NGS: Read Mapping to Reference Genome

Daniel Sobral / José Ferrão

Ministério da Saúde

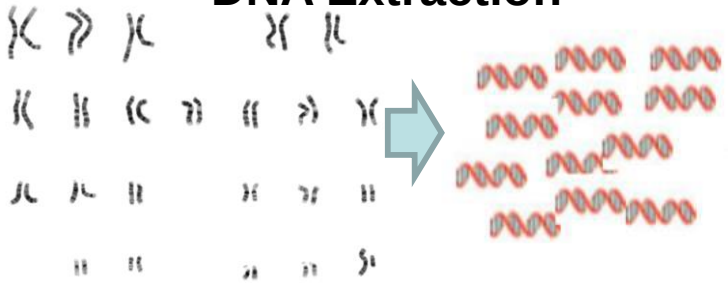
Instituto_Nacional de Saúde
Doutor Ricardo Jorge

Learning Objectives

- Overal understanding of the data preprocessing steps
  - Alignment of NGS reads (fastq) to a reference sequence
    - The SAM / BAM format
  - Marking Duplicates
  - Base Quality Recalibration
- Visualizing results and assessing quality

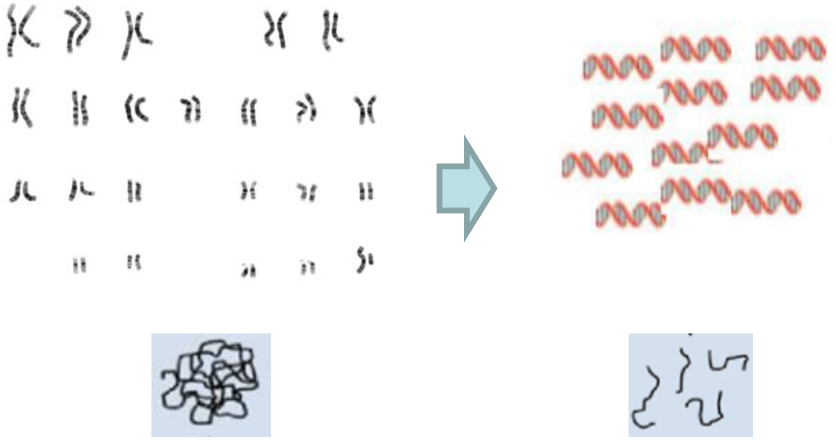
Data pre-processing for variant discovery

DNA Extraction

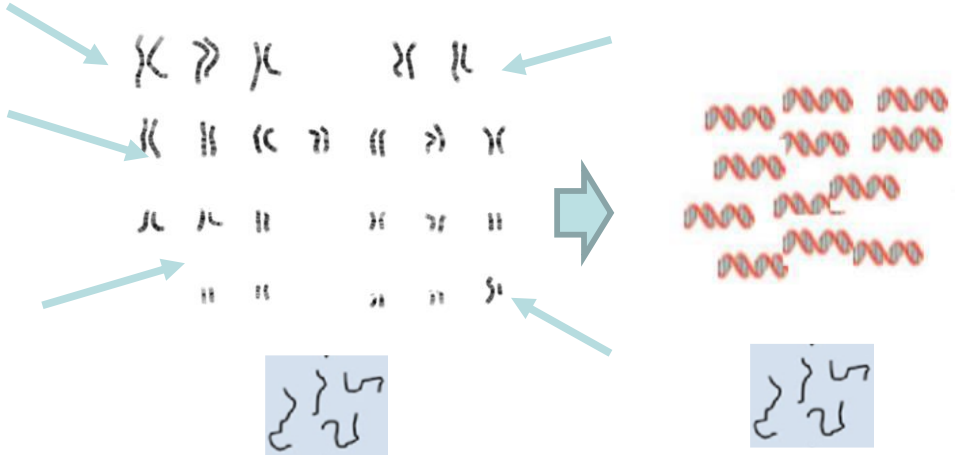


Data pre-processing for variant discovery

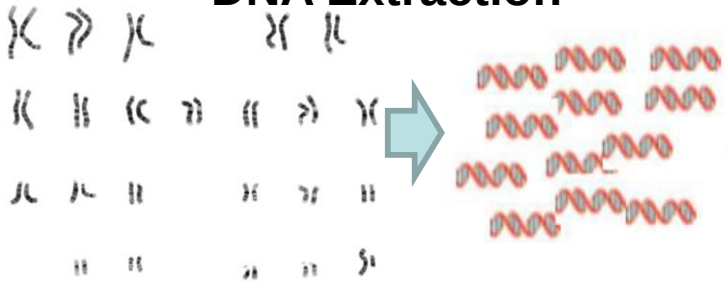
Whole Genome



Targeted (eg. WES)



DNA Extraction



Eg. TruSight One Enrichment Panel

How to calculate coverage

- WGS 30x coverage, 150bp read pairs
  - 30 x $3x10^9$ / (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

Raw Unmapped Reads

uBAM or FASTQ

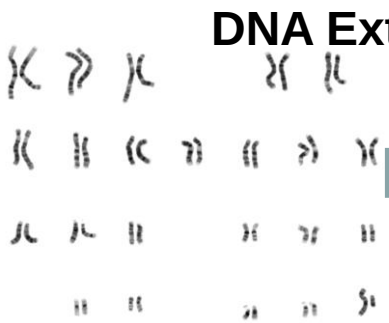
DNA Extraction **Sequencing**

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

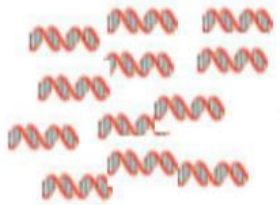
Data pre-processing for variant discovery

Raw Unmapped Reads

uBAM or FASTQ



DNA Extraction



Sequencing

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

QC and pre-processing

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

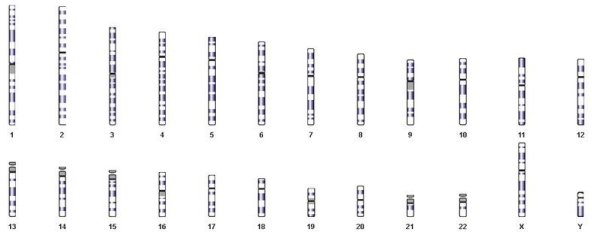
Data pre-processing for variant discovery

Raw Unmapped Reads

uBAM or FASTQ

(fasta file)

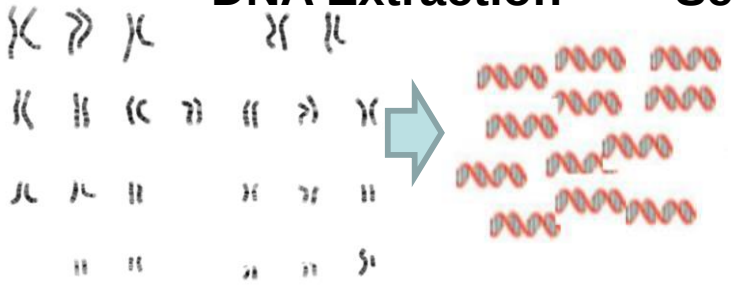
Reference Genome



geneA geneB geneC

AAGCGATGACTGCATGCACAGTTGTGTGTTTTCACGTGAC

DNA Extraction **Sequencing**



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

QC and pre-processing

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

What genome reference to use?

- GRCh37 (hg19)
  - The “classical”, for which most resources were built (1000 genomes, etc…)
- **GRCh38 (hg38)**
  - **The recommended one. Most resources have been converted to this now**
- T2T-CHM13
  - Most complete genome, but will take time for resources to be available

Future: pangenome graphs

A lot of investment is going here, but several challenges remain

Mostly relevant for Structural Variants

What genome reference to use?

Things to consider:

- Mask PAR regions in Y chromosome
- Include decoys (eg. EBV sequences)
- Include alternative sequences such as HLA loci
    - Though these may need to be analized specifically

https://gatk.broadinstitute.org/hc/en-us/articles/360035890951-Human-genome-reference-builds-GRCh38-or-hg38-b37-hg19

https://gatk.broadinstitute.org/hc/en-us/articles/360035890951

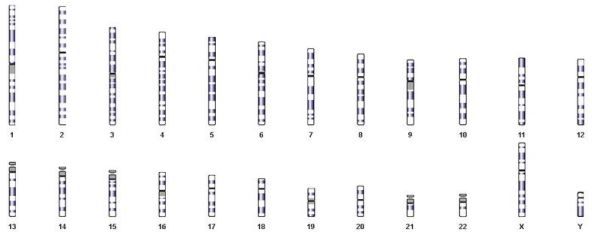
Data pre-processing for variant discovery

Raw Unmapped Reads

uBAM or FASTQ

(fasta file)

Reference Genome

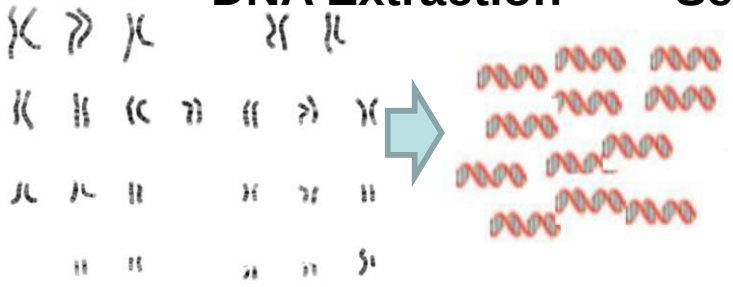


TGCTCAGTT

geneA geneB geneC

AAGCGATGACTGCATGCACAGTTGTGTGTTTTCACGTGAC

DNA Extraction **Sequencing**



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

QC and pre-processing

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

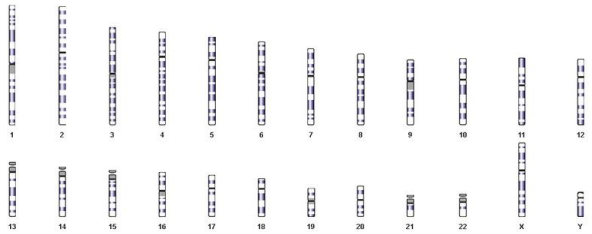
Data pre-processing for variant discovery

Raw Unmapped Reads

uBAM or FASTQ

(fasta file)

Reference Genome

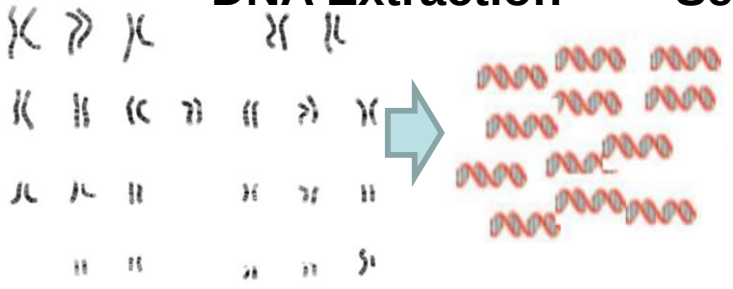


TGCTCAGTT

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)**

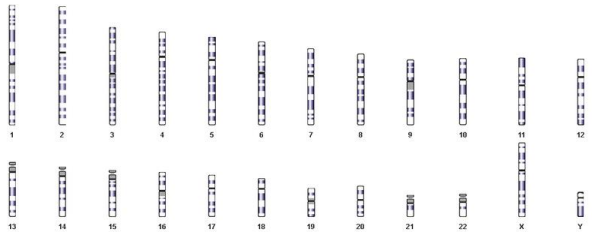
Data pre-processing for variant discovery

Raw Unmapped Reads

uBAM or FASTQ

(fasta file)

Reference Genome



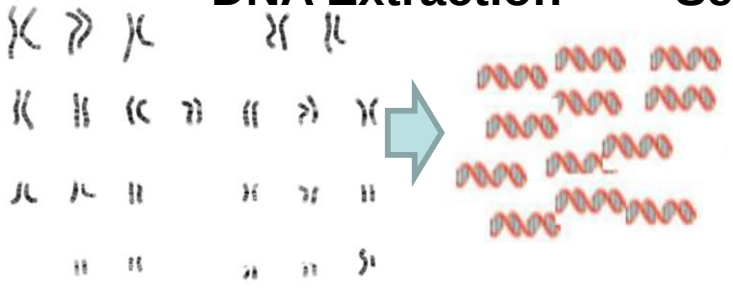
ACTCCAT GTTGT

TGCTCAGTT

geneA geneB geneC

AAGCGATGACTGCATGCACAGTTGTGTGTTTTCACGTGAC

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



TGCTCAGTT TGAC ATGGAGT GTTGT

(ACTCCAT)

QC and pre-processing

TGCTCAGTTA TGAC ATGGAGT TTT CGTTGT

(“clean” fastQ)

(“raw” fastQ)

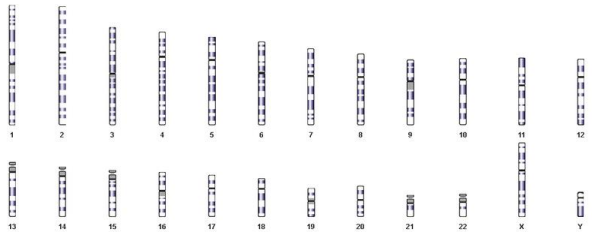
Data pre-processing for variant discovery

Raw Unmapped Reads

uBAM or FASTQ

(fasta file)

Reference Genome



TCCATGC AGTTGTGT

ACTCCAT GTTGT

AAGCGATG TGCTCAGTT

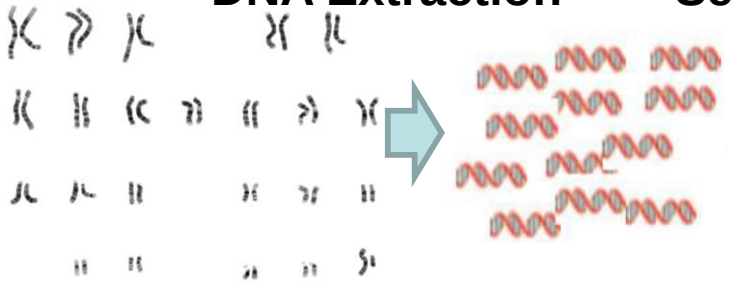
GTGTGTTT-CA

(SAM/BAM)

geneA geneB geneC

AAGCGATGACTGCATGCACAGTTGTGTGTTTTCACGTGAC

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



TGCTCAGTT TGAC ATGGAGT GTTGT

(ACTCCAT)

QC and pre-processing

(“clean” fastQ)

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
```

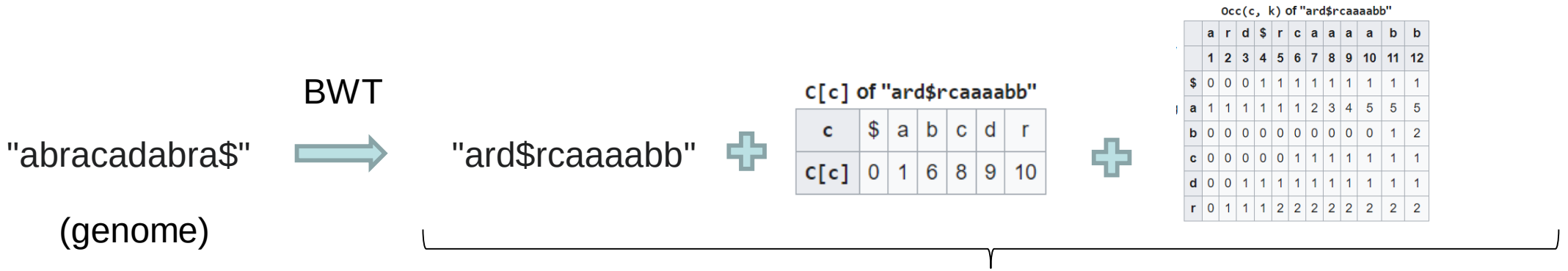
https://samtools.github.io/hts-specs/SAMv1.pdf

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



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

	a	r	d	$	r	c	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

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

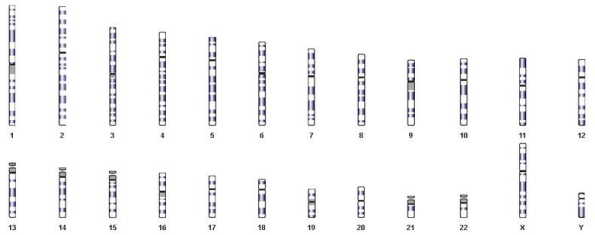
How alignment is made in practice

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

$ bwa index genome.fasta

(genome.fasta file)

Reference Genome



$3x10^9$ bases

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

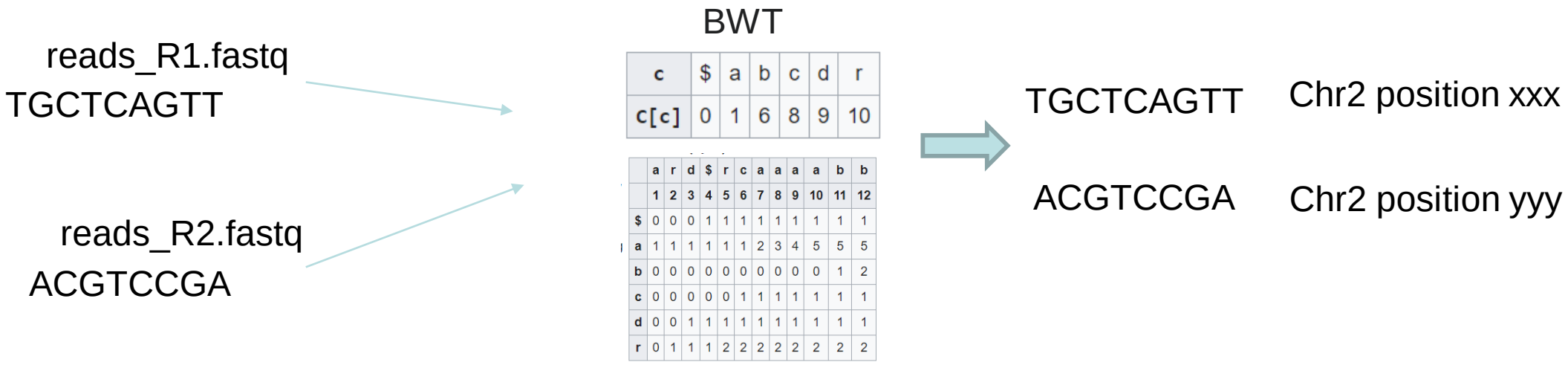
FM - index

How alignment is made in practice

Command to generate alignments with BWA:

$ bwa mem genome.fasta reads_R1.fastq(.gz) reads_R2.fastq(.gz)

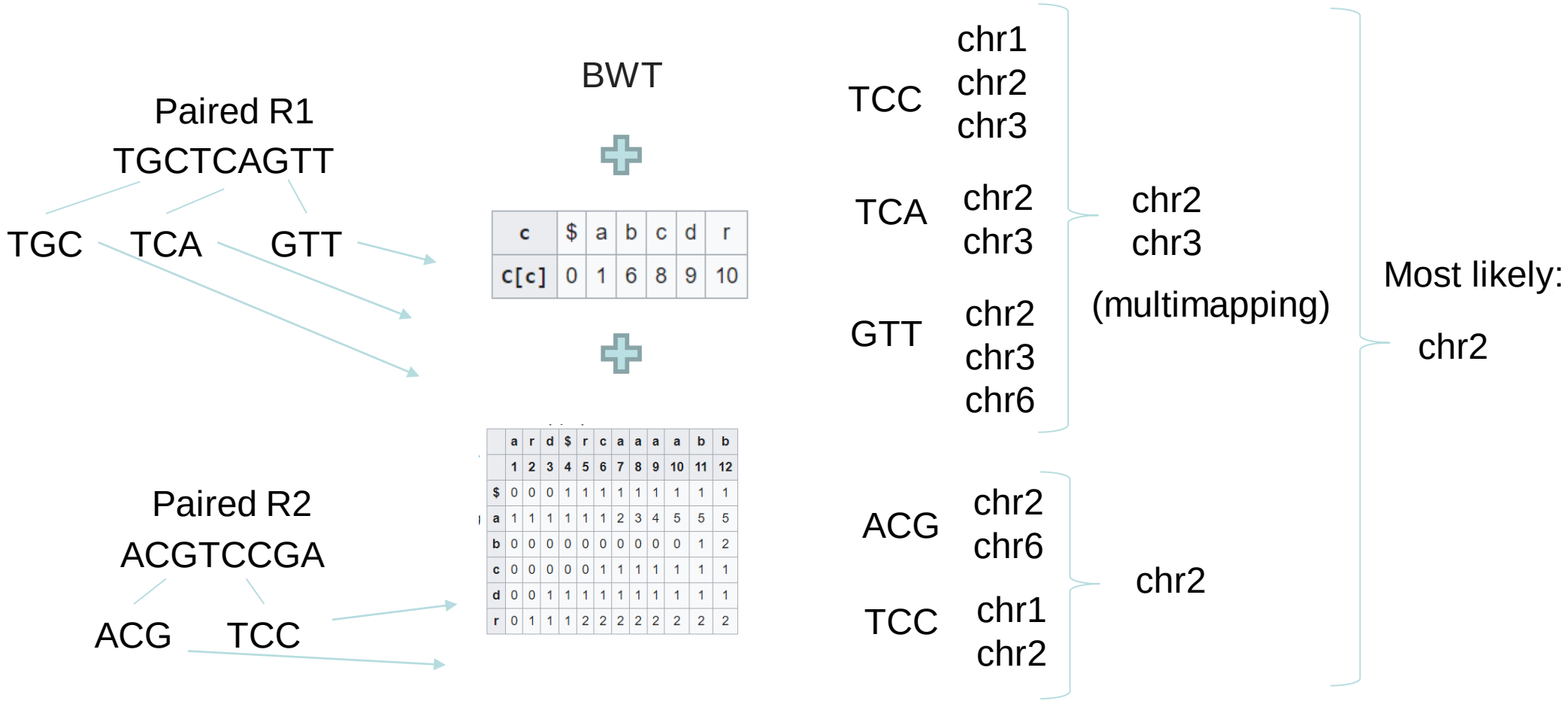
(BWT FM-index for genome.fasta needs to be created before)



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

How alignment is made in practice



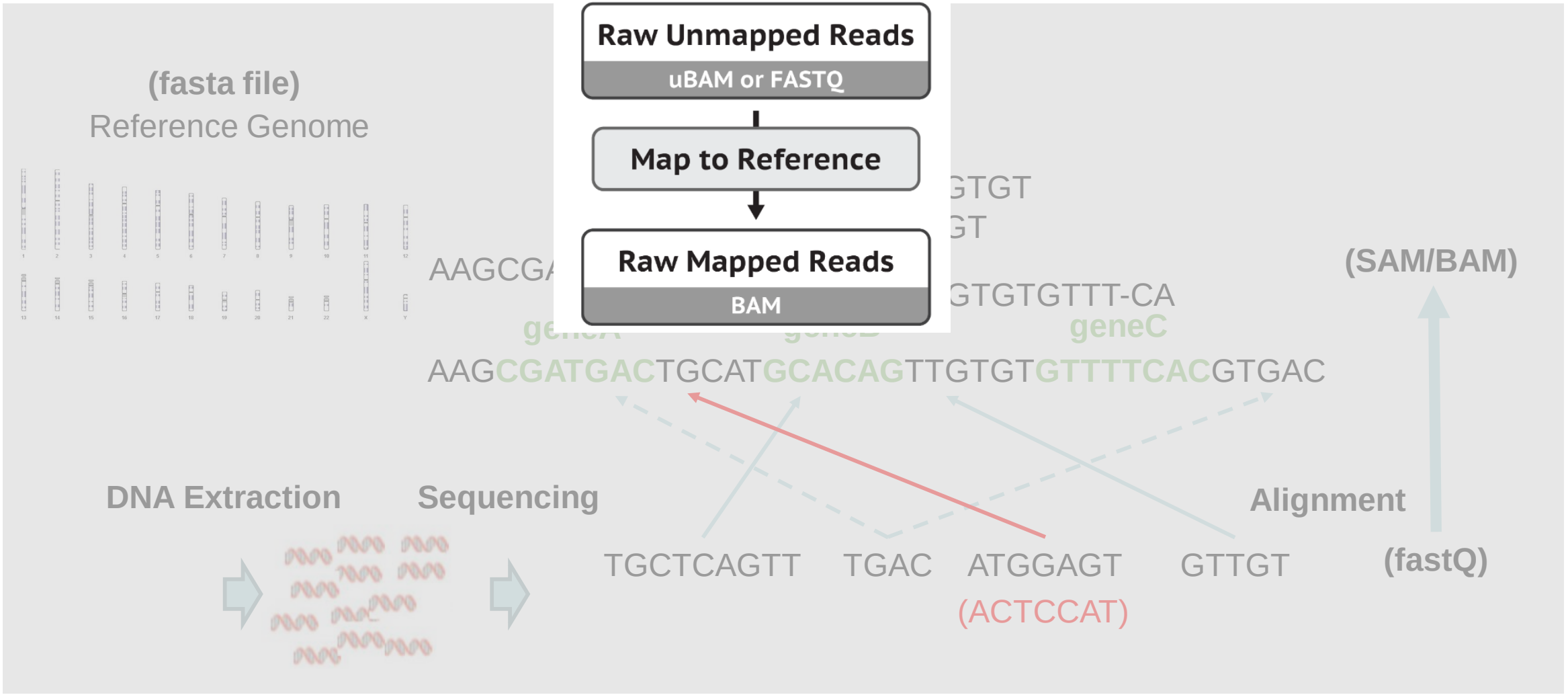
c	$	a	b	c	d	r
c[c]	0	1	6	8	9	10

	a	r	d	$	r	c	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

How alignment is made in practice

- Alignments are fast but may not be always perfect
  - Nonetheless it works very well, particularly for human
- Equally good mappings (multimappings) are marked as such
  - Mapping quality is set as 0 by default
- Alignments are made piece-wise
  - A read can be split in a principal and secondary alignment(s)
    - Eg. splicing in RNA-Seq; large insertions
  - Sometimes, only a part of the read is aligned (the rest is masked)
    - Particularly in repetitive areas this can lead to false alignments

Data pre-processing for variant discovery



Practical Example

GATK (Broad Institute) provides resources for their best practices protocol:
https://console.cloud.google.com/storage/browser/genomics-public-data/resources/broad/hg38/v0/
(already includes pre-built BWT FM-index)

> bwa index ficheiro_genoma.fasta

> bwa mem -t 1 -R '@RG\tID:sampleid\tPU:FlowCell.Lane.Sample\tSM:samplename\tPL:ILLUMINA\tLB:samplename\tCN:UTI' -M ficheiro_genoma.fasta amostrareads_R1.fastq.gz amostrareads_R2.fastq.gz > amostra.sam

ID: Identifier; PU: Platform Code (FlowCell.Lane.Sample); SM: Sample; PN: LB: Library; CN: Center Name (Sequencing Team) ; other fields possible (see SAM specification)

Exercise

- Each group should have one sample (1,2,3)
  - How many reads do you have in your sample?
  - Align selected reads against a specific chromosome
    - What is the identifier of the first aligned read? Where does it align?
    - How many entries are in the SAM file? Is it the number of reads?
    - Can you see in the header information about the program being run
    - Can you see the read group information you added?
  - Sort the aligned reads
    - What is the identifier of the first aligned read? Where does it align?
  - Convert sam to a bam and index the bam
    - Compare the size of the files. What other files were generated?

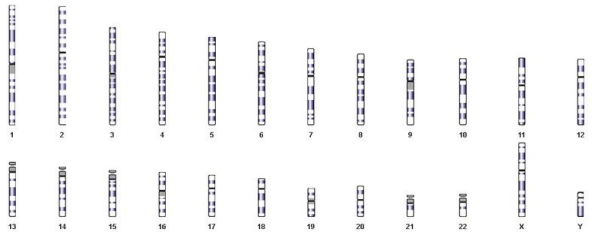
Data pre-processing for variant discovery

Raw Unmapped Reads

uBAM or FASTQ

(fasta file)

Reference Genome



TCCATGC AGTTGTGT

ACTCCAT GTTGT

AAGCGATG TGCTCAGTT

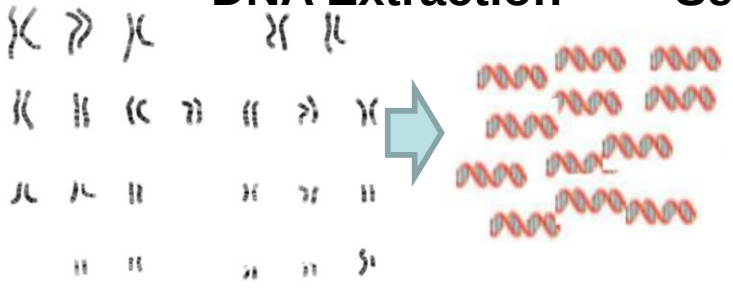
GTGTGTTT-CA

(SAM/BAM)

geneA geneB geneC

AAGCGATGACTGCATGCACAGTTGTGTGTTTTCACGTGAC

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



TGCTCAGTT TGAC ATGGAGT GTTGT

(ACTCCAT)

QC and pre-processing

(“clean” fastQ)

TGCTCAGTTA TGAC ATGGAGT TTT CGTTGT

(“raw” fastQ)

Data pre-processing for variant discovery
Duplicated Reads

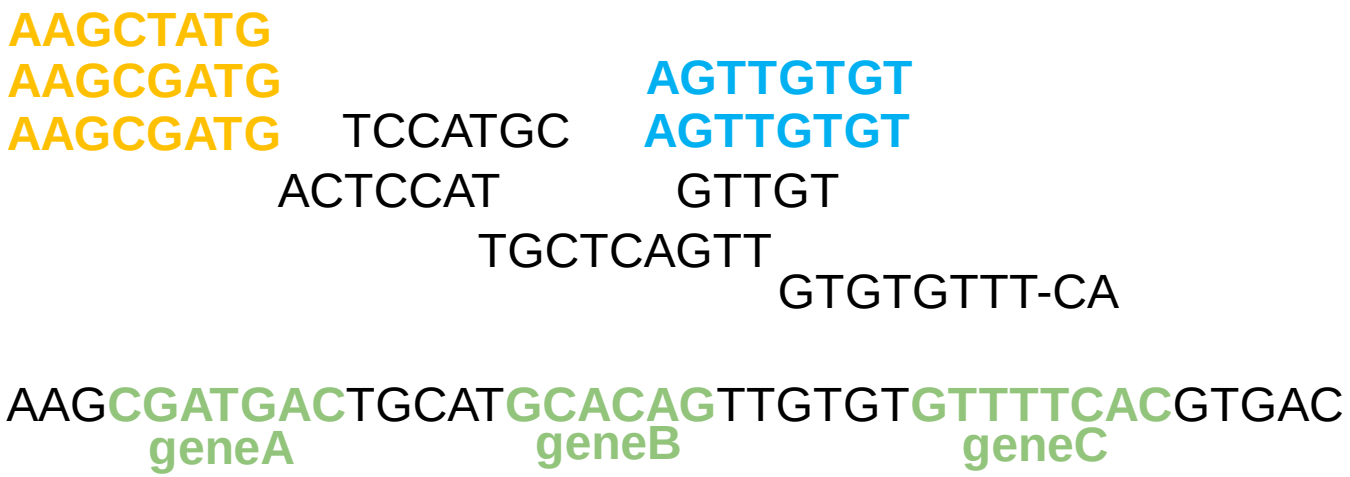
- Duplicate reads (same fragment) can appear
  - In library preparation during amplification (eg. WES)
  - In the amplification process while sequencing (optical duplicates)

Data pre-processing for variant discovery
Duplicated Reads

Read pairs need to be taken in consideration

Data pre-processing for variant discovery
Duplicated Reads

Duplicates are identified by position (alignment and/or flow-cell), not by sequence

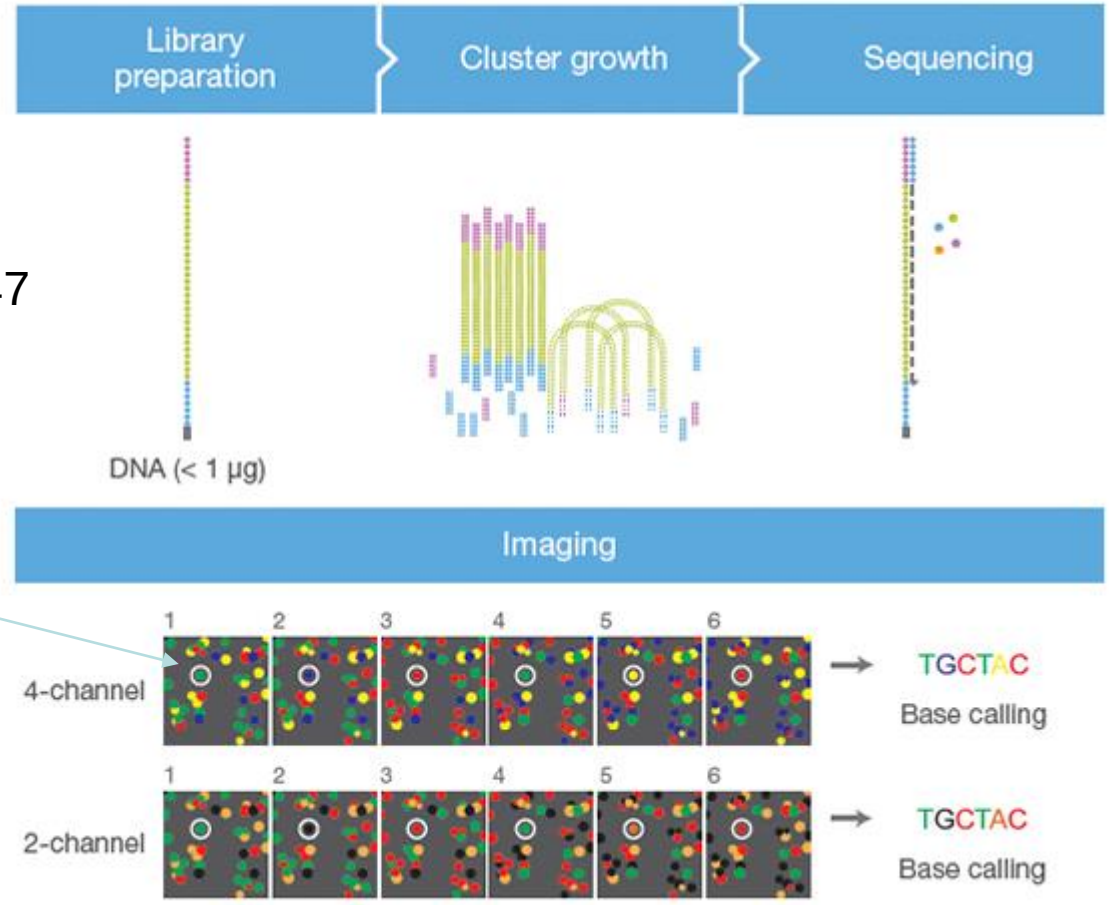


Data pre-processing for variant discovery
Duplicated Reads: optical duplicates

@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



Data pre-processing for variant discovery 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

Practical Exercise

> gatk MarkDuplicates -I=sample.bam -O=sample_marked.bam -M=sample_marked_metrics.txt

Alternative: samblaster

> samblaster sample.bam | samtools sort -o sample.marked.sorted.bam

Data pre-processing for variant discovery Duplicated Reads

Metrics

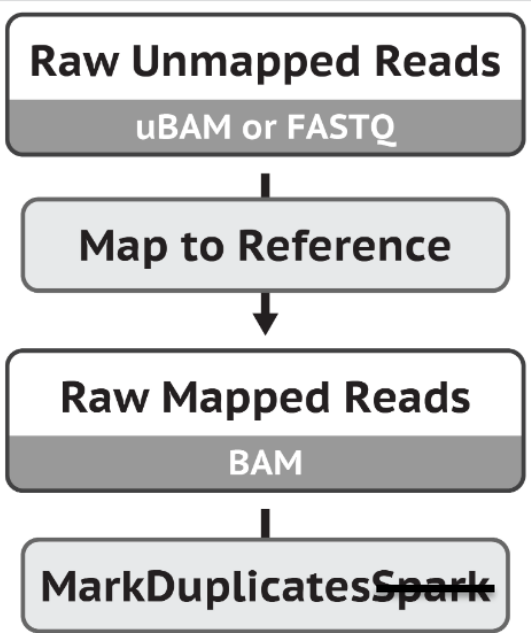
Field	Description
LIBRARY	The library on which the duplicate marking was performed.
UNPAIRED_READS_EXAMINED	The number of mapped reads examined which did not have a mapped mate pair, either because the read is unpaired, or the read is paired to an unmapped mate.
READ_PAIRS_EXAMINED	The number of mapped read pairs examined. (Primary, non-supplemental)
SECONDARY_OR_SUPPLEMENTARY_RDS	The number of reads that were either secondary or supplementary
UNMAPPED_READS	The total number of unmapped reads examined. (Primary, non-supplemental)
UNPAIRED_READ_DUPLICATES	The number of fragments that were marked as duplicates.
READ_PAIR_DUPLICATES	The number of read pairs that were marked as duplicates.
READ_PAIR_OPTICAL_DUPLICATES	The number of read pairs duplicates that were caused by optical duplication. Value is always < READ_PAIR_DUPLICATES, which counts all duplicates regardless of source.
PERCENT_DUPLICATION	The fraction of mapped sequence that is marked as duplicate.
ESTIMATED_LIBRARY_SIZE	The estimated number of unique molecules in the library based on PE duplication.

Exercise

- Run MarkDuplicates for your sample
  - There should be a metrics file generated and a new bam

    (we will discuss the content of the metrics file later)

Data pre-processing for variant discovery



(BAM)

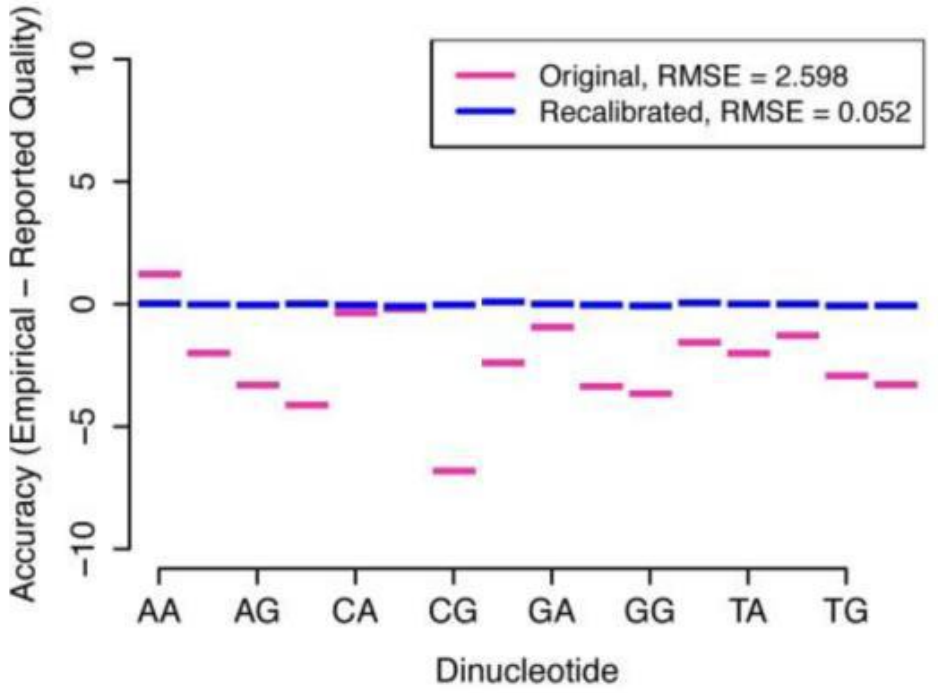
Now we have alignments where each read may have an extra annotation if it is recognized as being duplicate.

Data pre-processing for variant discovery Base Quality Recalibration

Base Quality not very precise

Depends on several factors:

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

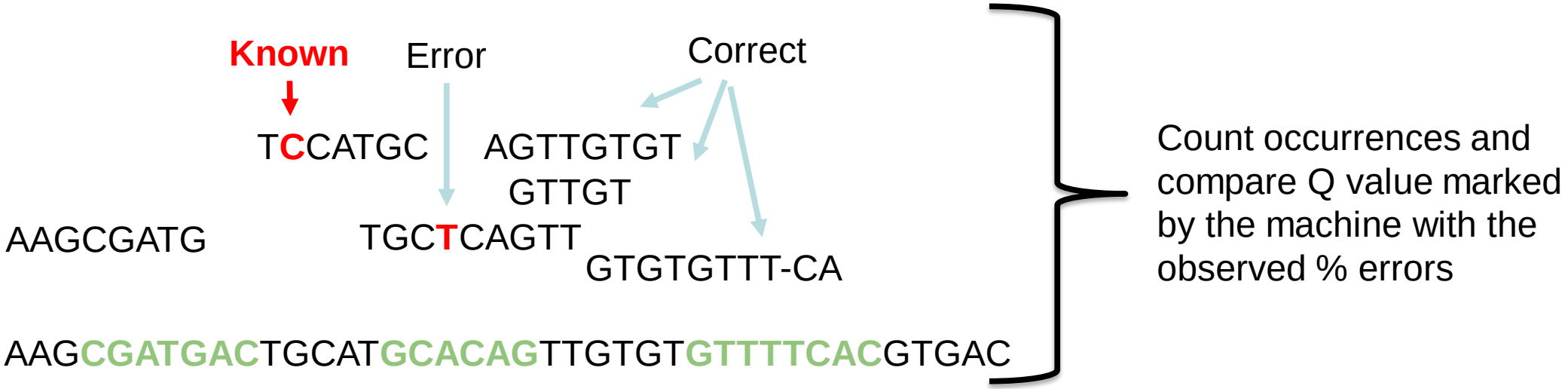


The machine oftens does not estimate well the base quality score

Data pre-processing for variant discovery 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

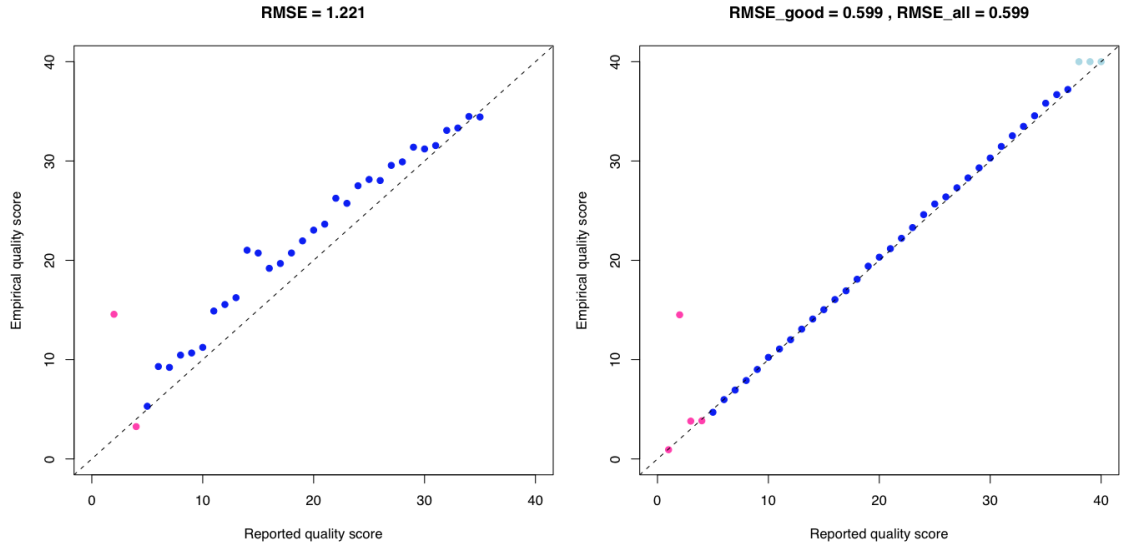


Data pre-processing for variant discovery Base Quality Recalibration

Base Quality Recalibration:

The covariates being used here:

- ReadGroupCovariate
- QualityScoreCovariate
- ContextCovariate
- CycleCovariate



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-

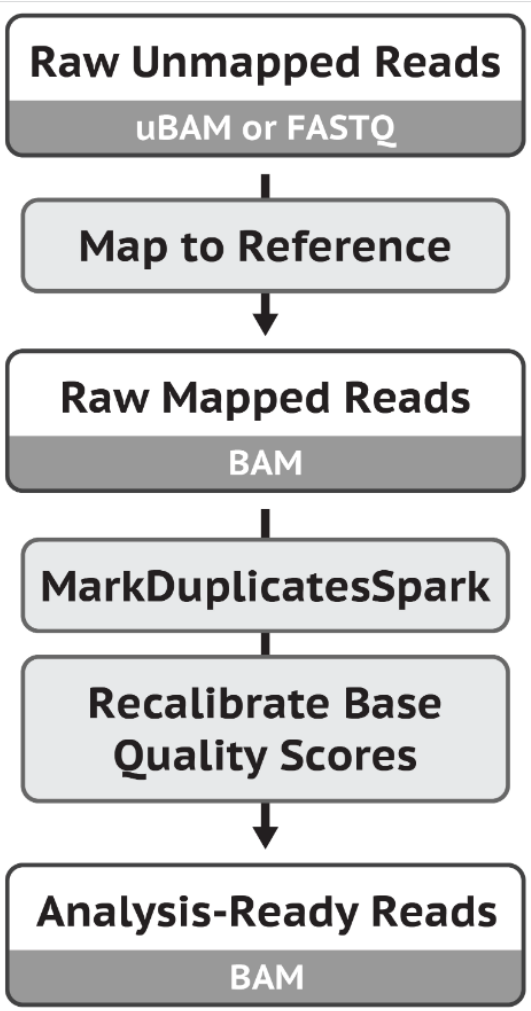
Data pre-processing for variant discovery
Base Quality Recalibration

> gatk BaseRecalibrator -I sample_marked.bam -R genome.fasta --known-sites known_snps.vcf --known-sites known_indels.vcf [--intervals target_positions.bed --interval-padding 100] -O sample_marked_baserecalibrator_report.txt

>gatk ApplyBQSR -I sample_marked.bam -R genome.fasta --bqsr-recal-file sample_marked_baserecalibrator_report.txt –O sample_marked_baserecalibrator.bam

Data pre-processing for variant discovery

GATK Best Practices



These steps can be used with other variant callers

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

NGS: Quality Assessment and Visualization of Read Mappings

Ministério da Saúde

Instituto_Nacional de Saúde
Doutor Ricardo Jorge

Quality Assessment of Read Mappings

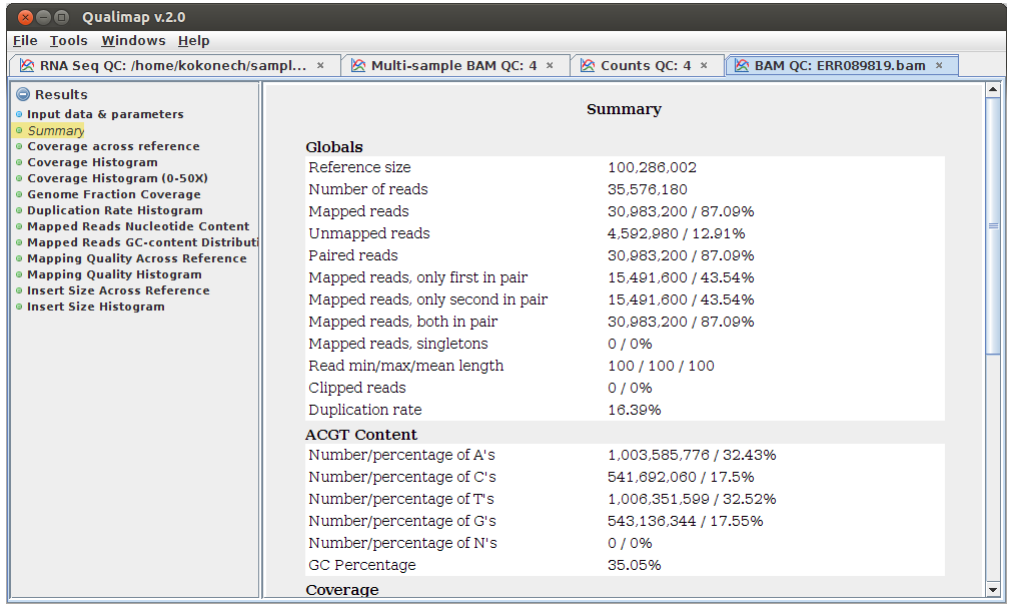
Measures to consider:

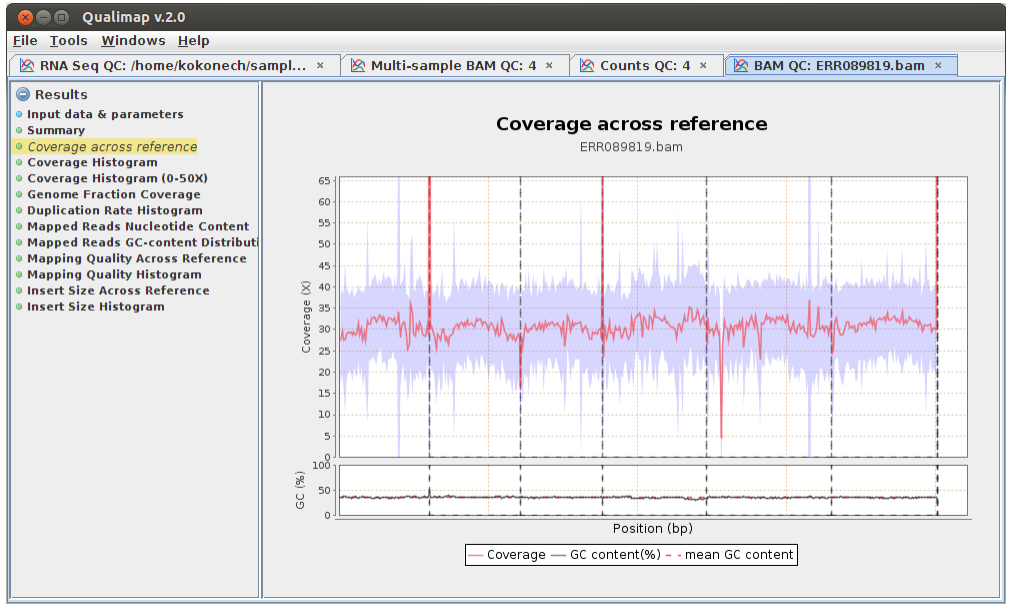
- Mapped Reads / Unmapped Reads
  - Usually >80% for WGS; >90% for WES
- Total Mapped Reads / Coverage
  - Presence of Duplicates (depends on coverage)
- Homogeneity in the coverage
- Distribution of Fragment Size (paired-end)
- Coverage of Targeted Regions (in case of eg. WES)

Quality Assessment of Read Mappings

- Qualimap

http://qualimap.conesalab.org/





Practical example

- Look at Metrics from GATK

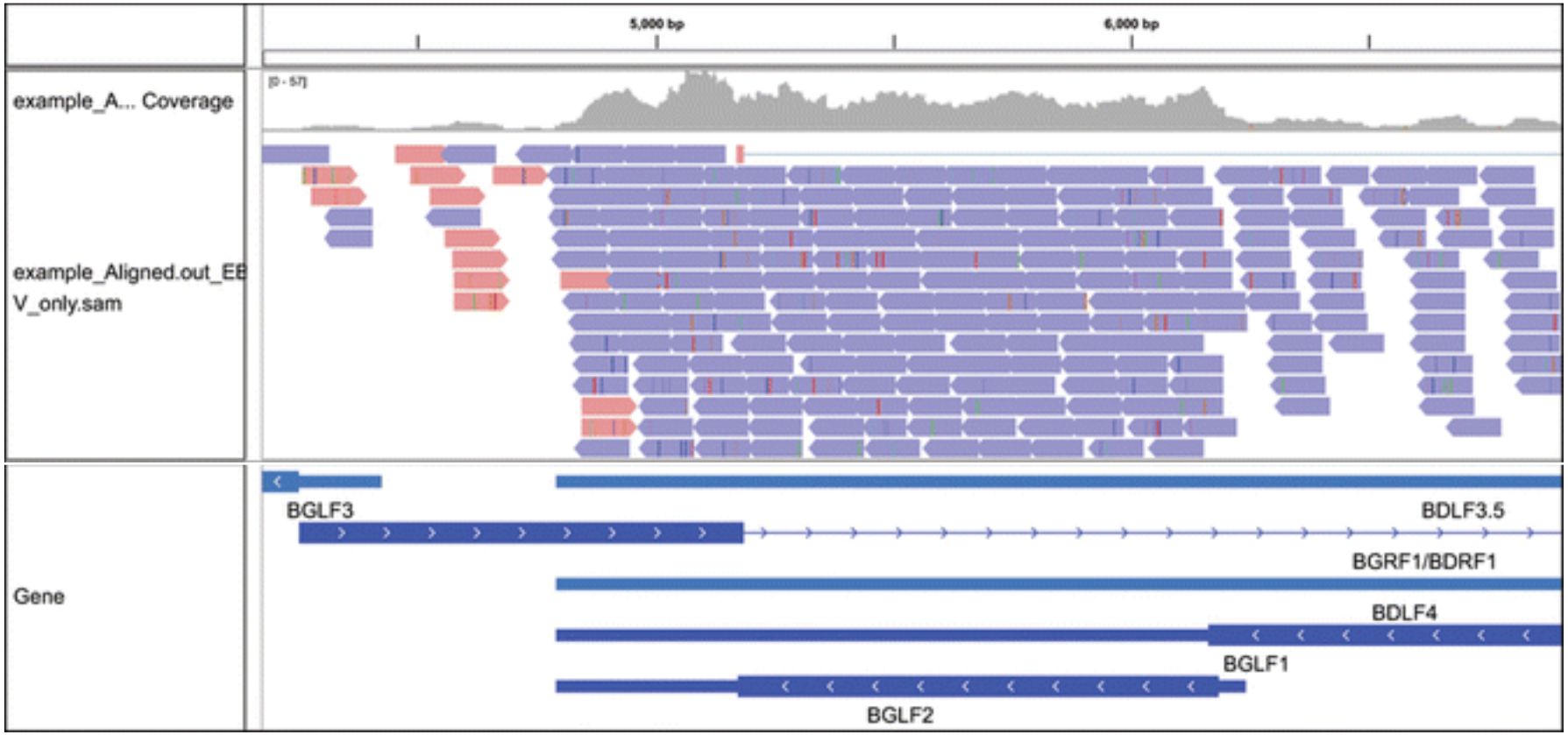
>qualimap bamqc –bam sample_marked_recalibrated.bam –outdir . –outfile sample_marked_recalibrated_qualimap.pdf –outformat PDF [-gff targets_position.bed]

Exercise

- Look at Metrics from GATK for your sample
- Run qualimap bamqc for your sample and see output

Visualization of Read Mappings

Integrative Genomics Viewer



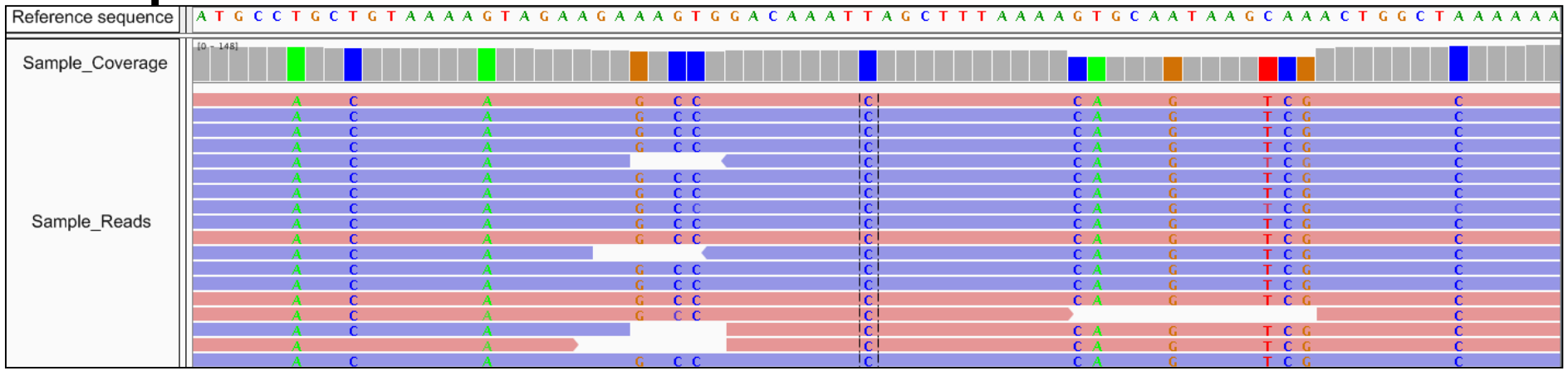
Coverage

Alignments

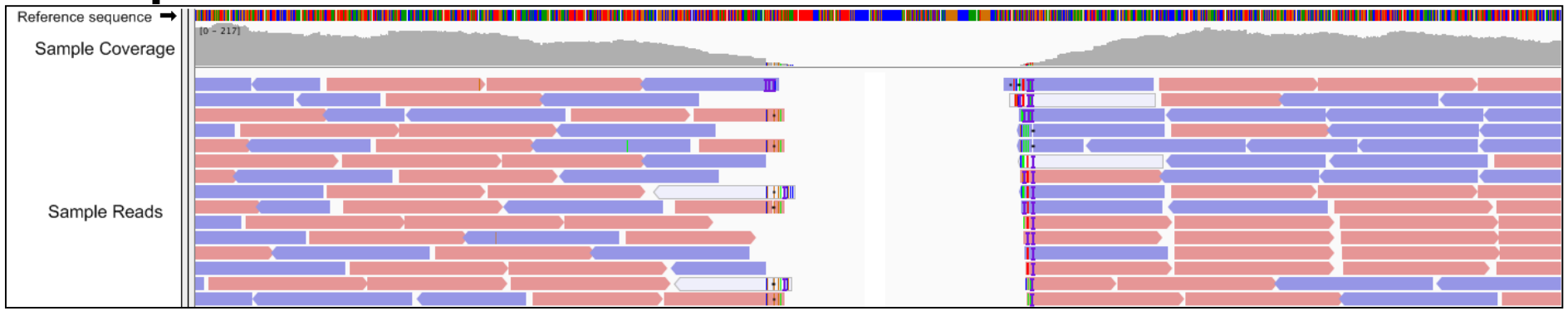
Annotation (genes)

Visualization of Read Mappings

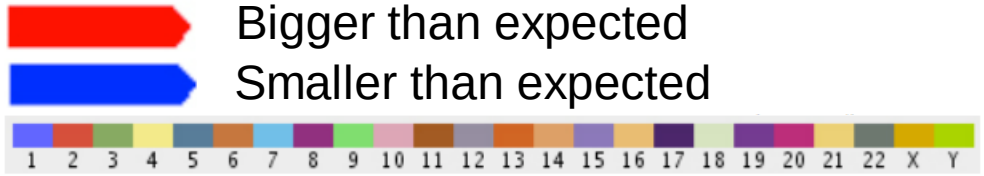
Example mutations



Example deletion

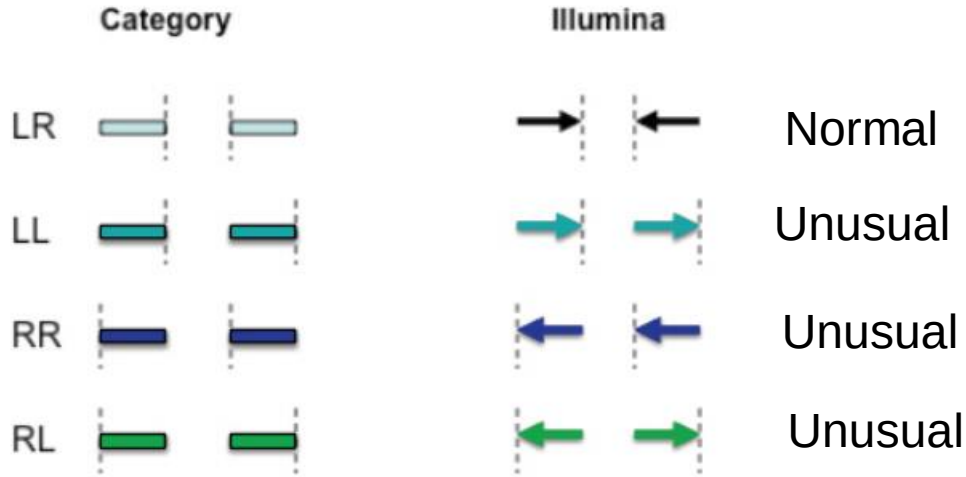


- IGV provides colors to signal unusual situations
  - Besides mutations, information from paired-end is also there



Pairs on different chromosomes

Insert Size Lengths

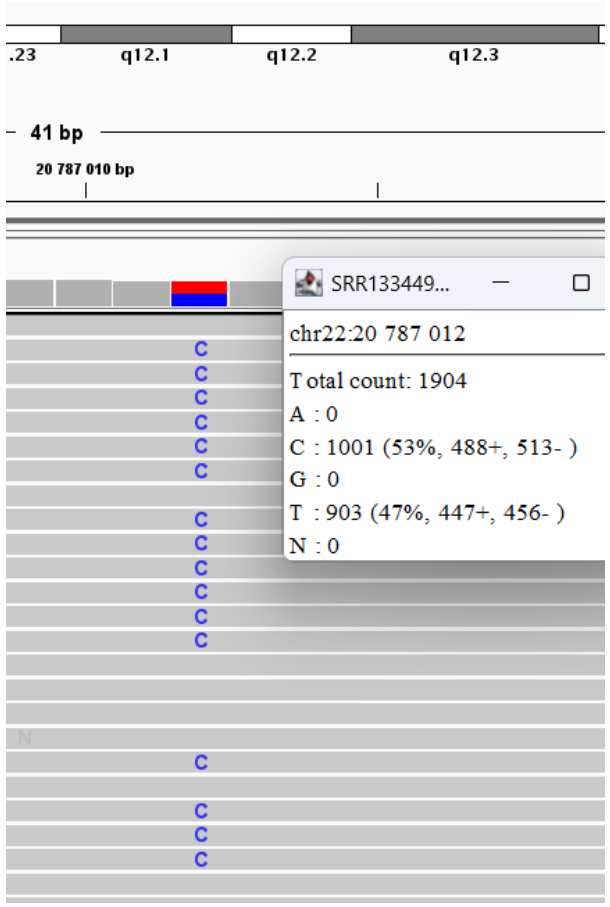


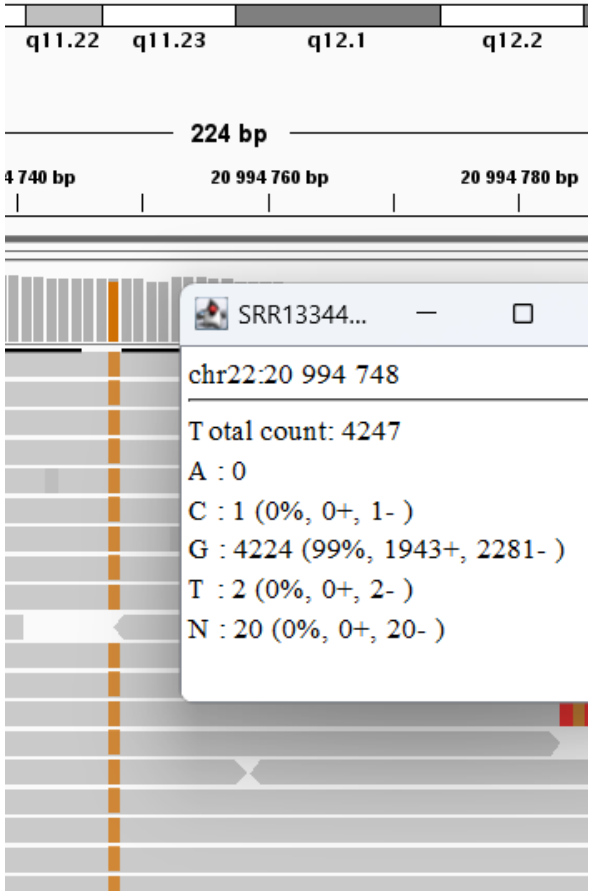
Pair orientation

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

Visualization of Read Mappings

- Example of mutations, showing evidence





Practical example

- Show alignments sample_example
  - Position 9:21968200

Gather evidence for a mutation

- Number of reads, strand of reads, quality etc…

- Exercise:
  - Sample 1: 2:171868791
  - Sample 2: 6:121216049
  - Sample 3: X:129746638

Gather evidence for a mutation

- Number of reads, strand of reads, quality etc…