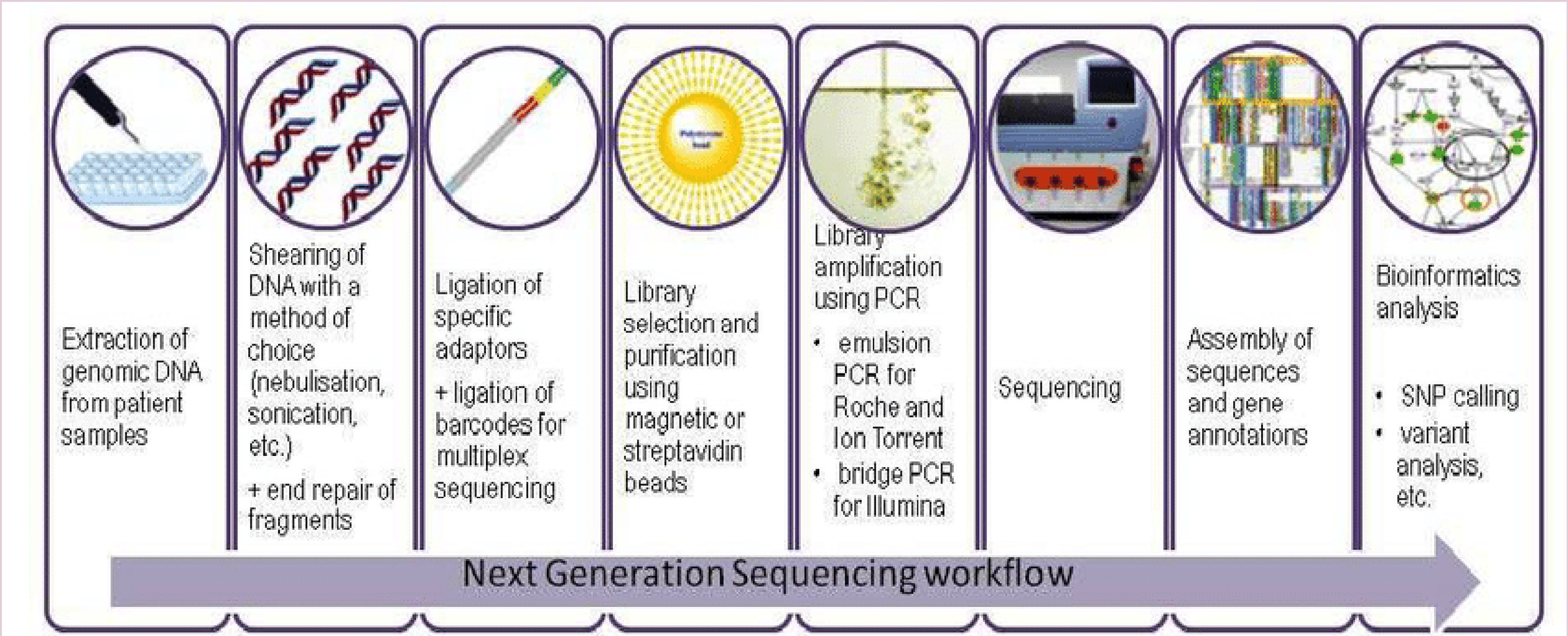
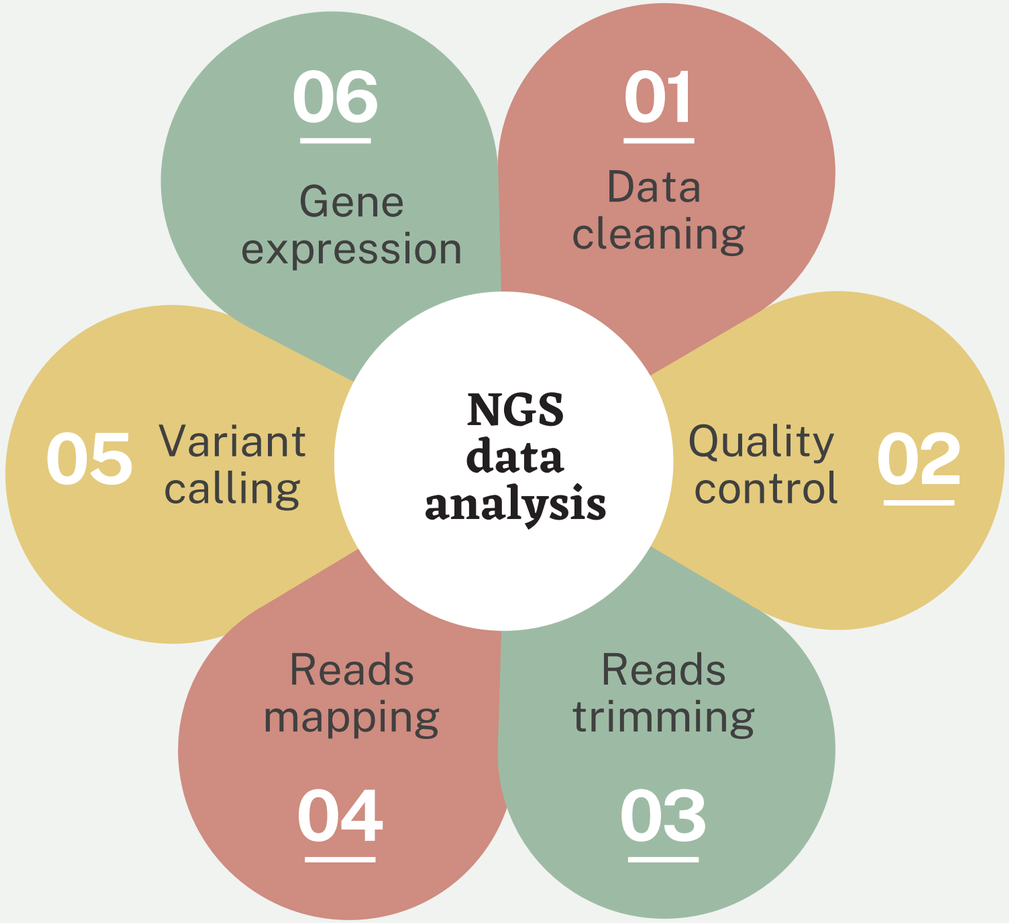
Review of available software tools for NGS data analysis

Jagoda Trzeciak

What is NGS?


Extraction of genomic DNA from patient samples
Shearing of DNA with a method of choice (nebulisation, sonication, etc.)
+ end repair of fragments
Ligation of specific adaptors
+ ligation of barcodes for multiplex sequencing
Library selection and purification using magnetic or streptavidin beads
Library amplification using PCR
• emulsion PCR for Roche and Ion Torrent
• bridge PCR for Illumina
Sequencing
Assembly of sequences and gene annotations
Bioinformatics analysis
• SNP calling
• variant analysis, etc.
Next Generation Sequencing workflow

Basic workflow



Data cleaning

Phred score

```
+SEQ_ID
!''*((((***+))%%%++)(%%%%).1**
```

A quality value Q is an integer representation of the probability *p* that the corresponding base call is incorrect.

$$Q = -10 \log_{10} P \quad \Longrightarrow \quad P = 10^{\frac{-Q}{10}}$$

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%
40	1 in 10000	99.99%
50	1 in 100000	99.999%

FastQC

FastQC Report

Mi 19 Sep 2018
L14-10_1.fastq.gz

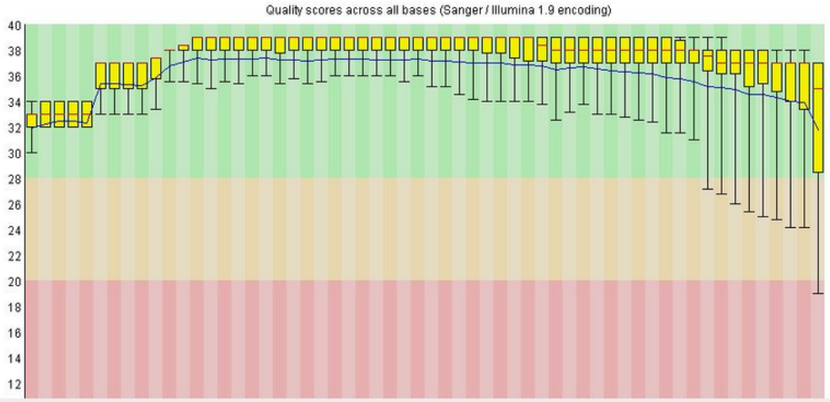
Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

Basic Statistics

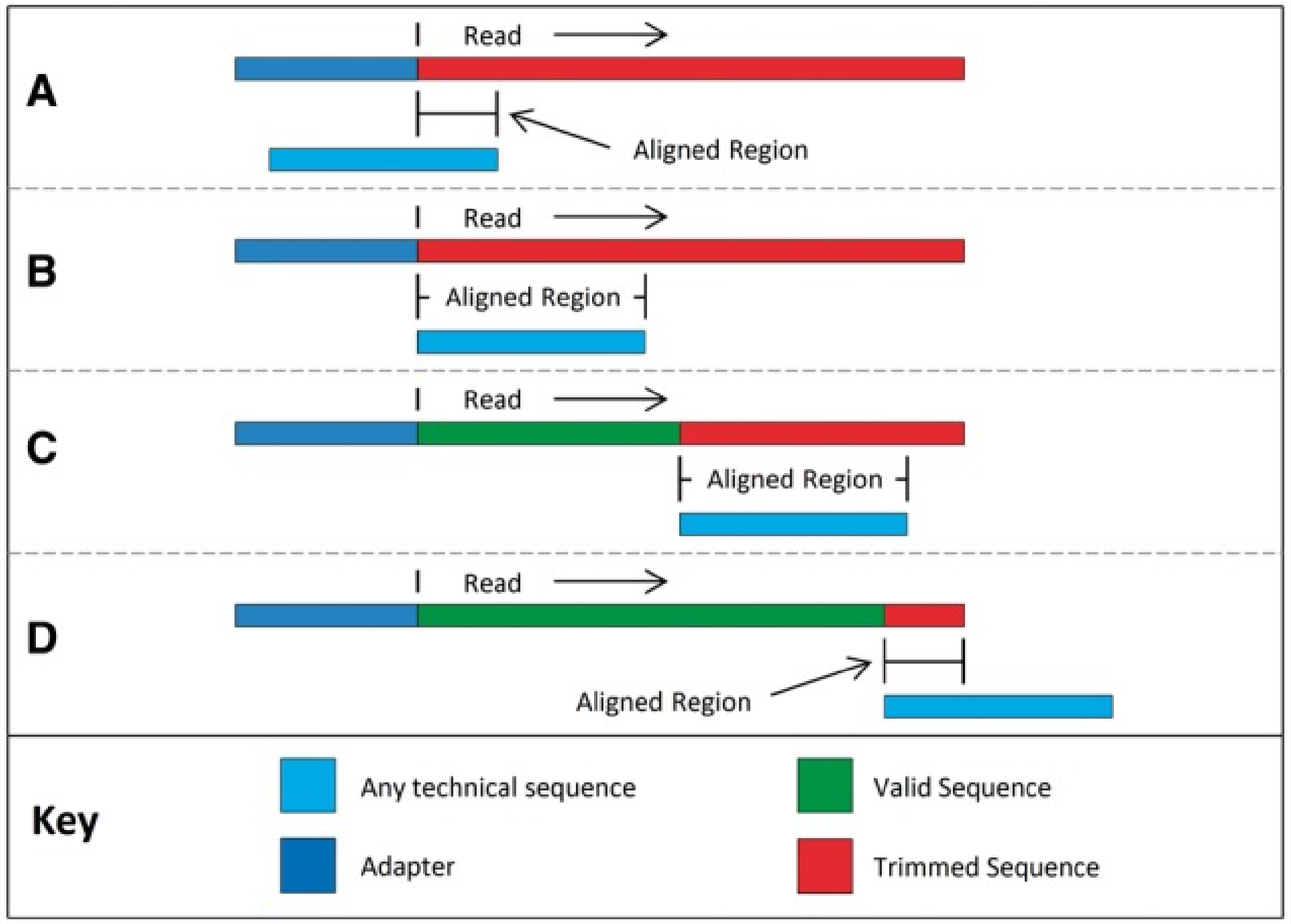
Measure	Value
Filename	L14-10_1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	4950
Sequences flagged as poor quality	0
Sequence length	35-251
%GC	39

Per base sequence quality



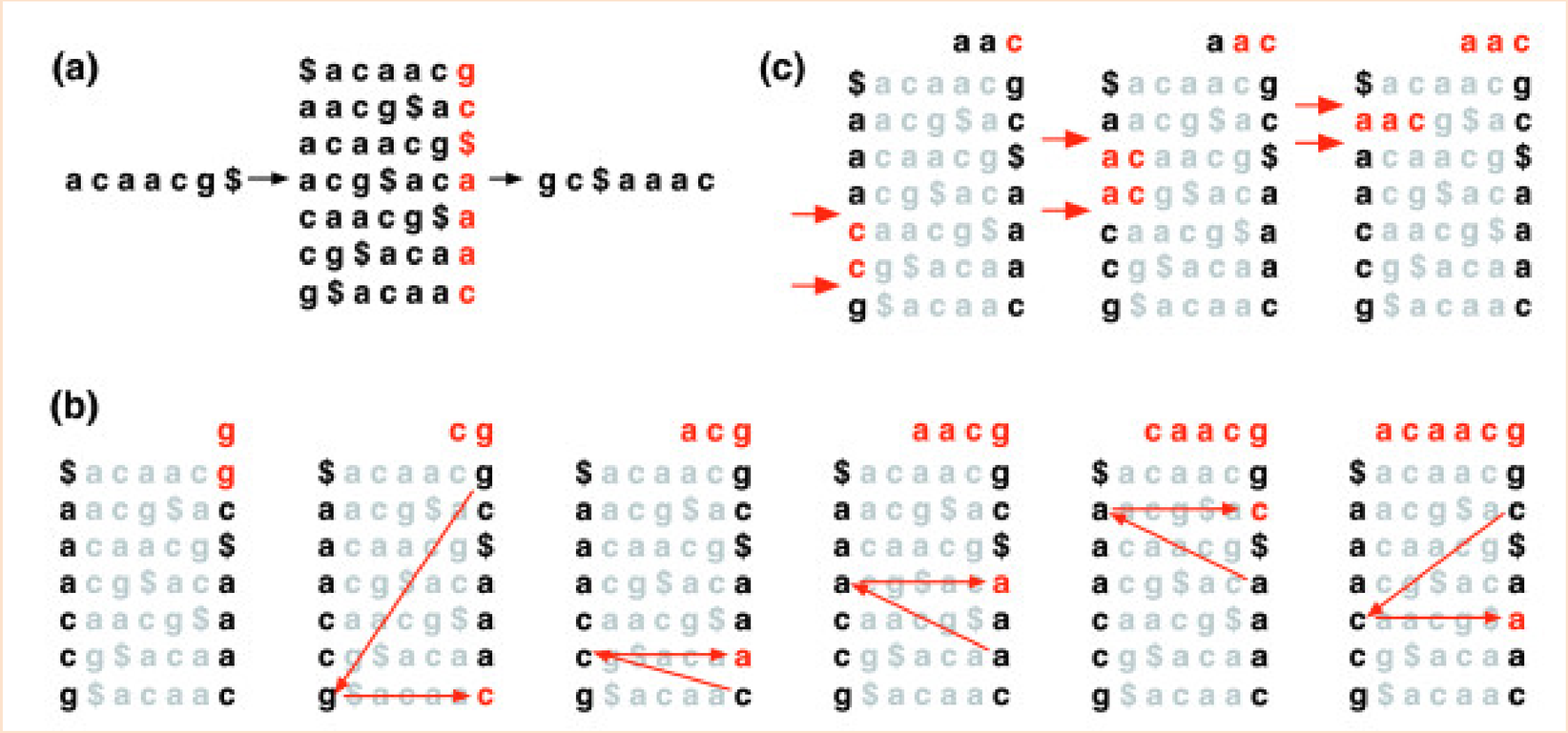
Produced by FastQC (version 0.11.7)

Trimmomatic

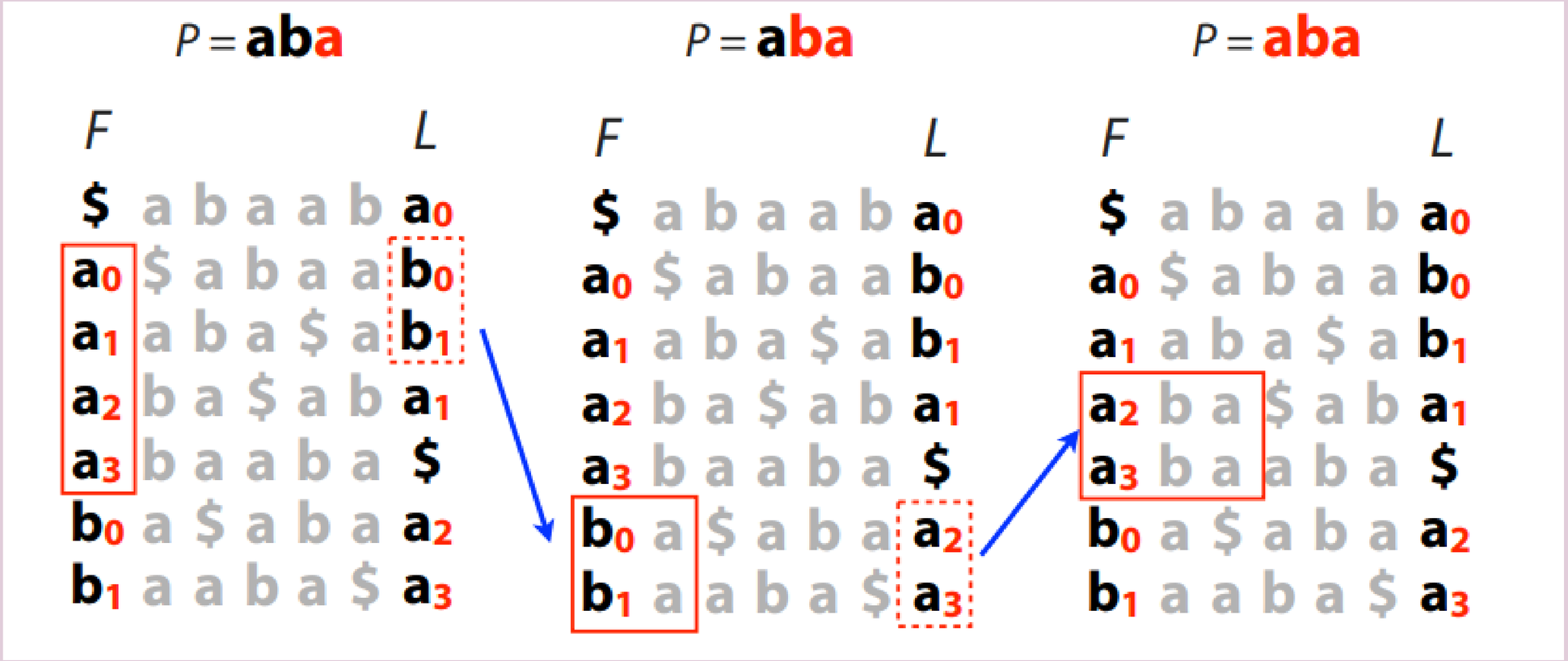


Reads mapping

Burrows-Wheeler Aligner



Bowtie2


P = aba
F L
$ a b a a b a0
a0 $ a b a a b0
a1 a b a $ a b1
a2 b a $ a b a1
a3 b a a b a $
b0 a $ a b a a2
b1 a a b a $ a3
P = aba
F L
$ a b a a b a0
a0 $ a b a a b0
a1 a b a $ a b1
a2 b a $ a b a1
a3 b a a b a $
b0 a $ a b a a2
b1 a a b a $ a3
P = aba
F L
$ a b a a b a0
a0 $ a b a a b0
a1 a b a $ a b1
a2 b a $ a b a1
a3 b a a b a $
b0 a $ a b a a2
b1 a a b a $ a3

SAMtools

view

The `view` command filters SAM or BAM formatted data. Using options and arguments it understands what data to select (possibly all of it) and passes only that data through. Input is usually a sam or bam file specified as an argument, but could be sam or bam data piped from any other command. Possible uses include extracting a subset of data into a new file, converting between BAM and SAM formats, and just looking at the raw file contents. The order of extracted reads is preserved.

sort

The `sort` command sorts a BAM file based on its position in the reference, as determined by its alignment. The element + coordinate in the reference that the first matched base in the read aligns to is used as the key to order it by. [TODO: verify]. The sorted output is dumped to a new file by default, although it can be directed to stdout (using the -o option). As sorting is memory intensive and BAM files can be large, this command supports a sectioning mode (with the -m options) to use at most a given amount of memory and generate multiple output file. These files can then be merged to produce a complete sorted BAM file [TODO - investigate the details of this more carefully].

index

The `index` command creates a new index file that allows fast look-up of data in a (sorted) SAM or BAM. Like an index on a database, the generated `*.sam.sai` or `*.bam.bai` file allows programs that can read it to more efficiently work with the data in the associated files.

tview

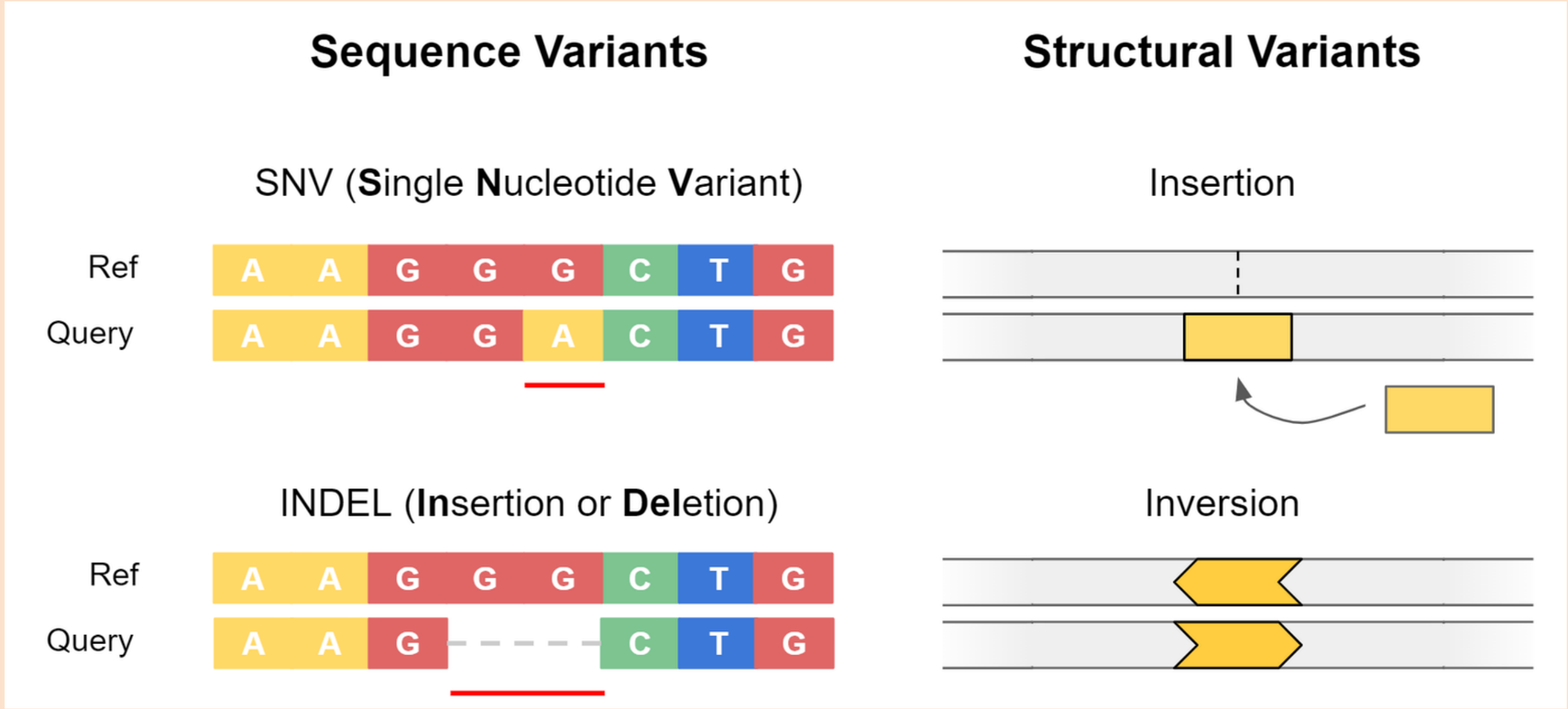
The `tview` command starts an interactive ascii-based viewer that can be used to visualize how reads are aligned to specified small regions of the reference genome. Compared to a graphics based viewer like IGV,[6] it has few features. Within the view, it is possible to jumping to different positions along reference elements (using 'g') and display help information ('?').

mpileup

The `mpileup` command produces a pileup format (or BCF) file giving, for each genomic coordinate, the overlapping read bases and indels at that

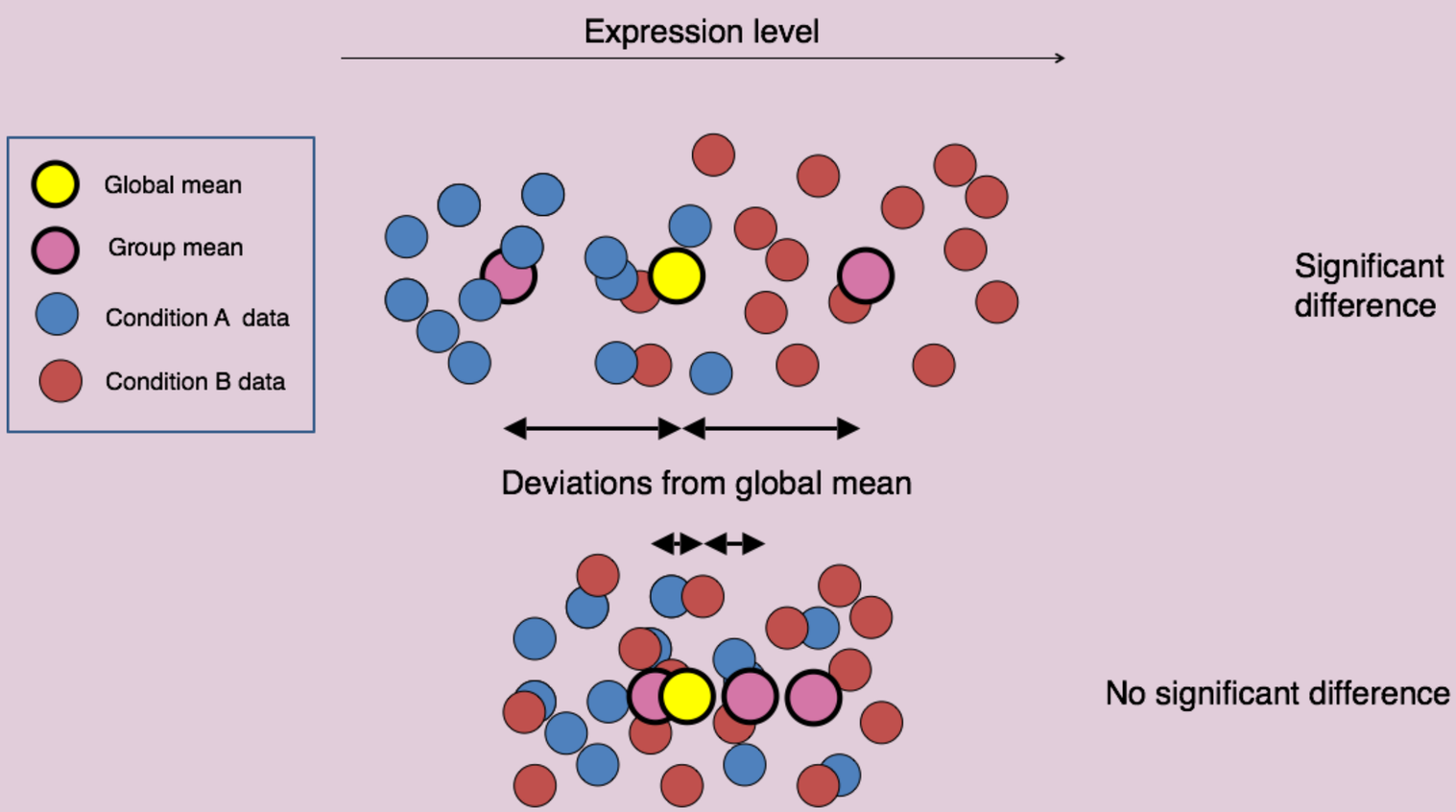
Variant calling

Genome Analysis Toolkit



Gene expression analysis

DESeq2


Expression level
Global mean
Group mean
Condition A data
Condition B data
Deviations from global mean
Significant difference
No significant difference

Summary

QUALITY CONTROL

- FastQC
- Trimmomatic

READS MAPPING

- BWA
- Bowtie2
- SAMtools

GENE EXPRESSION

- DESeq2

VARIANT CALLING

- Genome Analysis Toolkit

Bibliography

- Pereira R, Oliveira J, Sousa M. Bioinformatics and Computational Tools for Next-Generation Sequencing Analysis in Clinical Genetics. J Clin Med. 2020 Jan 3;9(1):132. doi: 10.3390/jcm9010132. PMID: 31947757; PMCID: PMC7019349.
- Qin D. Next-generation sequencing and its clinical application. Cancer Biol Med. 2019 Feb;16(1):4-10. doi: 10.20892/j.issn.2095-3941.2018.0055. PMID: 31119042; PMCID: PMC6528456.
- Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014 Aug 1;30(15):2114-20. doi: 10.1093/bioinformatics/btu170. Epub 2014 Apr 1. PMID: 24695404; PMCID: PMC4103590.
- https://github.com/s-andrews/FastQC
- https://github.com/lh3/bwa
- https://github.com/BenLangmead/bowtie2
- https://www.htslib.org/
- https://gatk.broadinstitute.org/hc/en-us