Introduction to Transcriptomics Analysis

Class 09 - Expression Quantification

INSTRUCTOR:
Aureliano Bombarely
Department of Bioscience
Universita degli Studi di Milano
aureliano.bombarely@unimi.it

Outline of Topics

1. Transcript assignments and discovery.

    1.1. Genome mapping approaches.

    1.2. Transcriptome mapping approaches.

    1.3. Reference free transcript assembly

2. Expression quantification

    2.1. Mapped read raw count approaches.

    2.2. Methods of normalisation

    2.3. Free mapping approaches

Outline of Topics

1. Transcript assignments and discovery.

    1.1. Genome mapping approaches.

    1.2. Transcriptome mapping approaches.

    1.3. Reference free transcript assembly

2. Expression quantification

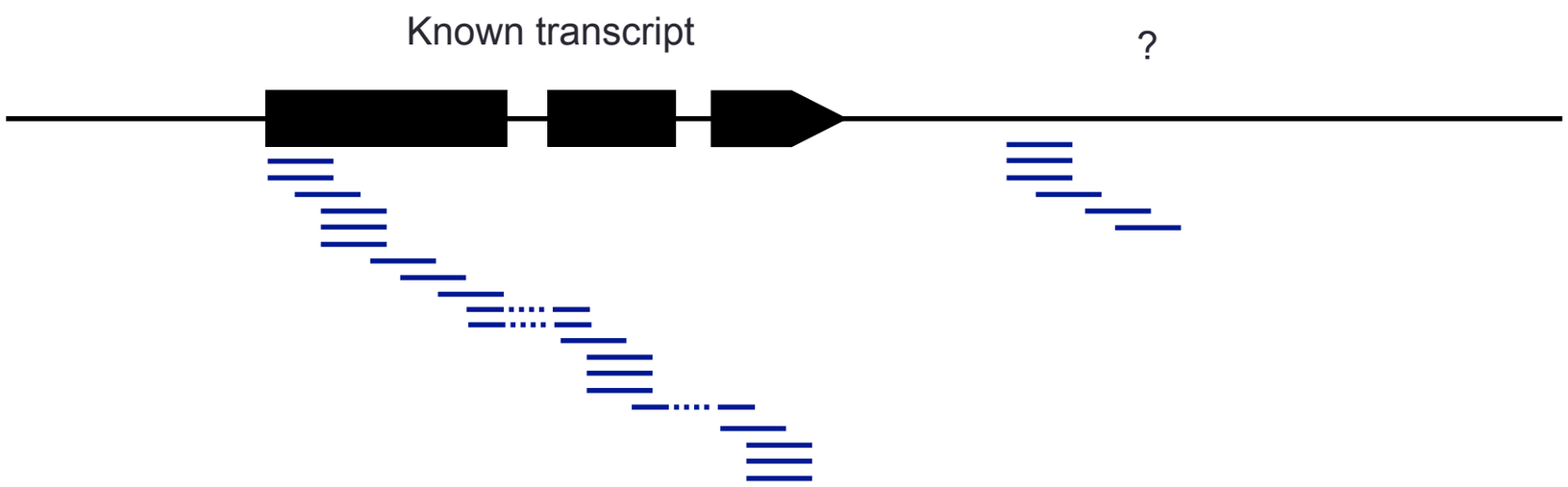
    2.1. Mapped read raw count approaches.

    2.2. Methods of normalisation

    2.3. Free mapping approaches

1. Transcript assignments and discovery.

Once the reads have been mapped to a reference, the next step is to identify the reads that are indeed from transcripts.



Usually only genome mapping approaches need to identify new transcripts. Transcriptome mapping and reference free transcript assembly assume that reads map to transcripts

1. Transcript assignments and discovery.

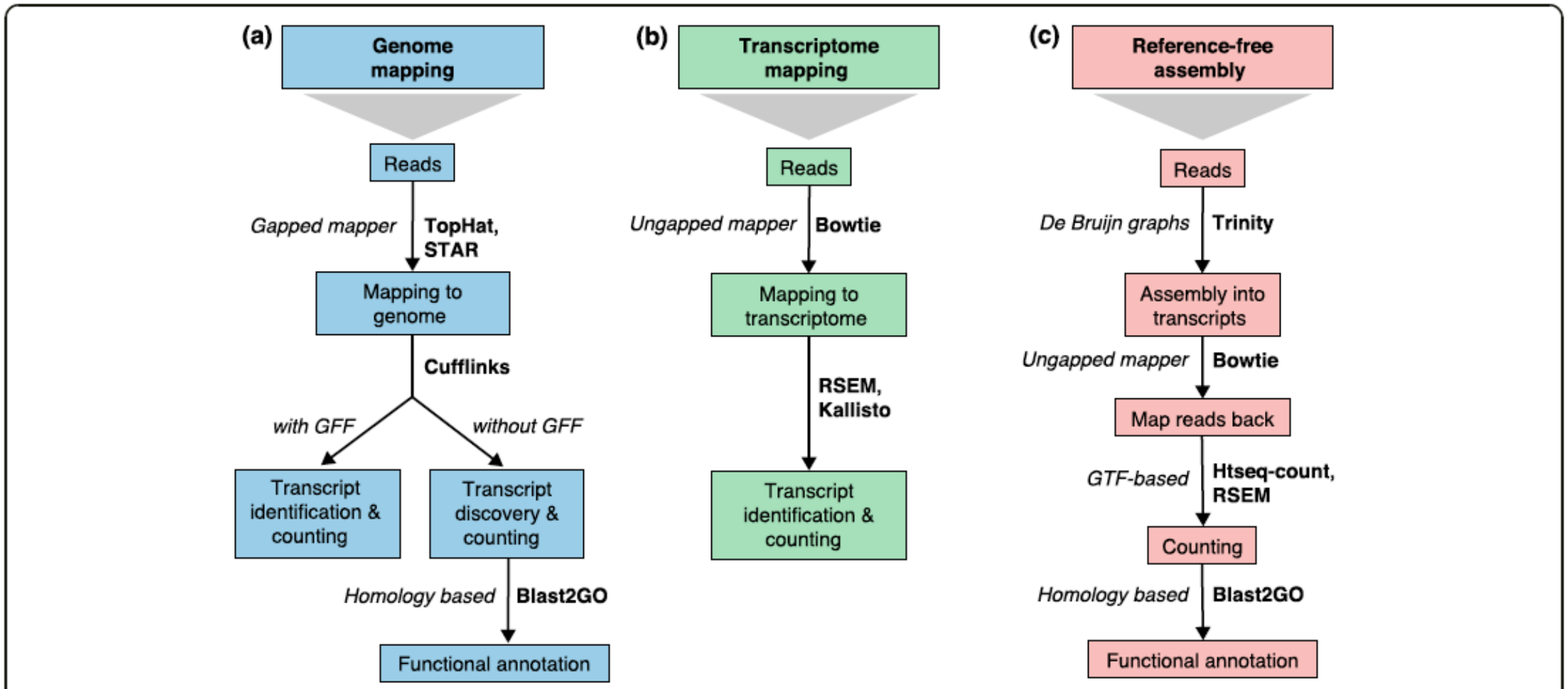


Fig. 2 Read mapping and transcript identification strategies. Three basic strategies for regular RNA-seq analysis. **a** An annotated genome is available and reads are mapped to the genome with a gapped mapper. Next (novel) transcript discovery and quantification can proceed with or without an annotation file. Novel transcripts are then functionally annotated. **b** If no novel transcript discovery is needed, reads can be mapped to the reference transcriptome using an ungapped aligner. Transcript identification and quantification can occur simultaneously. **c** When no genome is available, reads need to be assembled first into contigs or transcripts. For quantification, reads are mapped back to the novel reference transcriptome and further analysis proceeds as in (**b**) followed by the functional annotation of the novel transcripts as in (**a**). Representative software that can be used at each analysis step are indicated in *bold text*. Abbreviations: *GFF* General Feature Format, *GTF* gene transfer format, *RSEM* RNA-Seq by Expectation Maximization

Conesa, Ana, et al. "A survey of best practices for RNA-seq data analysis." *Genome biology* 17.1 (2016): 13.

Outline of Topics

1. Transcript assignments and discovery.

    1.1. Genome mapping approaches.

    1.2. Transcriptome mapping approaches.

    1.3. Reference free transcript assembly

2. Expression quantification

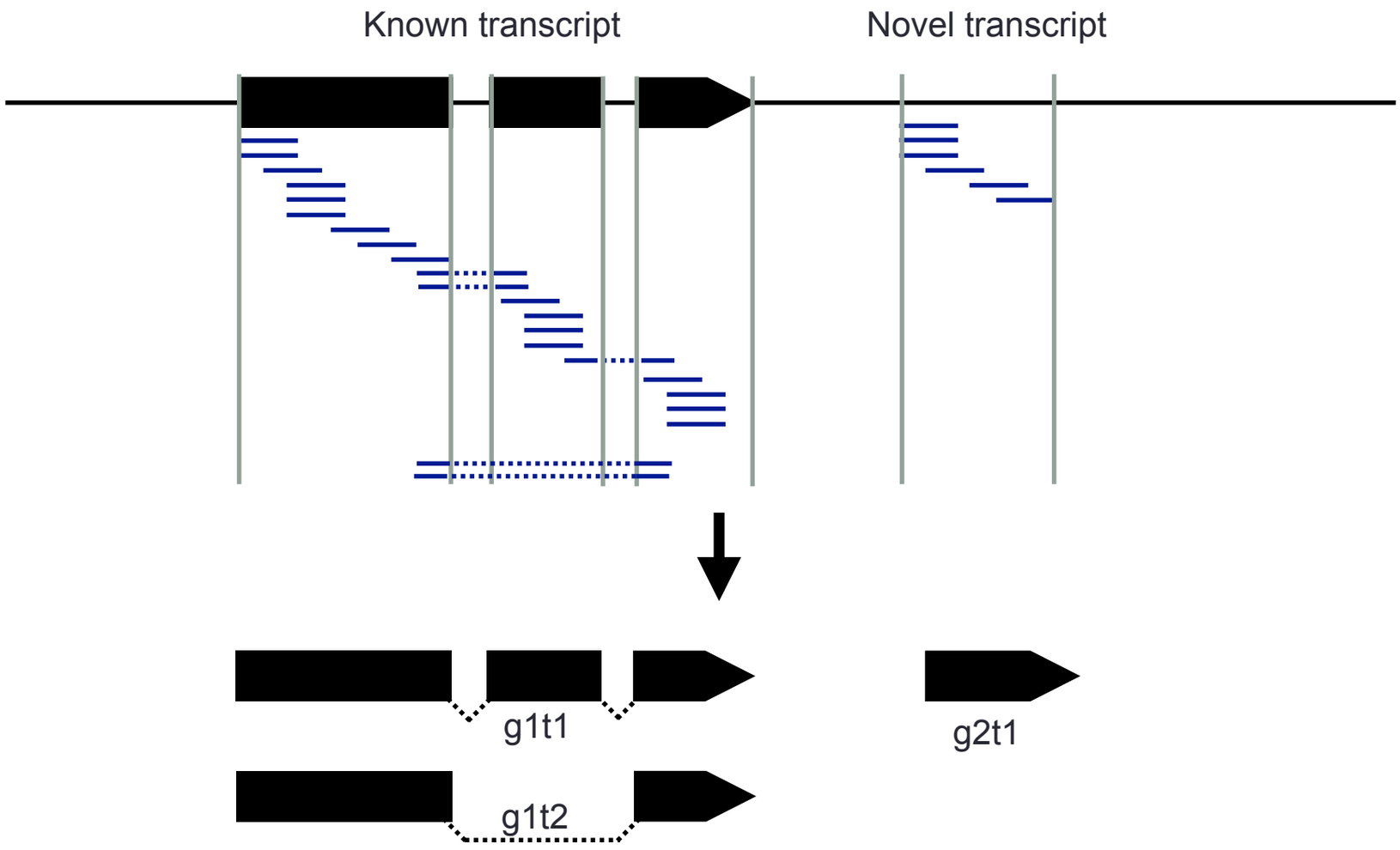
    2.1. Mapped read raw count approaches.

    2.2. Methods of normalisation

    2.3. Free mapping approaches

1.1. Genome mapping approaches.

Genome mapping approaches are based in the read mapping to a reference genome (DNA), comparison with the previous annotation (if they are available) and discovery of new transcripts if they have not been annotated.



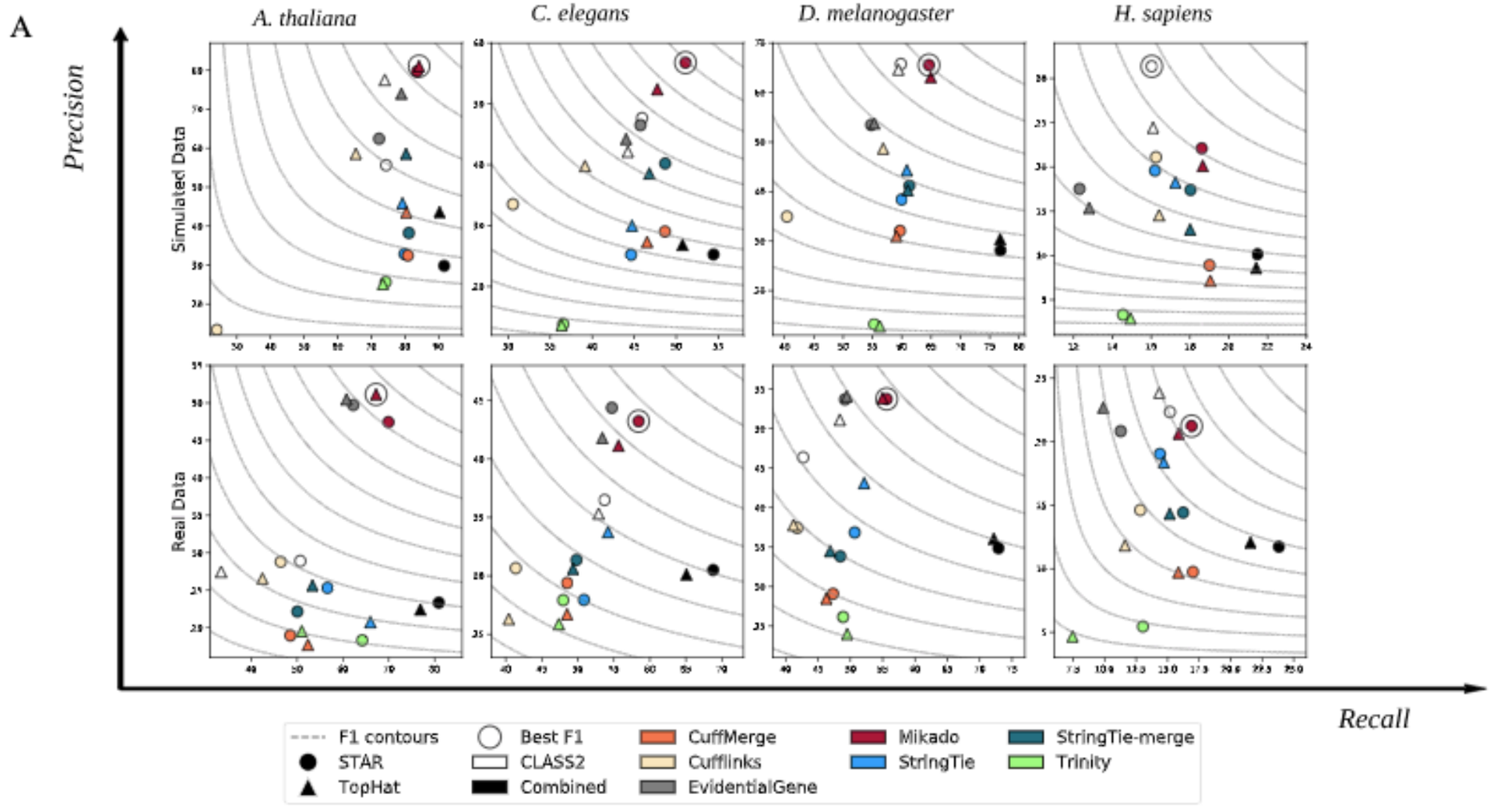
1.1. Genome mapping approaches.

Genome mapping approaches are based in the read mapping to a reference genome (DNA), comparison with the previous annotation (if they are available) and discovery of new transcripts if they have not been annotated.

Name	Source	Input	Output
CLASS2	http://ccb.jhu.edu/people/florea/research/CLASS2/	BAM	GTF
Cufflinks	https://github.com/cole-trapnell-lab/cufflinks	BAM	GTF
Mikado	https://mikado.readthedocs.io/en/latest/	BAM	GTF
Stringtie	https://ccb.jhu.edu/software/stringtie/	BAM	GTF
Trinity (Ref.guided mode)	https://github.com/trinityrnaseq/trinityrnaseq	BAM	FASTA, GTF

1.1. Genome mapping approaches.

Genome mapping approaches are based in the read mapping to a reference genome (DNA), comparison with the previous annotation (if they are available) and discovery of new transcripts if they have not been annotated.



Venturini, Luca, et al. "Leveraging multiple transcriptome assembly methods for improved gene structure annotation." *GigaScience* 7.8 (2018): giy093.

Outline of Topics

1. Transcript assignments and discovery.

    1.1. Genome mapping approaches.

    1.2. Transcriptome mapping approaches.

    1.3. Reference free transcript assembly

2. Expression quantification

    2.1. Mapped read raw count approaches.

    2.2. Methods of normalisation

    2.3. Free mapping approaches

1.2. Transcriptome mapping approaches.

Transcriptome mapping approaches are based in the read mapping to a reference transcriptome (DNA) in which there is not intron-exon boundaries (ungapped alignment). Any short read untapped alignment could work for this case.

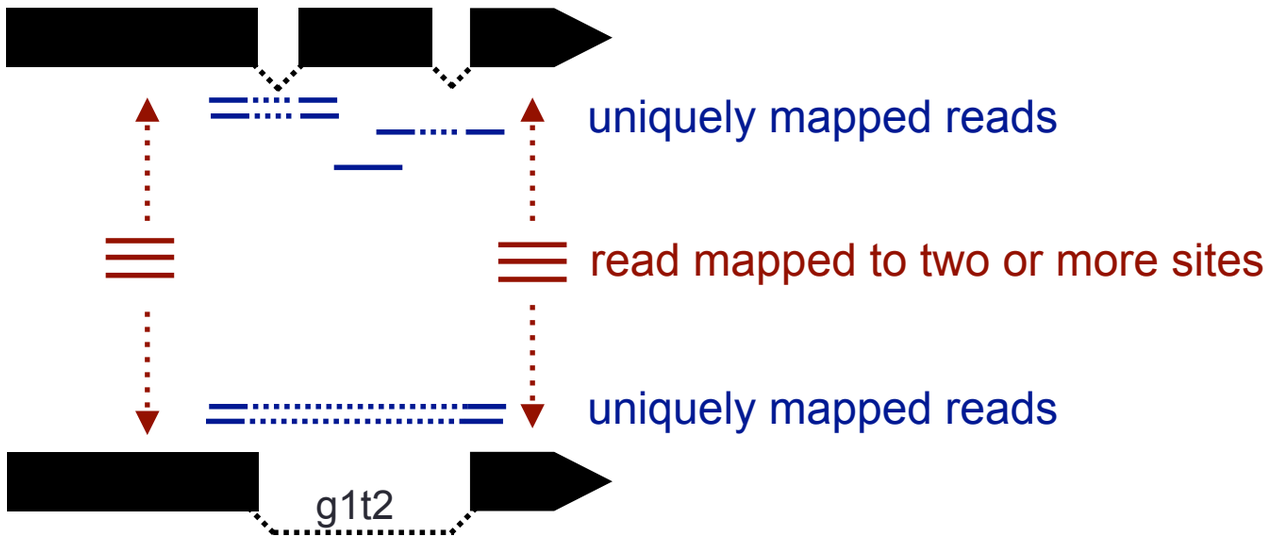
Name	Type	Input	Output	Pair ends
Bowtie/Bowtie2	Short sequences	Fasta, Fastq	Sam	Yes
BWA	Short sequences	Fasta, Fastq	Sam	Yes
GEM	Short sequences	Fasta, Fastq	Sam	Yes
Novoalign	Short sequences	Fasta, Fastq	Sam	Yes
SOAP	Short sequences	Fasta, Fastq	Sam	Yes
Stampy	Short sequences	Fasta, Fastq	Sam	Yes

1.2. Transcriptome mapping approaches.

Transcriptome mapping approaches are based in the read mapping to a reference transcriptome (DNA) in which there is not intron-exon boundaries (ungapped alignment). Any short read untapped alignment could work for this case.

↓

The use of a reference transcriptome involves to deal with possible alternative splicings.



1.2. Transcriptome mapping approaches.

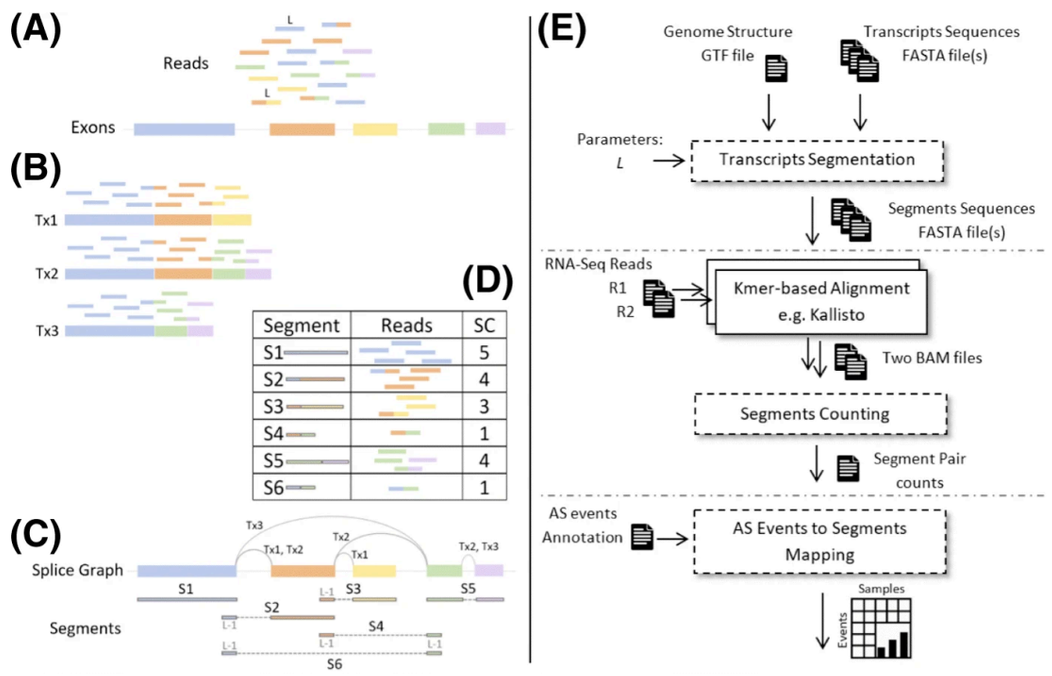
Transcriptome mapping approaches are based in the read mapping to a reference transcriptome (DNA) in which there is not intron-exon boundaries (ungapped alignment). Any short read untapped alignment could work for this case.

↓

The use of a reference transcriptome involves to deal with possible alternative splicings.

↓

New tools such as Yanagi



Gunady, Mohamed K., Stephen M. Mount, and Héctor Corrada Bravo. "Yanagi: Fast and interpretable segment-based alternative splicing and gene expression analysis." *BMC bioinformatics* 20.1 (2019): 1-19.

Outline of Topics

1. Transcript assignments and discovery.
    1.1. Genome mapping approaches.
    1.2. Transcriptome mapping approaches.
    1.3. Reference free transcript assembly
2. Expression quantification
    2.1. Mapped read raw count approaches.
    2.2. Methods of normalisation
    2.3. Free mapping approaches

1.3. Reference free transcript assembly

Reference free transcript assembly approaches involve the sequence assembly of the transcript. There are different tools for transcriptome assemblies.

Name	Source	Input	Output
BinPacker	https://github.com/macmanes-lab/BinPacker/	FASTA, FASTQ	FASTA
Bridger	https://github.com/fmaguire/Bridger_Assembler	FASTA, FASTQ	FASTA
Oases (Velvet)	https://www.ebi.ac.uk/~zerbino/oases/	FASTA, FASTQ	FASTA
SOAPdenovo-Trans	https://github.com/aquaskyline/SOAPdenovo-Trans	FASTA, FASTQ	FASTA
Trans-ABySS	https://github.com/bcgsc/transabyss	FASTA, FASTQ	FASTA
TransLiG	https://sourceforge.net/projects/transcriptomeassembly/files/	FASTA, FASTQ	FASTA
Trinity	https://github.com/trinityrnaseq/trinityrnaseq	FASTA, FASTQ	FASTA

↓

Trinity (most popular)

1.3. Reference free transcript assembly

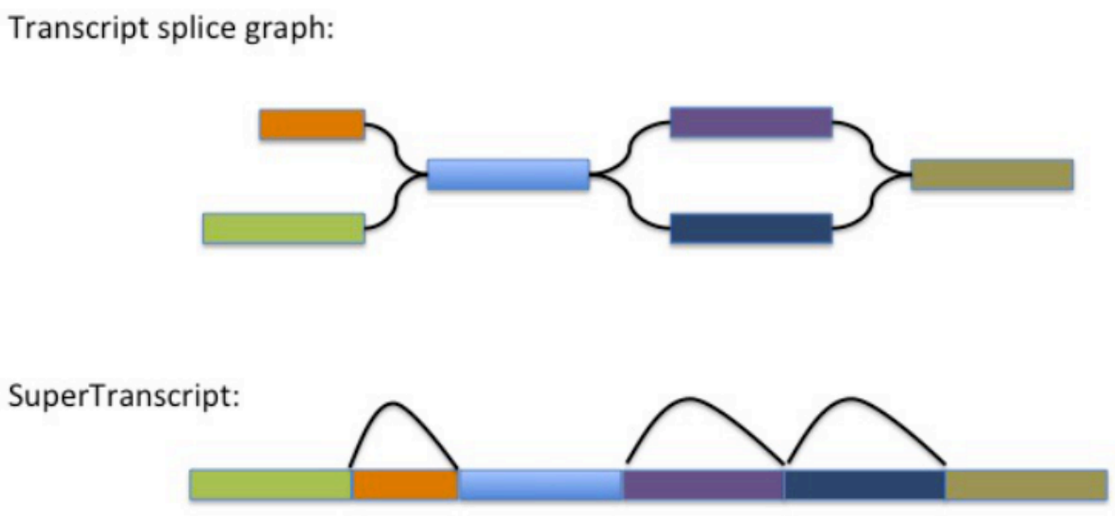
Reference free transcript assembly approaches involve the sequence assembly of the transcript. There are different tools for transcriptome assemblies.

↓

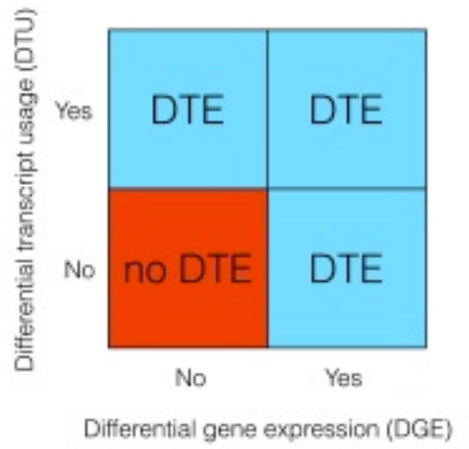
Once the transcriptome has been assembled, any transcriptome mapping approach can be used.

↓

Trinity pipeline propose the alternative use of Supertranscripts and Differential Transcript Usage (DTU)



https://github.com/trinityrnaseq/trinityrnaseq/wiki/SuperTranscripts



Soneson, Charlotte, Michael I. Love, and Mark D. Robinson. "Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences." *F1000Research* 4 (2015).

Outline of Topics

1. Transcript assignments and discovery.

    1.1. Genome mapping approaches.

    1.2. Transcriptome mapping approaches.

    1.3. Reference free transcript assembly

2. Expression quantification

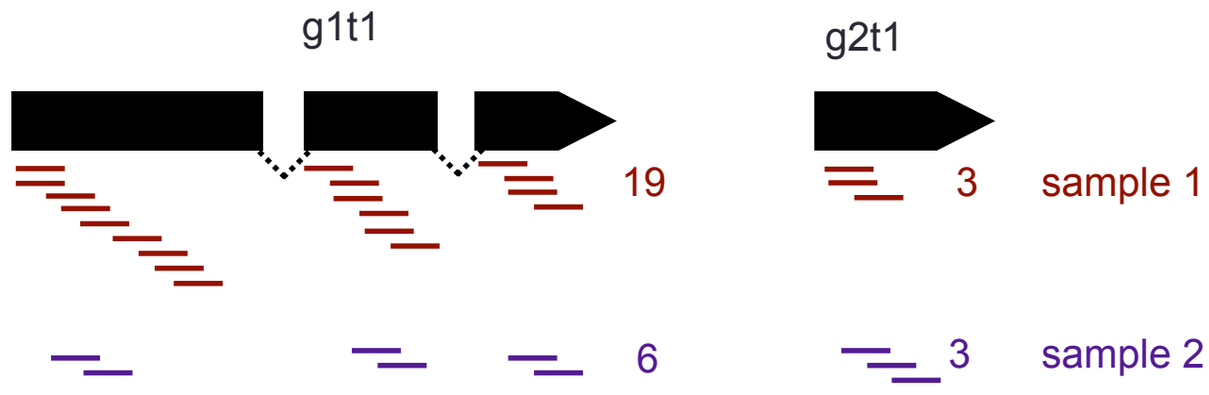
    2.1. Mapped read raw count approaches.

    2.2. Methods of normalisation

    2.3. Free mapping approaches

2. Expression quantification

The quantification of the gene expression in a RNA-Seq experiment is associated to the number of reads linked to a transcripts. More reads that the transcripts has more expressed it will be.



Nevertheless there are some complications and different ways to resolve them. These methodologies can be divided in three groups:

- Mapped read raw counts (e.g. HTSeq-Count).
- Mapped read normalised counts (e.g. Cufflinks).
- Kmer counts from non-mapped reads (e.g. Salmon).

Outline of Topics

1. Transcript assignments and discovery.
    1.1. Genome mapping approaches.
    1.2. Transcriptome mapping approaches.
    1.3. Reference free transcript assembly
2. Expression quantification
    2.1. Mapped read raw count approaches.
    2.2. Methods of normalisation
    2.3. Free mapping approaches

2.1. Mapped read raw count approaches.

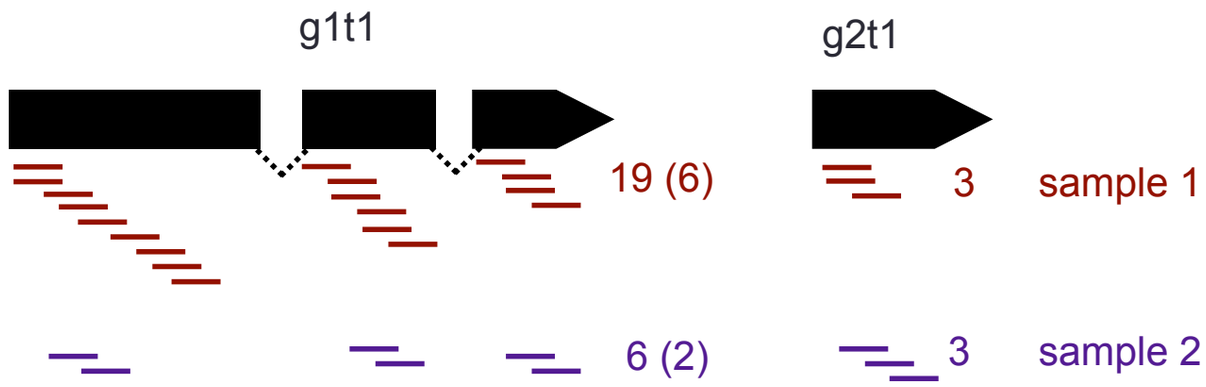
The most simple way to quantify the gene expression is to count reads.

The most popular tools to quantify gene expression with aggregate raw counts are:

- HTSeq-count
- featuresCount

Nevertheless the amount of short reads depends not only on the gene expression (***transcript abundance***) but also of the ***transcript length***. A longer transcript will produce more reads when it is fragmented. Additionally there are technical problems such as bias in the library preparation, mapping efficiency…

(both tools are more adequate for ChIP experiments)



Outline of Topics

1. Transcript assignments and discovery.

    1.1. Genome mapping approaches.

    1.2. Transcriptome mapping approaches.

    1.3. Reference free transcript assembly

2. Expression quantification

    2.1. Mapped read raw count approaches.

    2.2. Methods of normalisation

    2.3. Free mapping approaches

2.1. Methods of normalisation

Gene expression can not be measured as a simple read count comparison between different transcripts. Factors such as read length, sequencing depth… may influence the transcript abundance measures.

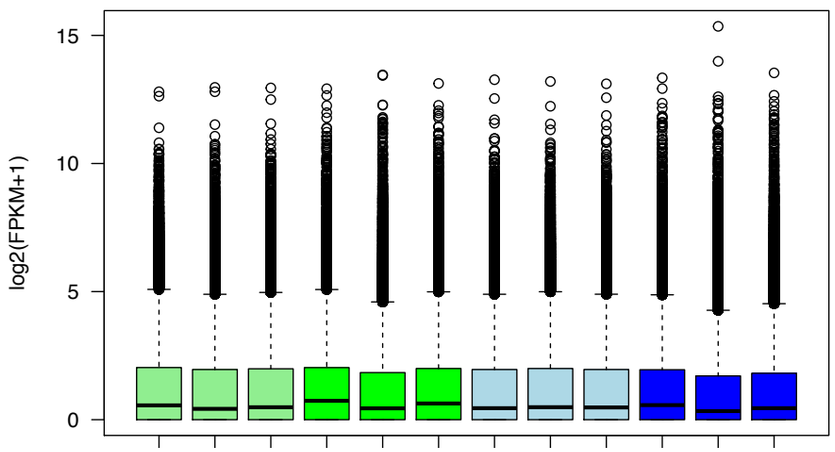
Most of the normalisation methods have some assumptions:

1. The majority of the transcripts are equally expressed in each experimental unit (e.g. the majority of expressed genes for control and stressed Arabidopsis leaves should have the same expression levels).

2. There is a symmetrical distribution of over and under expressed genes.

↓

Similar gene expression distributions



2.1. Methods of normalisation

Normalisation assumptions usually do not work under the following conditions:

1. Different tissues and different developmental stages usually have big differences in the transcriptome populations.
2. Over and under-expressed genes may do not have symmetrical patterns.
3. Small RNA molecules such as miRNA may do not have enough ocurrences to meet the statistical requirements.

Development of different methods of normalisation

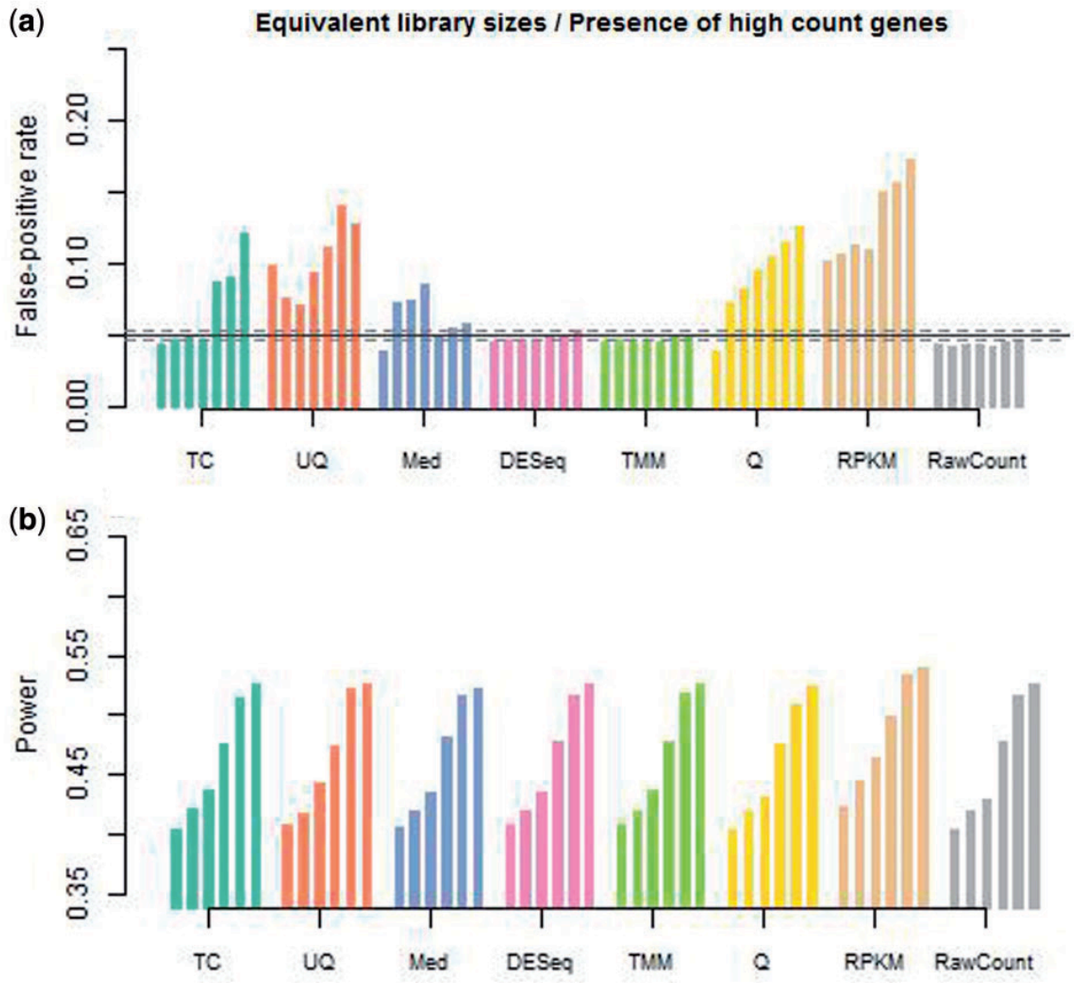
2.1. Methods of normalisation

The most common normalisation methods are:

Normalisation	Tools	Description	Library scalation	Transcript Length scalation	Sample Scaling factor
TC (Total Counts)	EDASeq	Gene count divided by the total number of reads	Yes	No	Yes
UQ (Upper Quartile)	EdgeR, Cufflinks, EDASeq	TC replaced by the upper quartile of counts	Yes	No	Yes
Med (Sample Median)	EDASeq	TC replaced by the median counts	Yes	No	Yes
DESeq	DESeq	Geometric mean and correction factor	Yes	No	Yes
TMM (Trimmed Mean of M values)	edgeR	Weighted mean of log ratios across samples	Yes	No	Yes
FQ (Full Quantile)	Aroma.light	Matching distribution across different samples	Yes	No	Yes
TPM (Transcripts per Million)	RSEM	Similar to RPKM	Yes	No	No
RPKM/FPKM (Reads/ Fragments Per Kilobase per Million Reads	Cufflinks, Stringtie	Introduce a bias for low expressed genes	Yes	Yes	No

2.1. Methods of normalisation

The most common normalisation methods are:



(A) Average false-positive rate over 10 independent datasets simulated with varying proportions of differentially expressed genes (from 0% to 30% for each normalization method). (B) Power over 10 independent datasets simulated with varying proportions of differentially expressed genes (from 5% to 30% for each normalization method).

DESeq and TMM are the most optimal normalisation methods

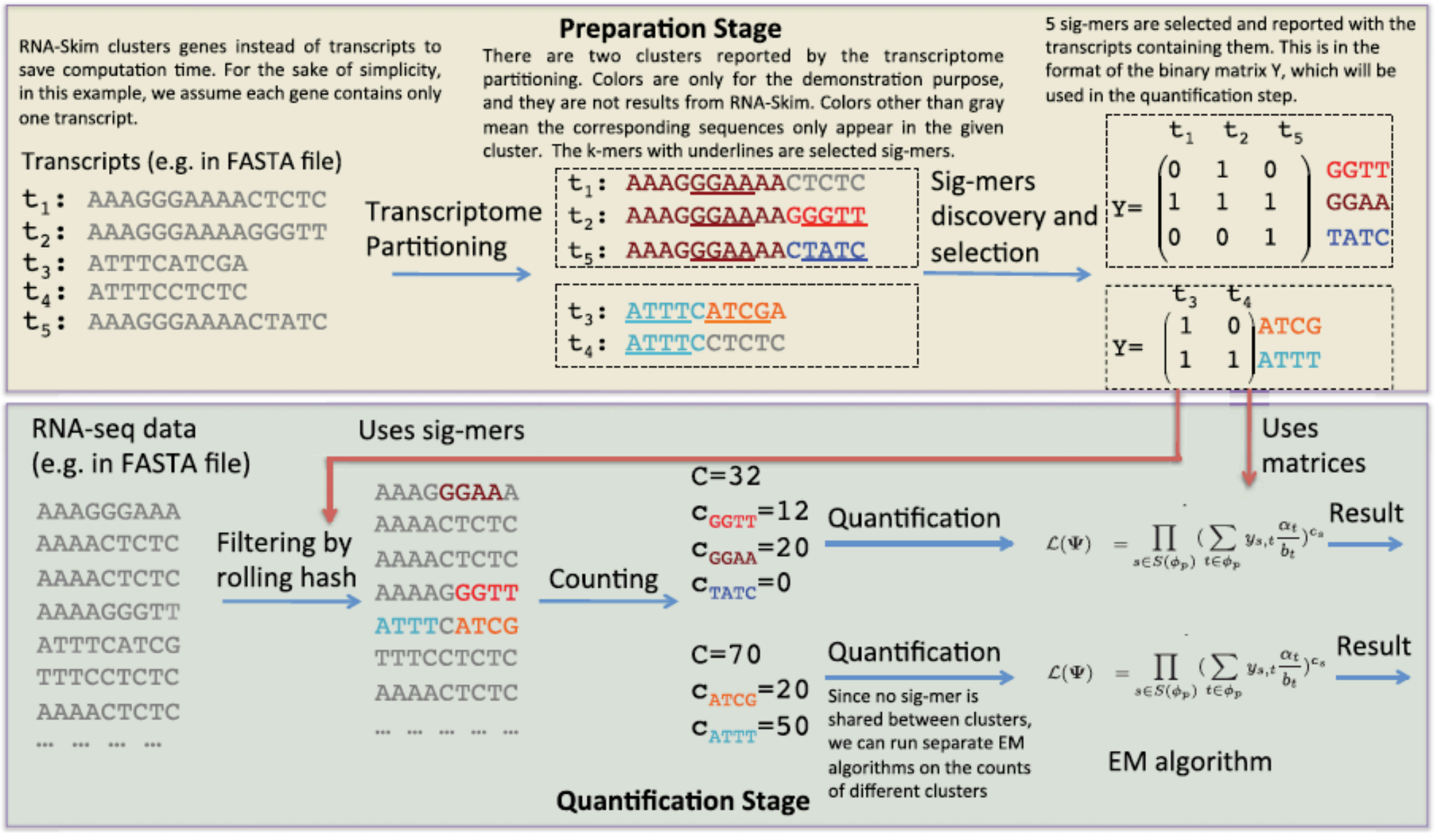
Dillies, Marie-Agnès, et al. "A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis." *Briefings in bioinformatics* 14.6 (2013): 671-683.

Outline of Topics

1. Transcript assignments and discovery.
    1.1. Genome mapping approaches.
    1.2. Transcriptome mapping approaches.
    1.3. Reference free transcript assembly
2. Expression quantification
    2.1. Mapped read raw count approaches.
    2.2. Methods of normalisation
    2.3. Free mapping approaches

2.3. Free mapping approaches

The free mapping quantification approaches are based in the Kmer counting algorithms.



Zhang, Zhaojun, and Wei Wang. "RNA-Skim: a rapid method for RNA-Seq quantification at transcript level." *Bioinformatics* 30.12 (2014): i283-i292.

2.3. Free mapping approaches

The most popular tools are:

Tool Name	Features	Source
Kallisto	Transcript abundance quantification from RNA-seq data (uses pseudoalignment for rapid determination of read compatibility with targets)	https://pachterlab.github.io/kallisto/
Sailfish	Estimation of isoform abundances from reference sequences and RNA-seq data (k-mer based)	http://www.cs.cmu.edu/~ckingsf/software/sailfish/
Salmon	Quantification of the expression of transcripts using RNA-seq data (uses k-mers)	https://combine-lab.github.io/salmon/
RNA-Skim	RNA-seq quantification at transcript-level (partitions the transcriptome into disjoint transcript clusters; uses sig-mers, a special type of k-mers)	http://www.csbio.unc.edu/rs/