Addis Ababa University
አዲስ አበባ ዩኒቨርሲቲ
SEEK WISDOM, ELEVATE YOUR INTELLECT AND SERVE HUMANITY !

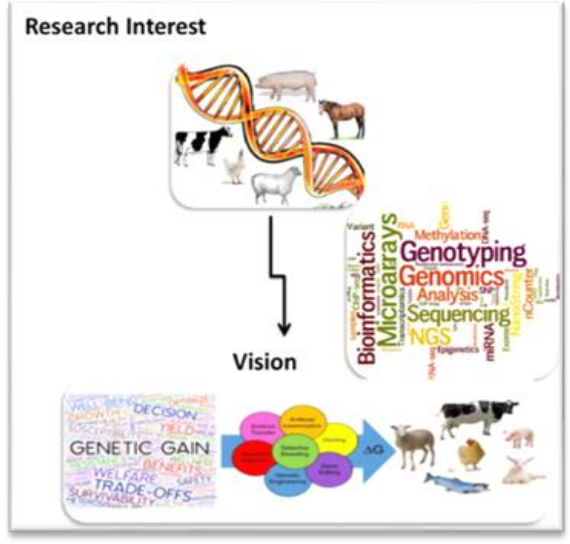
Who am I?
Helen Nigussie(PhD)

Background

Animal genetics and Breeding
(Sheep genetics /genomics)



- AfricanBP member & national node representative
- BixCoP2018 fellow
- PEER USAID mentoring fellow
- Board Member of SEWiST
- EC Member of ESAP
- ISAG member



- AfricanBP: https://africanbiogenome.org/national-nodes/
- ACaCCIA: http://acaciaafrica.org/bix-fellow/helen-nigussie/
- @helen_nigussie2
- https://www.linkedin.com/in/helen-nigussie-b3135536/
- https://www.researchgate.net/profile/Helen_Nigussie2

BIOTECNET Workshop

25-26 August, BETin, Addis Ababa

BIOTECHNET-ICG

Introduction to Transcriptomics

Helen Nigussie(PhD),
MCMB, AAU
helen.nigussie@aau.edu.et

Introduction to Transcriptomics

Contents

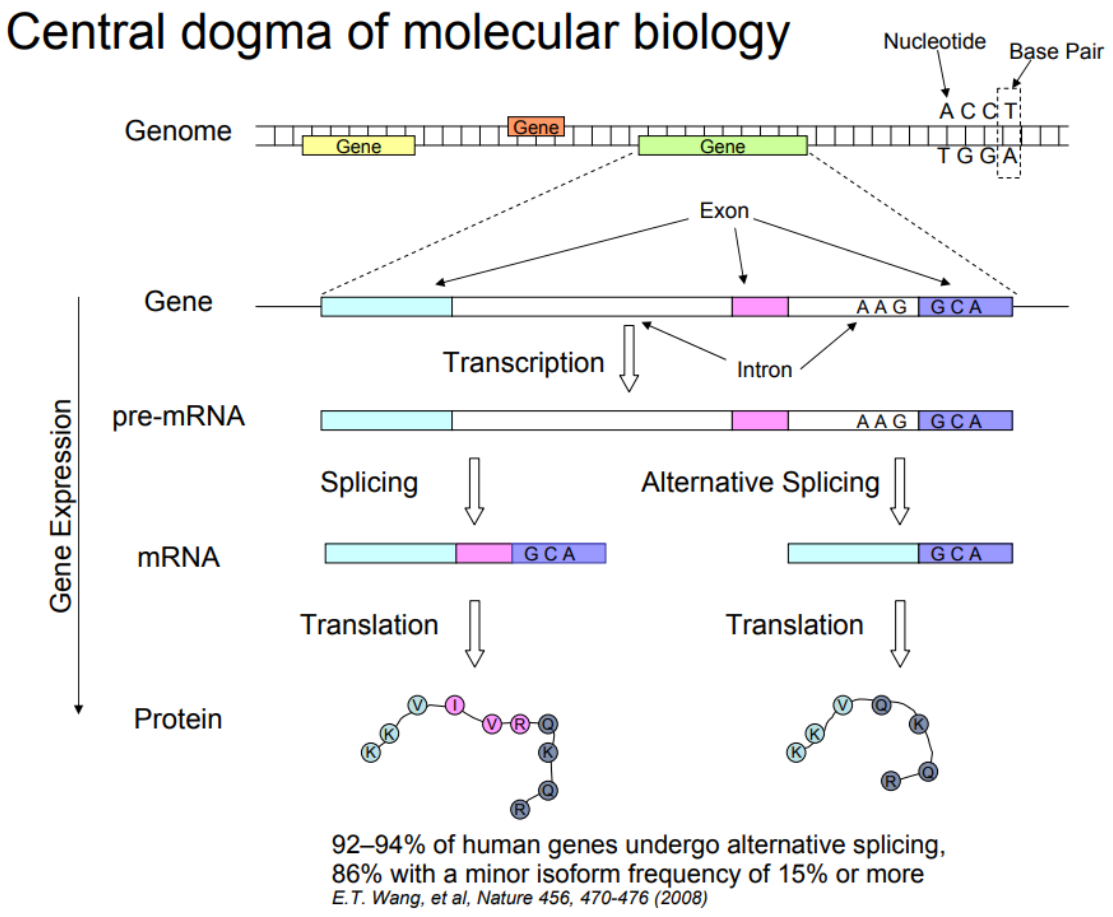
- Introduction
- Techniques of transcriptomics
- Quality Control and trimmomatic
- Exercise
- Transcript Assembly (de novo or reference based)

What is Transcriptomics?

Transcriptomics is :

- The study of all RNA molecules in a cell.
- a comprehensive analysis of whole sets of transcripts for a particular cell, tissue, organ, or
  - ✓ whole organism corresponding **to a particular time** or
  - ✓ **developmental stag**es or
  - ✓ may be under **some specific physiological conditions.**

Central Dogma of Molecular Biology and Transcriptome

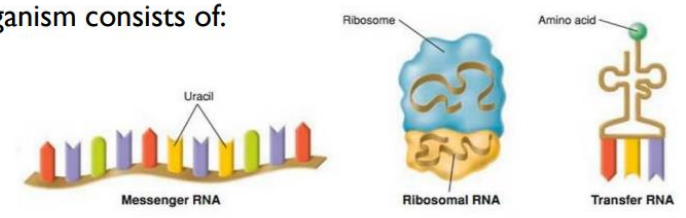


92–94% of human genes undergo alternative splicing, 86% with a minor isoform frequency of 15% or more
E.T. Wang, et al, Nature 456, 470-476 (2008)

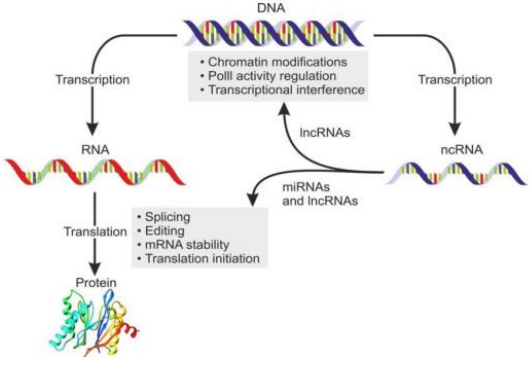
The transcriptome comprises all RNA that coding for genes and noncoding RNA.

The transcriptome of an organism consists of:

- **mRNA**
- **tRNA**
- **rRNA**

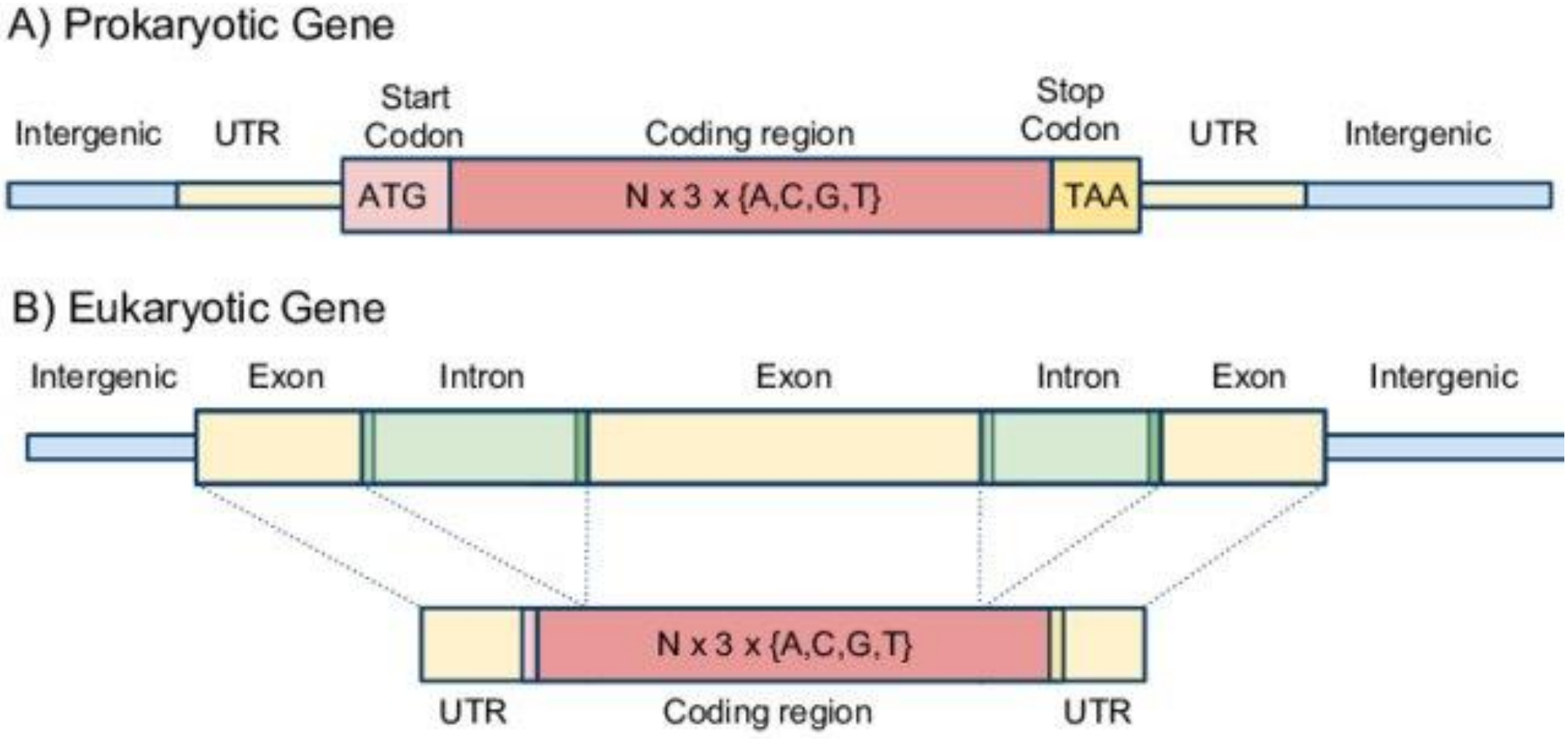


- **non-coding RNA** (ncRNA):
- micro RNA (miRNA),
- small interfering RNA (siRNA),
- long noncoding RNA (lncRNA)
- many others



Bianca Castiglioni

Gene structure



(Zeller et al. 2011)

Why Transcriptomics?

Transcriptomics can determine:-

- When and where each gene **is turned on or off** in the cells and tissues of an organism.
- **The amount of gene activity** - also called **gene expression** - in a certain cell or tissue type.
- **Level of gene expression:** behaviors of various cells and tissues, both in health and disease.

Application

- Differential expression
- Gene fusion
- Alternative splicing
- Novel transcribed regions
- Allele-specific expression
- RNA editing
- Transcriptome for non-model organisms
- Cataloguing all types of transcripts (mRNA, tRNA, rRNA, siRNA, noncoding RNAs, etc.)

Techniques of transcriptomics

Hybrdization based approach

- ✓ Florescent labeled cDNA with custom made microarrays
- ✓ Commercial high density Oligo microarrays

Sequence based approach

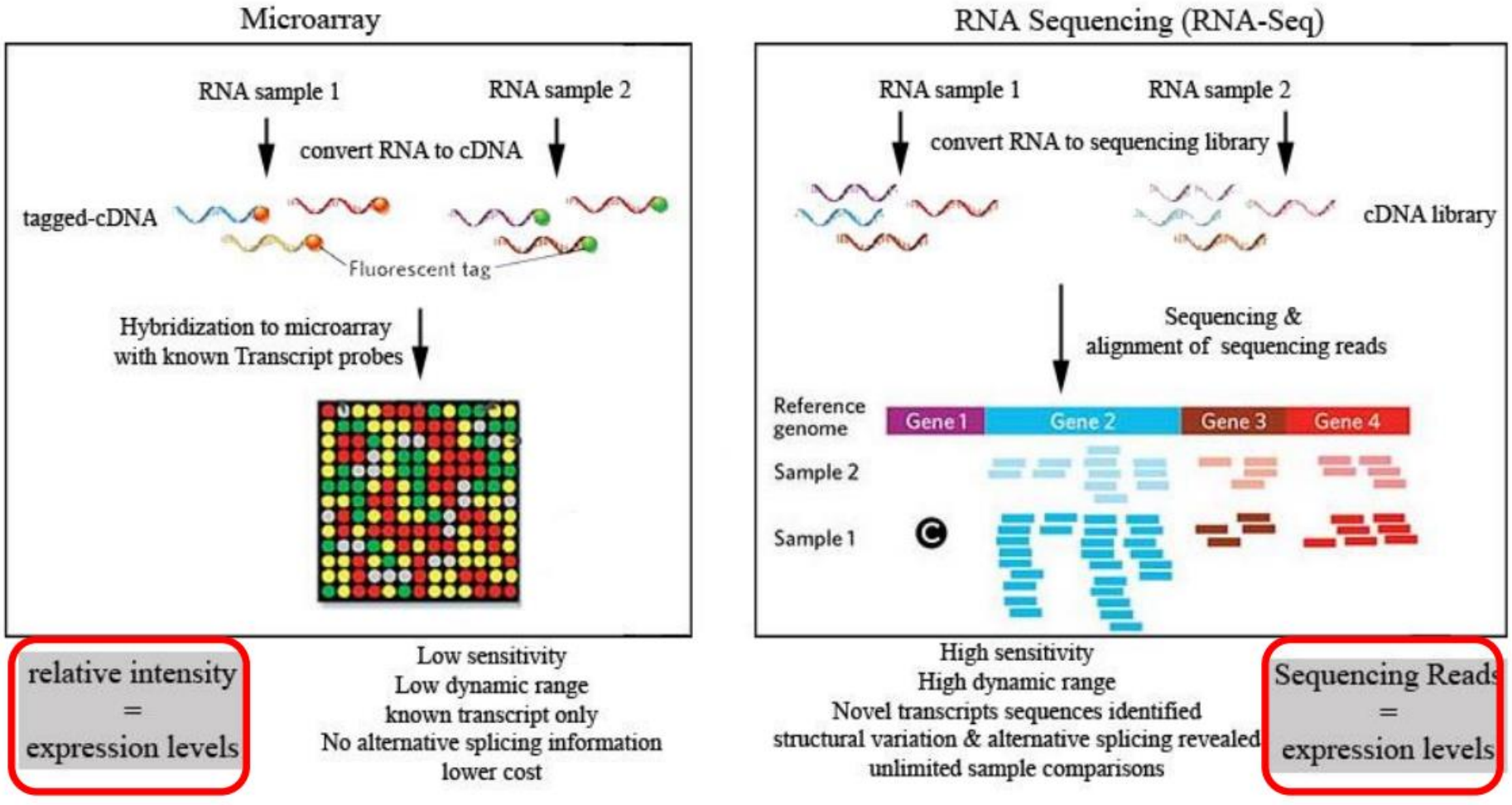
- ✓ Sangar Sequencing of cDNA or Expressed sequence tag (EST) libraries
- ✓ Serial and Cap analysis of gene expression (SAGE/CAGE)
- ✓ Whole exome sequencing
- ✓ RNA sequencing

Evolution of Transcriptomics Technologies

- Northern Blot
- RT-PCR
- Microarrays
- (NGS) RNA-seq

- Single Genes
- Multiple genes
- Whole Genomes
- Populations of genomes

Microarray vs RNA sequencing



https://www.otogenetics.com/rna-sequencing-vs-microarray/

Sequencing technologies

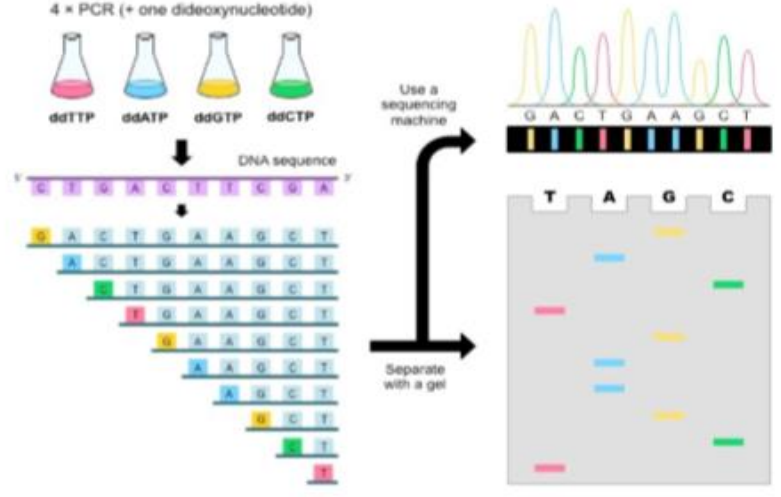
Sanger Sequencing

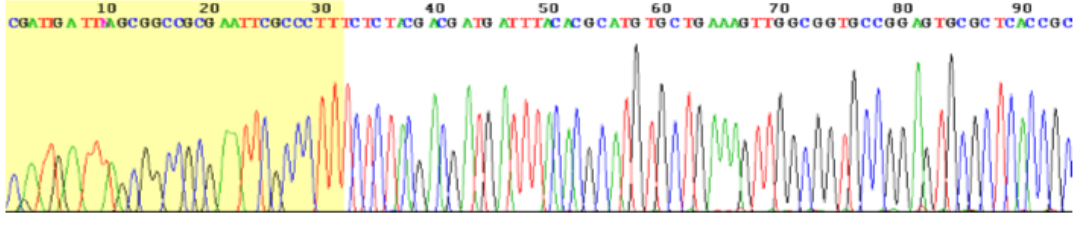


Fred Sanger 1918-2013

Double Nobel laureate and developer of the dideoxy sequencing method, first published in December 1977. [Credit: Wellcome Images]

- 1986 – Dye terminator Sanger sequencing (dideoxy chain termination) technology. Peaking at about 900kb/day





Second /Next Generation sequencing

The second generation sequencing technology are mainly classified into three major sequencing techniques:

- ✓Roche/454 pyrosequencing (2005),
- ✓Illumina/Solexa sequencing of polymerase synthesis (2006)
- ✓ and ABI/SOLiD sequencing ligase (2007) technology.

NGS….

Next Generation

- 2006: the second NGS platform, was Solexa (later acquired by Illumina). Now the dominant platform an estimated >90% of all bases sequenced are from an Illumina machine, Sequencing by Synthesis (200Gb/day)


Illumina MiniSeq


Illumina MiSeq


Illumina NextSeq


Illumina HiSeq


NovaSeq

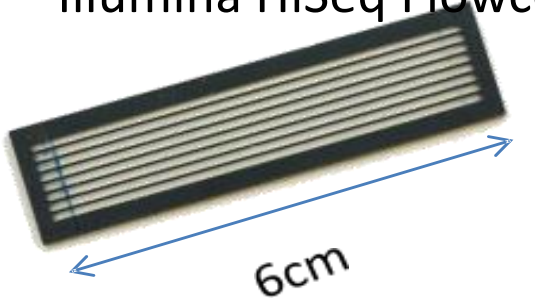
https://petitinstitute.gatech.edu/next-generation-sequencing

High Throughput DNA/RNA sequencing

Automated sample preparation



Illumina HiSeq Flowcell





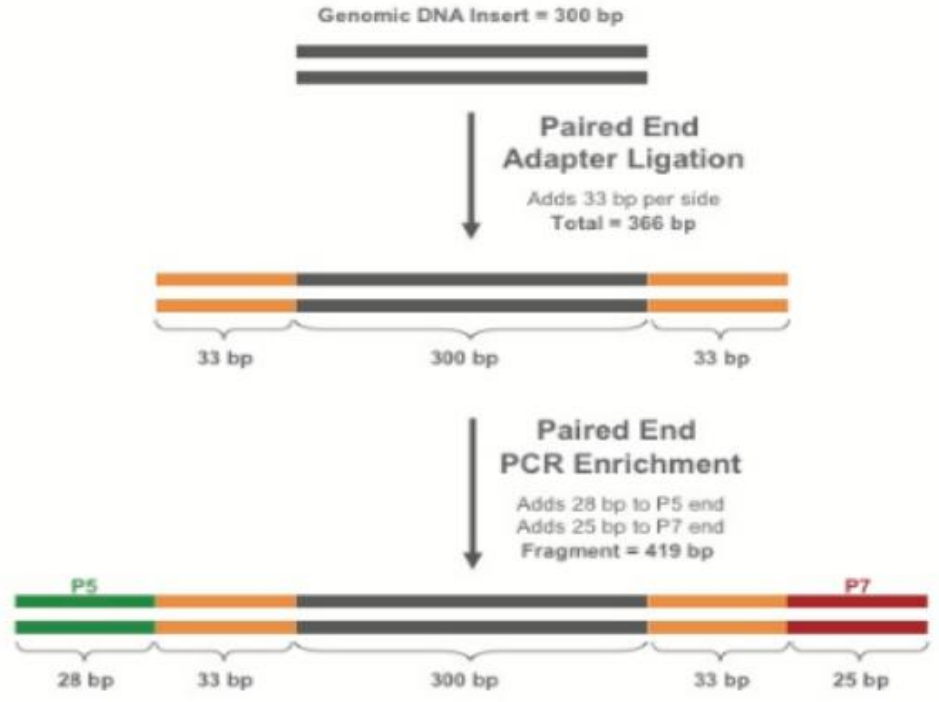
cBot Cluster generation



HiSeq 2500

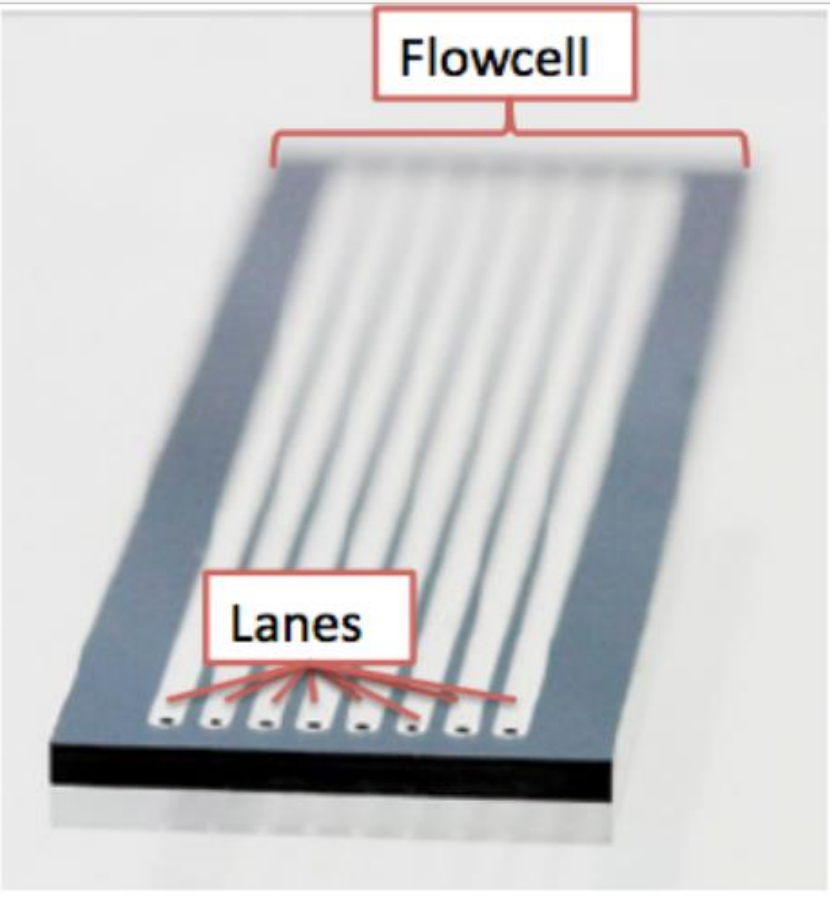
Amel Ghouila, 2018

Sequencing terminology



Library preparation, each reaction produces a library of DNA fragments

Source: GATK website



Each NGS machine has a flowcell with 8 independent lanes during a single sequencing run

Third generation sequencers

- Single-molecule sequencing
- Current market leaders:



PACBIO®

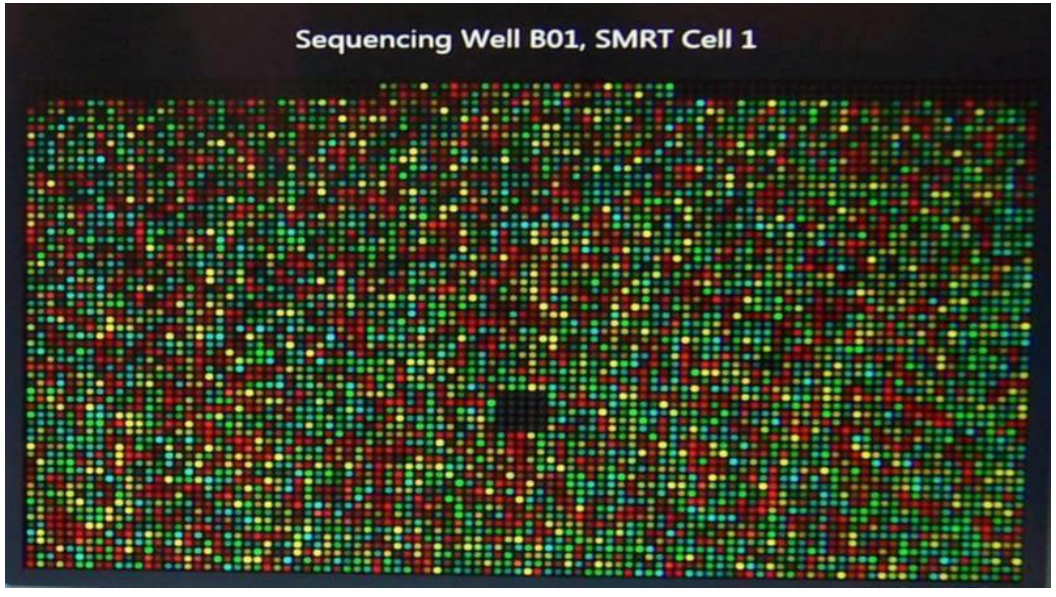


PacBio SMRT Sequencing



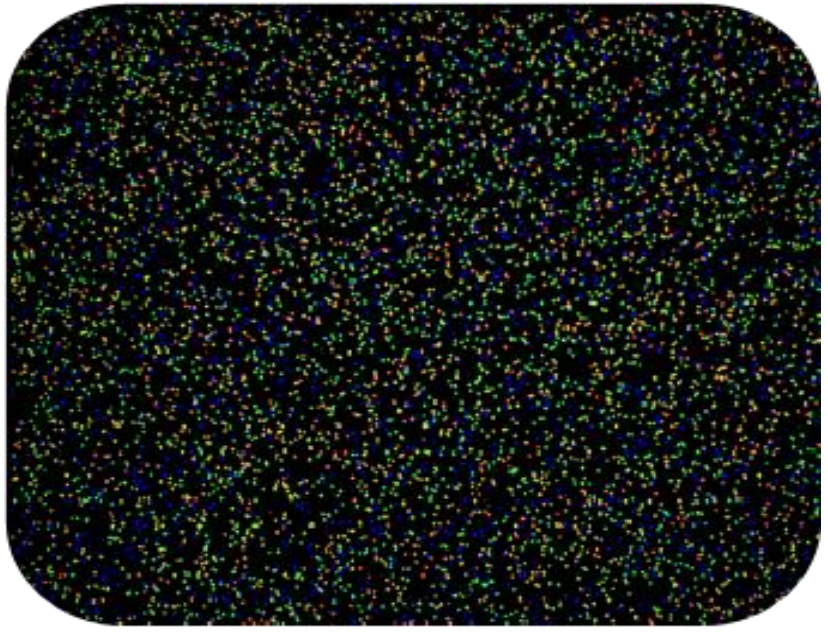
- Single Molecule Real Time sequencing (SMRT)
- Facilitates very long read lengths up to 90Kbp
- Can be used to sequence whole transcripts for RNA sequencing
- Simultaneous sequencing and methylation detection

What it looks like



PacBio ZMWs with single DNA strand

Ordered

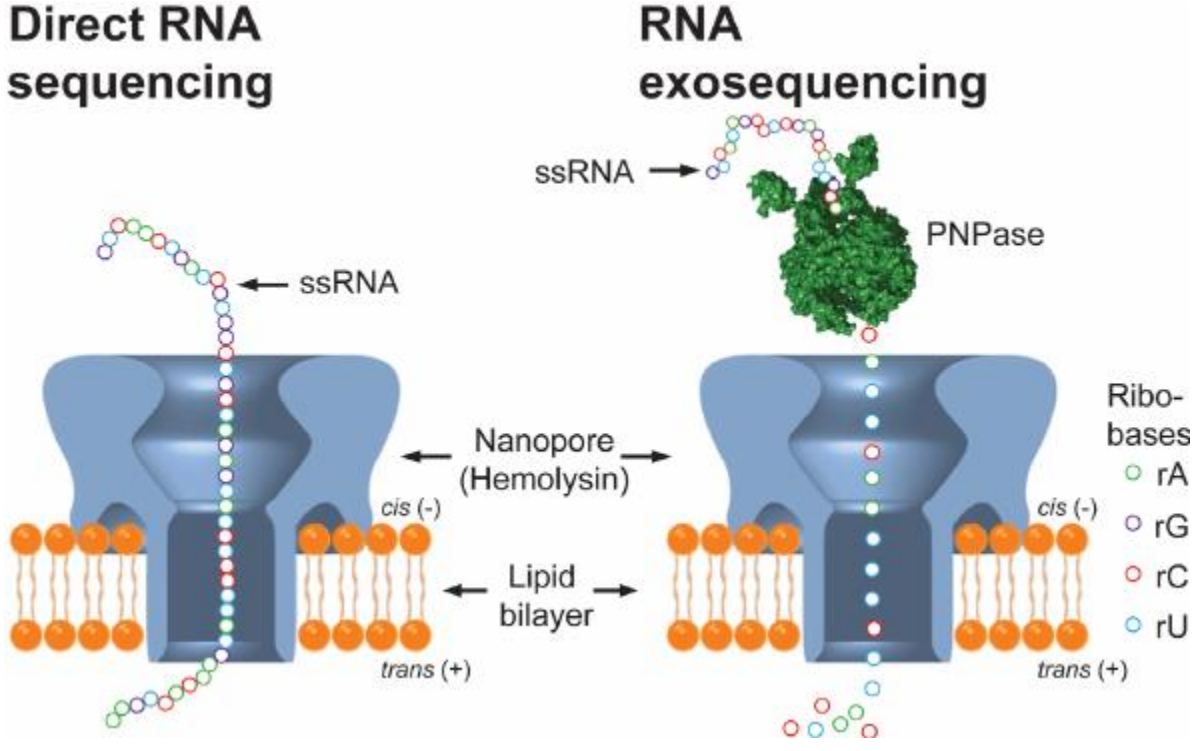


Illumina DNA mono-colonal clusters

Unordered

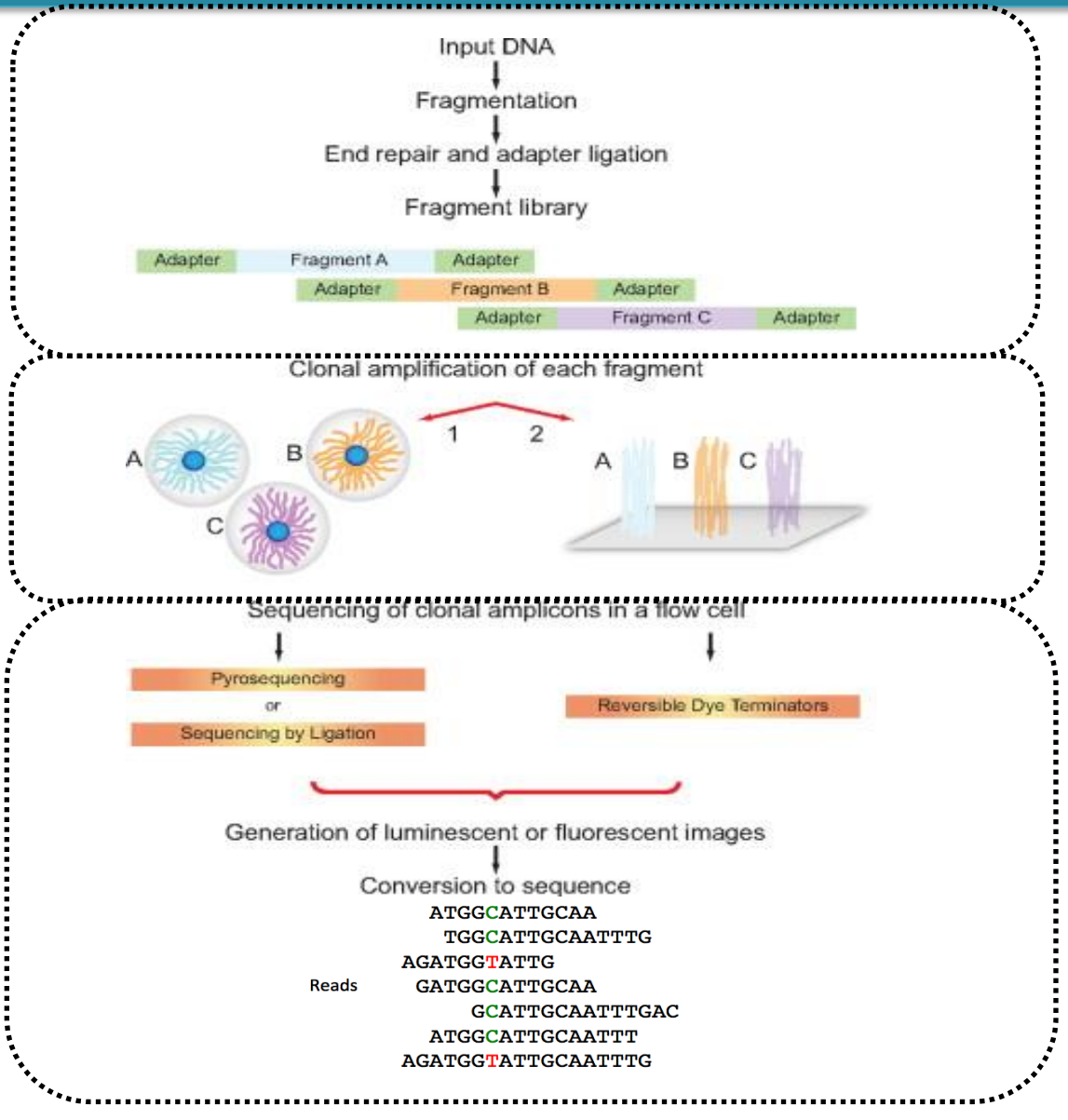
Oxford Nanopore Sequencing

- Sequences single DNA molecules



Saliba et al 2014; 10.1093/nar/gku555

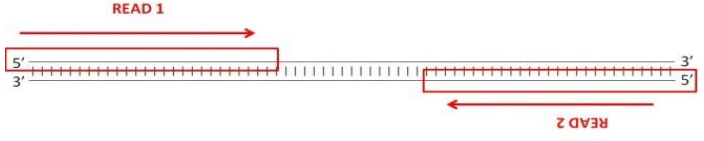
Sequencing steps



Library preparation

Library amplification

Parallel sequencing

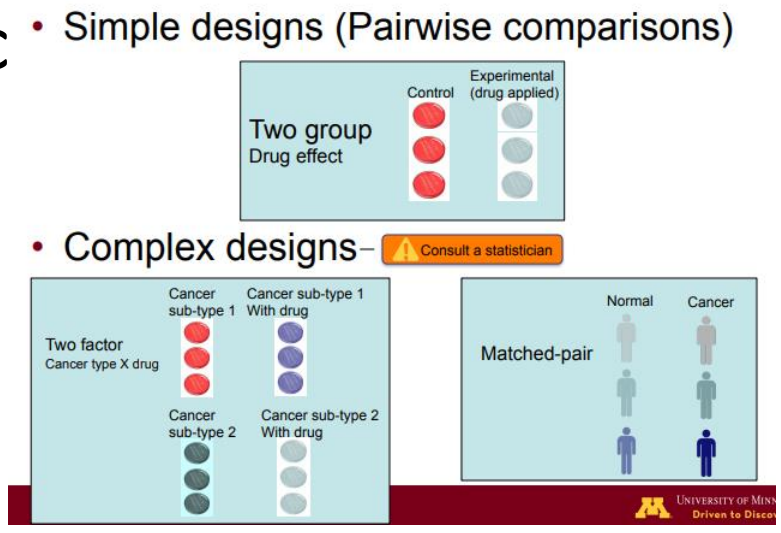


Voelkerding KV et al., J Mol Diagn (2010) 12,539-51.

Experimental Design

Important points needs to consider Biologic

- Paired end vs single end reads
- Read length
- Read depth
- Replicates
- Pooling

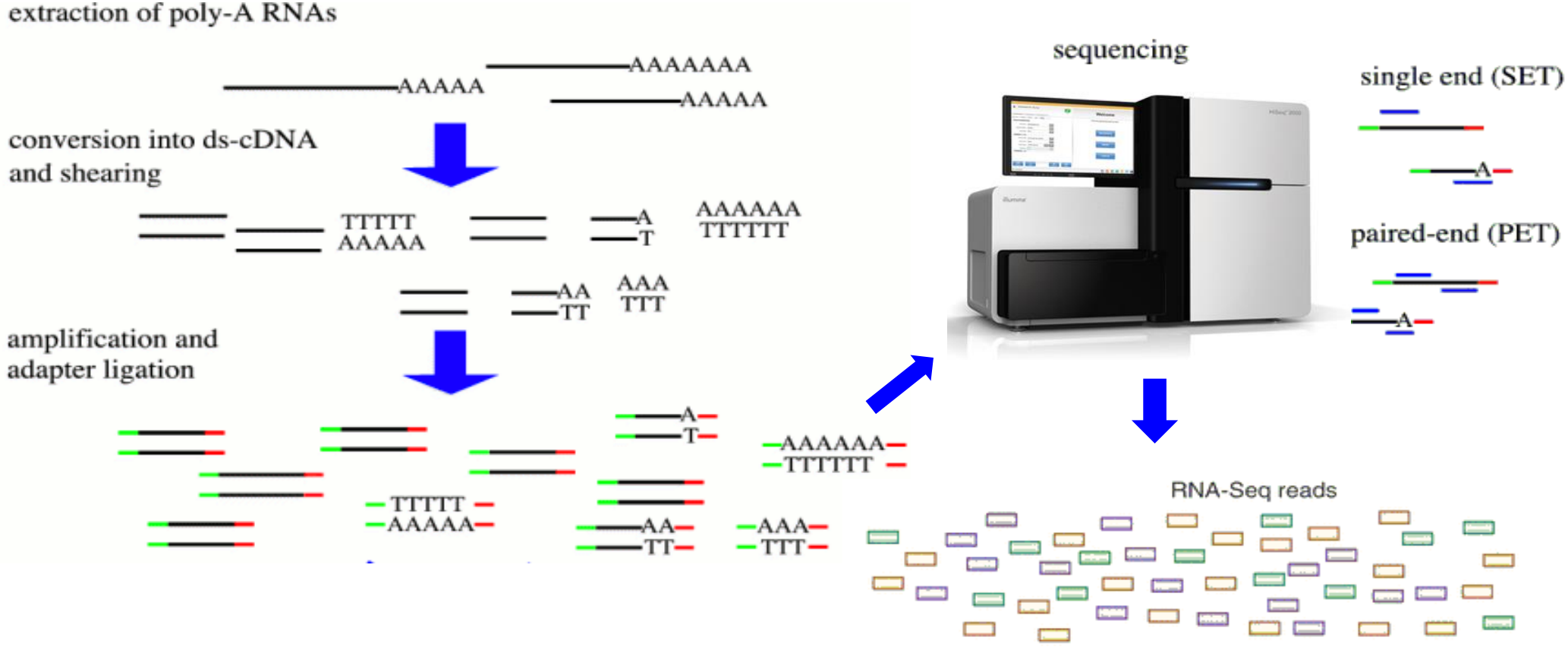


Experimental Design

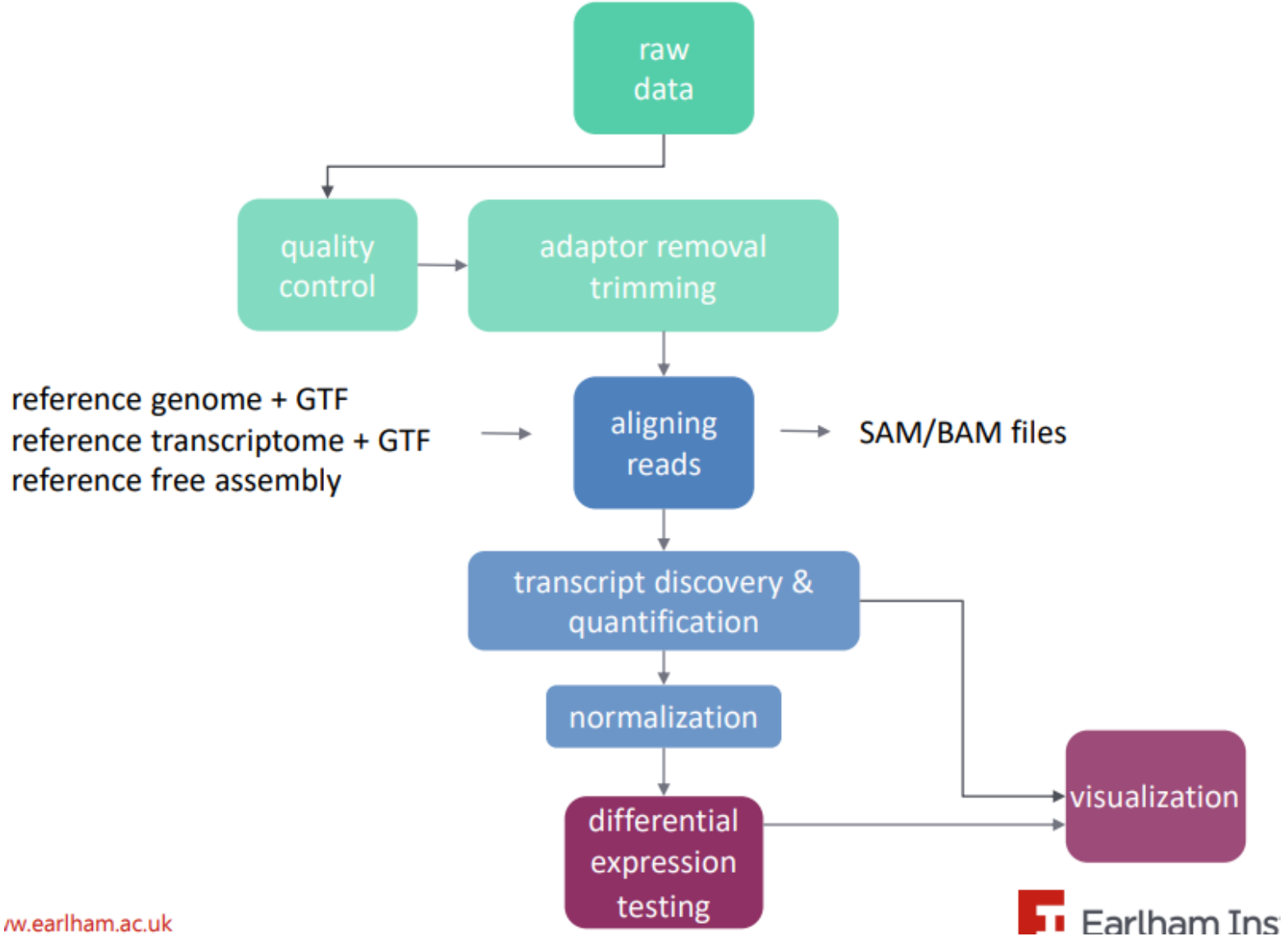
- What are my goals?
  - Transcriptome assembly?
  - Differential expression analysis?
  - Identify rare transcripts?
- What are the characteristics of my system?
  - Large, complex genome?
  - Introns and high degree of alternative splicing?
  - No reference genome or transcriptome?

Overview of RNA-Seq

Transcriptome profiling using NGS



RNA Sequence Work flow



Earlham Ins

File formats

- ✓ There are multiple file format used at various stages of NGS data processing.
- ✓ We can divide them into two basic types:
  - **Text based (FASTA, FASTQ, SAM, TF/GFF, BED, VCF)**
  - **Binary (BAM, BCF)**
- ✓ We can view and manipulate text based formats without special tools, but we will need these to access and view binary formats.

- ✓ The text--based format are often compressed to save space making them de facto binary, but still easy to read by eye using standard unix tools.

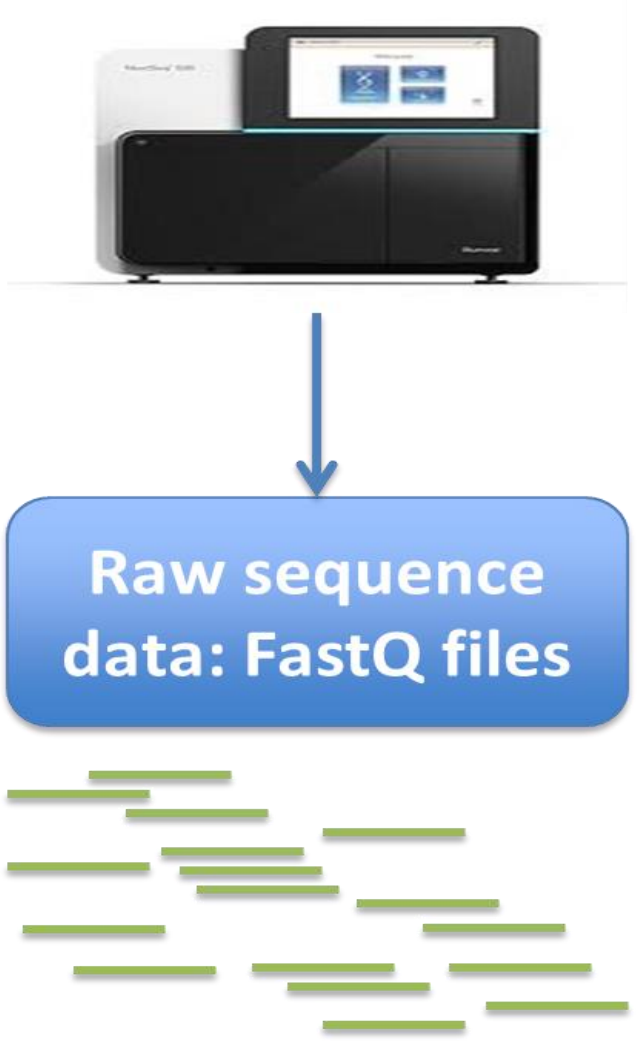
FASTQ File Format

What is a FastQ file?

FASTQ= FASTA + Quality

FastQ format is a text-based format for storing both a biological sequence and the corresponding per base quality scores

-> Most Common output provided by sequencing platforms



File formats

- A **FASTA** file uses at least 2 lines per sequence:

  Line 1 begins with a '>' and is followed by a sequence identifier

  Line 2 is the sequence letters

>HM236315.1_SA Ovis aries isolate D-loop, partial sequence; mitochondrial
AAGCCCACATAACAACCCATACAAGAAAAGCACAACCACCCACGGACATGAGCGTTCATAAACCC
AACATATCTTATGTCTGTC

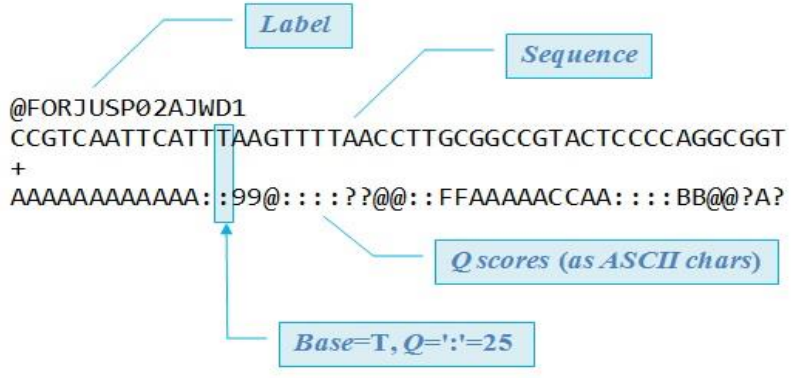
- A **FASTQ** file uses four lines per sequence:

  Line 1 begins with a '**@**' and is followed by a sequence identifier

  Line 2 is the **sequence letters**

  Line 3 begins with a '**+**' and is *rarely* followed by the sequence identifier

  Line 4 encodes th *ne 2*, and must contain the same number



Phred quality score

The quality score of a base is called the **Phred score** (or **Q score**)
It is an integer value defined as a property that is logarithmically related to the base calling error probabilities (P)

$$Q = -10 \log_{10}(P) \Leftrightarrow P = 10^{-Q/10}$$

- Phred or Q scores are often represented as ASCII characters
Starting in Illumina 1.8, the quality scores have basically returned to the use of the Sanger format (Phred+33)

Phred quality score

Note: Phred version is independent from the sequencing technology version

Illumina 1.8+

```
LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
|                         |    |        |                              |                     |
0.........................26...31.......41
```

Phred score

Table 1: Quality Scores and Base Calling Accuracy

Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

Position Quality Score:

30–40: ✔ Good

<20: Discard

Reads Quality Check
FastQC



FastQC: Quality Control for FastQ files

GUI, Command line, Available on Galaxy

Graphic reports

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

http://dnacore.missouri.edu/PDF/FastQC_Manual.pdf

- - Quality control checks on raw sequence data
  - **Input:** FastQ files (or BAM/SAM)
  - **Output:** Summary graphs allowing quick data assessment
- - FastQC can be run in 2 modes:
  - Stand alone interactive application ← small number of samples
  - Non-interactive mode ← large number

FastQC can be integrated in

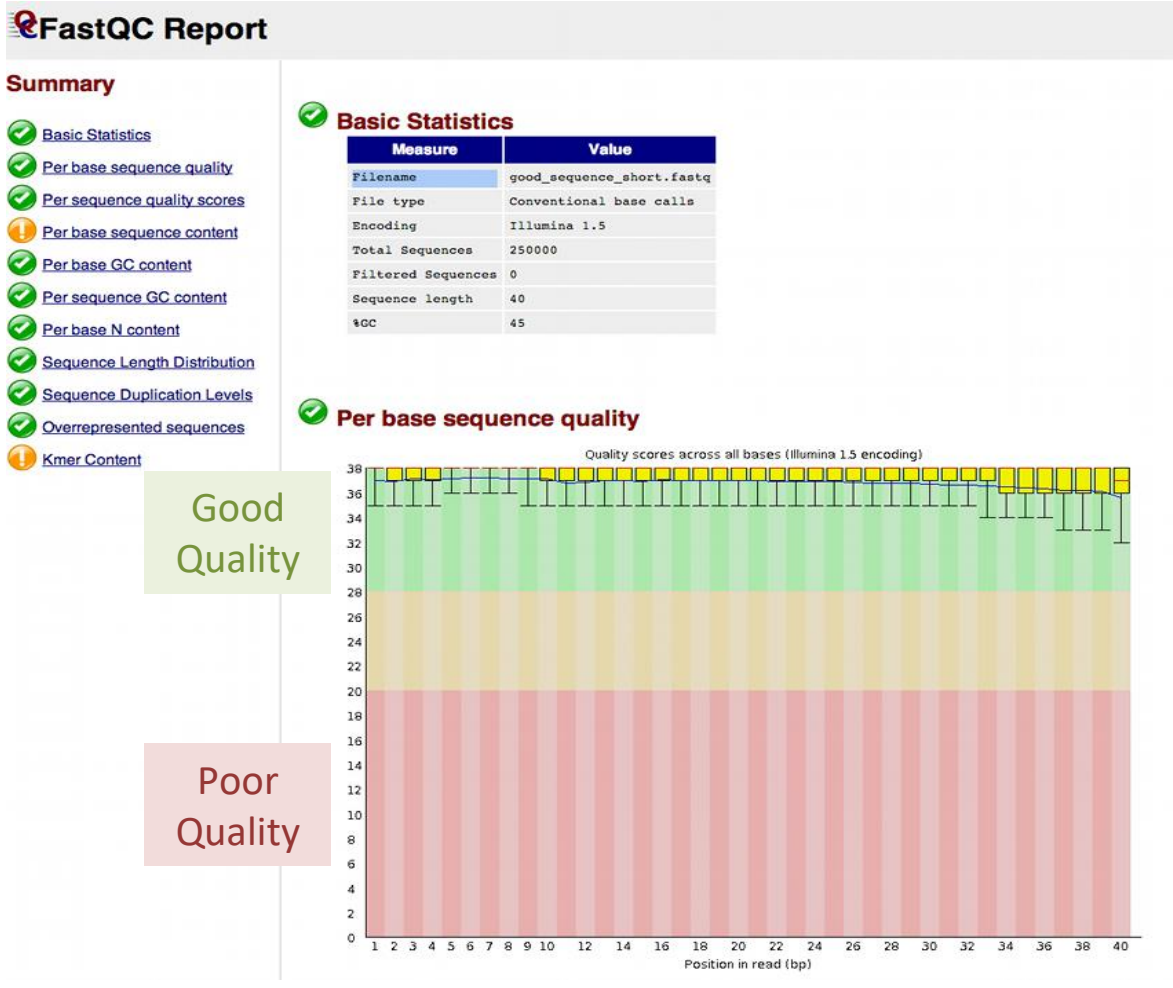
ANALYSIS PIPELINES

A summary of the modules which were run, and a quick evaluation of whether the results seem **entirely normal**, **slightly abnormal** or **very unusual**.

Normal

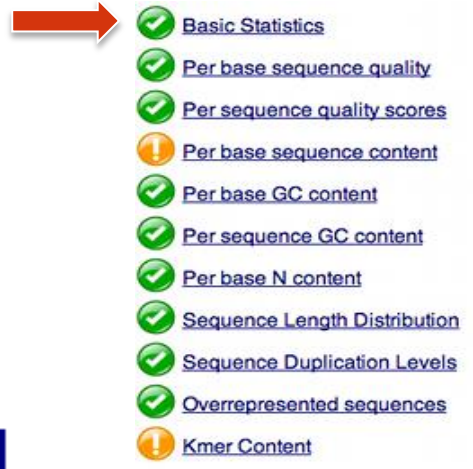
Reasonable

Unexpected



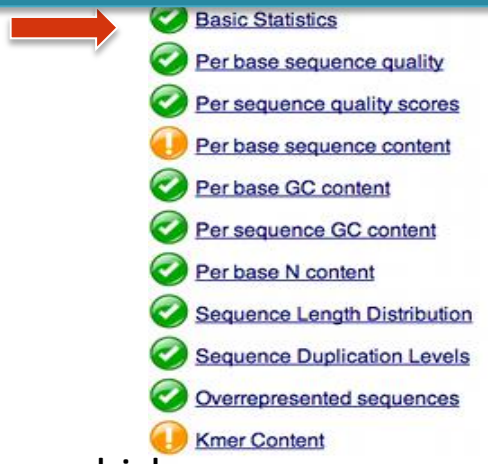
Basic Statistics

ASCII quality encoding format (*Phred* format)
Number of reads, Read length, …



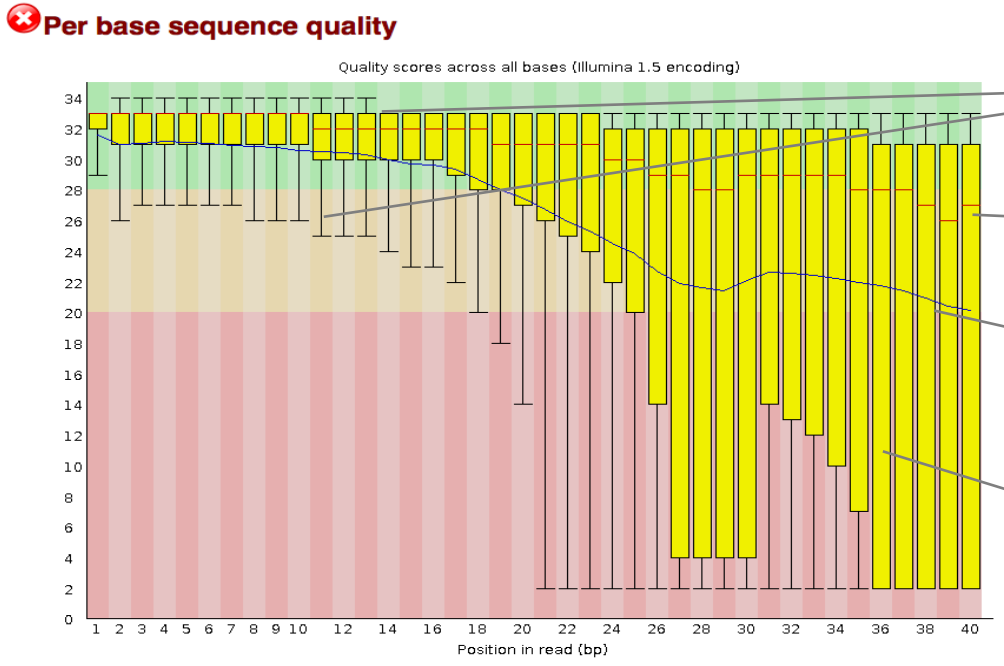
Basic Statistics

Measure	Value
Filename	good_sequence_short.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	250000
Sequences flagged as poor quality	0
Sequence length	40
%GC	45



Per base sequence quality

Overview of the range of quality values (*Phred* scores) across all bases from all reads



Upper and Lower whiskers: 10% and 90% points

Red line: **median** value

Blue line: **mean** quality

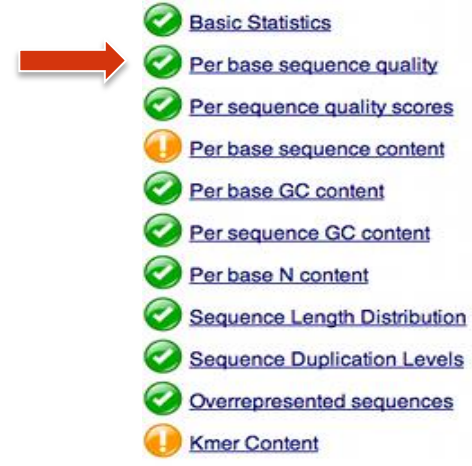
Yellow box: **inter-quartile** range (25-75%)

Position in read (bp)

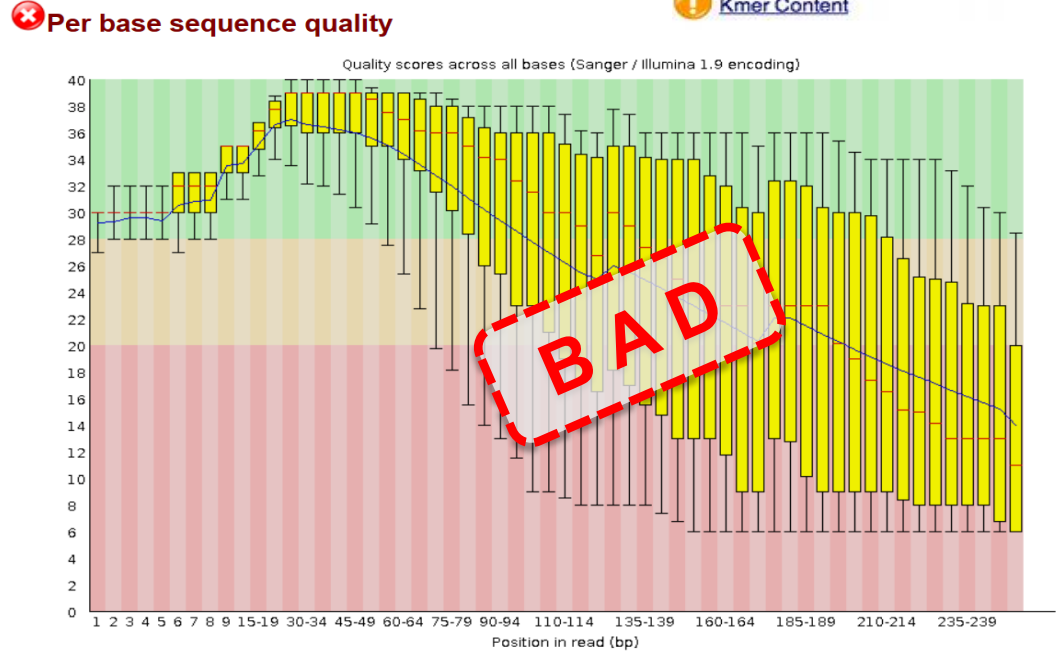
In most platforms, the quality is lower at the beginning/**end** of reads

-> This is common ! Short loss of quality early in the run and quality degrades as the run progresses

Per base sequence quality





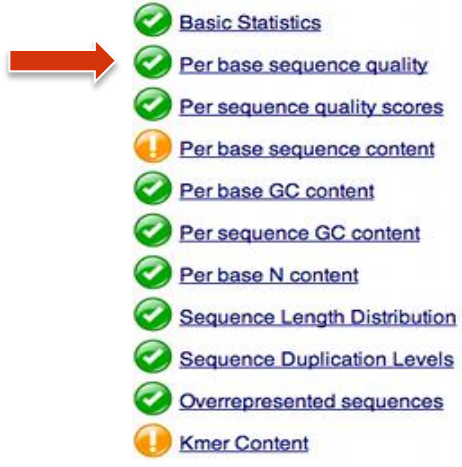


Phred Q Score

Position in read (bp)

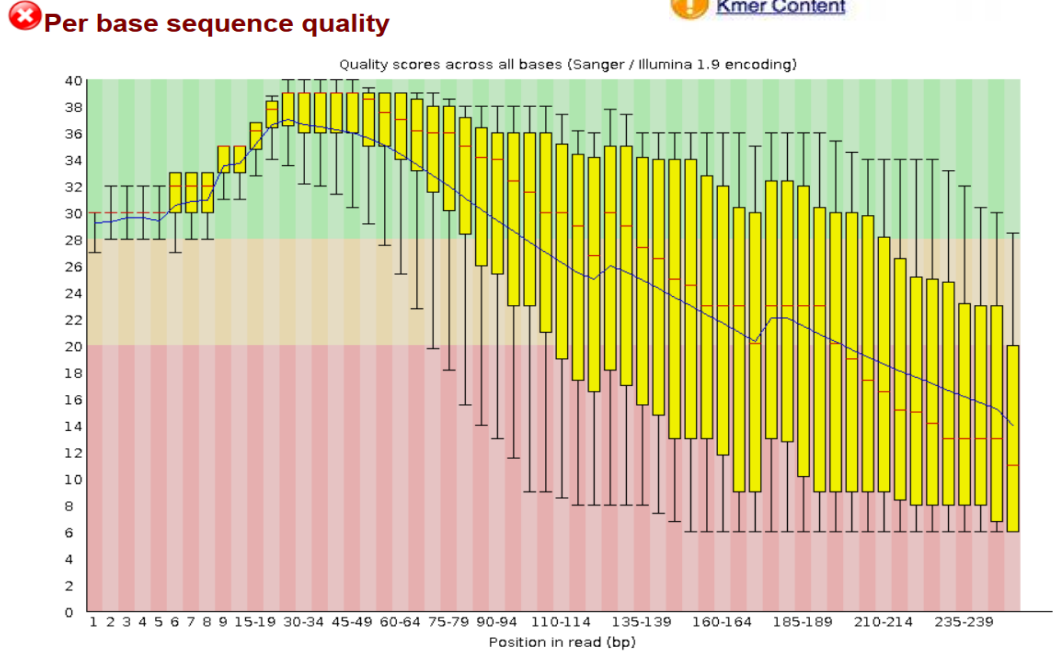
Amel Ghouila, 2018

Per base sequence quality



- High proportion of bad quality bases
- High variability
- High decrease in terms of quality towards the end of the read

→ **Important to remove bad quality data**



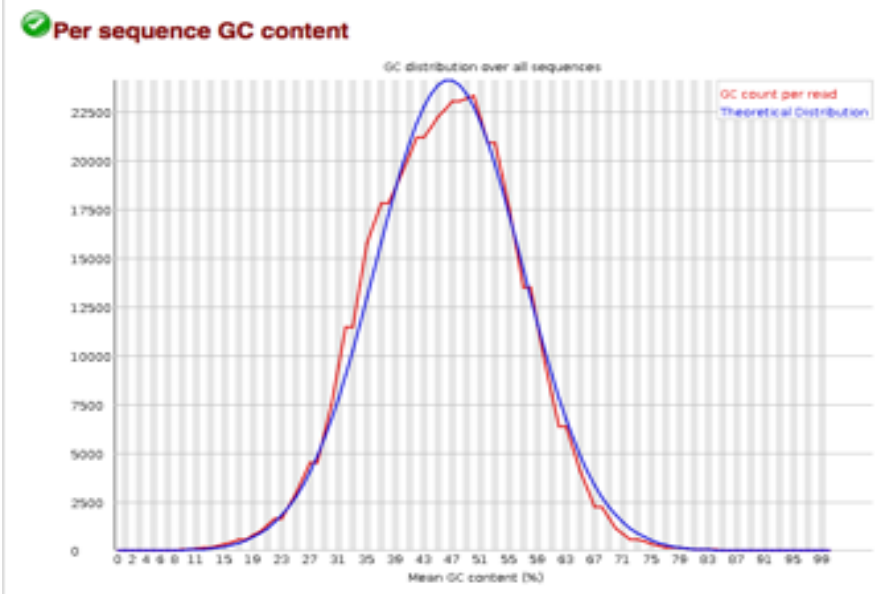
READS QUALITY CHECK
FastQC

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

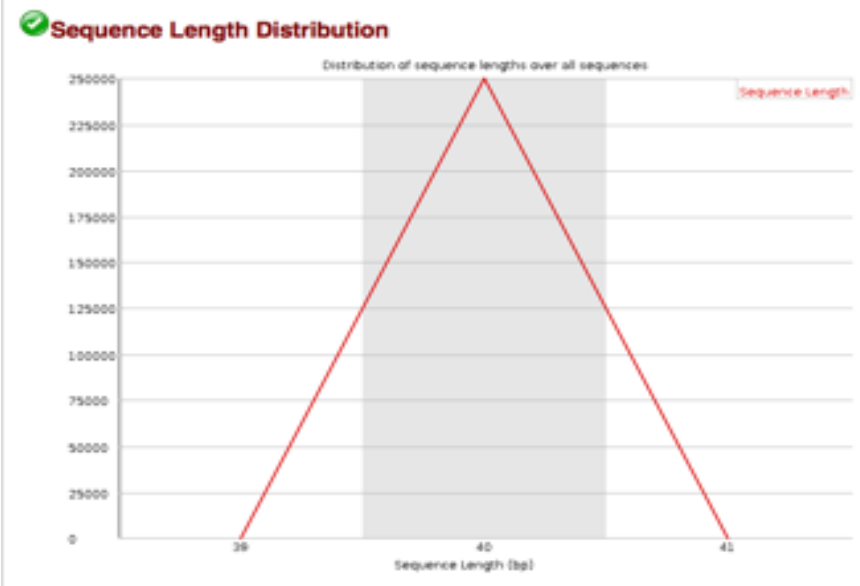


- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per base GC content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Kmer Content

Per Sequence GC Content



Sequence Length Distribution



Read Pre-Processing
Trimming low quality bases

Trimming low Quality Bases:

- Remove low quality bases, identified by the probability that they are called incorrectly.

- Widely but heterogeneously applied

- Keep in mind! Trimming might impact downstream analyses

Trimming is more important for denovo assembly

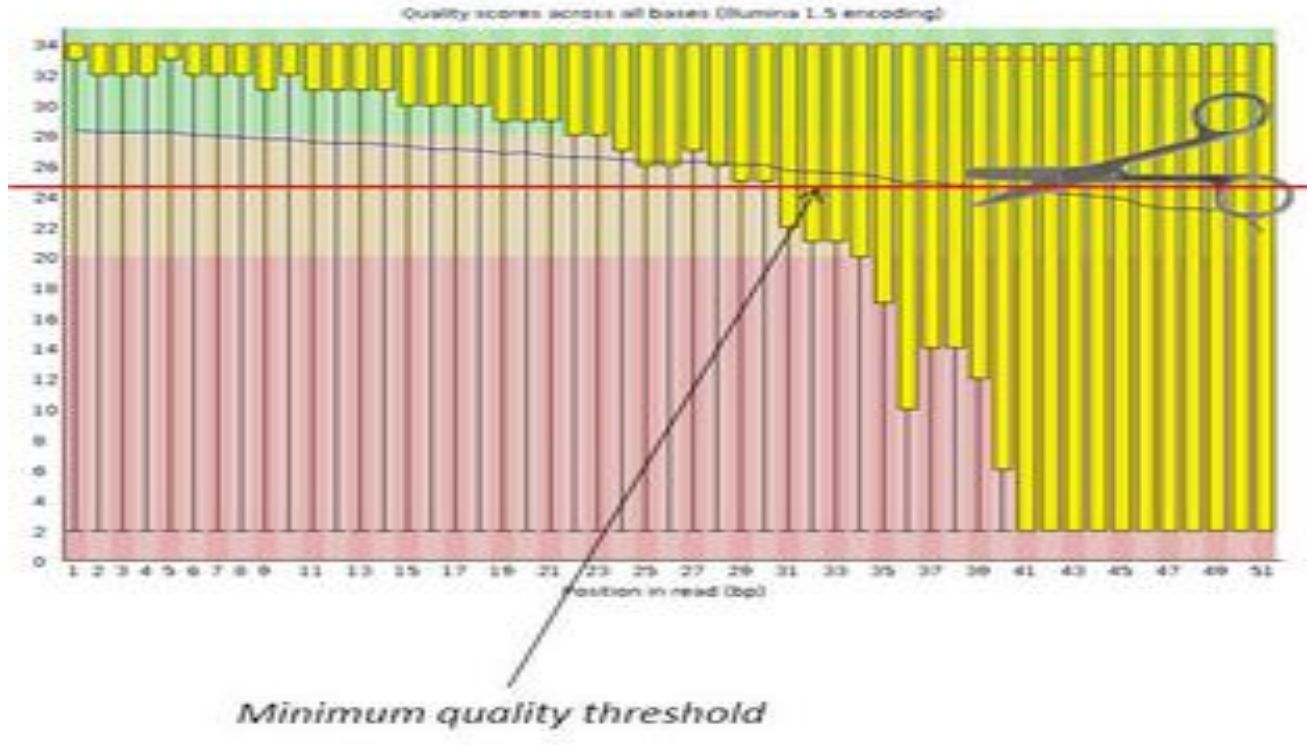
Trimming

Raw reads sequences



Our example sample analysed using FastQC:

- reads of 51bp
- average quality = 26



=> Poor quality in the ends of reads can be removed using reads **filtering** or reads **trimming** tools

Read Pre-processing
Trimming low quality bases

1) **READS FILTERING**

2) **READS TRIMMING**

3) **ADAPTIVE TRIMMING**

4) **ADAPTIVE TRIMMING** FOLLOWED BY **READ FILTERING**

Tools for trimming

Flexible tools for Illumina NGS data:

- Cutadapt
- Trimmomatic
- Seq TK, etc.

Interest:

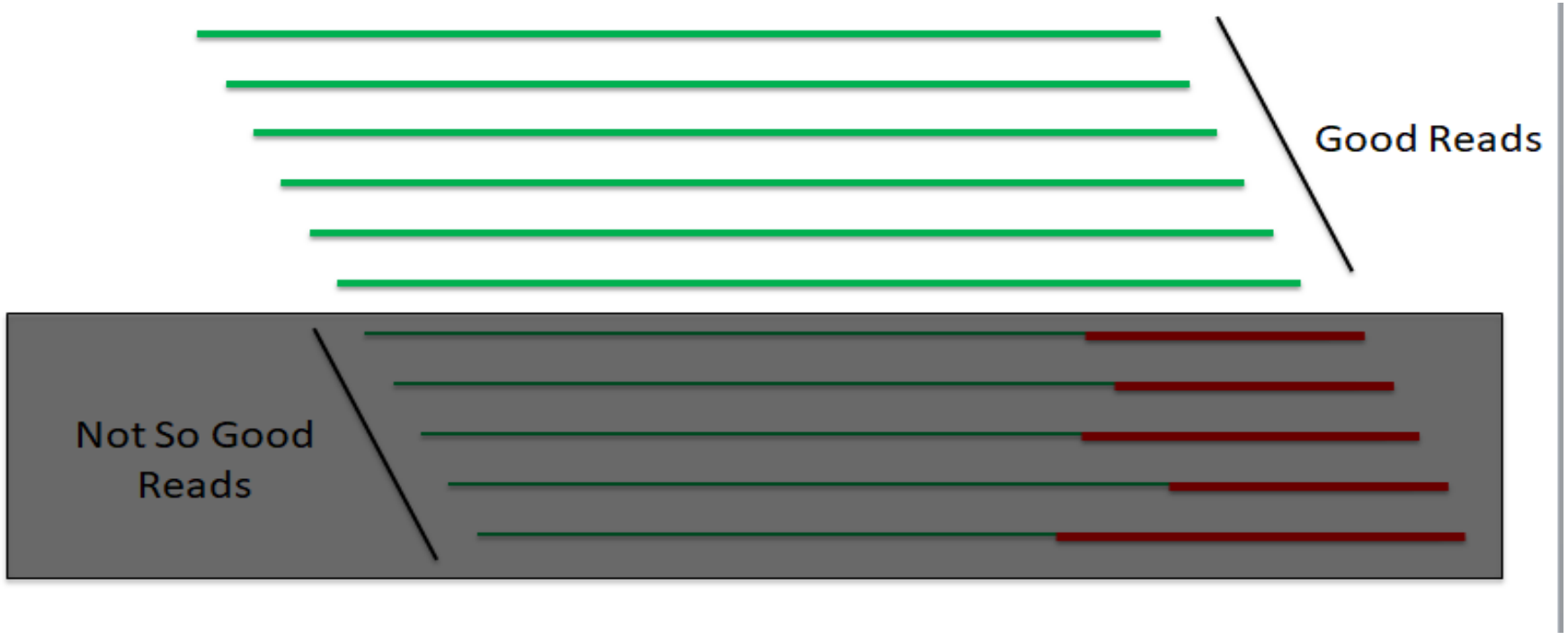
- Can align more reads: may increase mapping rates
- Improve probably de novo assembly

Warning:

- Loss of information if stringent trimming

1) READS FILTERING

Reads with poor quality ends can be discarded using **FILTERING** tools



Filtering

Reads with poor quality ends can be discarded using **FILTERING** tools

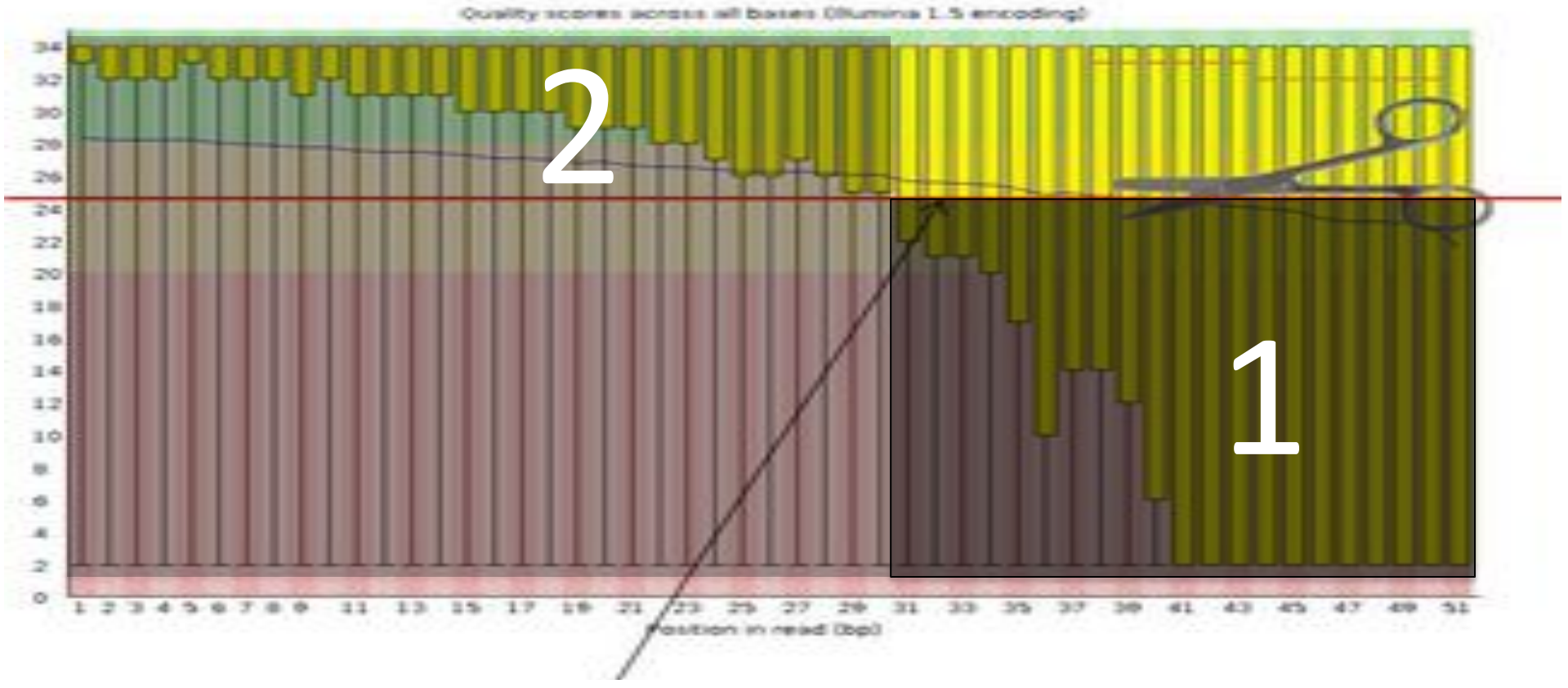


AFTER FILTERING:

Sample Overall Quality ↑

Number of Reads ↓

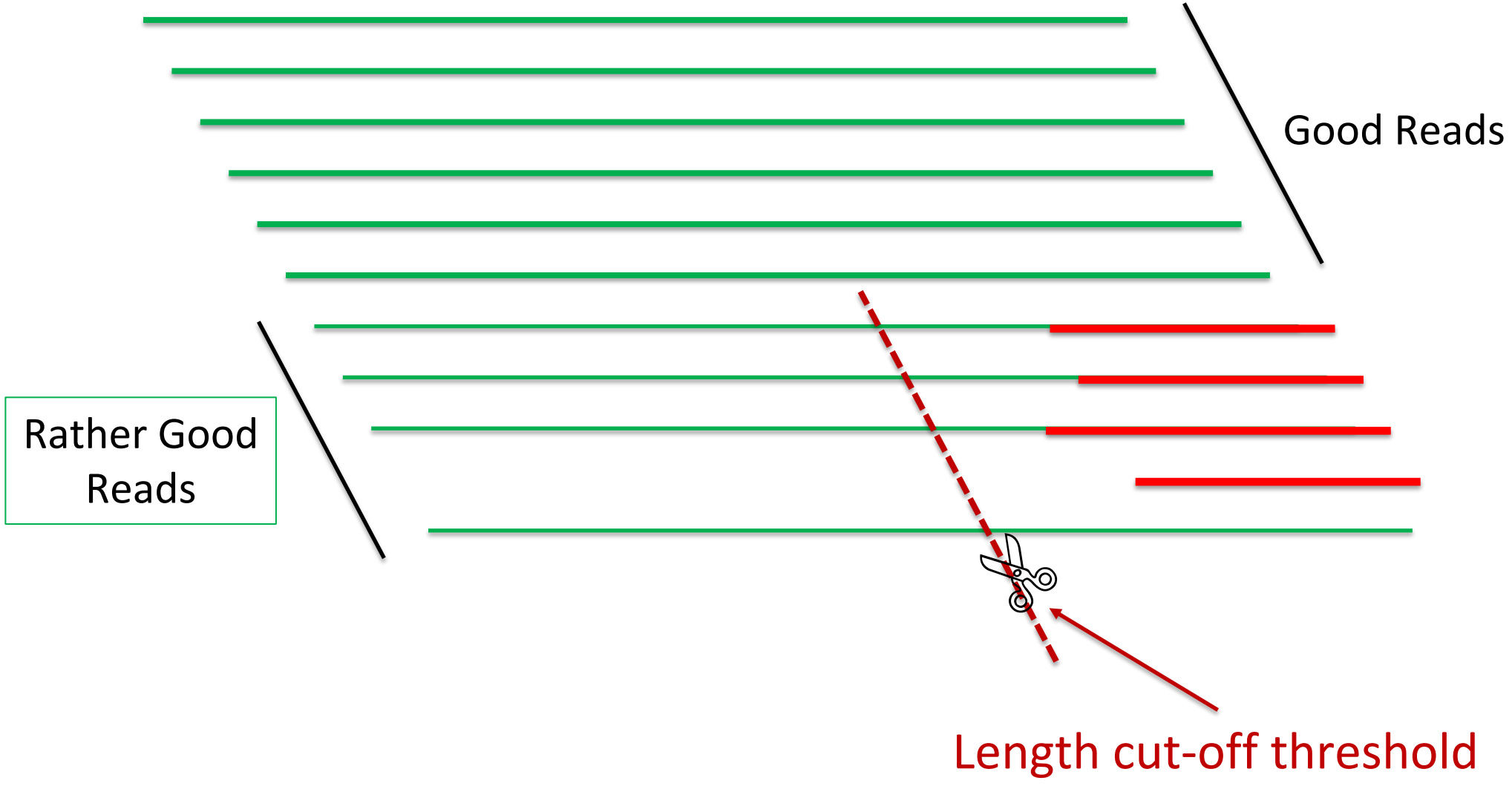
FastQC graph



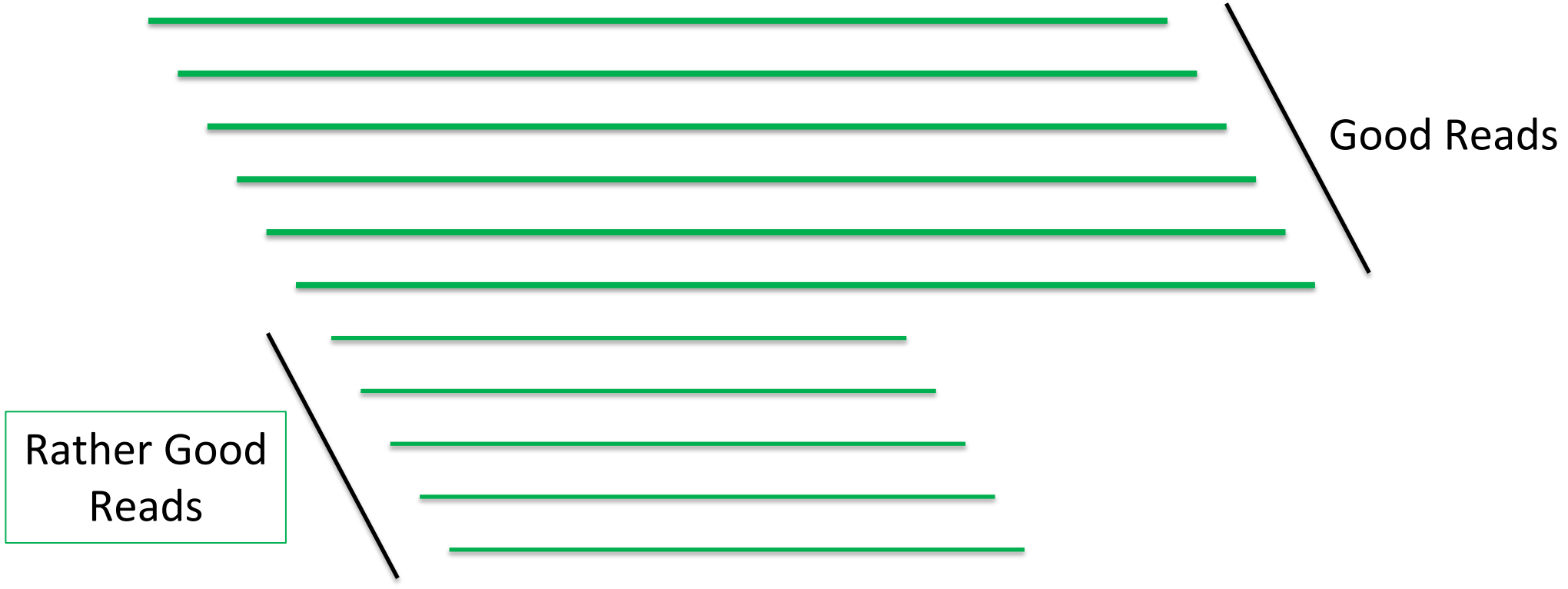
Reads with poor quality ends are removed from the dataset (1)
But some good quality positions are lost (2)

2)READS TRIMMING

Poor quality ends of reads can be cut using **TRIMMING** tools

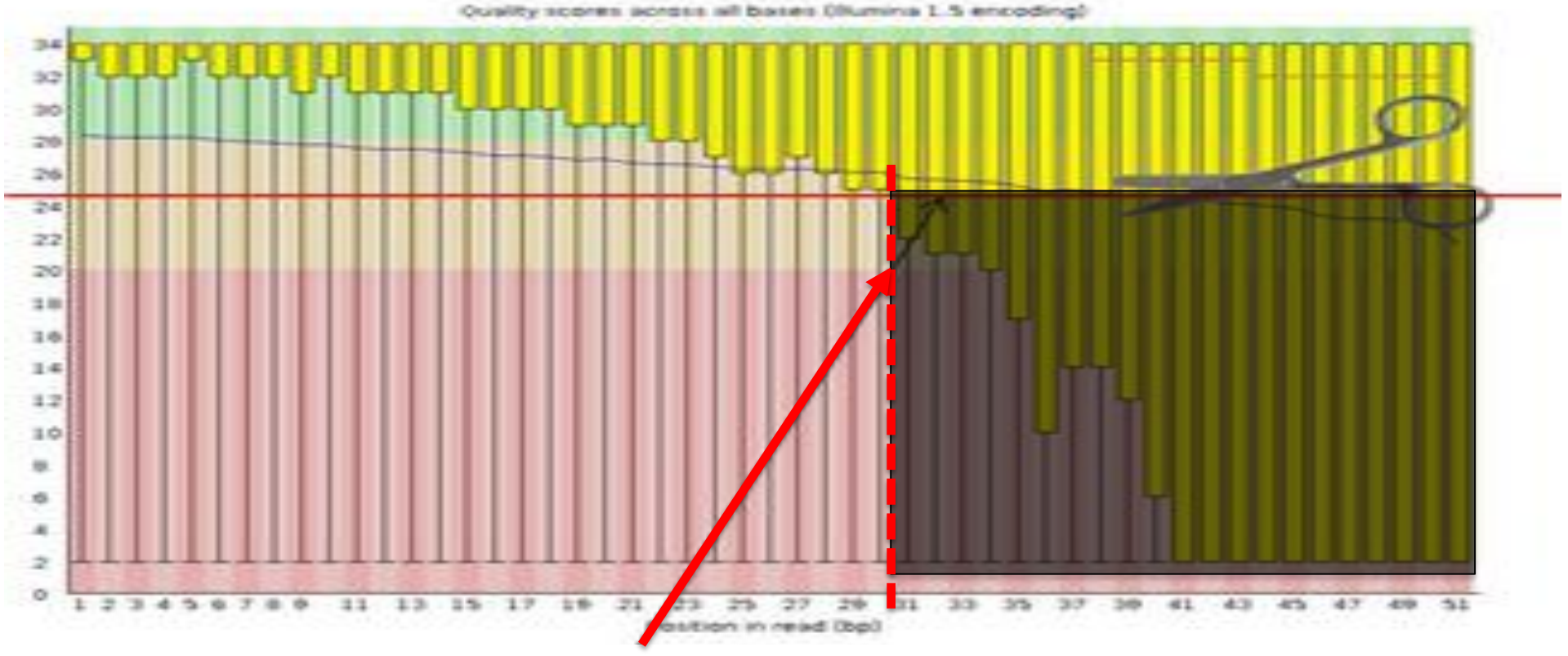


Poor quality ends of reads can be cut using **TRIMMING** tools



Trimmed reads are of better quality than raw reads but they are shorter!

FastQC graph



Cut-off threshold Length = 30 bp

Reads are now of different lengths

Number of raw reads = number of raw reads

Adaptive Trimming

A more precise **TRIMMING** approach



A more precise **TRIMMING** approach



Trimmed reads are of better quality than raw reads but they are of **variable** lengths.

Some reads can be extremely short!

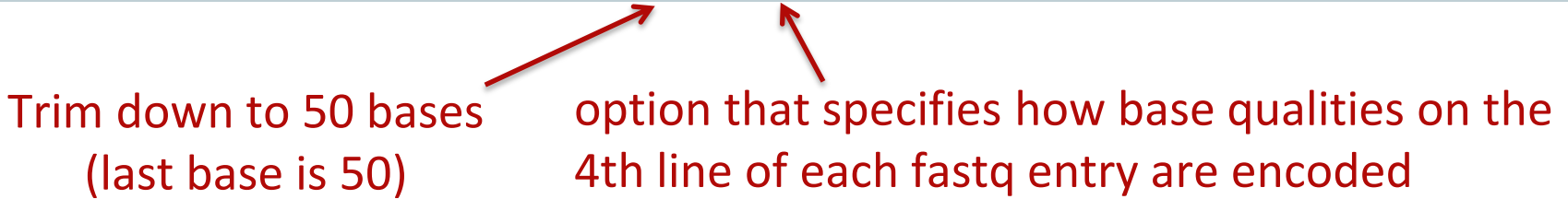
Low Quality read trimming

Low quality base reads from the sequencer can cause an otherwise mappable sequence not to align. There are a number of open source tools that can trim off 3' bases and produce a FASTQ file of the trimmed reads to use as input to the alignment program.

Manipulating tools

FASTX-Toolkit provides a set of command line tools for manipulating fasta and fastq files. The available modules include a fastx_trimmer utility for trimming fastq sequences (and quality score strings) before alignment

```
gunzip -c Sample.fastq.gz | fastx_trimmer -l 50 -Q 33 > Sample_trimmed.fq
```



Adapter trimming

A 3' adapter contamination can cause the insert sequence not to align (adapter sequence ≠ bases at the 3' end of the reference genome sequence). Unlike general fixed-length trimming, adapter trimming removes differing numbers of 3' bases depending on where the adapter sequence is found.

Manipulating tools
Cutadapt program is an excellent tool for removing adapter contamination.
Ex cutadapt on on a single-end read sample:

```
cutadapt -m 22 -O 33 -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC
```

discard any sequence that is smaller than 22 bases after trimming

No trimming of 3' adapter sequences unless at least the first 10 bases of the adapter are seen at the 3' end of the read

Exercise

- Check the quality of raw reads using fastqc
- Low quality trimming using trimmomatic
- Raw data/reads and command line for fastqc and trimmomatic will be provided

Requirements

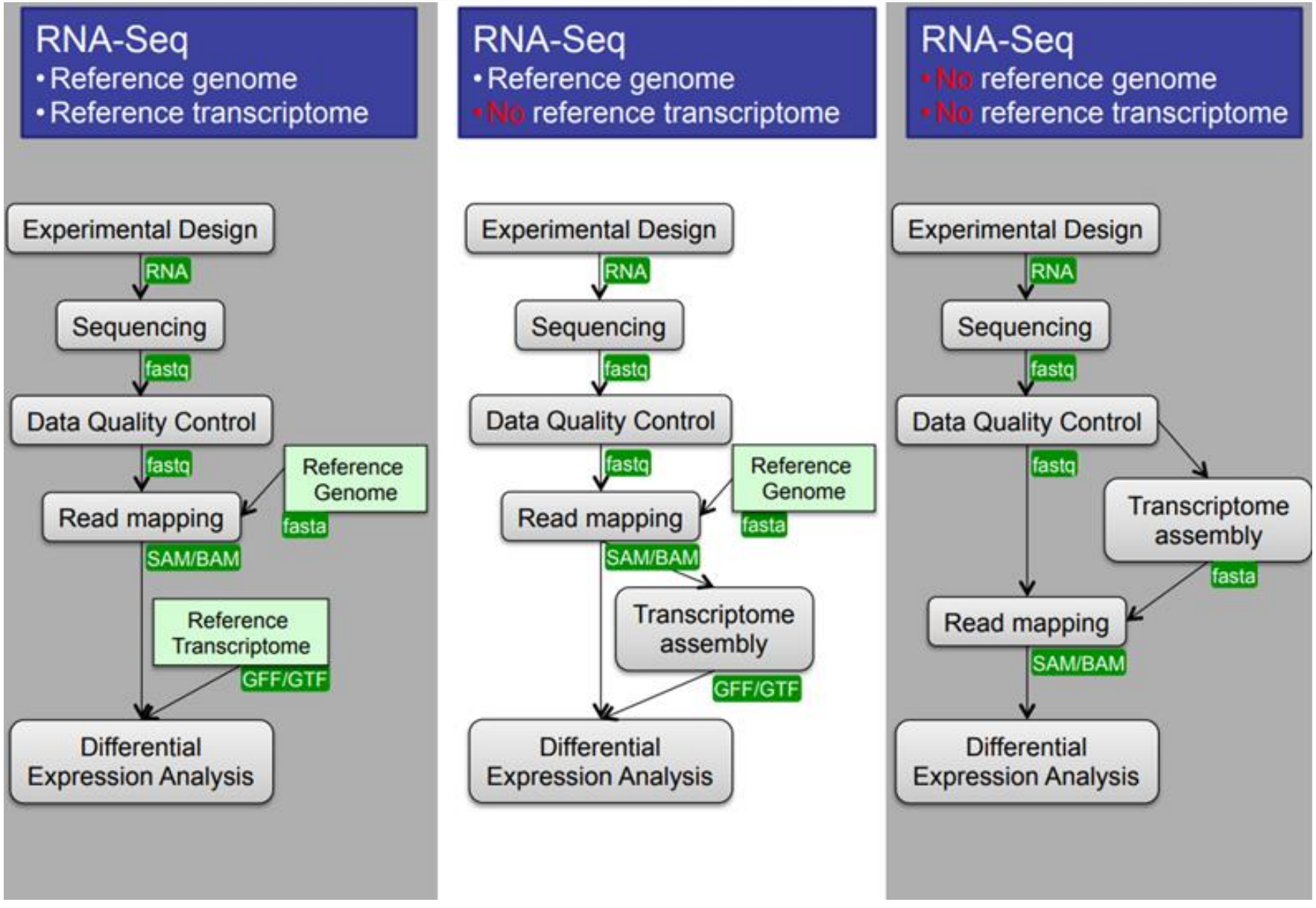
- Linux /ubuntu system

Transcript Assembly

Mapping / reference based

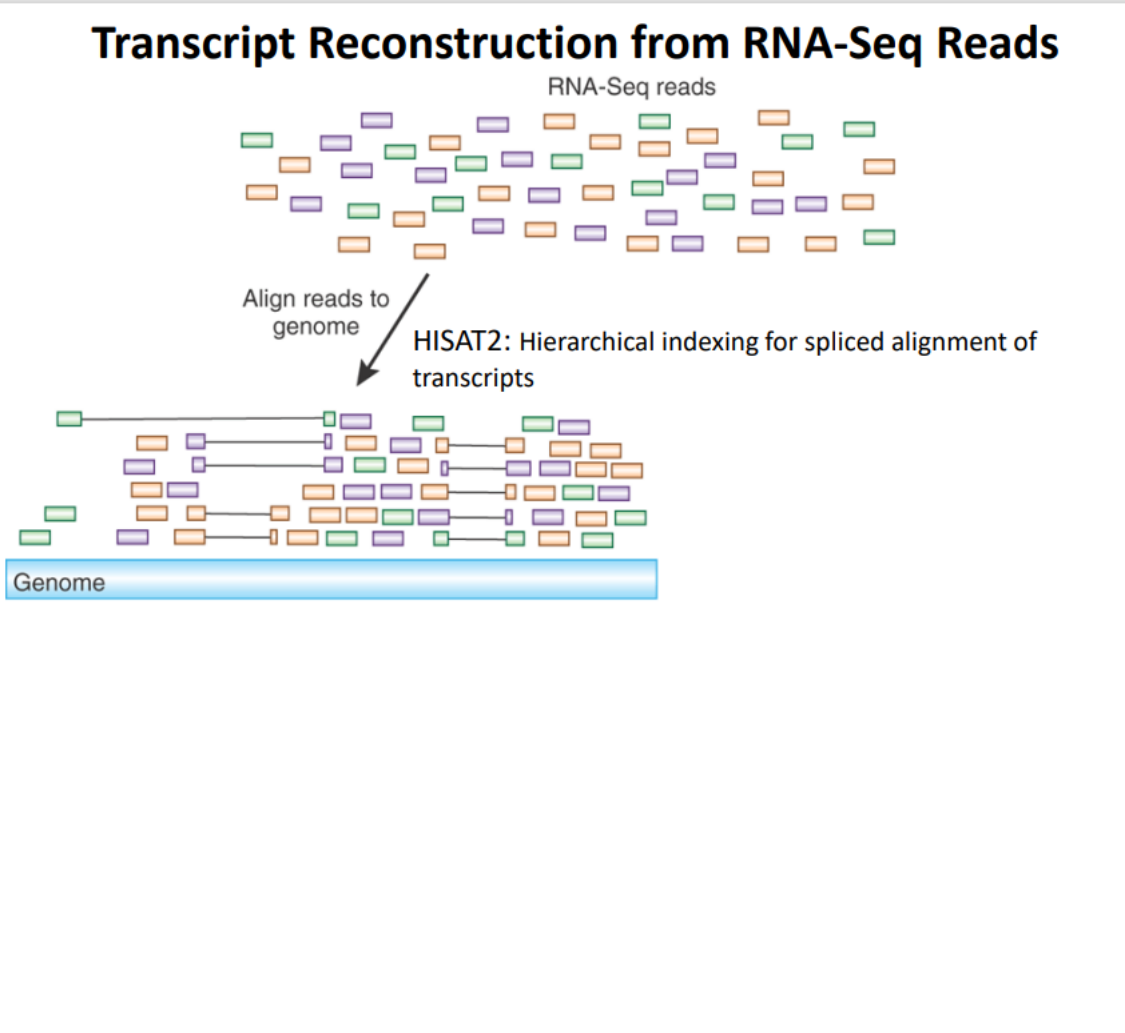
De novo assemby

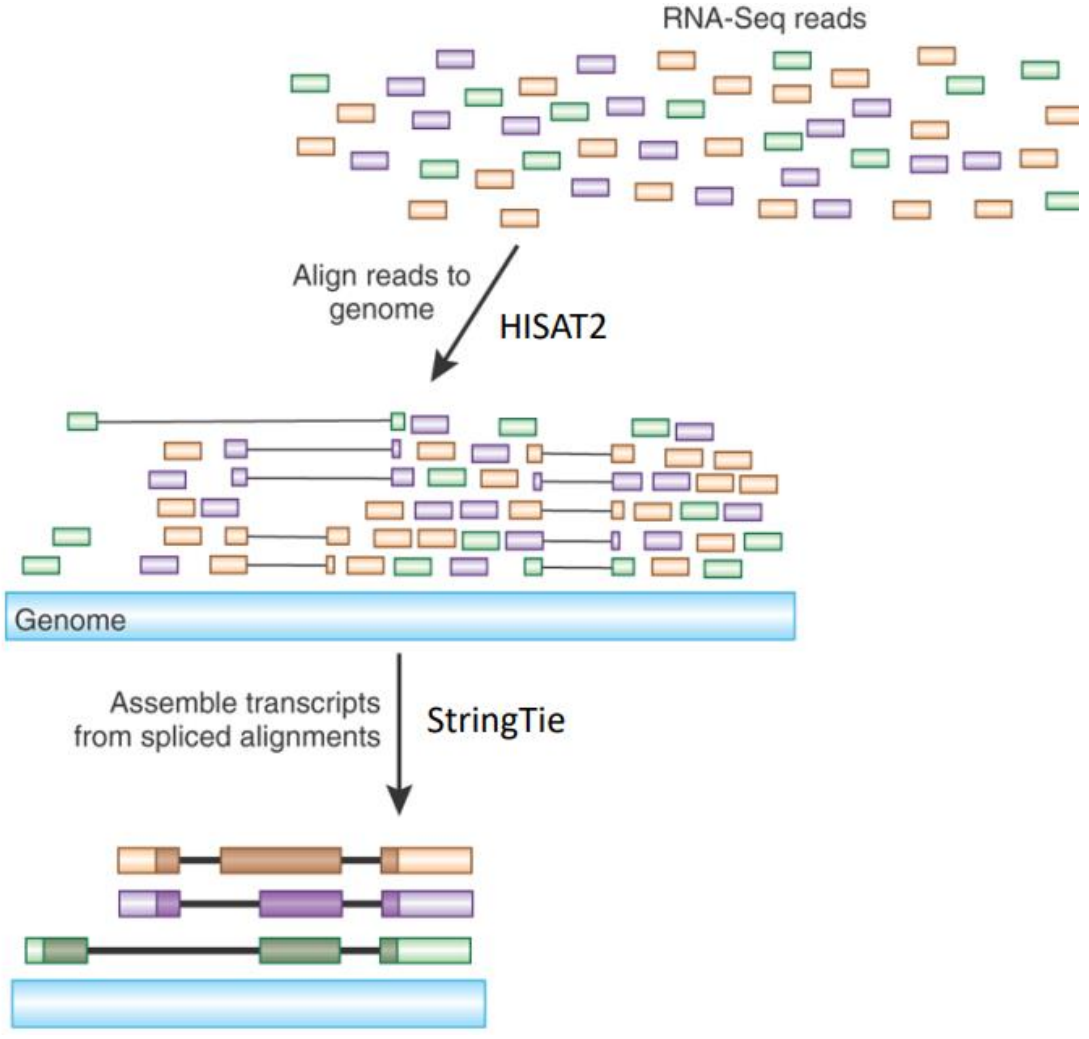
Overview of transcript Assembly



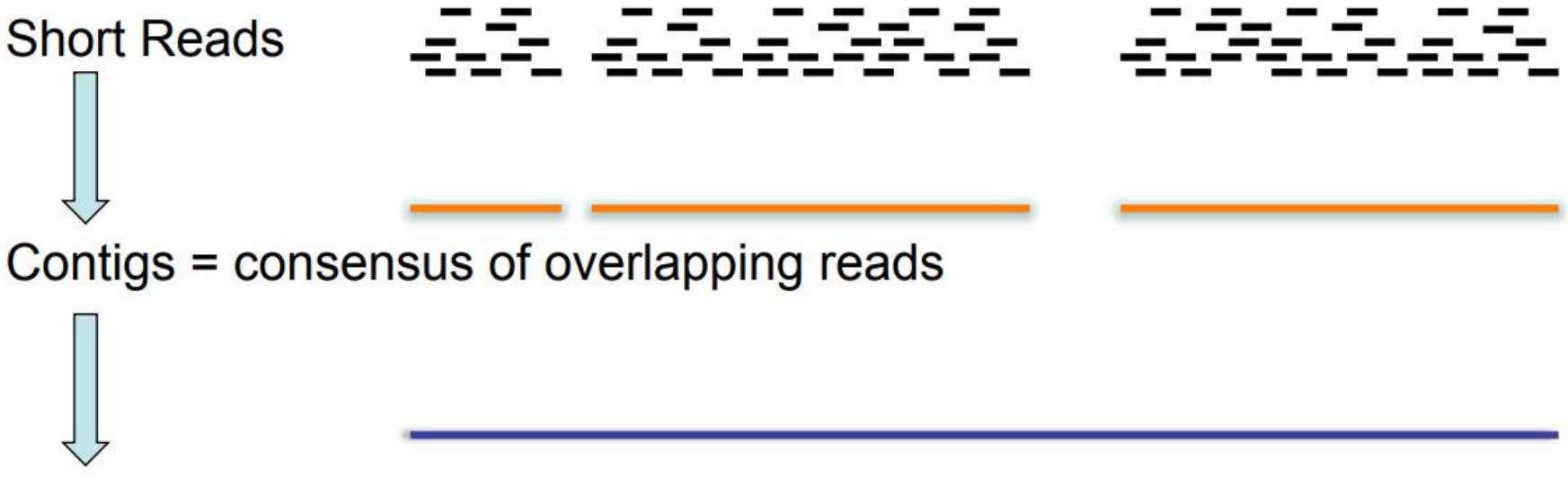
Zhang, 2012

Mapping reads to the genome





Key Definitions



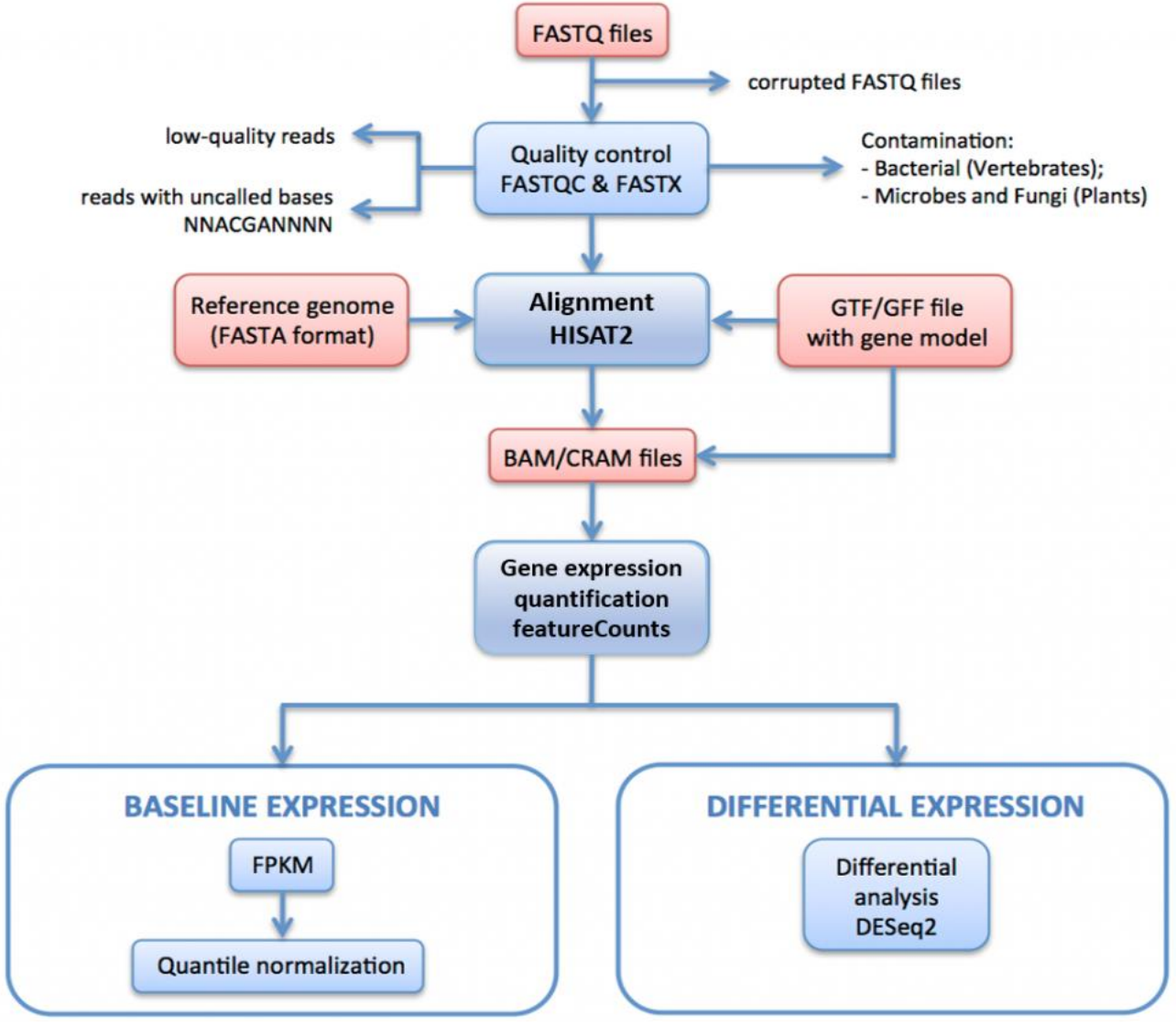
Draft transcriptome/genome = a collection of **non-ordered** scaffolds

Mapping reads to the genome/transcriptome

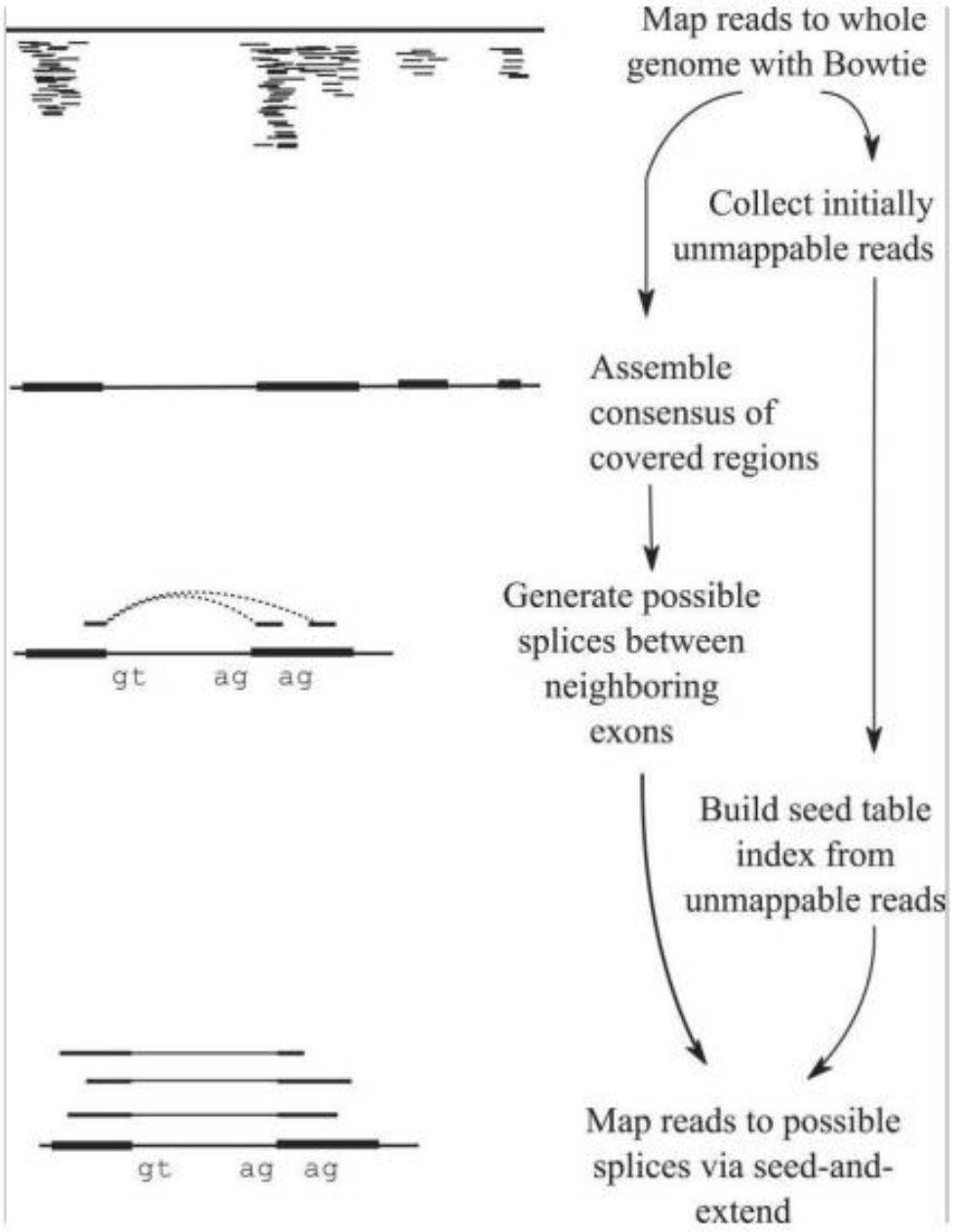
Requirements:

- A reference genome sequence (fasta), or a reference transcriptome (fasta)
- The reference genome/transcriptome is pre-processed by indexing, making mapping faster. • Mapping is guided by a genome annotation file (GFF3 or GTF).
- The trimmed/cleaned (pre-processed) reads (fastq

RNA Sequence Work flow



HISAT2 approach to alignment of reads



Output: SAM/BAM file

Mapping reads to the transcriptome

Pseudo-aligners (kallisto)

- Quasi-mapping
- Probabilistic
- Map to transcripts, not genome
- Does transcript quantifications (or gene)
- FAST and can run on most laptops
- Reports suggest differences between ‘traditional’ mappers are in the low abundance genes.

Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. 34, 525–527 (2016).

Mapping reads to the transcriptome

- Transcript-mapped read counts are normalized for both **length of the transcript and total depth of sequencing.**

Reported as:

- **HISAT2:** Number of RNA-Seq **F**ragments **P**er **K**ilobase of transcript
- per total **M**illion fragments mapped: **FPKM (PE**), or **RPKM** (reads per kb per M) (SE)
- **kallisto: T**ranscripts **Per M**illion
- These methods allow for comparison of features with different length
- **WITHIN a sample** but not BETWEEN samples

List of mapping methods

Class	Category	Package	Notes	Uses	Input
Read mapping					
Unspliced aligners[a]	Seed methods	Short-read mapping package (SHRiMP)[41]	Smith-Waterman extension	Aligning reads to a reference transcriptome	Reads and reference transcriptome
		Stampy[39]	Probabilistic model		
	Burrows-Wheeler transform methods	Bowtie[43]			
		BWA[44]	Incorporates quality scores		
Spliced aligners	Exon-first methods	MapSplice[52]	Works with multiple unspliced aligners	Aligning reads to a reference genome. Allows for the identification of novel splice junctions	Reads and reference genome
		SpliceMap[50]			
		TopHat[51]	Uses Bowtie alignments		
	Seed-extend methods	GSNAP[53]	Can use SNP databases		
		QPALMA[54]	Smith-Waterman for large gaps		

SAM/BAM format

Two section: header section, alignment section

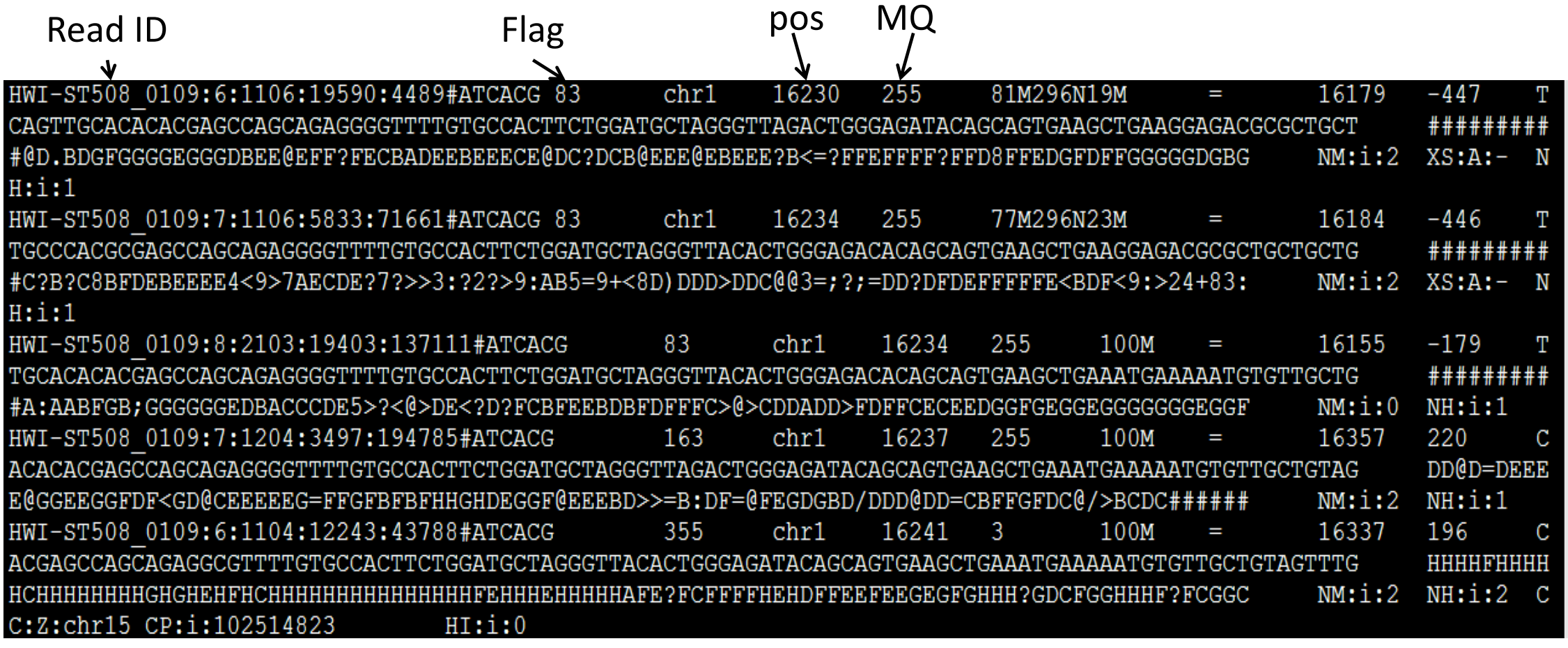
Each alignment line has 11 mandatory fields. These fields always appear in the same order and must be present, but their values can be '0' or '*' (depending on the field) if the corresponding information is unavailable. The following table gives an overview of the mandatory fields in the SAM format:

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,255}`	Query template NAME
2	FLAG	Int	$[0,2^{16}-1]$	bitwise FLAG
3	RNAME	String	`\*|[!-()+-<>-~][!-~]*`	Reference sequence NAME
4	POS	Int	$[0,2^{29}-1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0,2^{8}-1]$	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[!-()+-<>-~][!-~]*`	Ref. name of the mate/next segment
8	PNEXT	Int	$[0,2^{29}-1]$	Position of the mate/next segment
9	TLEN	Int	$[-2^{29}+1,2^{29}-1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	
11	QUAL	String	`[!-~]+`	

Bit	Description
0x1	template having multiple segments in sequencing
0x2	each segment properly aligned according to the aligner
0x4	segment unmapped
0x8	next segment in the template unmapped
0x10	SEQ being reverse complemented
0x20	SEQ of the next segment in the template being reversed
0x40	the first segment in the template
0x80	the last segment in the template
0x100	secondary alignment
0x200	not passing quality controls
0x400	PCR or optical duplicate

http://samtools.sourceforge.net/SAM1.pdf

One example: SAM file



83= 1+2+16+64

read paired; read mapped in proper pair; read reverse strand; first in pair

GFF/GTF files

The GFF (General Feature Format) format consists of one line per feature, each containing 9 columns of data (fields).

The GTF (General Transfer Format) is identical to GFF version 2.

Fields must be tab-separated and all fields must contain a value; "empty" fields should be denoted with a '.'.

chr12	unknown	exon	4382902	4383401	.	+	.	gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
chr12	unknown	CDS	4383207	4383401	.	+	.	gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
chr12	unknown	start_codon	4383207	4383209	.	+	.	gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
chr12	unknown	CDS	4385171	4385386	.	+	.	gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
chr12	unknown	exon	4385171	4385386	.	+	.	gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
chr12	unknown	CDS	4387926	4388085	.	+	.	gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
chr12	unknown	exon	4387926	4388085	.	+	.	gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
chr12	unknown	CDS	4398008	4398156	.	+	.	gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
chr12	unknown	exon	4398008	4398156	.	+	.	gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
chr12	unknown	CDS	4409026	4409172	.	+	.	gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
chr12	unknown	exon	4409026	4414522	.	+	.	gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";
chr12	unknown	stop_codon	4409173	4409175	.	+	.	gene_id "CCND2"; gene_name "CCND2"; p_id "P6197"; transcript_id "NM_001759"; tss_id "TSS231";

The left columns list source, feature type, and genomic coordinates

The right column includes attributes, including gene ID, etc.

Mapping Quality Checking

Information we need to check

- Percentage of reads properly mapped or uniquely mapped
- Among the mapped reads, the percentage of reads in exon, intron, and intergenic regions.
- 5' or 3' bias
- The percentage of expressed genes

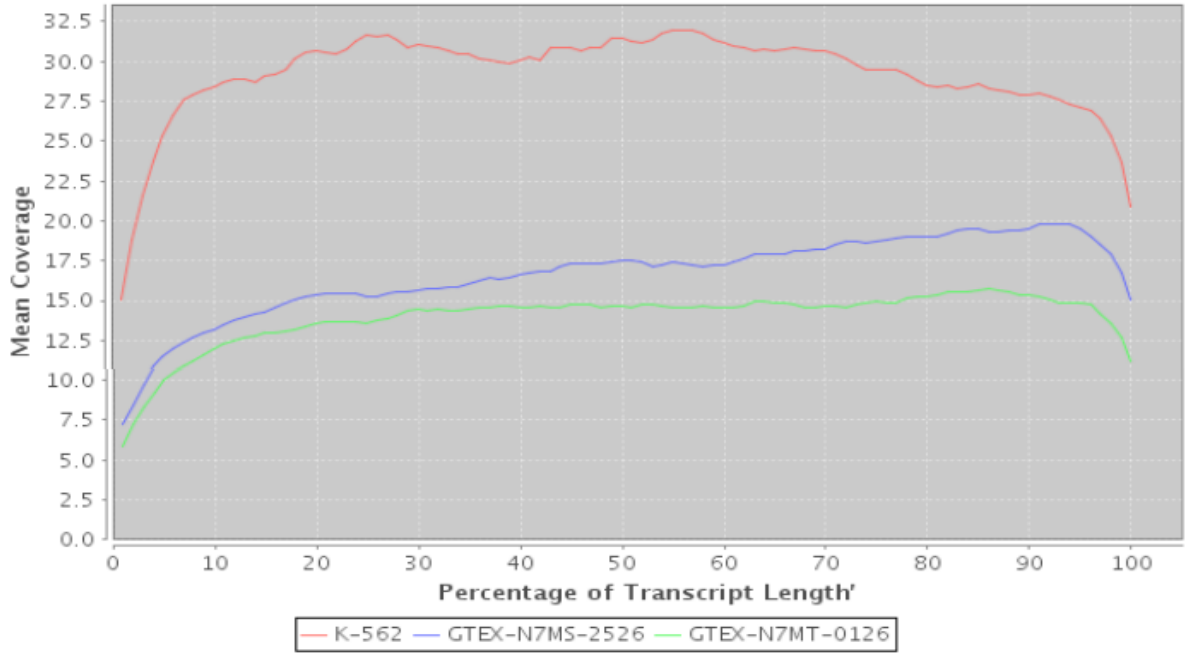
Genome analysis **2012, Bioinformatics** Advance Access publication April 25, 2012

RNA-SeQC: RNA-seq metrics for quality control and process optimization

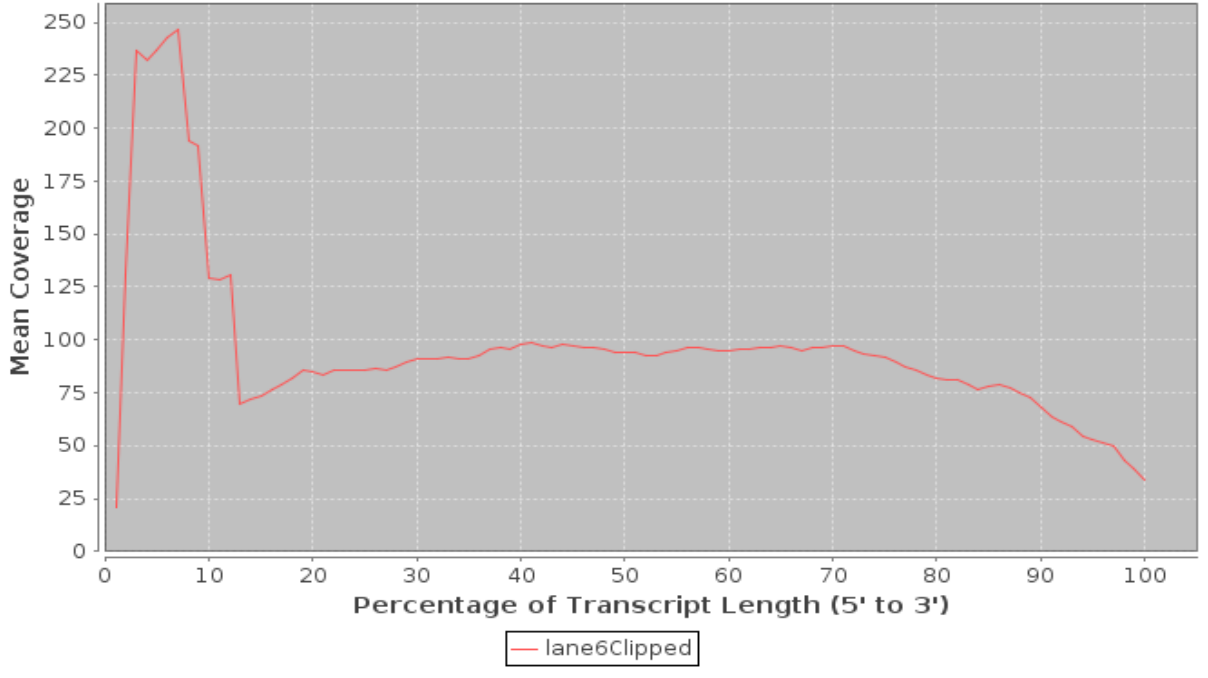
David S. DeLuca*, Joshua Z. Levin, Andrey Sivachenko, Timothy Fennell,
Marc-Danie Nazaire, Chris Williams, Michael Reich, Wendy Winckler and Gad Getz*
The Broad Institute of MIT and Harvard, Cambridge, MA, USA

- Read Metrics
  - Total, unique, duplicate reads
  - Alternative alignment reads
  - Read Length
  - Fragment Length mean and standard deviation
  - Read pairs: number aligned, unpaired reads, base mismatch rate for each pair mate, chimeric pairs
  - Vendor Failed Reads
  - Mapped reads and mapped unique reads
  - rRNA reads
  - Transcript-annotated reads (intragenic, intergenic, exonic, intronic)
  - Expression profiling efficiency (ratio of exon-derived reads to total reads sequenced)
  - Strand specificity
- Coverage
  - Mean coverage (reads per base)
  - Mean coefficient of variation
  - 5'/3' bias
  - Coverage gaps: count, length
  - Coverage Plots
- Downsampling
- GC Bias
- Correlation:
  - Between sample(s) and a reference expression profile

The following plot show the mean coverage for expressed transcripts from 5’ to 3’ end, with the lengths of transcripts normalized to 1-100.

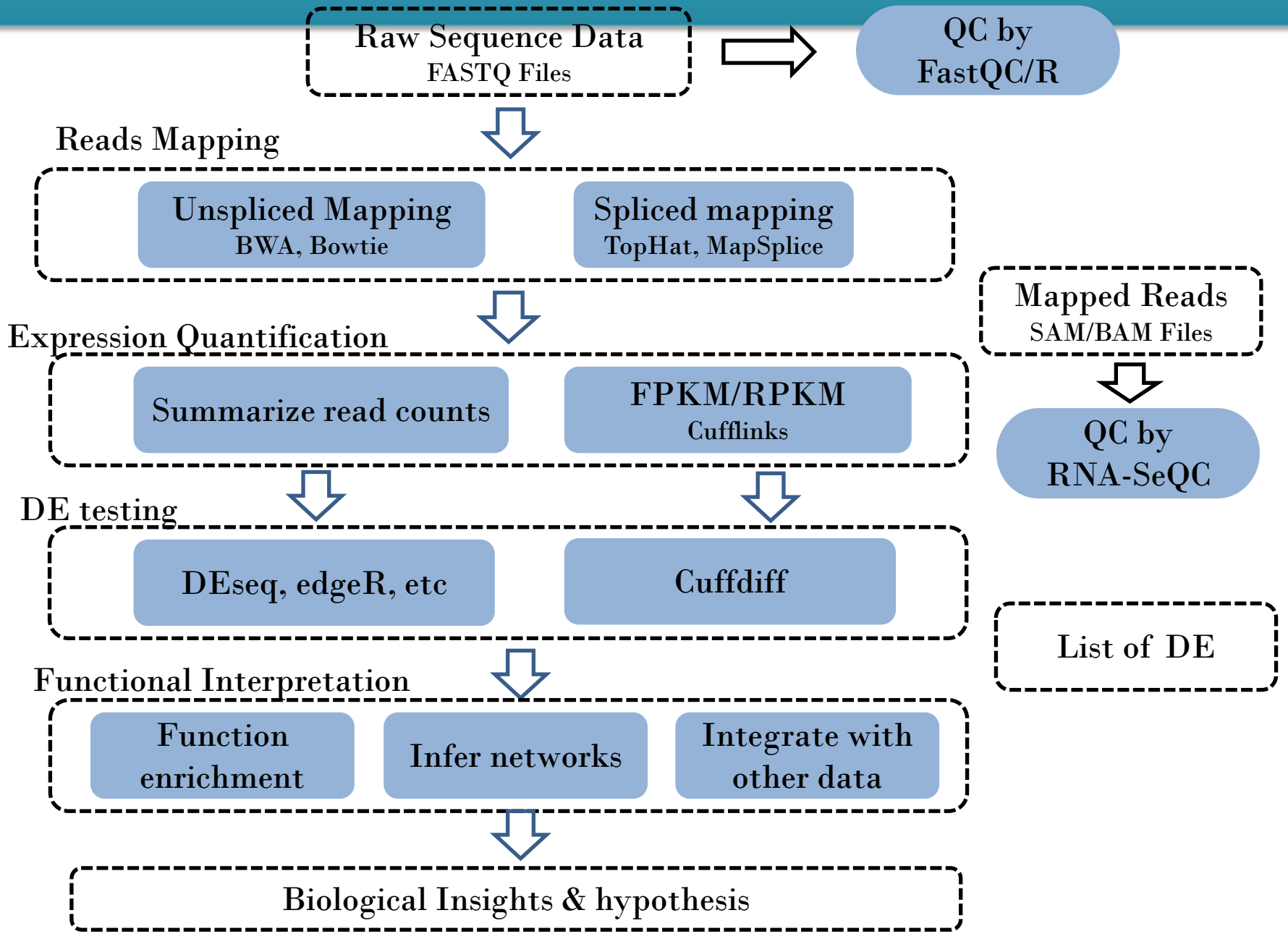


No 5' or 3' bias



5' bias

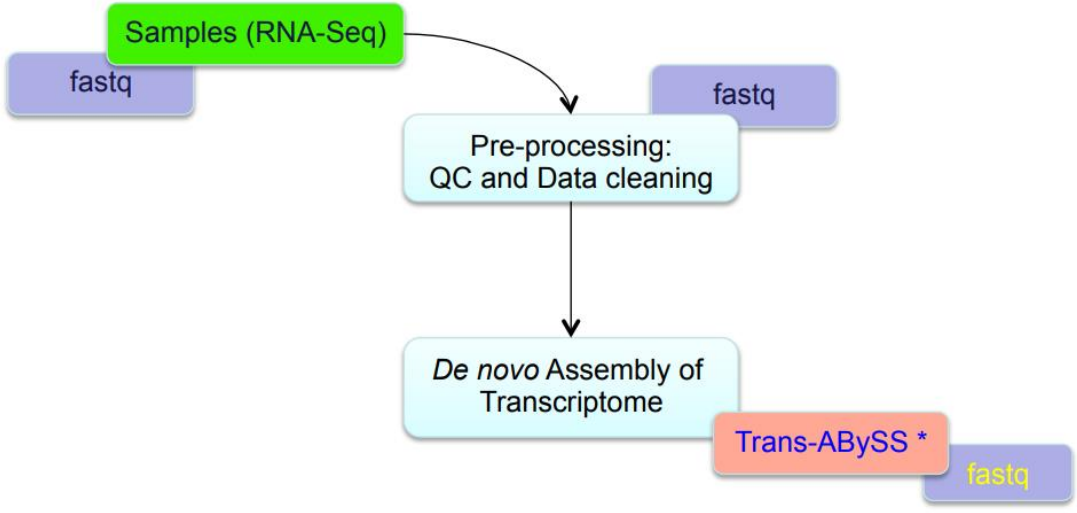
From reads to differential expression



Transcript Assembly (de novo)

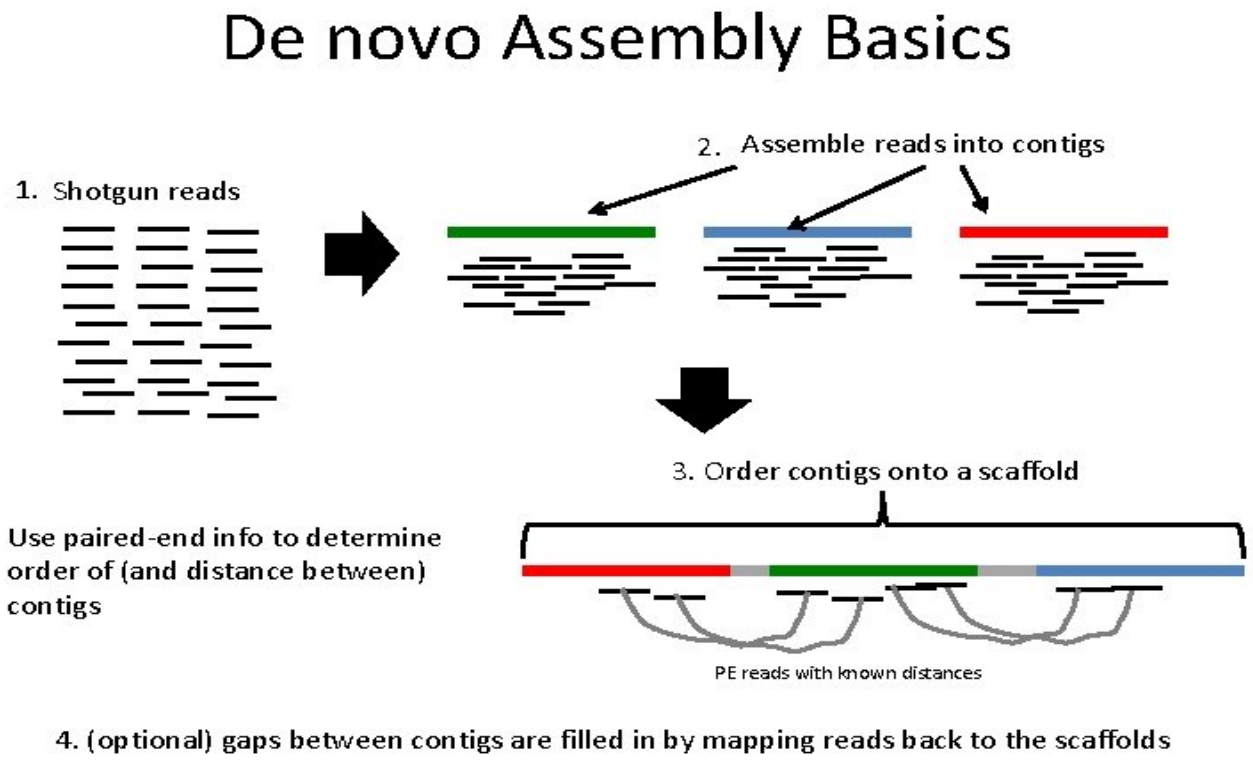
De novo assembly is a method for constructing genomes from a large number of (short- or long-) DNA fragments, with no a priori knowledge of the correct sequence or order of those fragments.

- without the use of a reference genome/ transriptome



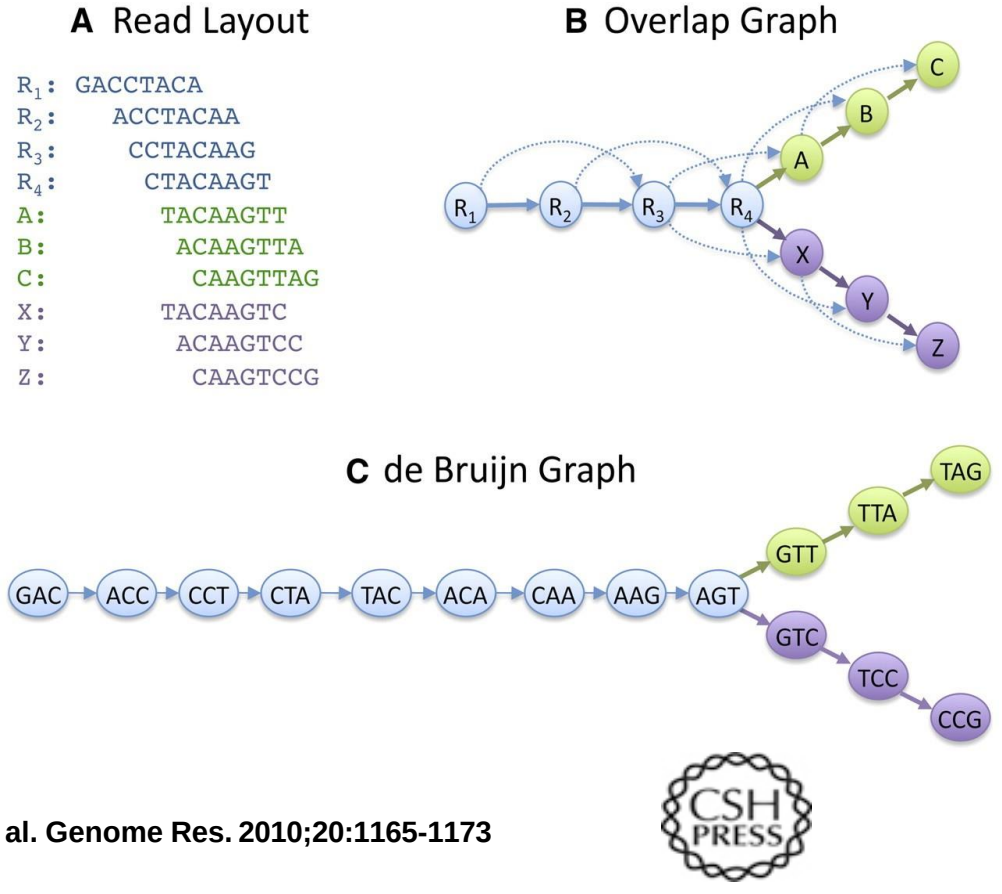
De novo assembly basics

- Find all overlaps between reads
- Build a graph
- Simplify the graph (sequencing errors)
- Traverse a graph to produce a consensus.



Assembly Algorithms

1. Greedy
2. Overlap-layout- consensus (OLC)
3. De Bruijn Graph



Schatz M C et al. Genome Res. 2010;20:1165-1173

De novo assembly tools

Some Assemblers



Canada's Michael Smith Genome Sciences Centre



Beijing Genomics Institute

ALLPATHS-LG





Platanus Kajitani et al., 2014

PLATform for Assembling NUcleotide Sequences

Meraculous – Joint Genomes Institute

MaSURCA – Maryland Genome Assembly Group

Trans-ABySS Facts

- ABySS is a *de novo*, parallel sequence assembler that is designed for short reads.
  - Can work on single end reads and paired end reads.
  - Is a de Bruijn graph assembler
  - It takes two steps:
    - Using all possible k-mers from the reads to build the initial contigs
    - Using mate-pair information to extend contigs
- **Trans-ABySS** is a pipeline for analyzing ABySS-assembled contigs from RNA-Seq data.
  - Use several k-mer length
- Availability: Command line
- Homepage:
  - http://www.bcgsc.ca/platform/bioinfo/software/abyss
  - http://www.bcgsc.ca/platform/bioinfo/software/trans-abyss

Quantification and Visualization

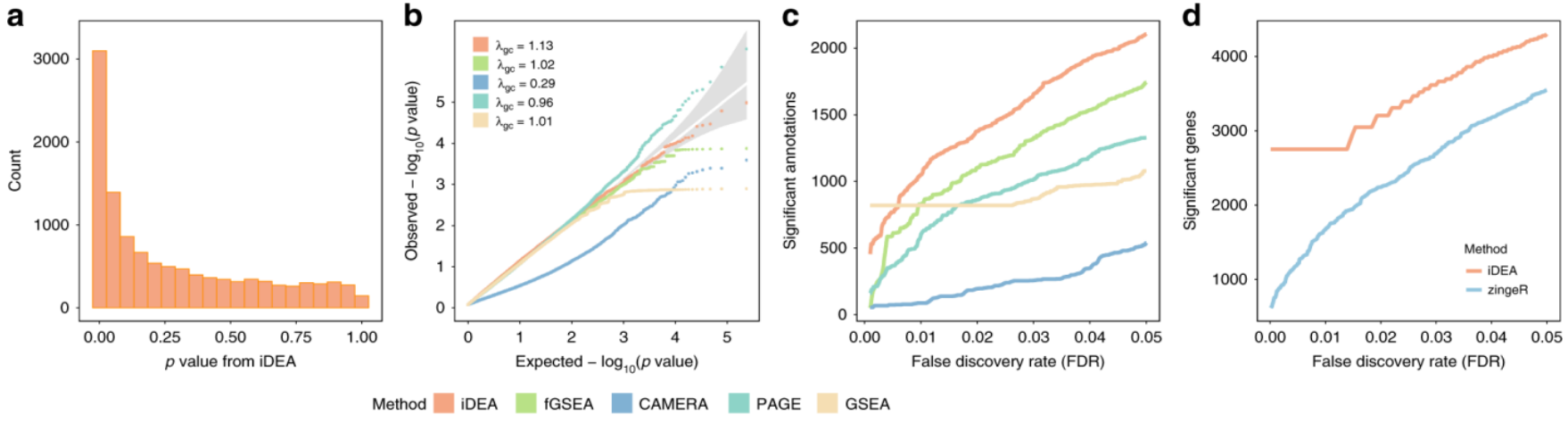
Differential gene expression analysis

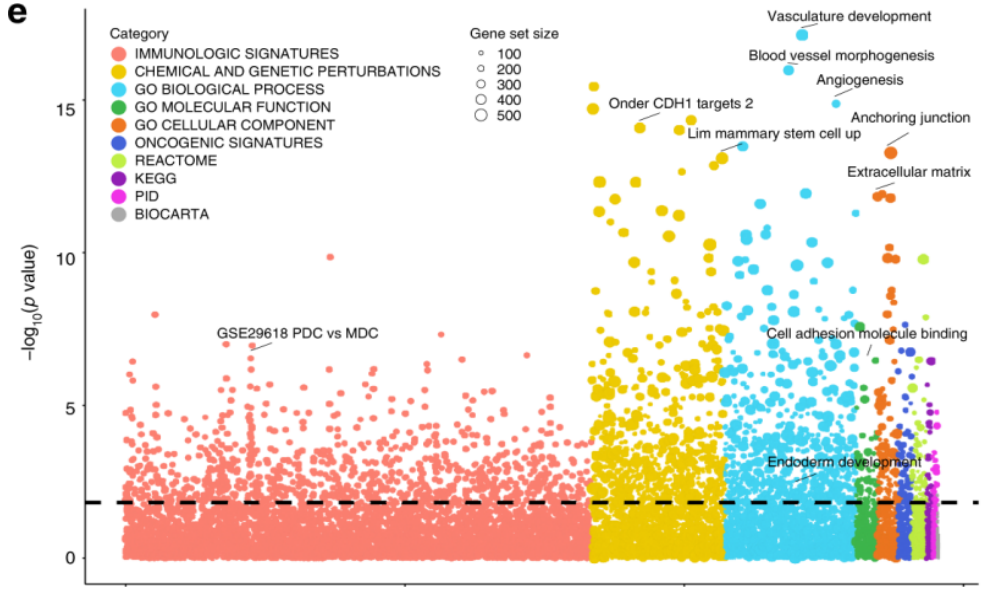
- edgeR and DESeq based on negative binomial (NB) distributions or
- baySeq and EBSeq which are Bayesian approaches based on a negative binomial model

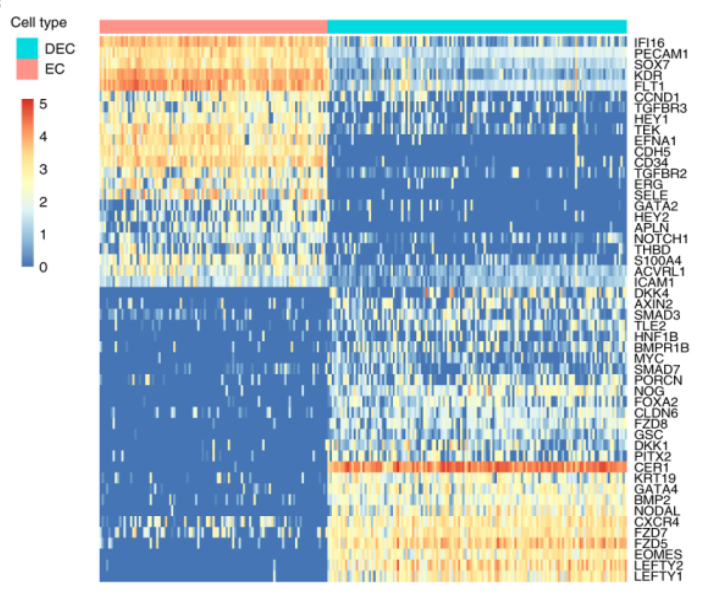
Biological interpretation of gene expression data

- ✓ Heat maps and Clustering
- ✓ Gene set enrichment analysis and pathway analysis
- ✓ Network Analysis

Examples of Transcrptomic visualization







Links and resources

- https://bioinformatics.uconn.edu/resources-and-events/tutorials-2/rna-seq-tutorial-with-reference-genome/
- https://www.uptodate.com/contents/tools-for-genetics-and-genomics-gene-expression-profiling
- https://www.arabidopsis.org/portals/expression/microarray/microarraySoftwareV2.jsp
- https://www.bio-rad.com/en-et/applications-technologies/what-gene-expression-analysis?ID=LUSNINKSY
- https://guides.library.yale.edu/bioinformatics/gene-prediction
-


Thank You