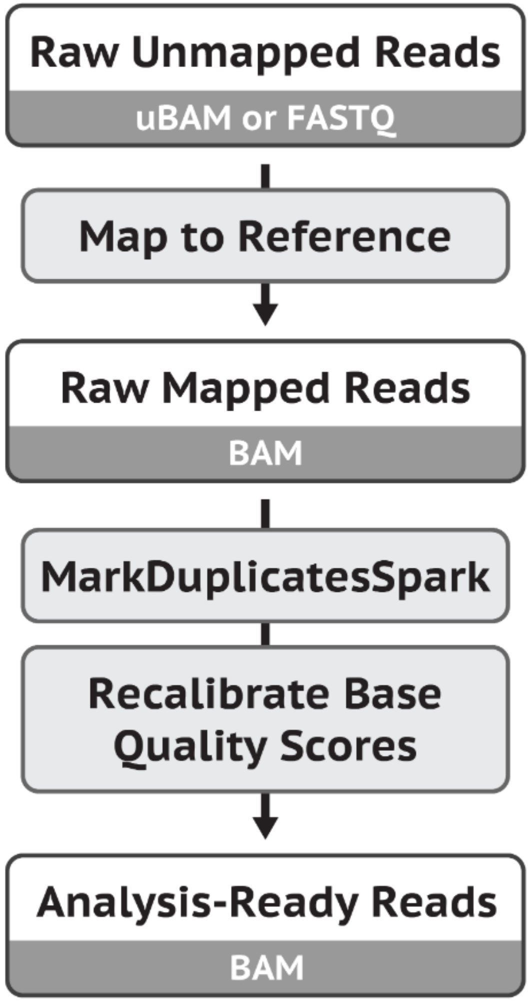
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3083463/

https://www.youtube.com/watch?v=L4D1dwES9s8

https://gatk.broadinstitute.org/hc/en-us/articles/360035890531-Base-Quality-Score-Recalibration-BQSR-

Data pre-processing for variant discovery

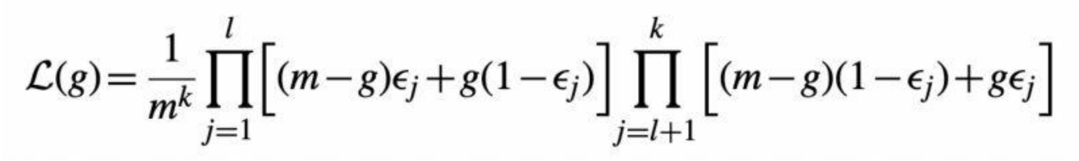
GATK
Best
Practices



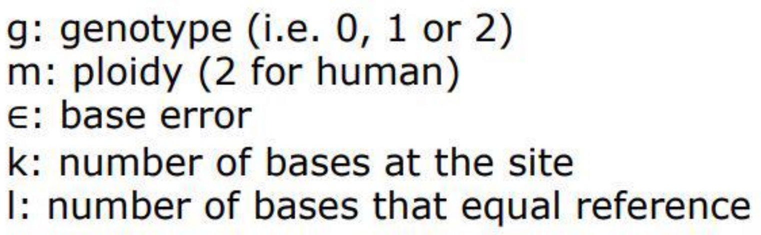
https://gatk.broadinstitute.org/hc/en-us/articles/360035535912-Data-pre-processing-for-variant-discovery

Example estimating Genotype Quality

- These calculations are software-dependent
  - Example Genotype Likelihood (GATK)



$$\mathcal{L}(g)=\frac{1}{m^k}\prod_{j=1}^{l}\left[(m-g)\epsilon_j+g(1-\epsilon_j)\right]\prod_{j=l+1}^{k}\left[(m-g)(1-\epsilon_j)+g\epsilon_j\right]$$

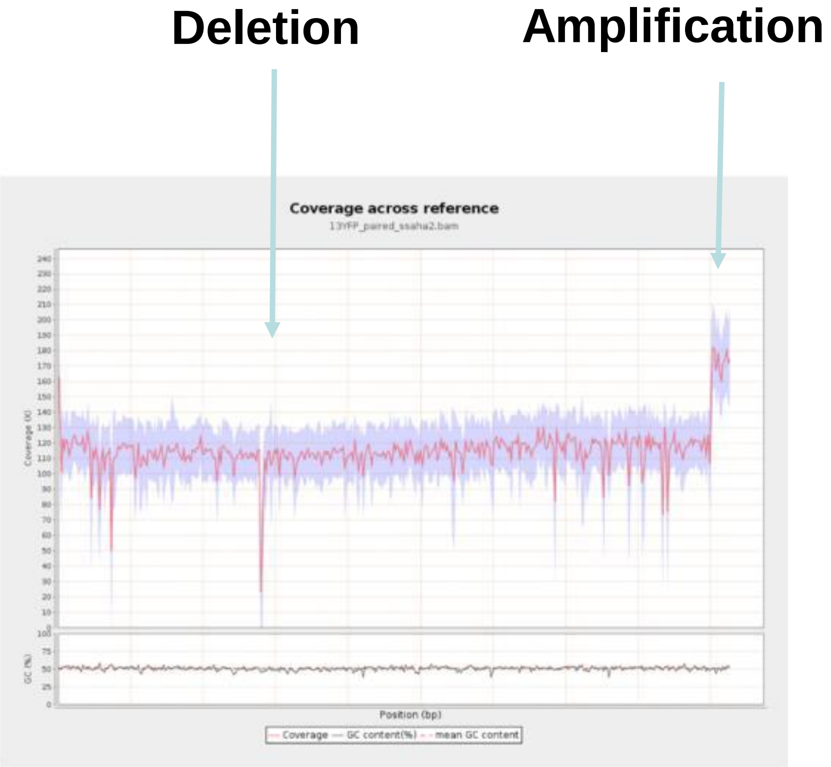


g: genotype (i.e. 0, 1 or 2)
m: ploidy (2 for human)
ϵ: base error
k: number of bases at the site
l: number of bases that equal reference

Large Structural Variants

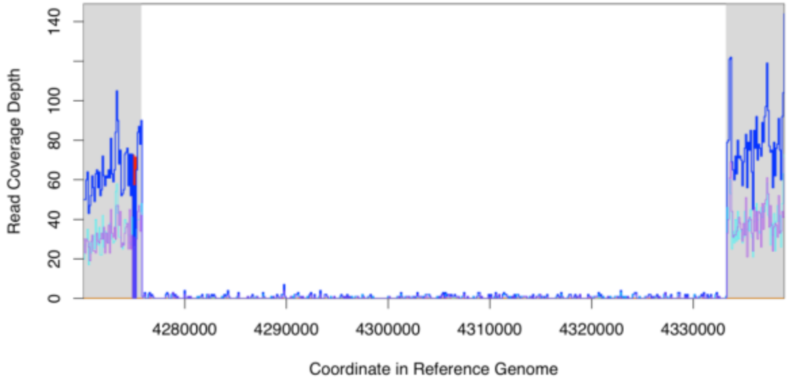
Variants larger than the size of reads

- Large Deletions and Amplifications
  - – Gene Deletions and Duplications
- Other Structural Variants
  - – Fusions; Inversions; Transposons...
- Horizontal Transfer
  - – Novel genomic regions



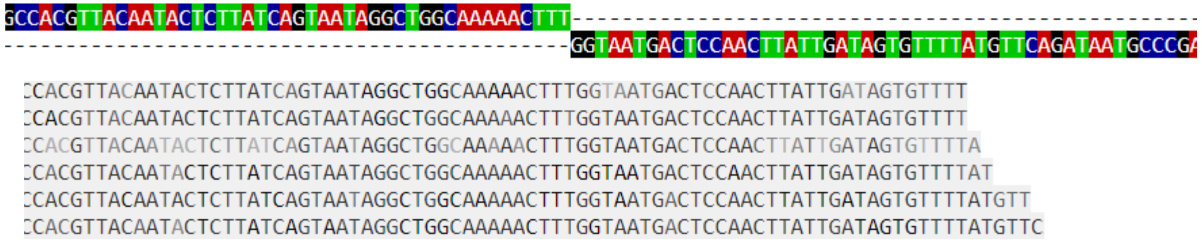
Large Structural Variants

- Evidence used to detect Structural Variants
  - – Differences in Coverage
    - Most commonly used
    - Particularly with targeted sequencing
    - Although there's still amplification bias



Large Structural Variants

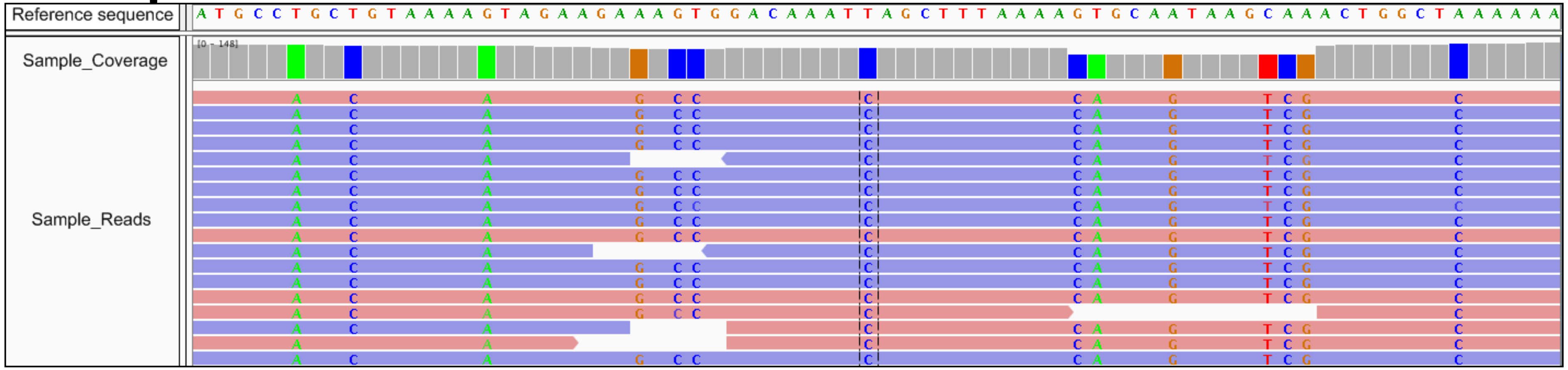
- Evidence used to detect Structural Variants
  - Junction evidence (difficult in targeted sequencing)
    - Can use paired read information (namely, expected fragment length –noisier)
    - Can use information within reads (more precise - requires bigger reads)



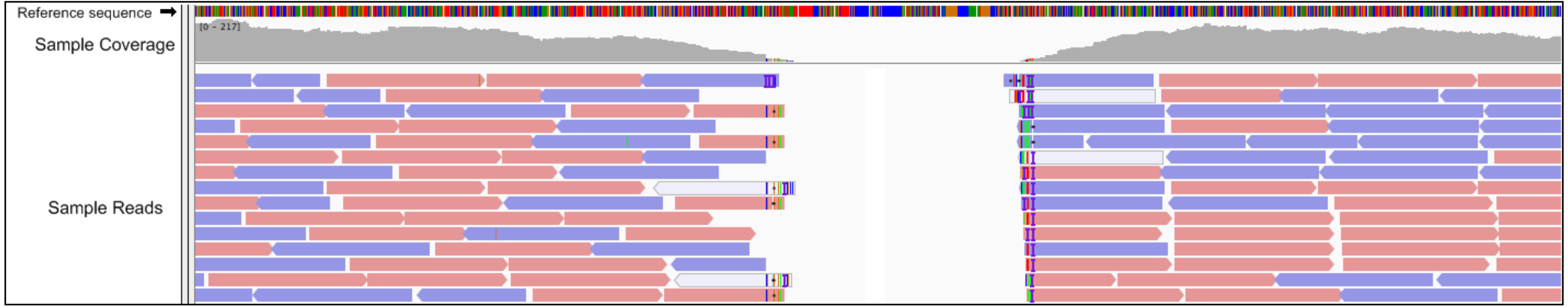
```
GCCACGTTACAATACTCTTATCAGTAATAGGCTGGCAAAAACTTT--------------------------------------------------
---------------------------------------------GGTAATGACTCCAACTTATTGATAGTGTTTTATGTTCAGATAATGCCCGA

 CCACGTTACAATACTCTTATCAGTAATAGGCTGGCAAAAACTTTGGTAATGACTCCAACTTATTGATAGTGTTTT
 CCACGTTACAATACTCTTATCAGTAATAGGCTGGCAAAAACTTTGGTAATGACTCCAACTTATTGATAGTGTTTT
 CCACGTTACAATACTCTTATCAGTAATAGGCTGGCAAAAACTTTGGTAATGACTCCAACTTATTGATAGTGTTTTA
 CCACGTTACAATACTCTTATCAGTAATAGGCTGGCAAAAACTTTGGTAATGACTCCAACTTATTGATAGTGTTTTAT
 CCACGTTACAATACTCTTATCAGTAATAGGCTGGCAAAAACTTTGGTAATGACTCCAACTTATTGATAGTGTTTTATGTT
 CCACGTTACAATACTCTTATCAGTAATAGGCTGGCAAAAACTTTGGTAATGACTCCAACTTATTGATAGTGTTTTATGTTC
```

Examples of variants

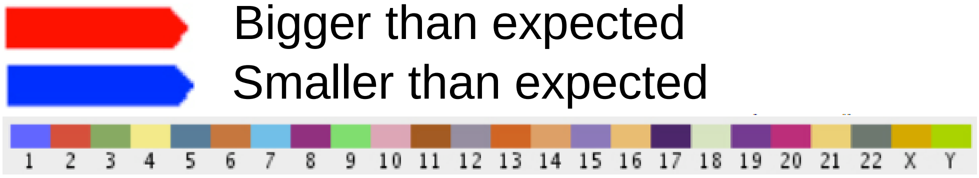
Example mutations


Reference sequence
A T G C C T G C T G T A A A A G T A G A A G A A A G T G G A C A A A T T A G C T T T A A A A G T G C A A T A A G C A A A C T G G C T A A A A A A
Sample_Coverage
[0 - 148]
Sample_Reads

Example deletion



- IGV provides colors to signal unusual situations
  - Besides mutations, information from paired-end is also there

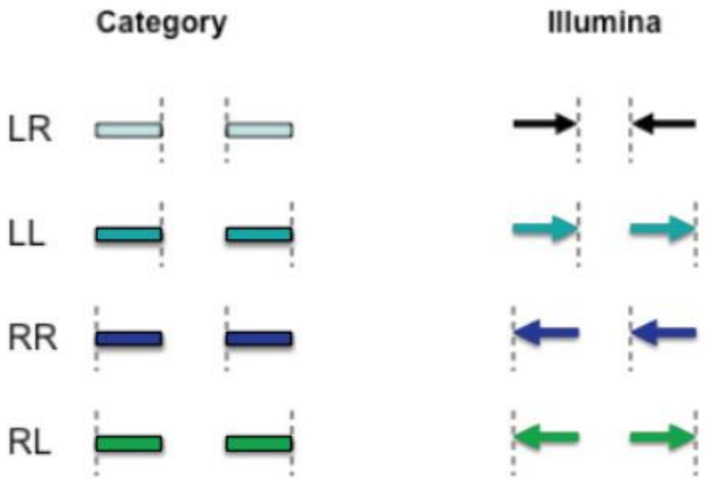


Bigger than expected

Smaller than expected

Pairs on different chromosomes

Insert Size Lengths



Normal

Unusual

Unusual

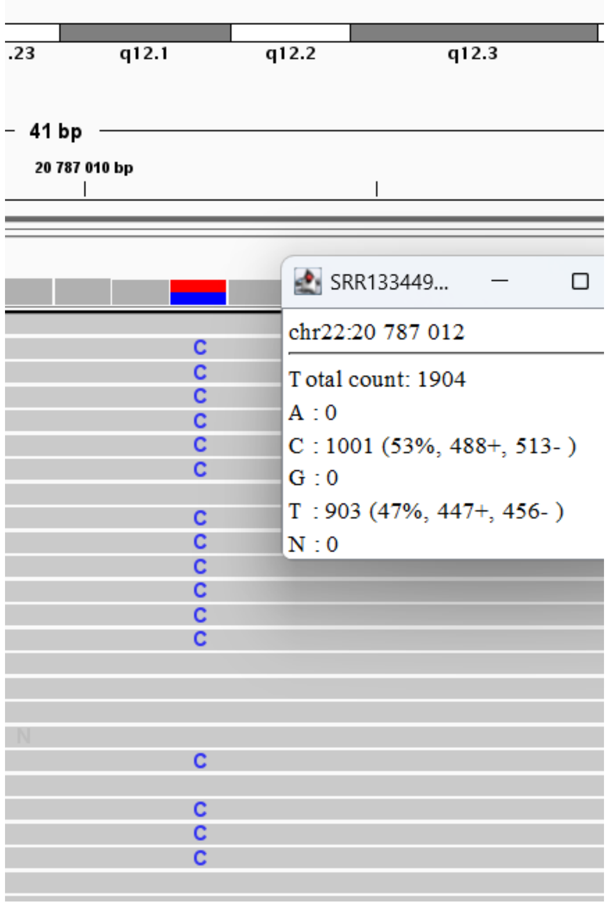
Unusual

Pair orientation

https://software.broadinstitute.org/software/igv/

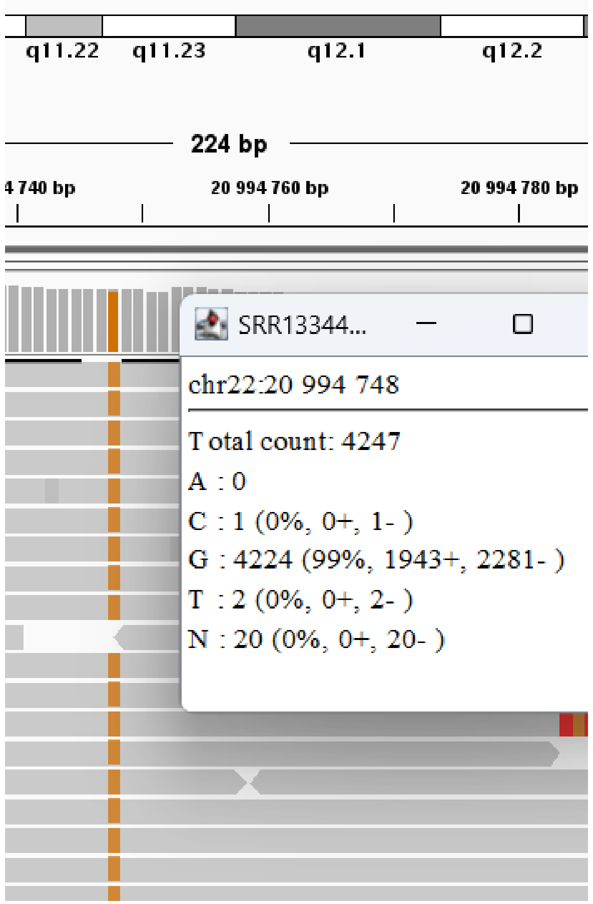
Visualization of Read Mappings

- Example of mutations



chr22:20 787 012

Total count: 1904
A : 0
C : 1001 (53%, 488+, 513-)
G : 0
T : 903 (47%, 447+, 456-)
N : 0



chr22:20 994 748

Total count: 4247
A : 0
C : 1 (0%, 0+, 1-)
G : 4224 (99%, 1943+, 2281-)
T : 2 (0%, 0+, 2-)
N : 20 (0%, 0+, 20-)

Variant Selection

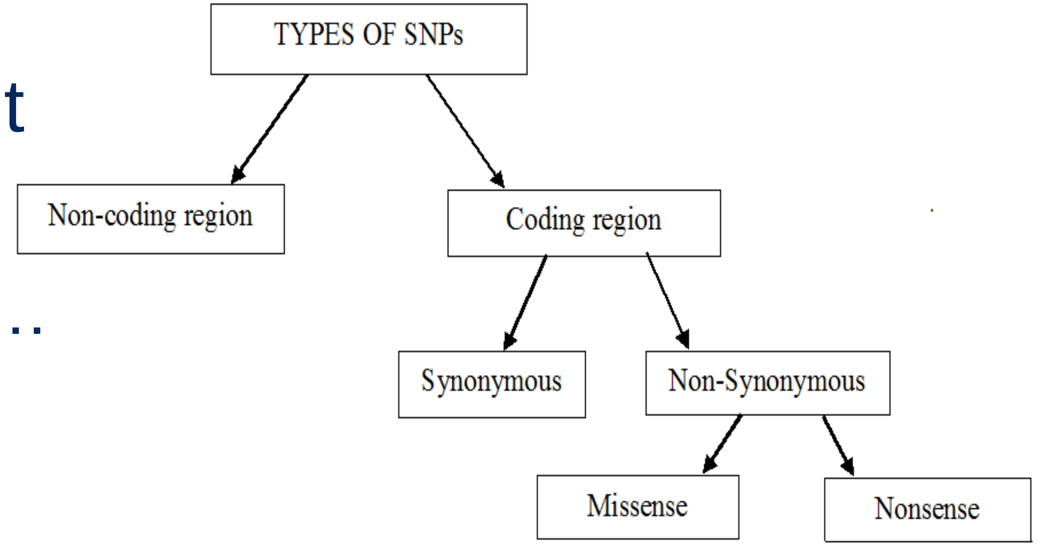
Criteria to select “valid” variants:

- Higher Number of reads (coverage supporting a variant)
- Low Bias in the Base quality supporting the variant
- Low bias in the strand of the reads supporting the variant
- Avoid variants only at the end of reads (repetitive areas)
- Avoid duplicate reads
- Etc…

Variant Annotation

Criteria to select “relevant” variants:

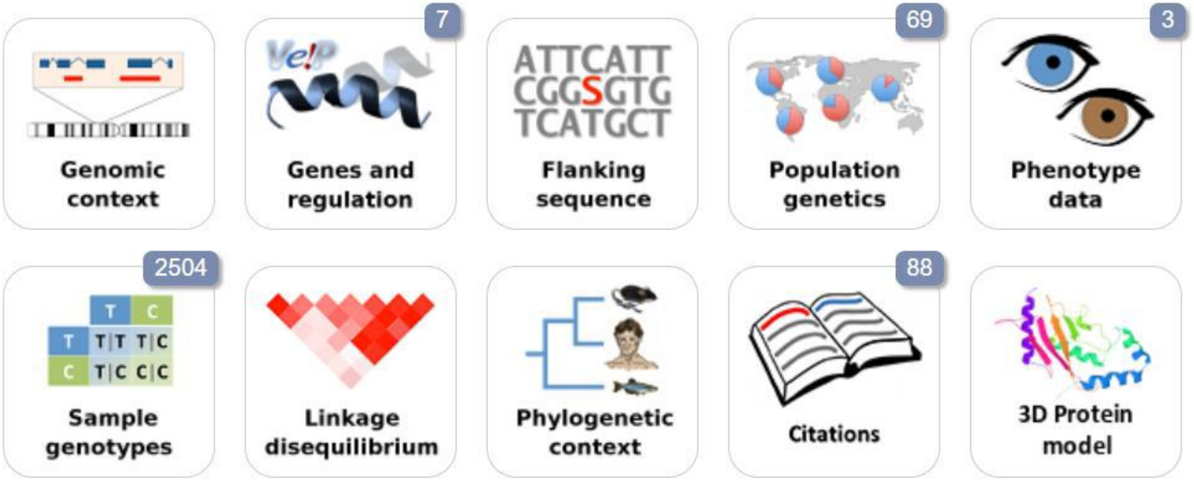
- Coding versus non-coding
    - Coding: Silent versus non-silent
    - Non-coding: can be complex
        - splice-sites, regulatory regions,…



Variant Annotation

Criteria to select “relevant” variants:

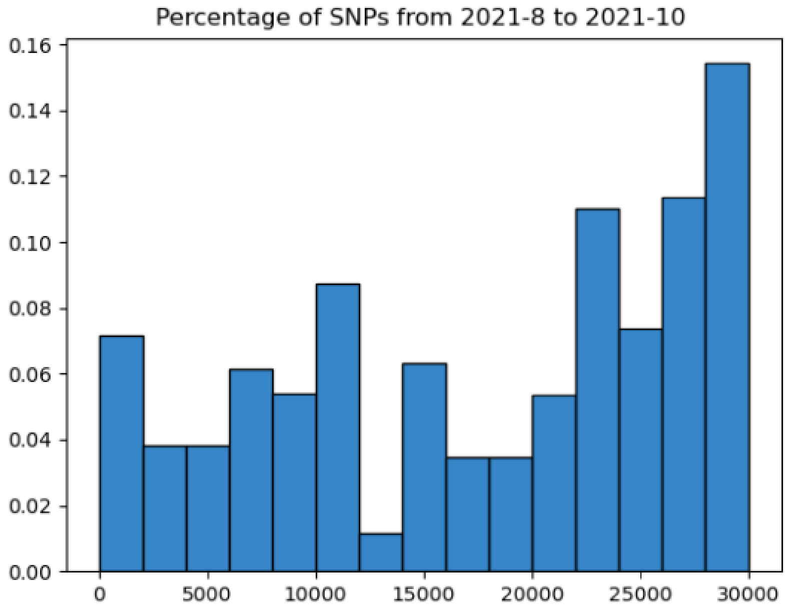
- Population frequency
- Disease-association

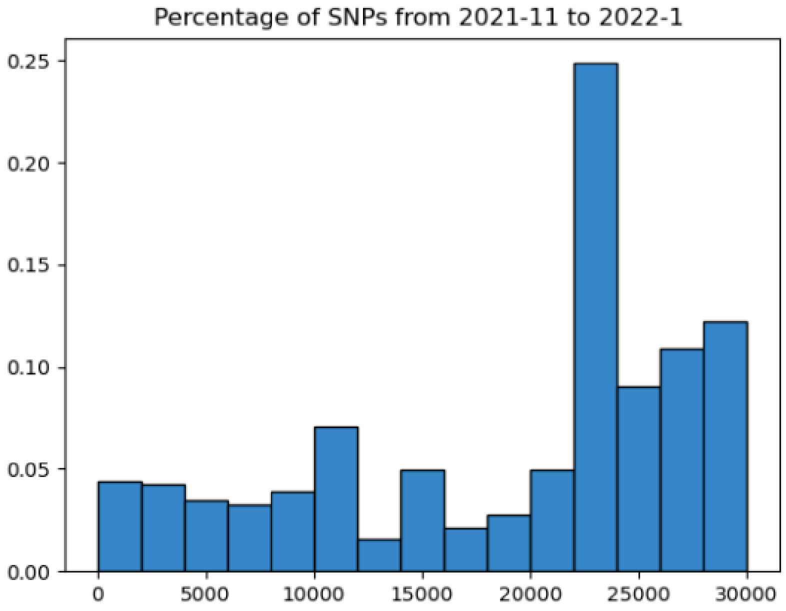


Only works for species with a lot of data, such as human

Variant Annotation

- Criteria to select “relevant” variants:
  - Simple exemple with SARS-CoV-2 (Omicron)





A few optional references:

- Single Nucleotide Variants and small Indels
  - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3083463/
- DeepVariant: Using AI to find variants
  - https://github.com/google/deepvariant; https://www.nature.com/articles/nbt.4235
- Copy Number Alterations and other Structural Variants
  - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4300727/
- Finding clonal vs subclonal variants
  - https://www.sciencedirect.com/science/article/pii/S2001037017300946
- Variant annotation
  - https://www.nature.com/articles/nprot.2015.105
  - https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0974-4

Practical Exercise

https://github.com/dsobral/MBCB/blob/main/README.md