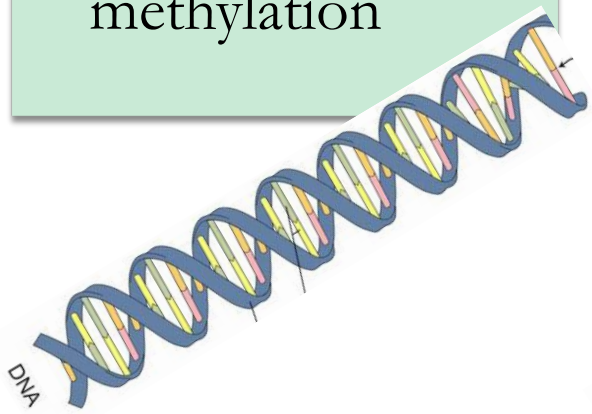
RNA-SEQ DATA ANALYSIS: TRANSCRIPTOME ASSEMBLY AND DIFFERENTIAL EXPRESSION ANALYSIS

Luigi Grassi (lg490@medschl.cam.ac.uk)
Guillermo Parada (guillermo.parada@sanger.ac.uk)

HTS Applications - Overview

DNA Sequencing

- Genome Assembly
- SNPs
- DNA methylation

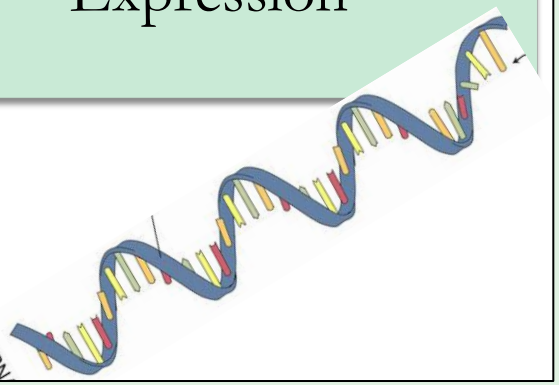


ChIP-sequencing

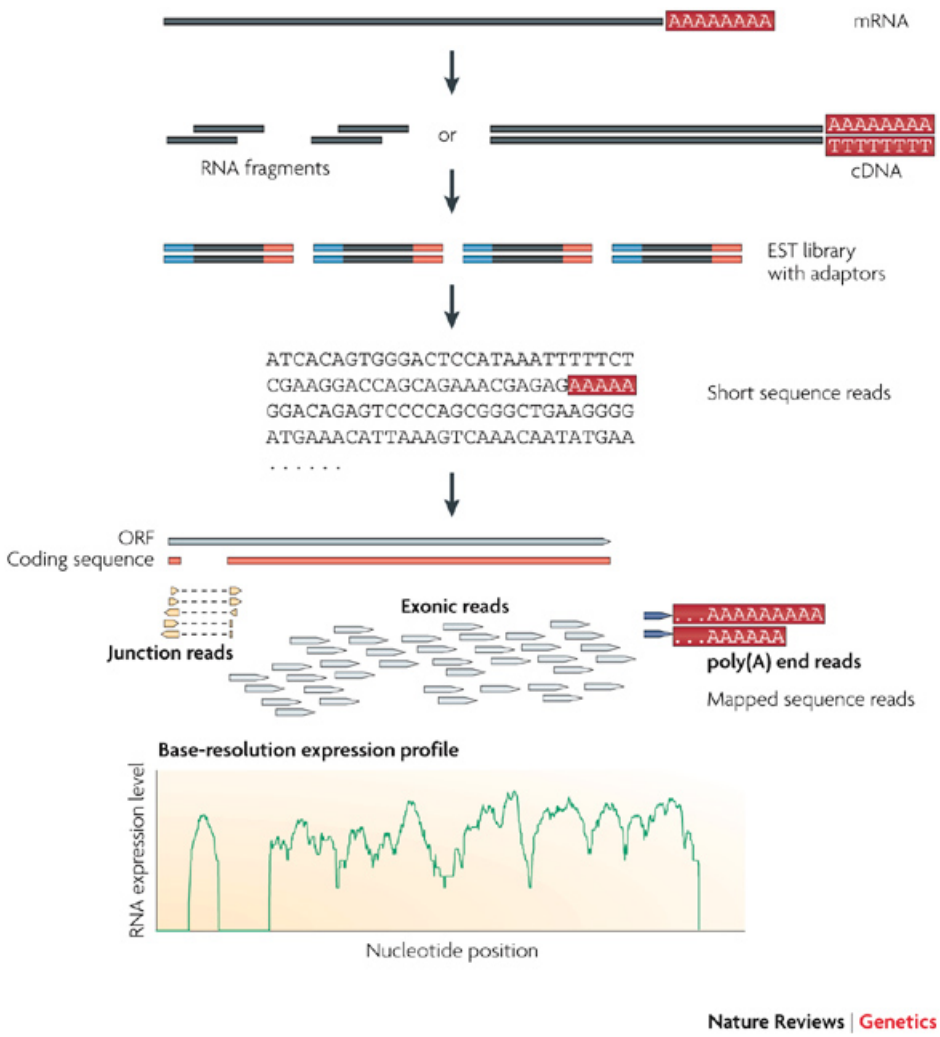
- Transcription Factor Binding Sites
- Chromatin Modification Regions

RNA-sequencing

- Transcriptome Assembly
- Gene Expression
- Differential Expression



RNA-seq workflow



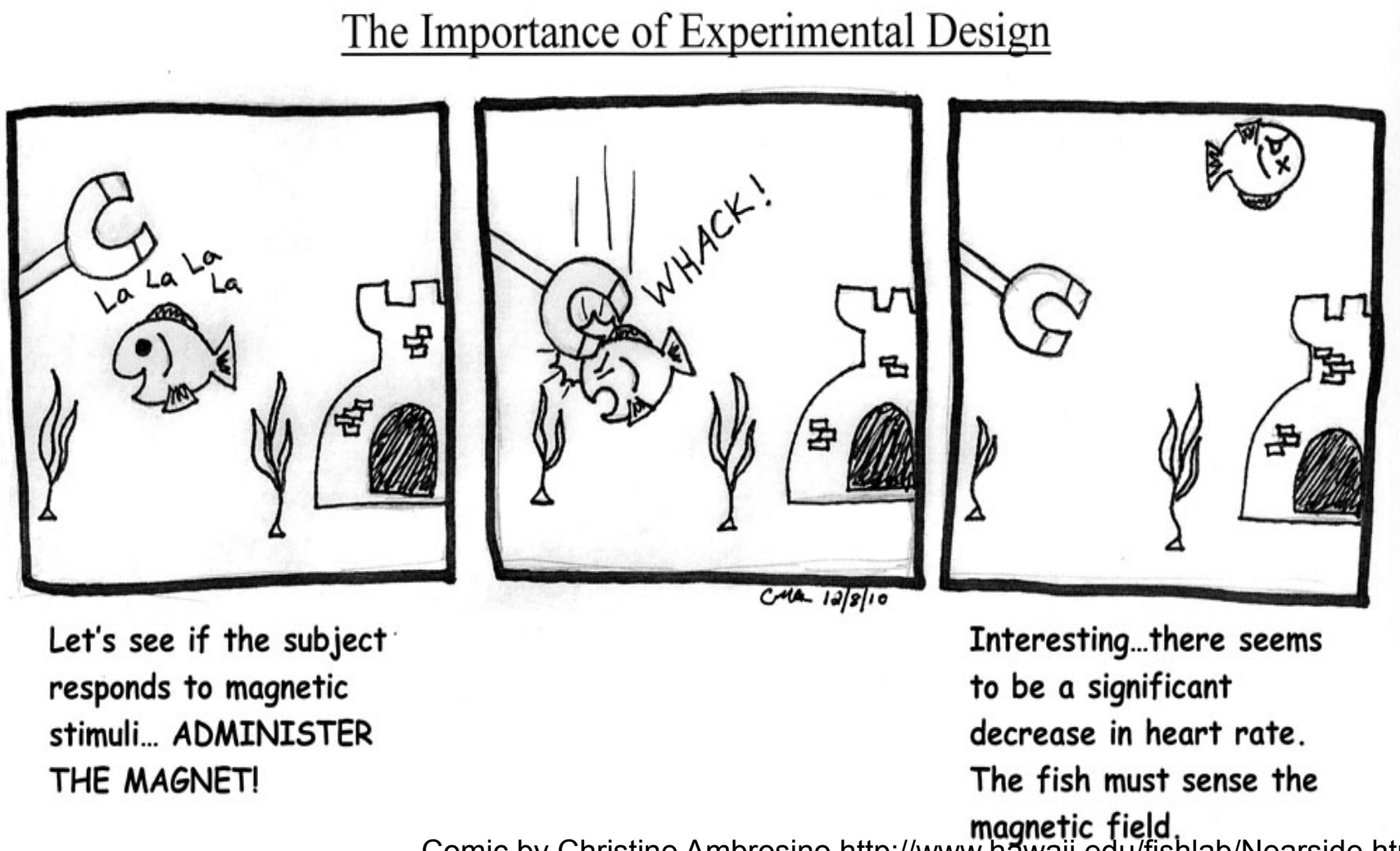
Step1 Library Preparation

Step2 Sequencing

Step3 Bioinformatics Analysis

Image adapted from: Wang, Z., et al. (2009), Nature Reviews Genetics, 10, 57–63.

Designing the right experiment



Comic by Christine Ambrosino http://www.hawaii.edu/fishlab/Nearside.htm

Designing the right experiment

- The design of the experiment is the first step and it is obviously determinant for all downstream analyses

- You have to evaluate all the eventualities and limitations of available technologies, designing the experiment according to your goals

Designing the right experiment

COVERAGE: How many reads do we need?

The coverage is defined as $C = (R_{length} \times R_{num}) / A_{length}$

R_{length} = length in nucleotides of the reads

R_{num} = number of sequenced reads

A_{length} = number of nucleotides of sequenced subject (genome, transcriptome, exome)

The amount of sequencing needed for a given sample is determined by the goals of the experiment and the nature of the RNA sample.

Designing the right experiment

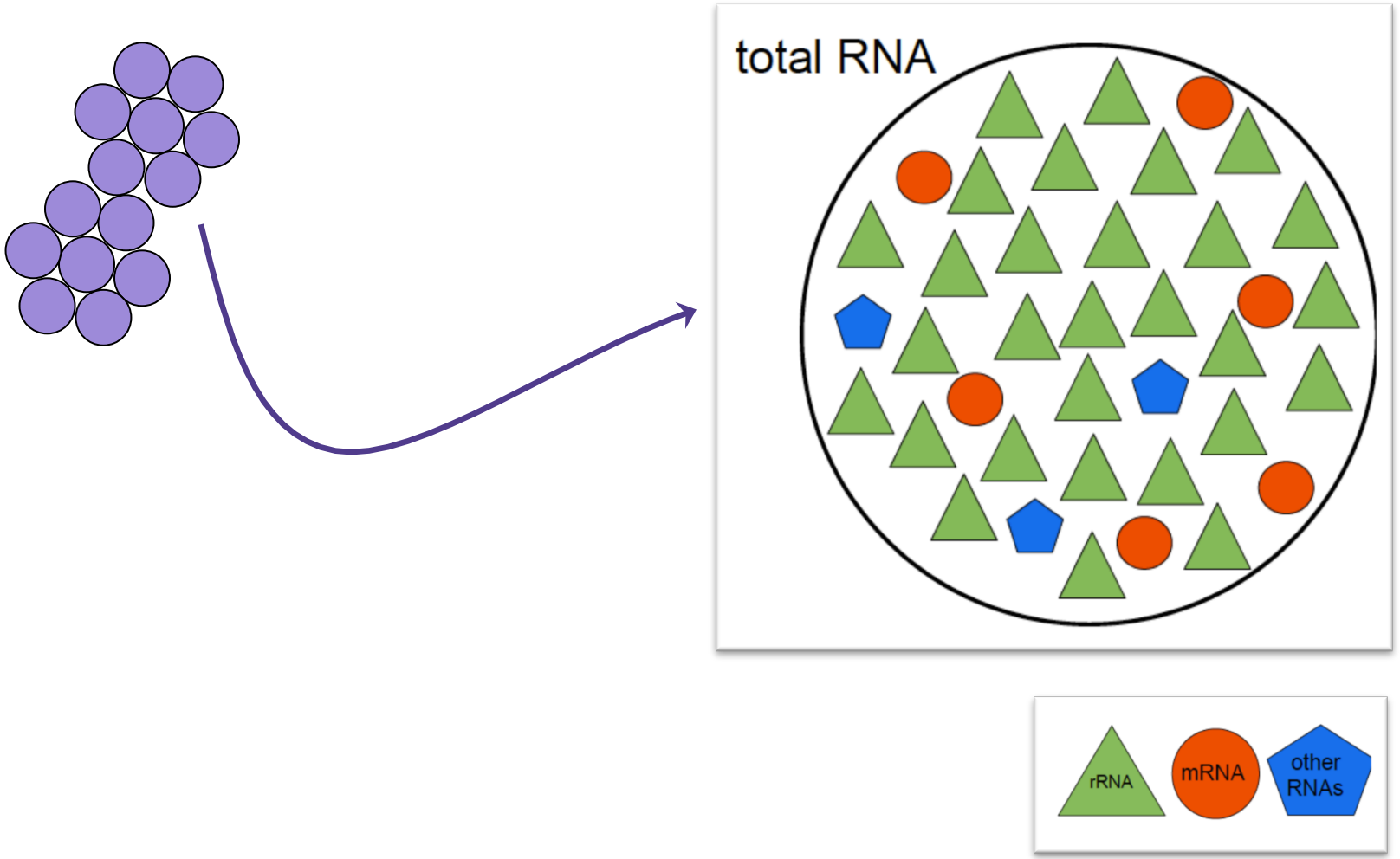
READ LENGTH: long or short reads?

The answer depends again on the experiment:

- GENOME RESEQUENCING
- De novo TRANSCRIPTOME
- TRANSCRIPTOME seq
- ChIP seq

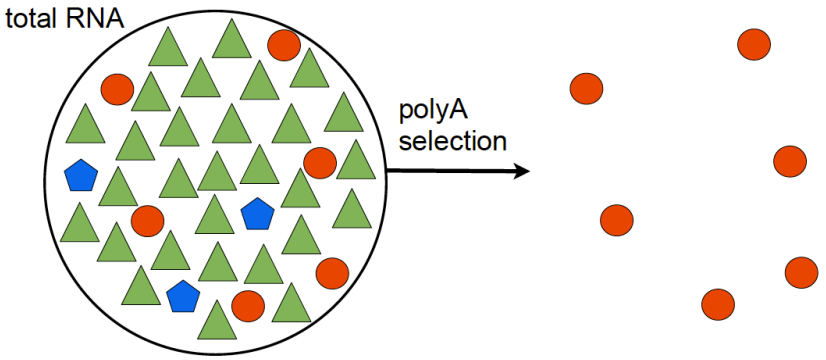
Read length is inversely proportional to the multi-mappability of a read, in a sample of 50 nt reads there is a small fraction (<0.01 %) that can be mapped to multiple positions of the human genome.

Step1 – Library Preparation



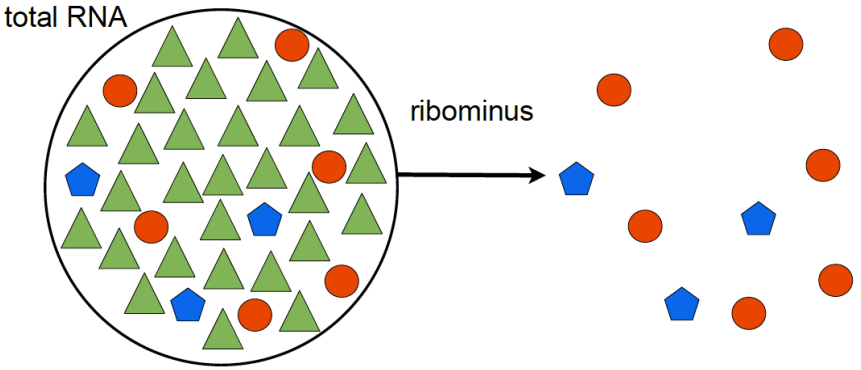
Step 1 – Library Preparation

polyA selection

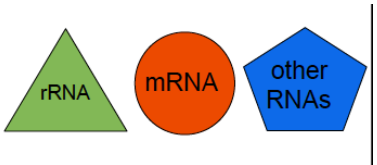


- poly(A+)-transcripts:
  - mRNAs
  - immature microRNAs
  - snoRNAs

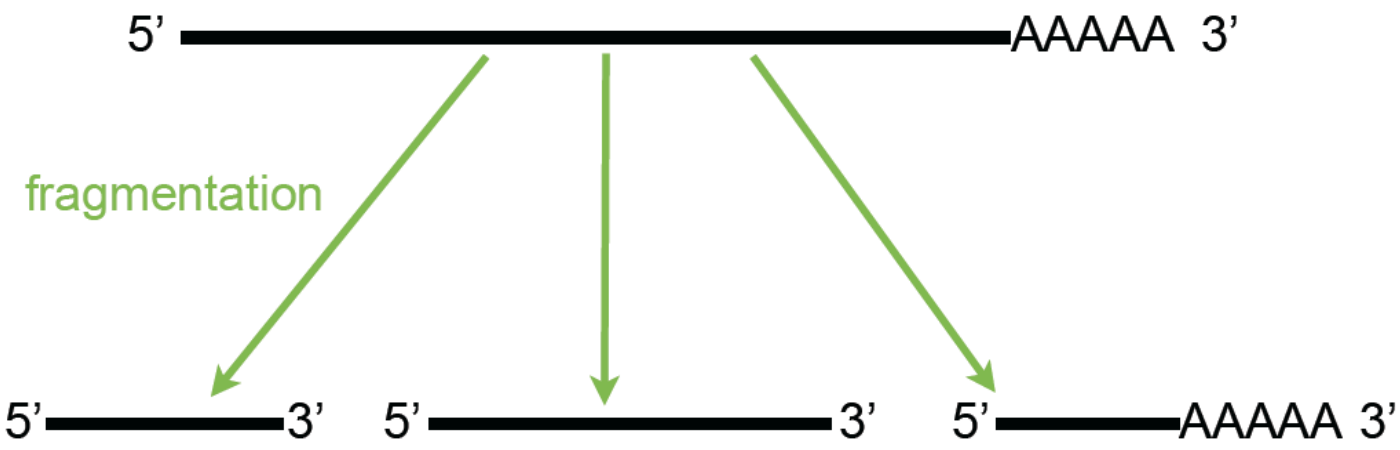
ribominus selection



- non poly(A+)-transcripts:
  - mRNAs
  - histone mRNAs
  - tRNAs
  - other small RNAs



Step 1 – Library Preparation



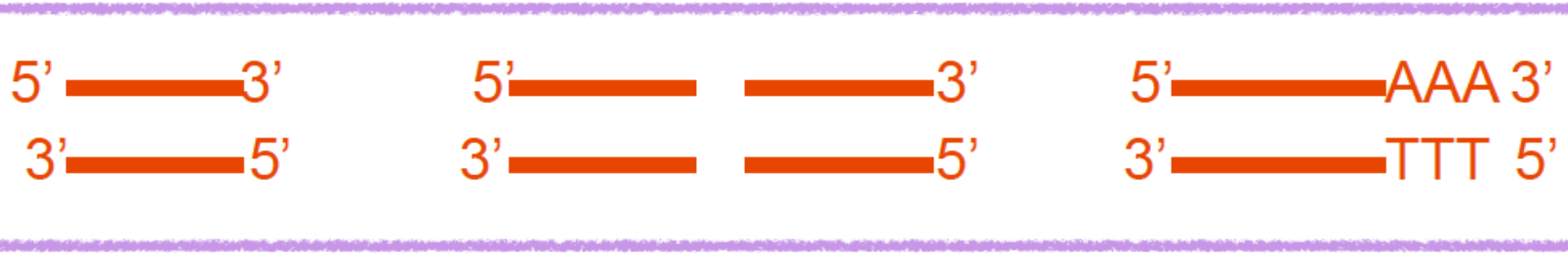
Step 1 – Library Preparation

1 strand cDNA synthesis

remove RNA strand

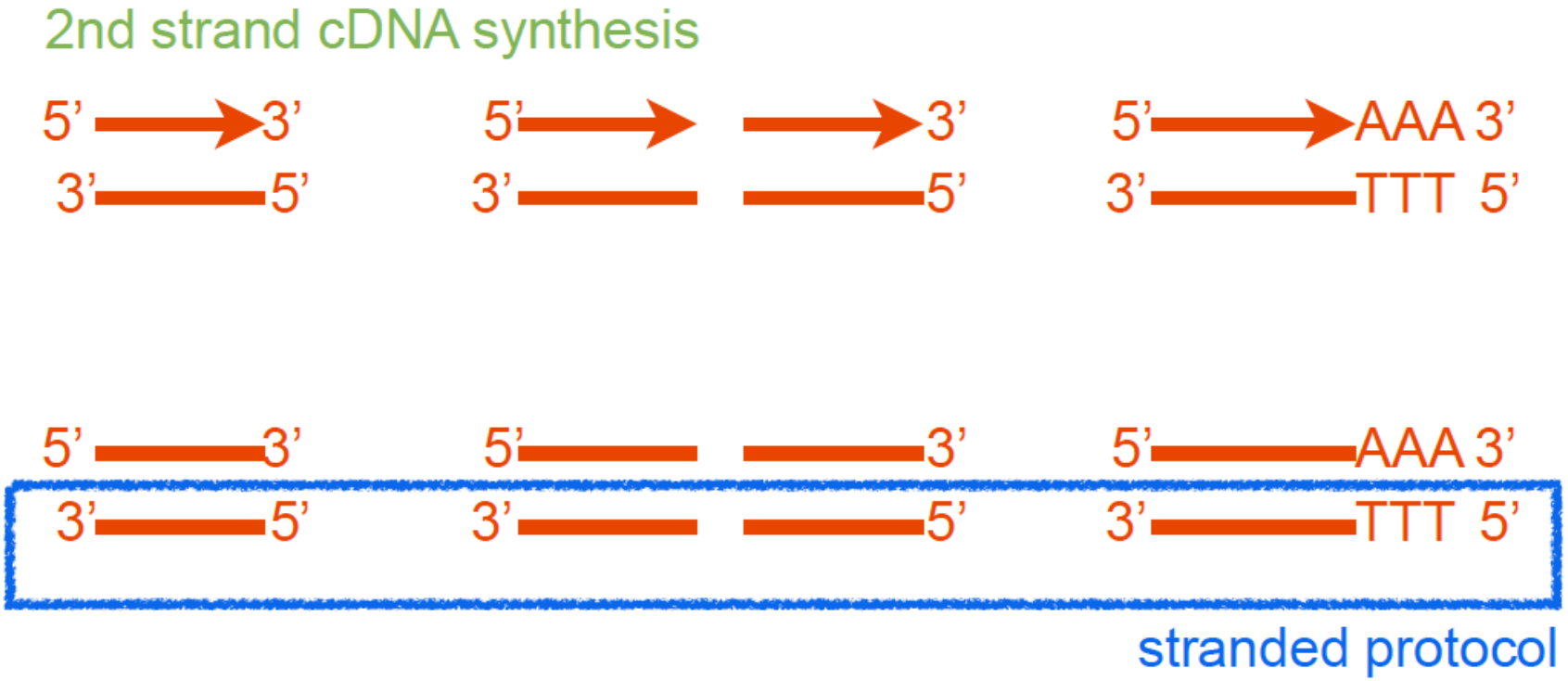
Step 1 – Library Preparation

2nd strand cDNA synthesis



Unstranded protocol

Step 1 – Library Preparation



Parkhomchuk, D., et al. (2009), Nucleic Acids Res, 37 (18), e123.

Step 1 – Library Preparation

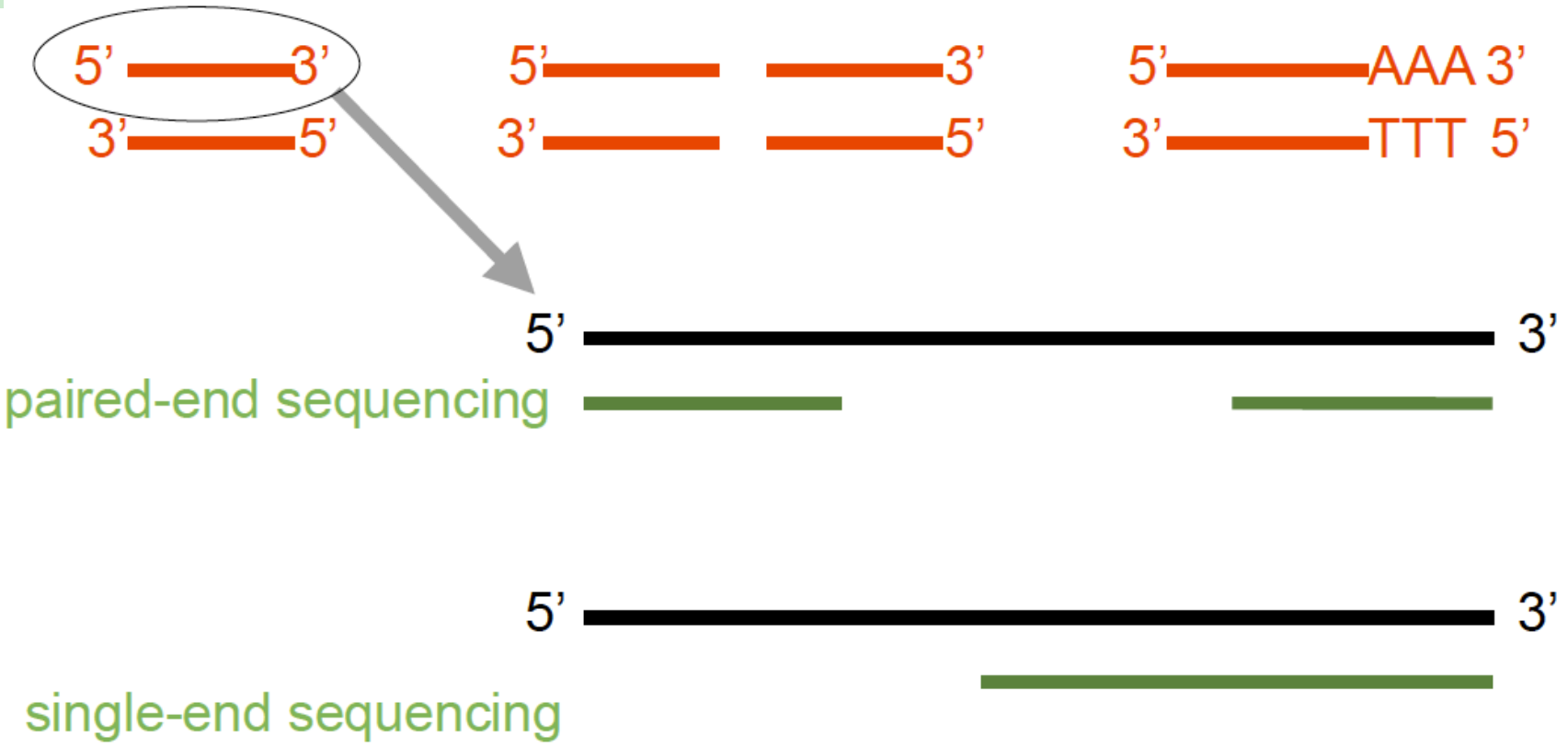
Fragmented cDNA

cDNA with adaptors

Size selection

PCR amplification

Step 2 – Sequencing



Single- vs paired-end sequencing

my_sequence.fastq

```
@HWI-BRUNOP16X_0001:1:1:1466:1018#0/1
AAGGAAGTGCTTGTCTGGCTAACACAGCNAGNCACGTGAC
+
aVfbe`^^^_TTTSSdffffdfffabbZbbfebafbbbbb
```

SE

my_sequence_1.fastq

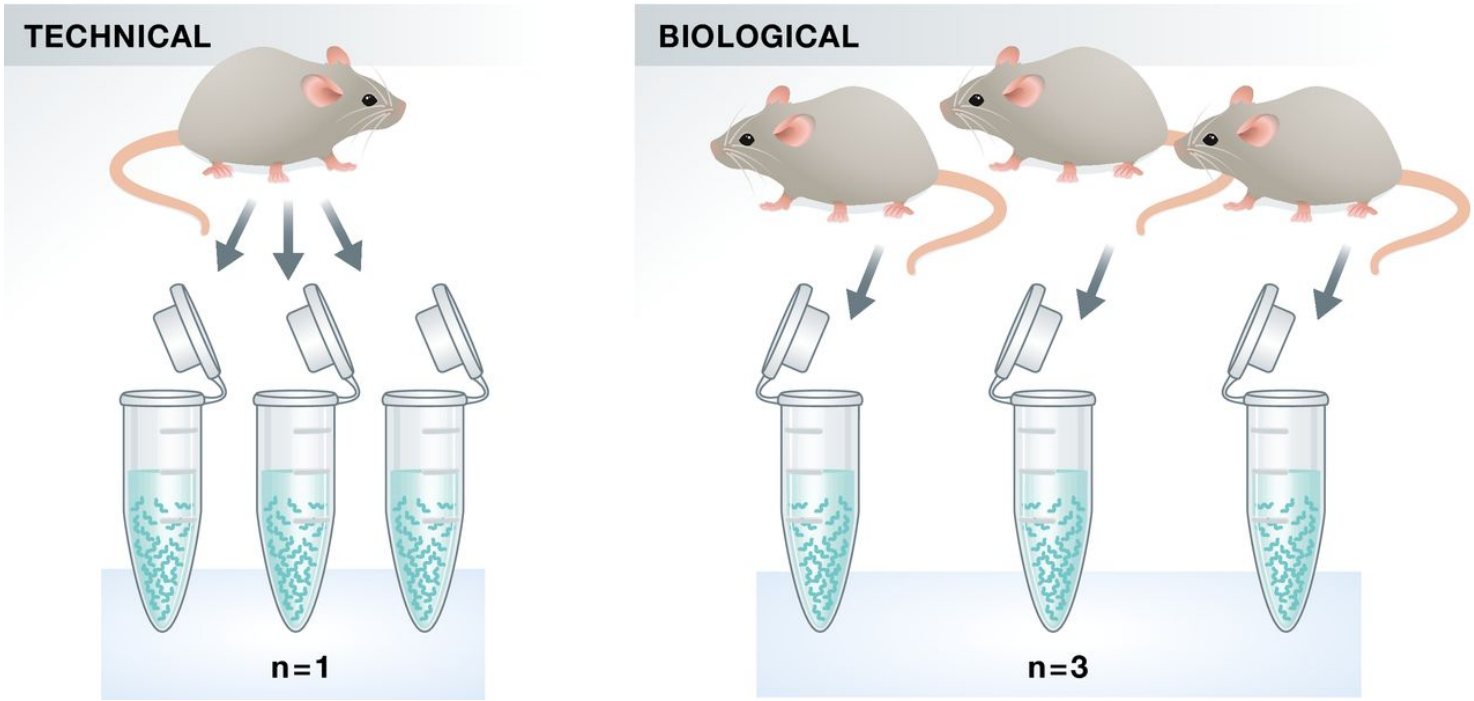
```
@HWI-BRUNOP16X_0001:1:1:1278:989#0/1
NAAATTTCGAATTTCTGTGAAGTAAGCATCTTCTTTGTCAT
+
BJJGGKIINN^^^^^QQNTUQOOTTTRTOTY^^Y^\\^^^\
```

my_sequence_2.fastq

```
@HWI-BRUNOP16X_0001:1:1:1278:989#0/2
AACCCACACAGGAGAGCAGCCTTACAGATGCAAATACTGTG
+
]K___fffffggghegggggggdggggfgggggeggggghh
```

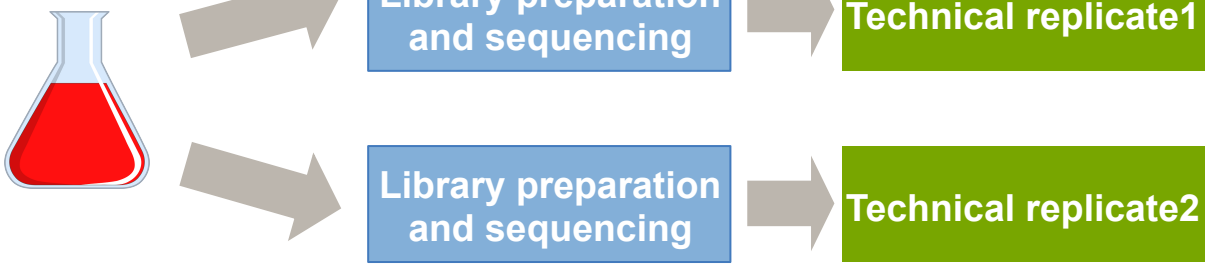
PE

Replicates – do I need them?



10.15252/embj.201592958 | Published online 21.09.2015 The EMBO Journal (2015) e201592958

Replicates – do I need them?

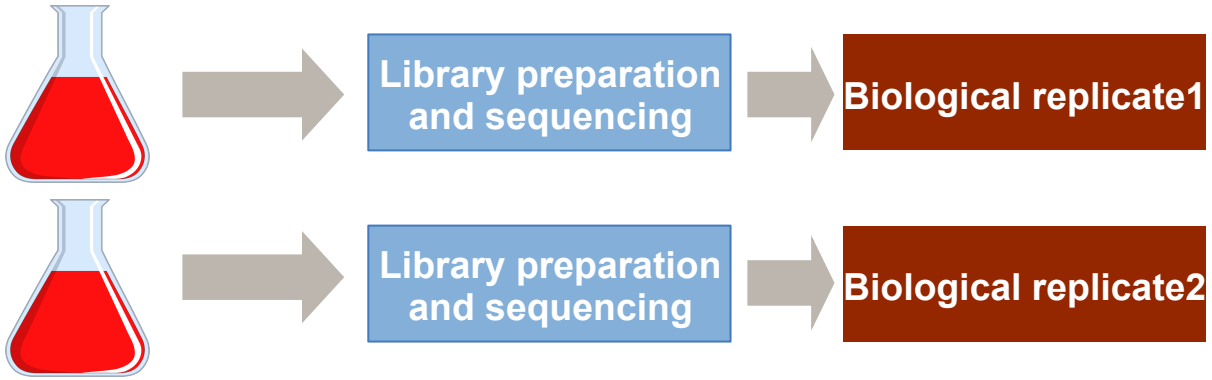


Technical

Same biological sample – same conditions

measure technical variation

Replicates – do I need them?



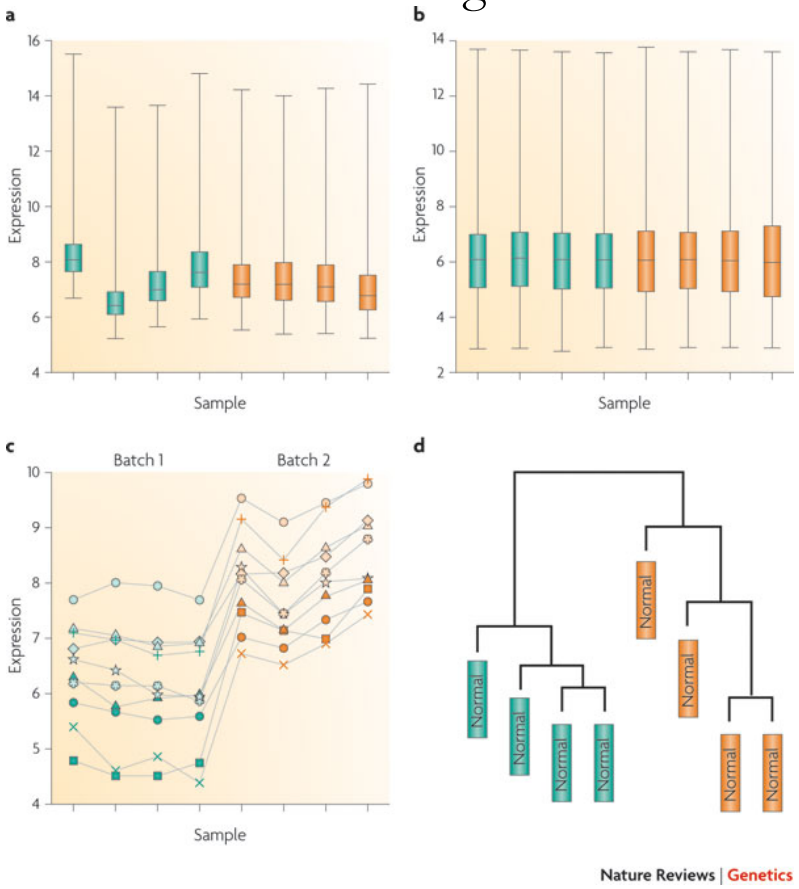
Biological

Different biological sample – same conditions

measure biological variation

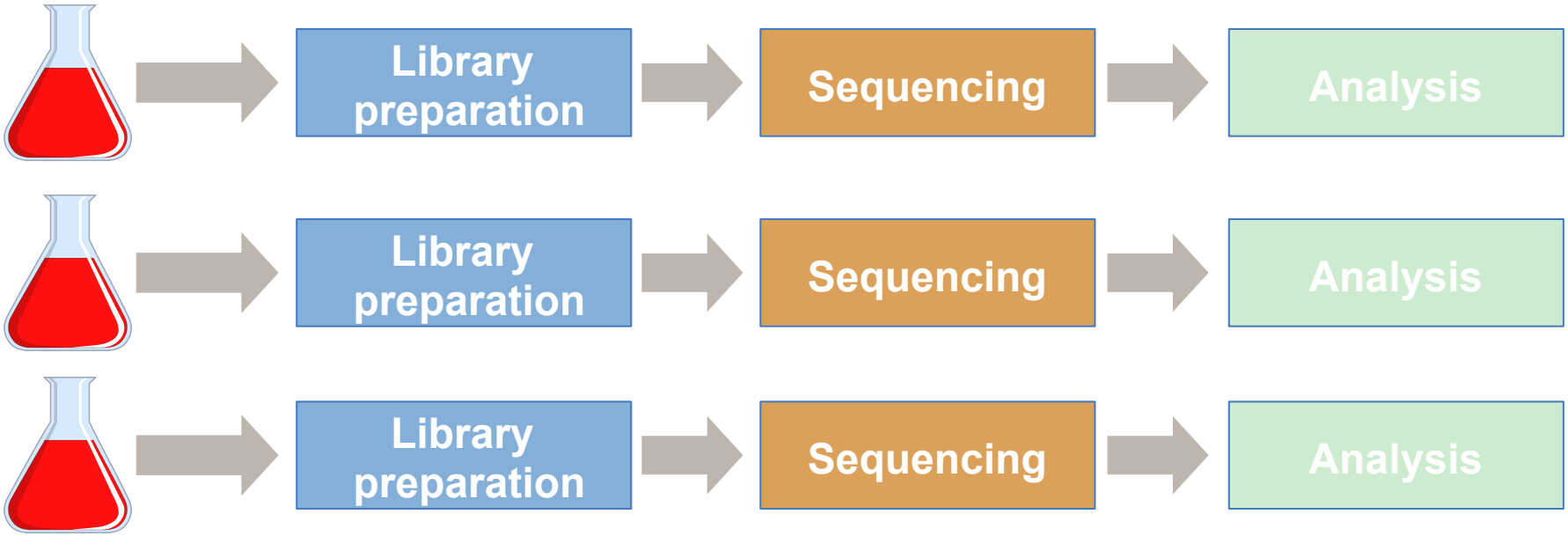
Controlling batch effects

Batch effects are sub-groups of measurements that have qualitatively different behavior across conditions and are unrelated to the biological or scientific variables in a study

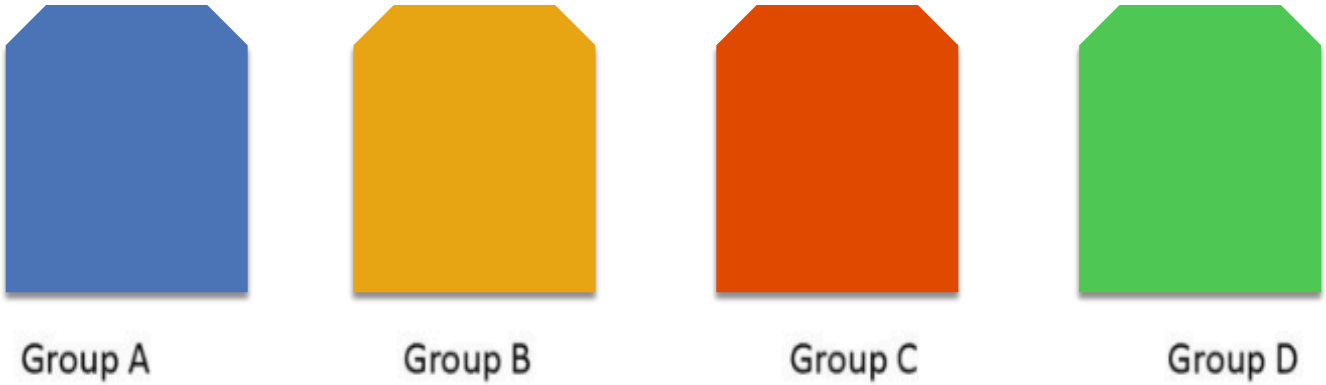


Leek et al. Nature Reviews Genetics 11, 733-739 (October 2010) | doi:10.1038/nrg2825

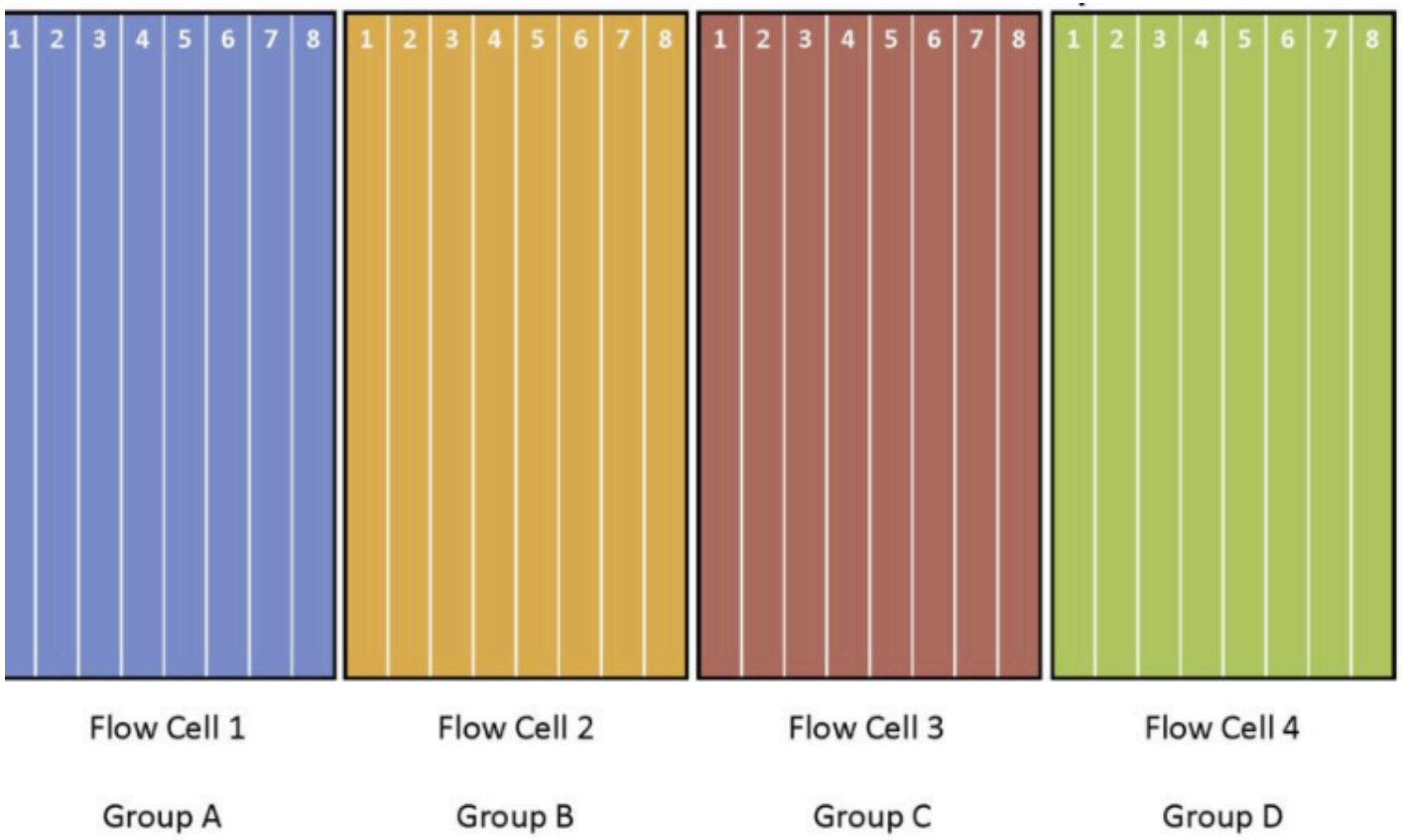
Controlling batch effects



Example of experimental design

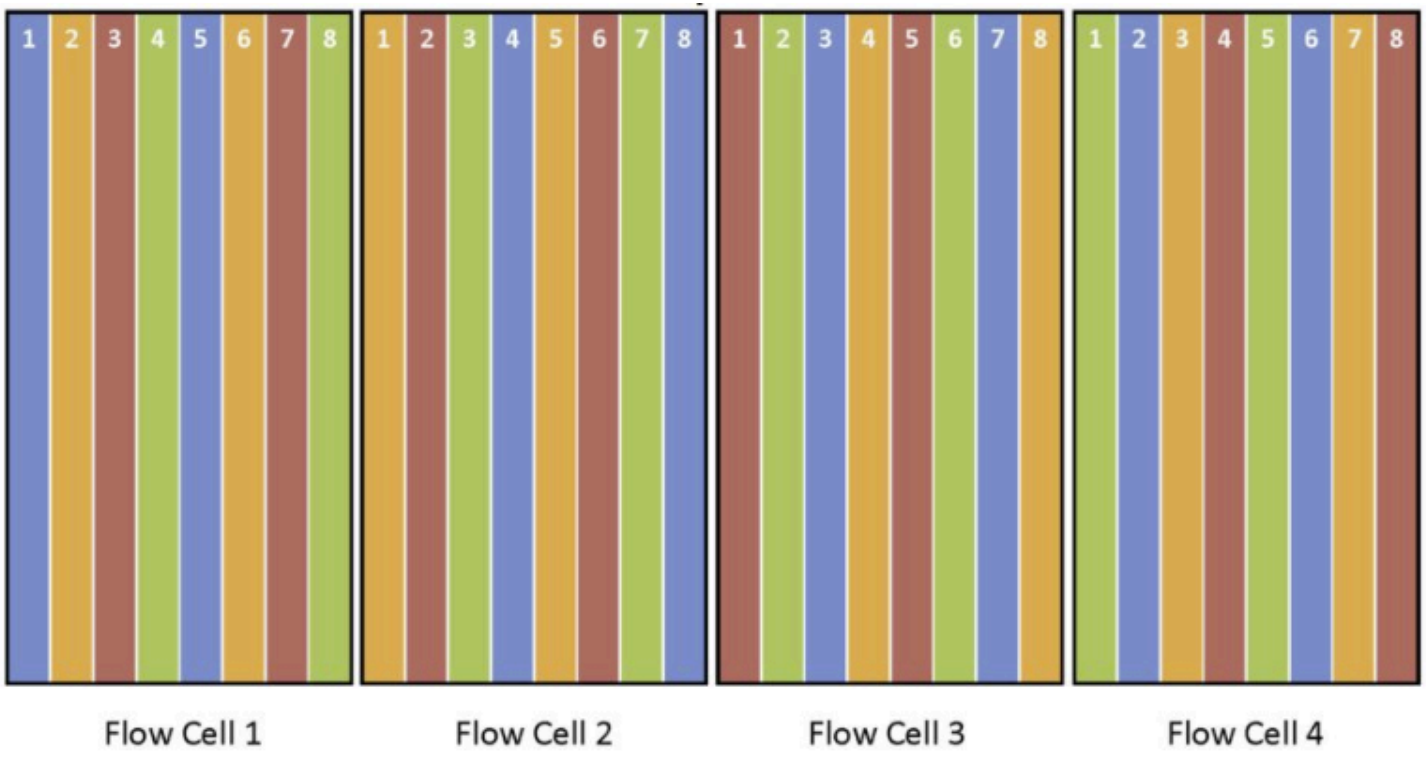


Example of experimental design

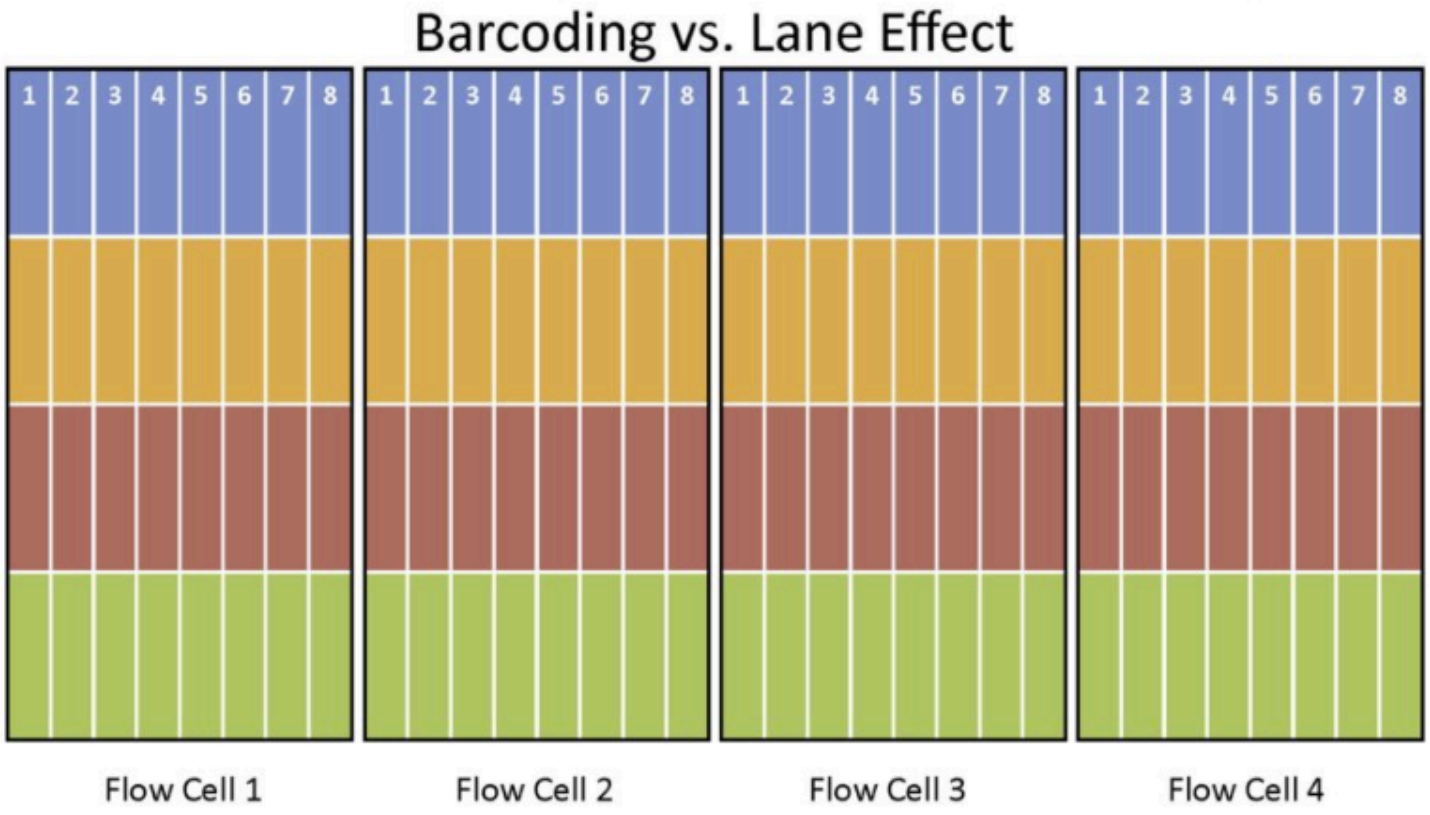


...better experimental design

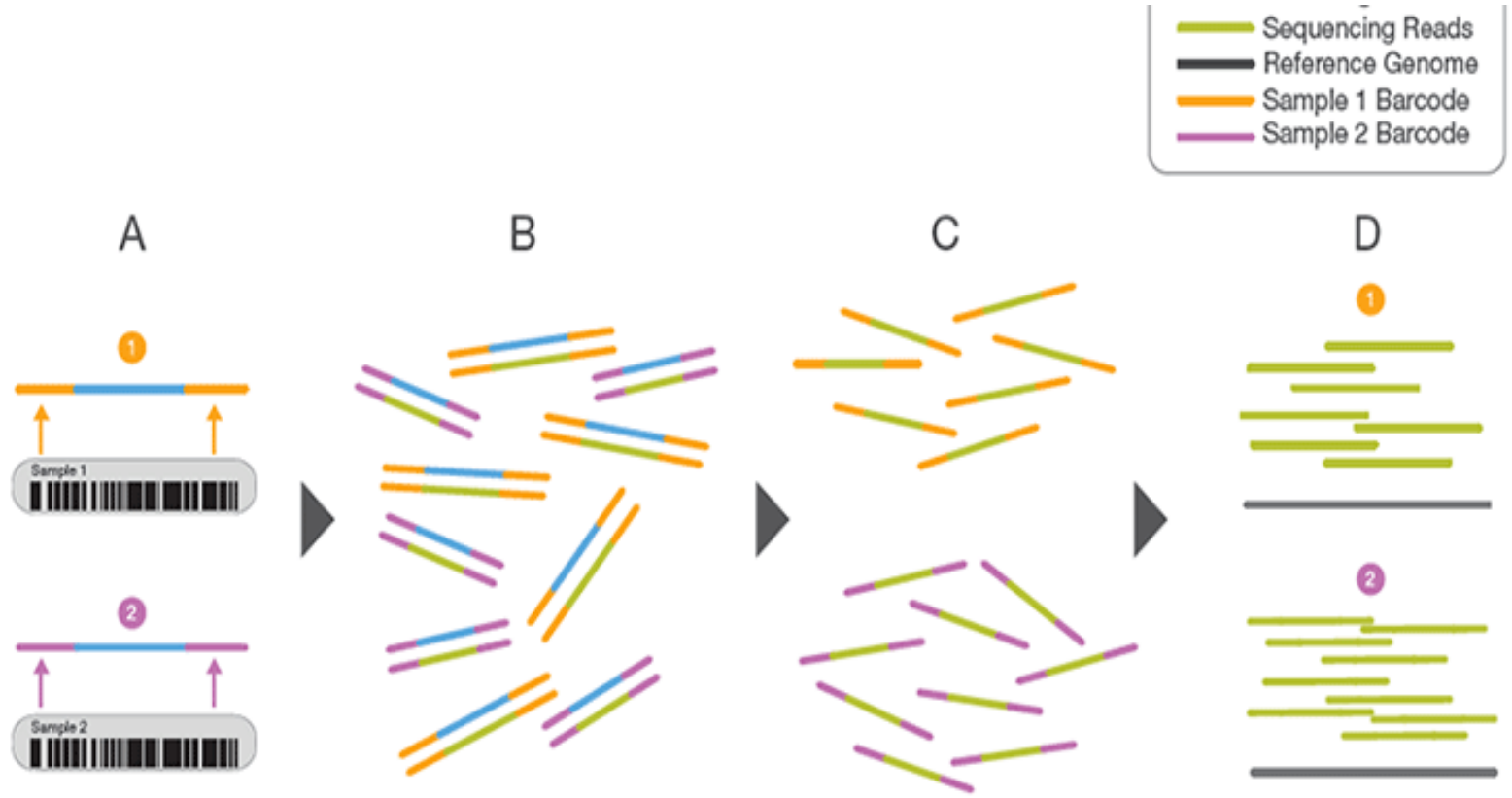
- Randomize samples with respect to the flow cell



...even better experimental design



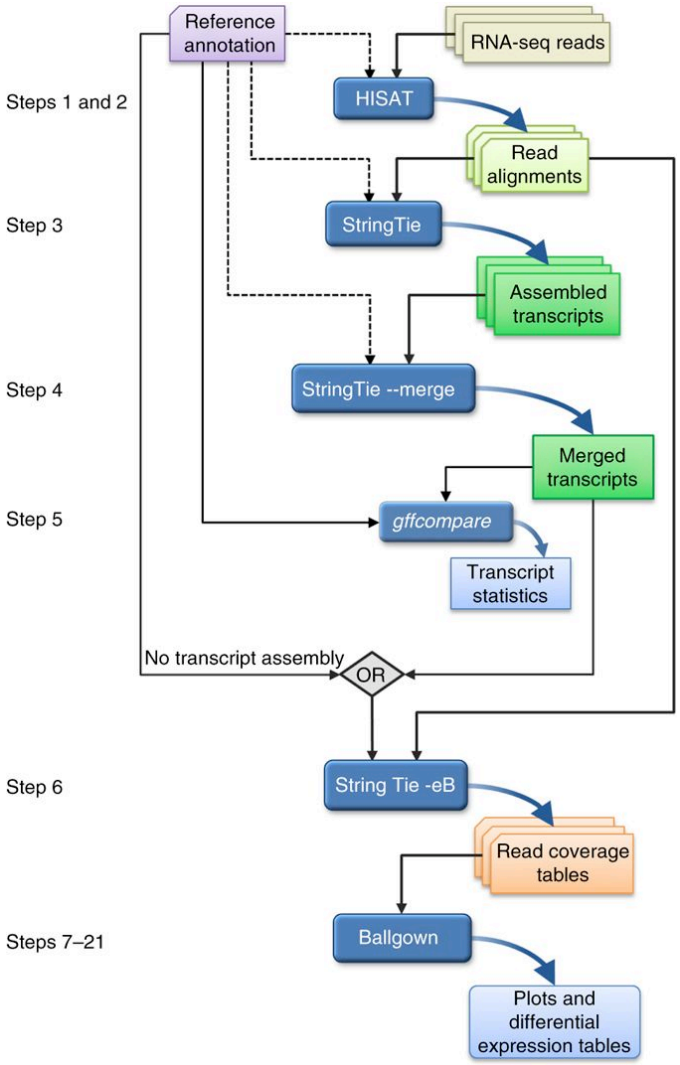
Multiplexing to prevent batch effects



A. Two representative DNA fragments from two unique samples, each attached to a specific barcode sequence that identifies the sample from which it originated.

B. Libraries for each sample are pooled and sequenced in parallel. Each new read contains both the fragment sequence and its sample-identifying barcode.

Step 3 - RNA-seq analysis workflow* (UPDATED)

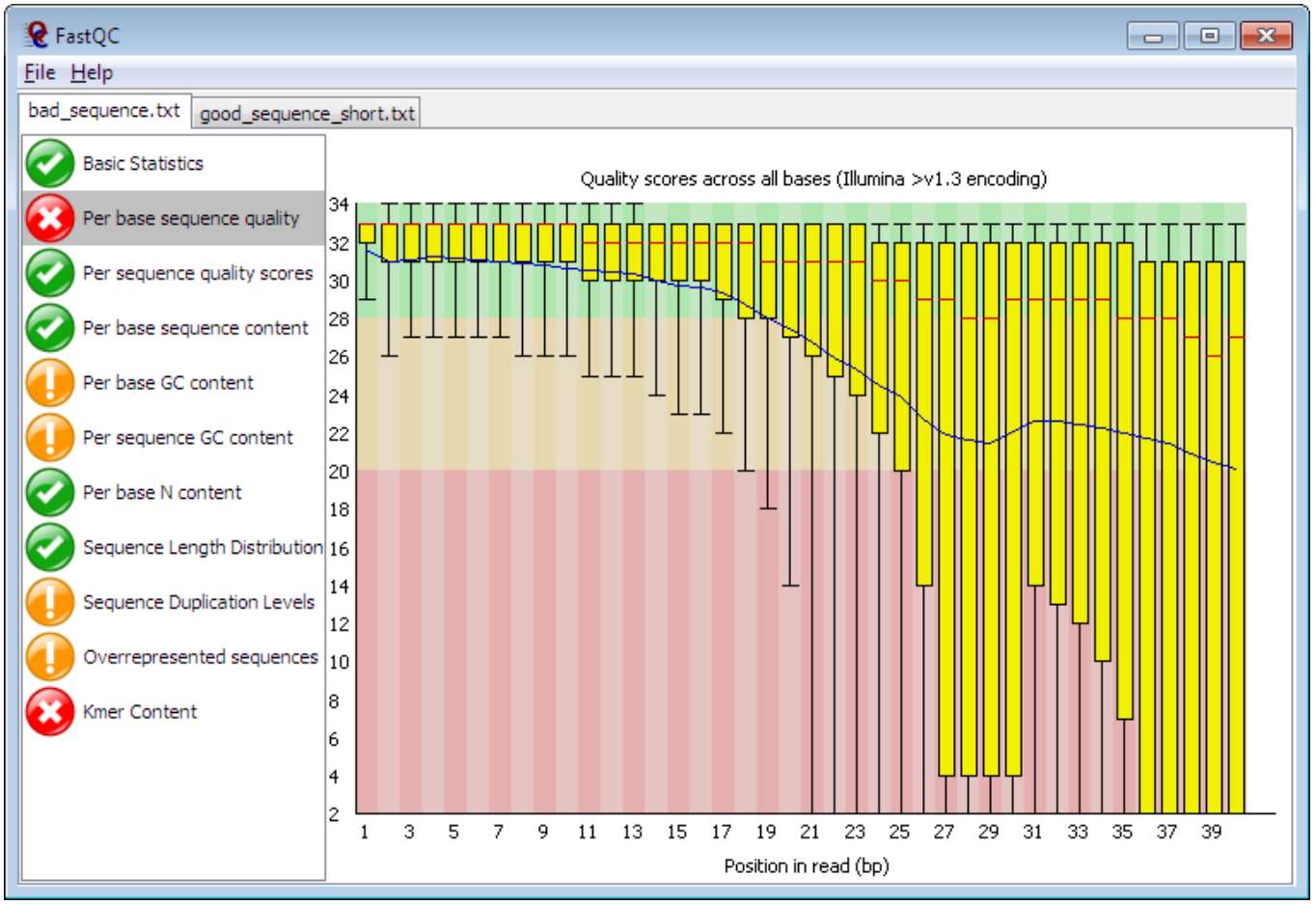


Pertea, M., et al. (2016), Nature protocols

Quality Control

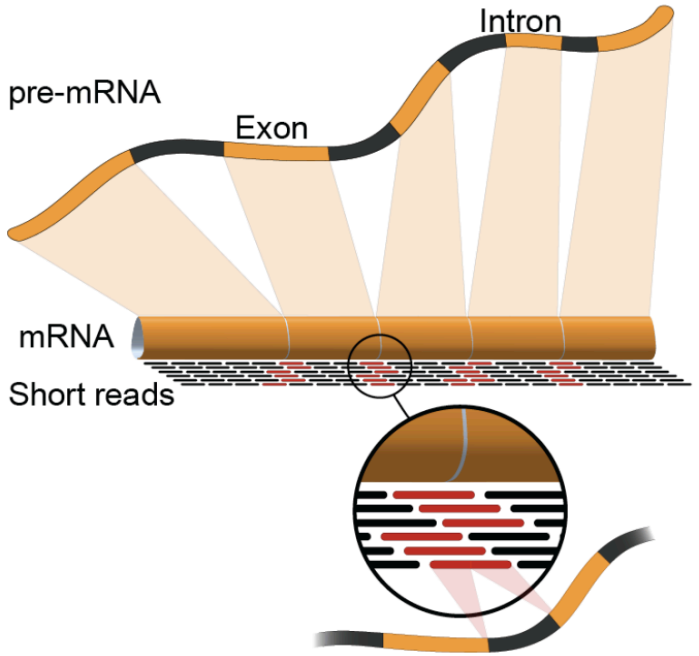
- Essential for downstream analysis.
- Decide sensibly on which data can be filtered out from the downstream analysis.
- You might find yourself going back to that step several times during downstream analysis.

FastQC


FastQC
File Help
bad_sequence.txt
good_sequence_short.txt
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
Quality scores across all bases (Illumina >v1.3 encoding)
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39
Position in read (bp)

Alignment

Class	Category	Package
Read mapping		
Unspliced aligners[a]	Seed methods	Short-read mapping package (SHRiMP)[41]
		Stampy[39]
	Burrows-Wheeler transform methods	Bowtie[43]
		BWA[44]
Spliced aligners	Exon-first methods	MapSplice[52]
		SpliceMap[50]
		TopHat[51]
	Seed-extend methods	GSNAP[53]
		OPAL MA[54]



Garber, M., et al. (2011), Nature Methods, 8(6), 469–477.

Aliments are reported as SAM

Most aligners used their own format to output the alignments.

Hence, downstream comparisons and analyses were difficult to perform.

To resolve this issue, Li et al. have suggested a standardized file format: the Sequence Alignment/Map (SAM) format

SAM is currently the standard format for alignment results

SAMtools is a suite of programs for interacting with high-throughput sequencing data. (http://www.htslib.org/)

SAM FORMAT

A SAM file consists of two parts:

- Header
  - -contains meta data (source of the reads, ref. genome, aligner, etc.)
  - -Header lines necessarily start with "@".
  - -Header fields have standardized two-letter codes for easy parsing

- Alignment section
  - -A tab-separated table with at least 11 columns
  - -Each line describes one alignment

@HD VN:1.4 SO:coordinate
@SQ SN:CHROMOSOME_I LN:15072423
@SQ SN:CHROMOSOME_II LN:15279345
@SQ SN:CHROMOSOME_III LN:13783700
@SQ SN:CHROMOSOME_IV LN:17493793
@SQ SN:CHROMOSOME_V LN:20924149
@SQ SN:CHROMOSOME_X LN:17718866
@SQ SN:CHROMOSOME_MtDNA LN:13794
@SQ SN:sensor_piRNA_mjls144 LN:1663
@CO user command line: /Users/berkyurekahmetcan/Desktop/Data_analysis/STAR-master/bin/MacOSX_x86_64/STAR --runThreadN 4 --genomeDir /Users/berkyurekahmetcan/Desktop/Data_analysis/pichip/totalRNA/data/reference/ --readFilesIn "/Users/berkyurekahmetcan/Dropbox (ericmiskalab)/cambridge-UK/NGS/piChIP/totalRNA/outfilterMultimapper_500/elution/wt1/elution_wt_1.fastq.gz" --readFilesCommand "gunzip -c" --outSAMtype BAM SortedByCoordinate --outMultimapperOrder Random --outFilterMultimapNmax 500 --alignIntronMax 1

Header

L180:540:HTMV2BCXY:1:1112:6120:12935 0 CHROMOSOME_I 3745 255 49M * 0 0
TAGAGGGTTAGACCCAAAATTCAGCCCGCGAAGGCATGACGTCAGCGCG
GGGGGGIIIIIGIIIIIIIIIIGIIIIIIIIIIIIIIIIIIIIIIIII NH:i:1 HI:i:1 AS:i:44 nM:i:2
L180:540:HTMV2BCXY:1:1113:17520:75431 0 CHROMOSOME_I 3745 255 49M * 0 0
TAGAGGGTTAGACCCAAAATTCAGCCCGCGAAGGCATGACGTCAGCGCG
GGGGGIIIIIGIIIIIIIIIIIGIIIIIIIIGIIIIIIIIIIIGIIII NH:i:1 HI:i:1 AS:i:44 nM:i:2
L180:540:HTMV2BCXY:1:2111:6429:77948 0 CHROMOSOME_I 3745 255 49M * 0 0
TAGAGGGTTAGACCCAAAATTCAGCCCGCGAAGGCATGACGTCAGCGCG
GGGGGIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII NH:i:1 HI:i:1 AS:i:44 nM:i:2

Aligments

L180:540:HTMV2BCXY:1:2202:18496:19290 0 CHROMOSOME_I 3745 255 49M * 0 0
TAGAGGGTTAGACCCAAAATTCAGCCCGCGAAGGCATGACGTCAGCGCG
GGGGGIIIIIIIIIIIIIIIIIGIIIIIIIIGIIIIIIIIIGIIIII NH:i:1 HI:i:1 AS:i:44 nM:i:2
L180:540:HTMV2BCXY:1:2205:19372:73018 0 CHROMOSOME_I 3745 255 49M * 0 0
TAGAGGGTTAGACCCAAAATTCAGCCCGCGAAGGCATGACGTCATCGCG
GGGGGIIIIIGIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.GGII NH:i:1 HI:i:1 AS:i:42 nM:i:3
L180:540:HTMV2BCXY:1:2213:1720:60756 0 CHROMOSOME_I 3745 255 49M * 0 0
TAGAGGGTTAGACCCAAAATTCAGCCCGCGAAGGCATGACGTCAGCGCG

SAM FORMAT

1)QNAME: ID of the read (“query”)

2)**FLAG**: alignment flags

3)RNAME: ID of the reference (typically: chromosome name)

4)POS: Position in reference (1-based, left side)

5)MAPQ: Mapping quality (as Phred score)

6)**CIGAR**: Alignment description (gaps etc.) in CIGAR format

7)MRNM: Mate reference sequence name [for paired end data]

8)MPOS: Mate position [for paired end data]

9)ISIZE: inferred insert size [for paired end data]

10)SEQ: sequence of the read

11)QUAL: quality string of the read

N)EXTRA fields

SAM FORMAT

The flag (F2) is a number that gives precise information about the alignment:
https://broadinstitute.github.io/picard/explain-flags.html

Field	Description
0x0001	the read is paired in sequencing
0x0002	the read is mapped in a proper pair
0x0004	the query sequence itself is unmapped
0x0008	the mate is unmapped
0x0010	seq being reverse complemented
0x0020	mate seq being reversed
0x0040	the first segment in the template
0x0080	the last segment in the template
0x0100	secondary alignment
0x0200	not passing quality controls
0x0400	PCR or optical duplicate
0x0800	supplementary alignment

The information can be summed: E.G. an unpaired read that aligns to the reverse reference strand will have flag 16. A paired-end read that aligns and is the first mate in the pair will have flag 83 (= 64 + 16 + 2 + 1).

SAM FORMAT

The CIGAR (F6) is a representation of the alignment:

```
RefPos:     1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19
Reference:  C  C  A  T  A  C  T  G  A  A  C  T  G  A  C  T  A  A  C

Read: ACTAGAATGACT
```

```
RefPos:     1  2  3  4  5  6  7     8  9 10 11 12 13 14 15 16 17 18 19
Reference:  C  C  A  T  A  C  T     G  A  A  C  T  G  A  C  T  A  A  C
Read:                  A  C  T  A  G  A  A     T  G  A  C  T
```

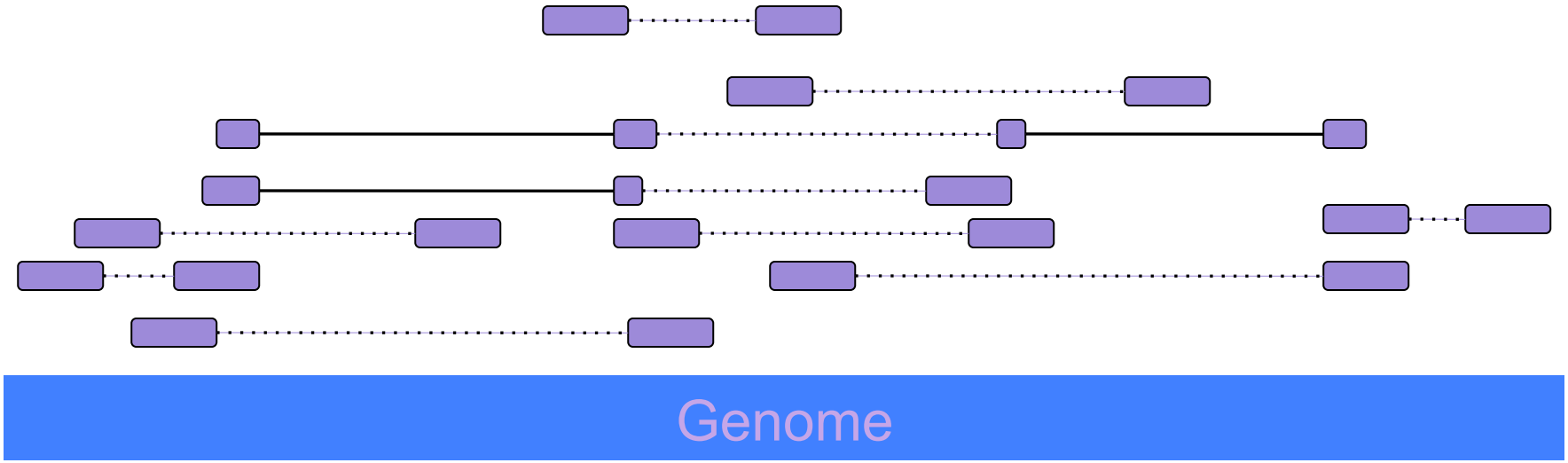
POS: 5
CIGAR: 3M1I3M1D5M

SAMtools

SAMtools are a set of simple tools useful to:

- -convert between SAM and BAM
  - -SAM: a human-readable text file
  - -BAM: a binary version of a SAM file, suitable for fast processing
- -sort and merge SAM files
- -index SAM and FASTA files for fast access
- -view alignments (“tview”)
- -produce a “pile-up”, i.e., a file showing
  - -local coverage
  - -mismatches and consensus calls
  - -indels

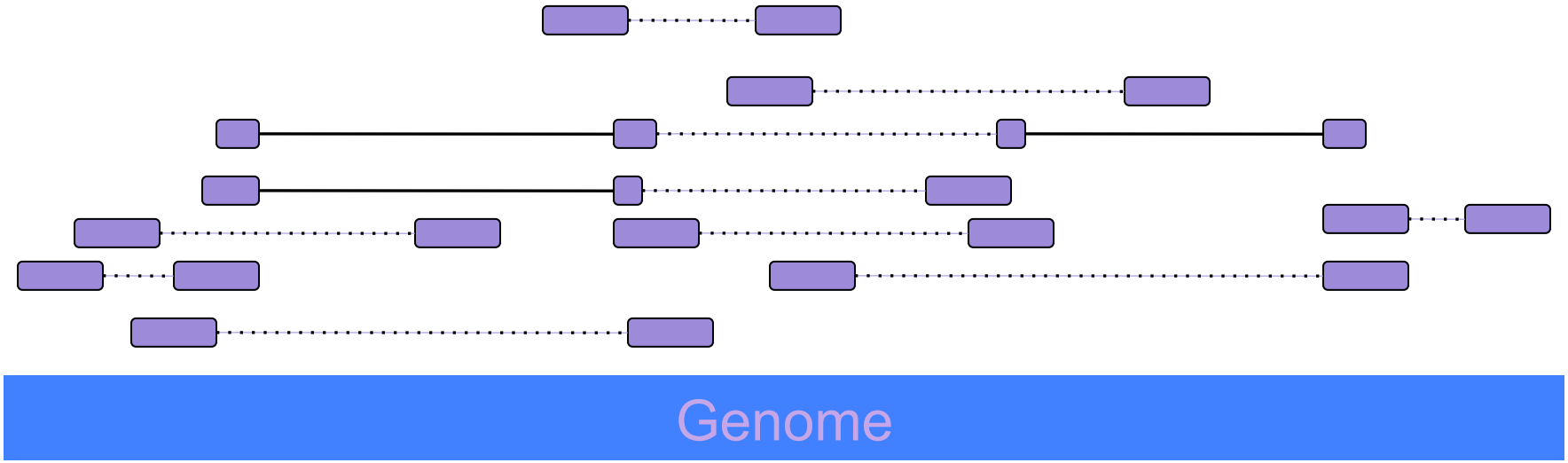
Transcriptome assembly



Transcriptome assembly

STEP 1: Identify fragments that cannot have originated from the same transcript

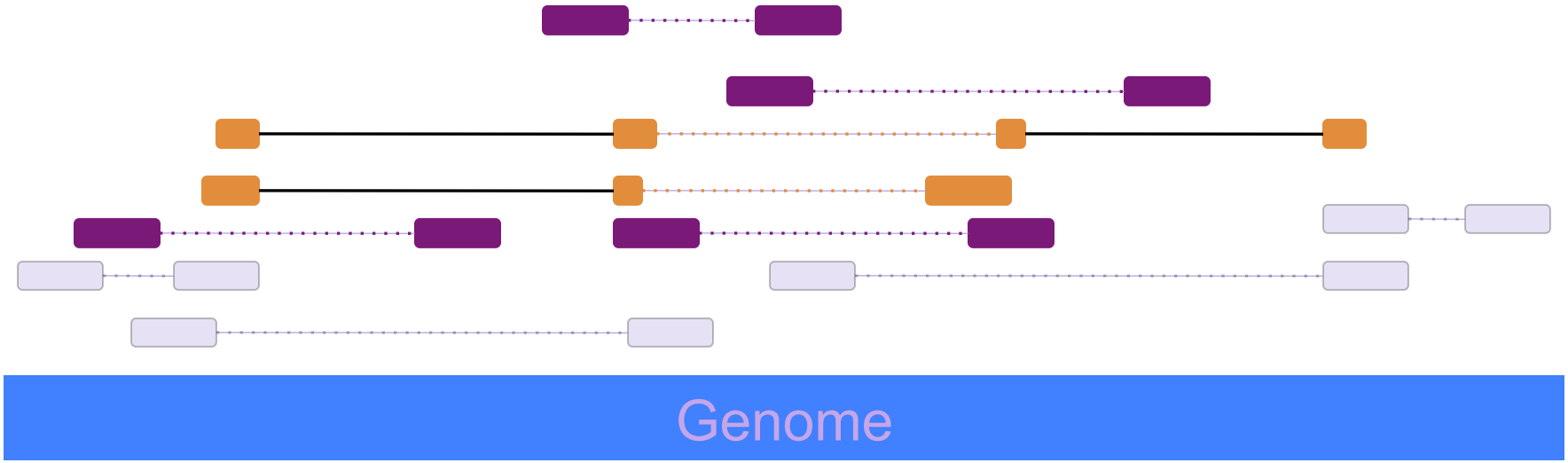
(Hint: Use spliced reads)



Transcriptome assembly

STEP 1: Identify fragments that cannot have originated from the same transcript

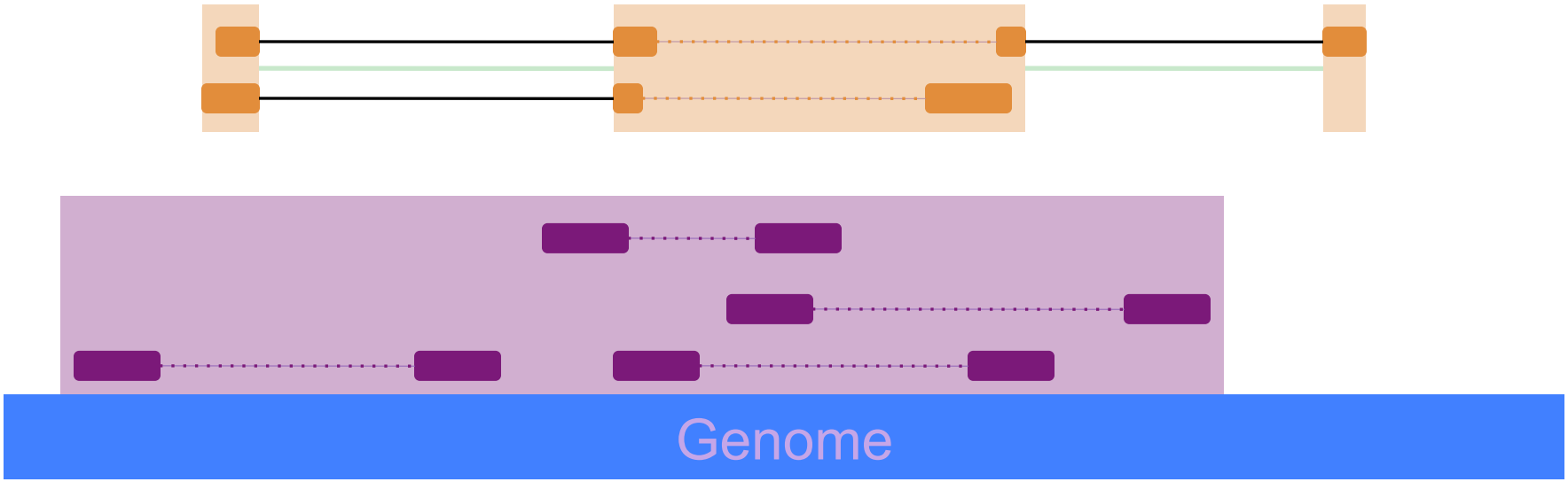
(Hint: Use spliced reads)



Transcriptome assembly

STEP 2: Connect ‘incompatible’ fragments into directed graphs

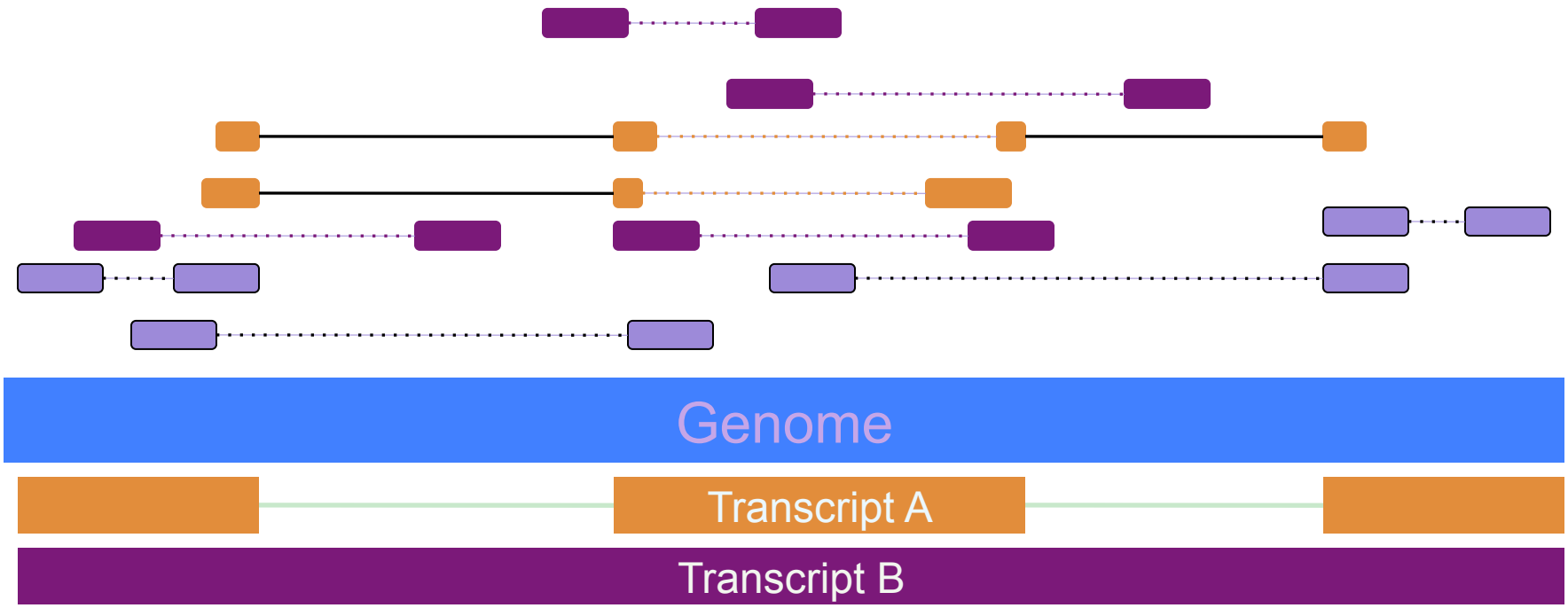
(Hint: Use paired reads)



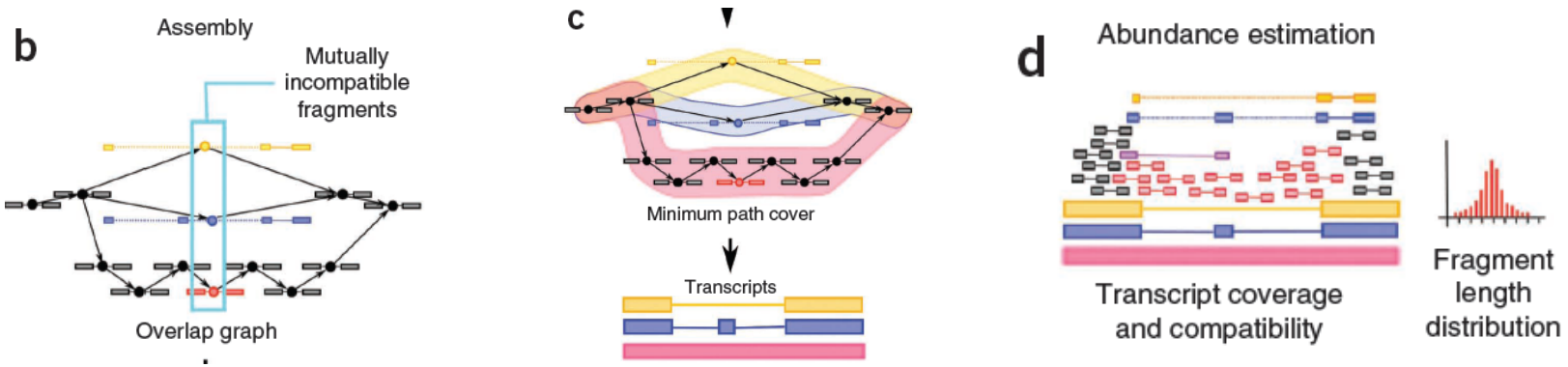
Transcriptome assembly

STEP 3: Assemble transcripts

STEP 4: Quantify transcript expression



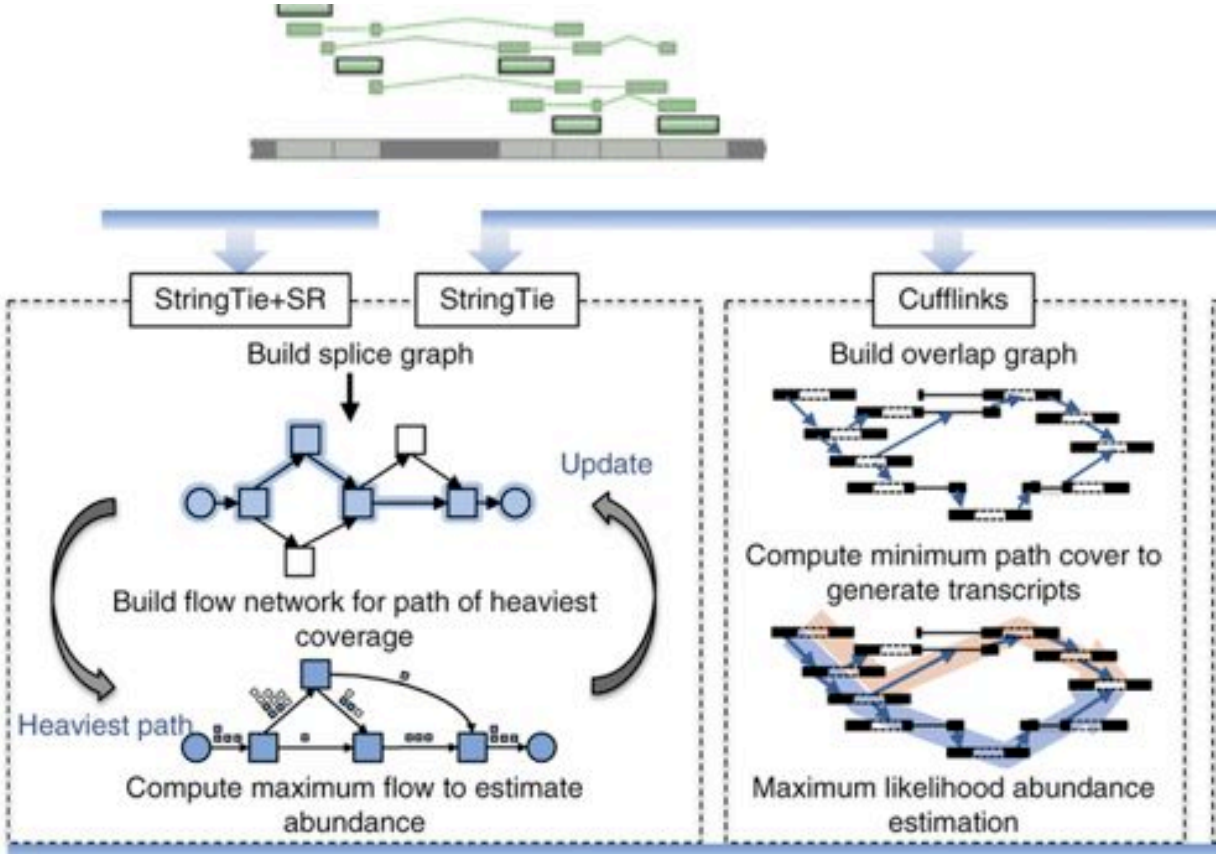
Cufflinks

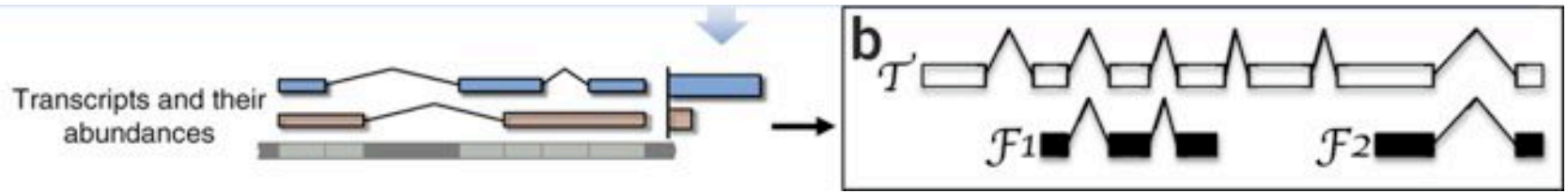


Transcript abundance is estimated in FPKMs (Fragments Per Kilobase of exon per Million fragments mapped)

Trapnell, C., et al. (2010), Nature biotechnology, 28(5), 511–515

StringTie


StringTie+SR
StringTie
Cufflinks
Build splice graph
Update
Build flow network for path of heaviest coverage
Heaviest path
Compute maximum flow to estimate abundance
Build overlap graph
Compute minimum path cover to generate transcripts
Maximum likelihood abundance estimation


Transcripts and their abundances
b
$\mathcal{T}$
$\mathcal{F}1$
$\mathcal{F}2$

Reference sequence: NOT FOUND

- *De novo* transcriptome assembly

- Requirements:
  - Deep sequencing and/or longer reads
  - Thorough quality control
  - Large memory/Multiple processors
  - Patience

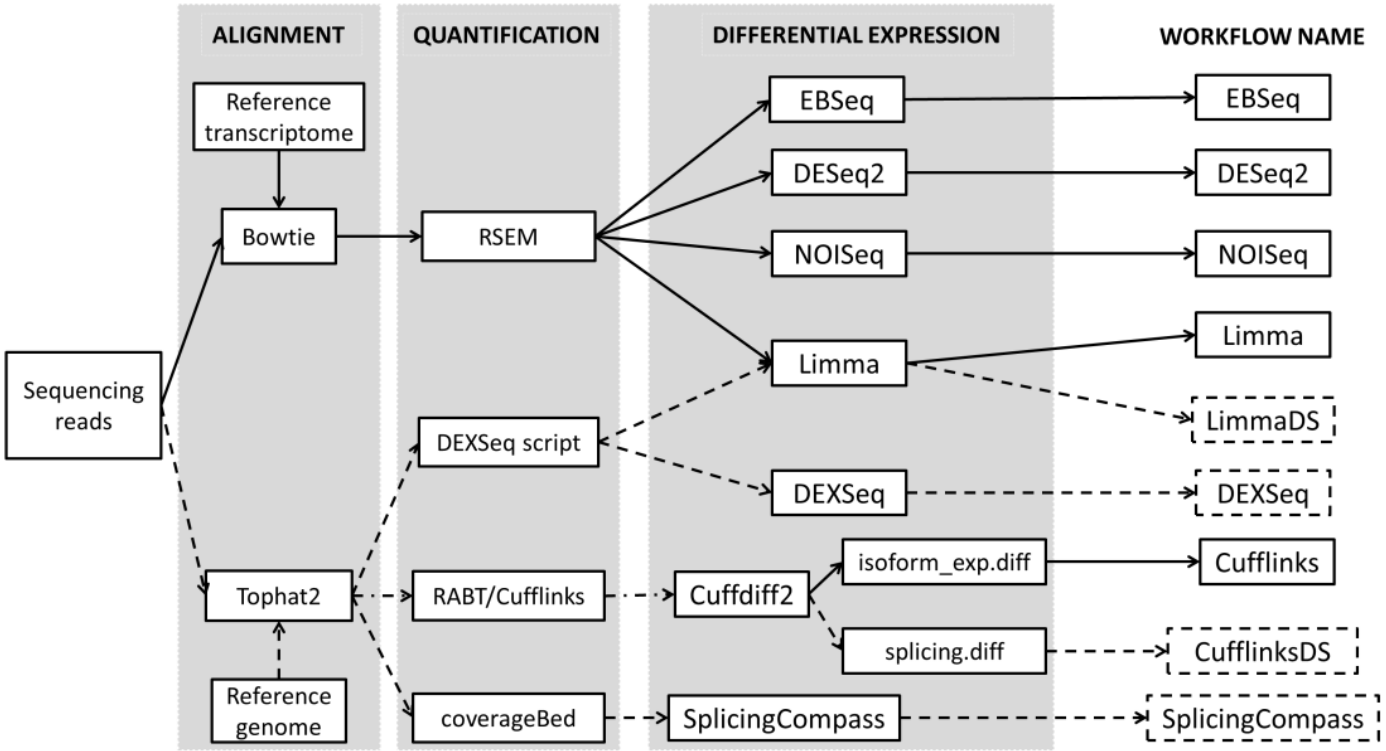
- Tools:
  - Velvet/Oases: http://www.ebi.ac.uk/~zerbino/oases/
  - Trinity: http://trinityrnaseq.sourceforge.net/
  - Trans-ABySS: http://www.bcgsc.ca/platform/bioinfo/software/trans-abyss
  - MIRA, CLC etc.

Differential Expression Analysis

- Use statistical testing to decide whether an observed difference in read counts is significant.

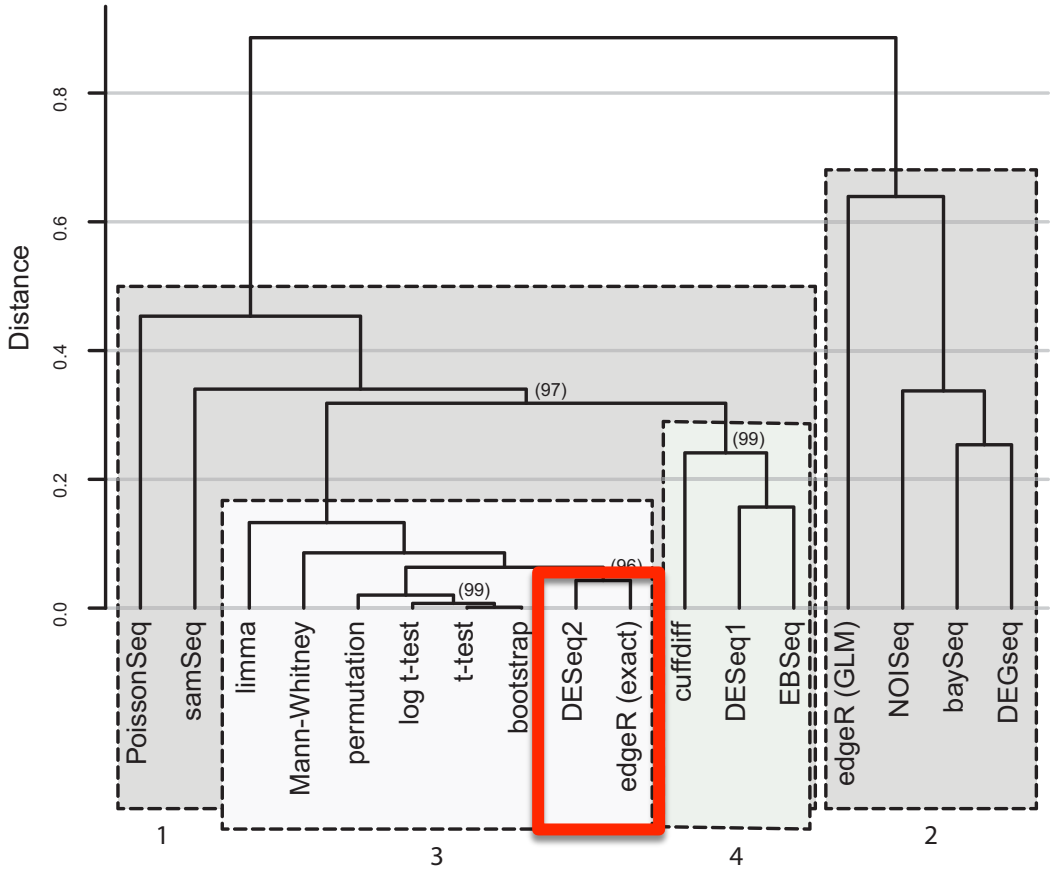
- Which genes/isoforms are being expressed at different levels in different conditions?

Differential Expression Tools



Merino, GA., et al. (2017), bioRxiv

Different tools outcome different results

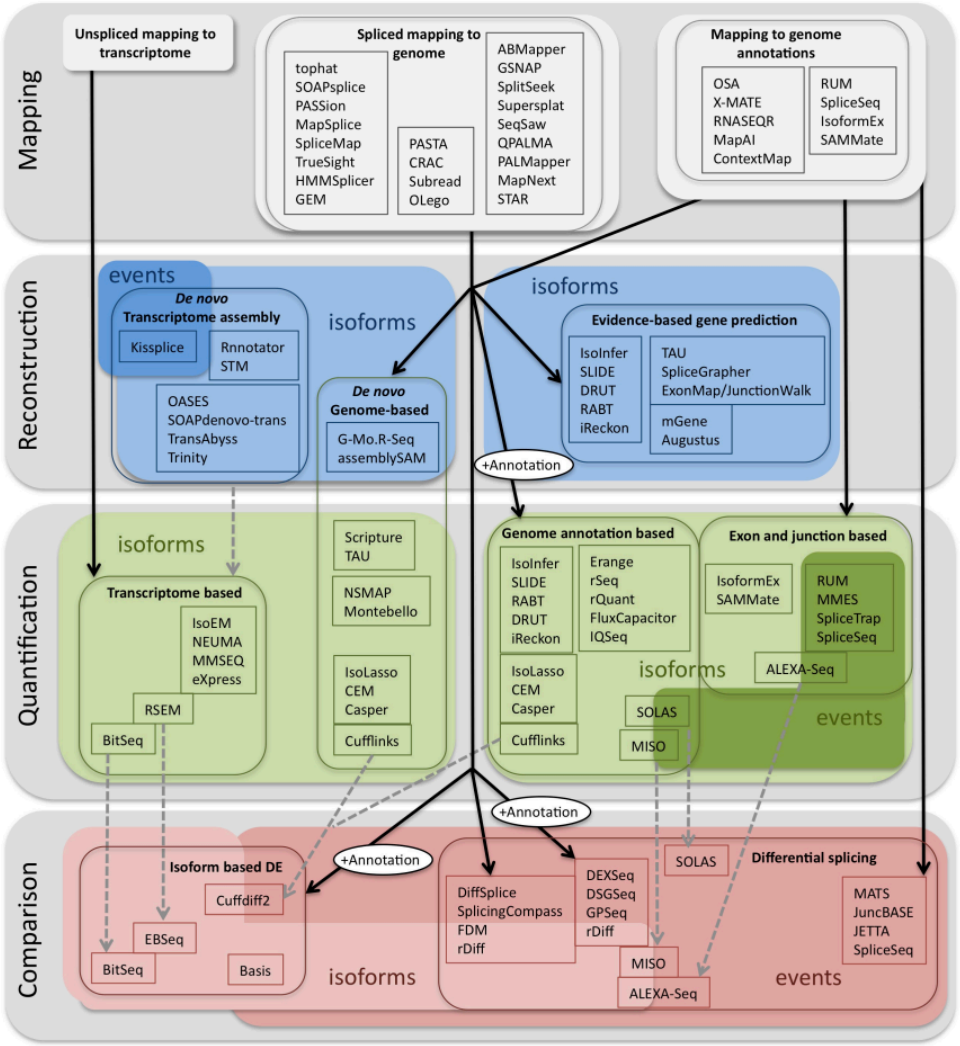


For experiments with <12 replicates per condition; use *edgeR (exact)*.

For experiments with >12 replicates per condition; use *DESeq*.

Schurch, NJ., et al. (2016), RNA,

Methods to study splicing from RNA-seq data



Clifton, NJ., et al. (2014), Method in molecular biology, 1126, 357-97.

Multiple other applications

- Allele specific expression
- RNA editing
- SNP analysis
- Small RNA profiling …

How do I choose the right tool ?

- Understand each tool's requirements
  - e.g. MMSEQ requires alignment onto transcriptome
- Identify how tools behave differently and pick one accordingly
  - e.g. Cufflinks (Mapping-first) vs. ABySS (Assembly-first)
- Pick commonly used tools
  - cause there's online help
- ..or tools implemented by someone in the lab/institute
  - cause you can always poke him when it doesn't work

Quantitative analyses using RNA-seq data

Luigi Grassi < *lg*490@*medschl.cam.ac.uk* >
Guillermo Parada < *gp*7@*sanger.ac.uk* >

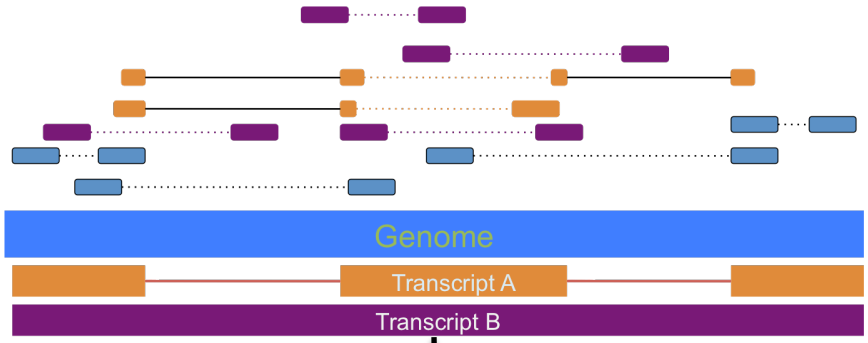
Classic quantification of gene expression using RNA-seq

Mapping

Alignment to genome

-Hisat2

-STAR



Counts reads per transcript

Normalization

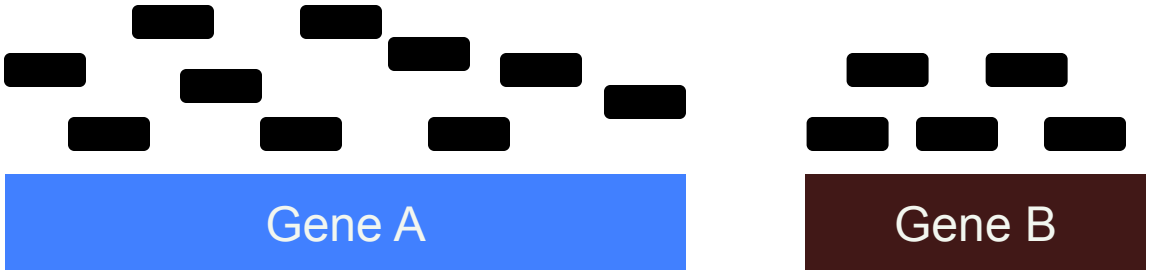
Read counts tables

FPKM

TPM

Normalised expression values

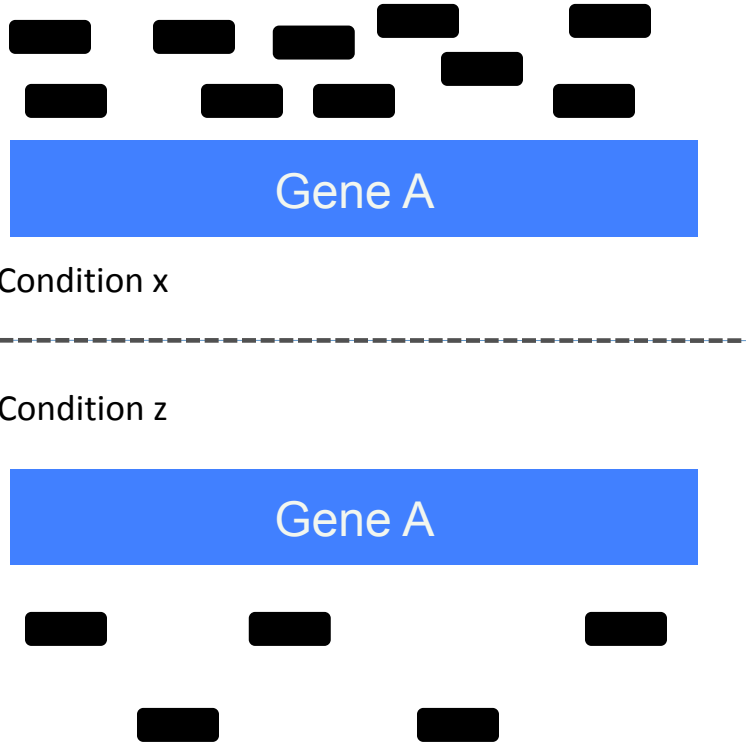
- For gene/isoform length



Gene	Raw reads	Length	Normalised Reads
A	10	2	5
B	5	1	5

Normalised expression values

- For total number of mapped reads



Condition	Raw reads	Total mapped reads	Normalised Reads
x	10	1000	0.01
z	5	500	0.01

FPKM (Fragment Per Kilobase Million)

I STEP: normalize by depth

GENE	REP1	REP2	REP3
A1 (2kb)	10	12	30
A2 (4kb)	20	25	60
A3 (1kb)	5	8	15
A4 (10kb)	0	0	1

FPKM (RPKM)

I STEP: normalize by depth

GENE	REP1	REP2	REP3
A1 (2kb)	10	12	30
A2 (4kb)	20	25	60
A3 (1kb)	5	8	15
A4 (10kb)	0	0	1
Sum all the counts	35	45	106
Scale by 1M (10)	3.5	4.5	10.6

FPKM (RPKM)

II STEP: divide counts by scaling factor

SCALING FACTOR 3.5 4.5 10.6

GENE	REP1	REP2	REP3
A1 (2kb)	2.86	2.67	2.83
A2 (4kb)	5.71	5.56	5.66
A3 (1kb)	1.43	1.78	1.43
A4 (10kb)	0	0	0.09

COUNTS -> FPM

FPKM (RPKM)

III STEP: divide counts by length (kb)

GENE	REP1	REP2	REP3
A1 (2kb)	1.43	1.33	1.42
A2 (4kb)	1.43	1.39	1.42
A3 (1kb)	1.43	1.78	1.42
A4 (10kb)	0	0	0.009

FPM -> FPKM

TPM (Transcripts Per Million)

TPM is similar to FPKM and RPKM but it is calculated in a different order

GENE	REP1	REP2	REP3
A1 (2kb)	10	12	30
A2 (4kb)	20	25	60
A3 (1kb)	5	8	15
A4 (10kb)	0	0	1

TPM (Transcripts Per Million)

I STEP: normalize by gene length

GENE	REP1	REP2	REP3
A1 (2kb)	5	6	15
A2 (4kb)	5	6.25	15
A3 (1kb)	5	8	15
A4 (10kb)	0	0	0.1

COUNTS -> FPK

TPM (Transcripts Per Million)

II STEP: normalize by sequencing depth

GENE	REP1	REP2	REP3
A1 (2kb)	5	6	15
A2 (4kb)	5	6.25	15
A3 (1kb)	5	8	15
A4 (10kb)	0	0	0.1
Sum all the FPKs	15	20.25	45.1
Scale by 1M (10)	1.5	2.025	4.51

TPM (Transcripts Per Million)

II STEP: normalize by sequencing depth

GENE	REP1	REP2	REP3
A1 (2kb)	3.33	2.96	3.326
A2 (4kb)	3.33	3.09	3.326
A3 (1kb)	3.33	3.95	3.326
A4 (10kb)	0	0	0.02

FPK -> TPM

FPKM VS TPM

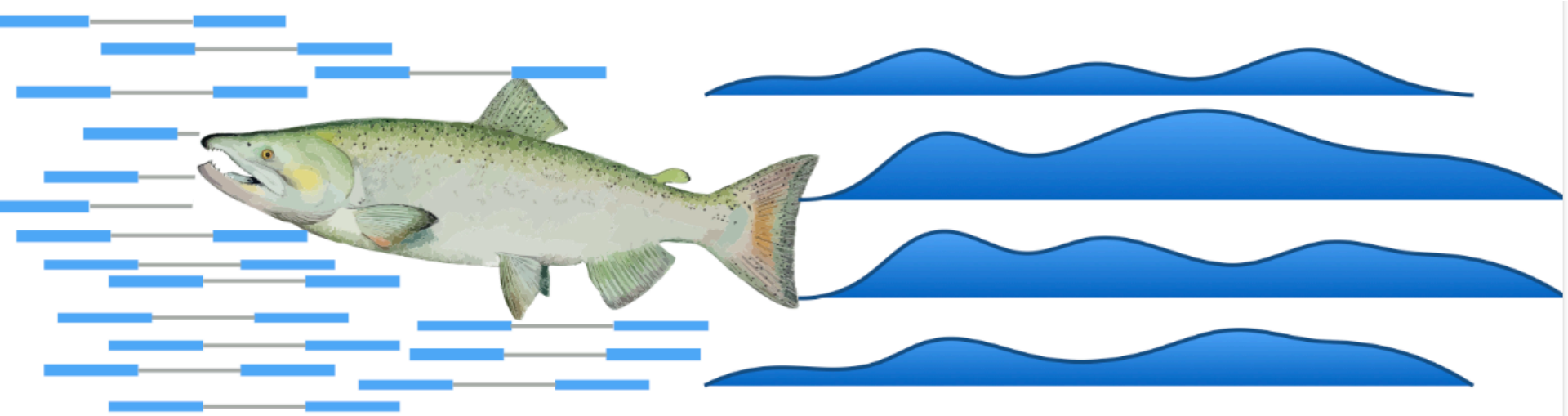
FPKM

GENE	REP1	REP2	REP3
A1 (2kb)	1.43	1.33	1.42
A2 (4kb)	1.43	1.39	1.42
A3 (1kb)	1.43	1.78	1.42
A4 (10kb)	0	0	0.009
	4.29	4.5	4.25

TPM

GENE	REP1	REP2	REP3
A1 (2kb)	3.33	2.96	3.326
A2 (4kb)	3.33	3.09	3.326
A3 (1kb)	3.33	3.95	3.326
A4 (10kb)	0	0	0.02
	10	10	10

Defying the paradigm of transcript quantification



Salmon —*Don't count . . . quantify!*

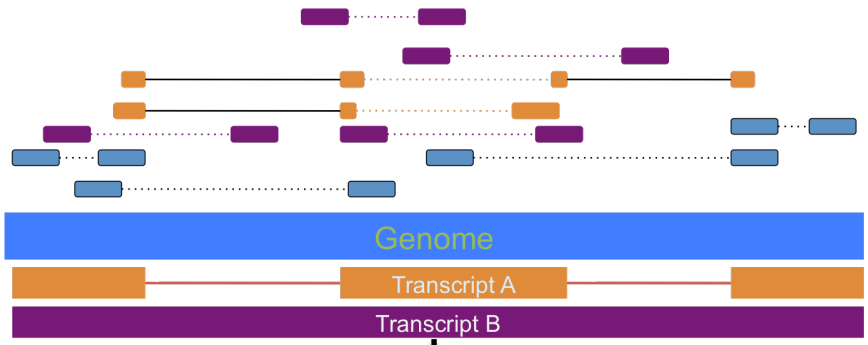
Classic quantification of gene expression using RNA-seq

Mapping

Alignment to genome

-Hisat2

-STAR



Salmon

Quasi-mapping to transcriptome

Counts reads per transcript

Normalization

Bias correction and Quantification

Read counts tables

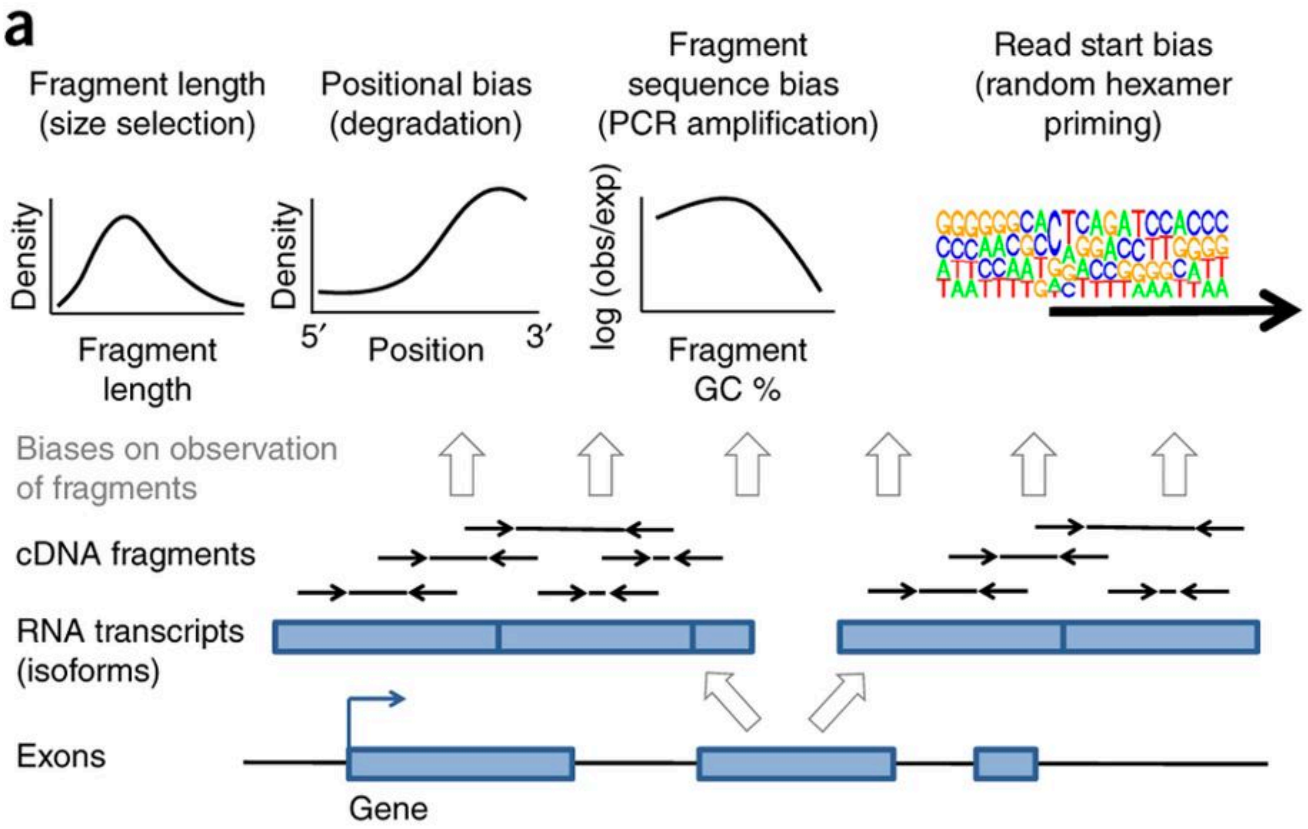
TPM

TPM

Quasi-mapping: Let speed up!

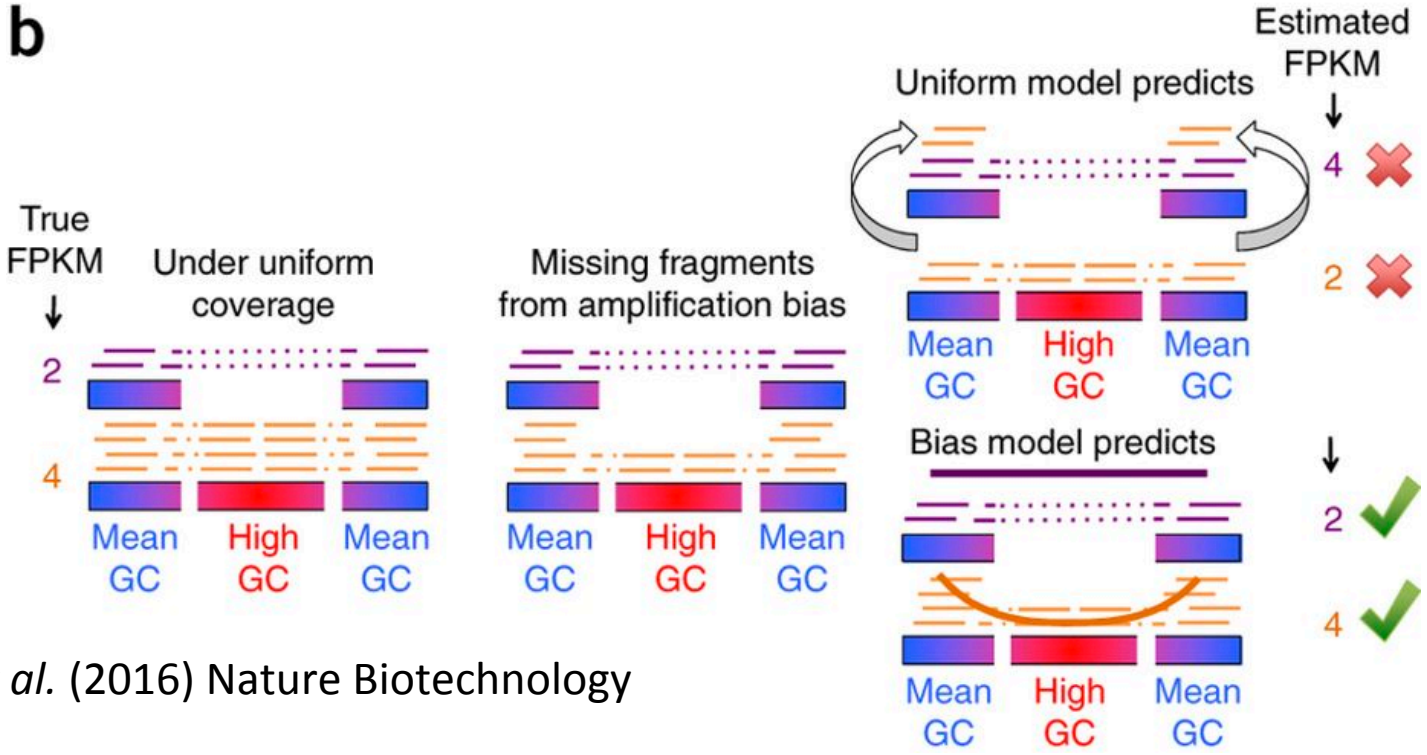
- In many cases all the information provided for the alignment is not necessary.
- **Base-to-base alignment** is **slow** and to quantify we just need to know the position where the reads map.
- Quasi-mapping (**RapMap**)
  - **Faster**!!!
  - Produces mapping that meet or exceed the accuracy of existing popular aligners

RNA-seq biases



Love *et al.* (2016) Nature Biotechnology

Salmon: Accounting for fragment sequence bias

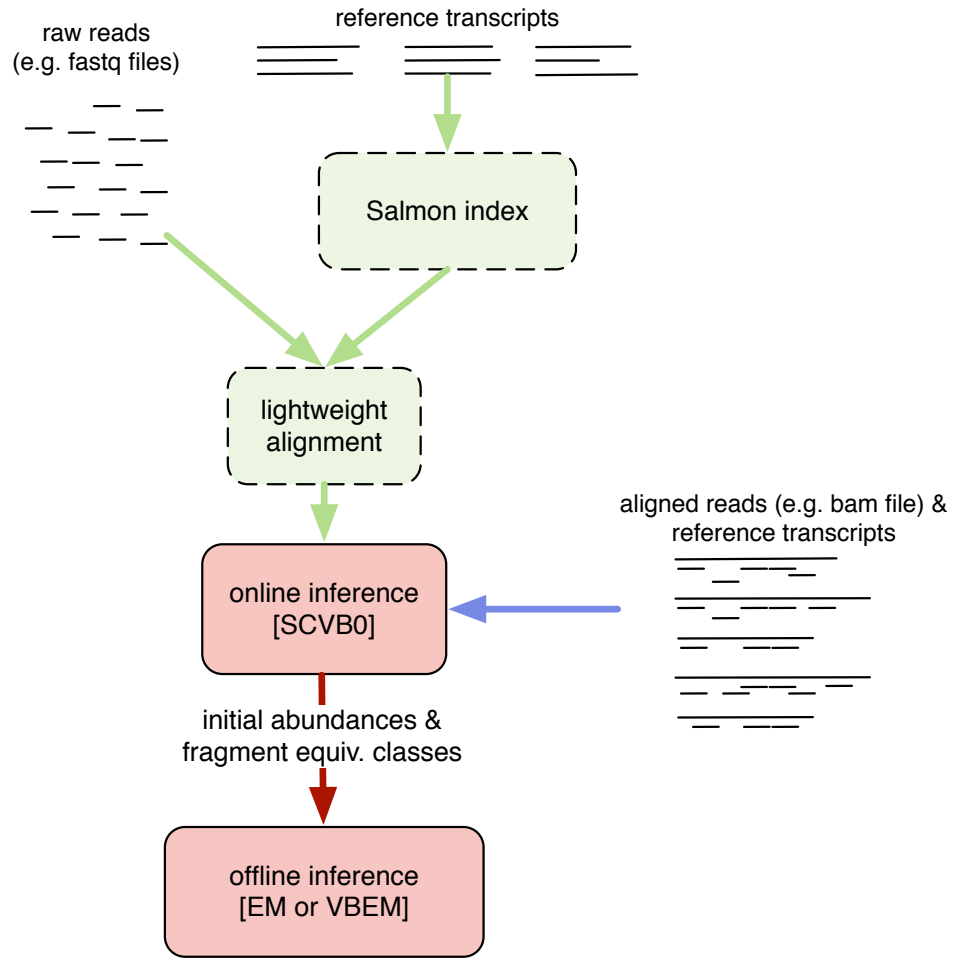


Love *et al.* (2016) Nature Biotechnology

[Salmon] "It is the first transcriptome-wide quantifier to correct for fragment GC-content bias"

Patro *et al.* (2017) Nature Methods

Supplementary Figure 1: Overview of Salmon's method and components.



Supplementary Figure 1: Overview of Salmon's method and components. Salmon excepts either raw (green arrows) or aligned reads (blue arrow) as input, performs an online inference when processing fragments or alignments, builds equivalence classes over these fragments and subsequently refines abundance estimates using an offline inference algorithm on a reduced representation of the data.