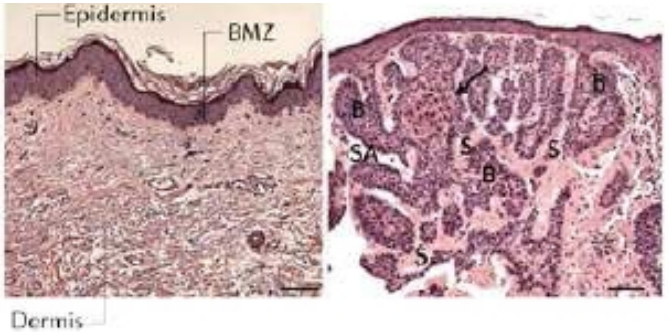
RNA-seq experiment overview

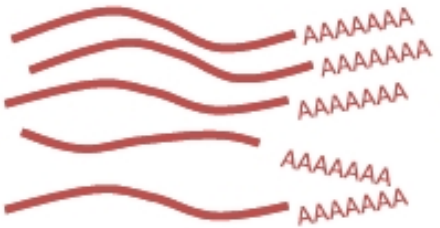
mRNA-seq experiment workflow

A

B



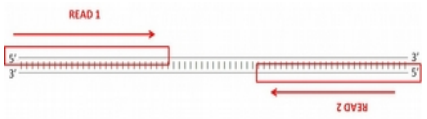
mRNA



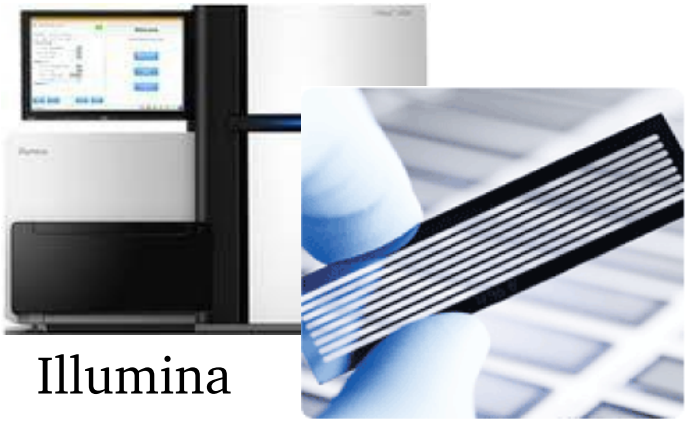
fragmentation

conversion into cDNA and adapter ligation

cDNA fragment



reads



Illumina sequencer

flowcell (8 lanes)

library

RNA-seq single- or paired- end (library preparation)

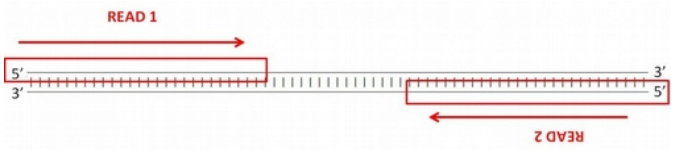
a population of RNA is converted to a library of cDNA fragments

each molecule is then sequenced in a high-throughput parallel manner to obtain short sequences (reads) from one end (single-end sequencing) or both ends (paired-end sequencing)

RNA-seq single- or paired- end (library preparation)

a population of RNA is converted to a library of cDNA fragments

each molecule is then sequenced in a high-throughput parallel manner to obtain short sequences (reads) from one end (single-end sequencing) or both ends (paired-end sequencing)

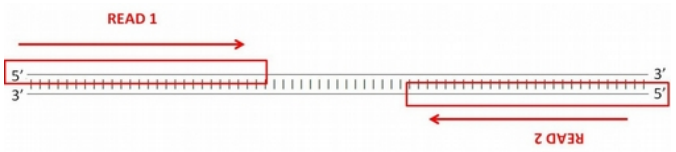


cDNA fragment
(200-300 bp)

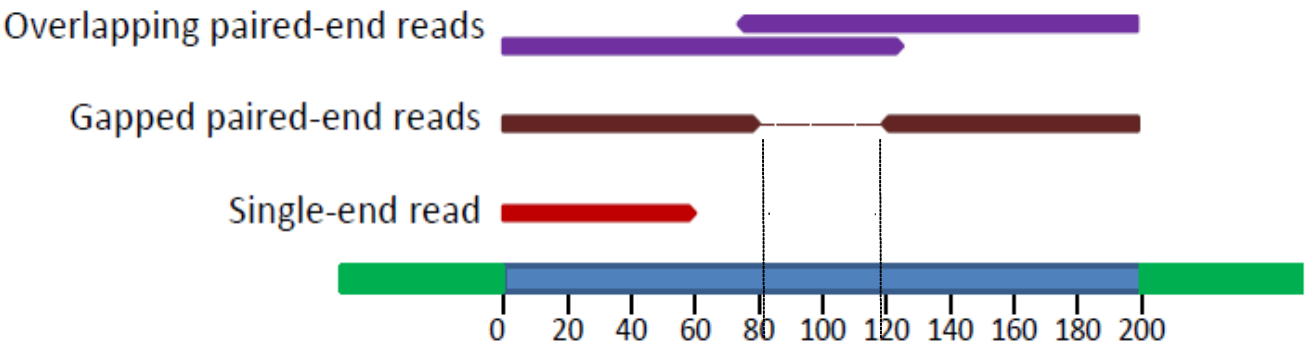
RNA-seq single- or paired- end (library preparation)

a population of RNA is converted to a library of cDNA fragments

each molecule is then sequenced in a high-throughput parallel manner to obtain short sequences (reads) from one end (single-end sequencing) or both ends (paired-end sequencing)



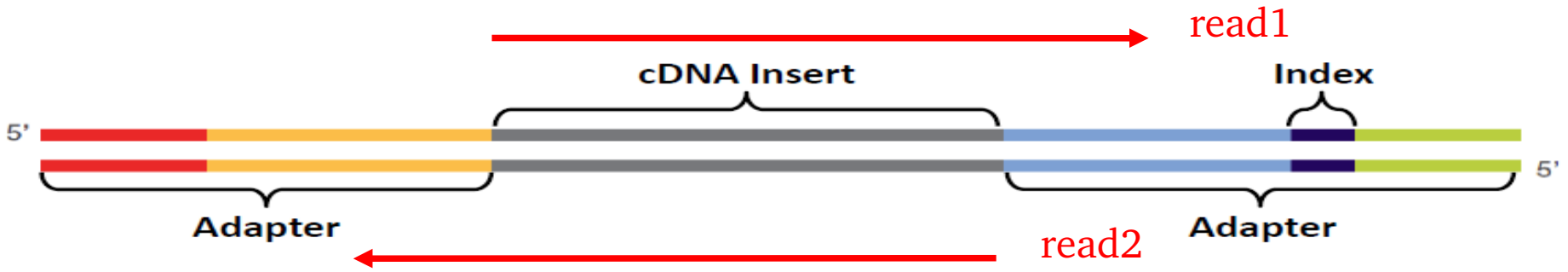
cDNA fragment
(200-300 bp)



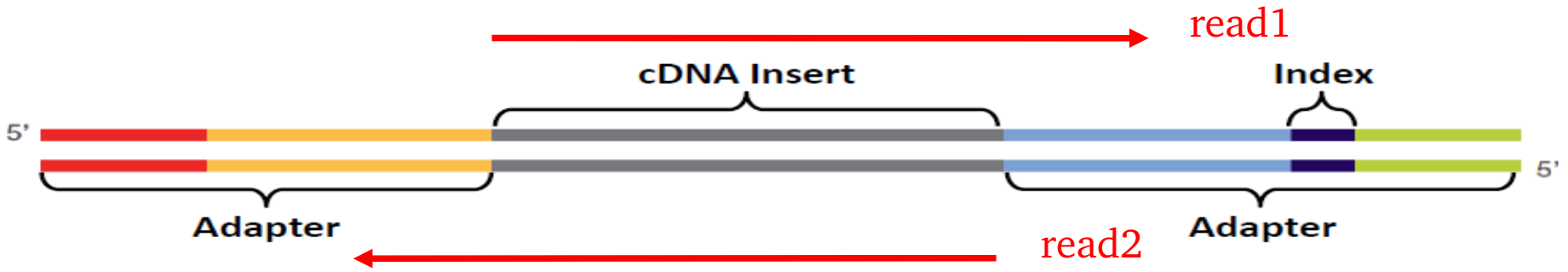
inner distance
(+ or -)

adapter cDNA fragment or insert adapter

RNA-seq adapter artifacts or contaminations



RNA-seq adapter artifacts and contaminations



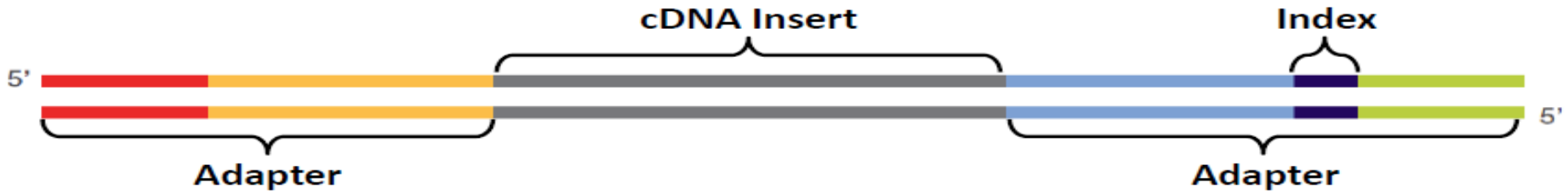
cDNA insert size is a distribution (not a constant value)

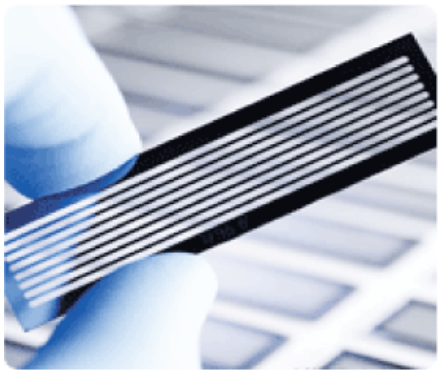
- reads can contain adapter sequence at 3' end

Removal of adapter can improve:

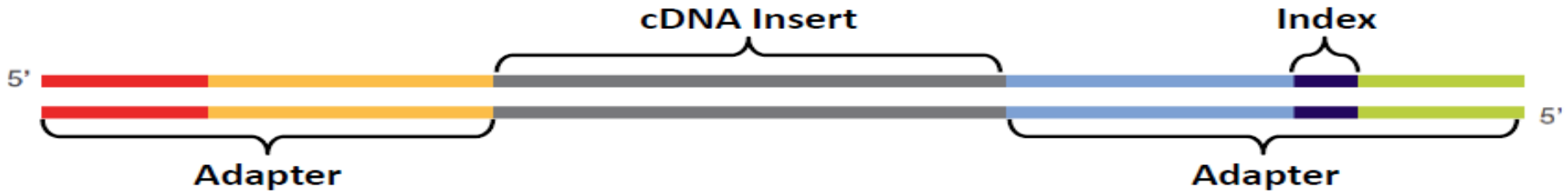
- alignment to the reference
- very important for de novo assemblies

RNA-seq multiplexing (library preparation)





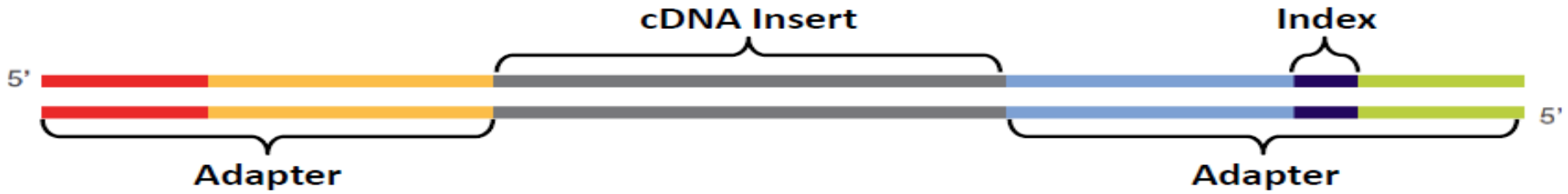
RNA-seq multiplexing (library preparation)



short (6-8 nt), unique barcodes (index) as part of adapters

a unique identifier for each sample

RNA-seq multiplexing (library preparation)



short (6-8 nt), unique barcodes (index) as part of adapters

a unique identifier for each sample

allows pooling samples to mitigate biases effects

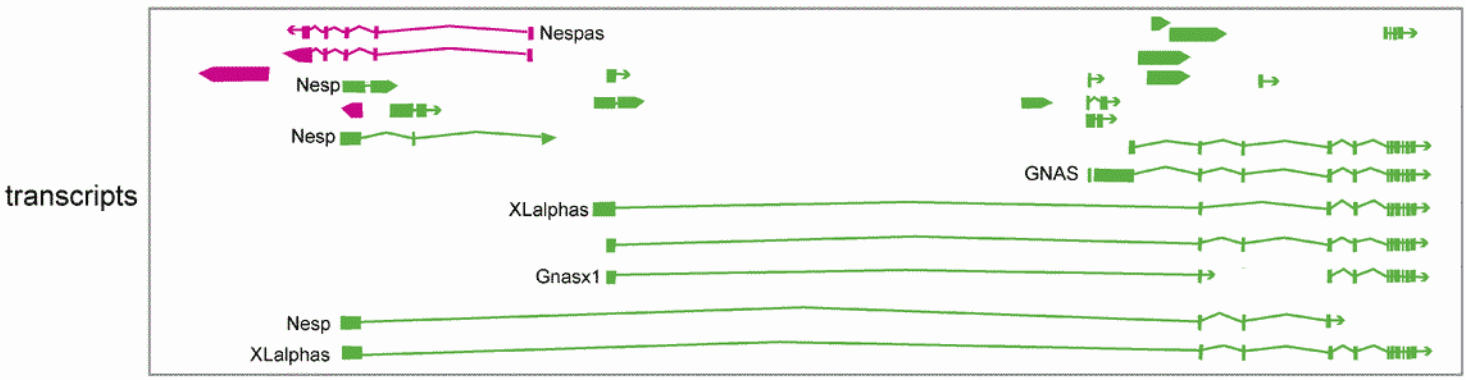
allows sequencing capacity to be used efficiently

dual barcodes allow deep multiplexing
(e.g. 96 samples)

Strand-specific/directional RNA-seq (library preparation)

Orientation of RNA is **mantained** after reverse transcription to cDNA

Useful to

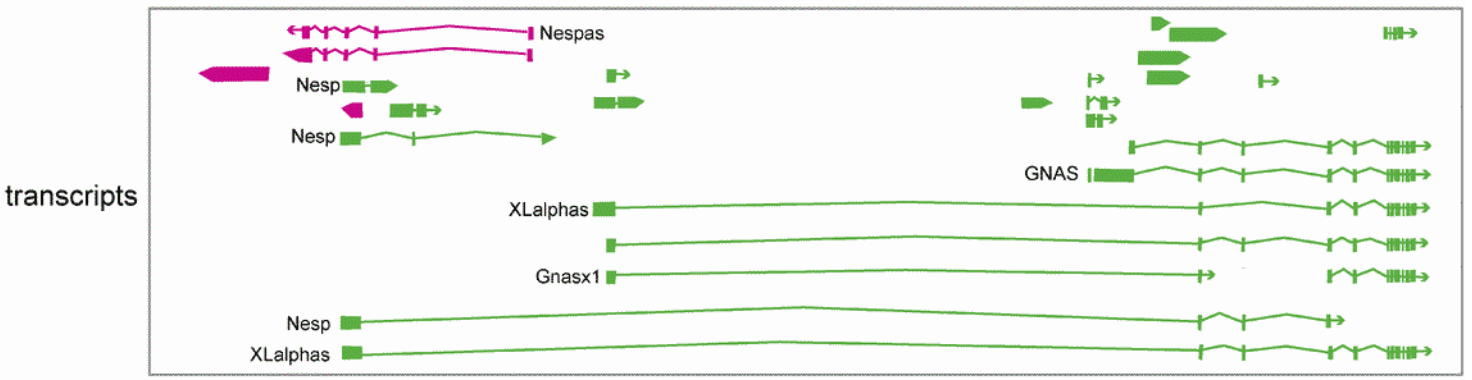


Strand-specific/directional RNA-seq (library preparation)

Orientation of RNA is mantained after reverse transcription to cDNA

Useful to discern among overlapping transcripts in different directions

- identify sense and anti-sense transcripts
- quantify more precisely
- useful in de novo transcriptome assembly



Basic workflow to generate RNA-seq data

Illumina platform

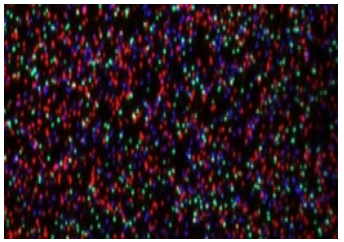


image production
(one per cycle)

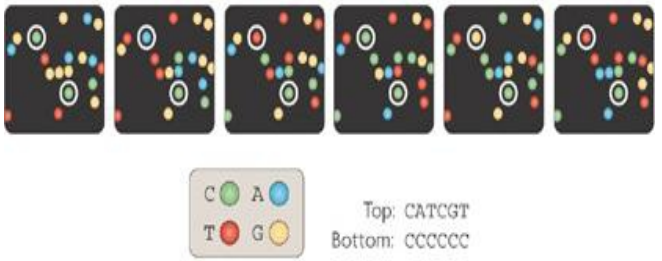


image analysis
and
intensity extraction
base calling
(**.bcl**)

raw data
output file
(**.fastq**)

```
@ILLUMINA-C3C24B:1:1:0:1478#0/1
NACTAATCCTGTGGGAAGGAGCTGGGCCCTGGAACA
+ILLUMINA-C3C24B:1:1:0:1478#0/1
B`_a___``bba^_`ba_aWb`ba^b`V``a\b]`_
@ILLUMINA-C3C24B:1:1:0:95#0/1
NGAGAGGGGTAGGGATTATCTTCAAAGCACCCCAGC
+ILLUMINA-C3C24B:1:1:0:95#0/1
BaaaX___aaTaaaabbbabbabaaba`bbYa^ab[
```

NGS FORMAT FILES

Fasta

Fastq

NGS FORMAT FILES

Fastq

@EAS139:136:FC706VJ:2:5:1000:12850 1:Y:18:**ATCACG**
GATTTGGGGTTCAAAGCANNNTCGNTCAAATAGTAAATCCATTTGTT
+
!"*((((***+))%%%++)(%%%###).1#**-+*"))**55CCF>>>>>>CCCCCCC6

Line 1: begins with a '**@**' character and is followed by a sequence identifier and an optional description (like a FASTA header line).
Line 2: the sequence.
Line 3: '+' character, optionally followed by the same sequence identifier (and description, if any) as in line 1 after the '@' sign
Line 4: encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence

NGS FORMAT FILES

Fastq

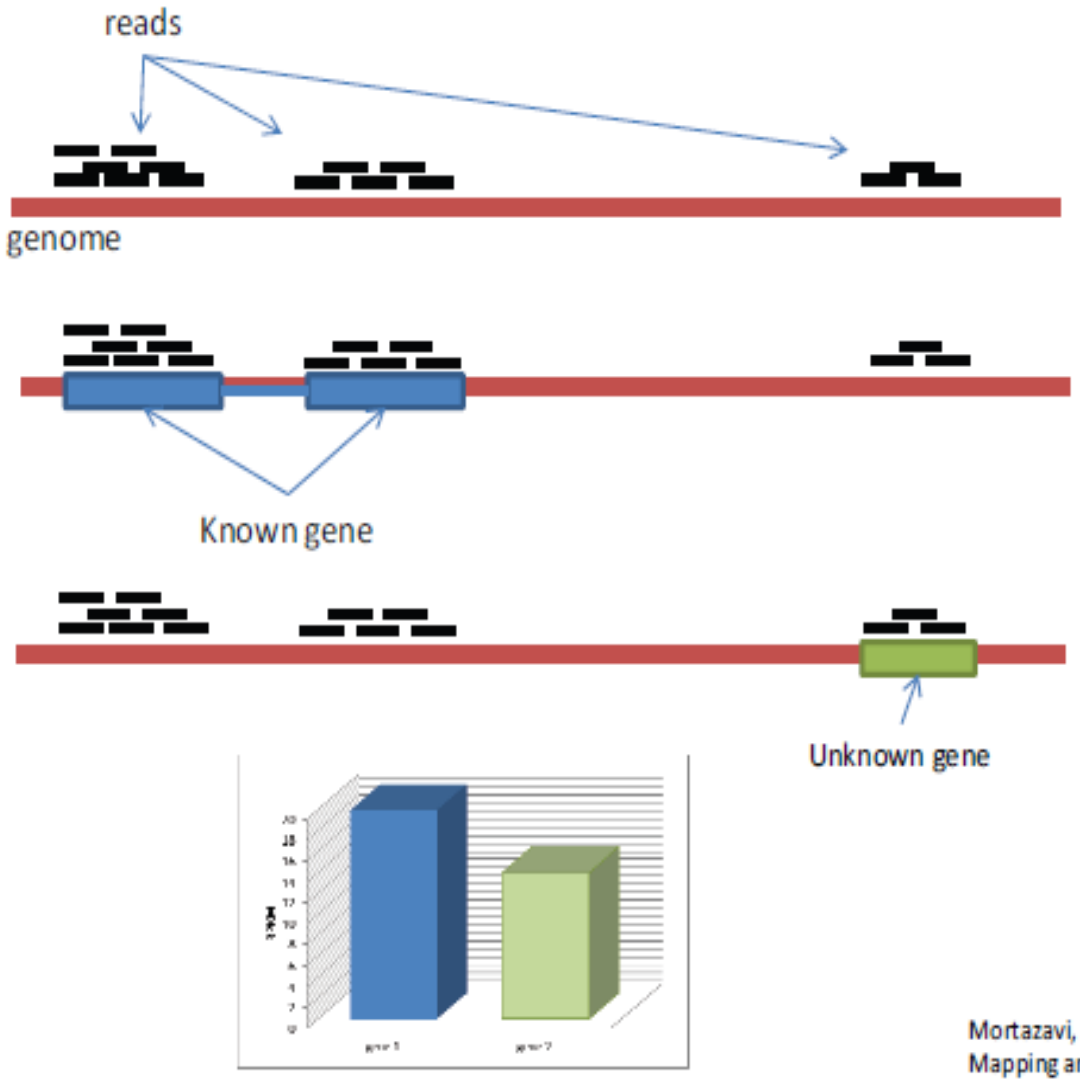
Illumina optional description

Instrument ID | Run number on instrument | flowcell ID | lane | tile | X | Y | single read (1) or read 1 or 2 of paired-end | **Index**

@EAS139:136:FC706VJ:2:5:1000:12850 1:Y:18:**ATCACG**
GATTTGGGGTTCAAAGCANNNTCGNTCAAATAGTAAATCCATTTGTT
+
!"*((((***+))%%%++)(%%%###).1#**-+*"))**55CCF>>>>>>CCCCCCC6

Line 1: begins with a '**@**' character and is followed by a sequence identifier and an optional description (like a FASTA header line).
Line 2: the sequence.
Line 3: '+' character, optionally followed by the same sequence identifier (and description, if any) as in line 1 after the '@' sign
Line 4: encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence

Basic steps to analyse RNA-seq data



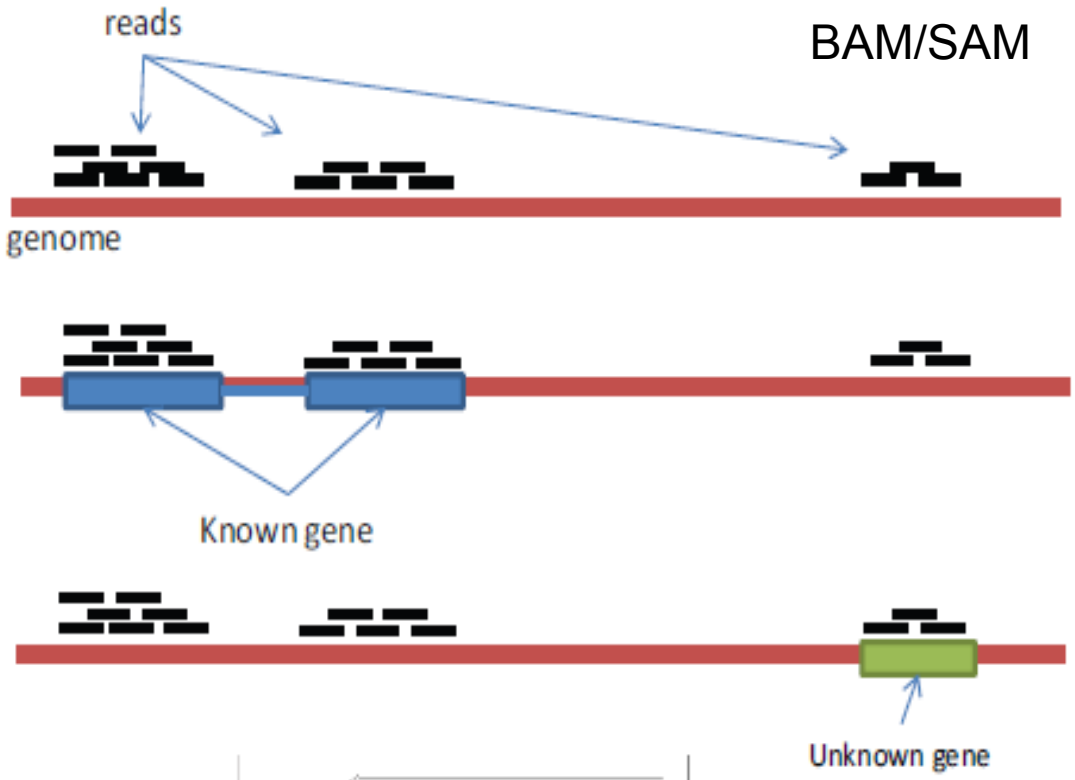
Alignment

Transcriptome reconstruction and quantification

Differential analysis

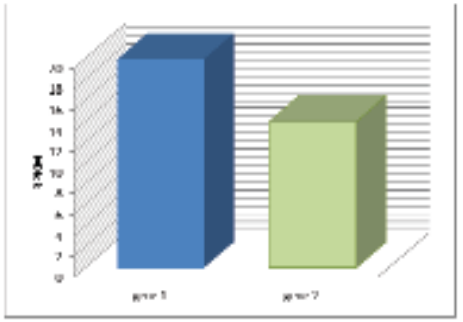
Mortazavi, A.,
Mapping and
methods, 5(7)

Basic steps to analyse RNA-seq data



Alignment

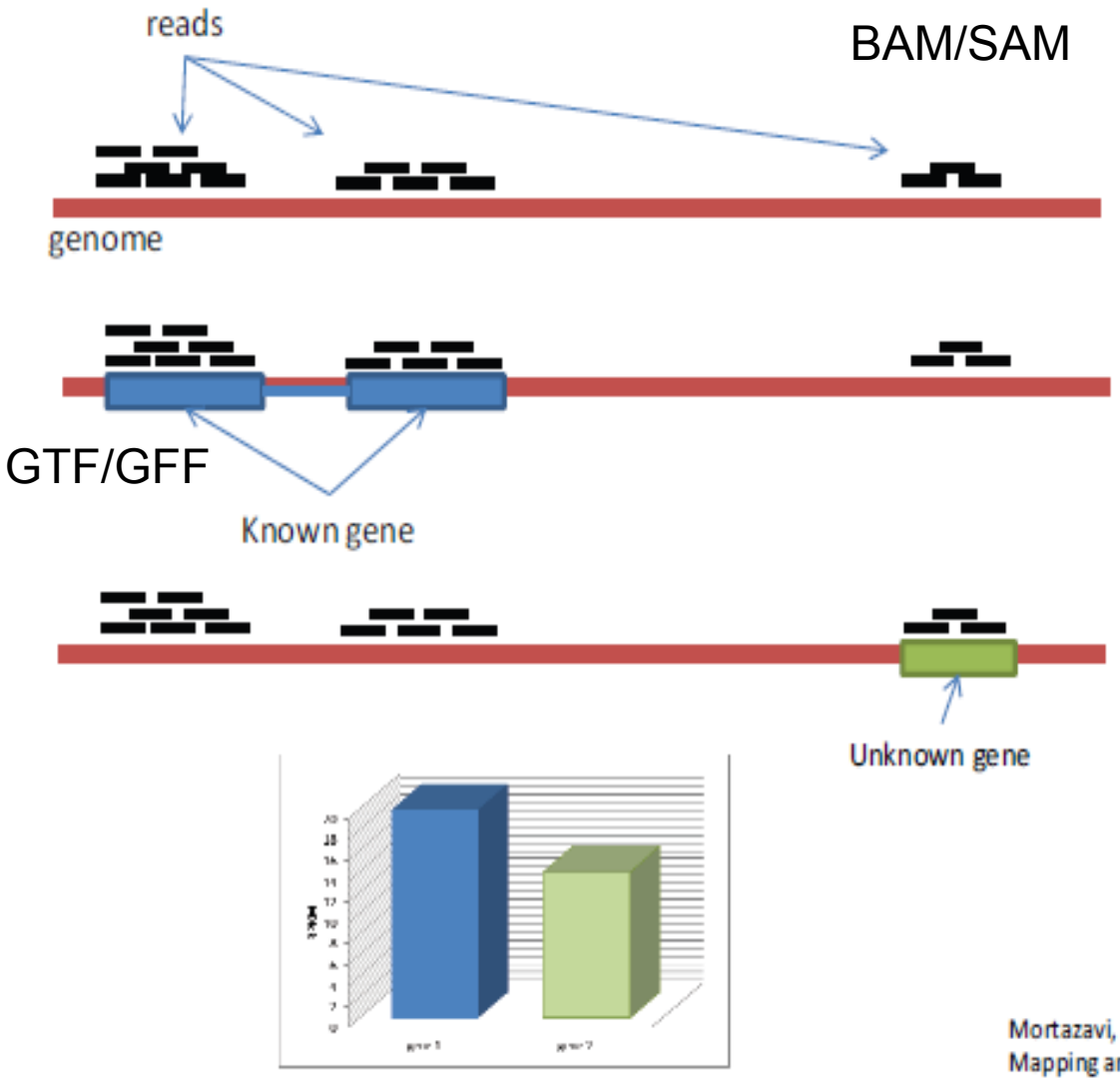
Transcriptome reconstruction and quantification



Differential analysis

Mortazavi, A.,
Mapping and
methods, 5(7)

Basic steps to analyse RNA-seq data



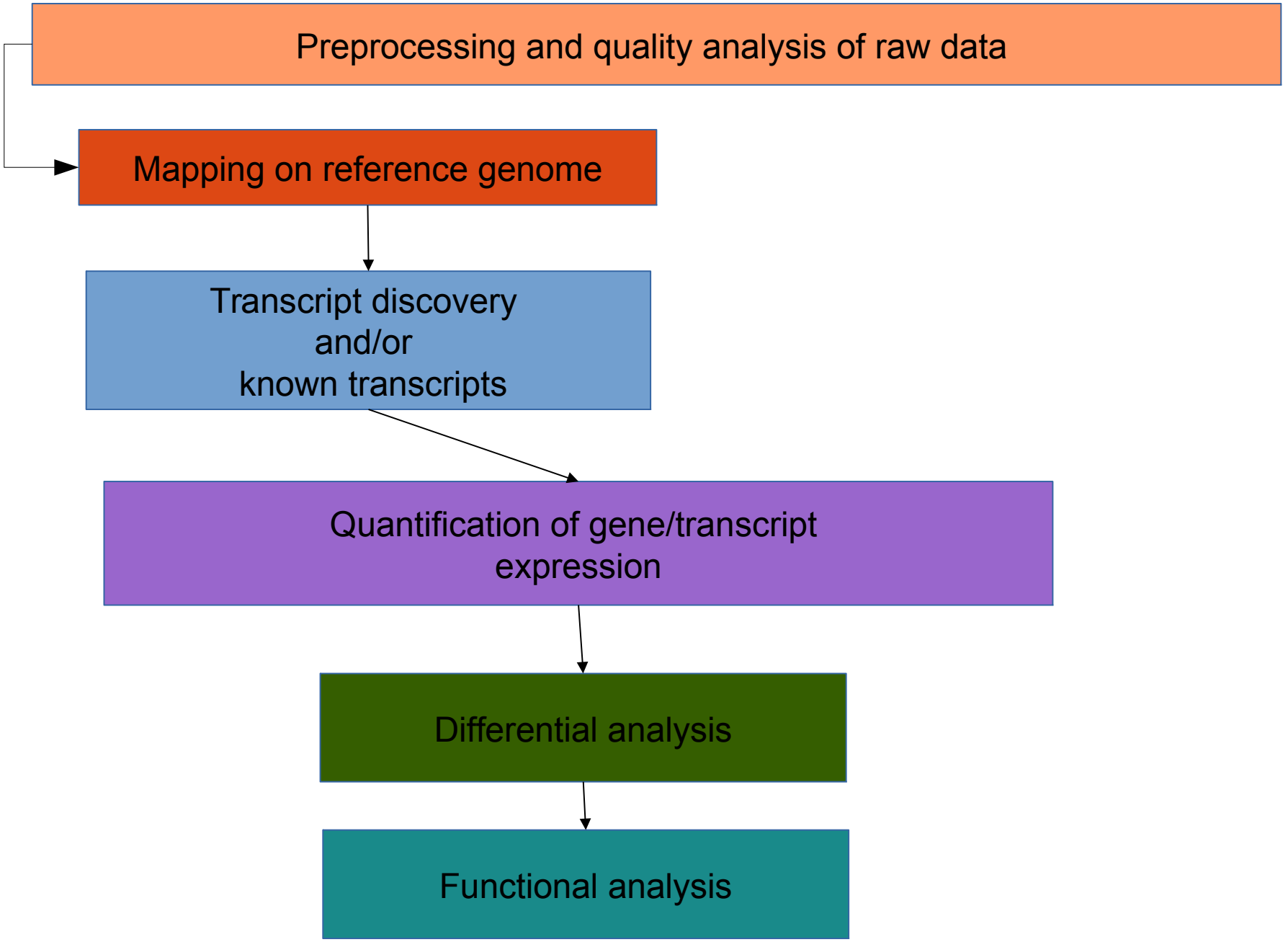
Alignment

Transcriptome reconstruction and quantification

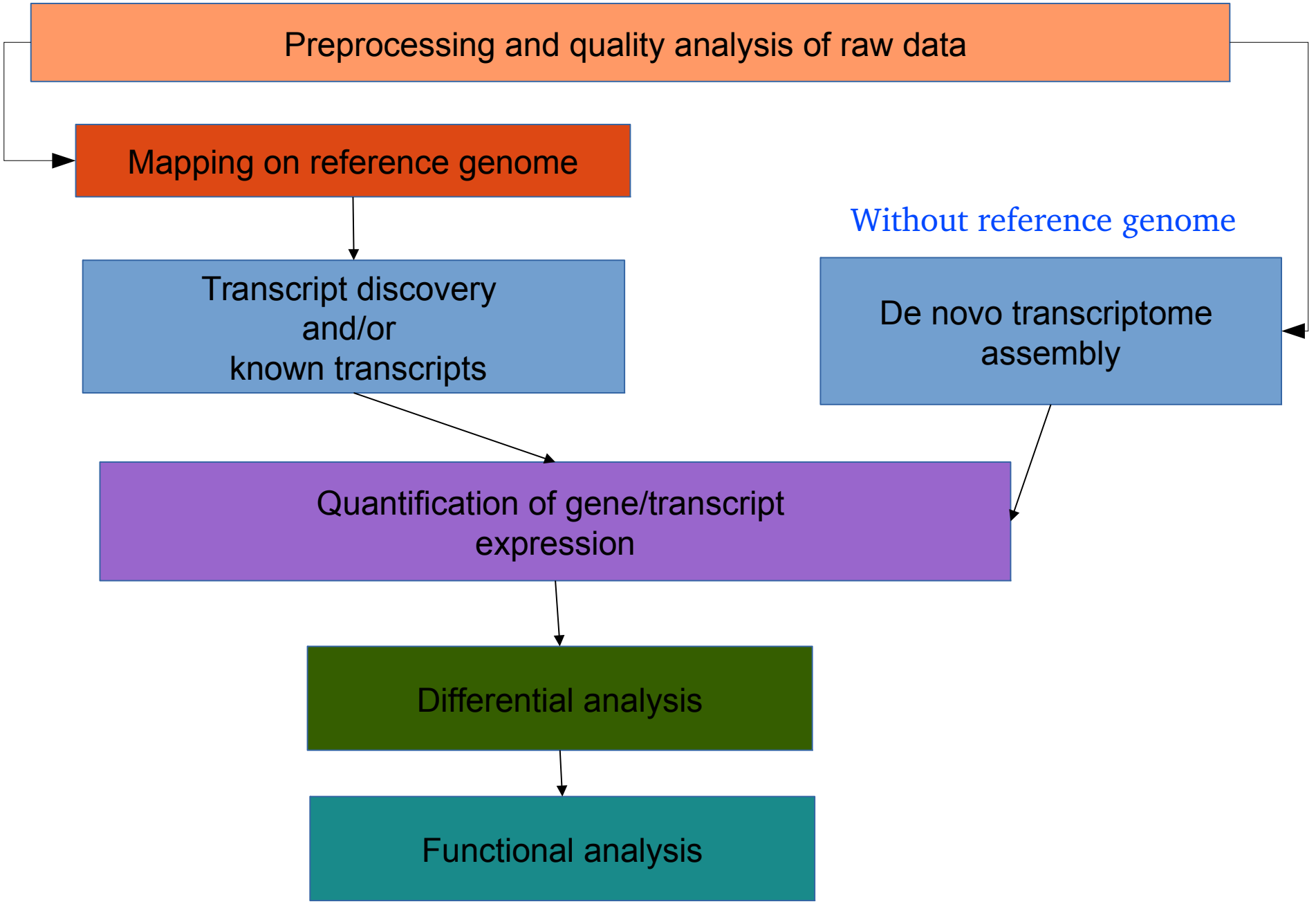
Differential analysis

Mortazavi, A.,
Mapping and
methods, 5(7)

Basic steps to analyse RNA-seq data



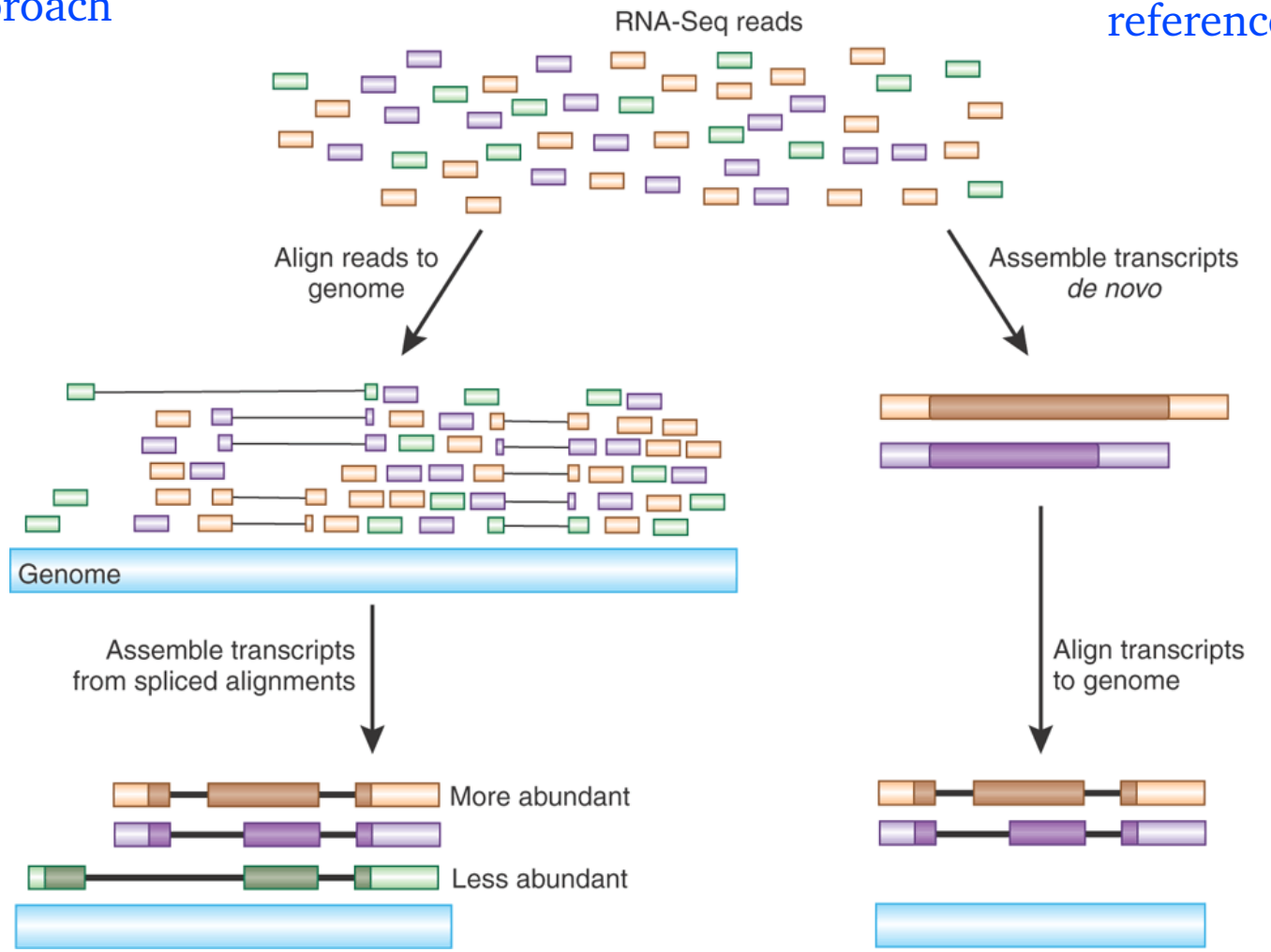
Basic steps to analyse RNA-seq data



Read mapping vs. de novo assembly

Genome guided approach

Without reference genome



Haas and Zody, Nature Biotechnology 28, 421–423 (2010)

Basic steps to analyse RNA-seq data

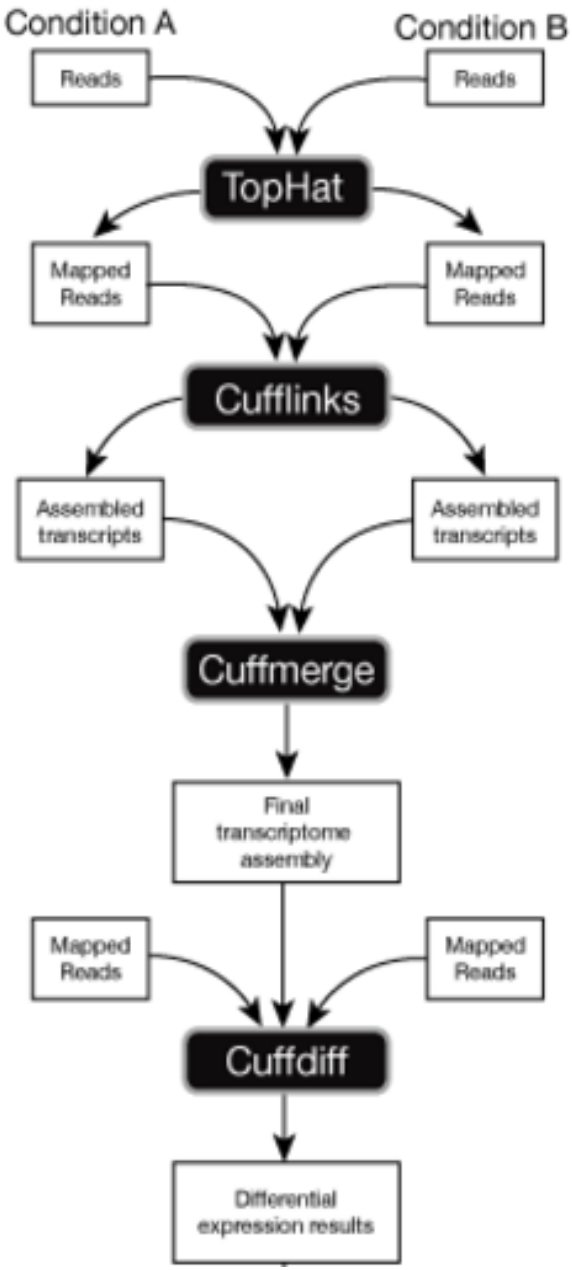
Mapping on reference genome

Transcript discovery
and/or
known transcripts

Quantification of gene/transcript
expression

Differential analysis

Tuxedo Suite



Basic steps to analyse RNA-seq data

Mapping on reference genome

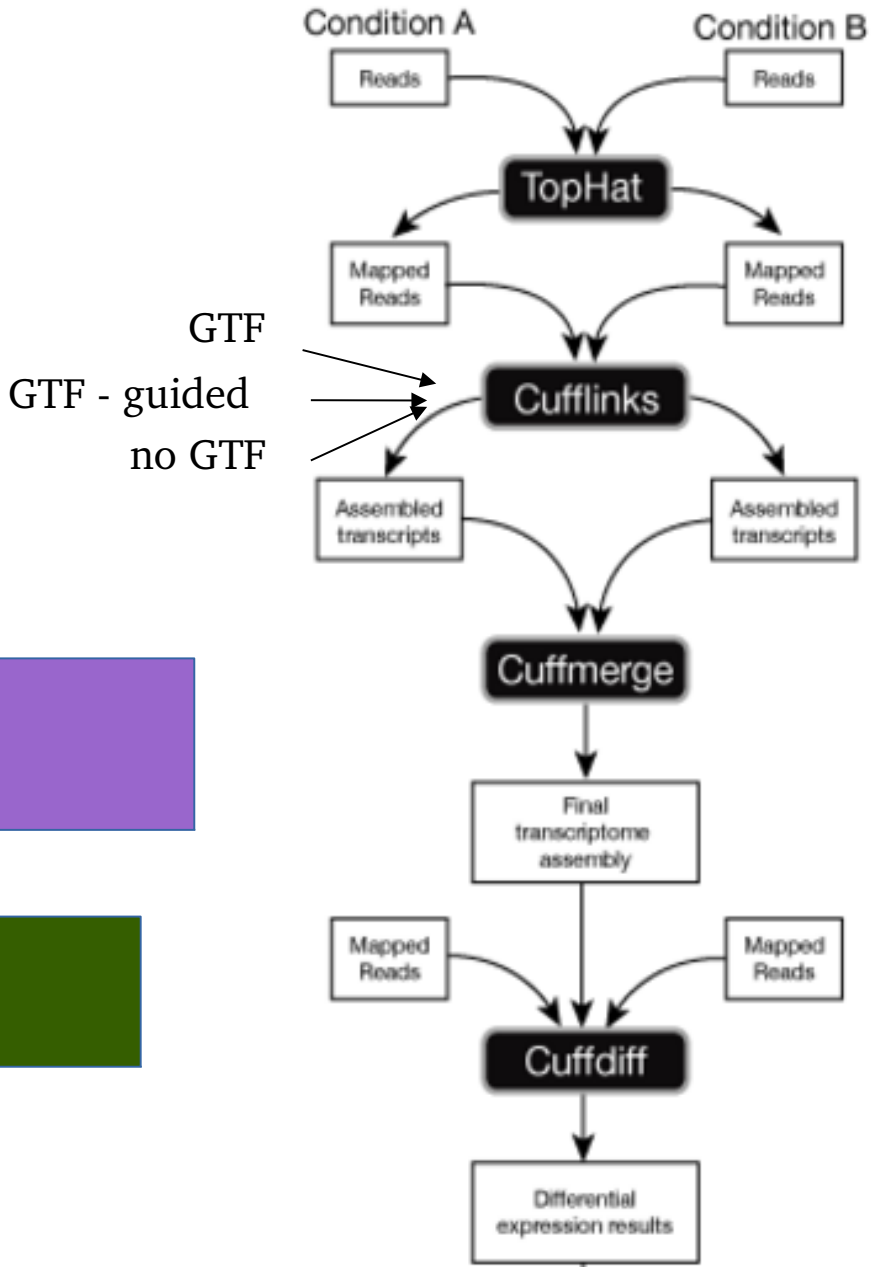
Transcript discovery
and/or
known transcripts

Quantification of gene/transcript
expression

Differential analysis

GTF = known/annotated transcriptome

Tuxedo Suite



Cufflinks

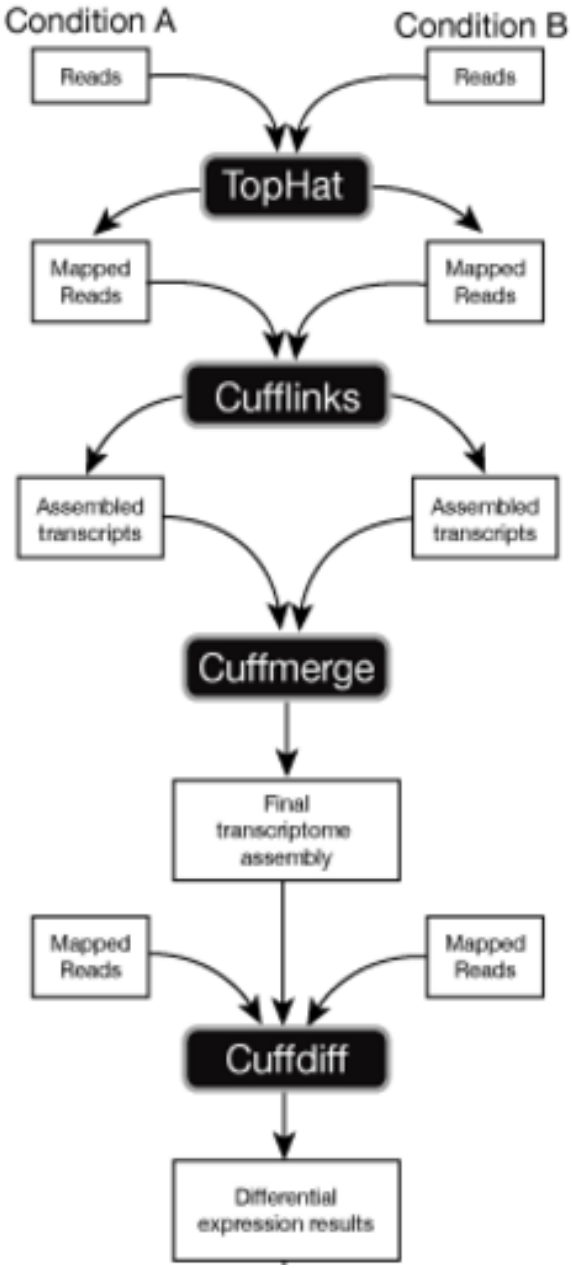
with GTF:
only restricted to known annotation
(no dicovery)

BUT the assumptions are:
- the GTF file contains annotation for ALL transcripts and genes
- all splice sites, start/stop codons, etc. are correct

Are these assumptions correct for every organism?

GTF = known/annotated transcriptome

Tuxedo Suite



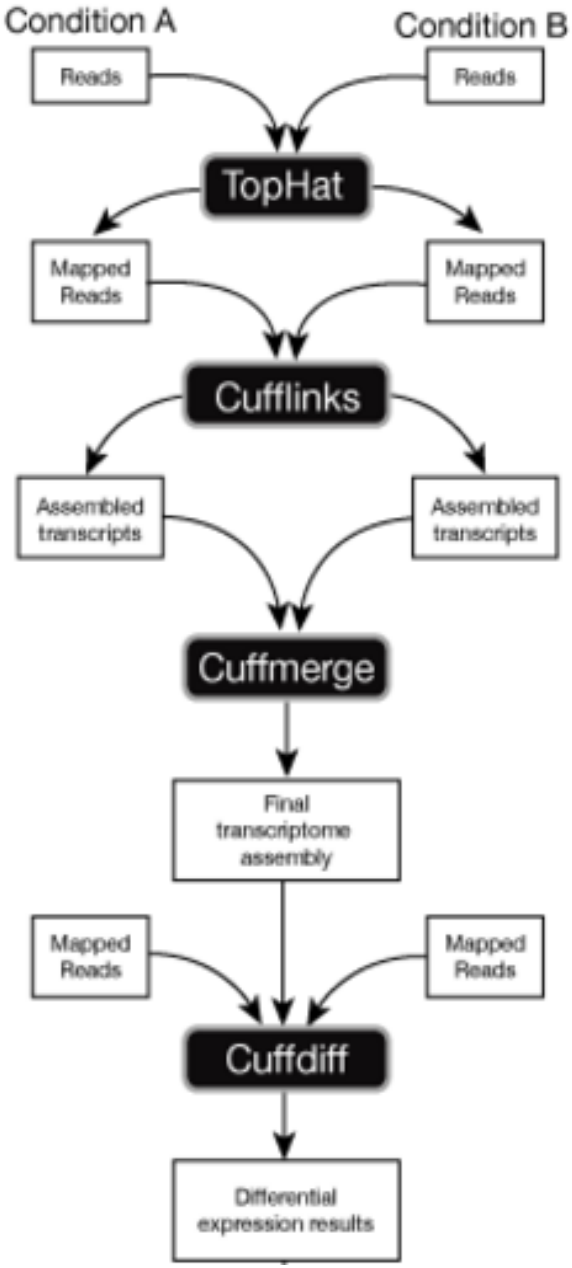
Cufflinks

GTF – guided:
also novel genes and isoforms

no GTF:
only transcript discovery

GTF = known/annotated transcriptome

Tuxedo Suite



Basic steps to analyse RNA-seq data

STAR

Mapping on reference genome

Transcript discovery
and/or
known transcripts

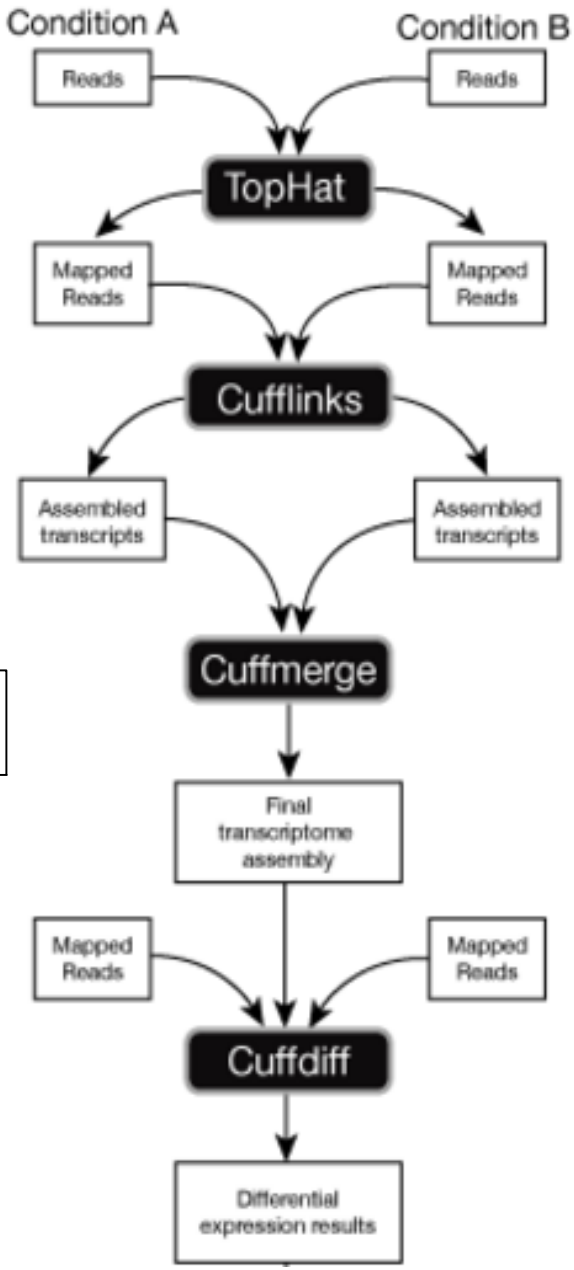
Quantification of gene/transcript expression

Differential analysis

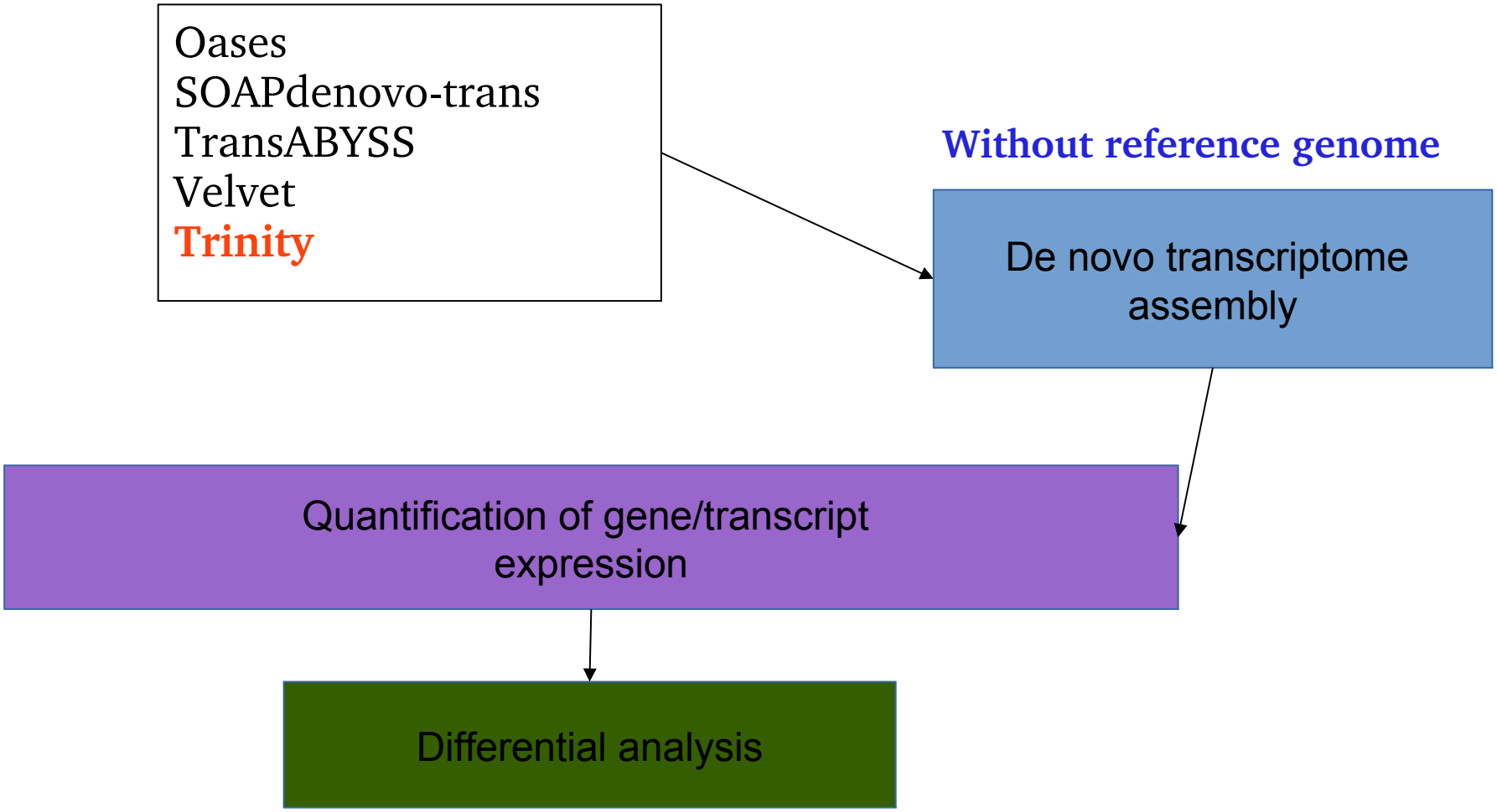
featureCounts

EdgeR,
DESeq,

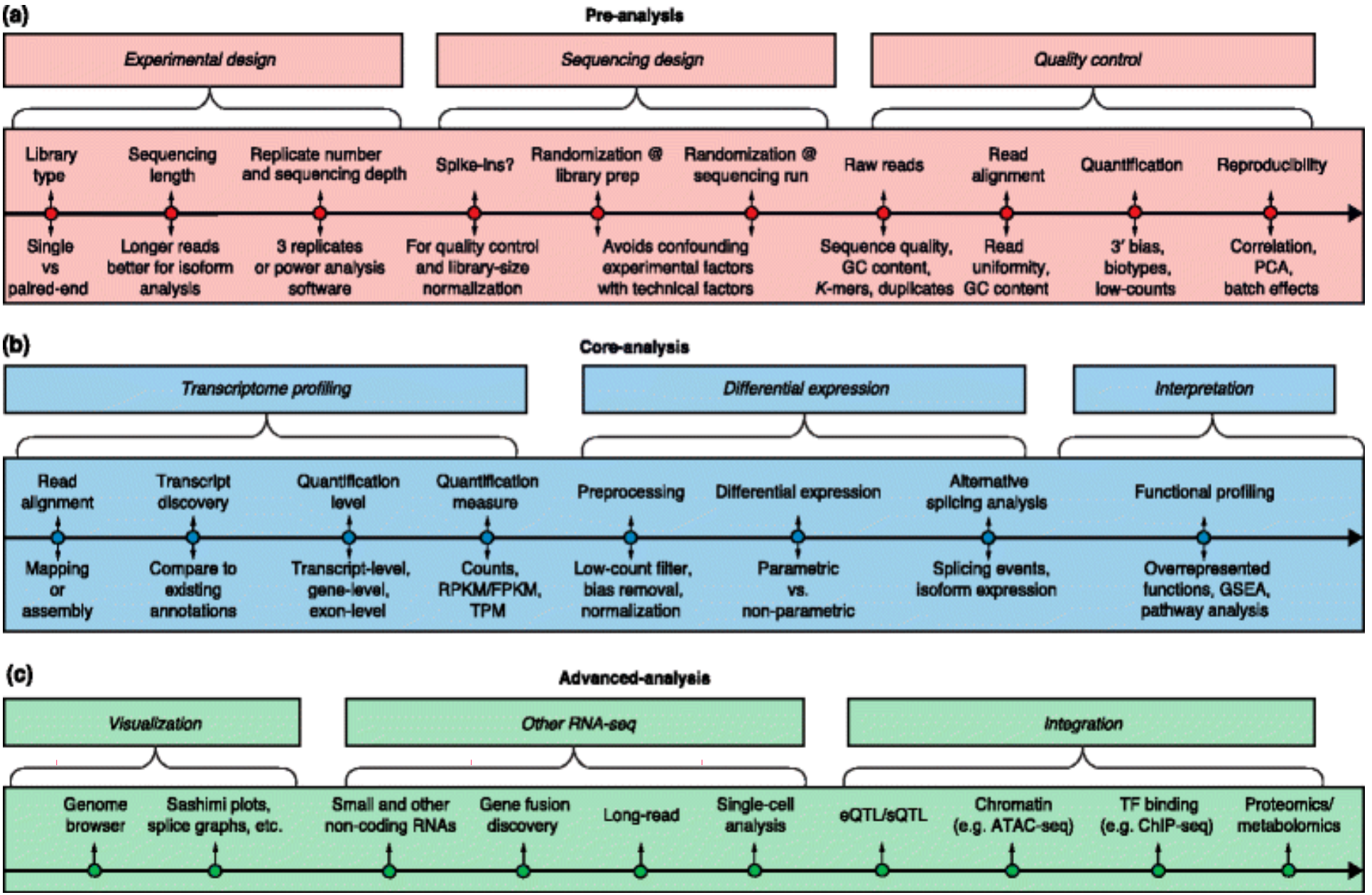
Tuxedo Suite



Basic steps to analyse RNA-seq data



A generic roadmap for RNA-seq data analyses



Conesa et al. - Genome Biology 2016

A well-designed experiment

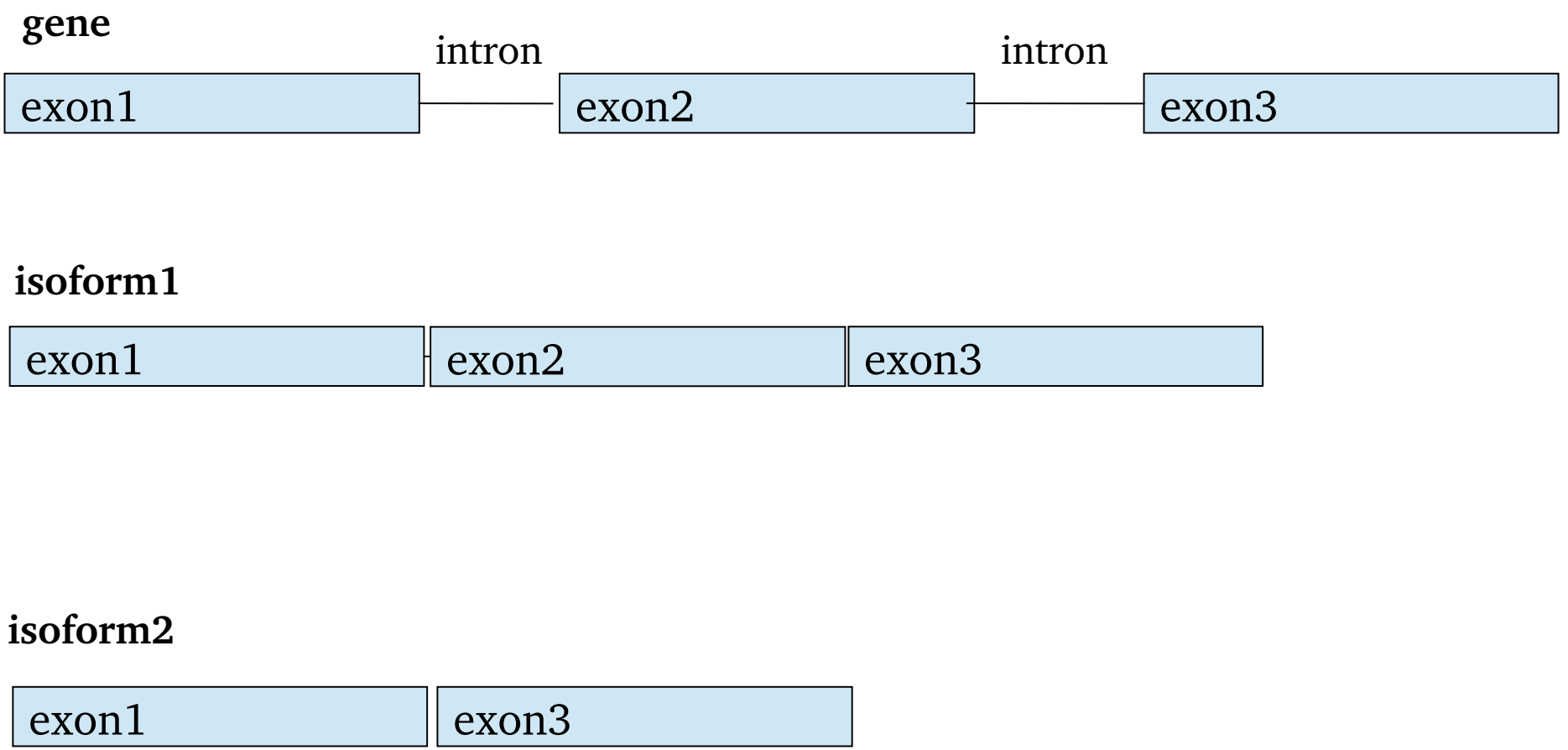
- clear objectives/questions
- focused (library preparation)
- sufficient statistical power (replicates/sample size)
- unbiased (randomization/blocking)

Experimental design

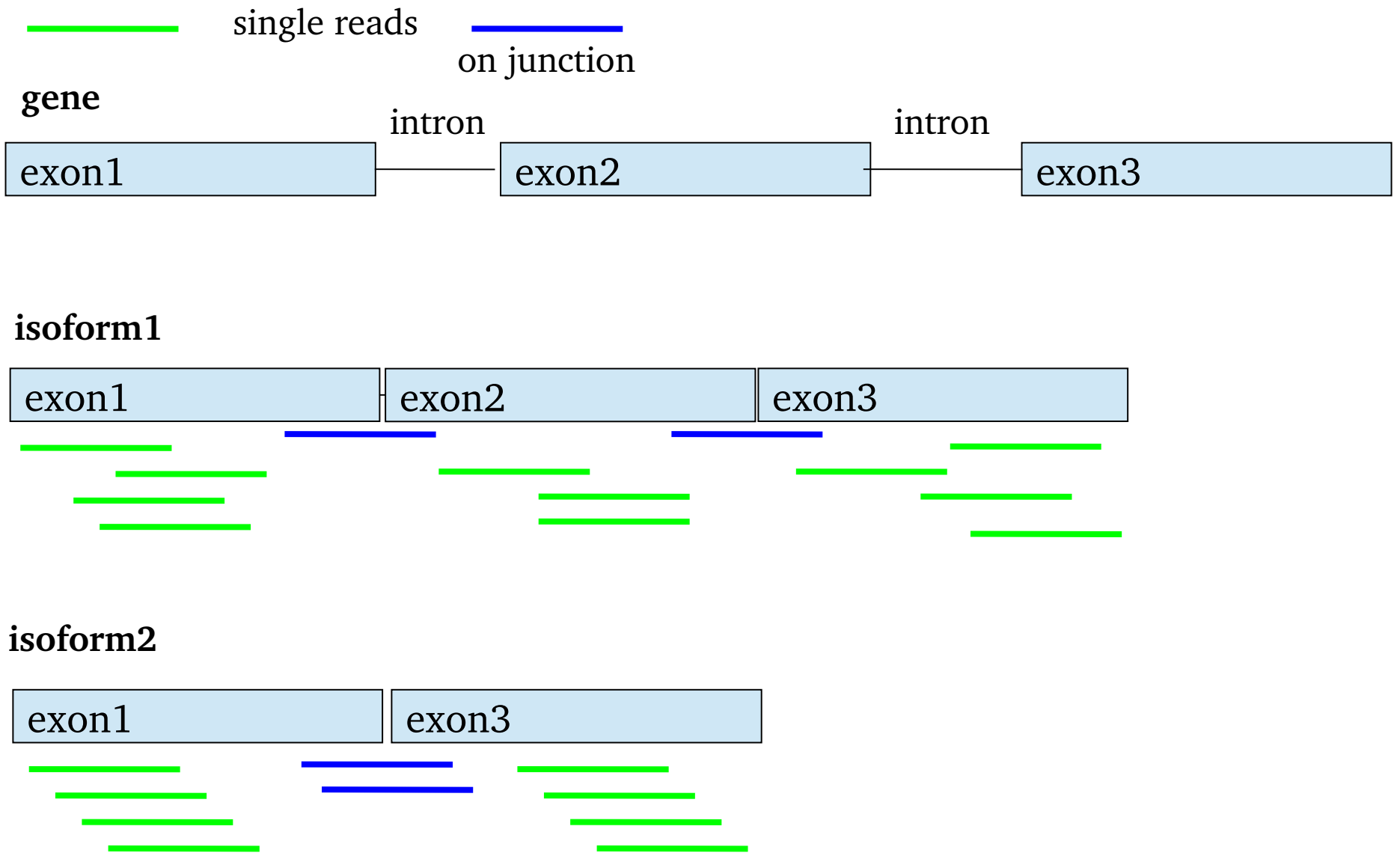
focused (library preparation)

- single or paired-end?

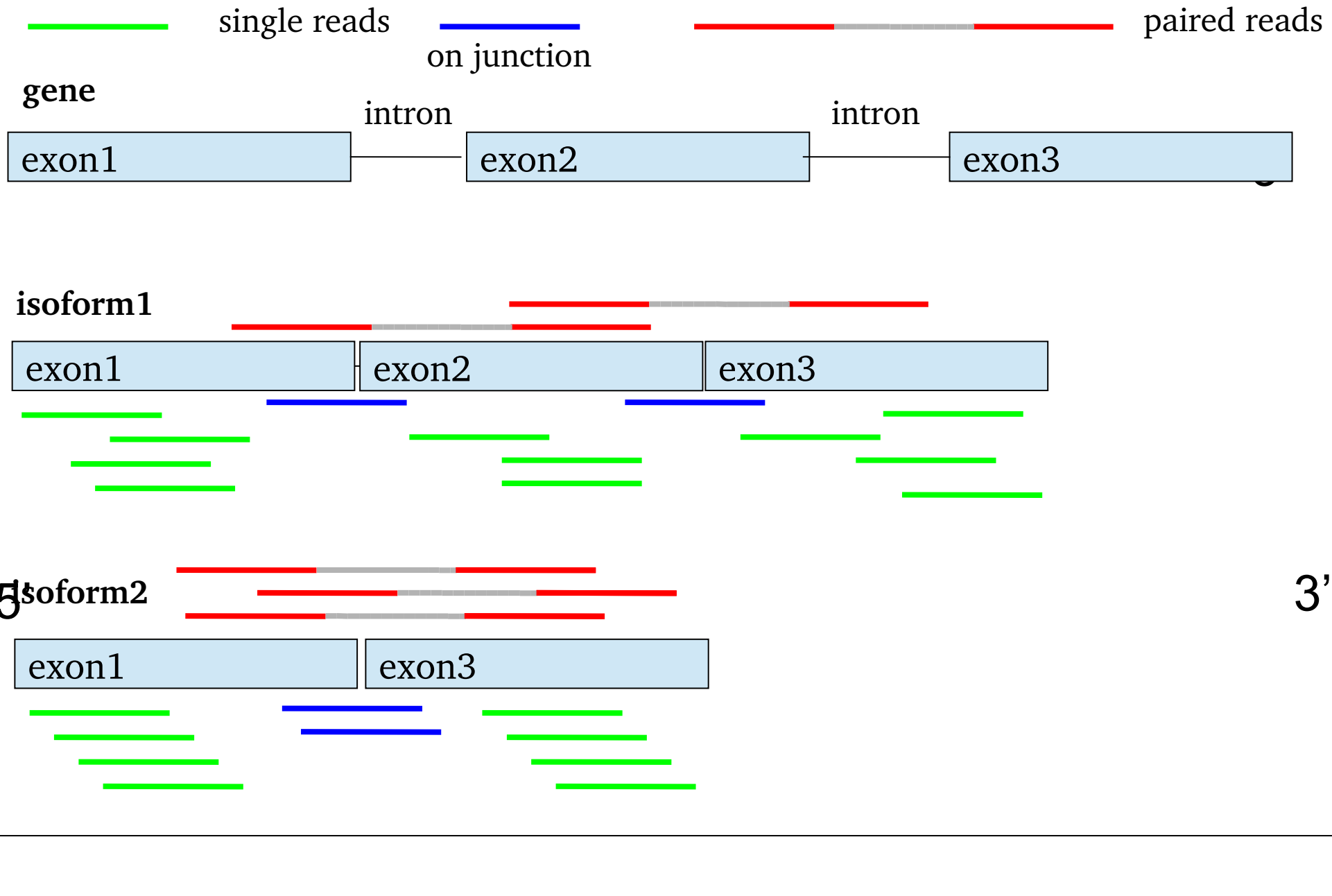
mRNA-Seq can identify isoforms (single- or paired-end reads)



mRNA-Seq can identify isoforms (single- or paired-end reads)



mRNA-Seq can identify isoforms (single- or paired-end reads)



Experimental design
library type: single-end (SE) or paired-end (PE)?

short SE reads (very cheap):

gene expression levels in well-annotated organisms

longer reads:

improve mappability and transcript identification

PE reads:

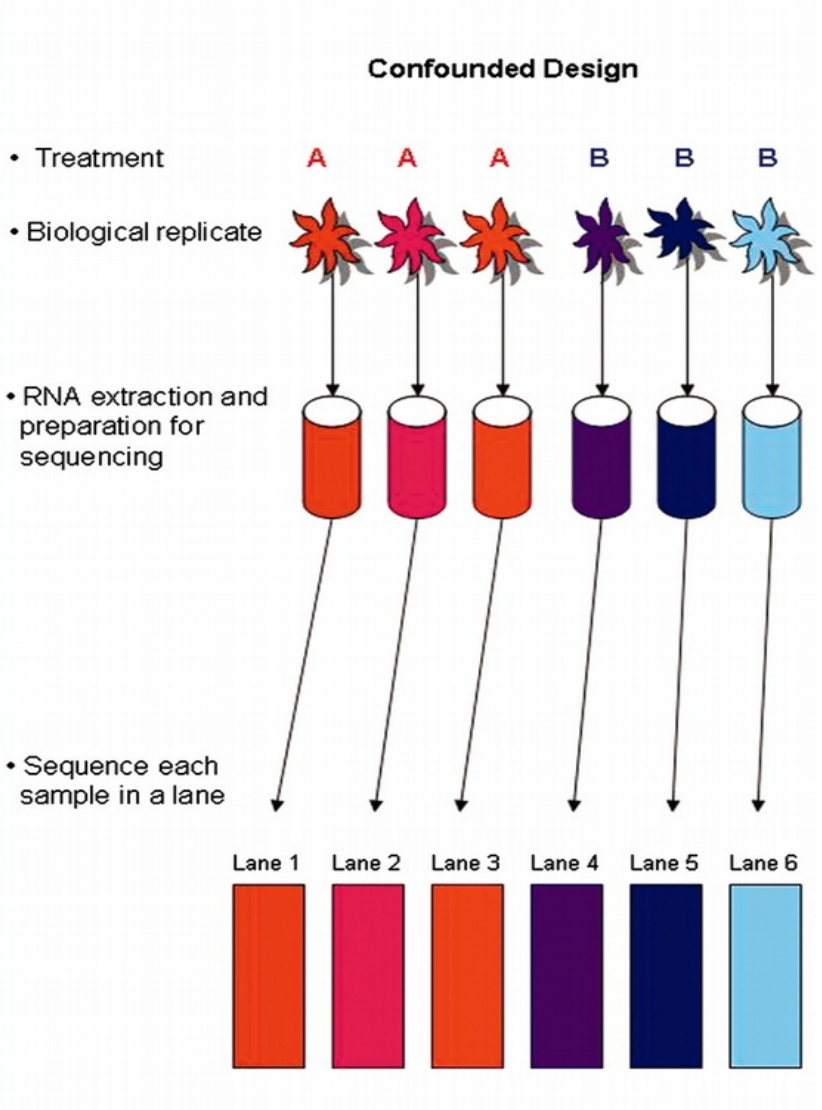
improve isoform expression analysis and transcript discovery

longer PE reads (very expensive):

for poorly annotated transcriptomes

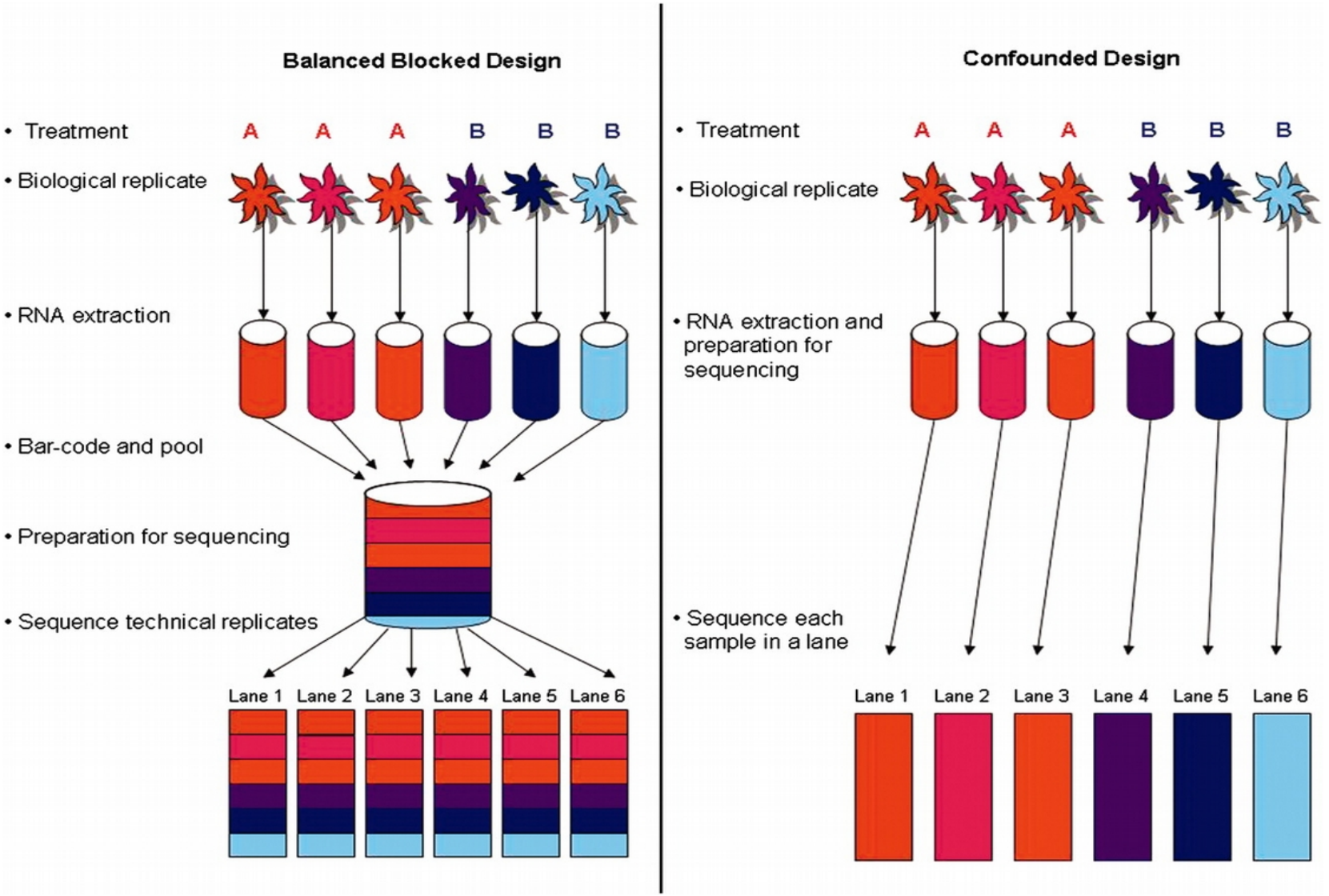
Experimental design – Bias issue

Comparison of two designs for testing differential expression between treatments A and B. Treatment A is denoted by red tones and treatment B by blue tones.



Experimental design – Bias issue

Comparison of two designs for testing differential expression between treatments A and B. Treatment A is denoted by red tones and treatment B by blue tones.



Auer P L , and Doerge R W Genetics 2010;185:405-416

Raw data Quality control

Raw data Quality control - Tools

FastQC is the most popular tool to perform these analyses
(on Illumina reads at least)

NGSQC can be applied to any platform

FastQC

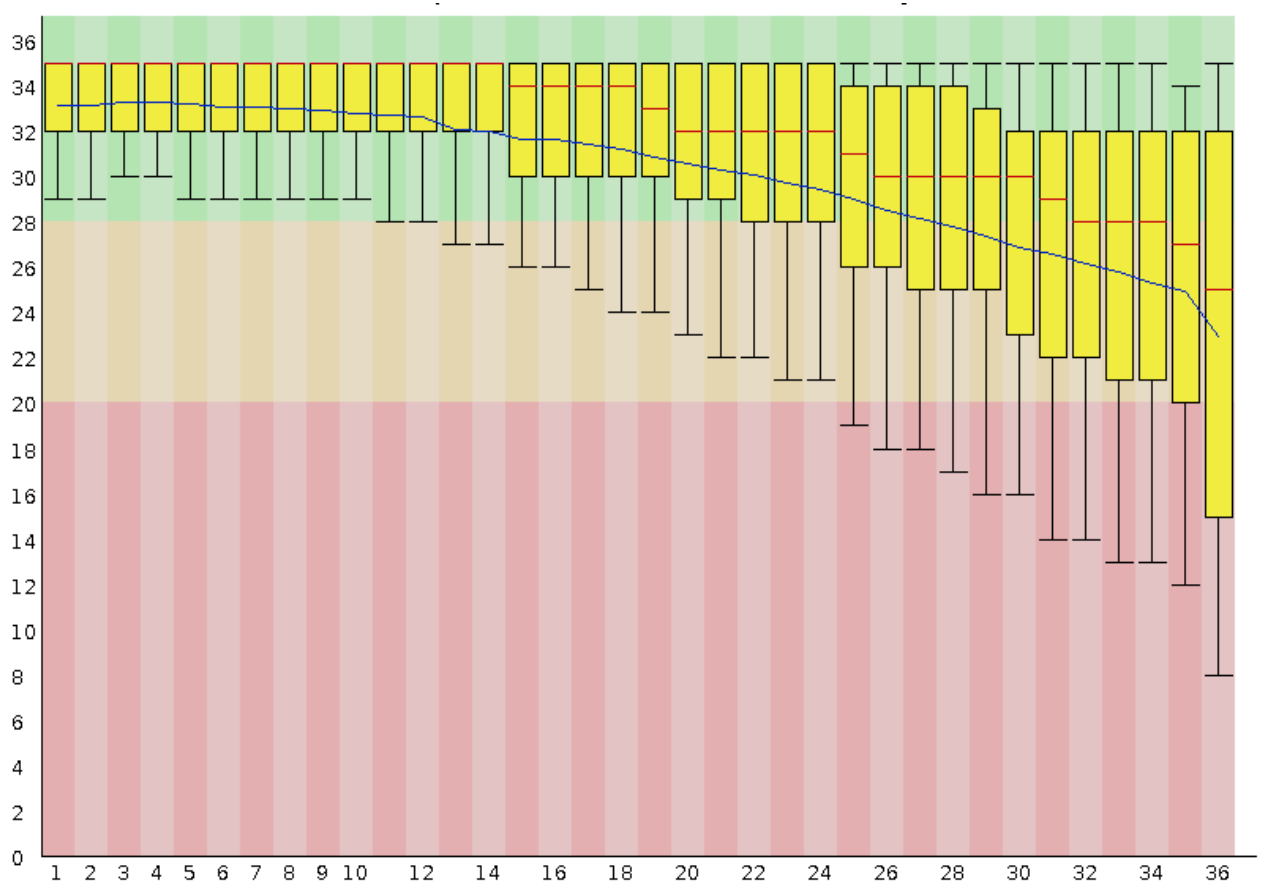
WARNING: the analysis results appear to give a **pass**/**fail** result

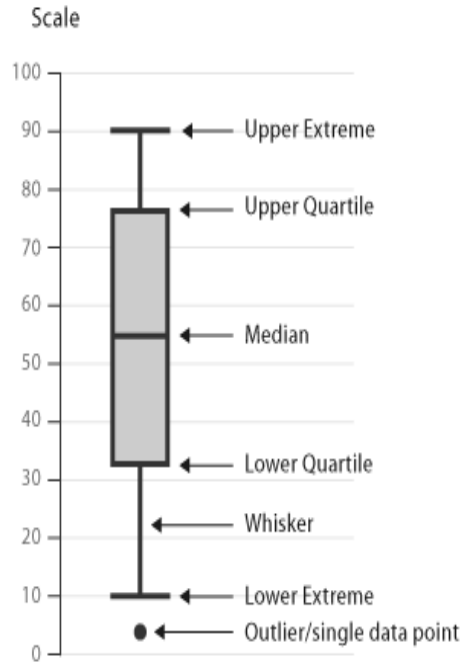
but these evaluations must be taken in the context of what you expect from your library

some experiments may be expected to produce libraries which are biased in particular ways

FastQC

Per Base Sequence Quality



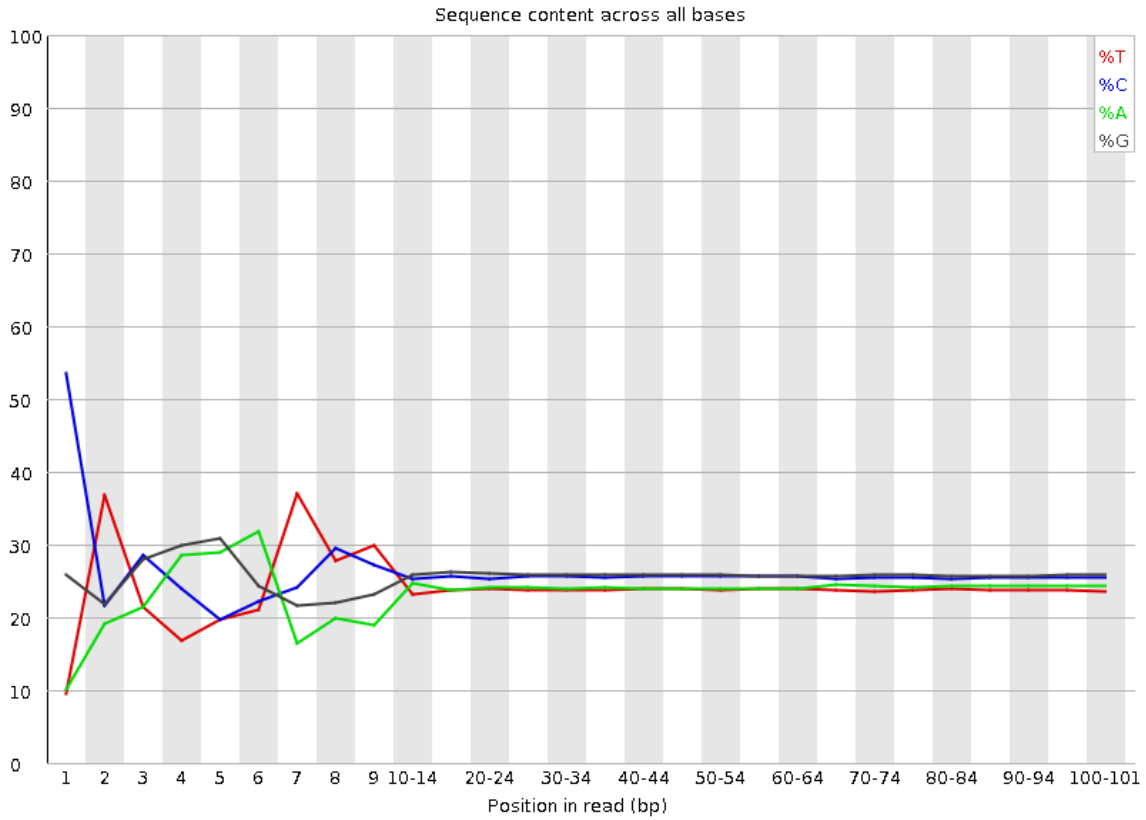


The blue line is the mean quality

As a general rule, read quality decreases towards the 3' end of reads, because chemistry degrades with increasing read length and if it becomes too low, bases should be removed to improve mappability.

FastQC

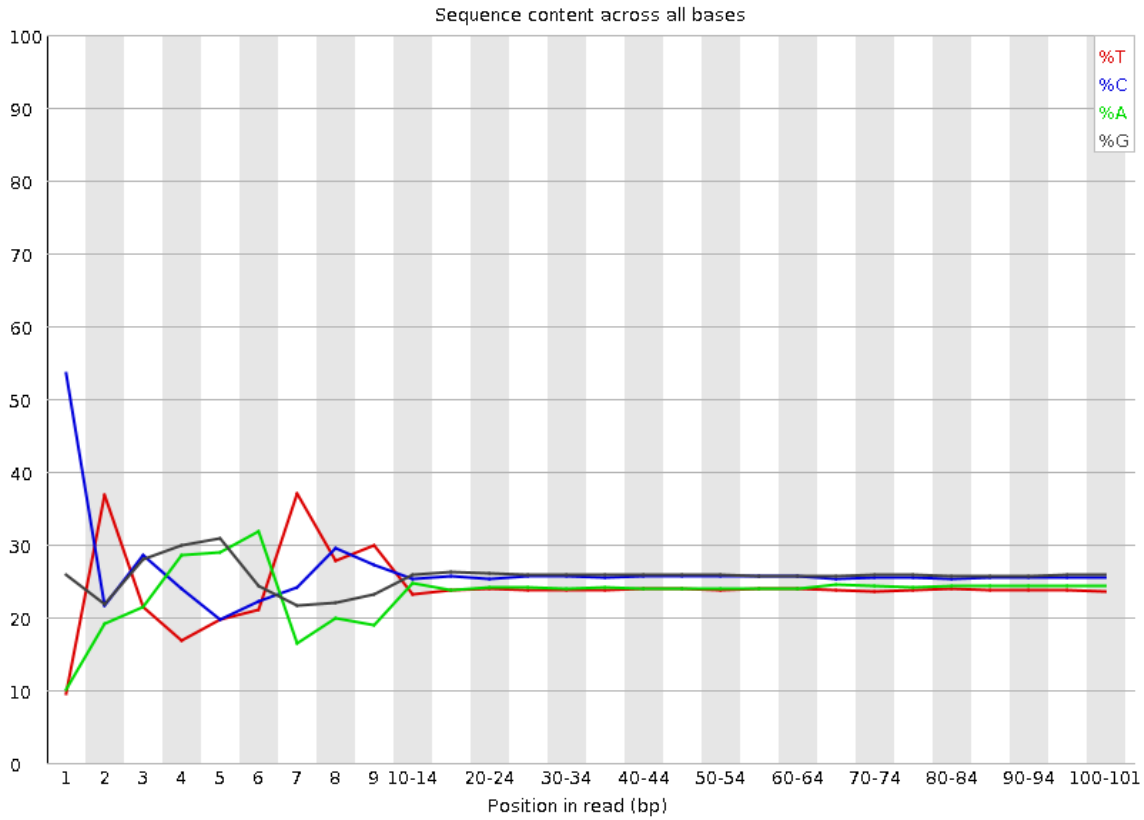
Per Base Sequence Content



the proportion of each base in each position

FastQC

Per Base Sequence Content

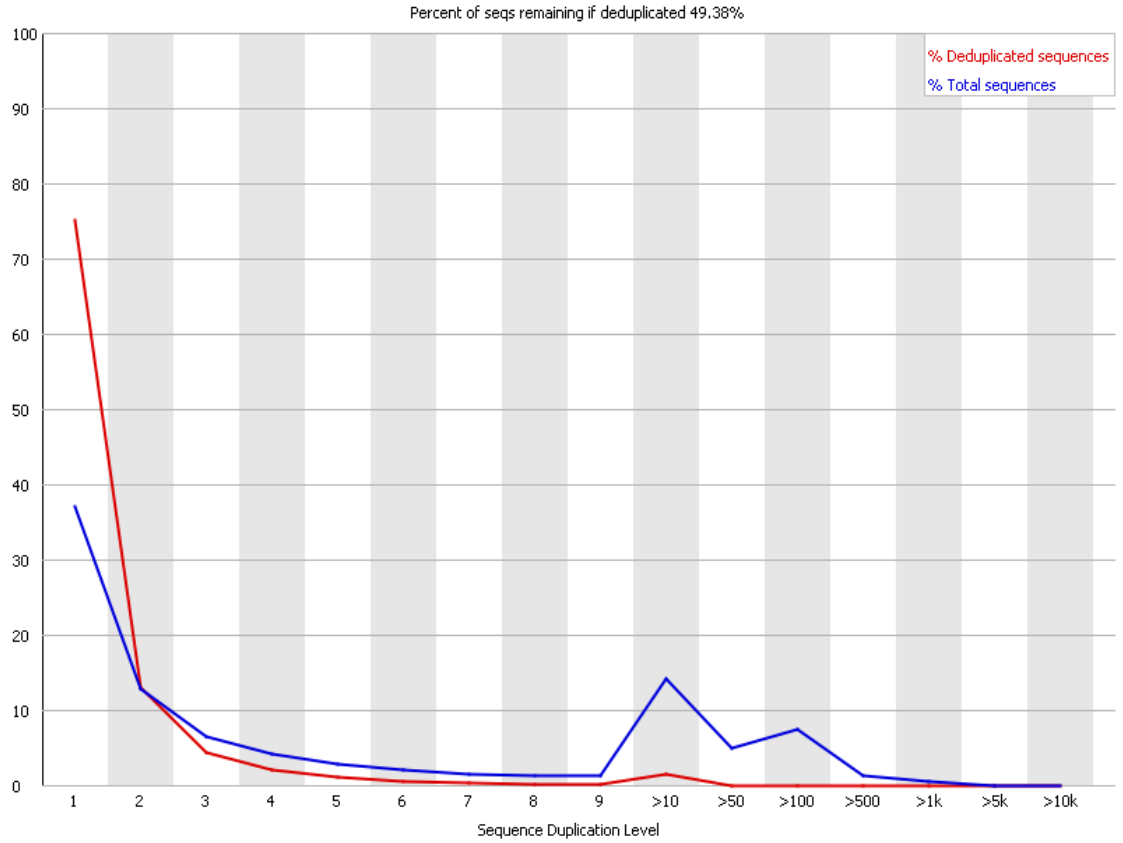


the proportion of each base in each position

In a random library you would expect that there would be little to no difference between the different bases of a sequence run, so the lines in this plot should run parallel with each other. Libraries produced by priming using random hexamers (including nearly all RNA-Seq libraries) produce biased sequence composition at the start of the read

FastQC

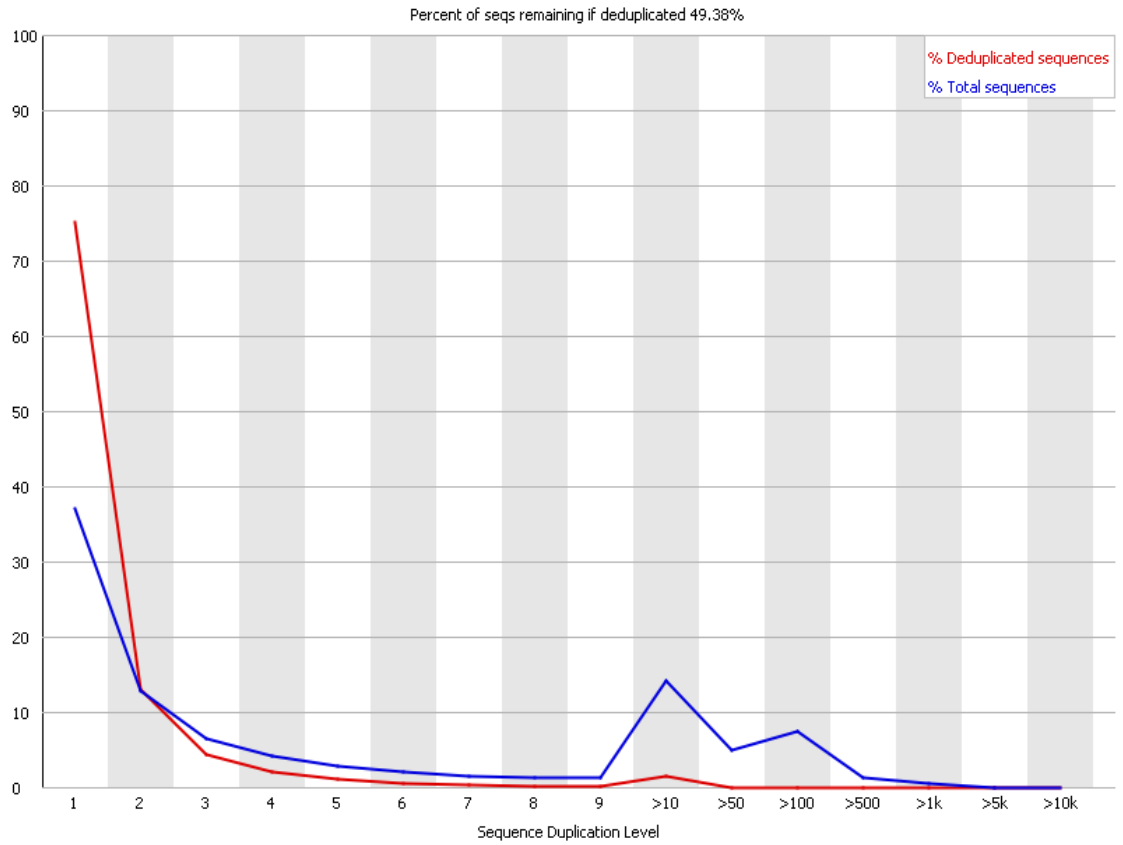
Duplicate sequences



the percentage of reads with different degrees of duplication

FastQC

Duplicate sequences

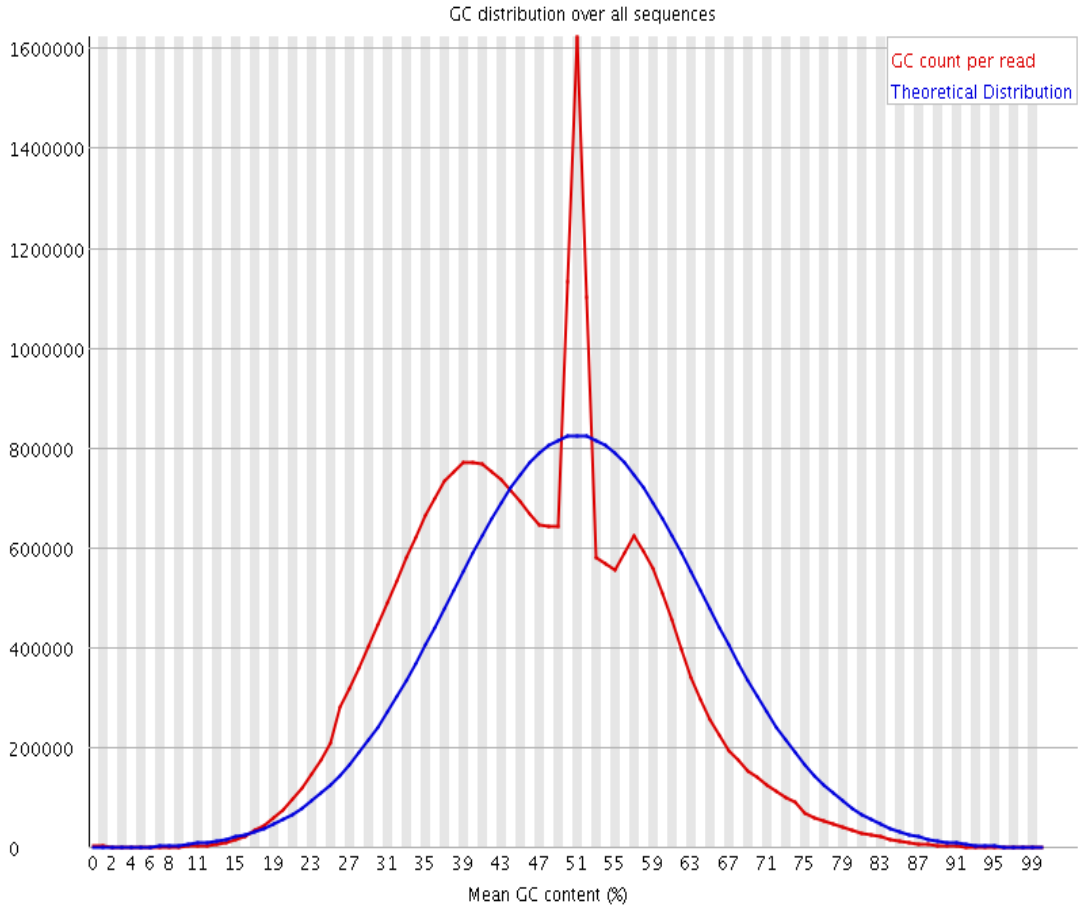


the percentage of reads with different degrees of duplication

More specific enrichments of subsets (PCR over amplification) or the presence of low complexity contaminants will tend to produce spikes towards the right of the plot.
In RNA-seq it's also generated by high level expressed genes.
The duplication level can be better measured and removed after alignment.

FastQC

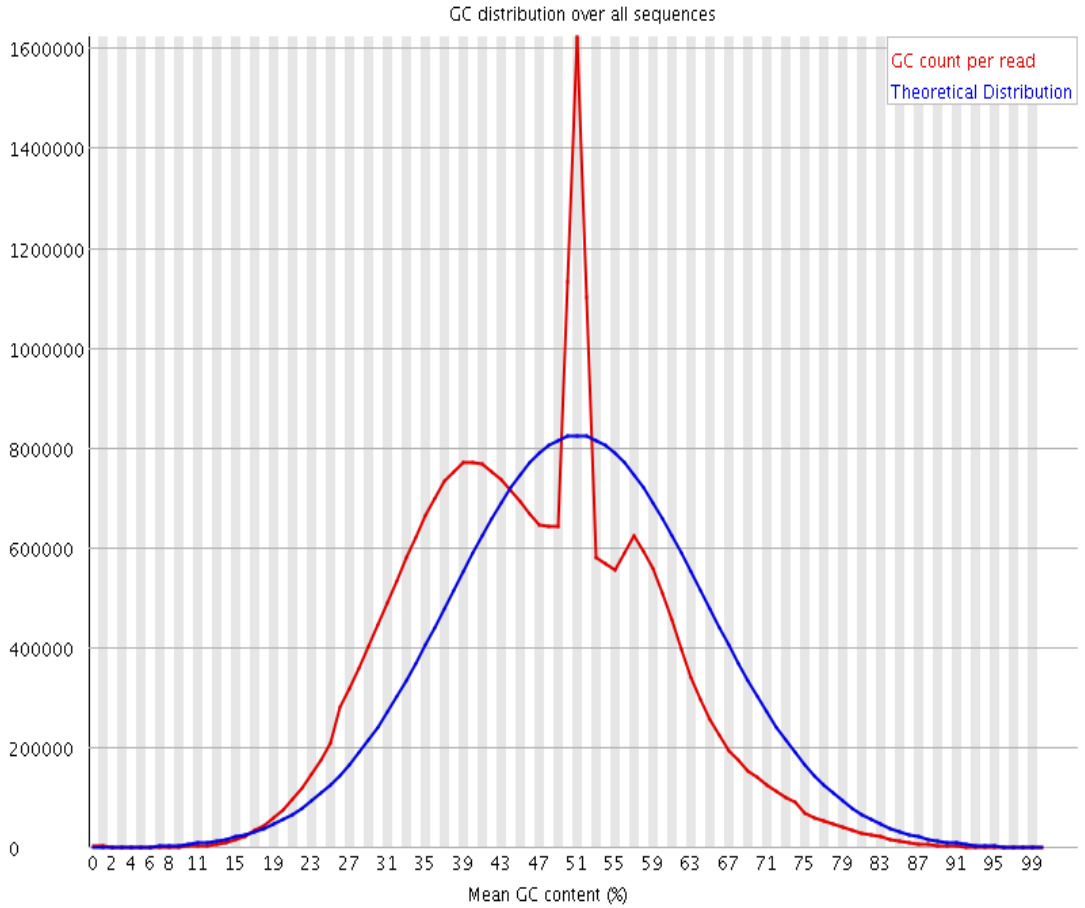
GC content



GC content distribution

FastQC

GC content



GC content distribution

An unusually shaped distribution could indicate a contaminated library or some other kinds of biased subset.

FastQC

Overrepresented sequences

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGC	47120	0.14460904567021887	Illumina3truseq1 (100% over 50bp)

adapters, contaminants (rRNA, vruses, …) typically

Quality filtering and trimming - Tools

Quality filtering and trimming - Tools

Software tools that can be used to discard low-quality reads, trim adaptor sequences, and eliminate poor-quality bases are:

FASTX-Toolkit: a collection of command line tools for Short-Reads FASTA/FASTQ files preprocessing

Trimmomatic: a flexible trimmer for Illumina sequence data

TrimGalore!: is a wrapper script to automate quality and adapter trimming as well as quality control

TrimGalore

Trim Galore!

Function	A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, with some extra functionality for MspI-digested RRBS-type (Reduced Representation Bisufite-Seq) libraries.
Language	Perl
Requirements	A functional version of Cutadapt and optionally FastQC are required.
Code Maturity	Stable.
Code Released	Yes, under GNU GPL v3 or later.
Initial Contact	Felix Krueger
	Download Now

https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

TrimGalore

The trim_galore perl wrapper itself **consumes just a few megabytes of RAM**

- in the first step, **low-quality base calls are trimmed off from the 3' end** of the reads before adapter removal

- in the next step, **Cutadapt** finds and **removes adapter sequences from the 3' end** of reads

if no sequence was supplied, it will attempt to auto-detect the adapter analysing the first 1 million sequences and attempt to find the first 12 or 13bp of the following standard adapters:

Illumina:	AGATCGGAAGAGC
Small RNA:	TGGAATTCTCGG
Nextera:	CTGTCTCTTATA

If no adapter can be detected within the first 1 million sequences, its default is --illumina

- in the last step, **FastQC** is automatically launched on the trimmed sequences

TrimGalore

trim_galore --quality 28 --phred33 --dont_gzip --stringency 4 --length 15 --fastqc --output_dir dir_name --retain_unpaired --trim-n --paired input_R1.fastq input_R2.fastq

--quality: the algorithm is the same as the one used by BWA (subtract (28) from all qualities; compute partial sums from all indices to the end of the sequence; cut sequence at the index at which the sum is minimal)

--stringency: the minimum number (4) of required overlap with the adapter sequence

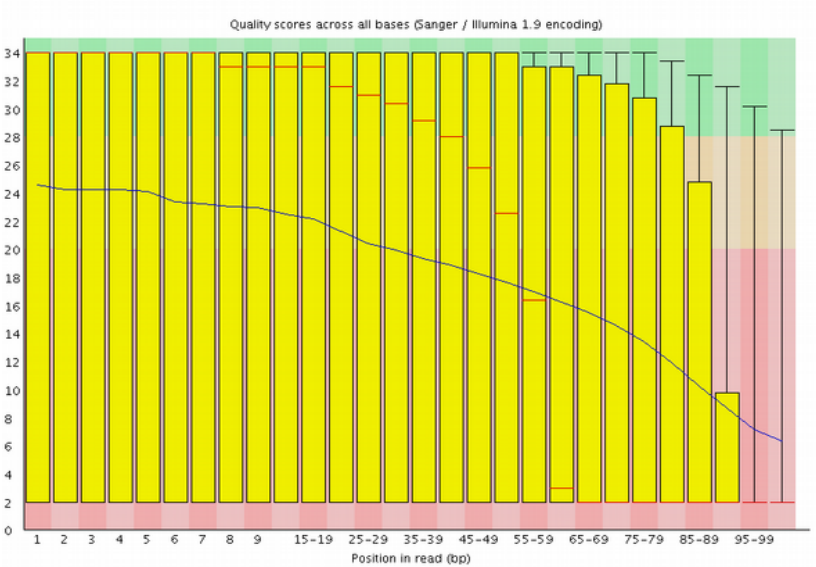
--length: discard reads that are shorter than (15) bp after trimming

--trim-n: removes Ns from either side of the read

--retain_unpaired: if only one of the two sequences became shorter than the threshold, the mate is retained and may be aligned in a single-end manner

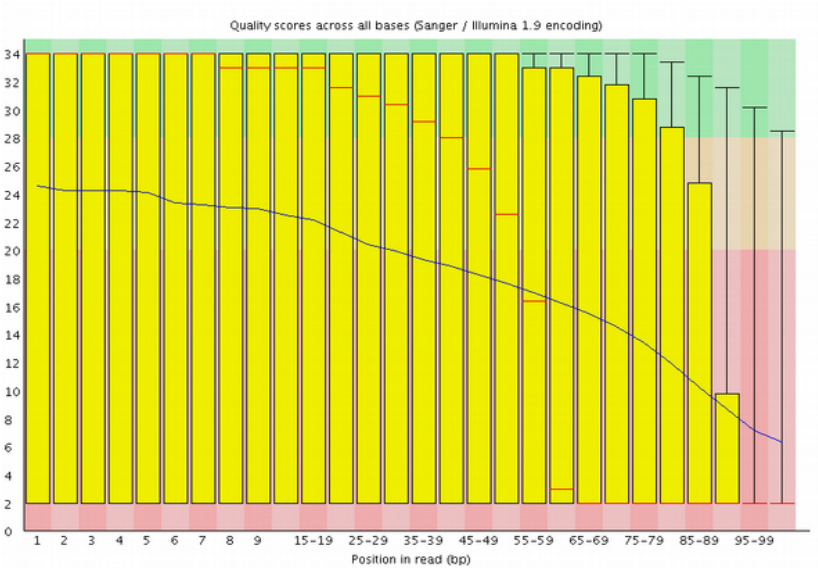
--fastqc: run FastQC once trimming is complete

TrimGalore



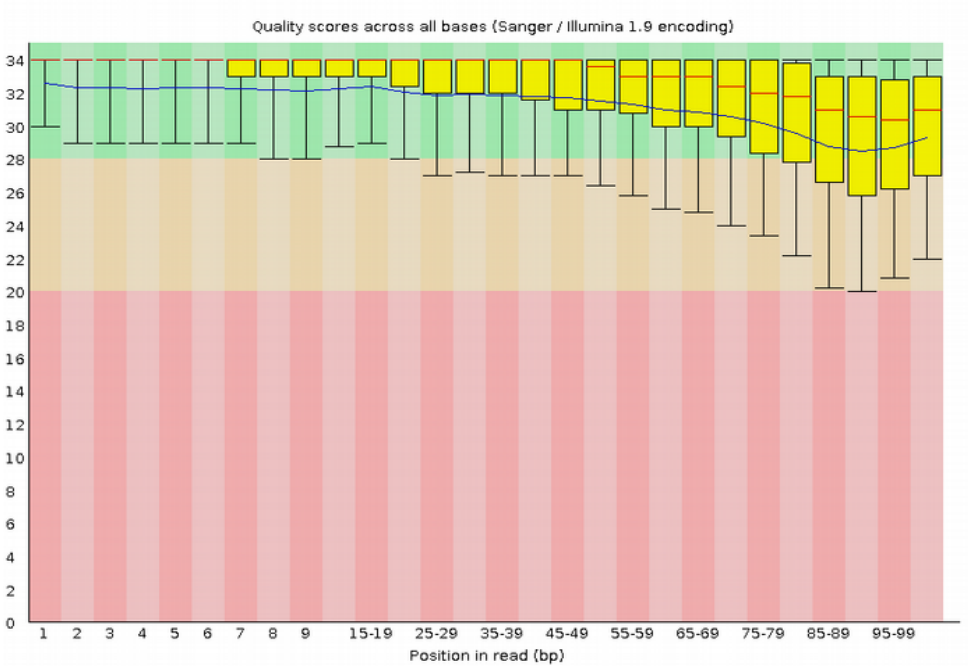
Before quality trimming

TrimGalore

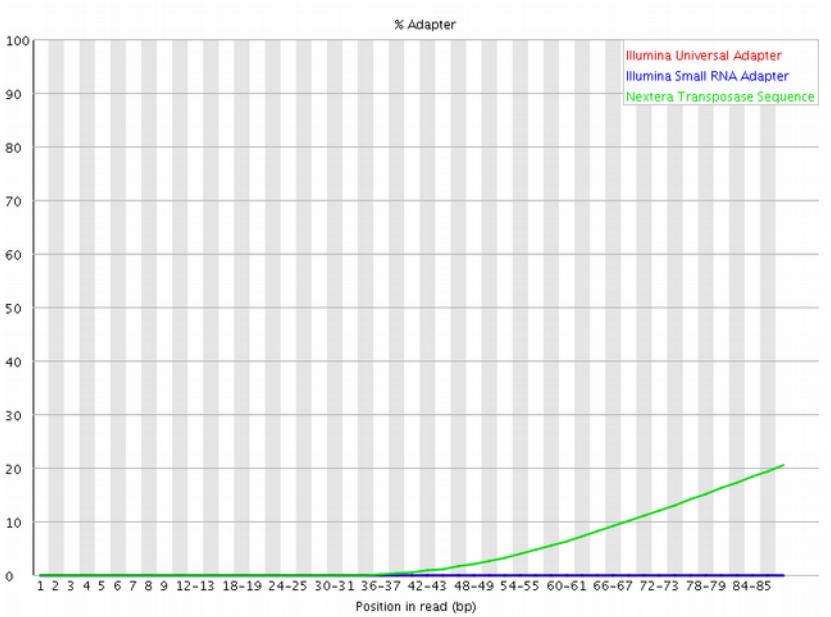


Before quality trimming

After quality trimming

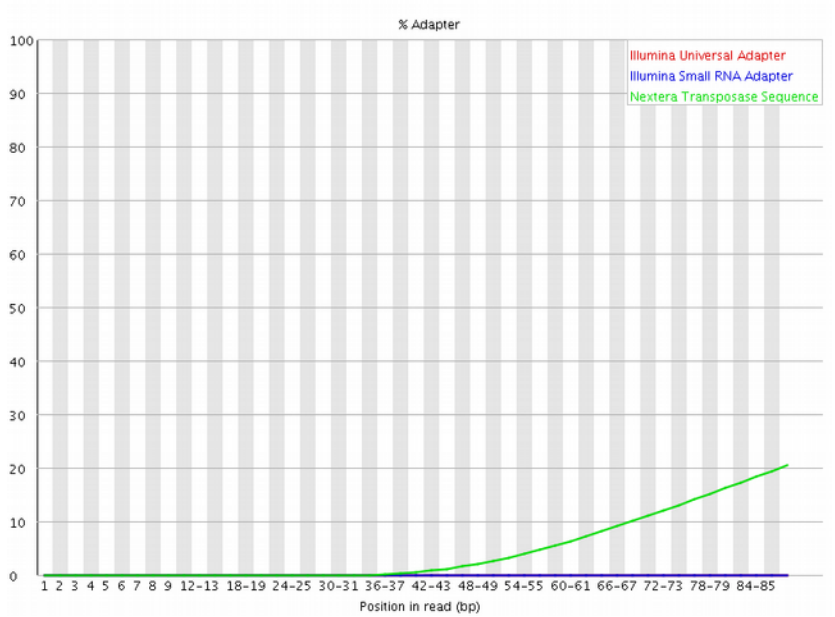


TrimGalore



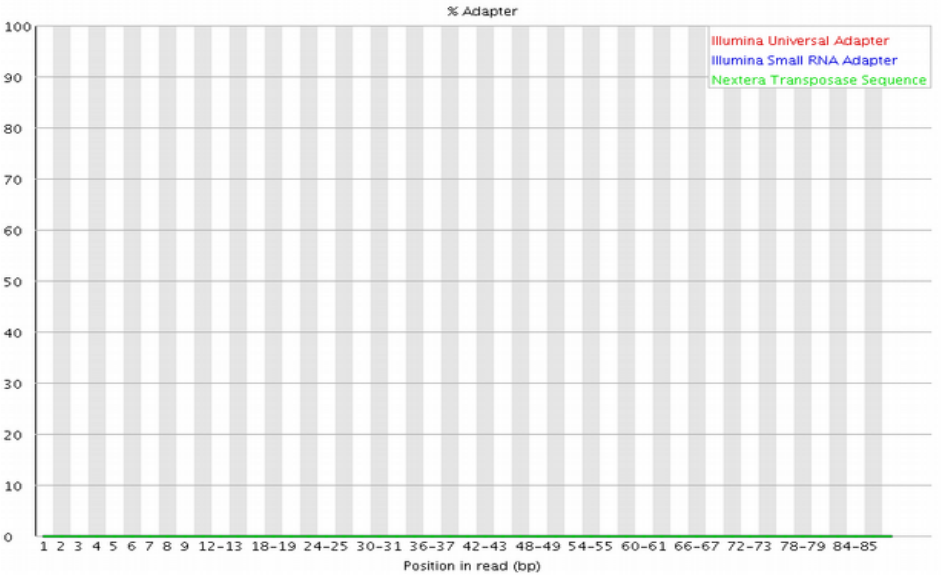
Before adapter trimming

TrimGalore

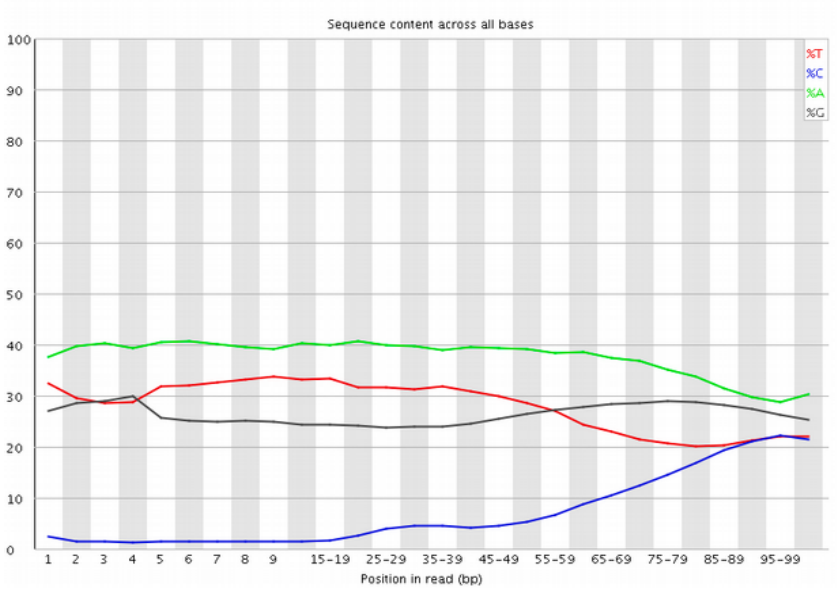


Before adapter trimming

After adaper trimming

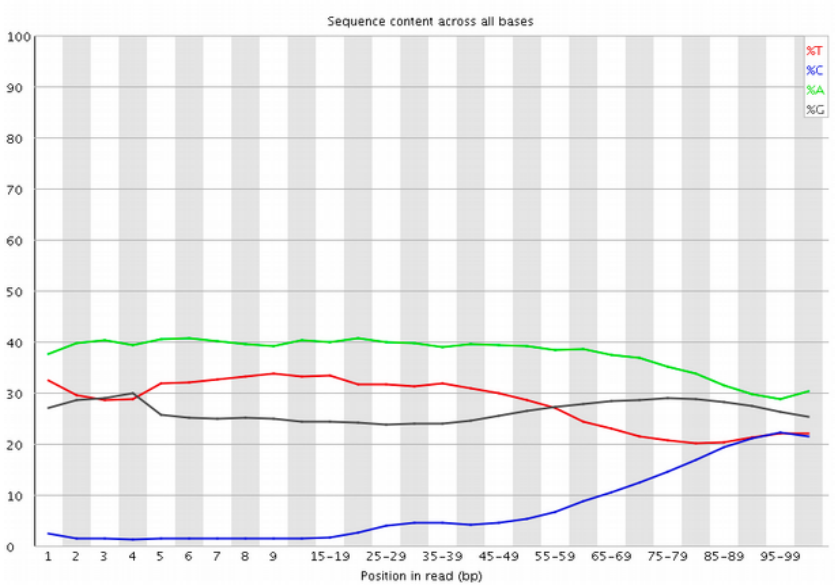


TrimGalore



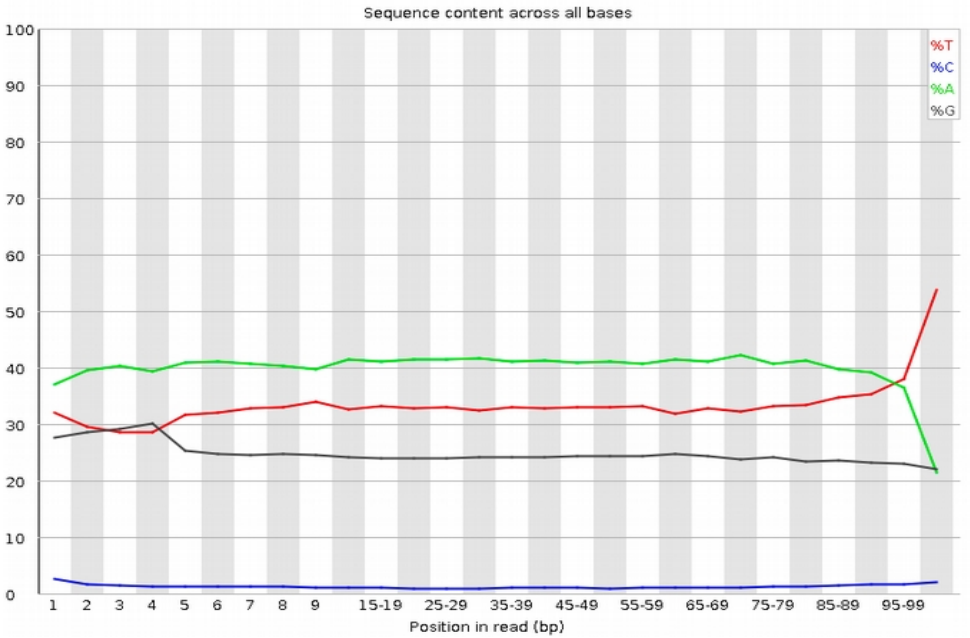
Before adapter trimming

TrimGalore



Before adapter trimming

After adaper trimming



Cuffmerge – Final Transcriptome