Session 10 –Theory and Exercises

Read Alignment II: BWA mem

ESCI upf.
School of International Studies
BelongBecome

Date: 14/02/2024, 15:00-19:30
Teacher: **Fernando Cruz** (CNAG)
fernando.cruz@prof.esci.upf.edu

Bachelor's Degree in Bioinformatics
Course 2023-2024
52115 - Algorithms for sequence analysis in Bioinformatics (**ASAB**)

Read

CTCAAACTCCTGACCTTTGGTGATCCACCCGCCTNGGCCTTC

Reference

Sequence differences occur because of...

1. Sequencing error
2. Genetic variation

https://langmead-lab.org/teaching-materials/

What do we need to determine the original genomic location of these reads?

Global Alignment with Respect to the Reads

Local Alignment

Target Sequence

```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
             |||| |||||| ||||||||||||||||
```

Query Sequence

```
5' TACTCACGGATGAGGTACTTTAGAGGC 3'
```

Global Alignment

Target Sequence **Reference**

```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
   ||||||||||||     |||||||  |||||||||||||| ||||||||
```

```
5' ACTACTAGATT----ACGGATC--GTACTTTAGAGGCTAGCAACCA 3'
```

Query Sequence **Reads**

BWA in action:
Short Reads Alignment

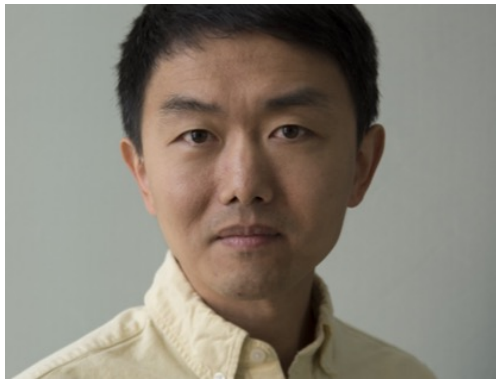
BWA

Burrows Wheeler Aligner(s)

Is a software package for mapping low-divergent sequences against a large ***reference*** genome, such as the human genome.

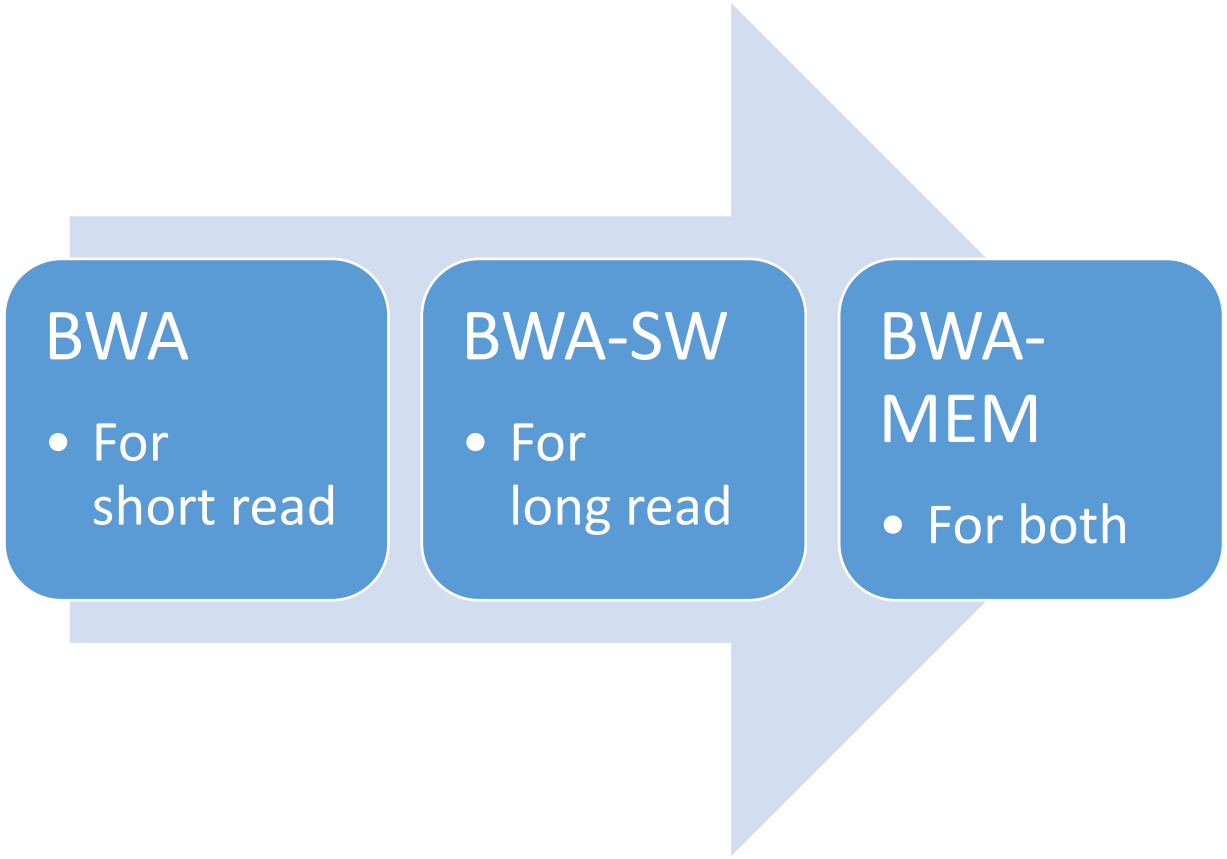
Heng Li (Broad Institute at MIT)

http://bio-bwa.sourceforge.net/bwa.shtml



BROAD INSTITUTE

BWA
Burrows Wheeler Aligner(s)



BWA

Burrows Wheeler Aligner(s)

It consists of three algorithms:

- **BWA-backtrack** (Illumina sequencing reads ≤ 100bp)
- BWA-SW (more sensitive when alignment gaps are frequent)
- BWA-MEM (maximum exact matches)

BWA-SW and BWA-MEM can map “longer” reads from 75 bp to 1Mb

BWA Fundamentals

- Burrow-Wheeler Transform (**BWT**) to Construct a Suffix Array (SA) of the **reference string X**
- Backward search to build the **FM-index**
- Knowing the intervals in the Suffix-Array we can get the positions in the genome.
- Sequence alignment searching for the SA intervals of substrings of X that match the query (i.e. our read).

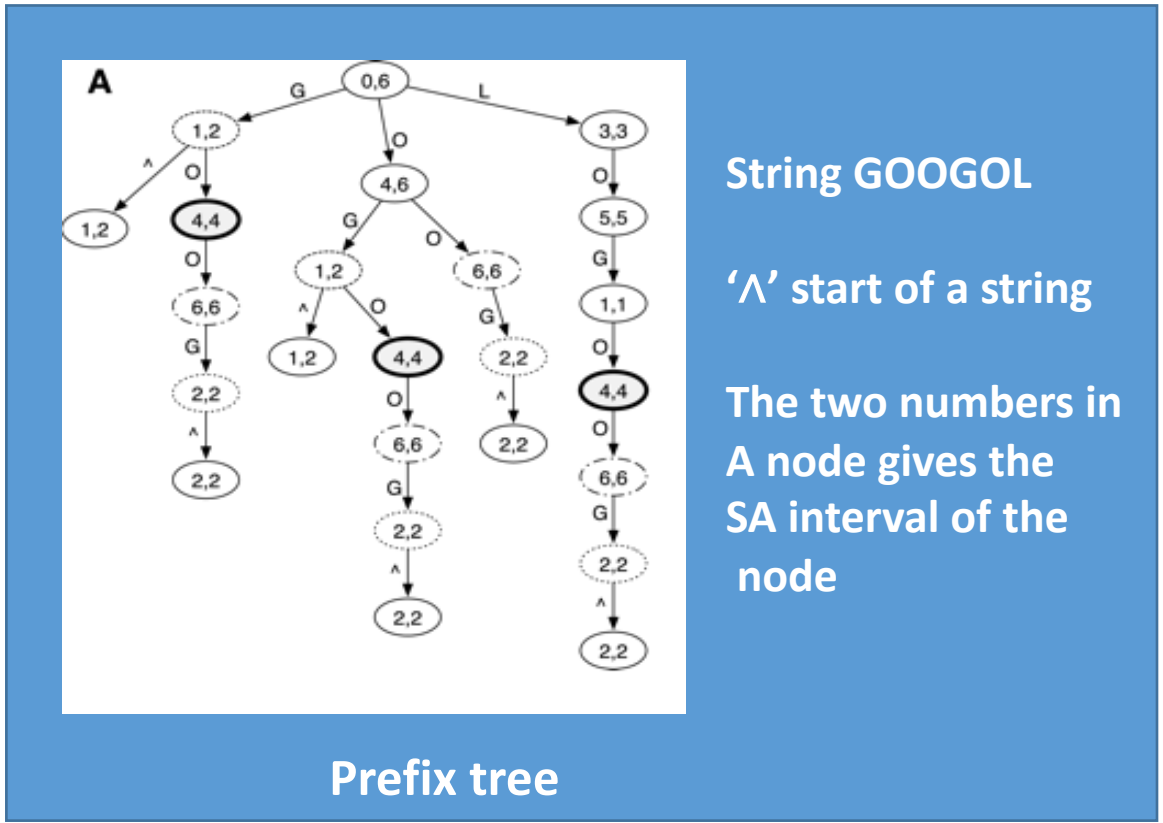
Li, H., and R. Durbin, 2009 Fast and accurate **short read** alignment with Burrows-Wheeler transform. *Bioinformatics* 25 (14):1754-1760.

BWA Algorithm Overview

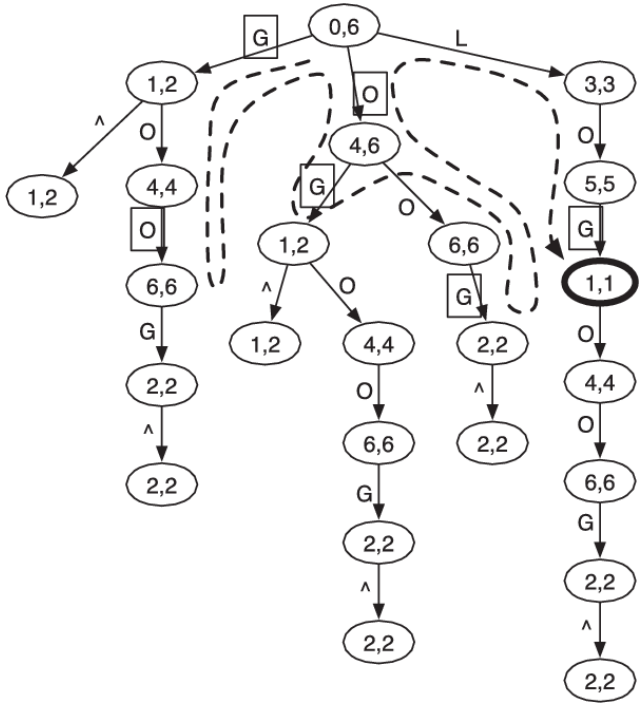
(1) Build **FM-indices** for reference and query sequences

(2) Represent reference in a prefix trie

(3) Mapping: searching for the SA intervals of substrings of X (reference) that match the query (read).



BWA Algorithm Overview



Suffix "*Trie*"

Alignment with BWT

0	googol$
1	oogol$g
2	ogol$go
3	gol$goo
4	ol$goog
5	l$googo
6	$googol

Alphabetical String Sorting

0	6	$googo l
1	3	gol$go o
2	0	googol $
3	5	l$goog o
4	2	ogol$g o
5	4	ol$goo g
6	1	oogol$ g

Pos

X = googol$

i S(i) B[i]

lo$oogg

(6,3,0,5,2,4,1)

Suffix Array

BWT
(+ FM Index)

Li, H., and R. Durbin, 2009 Fast and accurate **short read** alignment with Burrows-Wheeler transform. *Bioinformatics* 25 (14):1754-1760.

Note: Updated after uploading Moodle

Fig. 3. Algorithm for inexact search of SA intervals of substrings that match W. Reference X is $ terminated, while W ...

Mapping Algorithm

Precalculation:
Calculate BWT string B for reference string X
Calculate array $C(\cdot)$ and $O(\cdot,\cdot)$ from B
Calculate BWT string B' for the reverse reference
Calculate array $O'(\cdot,\cdot)$ from B'
Procedures:

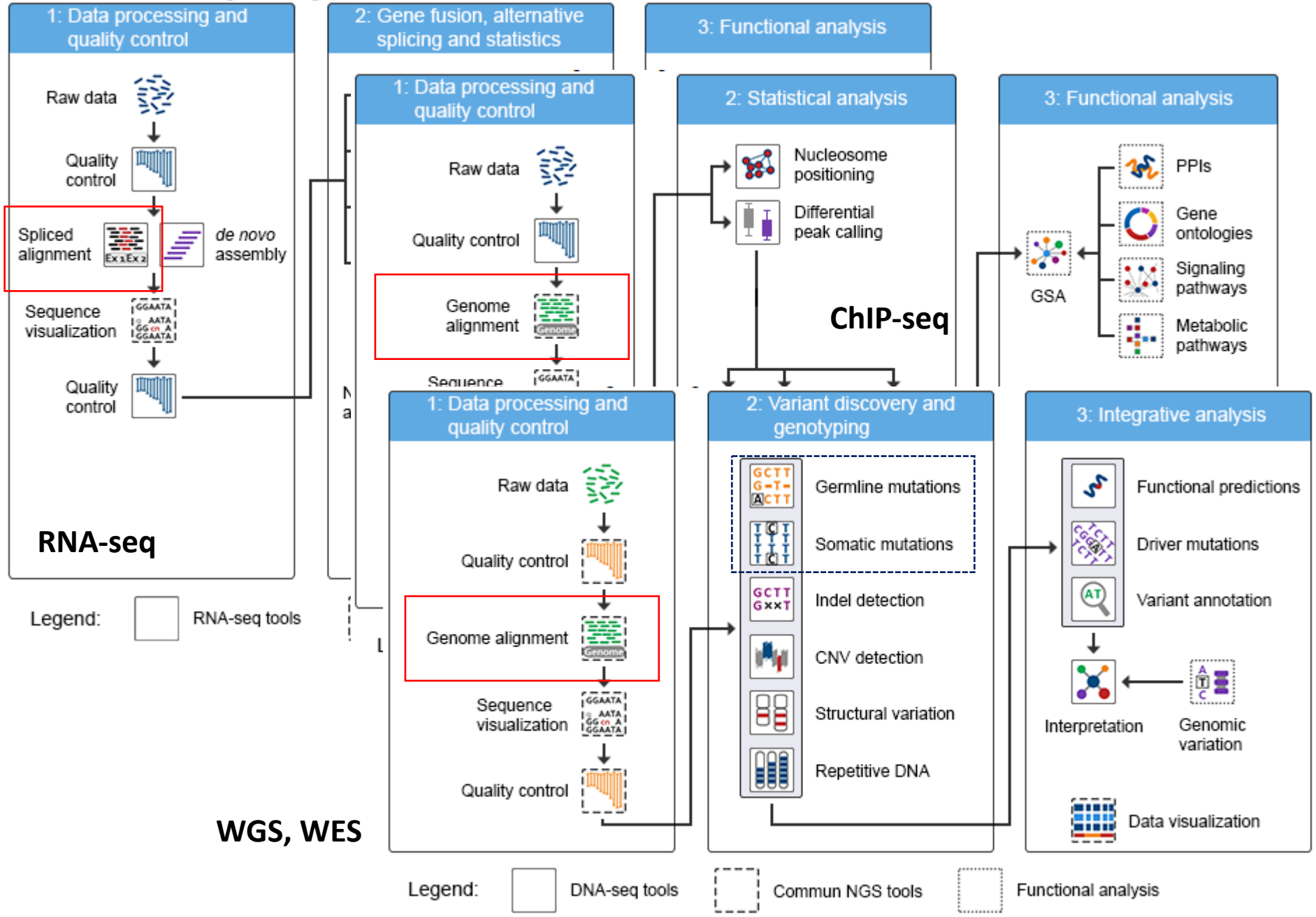
```
InexactSearch(W, z)
    CalculateD(W)
    return InexRecur(W, |W|−1, z, 1, |X|−1)

CalculateD(W)
    k ← 1
    l ← |X|−1
    z ← 0
    for i = 0 to |W|−1 do
        k ← C(W[i]) + O'(W[i], k−1) + 1
        l ← C(W[i]) + O'(W[i], l)
        if k > l then
            k ← 1
            l ← |X|−1
            z ← z+1
        D(i) ← z

InexRecur(W, i, z, k, l)
    if z < D(i) then
        return ∅
    if i < 0 then
        return {[k, l]}
    I ← ∅
*   I ← I ∪ InexRecur(W, i−1, z−1, k, l)
    for each b ∈ {A, C, G, T} do
        k ← C(b) + O(b, k−1) + 1
        l ← C(b) + O(b, l)
        if k ≤ l then
*           I ← I ∪ InexRecur(W, i, z−1, k, l)
            if b = W[i] then
                I ← I ∪ InexRecur(W, i−1, z, k, l)
            else
                I ← I ∪ InexRecur(W, i−1, z−1, k, l)
    return I
```

Bioinformatics, Volume 25, Issue 14, 15 July 2009, Pages 1754–1760, https://doi.org/10.1093/bioinformatics/btp324

BWA in action:
Mapping-based Applications In Genomics


1: Data processing and quality control
Raw data
Quality control
Spliced alignment
de novo assembly
Sequence visualization
Quality control
RNA-seq
2: Gene fusion, alternative splicing and statistics
3: Functional analysis
1: Data processing and quality control
Raw data
Quality control
Genome alignment
Sequence
2: Statistical analysis
Nucleosome positioning
Differential peak calling
ChIP-seq
3: Functional analysis
GSA
PPIs
Gene ontologies
Signaling pathways
Metabolic pathways
1: Data processing and quality control
Raw data
Quality control
Genome alignment
Sequence visualization
Quality control
2: Variant discovery and genotyping
Germline mutations
Somatic mutations
Indel detection
CNV detection
Structural variation
Repetitive DNA
3: Integrative analysis
Functional predictions
Driver mutations
Variant annotation
Interpretation
Genomic variation
Data visualization
WGS, WES
Legend:
RNA-seq tools
Legend:
DNA-seq tools
Commun NGS tools
Functional analysis

Mapping Quality

Mapping short DNA sequencing reads and calling variants using mapping quality scores

Heng Li,[1] Jue Ruan,[2] and Richard Durbin[1,3]

[1]*The Wellcome Trust Sanger Institute, Hinxton CB10 1SA, United Kingdom;* [2]*Beijing Genomics Institute, Chinese Academy of Science, Beijing 100029, China*

New sequencing technologies promise a new era in the use of DNA sequence. However, some of these technologies produce very short reads, typically of a few tens of base pairs, and to use these reads effectively requires new algorithms and software. In particular, there is a major issue in efficiently aligning short reads to a reference genome and handling ambiguity or lack of accuracy in this alignment. Here we introduce the concept of *mapping quality*, a measure of the confidence that a read actually comes from the position it is aligned to by the mapping algorithm. We describe the software MAQ that can build assemblies by mapping shotgun short reads to a reference genome, using quality scores to derive genotype calls of the consensus sequence of a diploid genome, e.g., from a human sample. MAQ makes full use of mate-pair information and estimates the error probability of each read alignment. Error probabilities are also derived for the final genotype calls, using a Bayesian statistical model that incorporates the mapping qualities, error probabilities from the raw sequence quality scores, sampling of the two haplotypes, and an empirical model for correlated errors at a site. Both read mapping and genotype calling are evaluated on simulated data and real data. MAQ is accurate, efficient, versatile, and user-friendly. It is freely available at http://maq.sourceforge.net.

Li H, Ruan J, Durbin R: **Mapping short DNA sequencing reads and calling variants using mapping quality scores.** *Genome Res* 2008, **18**:1851-1858

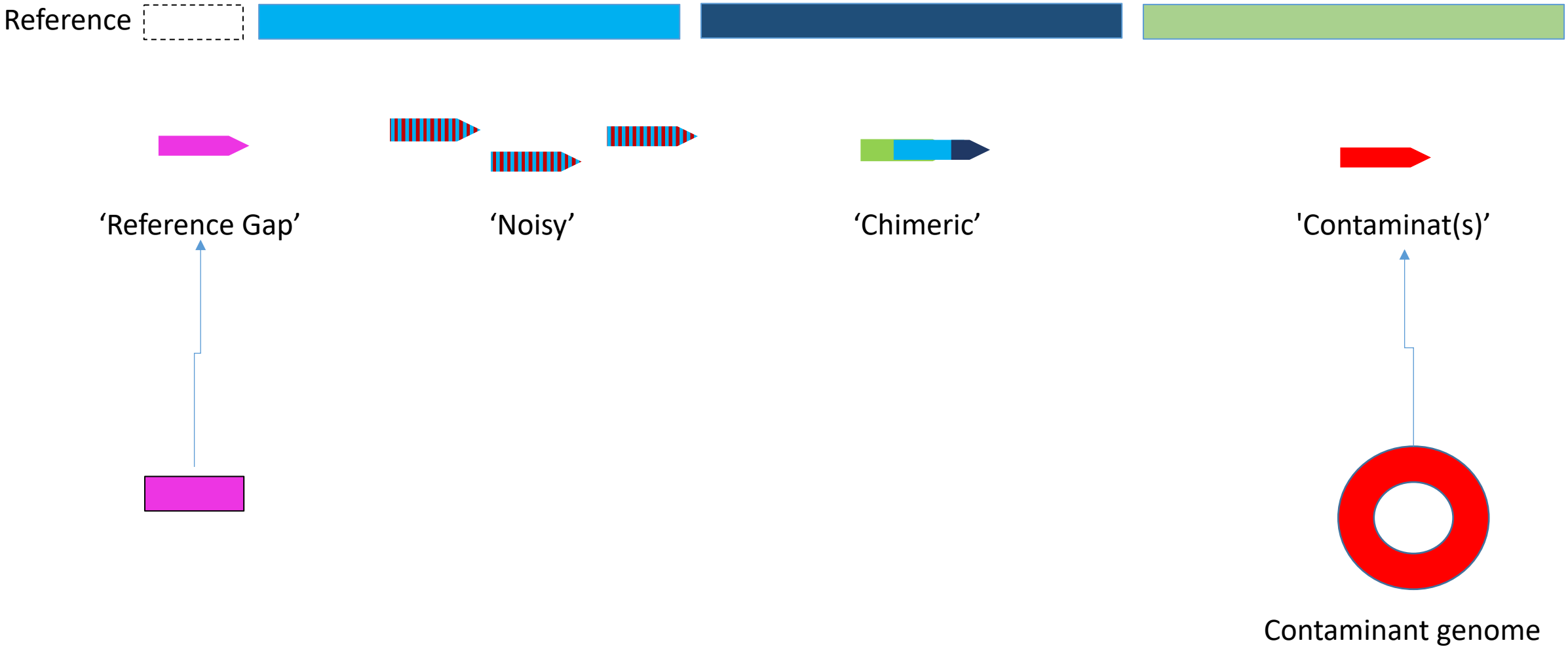
Mapping Quality

Probability of a Mapping to be incorrect = $10^{(-MQ/10)}$

MQ	P mapping error
0	1 in 1
10	1 in 10
20	1 in 100
30	1 in 1000
40	1 in 10,000
50	1 in 100,000
60	1 in 1,000,000

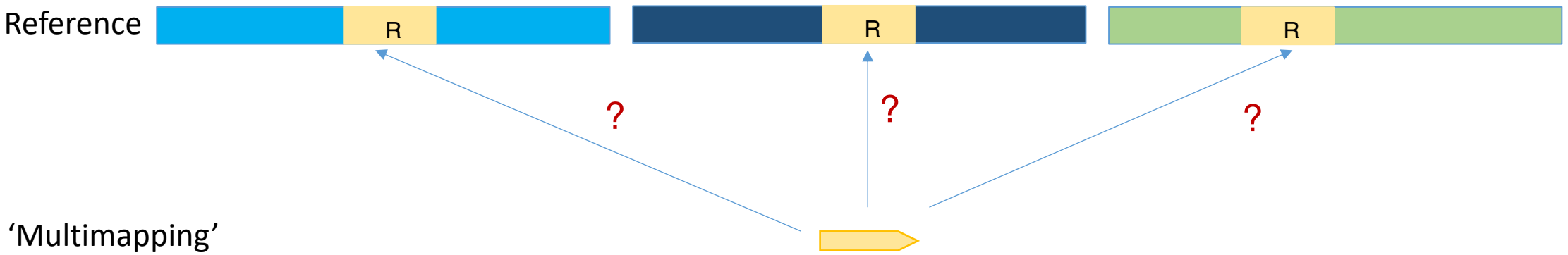
Unmapped Reads (by default MQ=0)

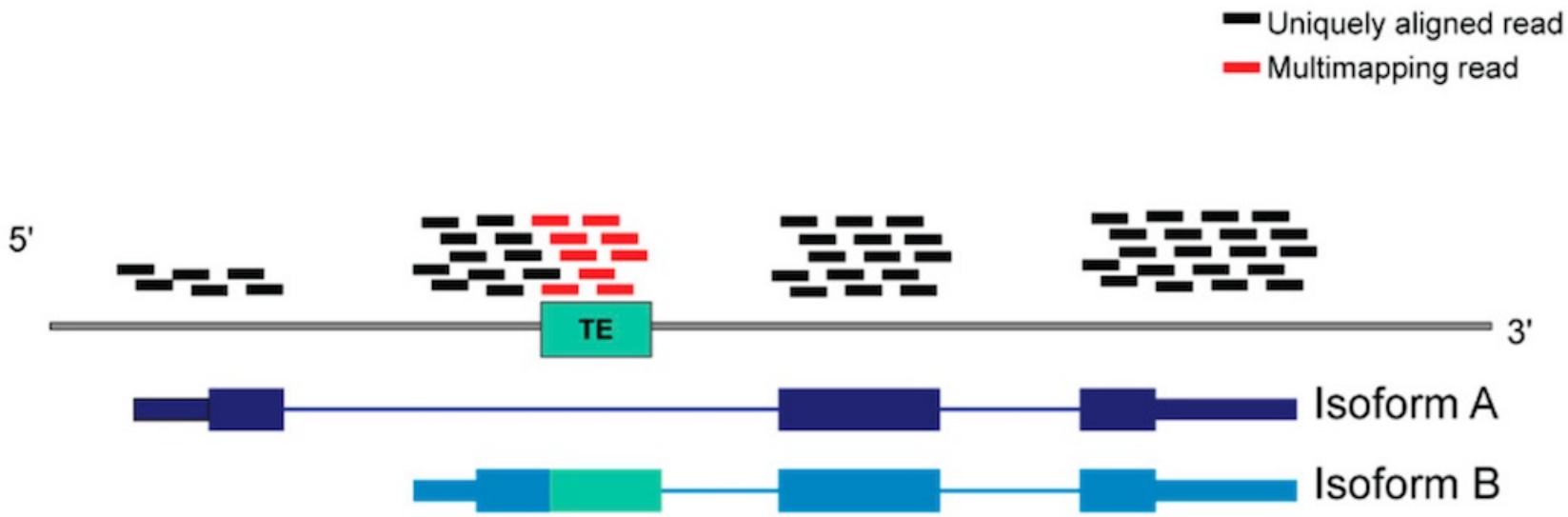
Reads that do not map to our reference genome



Multimappings (MQ=0-10)

Reads difficult to place into a unique location in the genome



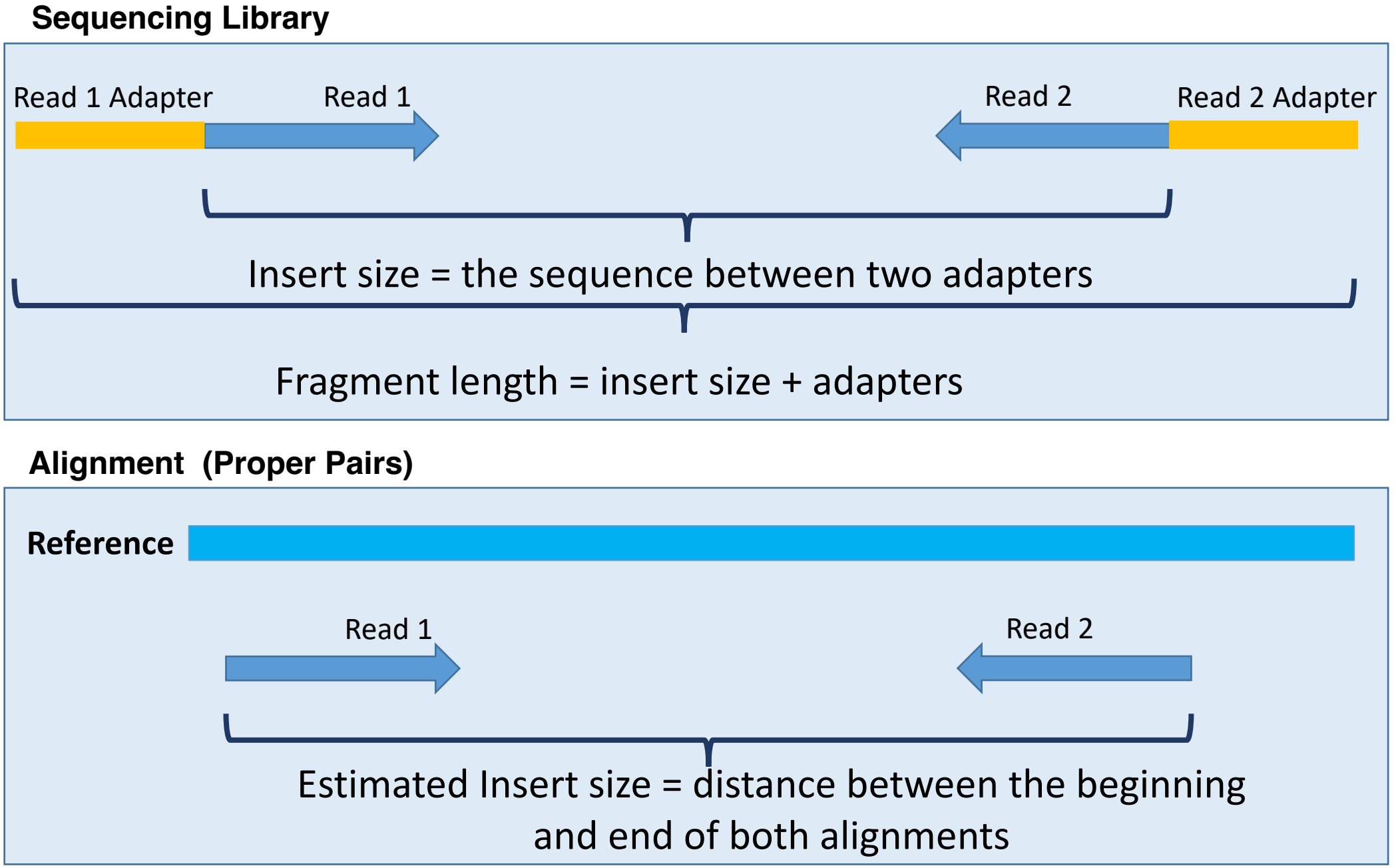


Mapping Quality

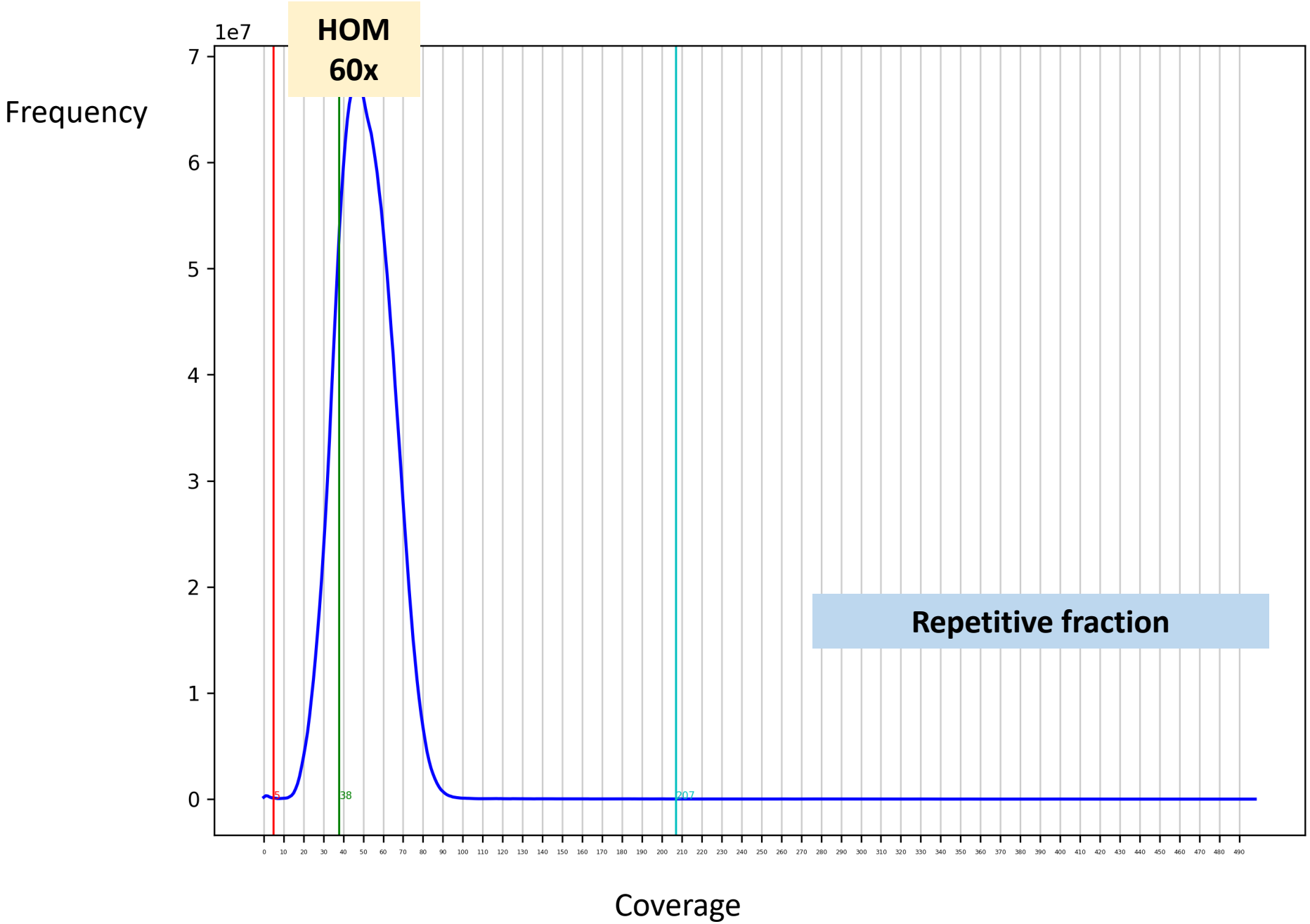
MQ	Interpretation
0	Unmapped/highly multimapping reads
1-10	Multimapping
11-39	Slightly ambiguous mappings
40-59	Almost Unique
60	Unique mapping

Note: Updated after uploading Moodle

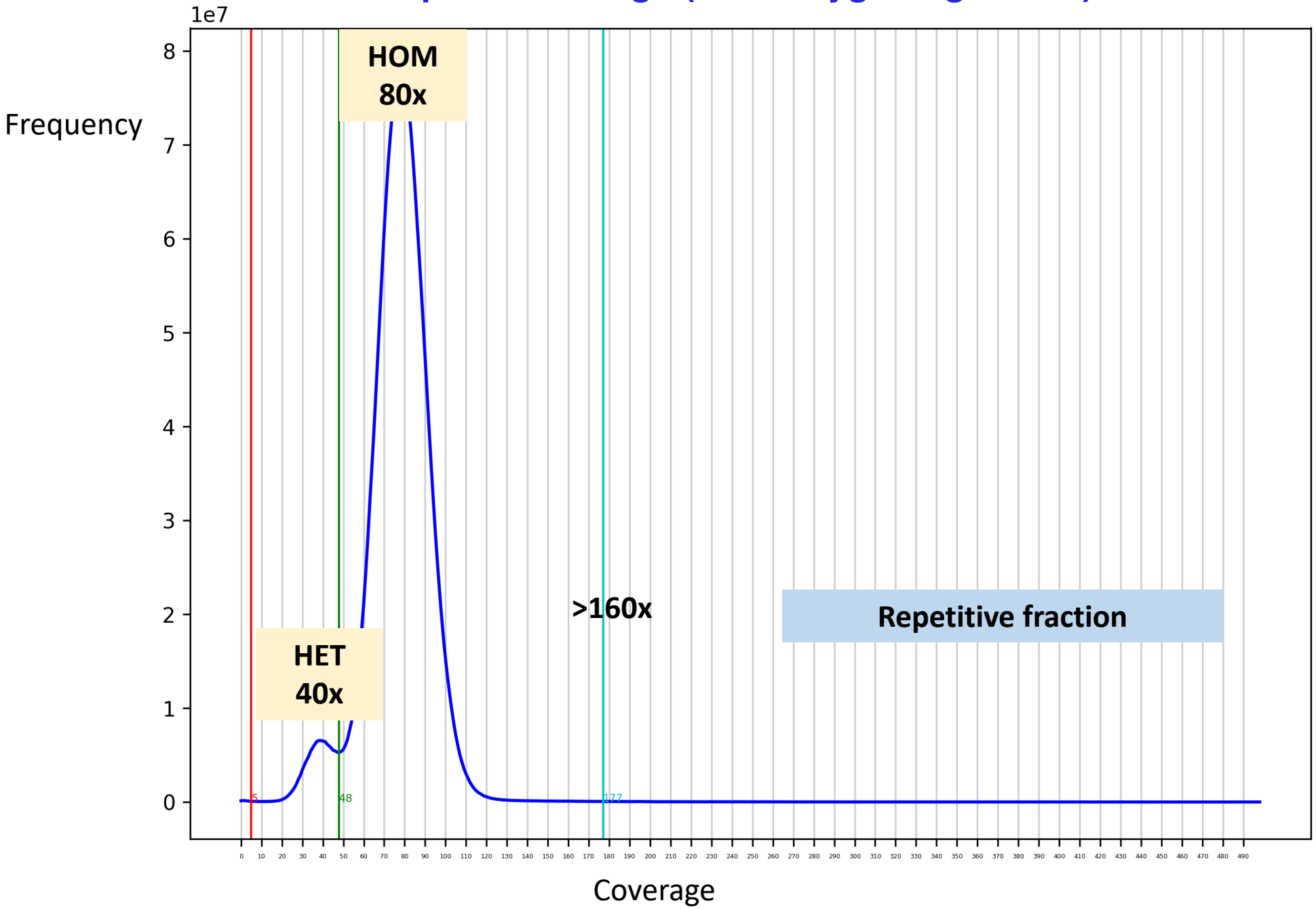
Paired-End (PE) Reads



Compute Coverage (Homozygous genome)

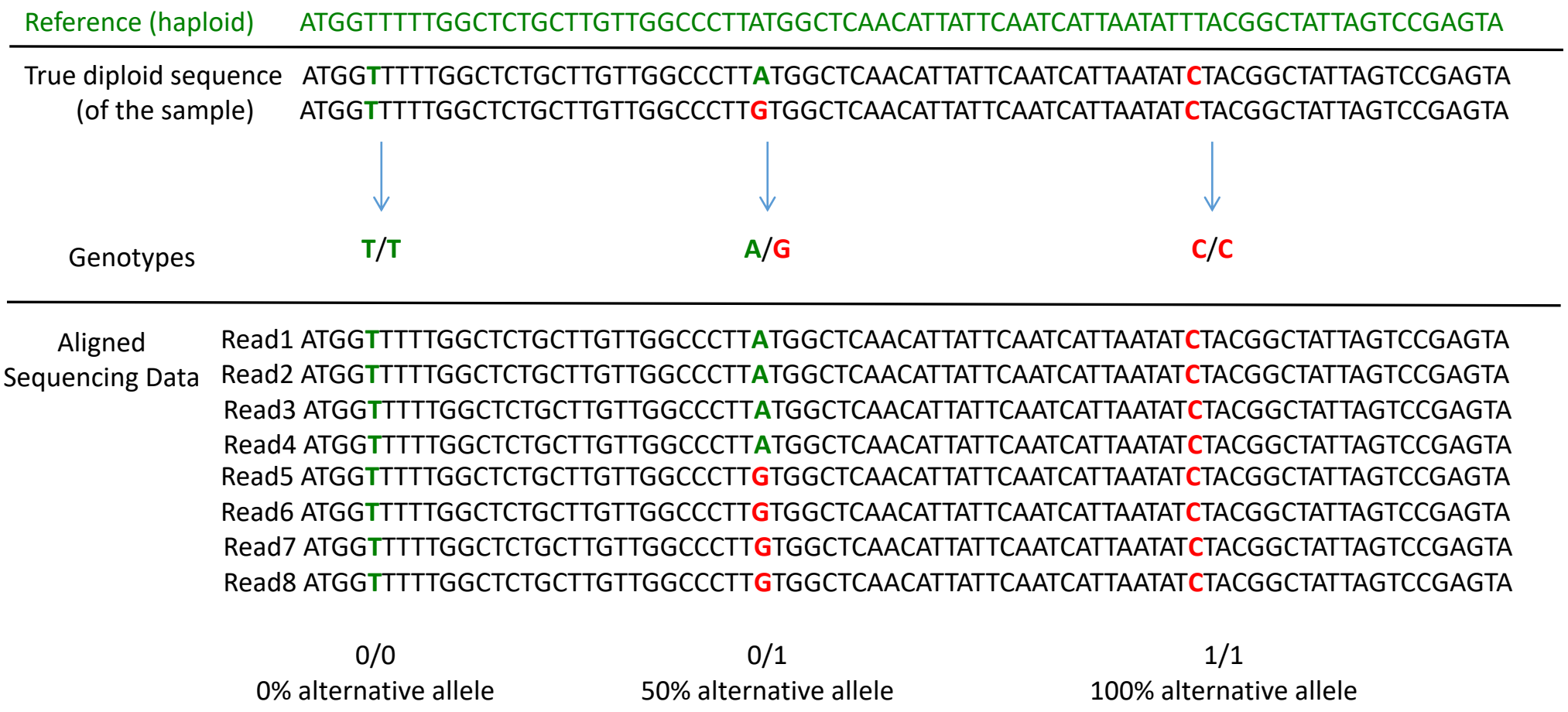


Compute Coverage (Heterozygous genome)

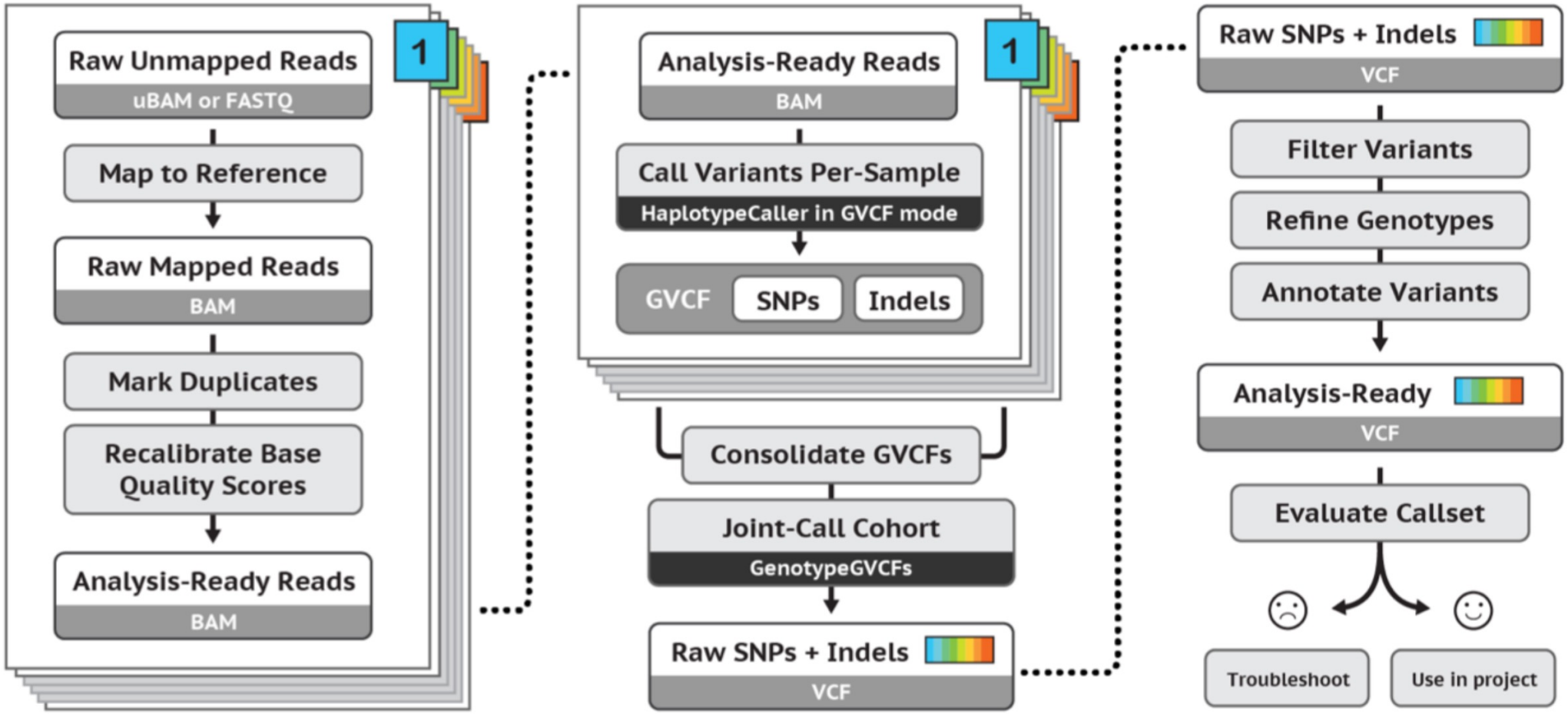


Variant Calling - from Sequence Alignments to Genomic Variants

Genetic difference identified by comparison to an haploid reference:



Variant Calling



https://gatk.broadinstitute.org/hc/en-us/sections/360007226651-Best-Practices-Workflows

Gene Expression

GATK / Tool Index / 4.1.9.0

GeneExpressionEvaluation (BETA) Follow

GATK Team
1 year ago · Updated

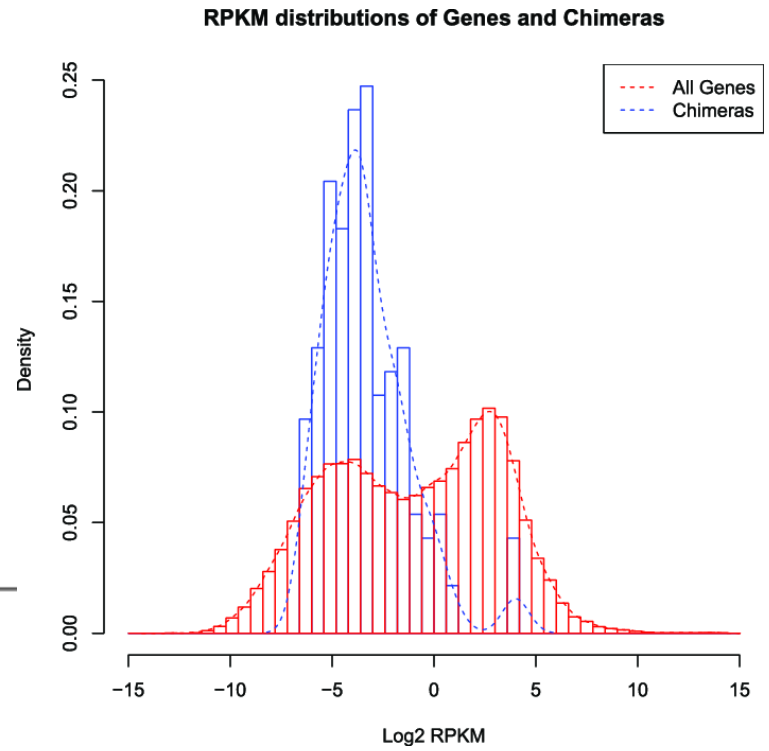
Evaluate gene expression from RNA-seq reads aligned to genome.

Category Coverage Analysis

Overview

Evaluate gene expression from RNA-seq reads aligned to genome.

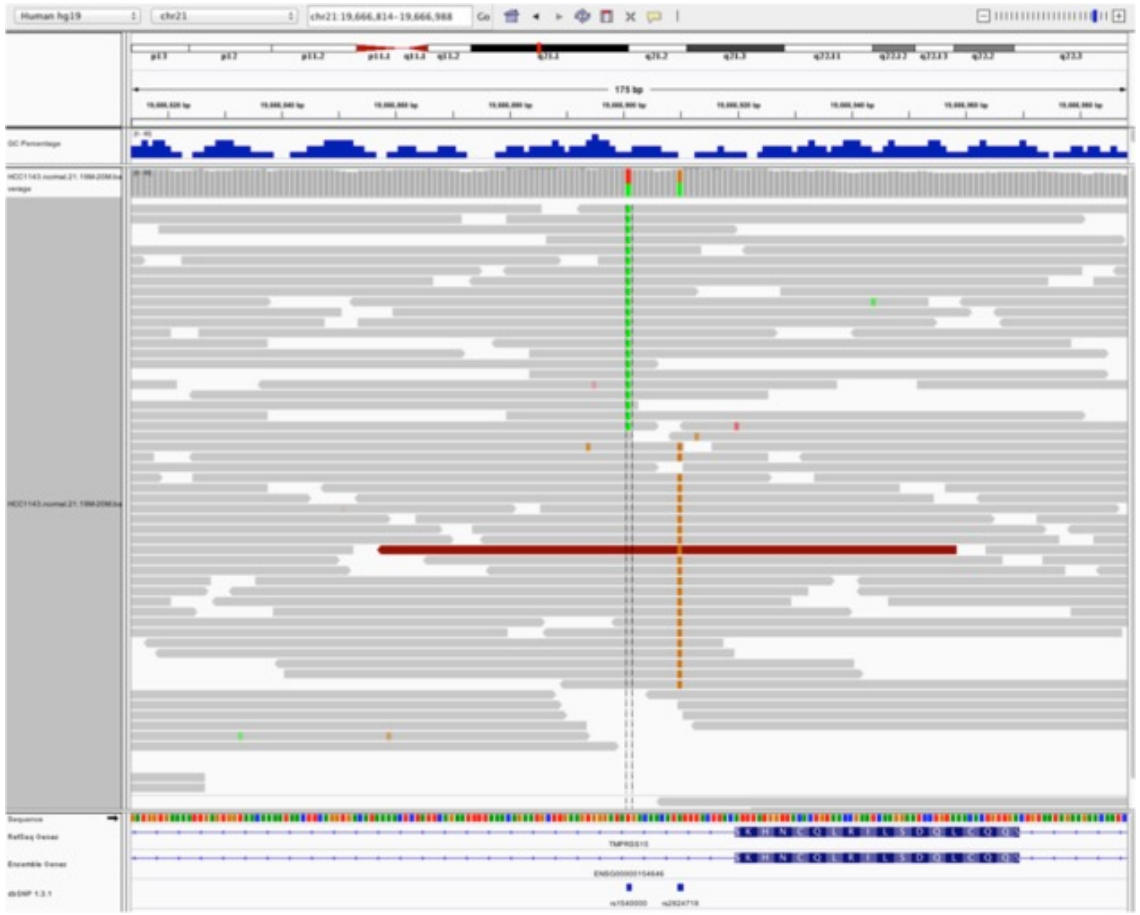
This tool counts fragments to evaluate gene expression from RNA-seq reads aligned to the genome. Features to evaluate expression over are defined in an input annotation file in gff3 fomat (https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md). Output is a tsv listing sense and antisense expression for all stranded grouping features, and expression (labeled as sense) for all unstranded grouping features.

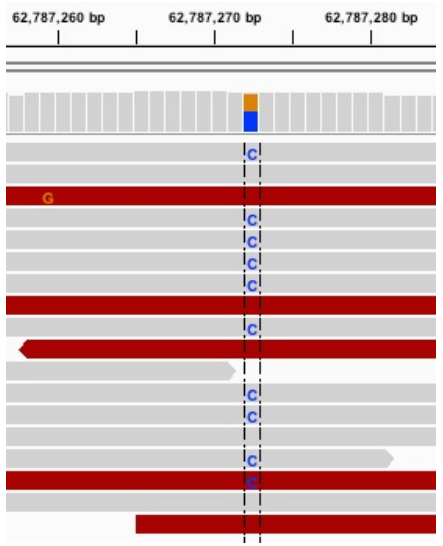


https://gatk.broadinstitute.org/hc/en-us/articles/360051304891-GeneExpressionEvaluation-BETA-

Visualizing Alignments with IGV

Integrative Genomics Viewer





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

Hitchhicker's Guide for Mapping-Based Applications

Read Alignments/Mappings

- **BWA-MEM** : map illumina reads
- **Minimap2** : map long noisy reads (Pacbio, Nanopore)

Manipulating Mappings

- **Samtools** : SAM/BAM conversión, view BAMs, select alignments for genomic intervals…
- **Picard** : Process alignments (e.g. Mark PCR Duplicates), get coverage etc…

Pairwise Alignments

- **MUMMER package** : align with nucmer4, produce Dot Plots with mummerplot, etc.
- **Minimap2** : fast genome alignments in PAF format

Preprocess illumina reads

- **Cutadapt** : detect and remove adaptors
- **FastQC** : quality report of fastQ files

Compute coverage

- **Deeptools** : PlotCoverage, etc
- **Bedtools** : coverage per-site, per-window. Also to manipulate genomic intervals, merge and intersect them.

Variant Calling

- GATK package: Variant and Haplotype Caller, etc.

Expression Levels

- GATK GeneExpression evaluation, etc…