McGill initiative in Computational Medicine

Intro to NGS data processing and formats

Georgette Femerling
Population and Statistical Genomics Laboratory
Department of Human Genetics

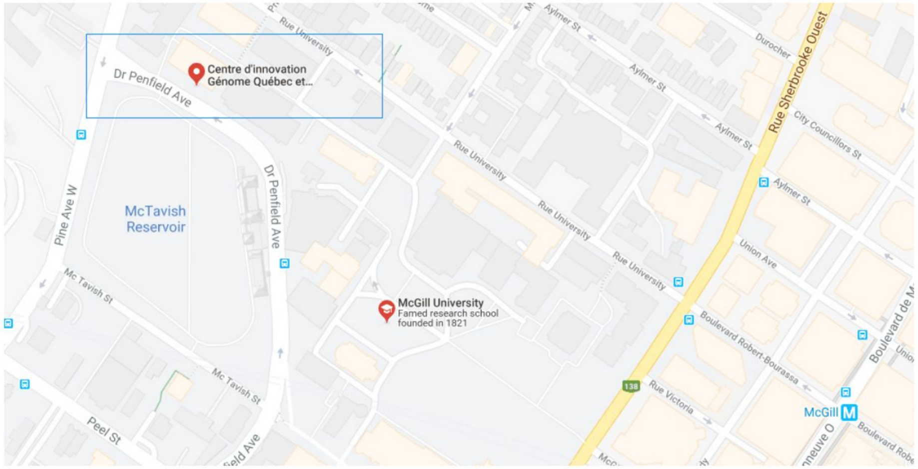
July-7th-2022

Mission : aims to deliver inter-disciplinary research programs and empower the use of data in health research and health care delivery

https://www.mcgill.ca/micm

McGill.CA / MCGILL INITIATIVE IN COMPUTATIONAL MEDICINE

Contact



McGill initiative in Computational Medicine

McGill initiative in Computational Medicine
740, Dr. Penfield Avenue, Montreal, Quebec,
Canada, H3A 0G1
email: info-micm@mcgill.ca

Signup to our newsletter to receive the latest news

Workshop outline

1. Intro to Next Generation Sequencing
2. NGS data quality control and preprocessing
3. Mapping to a reference genome

Part I: Intro to NGS

What is Next Generation Sequencing?

- Sequencing is *reading* the sequence of nucleotides in an RNA/DNA molecule.
- NGS is the highly parallelized sequencing of millions of DNA/RNA fragments at the same time.
- Several methods developed
  - Pyrosequencing
  - Sequencing by ligation (SOLiD)
  - **Sequencing by synthesis (Illumina)**
  - Ion Torrent

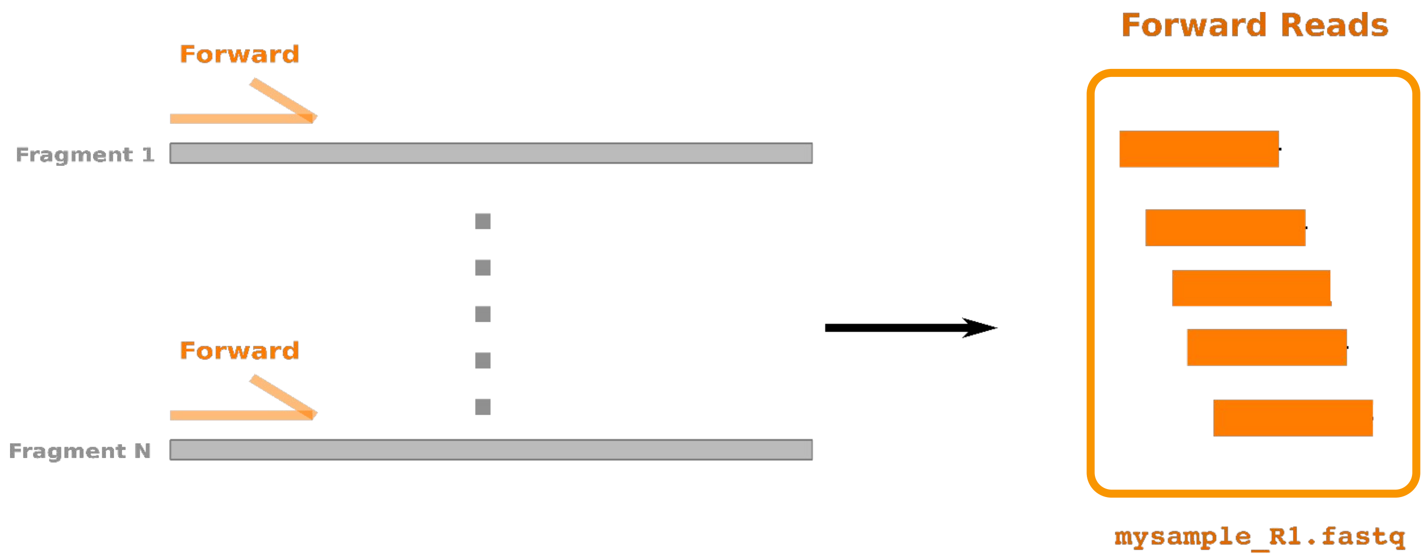
Illumina's HiSeq2000

Short vs Long Read

- There are two approaches for NGS technologies:

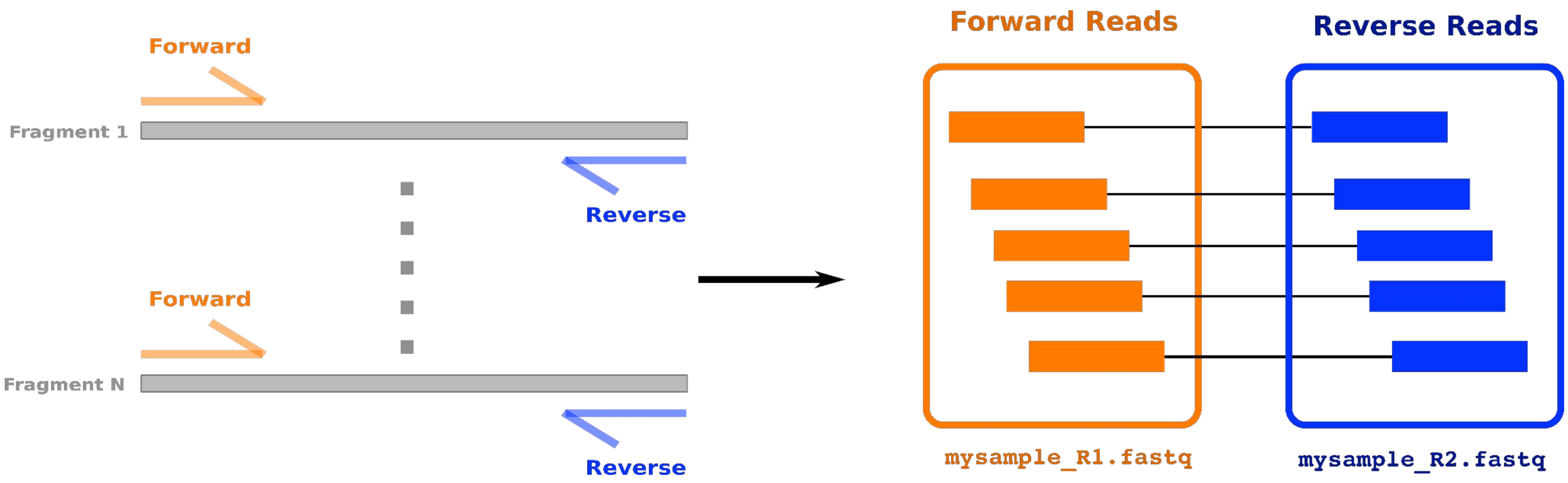
Short-read sequencing	· Higher sequence fidelity · Cheap · Can sequence fragmented DNA
Long-read sequencing	· Able to sequence genetic regions that are difficult to characterize with short-read seq due to repeat sequences · Able to resolve structural rearrangements or homologous regions · Able to read through an entire RNA transcript to determine the specific isoform · Assists *de novo* genome assembly

Single-end sequencing



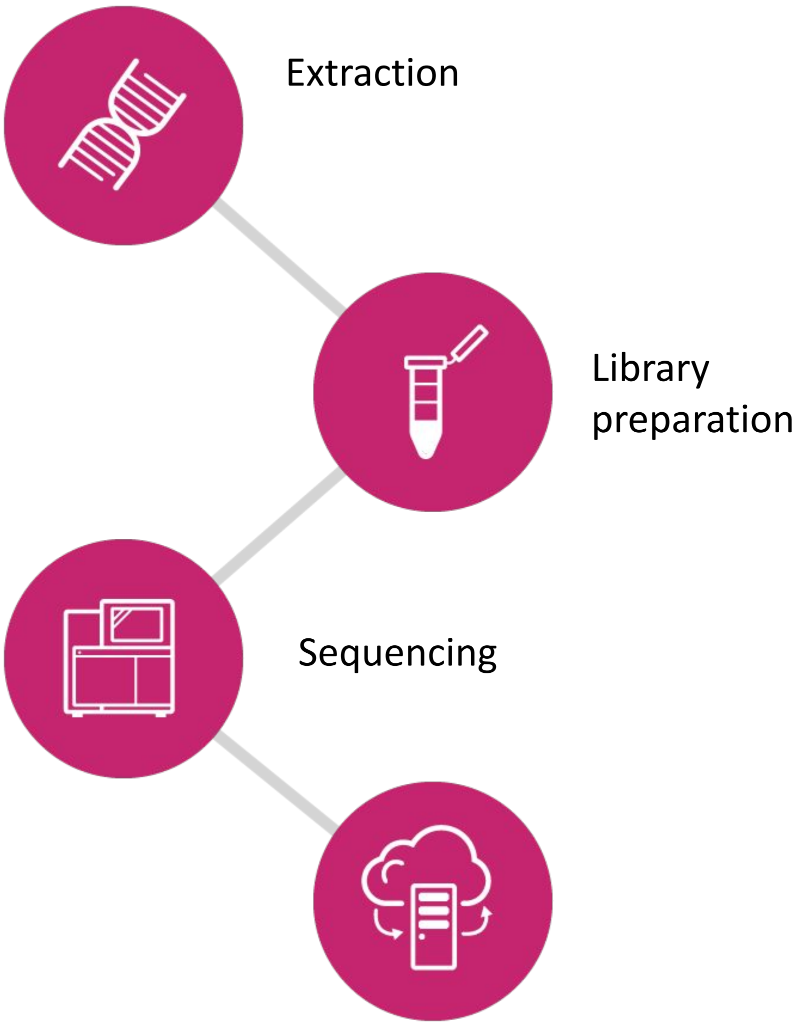
- Single-read sequencing involves sequencing DNA from only one end.
- The reads are independent of one another
- Tends to be cheaper as it yields high volume of data

Paired-end sequencing



- DNA Fragments sequenced from both sides, yielding two reads per fragment (first in forward and second in reverse).
- If you know your fragment size, the distance between both reads is known and therefore is additional information that can improve read mapping.

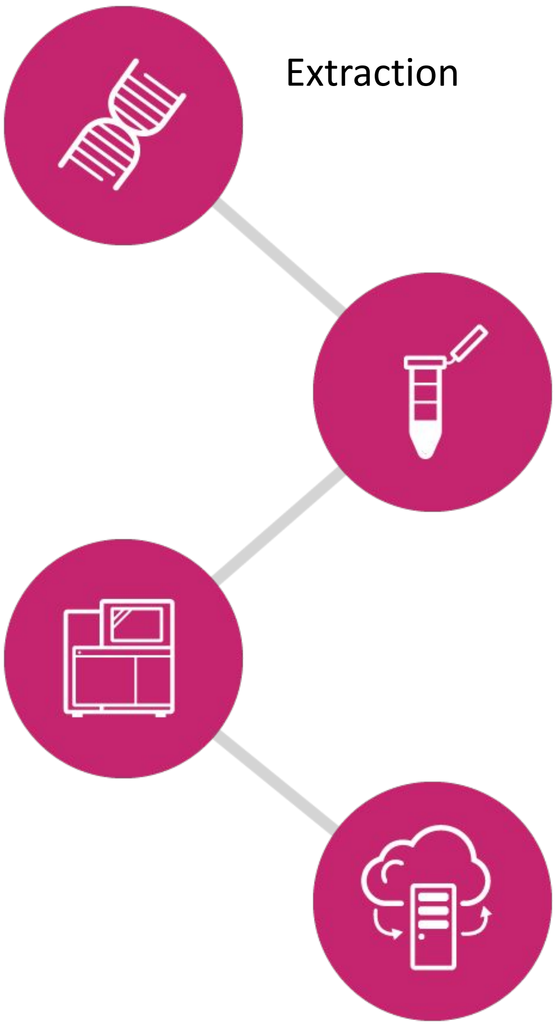
Data Generation



Steps in generating data vary depending on the protocol used, but can be summarized in three steps:

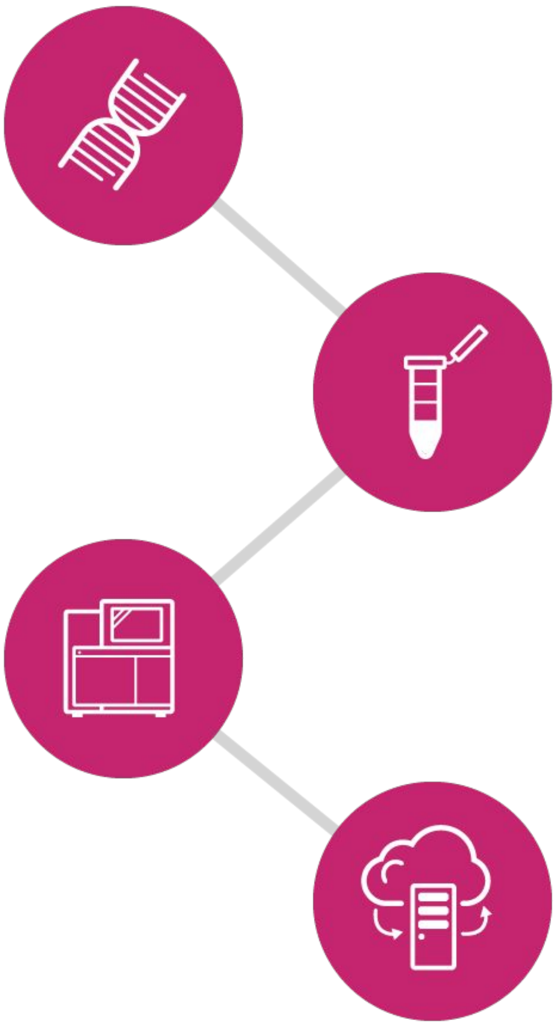
1. Extraction
2. Library Preparation
3. Sequencing

Data Generation - Extraction



- Is the processes of isolating/purifying DNA or RNA from a sample tissue like blood, bone marrow, skin, etc.

Data Generation - Library Preparation

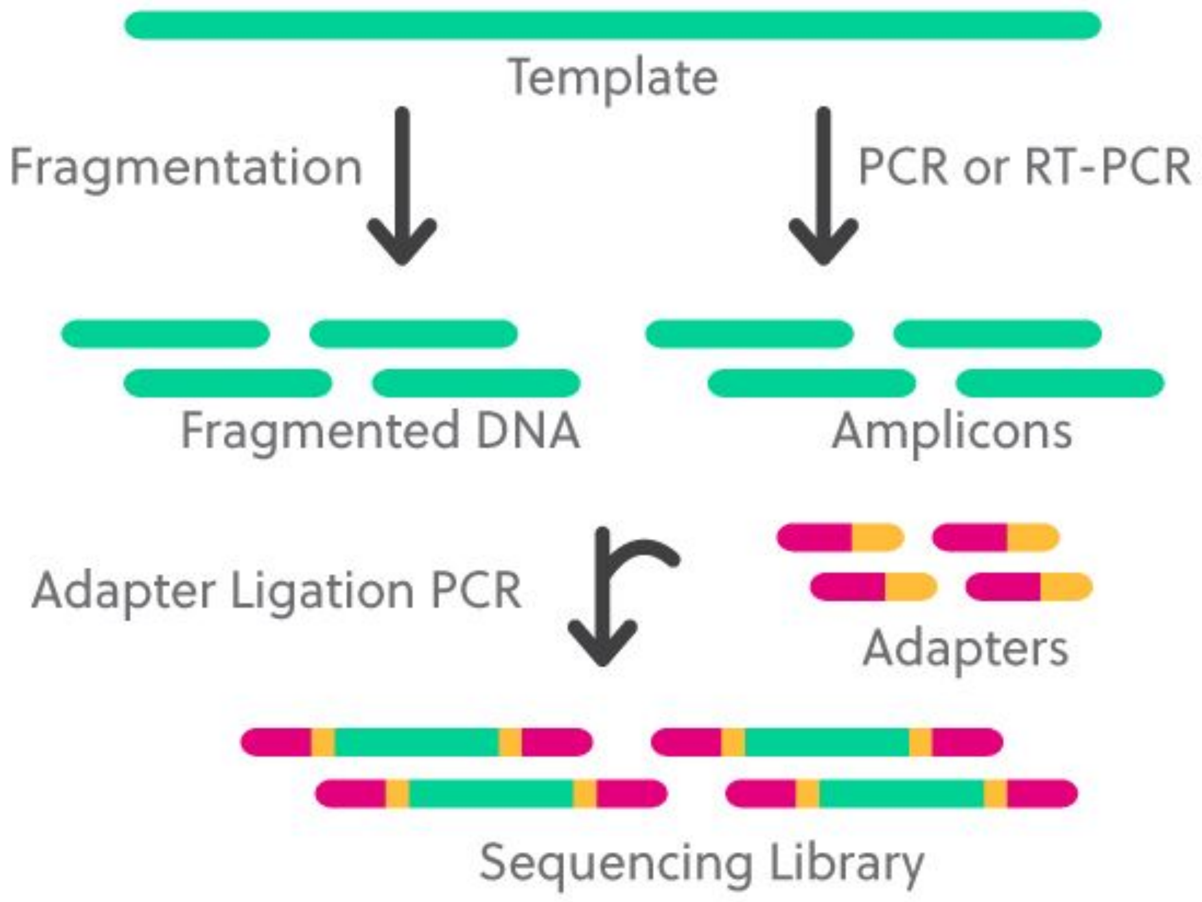


Library preparation

During library prep, we prepare the DNA/RNA molecules to interact with the sequencing machine.

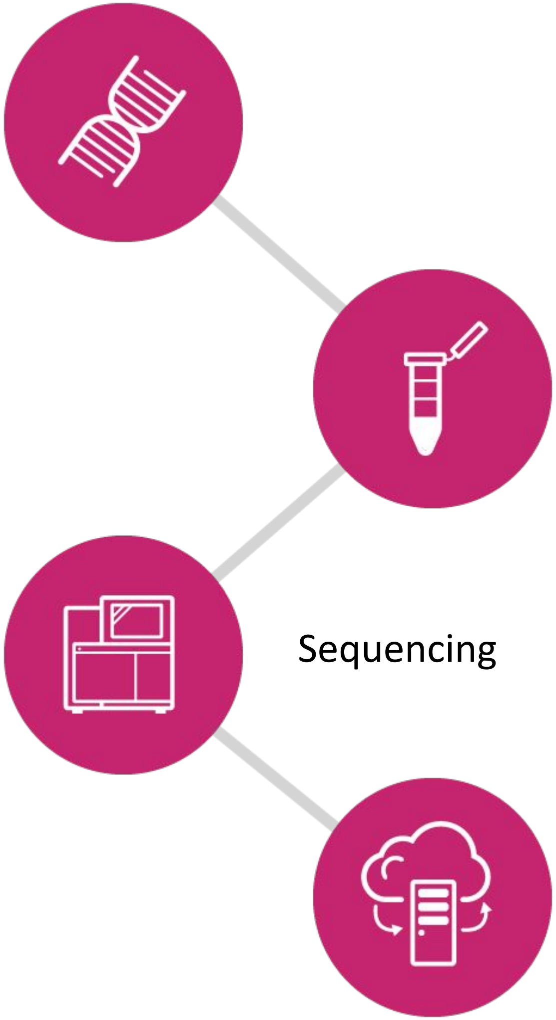
1. Fragmentation
2. Adapter Ligation
3. Size selection (optional)
4. Amplification
5. Clean up

Data Generation - Library Preparation



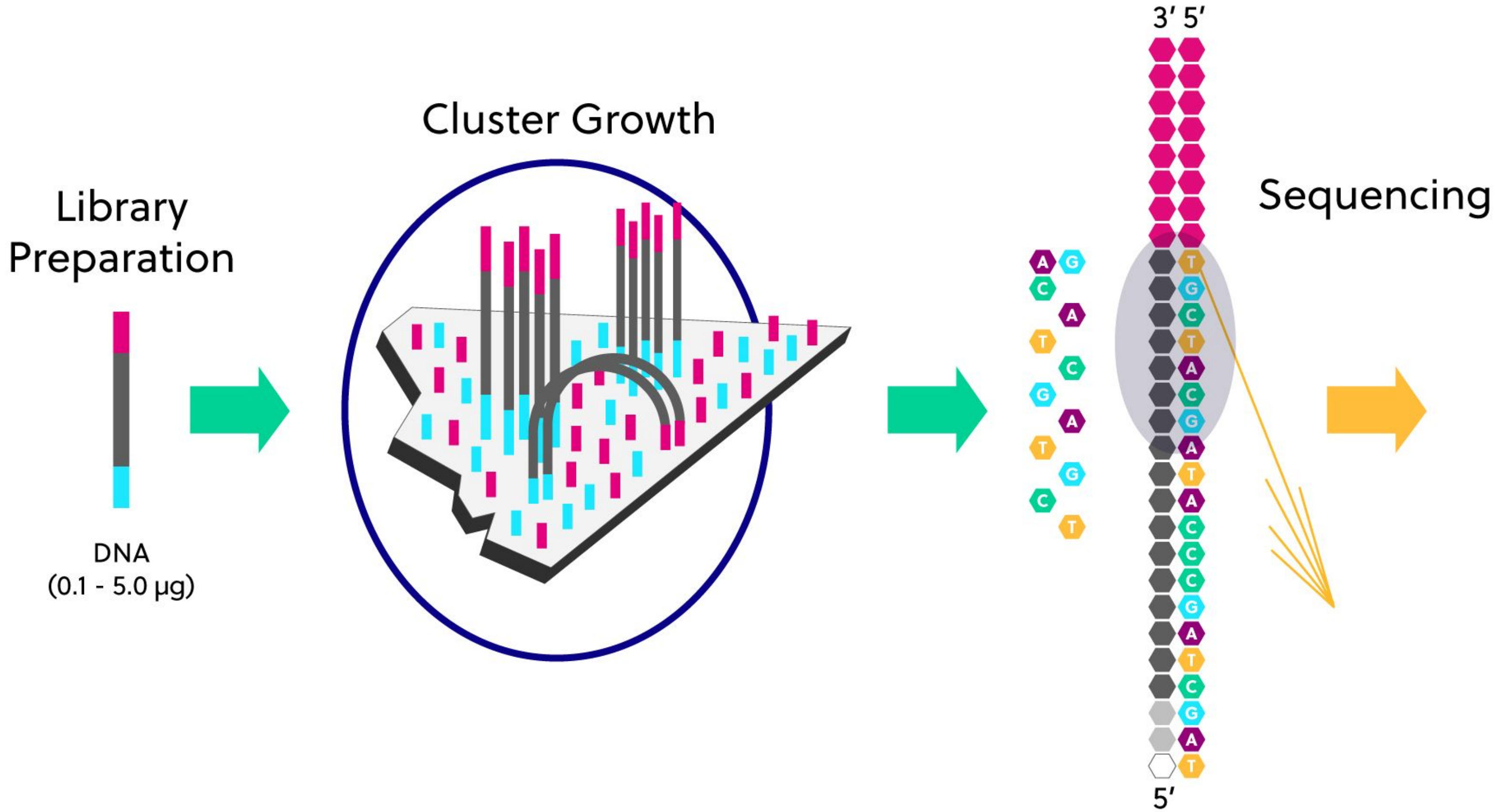
McGill initiative in Computational Medicine

Data Generation - Sequencing



The library is loaded onto the sequencer which then “reads” the nucleotides one by one.

Data Generation - Sequencing



Sequence repositories

-

Genome Sequence Archive

European Nucleotide Archive

National Library of Medicine
National Center for Biotechnology Information

Sequence Read Archive

DDBJ Sequence Read Archive (DRA)

Polls: True or False!

Part II: NGS Quality Control

2

NGS data quality control and preprocessing

- Assess short read FASTQ quality using FastQC
- Perform quality correction with Cutadapt (short reads)
- Summarise quality metrics using MultiQC
- Process single-end and paired-end data

Why Quality Control?

- No sequencing technology is perfect
- Different types and amounts of errors associated with each instrument.

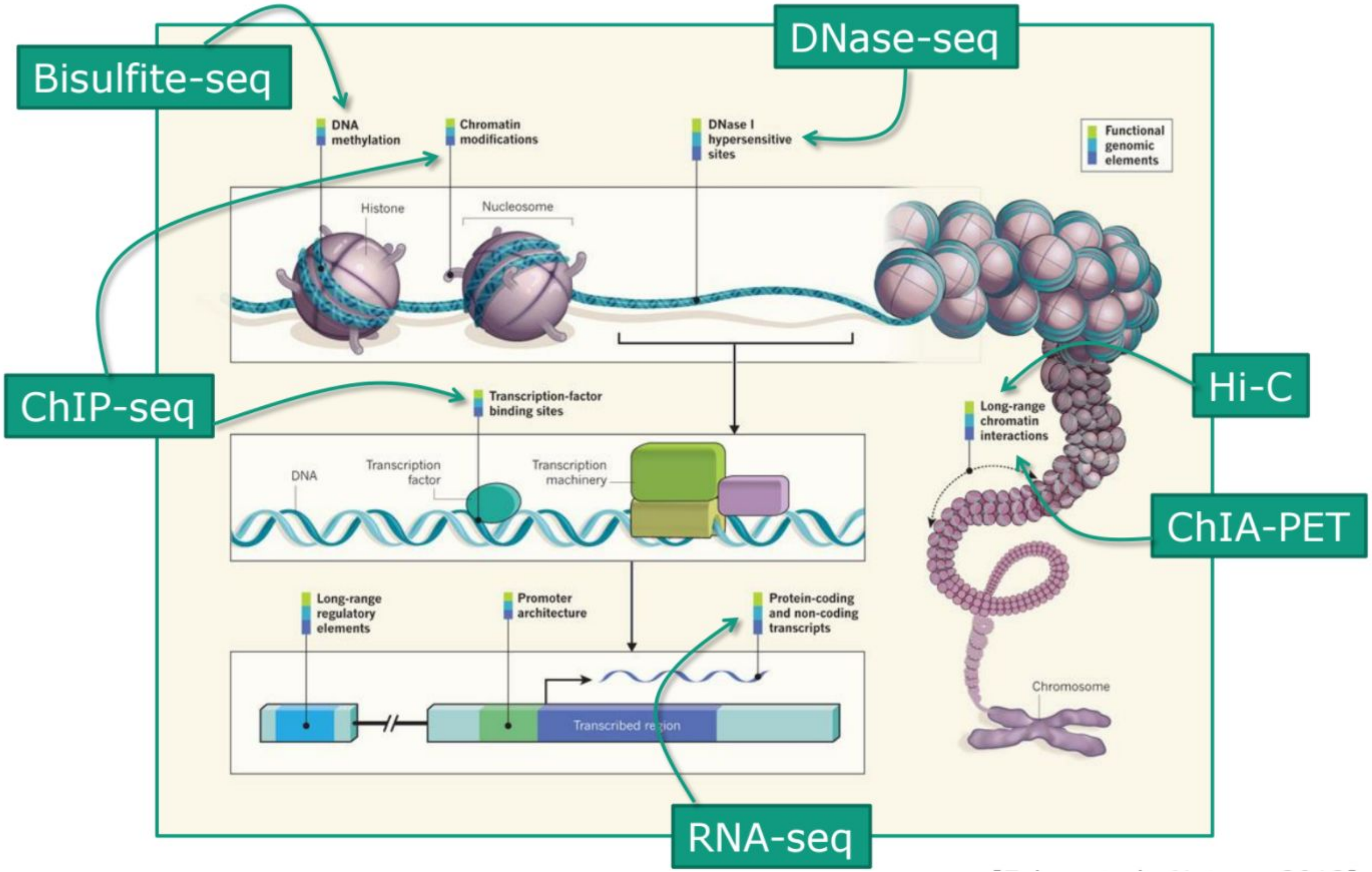
- QC is the process of removing low quality sequences and adaptors that might corrupt downstream analyses

Why Quality Control?

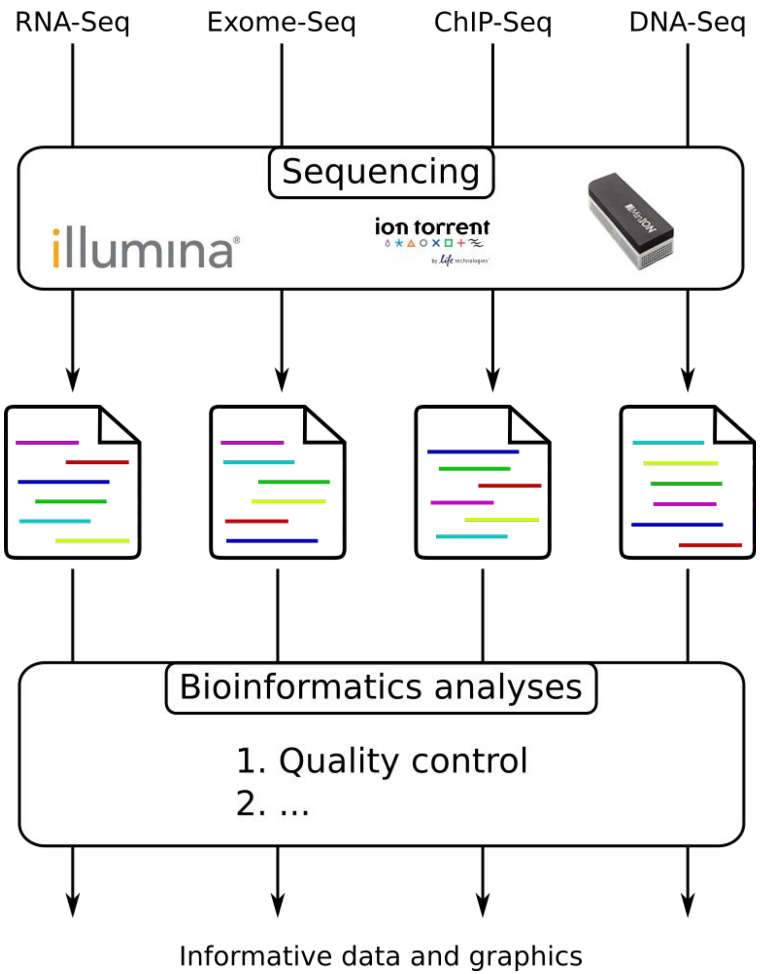
- Are the generated sequences conform to the expected level of performance?
    - Size
    - Number of reads
    - Quality
- Residual presence of adapters or indexes ?
- Are there (un)expected technical biases?
- Are there (un)expected biological biases?

Quality control without context leads to misinterpretation

Where is the data coming from?



From experiments to data



Quality control = First step of the bioinformatics analyses

McGill initiative in Computational Medicine

Sequence file formats

Fasta Format

>Identifier1 (comment)

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

>Identifier2 (comment)

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

XX

Fasta Format

```
>KRN06561.1 heat shock [Lactobacillus sucicola DSM 21376 = JCM 15457]
MSLVMANELTNRFNNWMKQDDFFGNLGRSFFDLDNSVNRALKTDVKETDKAYEVRIDVPGIDKKDITVDY
HDGVLSVNAKRDSFNDESDSEGNVIASERSYGRFARQYSLPNVDESGIKAKCEDGVLKLTLPKLAEEKIN
GNHIEIE
>3HHU_A Chain A, Human Heat-Shock Protein 90 (Hsp90)
MPEETQTQDQPMEEEEVETFAFQAEIAQLMSLIINTFYSNKEIFLRELISNSSDALDKIRYESLTDPSKL
DSGKELHINLIPNKQDRTLTIVDTGIGMTKADLINNLGTIAKSGTKAFMEALQAGADISMIGQFGVGFYS
AYLVAEKVTVITKHNDDEQYAWESSAGGSFTVRTDTGEPMGRGTKVILHLKEDQTEYLEERRIKEIVKKH
SQFIGYPITLFVEK
>KX580312.1 Homo sapiens truncated breast cancer 1 (BRCA1) gene, exon 15
and partial cds
GTCATCCCCTTCTAAATGCCCATCATTAGATGATAGGTGGTACATGCACAGTTGCTCTGGGAGTCTTCAG
AATAGAAACTACCCATCTCAAGAGGAGCTCATTAAGGTTGTTGATGTGGAGGAGTAACAGCTGGAAGAGT
CTGGGCCACACGATTTGACGGAAACATCTTACTTGCCAAGGCAAGATCTAG
```

FastQ Format

@Identifier1 (comment)

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XX

+

QQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQ
QQQQQQQQQQQQQQ

@Identifier2 (comment)

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXX

+

QQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQ
QQQQQQQQQQQQQQQQQQQQQQQQQ

FastQ Format

@ + identifier

Comments

sequence

Quality Score characters

```
1 @SRR031716.1 HWI-EAS299_4_30M2BAAXX:3:1:944:1798
2 AAAAAAAATACAAAAAAAAAAAAAAACCGGAAAAGTA
3 +SRR031716.1 HWI-EAS299_4_30M2BAAXX:3:1:944:1798
4 IIIIIIIIIIIIIIIIIIIIIIIII/C+I-III/=((
  @SRR031716.2 HWI-EAS299_4_30M2BAAXX:3:1:1768:1023
  TGGAAACTTGTCATGCATTGATTTCATCAAGATTTCG
  +SRR031716.2 HWI-EAS299_4_30M2BAAXX:3:1:1768:1023
  IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIDIIIII
```

Phred quality Score

```
                                 33                       59    64       73                               104                      126
                                 |                         |     |        |                                 |                        |
             ASCII code  !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
                                 |                         |     |        |                                 |                        |
Sanger       Phred+33    0........................26...31.......40                                          |                        |
Solexa       Solexa+64   |                        -5....0........9.............................40                                   |
Illumina 1.3+ Phred+64   |                         |     0........9............................40                                    |
Illumina 1.5+ Phred+64   |                         |     |   3.....9...........................41                                    |
Illumina 1.8+ Phred+33   0.2......................26...31........41                                         |                        |
PacBio       Phred+33    0....................20........30........40........50..........................................93
```

Measure of the quality of the identification of the nucleobases

generated by automated DNA sequencing

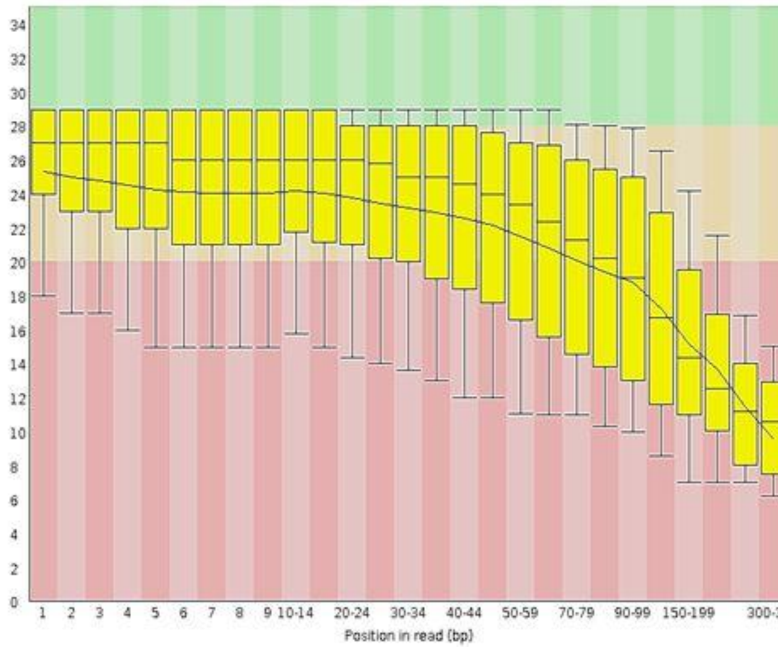
Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

What is a good quality read?

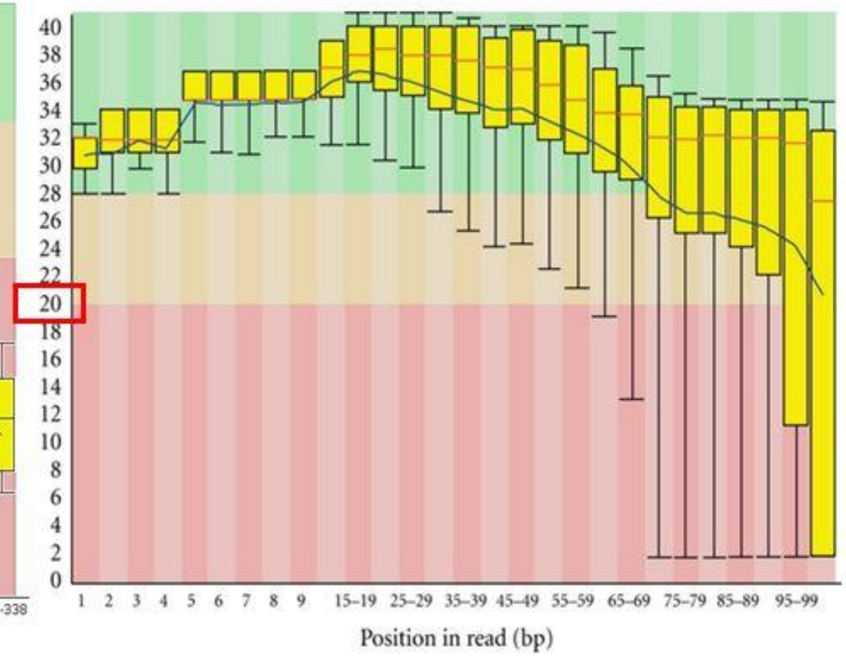
Different technologies *mean* ***different quality***

Comparison in sequencing quality

Ion Torrent PGM

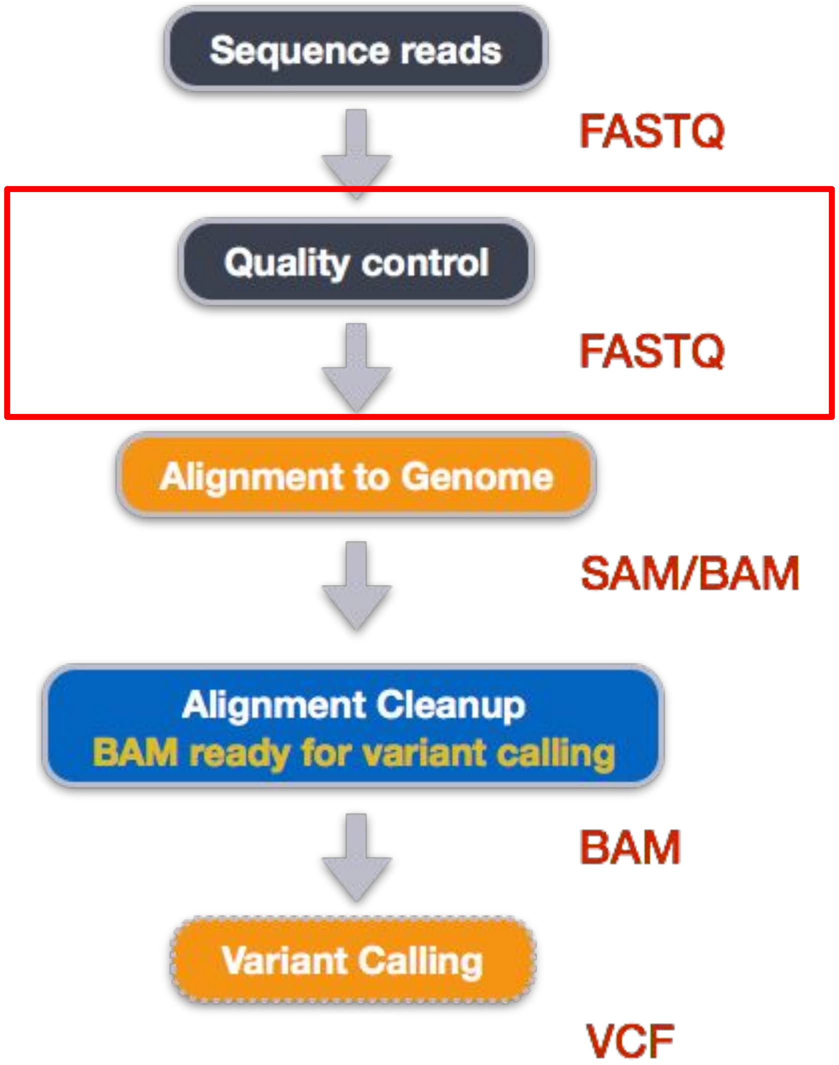


Illumina Hiseq 2000



Per base sequence quality of samples generated by FASTQC. The yellow box show the base-calling quality scores across all sequencing reads. The blue line indicates the mean quality score. Q20=99% accuracy. Q30=99.9% accuracy…

Example NGS pipeline



Assessing Quality of reads

FastQC → A tool for short and long reads quality control

FastQC

FastQC Report

Mon 6 May 2019
MCL1-DK.gz

Summary

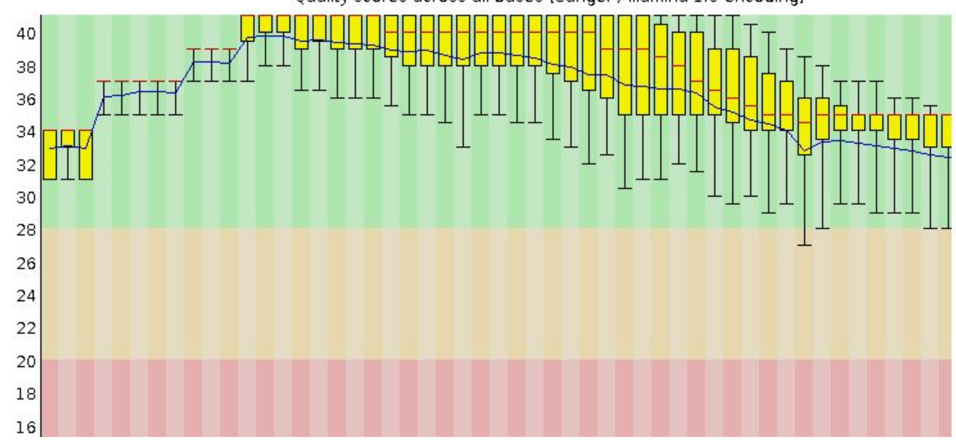
- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

Basic Statistics

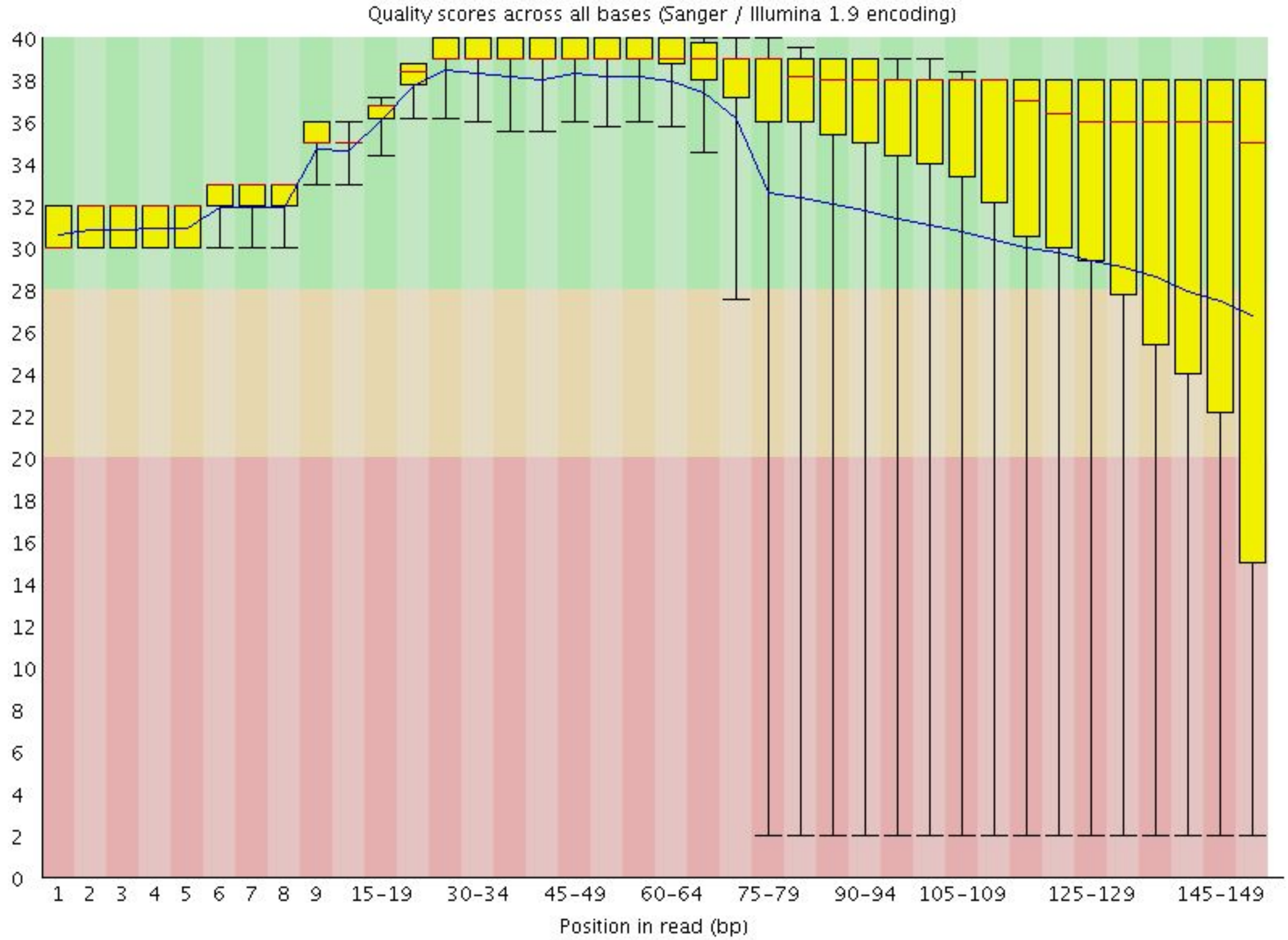
Measure	Value
Filename	MCL1-DK.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1000
Sequences flagged as poor quality	0
Sequence length	100
%GC	50

Per base sequence quality

Quality scores across all bases (Sanger / Illumina 1.9 encoding)

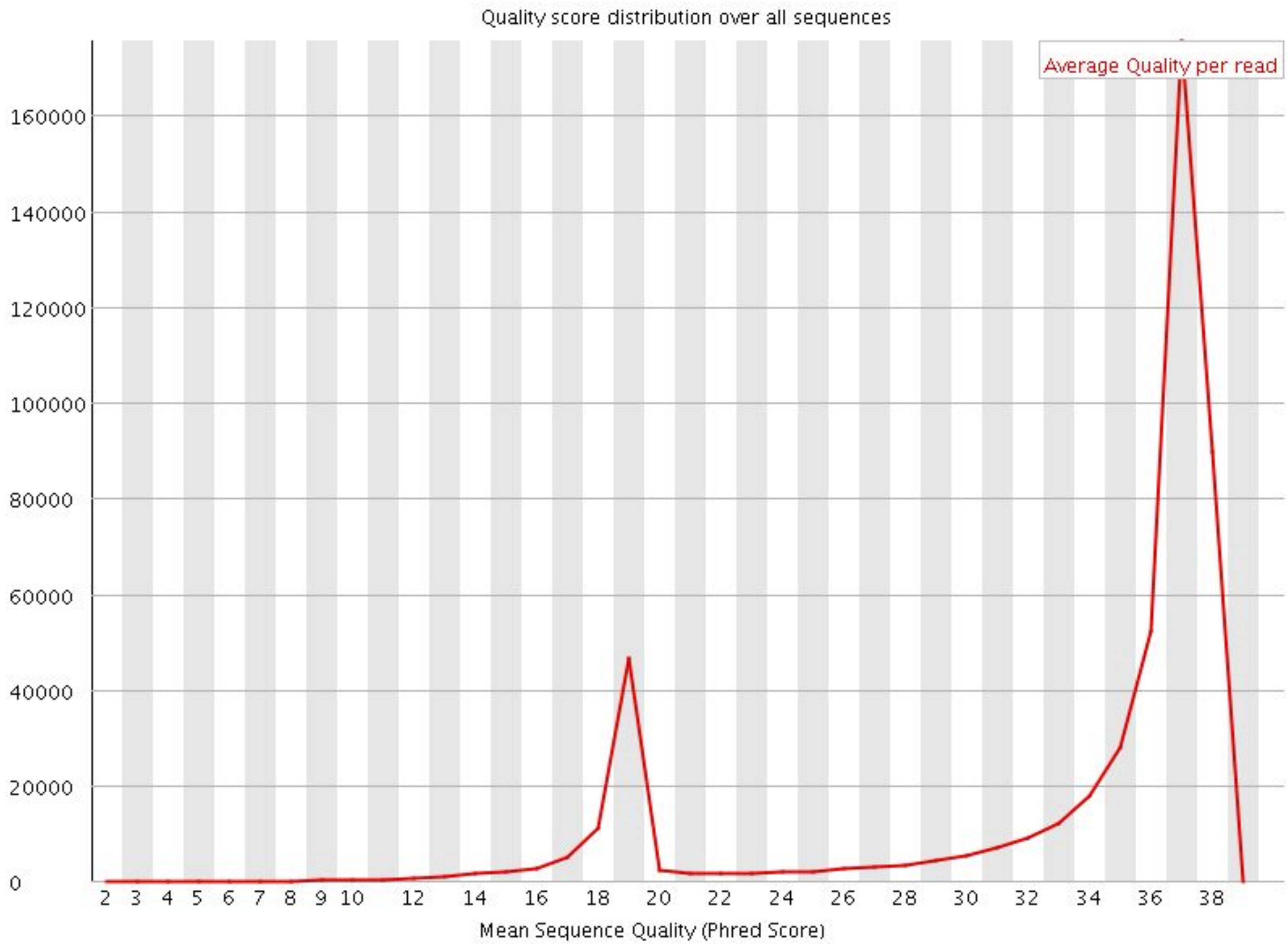


Per base sequence quality



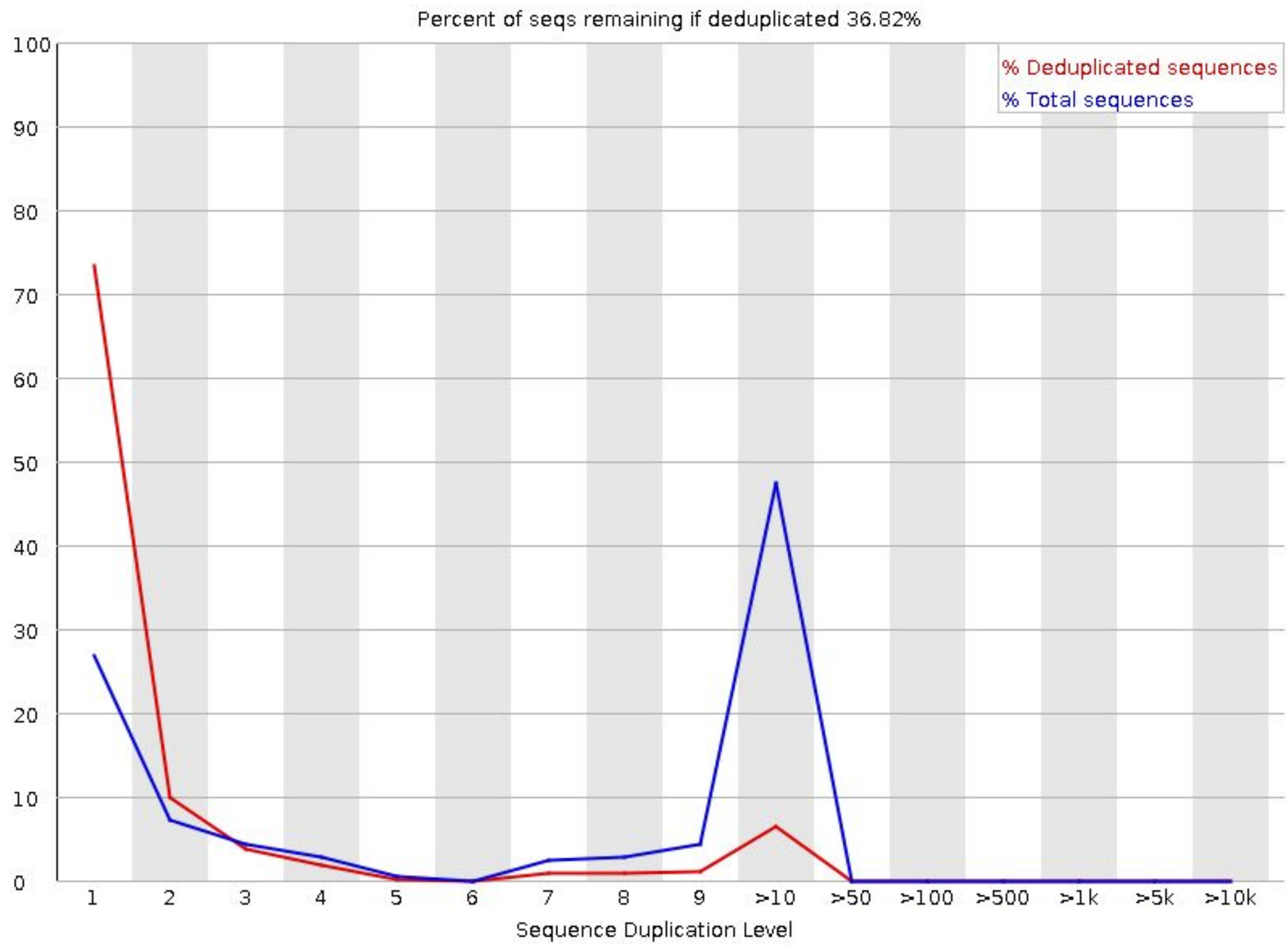
McGill initiative in Computational Medicine

Per sequence quality scores



McGill initiative in Computational Medicine

Duplication Levels



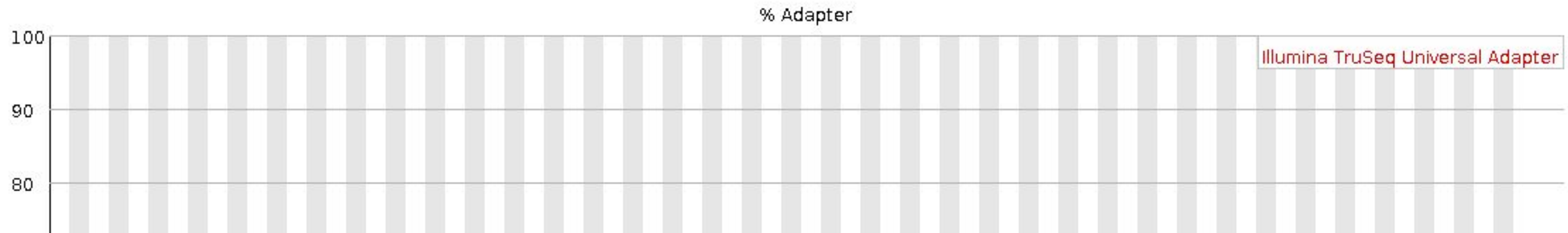
Sources of duplicate reads

Two sources of duplicate reads can be found:

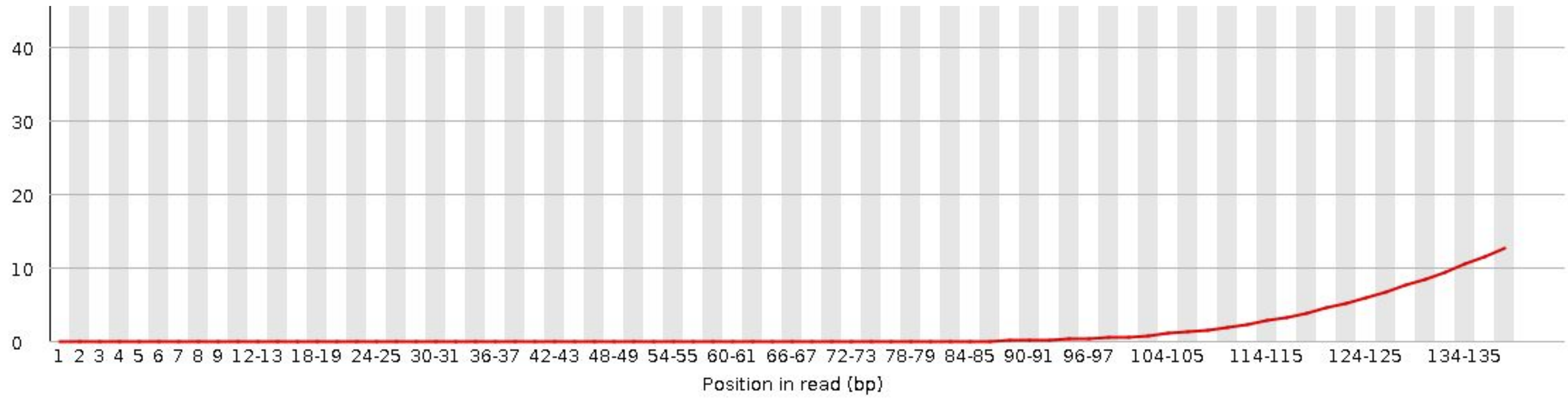
- PCR duplication in which library fragments have been over-represented due to biased PCR enrichment

- Truly over-represented sequences such as very abundant transcripts in an RNA-Seq library or in amplicon data (like in the previous plot)

Overrepresented sequences and Adapter content



Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCACACGTCTGAACTCCAGTCACAGCGCTATCTCGTATGC	66058	13.2116	TruSeq Adapter, Index 1 (97% over 36bp)

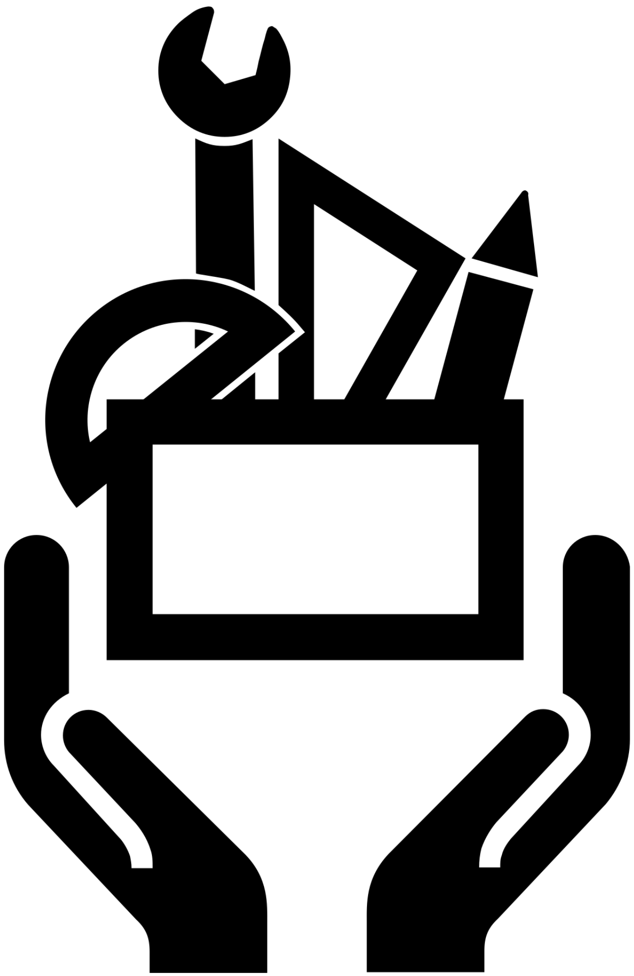


McGill initiative in Computational Medicine

Toolbox: Quality Control

Different software to asses QC in fastqc files!

- FastQC
- fastx-toolkit
- NGS QC toolkit
- Nanoplot * long reads only
- PycoQC * Nanopore only
- FastQe



5 minute Break?

Hands on: Assessing Quality Control

Hands on 1: Quality control

Download the data set and run FastQC on our Fastq files, inspect the output.

1. How does the mean quality score change along the sequence?
2. When comparing Forward and Reverse reads, what do you notice?
3. Given the Quality of our reads what should we do next?

Improving Quality

How can we improve the quality of of reads?

- Filtering of sequences
  - with small mean quality score
  - that are too small
  - with too many N bases
  - based on their GC content

- Cutting/Trimming sequences
  - from low quality score parts
  - tails
  - removing barcode and adapter sequence

Which trimming threshold?

It depends on your initial question, and also on the sequencing depth

RNAseq

- Gentle trimming
- Too aggressive trimming => losing part of the dataset

SNP calling

- You need to be sure of the bases
- Normally Q > 20

MacManes, M. D. (2014). On the optimal trimming of high-throughput mRNA sequence data. *Frontiers in genetics*, *5*, 13.

Toolbox: Improving quality

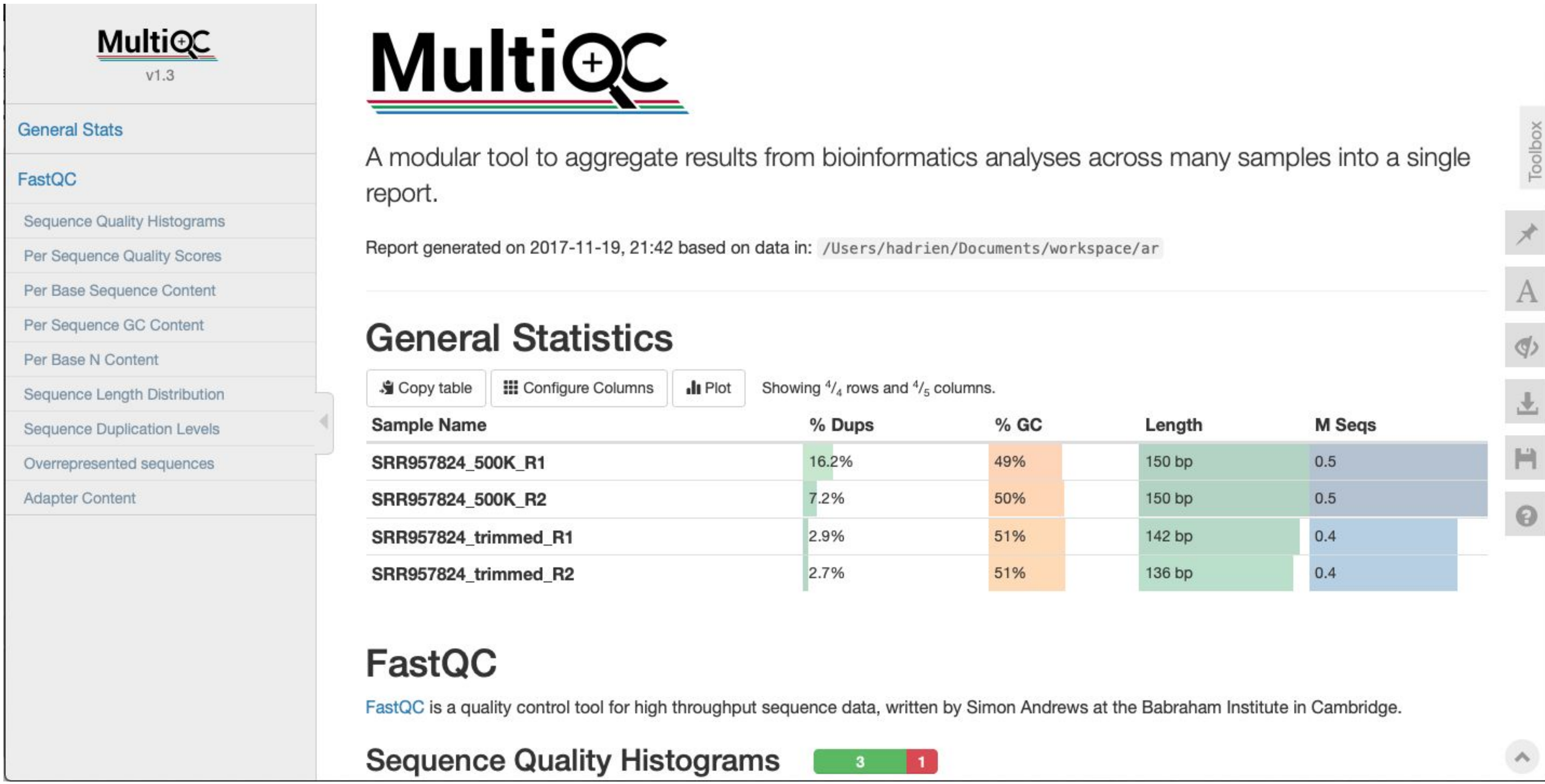
Many different software to trim/filter Fasta/Fastq files!

- Trimmomatic
- **Cutadapt**
- AdapterRemoval
- Scythe
- Sickle
- Atropos
- Fastp
- Trim Galore

Producing QC reports with MultiQC

To aggregate the results of bioinformatics tools across many samples/files

MultiQC


MultiQC
v1.3
General Stats
FastQC
Sequence Quality Histograms
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
MultiQC
A modular tool to aggregate results from bioinformatics analyses across many samples into a single report.
Report generated on 2017-11-19, 21:42 based on data in: /Users/hadrien/Documents/workspace/ar
General Statistics
Copy table
Configure Columns
Plot
Showing 4/4 rows and 4/5 columns.
Sample Name
% Dups
% GC
Length
M Seqs
SRR957824_500K_R1 16.2% 49% 150 bp 0.5
SRR957824_500K_R2 7.2% 50% 150 bp 0.5
SRR957824_trimmed_R1 2.9% 51% 142 bp 0.4
SRR957824_trimmed_R2 2.7% 51% 136 bp 0.4
FastQC
FastQC is a quality control tool for high throughput sequence data, written by Simon Andrews at the Babraham Institute in Cambridge.
Sequence Quality Histograms
3
1
Toolbox

Hands on: Improving the quality

Hands on 2: Improving our sequence quality

Use Cutadapt to remove adapters and low-quality ends.

1. Inspect the log file. How many times did the adapter get removed from the Forward reads?
2. Why do you think we need to run cutadapt in a paired-end mode instead of independently for each file?

Hands on 2: Improving our sequence quality

Run FastQC on the trimmed reads and aggregate the results using multiqc to compare the before and after. Open the multiQC report.

1. What differences do you notice?
2. Do you think trimming improved the overall quality in other aspects apart from adapter content and quality score?

Key points!

- Know your data first! Different techniques will yield different results in the FastQC plots, just check if it fits what you were expecting.
- Not because FastQC gives a warning it means you need to throw away your data.
- Run QC on every sequencing dataset before any other analysis.
- Assess the quality metrics and improve quality if necessary.
- You will not always want to trim based on quality. This will also depend on your type and amount of data and the downstream analyses you want to do.
- Re-run FastQC after filtering/trimming to assess the impact of the quality control step.
- For paired-end reads analyze the Fwd and Rv reads together.

10 minute Break

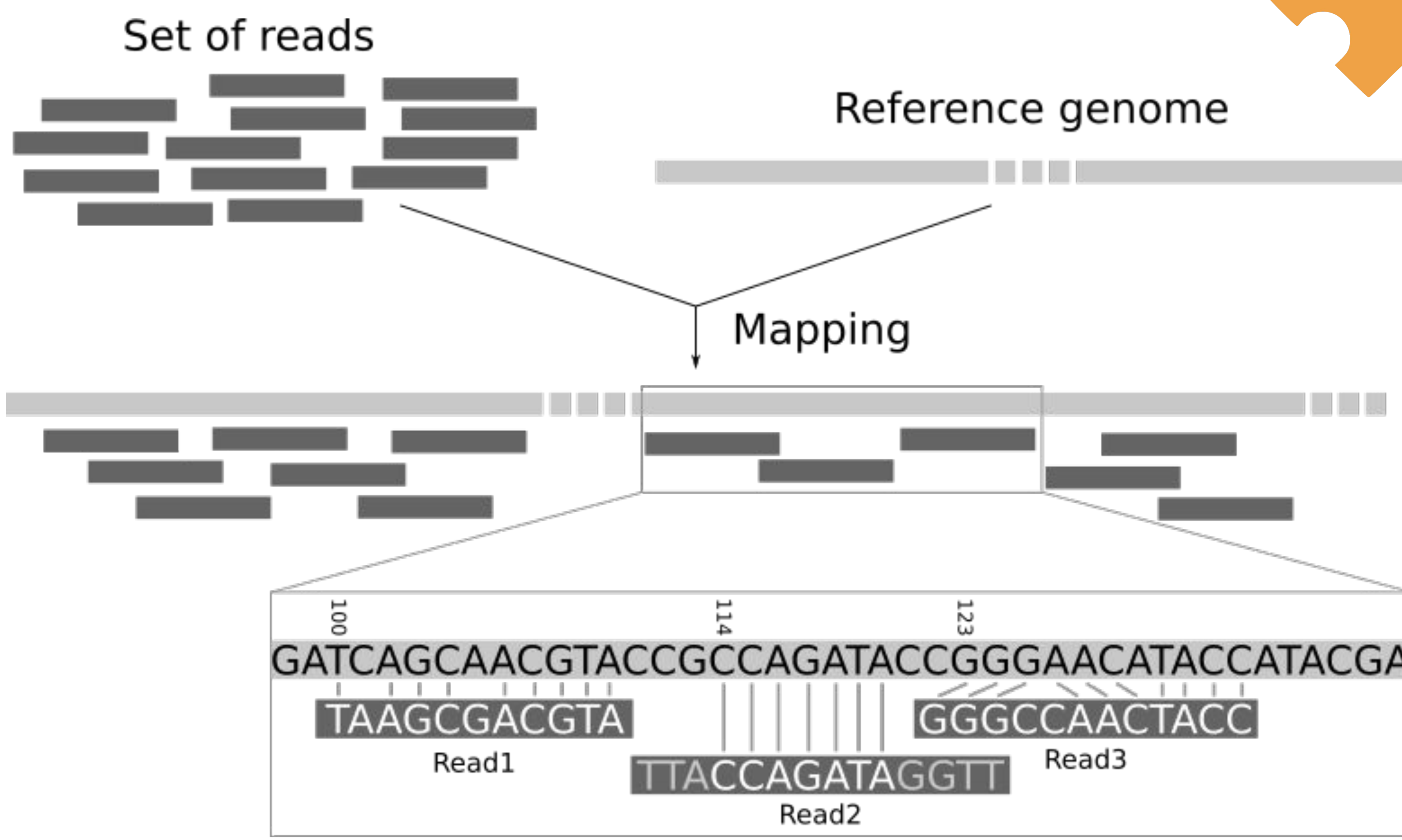
Part III: Mapping to a Reference

3

Mapping and manipulation of alignment files

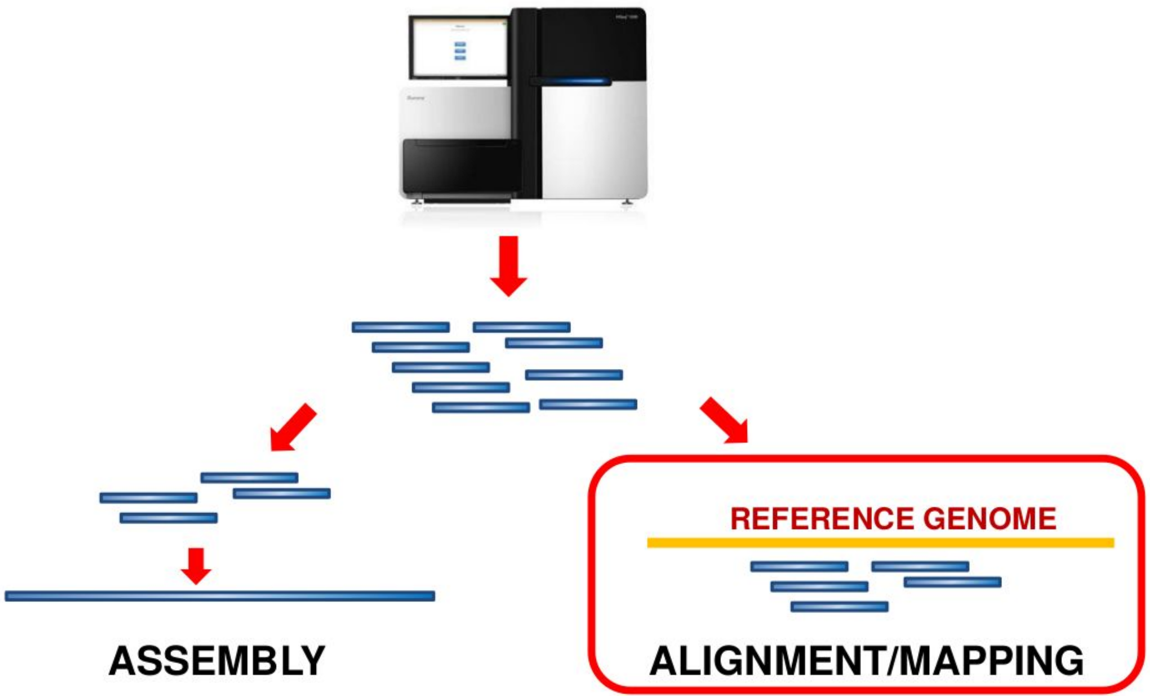
- The concept of mapping
- BAM/SAM formats
- Find a reference genome
- Choosing an alingner
- Conduct the mapping to the reference

What is mapping?



McGill initiative in Computational Medicine

Mapping vs Assembly



- **Mapping:** use a reference genome as a guide
- **De-novo assembly:** without reference genome

Mapping vs Assembly

- De novo assembly focuses on reconstructing the original sequence by aligning and merging reads.

- Mapping is used when you want to assemble sequences to a known reference. The main objective is to know the genome position each read comes from and it is used when you want to spot differences between the sequenced sample and the reference (for example, SNPS)

Sequence Alignment

- Determine position of a short read on a reference genome

```
Reference: . . . A A C G C C T T . . .
Read:            A G G G G C C T T
```

McGill initiative in Computational Medicine

Sequence Alignment

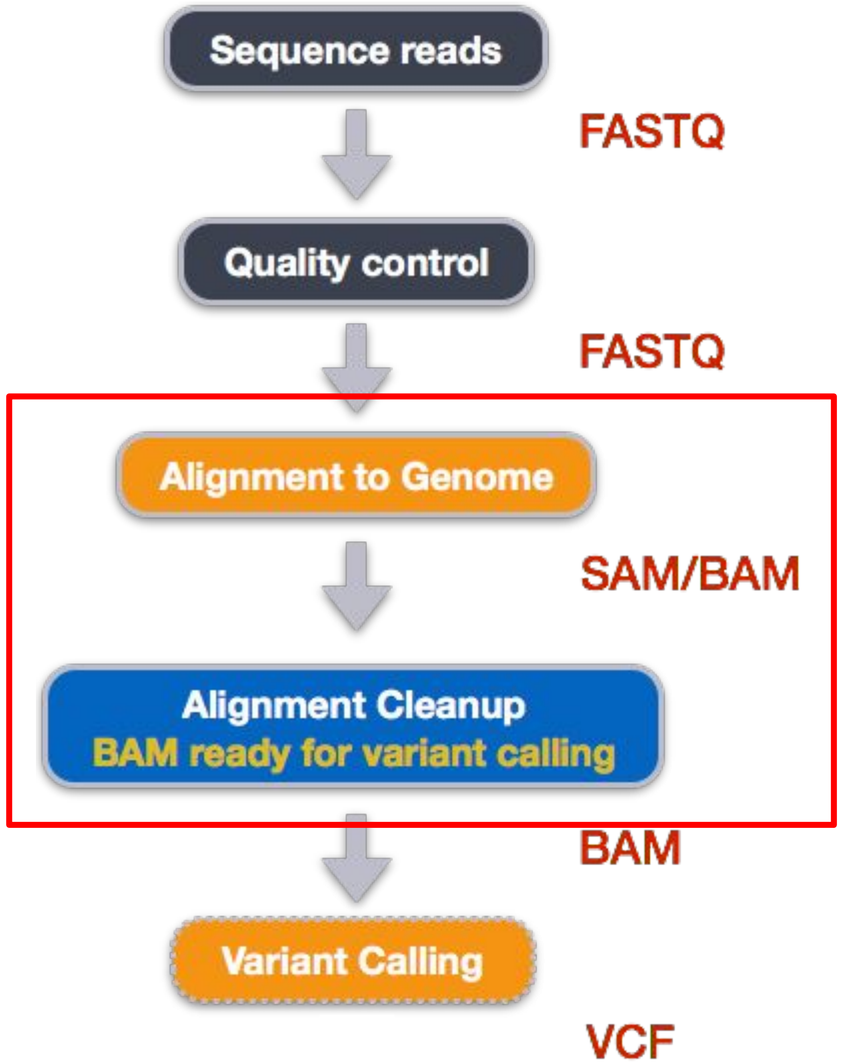
- Determine position of a short read on a reference genome

```
Reference: . . . A A - C G C C T T . . .          | = match
.                | : - : | | | | |                : = mismatch
Read:            A G G G G C C T T                - = gap
```

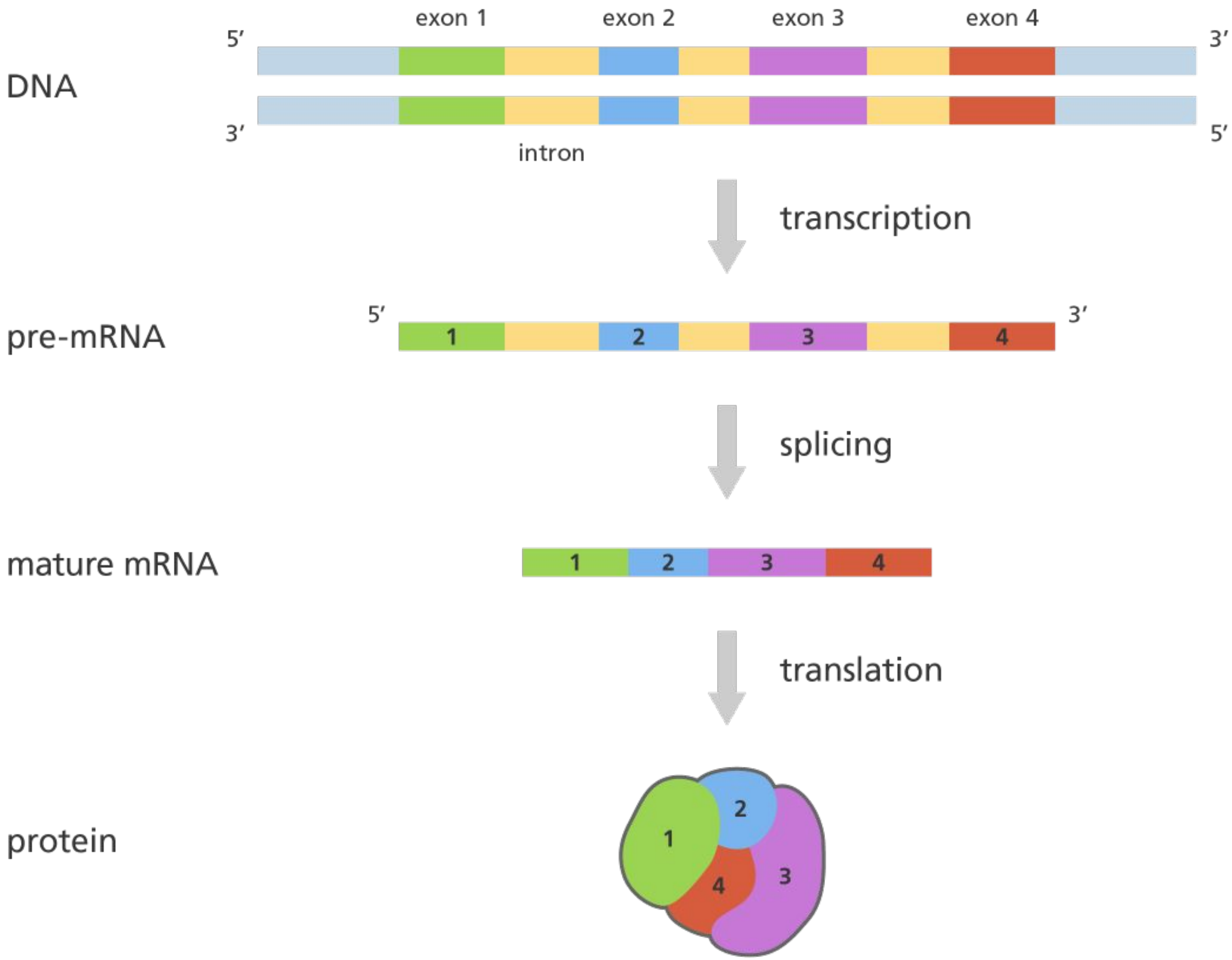
In order to make the most sense of the mapping, aligners can introduce gaps. Each position gets assigned as a match, mismatch or indel

- **Match:** Nucleotide at the read same as in reference
- **Mismatch**: Nucleotide at the read different than in reference
- **Gap/indel:** nucleotide that was inserted/deleted from the read or reference

Example NGS pipeline

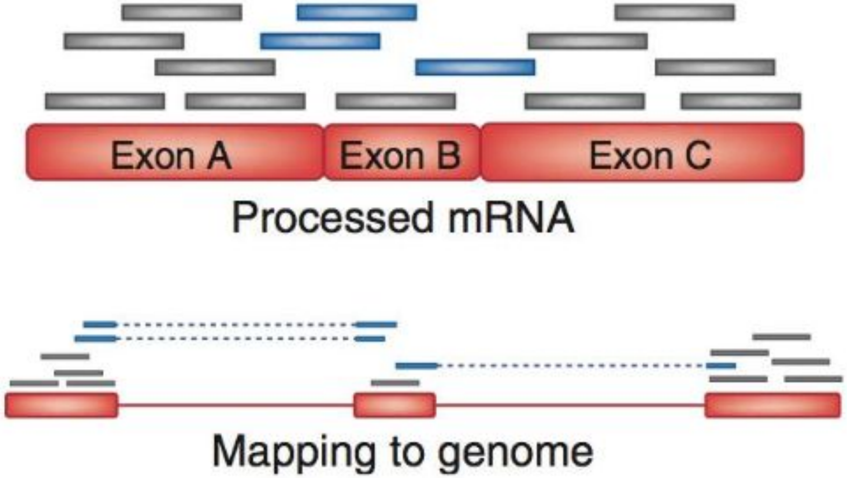


RNASeq and Splicing

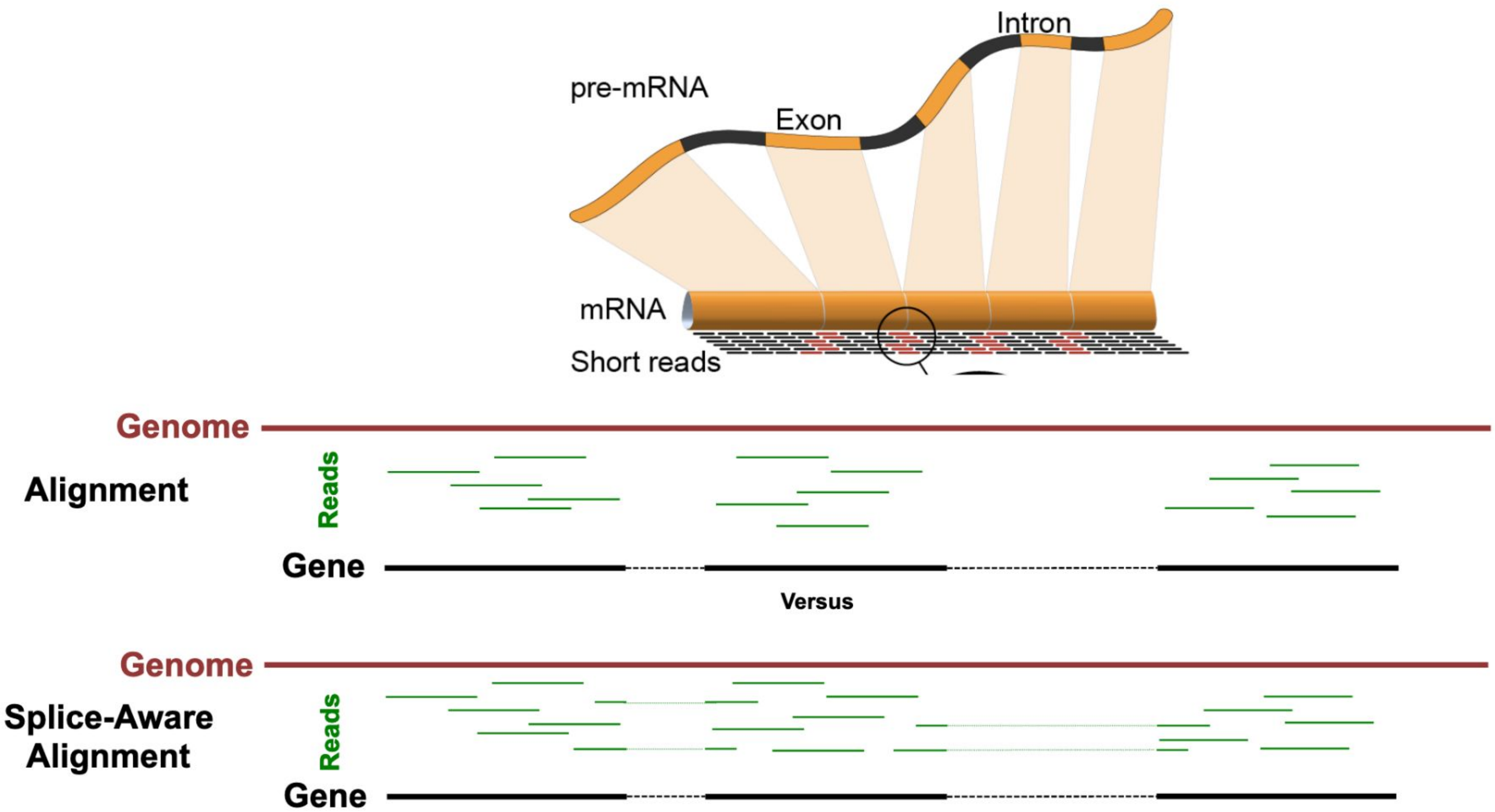


Choosing an Aligner

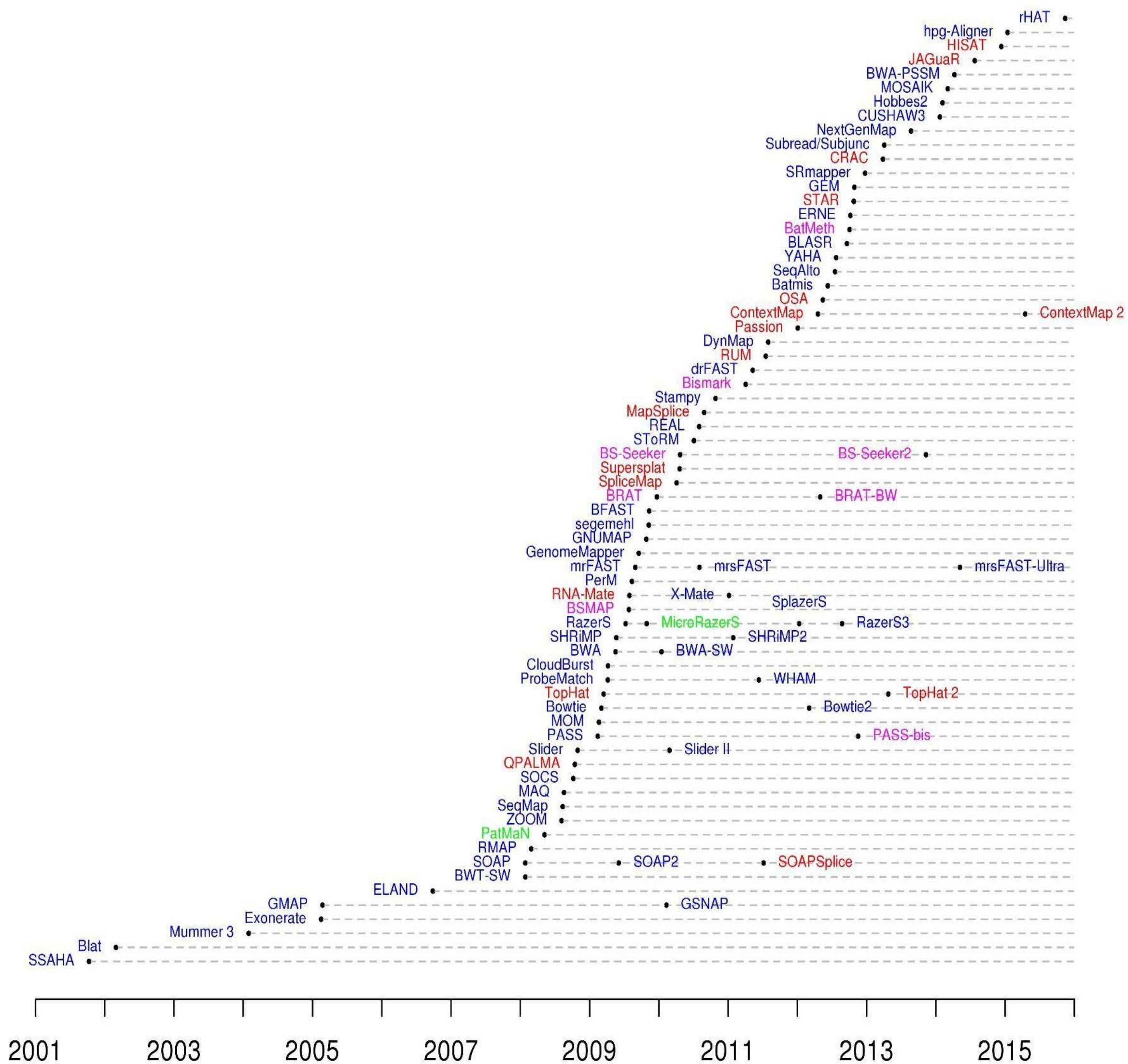
- Each tool makes **different choices** during alignment
  - Choice of aligner may **affect downstream results**
  - Default options may not be best for your data!
- Best tool for your data **depends on many factors**
  - Type of experiment (e.g. DNA, RNA, Bisulphite)
  - Sequencing platform
  - Compute resources vs sensitivity
  - Read characteristics (paired/single end, read length)



Splice-aware vs Splice unaware



Mapping Tools



60+ different mappers, many comparison papers. Figure from 10.1093/bioinformatics/bts605

McGill initiative in Computational Medicine

Some mapping tools

Mapping tool	Uses	Characteristics
HISAT2	DNA/RNA	Short reads. Based on GCSA. Reference.
RNASTAR	RNA	Short reads. Extremely fast. High sensitive and accuracy. Based on Maximal Mappable Prefixes (MMPs). Reference.
BWA-MEM2	DNA	Short reads. Twice as faster as BWA-MEM. Memory efficient. Based on Burrows-Wheeler. Reference.
Minimap2	DNA/RNA	Long reads (PacBio and ONT). Extremely fast. Based on DALIGN and MHAP. Reference.
Bismark	DNA/RNA	Short reads. Bisulfite treated sequencing. Based on GCSA. Reference.
BBMap	DNA/RNA	Short and long reads (PacBio and ONT). Memory demanding. Reference.
Whisper 2	DNA	Short reads. Indel sensitive. Variant-calling oriented. Reference.
S-conLSH	DNA	Long reads (ONT). High sensitivity and accuracy. Reference.

Alignment file format

SAM/BAM format

```
1:497:R:-272+13M17D24M 113    chr1 497   37         37M        15         100338662 0
CGGGTCTGACCTGAGGAGAACTGTGCTCCGCCTTCAG   0;==-==9;>>>>>=>>>>>>>>>>>>=>>>>>>>>>   XT:A:U     NM:i:0
SM:i:37    AM:i:0     X0:i:1     X1:i:0     XM:i:0     XO:i:0     XG:i:0     MD:Z:37
19:20389:F:275+18M2D19M         99         chr1       17644      0          37M        =          17919
314        TATGACTGCTAATAATACCTACACATGTTAGAACCAT   >>>>>>>>>>>>>>>>>>>><<>>><<>>4::>>:<9
RG:Z:UM0098:1         XT:A:R     NM:i:0     SM:i:0     AM:i:0     X0:i:4     X1:i:0     XM:i:0     XO:i:0
XG:i:0     MD:Z:37
19:20389:F:275+18M2D19M         147        chr1       17919      0          18M2D19M   =
17644      -314       GTAGTACCAACTGTAAGTCCTTATCTTCATACTTTGT   ;44999;499<8<8<<<8<<><<<<><7<;<<<><<
XT:A:R     NM:i:2     SM:i:0     AM:i:0     X0:i:4     X1:i:0     XM:i:0     XO:i:1     XG:i:2     MD:Z:
18^CA19
9:21597+10M2I25M:R:-209         83         chr1       21678      0          8M2I27M    =
21469      -244       CACCACATCACATATACCAAGCCTGGCTGTGTCTTCT   <;9<<5><<<<><<<<><<><>><9>><>>>9>>><>
XT:A:R     NM:i:2     SM:i:0     AM:i:0     X0:i:5     X1:i:0     XM:i:0     XO:i:1     XG:i:2     MD:Z:35
```

SAM: **S**equence **A**lignment **M**ap
BAM: Binary (compressed) SAM

Mandatory Fields

Col	Field	Type	Brief Description
1	QNAME	String	Query template NAME
2	FLAG	Integer	Bitwise FLAG
3	RNAME	String	References sequence NAME
4	POS	Integer	1- based leftmost mapping POSition
5	MAPQ	Integer	MAPping Quality
6	CIGAR	String	CIGAR String
7	RNEXT	String	Ref. name of the mate/next read
8	PNEXT	Integer	Position of the mate/next read
9	TLEN	Integer	Observed Template LENgth
10	SEQ	String	Segment SEQuence
11	QUAL	String	ASCII of Phred-scaled base QUALity+33

- The flag gives important information of your read!

SAM/BAM format

```
@HD VN:1.5 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       * 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
```

Header section

Alignment section

Optional fields in the format of TAG:TYPE:VALUE

QUAL: read quality; * meaning such information is not available

SEQ: read sequence

TLEN: the number of bases covered by the reads from the same fragment. Plus/minus means the current read is the leftmost/rightmost read. E.g. compare first and last lines.

PNEXT: Position of the primary alignment of the NEXT read in the template. Set as 0 when the information is unavailable. It corresponds to POS column.

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.

CIGAR: summary of alignment, e.g. insertion, deletion

MAPQ: mapping quality

POS: 1-based position

RNAME: reference sequence name, e.g. chromosome/transcript id

FLAG: indicates alignment information about the read, e.g. paired, aligned, etc.

QNAME: query template name, aka. read ID

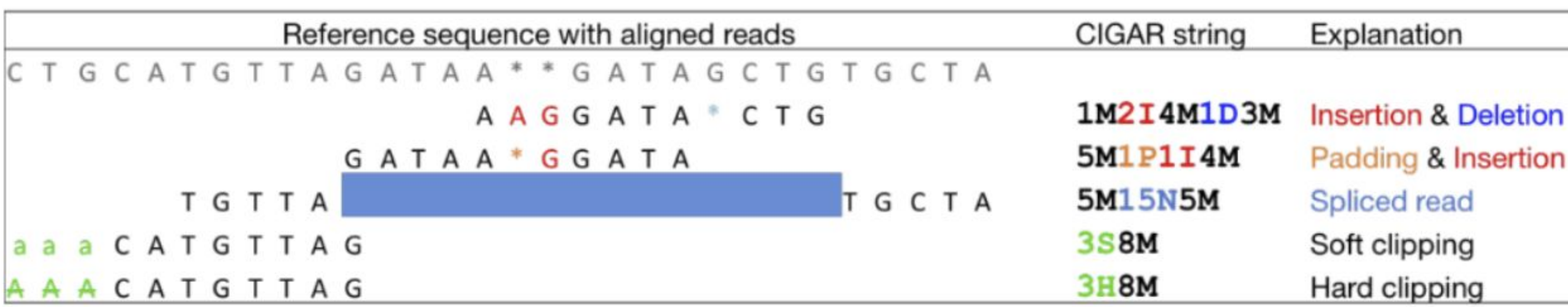
The CIGAR String

`CIGAR` stands for *Concise Idiosyncratic Gapped Alignment Report*. This sixth field of a SAM file contains a so-called CIGAR string indicating which operations were necessary to map the read to the reference sequence at that particular locus.

The following operations are defined in CIGAR format (also see figure below):

- **M** - Alignment (can be a sequence match or mismatch!)
- **I** - Insertion in the read compared to the reference
- **D** - Deletion in the read compared to the reference
- **N** - Skipped region from the reference. For mRNA-to-genome alignments, an N operation represents an intron. For other types of alignments, the interpretation of N is not defined.
- **S** - Soft clipping (clipped sequences are present in read); S may only have H operations between them and the ends of the string
- **H** - Hard clipping (clipped sequences are NOT present in the alignment record); can only be present as the first and/or last operation
- **P** - Padding (silent deletion from padded reference)
- **=** - Sequence match (not widely used)
- **X** - Sequence mismatch (not widely used)

The sum of lengths of the **M**, **I**, **S**, **=**, **X** operations must equal the length of the read. Here are some examples:



Reference Genomes

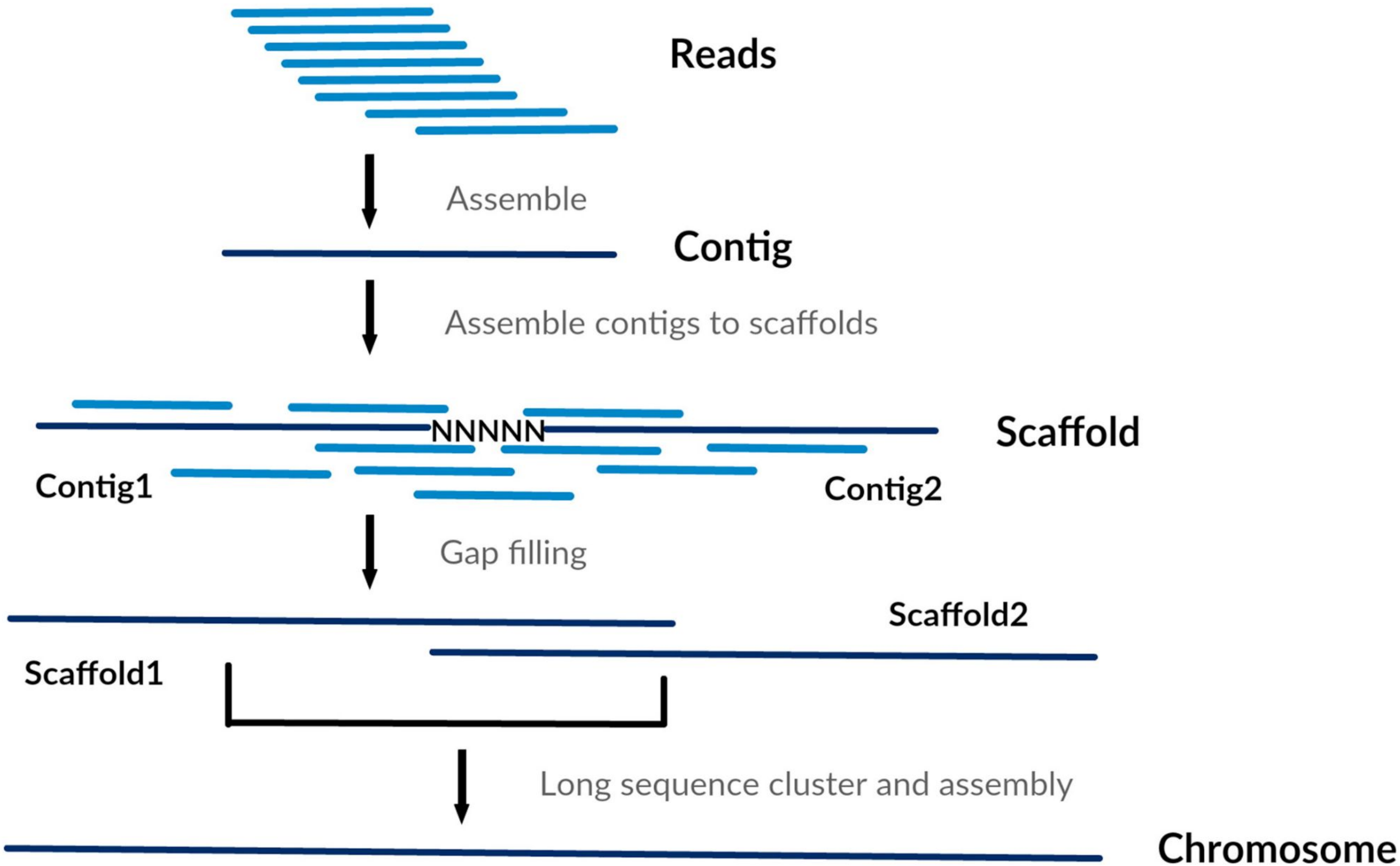
What is a Reference?

It can be everything containing DNA information:

- Complete genome
- Assembly
- Set of contigs
- Set of sequences
- Genes, non-coding RNA...

For mapping, references have to be stored in a FASTA file.

Contigs → Scaffolds → Chromosomes



(Guo et al., 2017)

Where can you find reference genomes/assemblies?

e! Ensembl

UCSC Genome Browser

ENA
European Nucleotide Archive

NCBI

NIH National Library of Medicine
National Center for Biotechnology Information

GenBank

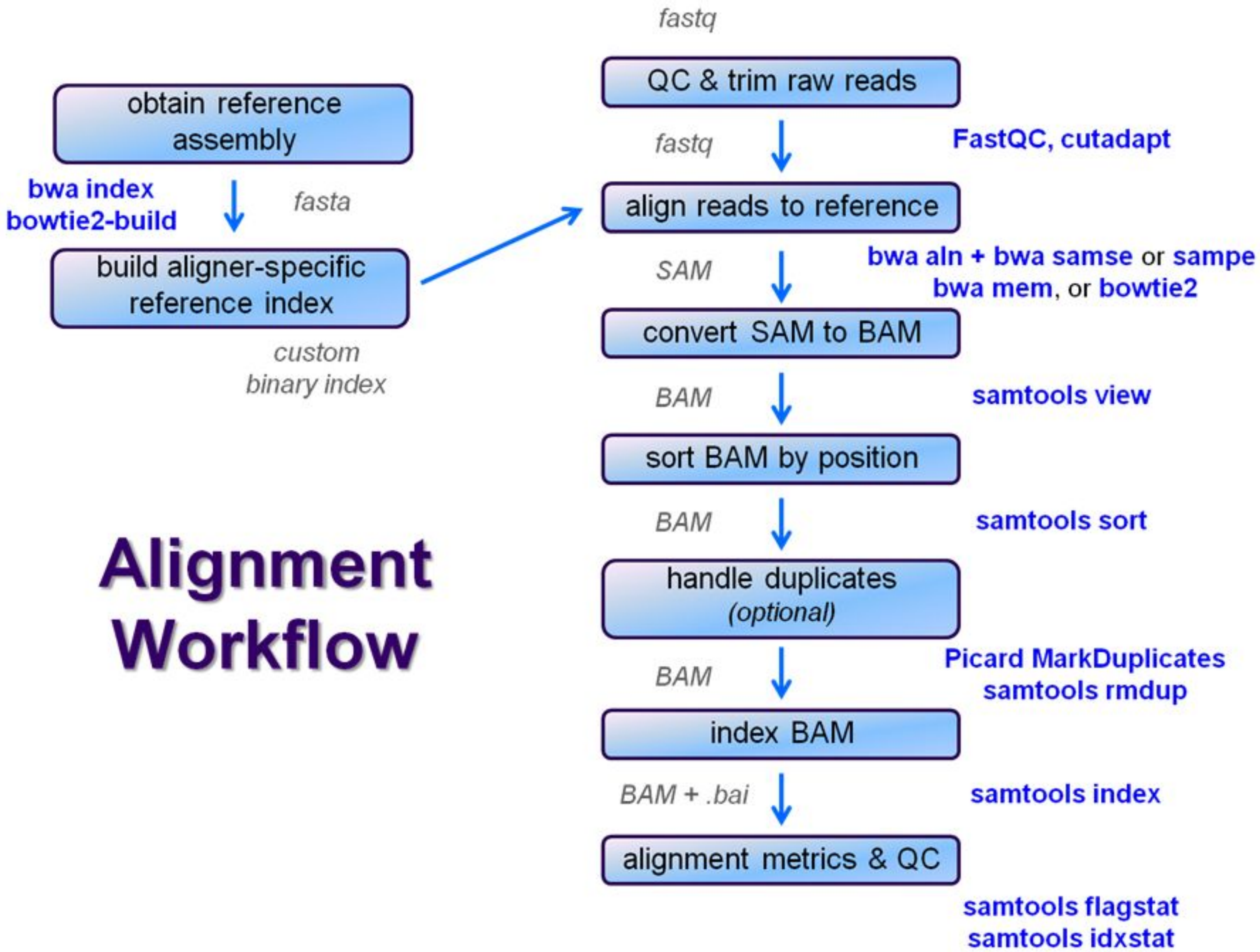
RefSeq

Some common assembly statistics

- Level of assembly: Chromosome, scaffold or contig
- # of bases (bp) - total length of the assembly
- Number of contigs/scaffolds
- Average contig length/size
- Longest contig size
- **N50** - the sequence length of the shortest contig at 50% of the total genome length.
- **L50** - count of smallest number of contigs whose length sum makes up half of genome size.

Sequence mapping Workflow

Alignment Workflow



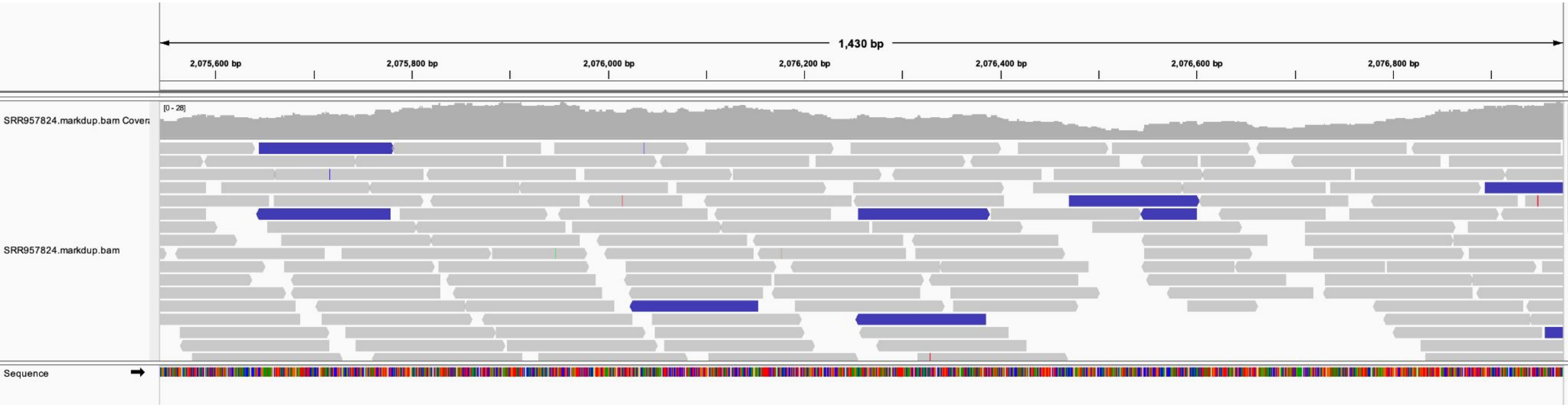
5 minute Break?

Hands on: Mapping reads to a Reference Genome

Visualizing alignments with IGV

Visualizing alignments with IGV

- IGV is a graphic and interactive tool to visualize genomic data. Not just bams, also annotations, variant calls, peaks, and more



McGill initiative in Computational Medicine

Today we learned

1

What is NGS and the steps to produce the data
Databases to find and deposit raw sequence data

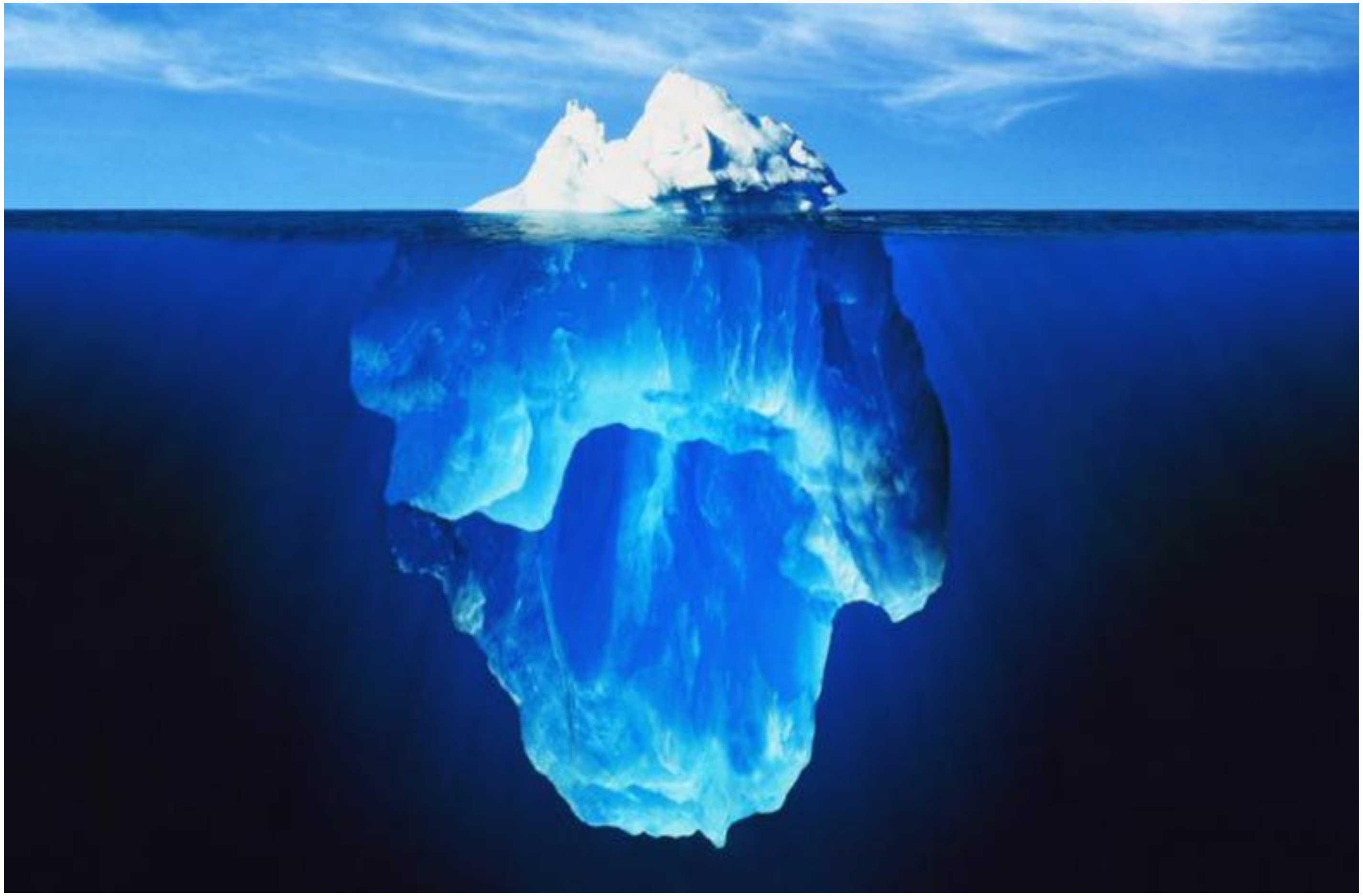
2

Sequence file formats fasta/fastq
How to read and interpret QC plots
How to remove adapters and trim by quality
How to aggregate QC results in one report

3

Differences between mapping and assembling
Different types of aligners
Aligned sequence format SAM/BAM
What is a reference genome and where to find them.
Map and handle bams using bowtie2 and samtools
Post-alignment processing and filtering.
Visualize an alignment

This is really just the tip of the iceberg!



Thanks for your attention!

Keep an eye for the workshops offered by the MiCM!

maria.femerlingromero@mail.mcgill.ca
workshop-micm@mcgill.ca