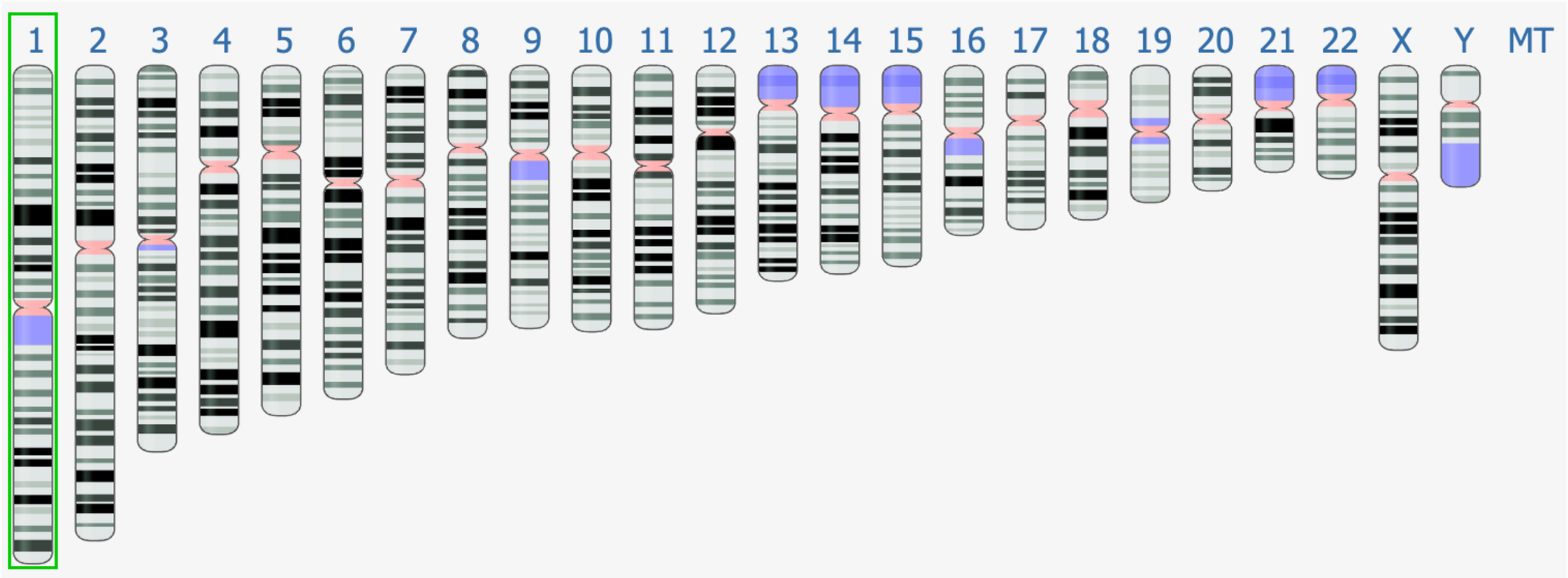


Genomics109

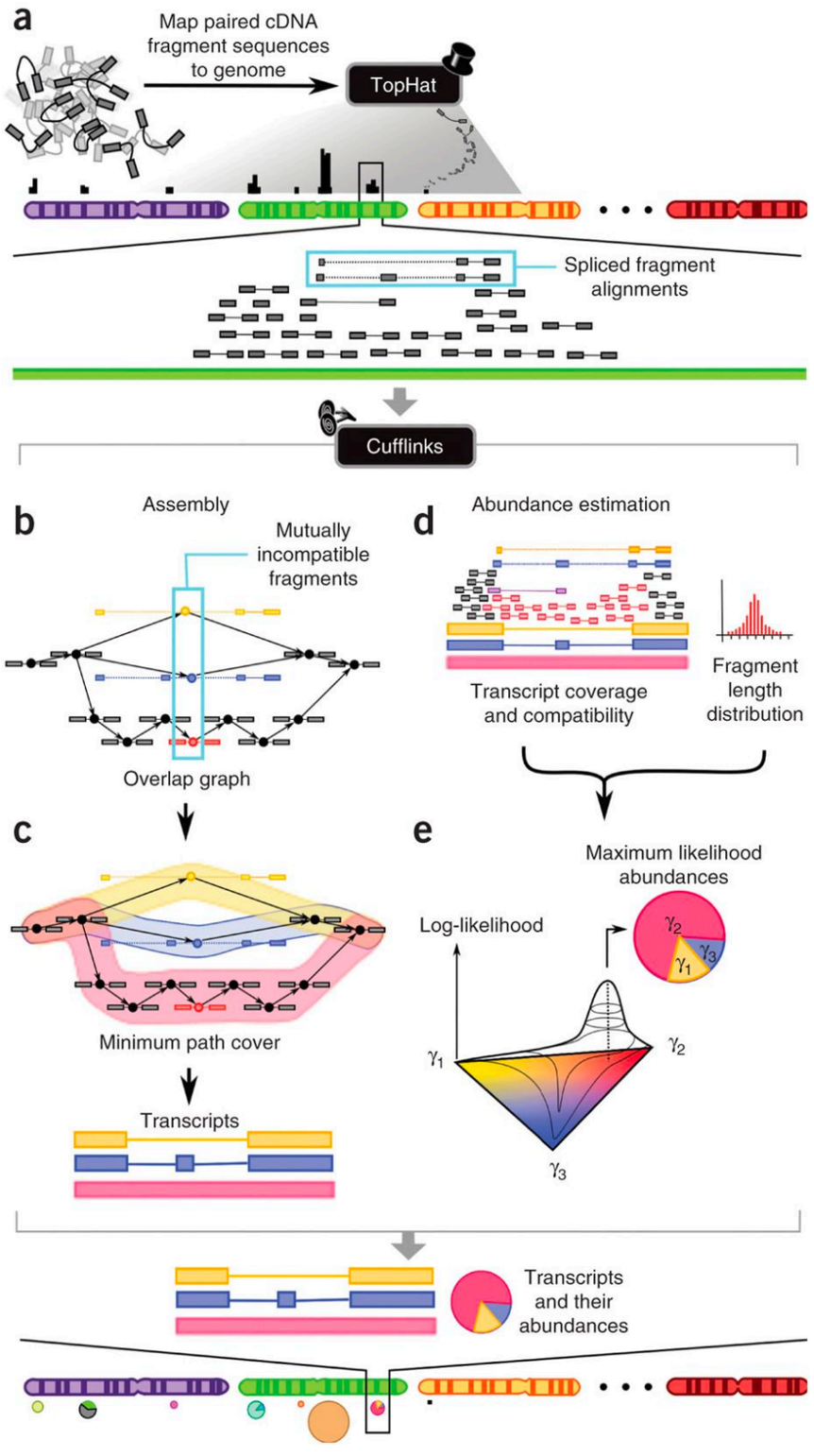
BIOL647
Digital Biology

Rodolfo Aramayo

Genomics109

Fundamentals of Transcriptome Mapping and Analysis

Outline of The Tuxedo Pipeline



Genomics109

Mapping RNASeq reads to the genome

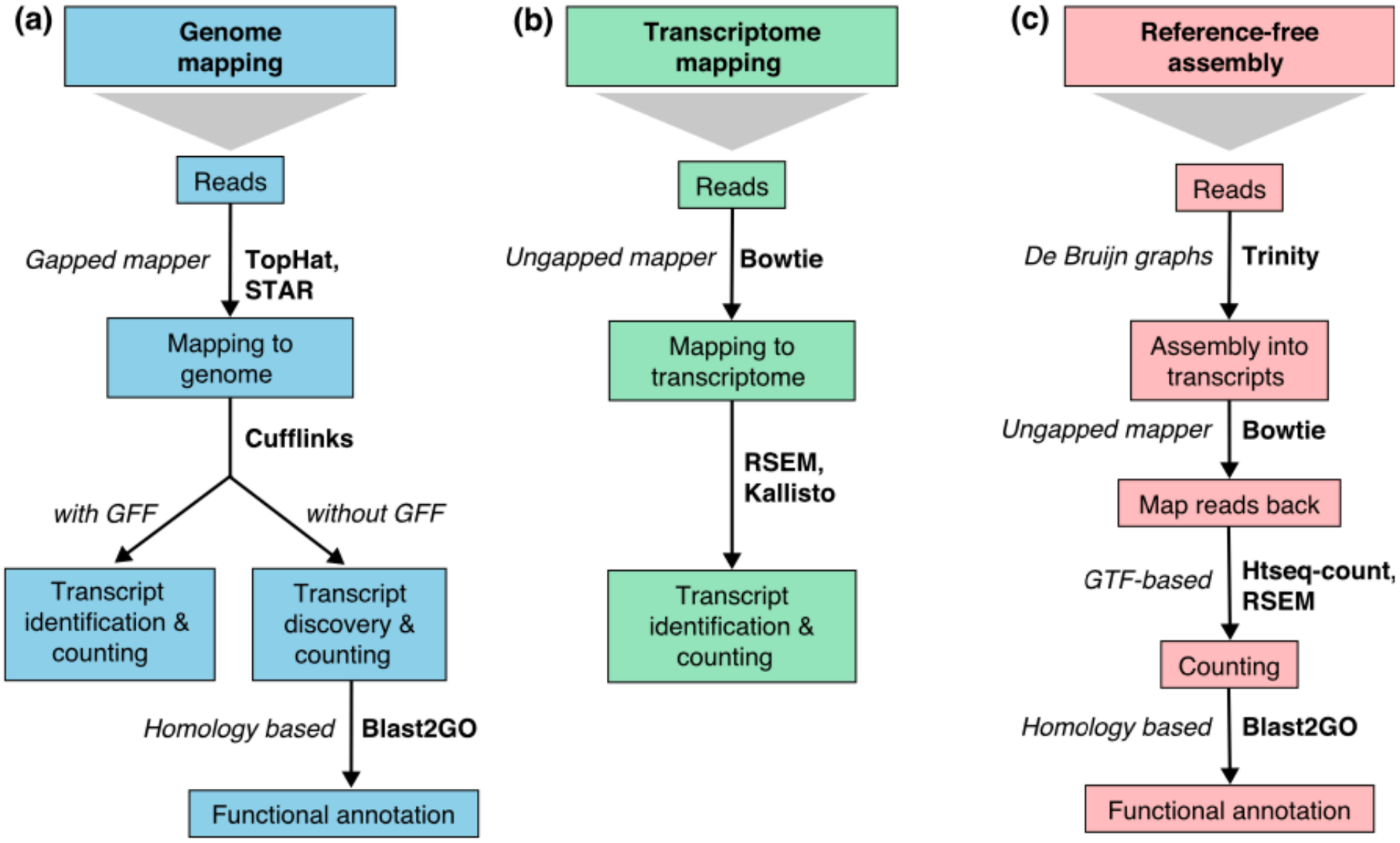
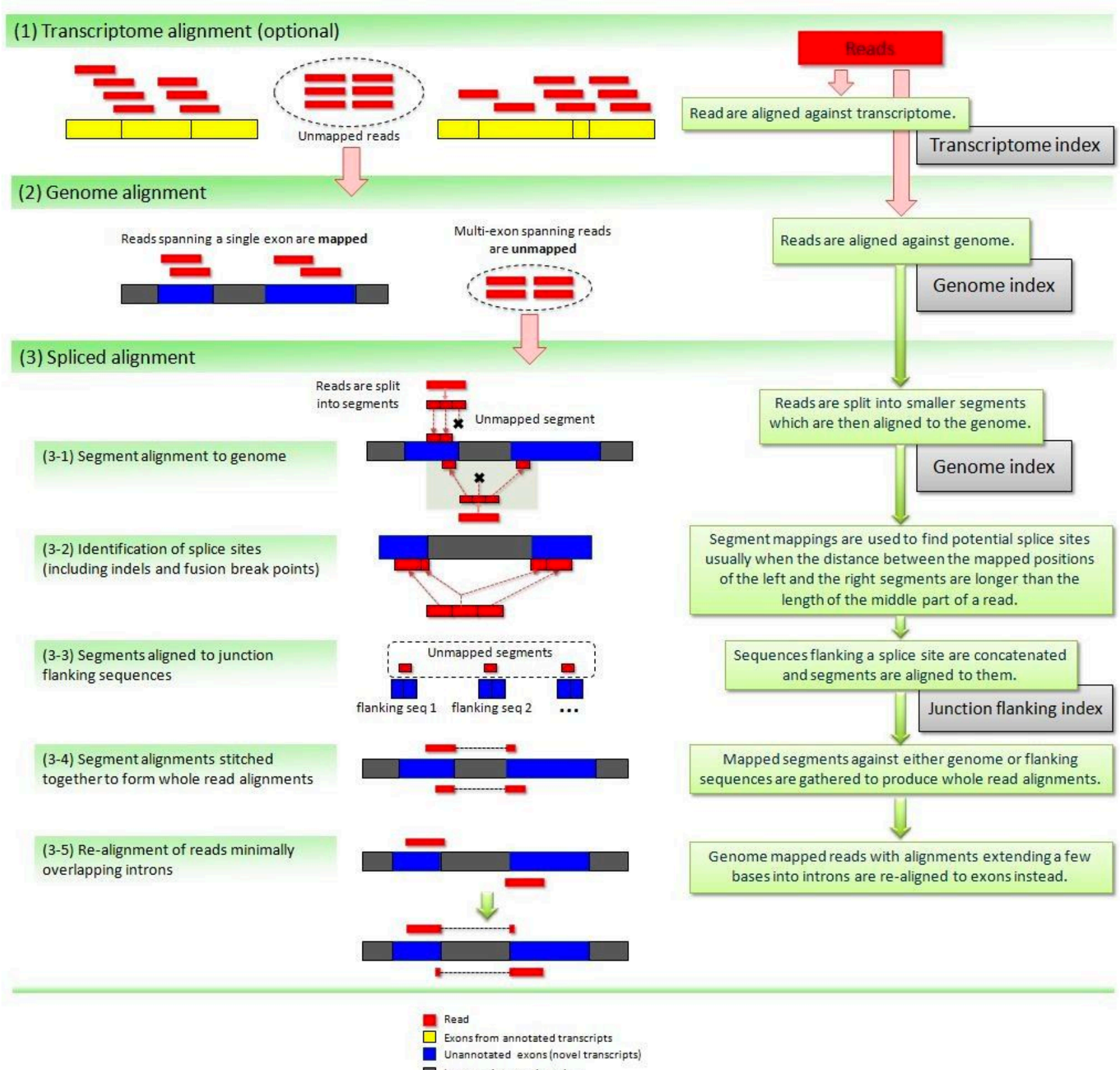


Fig. 2 Read mapping and transcript identification strategies. Three basic strategies for regular RNA-seq analysis. **a** An annotated genome is available and reads are mapped to the genome with a gapped mapper. Next (novel) transcript discovery and quantification can proceed with or without an annotation 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

Genomics109

Mapping RNASeq reads to the genome



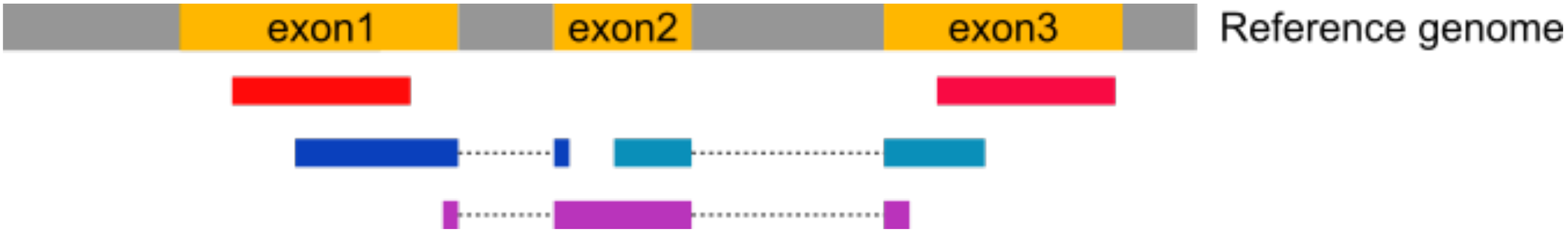
Genomics109

Mapping RNASeq reads to the genome

- To make sense of the data, we need to first figure out where the sequences originated from in the genome, so we can then determine to which genes they belong.

- When a reference genome for the organism is available, this process is known as aligning or “mapping” the reads to the reference. This is equivalent to solving a jigsaw puzzle, but unfortunately, not all pieces are unique.

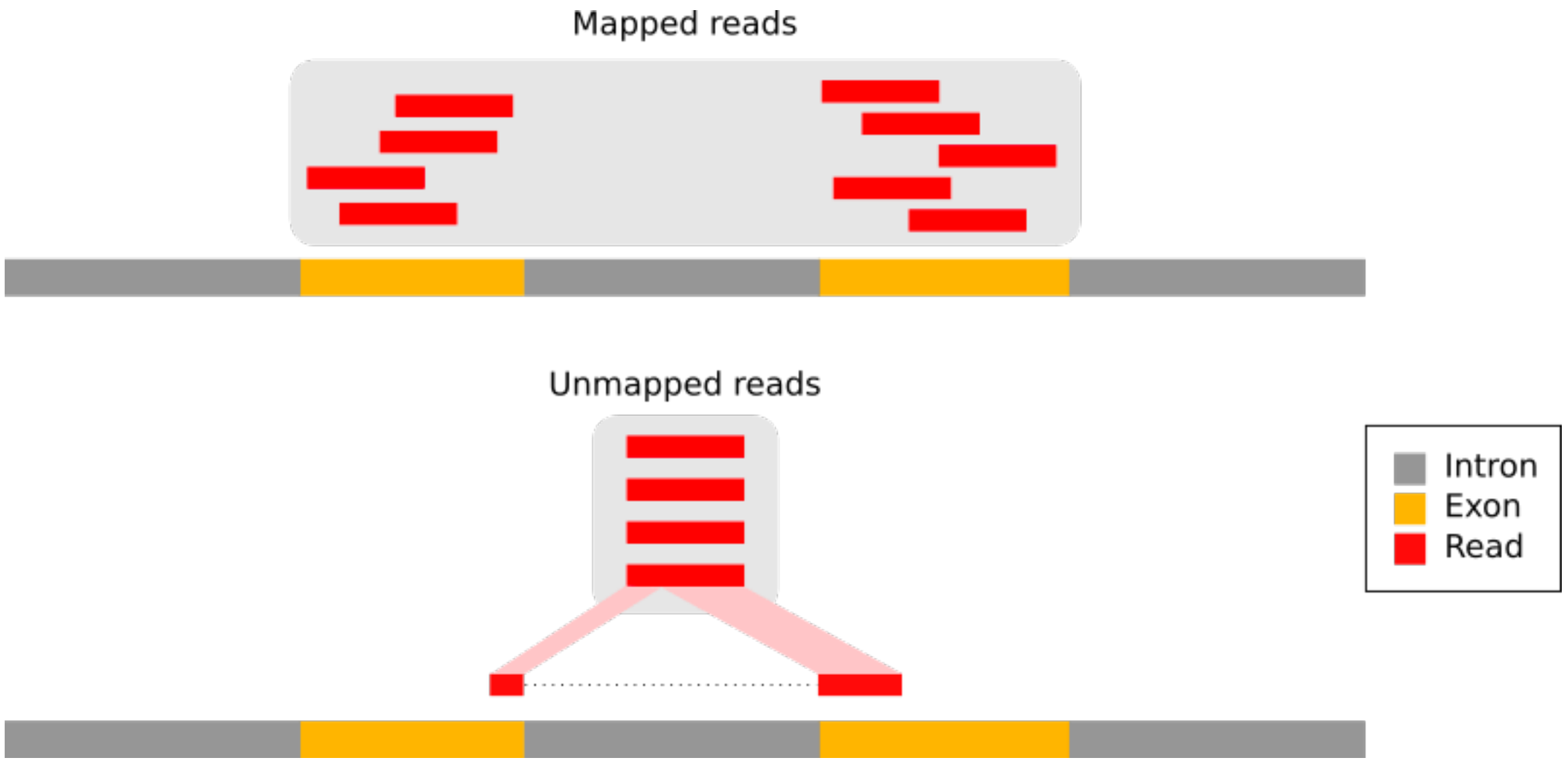
- With eukaryotic transcriptomes most reads originate from processed mRNAs lacking introns:



Genomics109

Mapping RNASeq reads to the genome

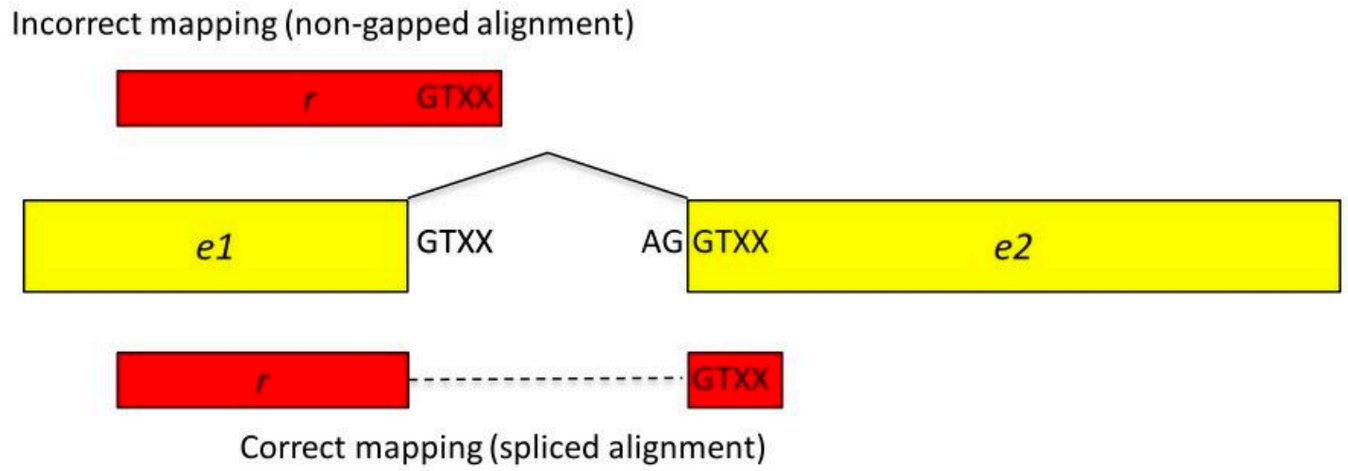
- **Therefore they cannot be simply mapped back to the genome as we normally do for DNA data. Spliced-awared mappers have been developed to efficiently map transcript-derived reads against a reference genome:**



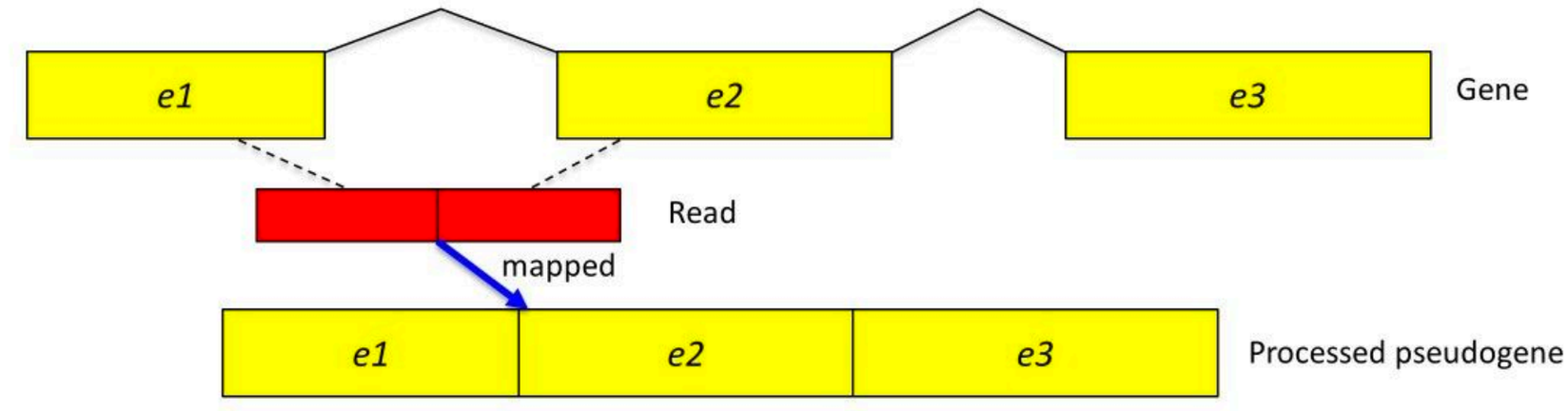
Genomics109

Mapping RNASeq reads to the genome

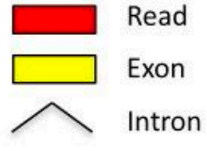
- **This process is called Split Alignment:**



(1) Read *r* may be incorrectly mapped to the intron between exons *e1* and *e2*.



(2) Here, the read shown in red, which spans a splice junction, can be aligned end-to-end to a processed pseudogene.



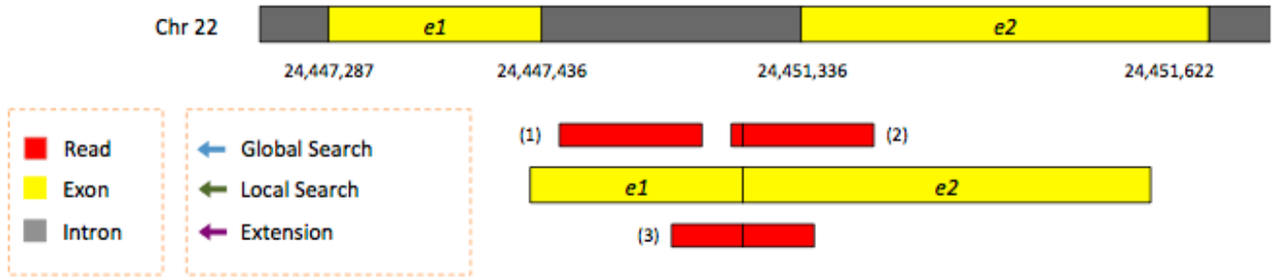
Genomics109

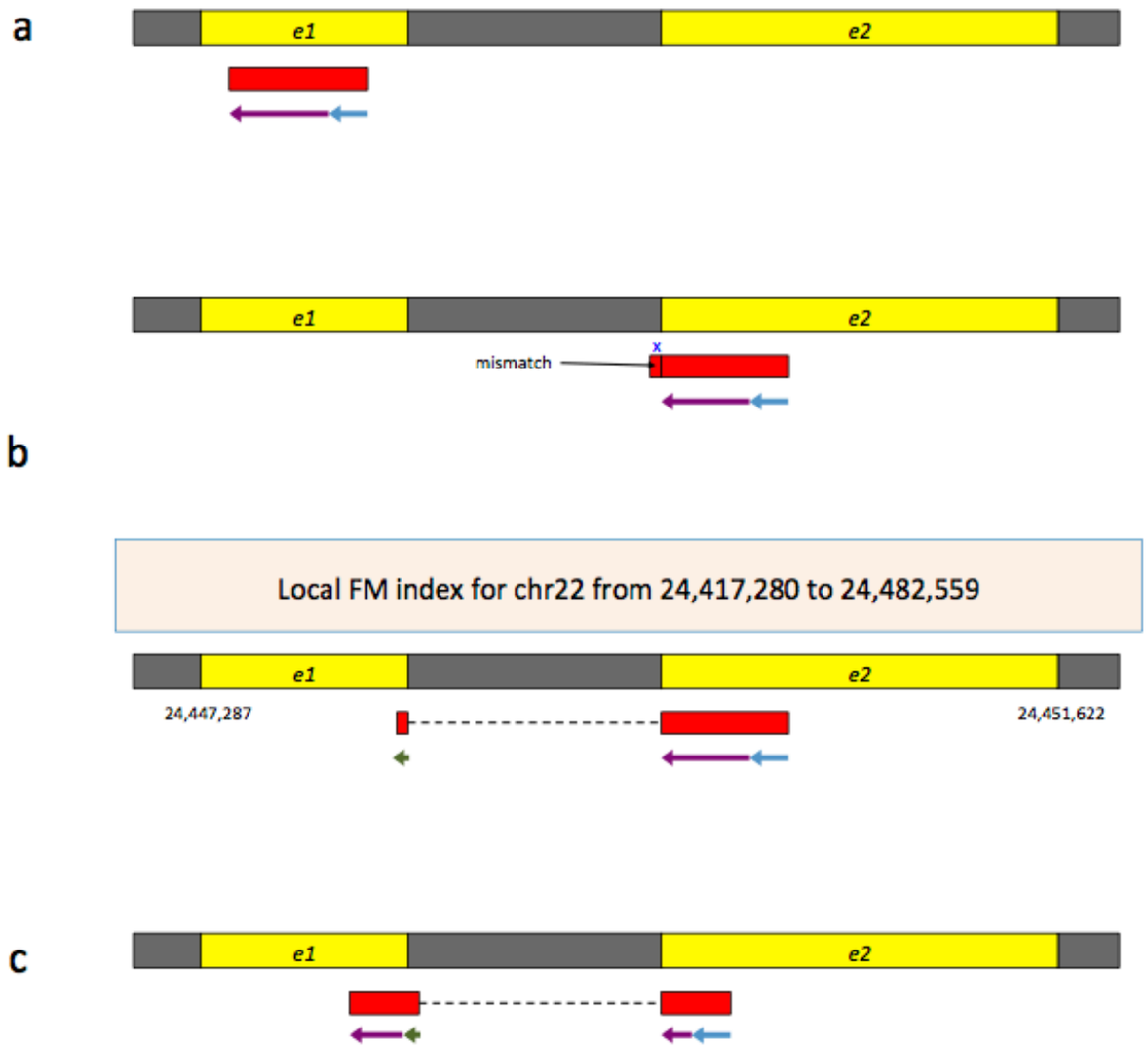
Mapping RNASeq reads to the genome

- To further optimize and speed up spliced read alignment, HISAT2 was developed.

- It uses a hierarchical graph FM (HGFM) index, representing the entire genome and eventual variants, together with overlapping local indexes (each spanning ~57kb) that collectively cover the genome and its variants.

- This allows to find initial seed locations for potential read alignments in the genome using global index and to rapidly refine these alignments using a corresponding local index:

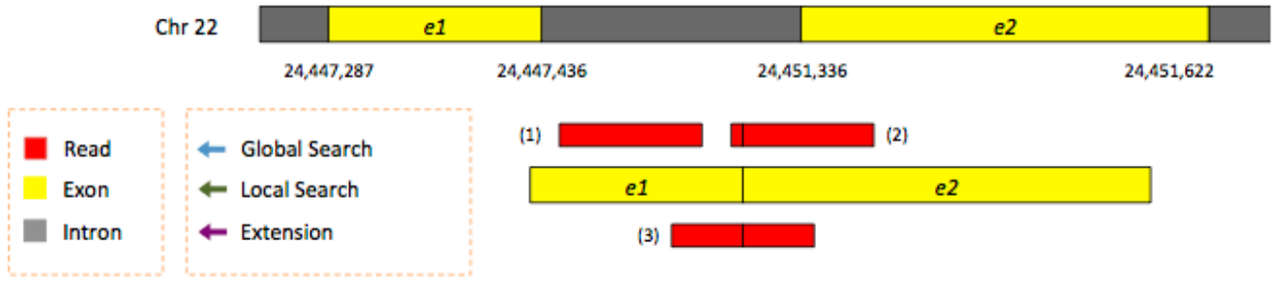


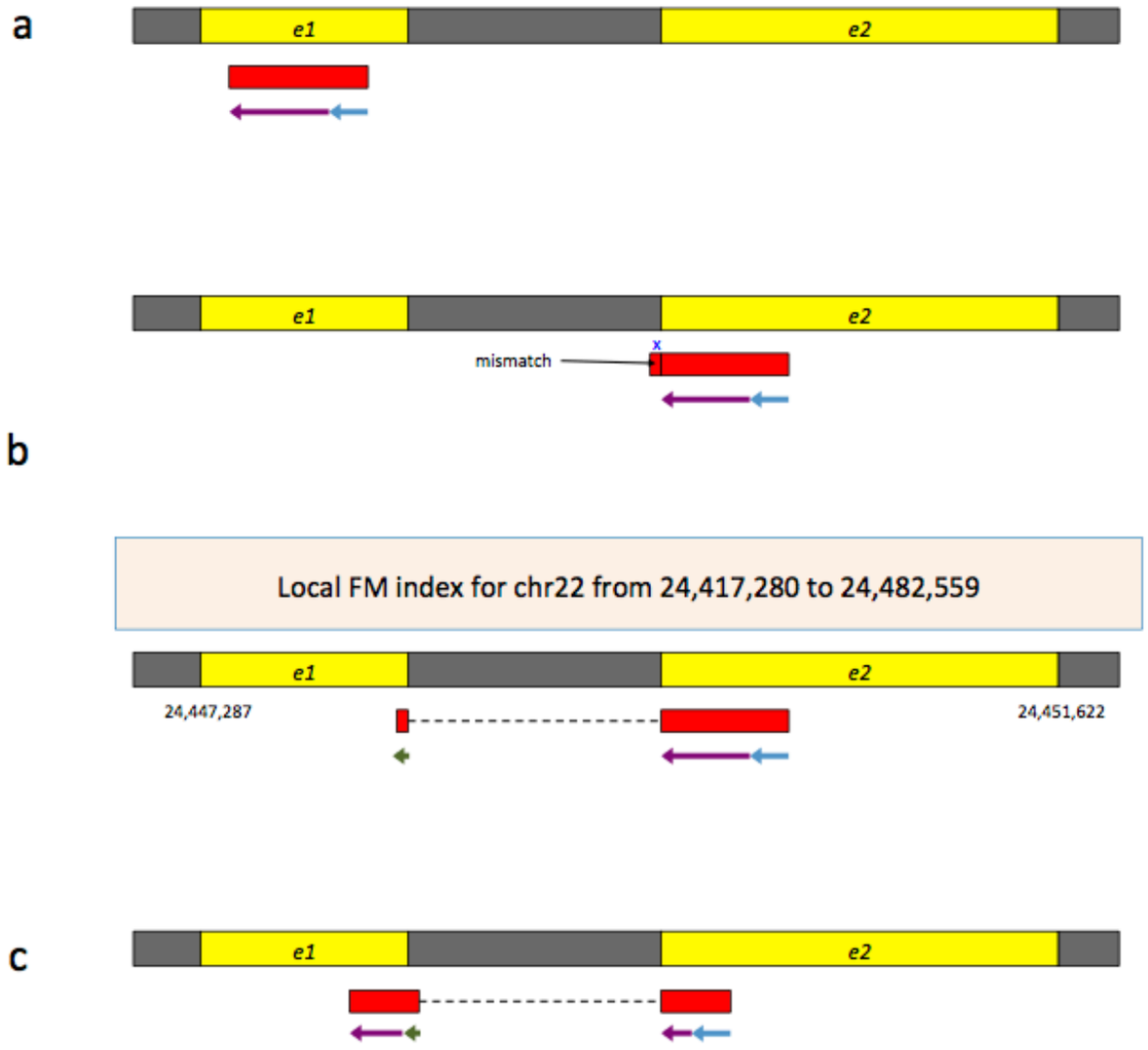


Genomics109

Mapping RNASeq reads to the genome

- A part of the read (blue arrow) is first mapped to the genome using the global FM index.
- HISAT2 then tries to extend the alignment directly utilizing the genome sequence (violet arrow).
- In (a) it succeeds and this read is aligned as it completely resides within an exon.
- In (b) the extension hits a mismatch. Now HISAT2 takes advantage of the local FM index overlapping this location to find the appropriate mapping for the remainder of this read (green arrow).
- The (c) shows a combination these two strategies: the beginning of the read is mapped using global FM index (blue arrow), extended until it reaches the end of the exon (violet arrow), mapped using local FM index (green arrow) and extended again (violet arrow).

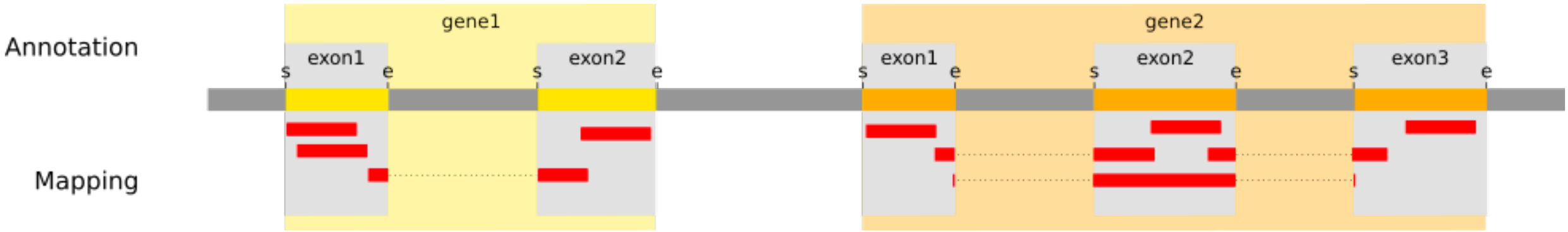




Genomics109

Counting the number of reads per annotated gene

- To compare the expression of single genes between different conditions (e.g. with or without PS depletion), an essential first step is to quantify the number of reads per gene, or more specifically the number of reads mapping to the exons of each gene.

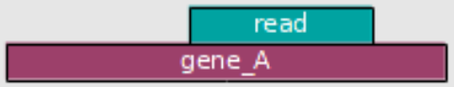
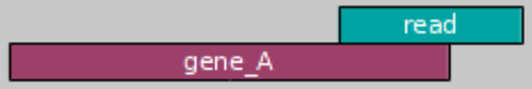
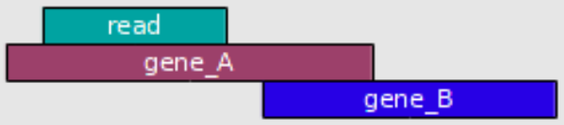
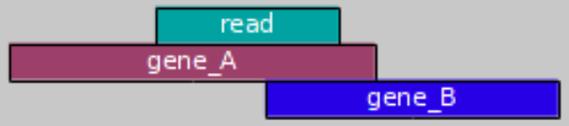
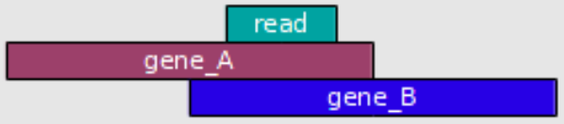
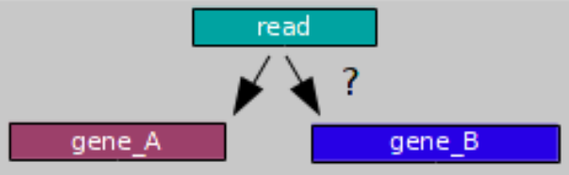


- Two main tools could be used for that:
  - HTSeqCount
  - Subread - featureCounts

Genomics109

Counting the number of reads per annotated gene

- **HTSeqCount:**

	union	intersection_strict	intersection_nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous (both genes with --nonunique all)	gene_A	gene_A
	ambiguous (both genes with --nonunique all)		
	alignment_not_unique (both genes with --nonunique all)		

Genomics109

Counting the number of reads per annotated gene

- **Subread - featureCounts:**
  - **FeatureCounts is considerably faster and requires far less computational resources.**
  - **In principle, the counting of reads overlapping with genomic features is a fairly simple task.**
  - **But there are some details that need to be given to featureCounts, e.g. the strandness.**

SUBREAD

Subread package: high-performance read alignment, quantification and mutation discovery

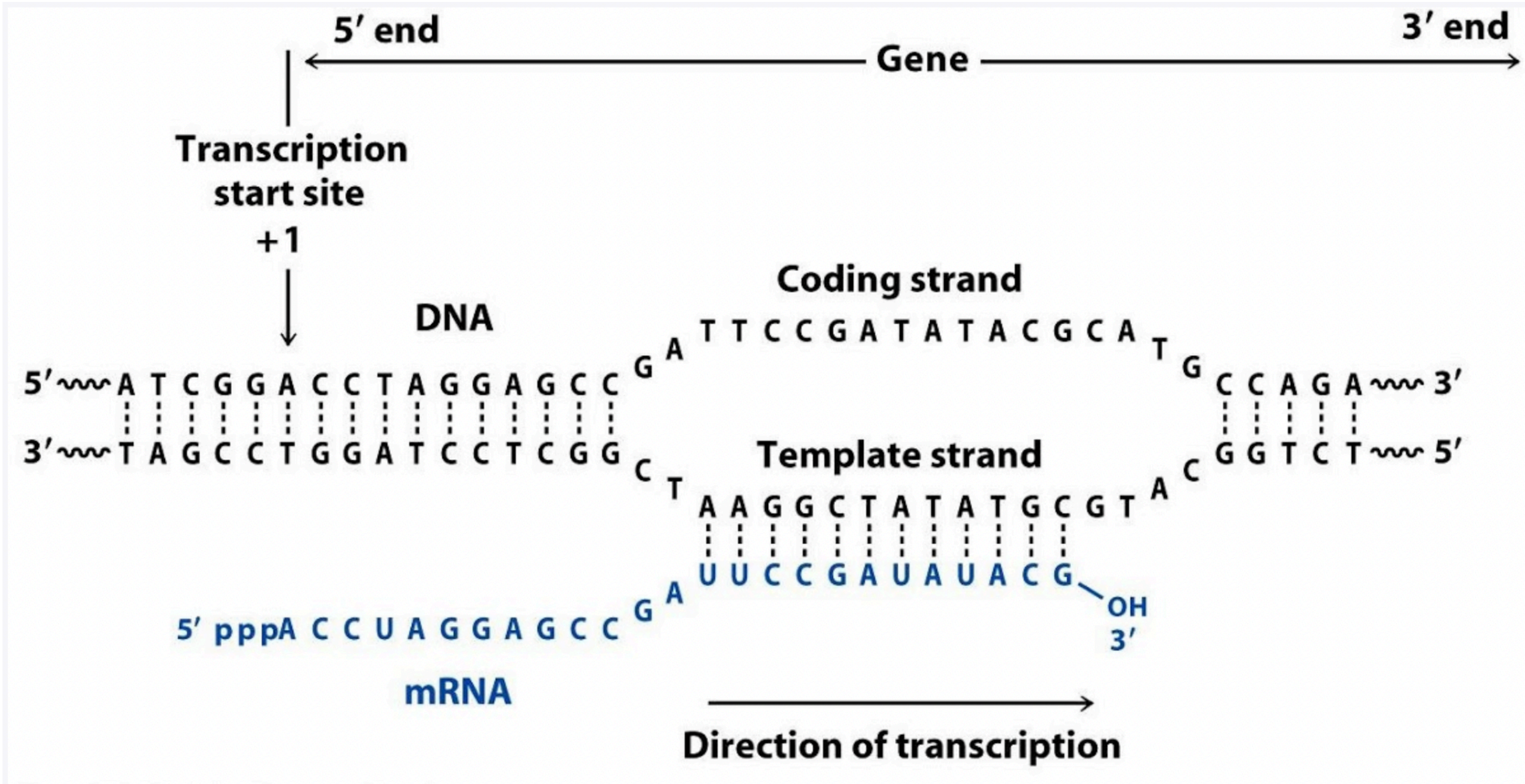
The Subread package comprises a suite of software programs for processing next-gen sequencing read data including:

- **Subread**: a general-purpose read aligner which can align both genomic DNA-seq and RNA-seq reads. It can also be used to discover genomic mutations including short indels and structural variants.
- **Subjunc**: a read aligner developed for aligning RNA-seq reads and for the detection of exon-exon junctions. Gene fusion events can be detected as well.
- **featureCounts**: a software program developed for counting reads to genomic features such as genes, exons, promoters and genomic bins.
- **Sublong**: a long-read aligner that is designed based on seed-and-vote.
- **exactSNP**: a SNP caller that discovers SNPs by testing signals against local background noises.

These programs were also implemented in Bioconductor R package Rsubread.

Genomics109

Estimation of the strandness

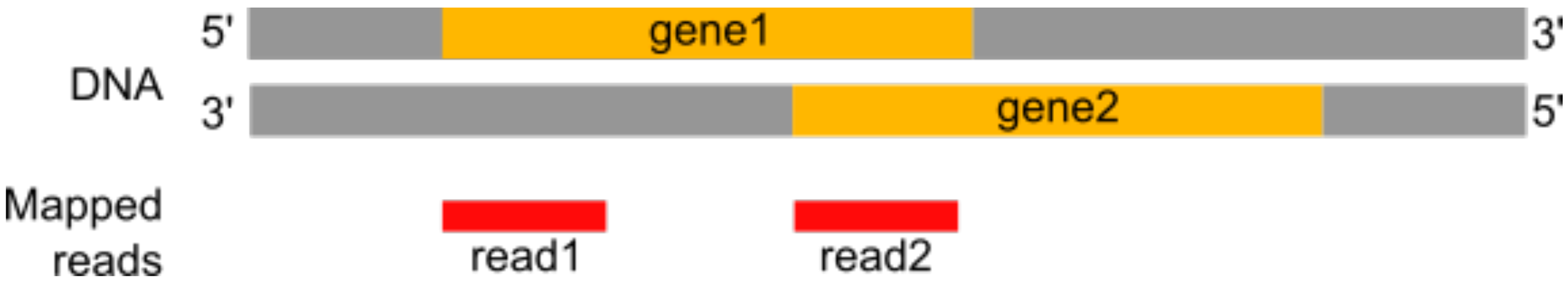


Genomics109

Estimation of the strandness

- **RNAs that are typically targeted in RNA-Seq experiments are single stranded (e.g., mRNAs) and thus have polarity (5’ and 3’ ends that are functionally distinct).**

- **During a typical RNA-Seq experiment the information about strandness is lost after both strands of cDNA are synthesized, size selected, and converted into a sequencing library. However, this information can be quite useful for the read counting step, especially for reads located on the overlap of 2 genes that are on different strands.**



Genomics109

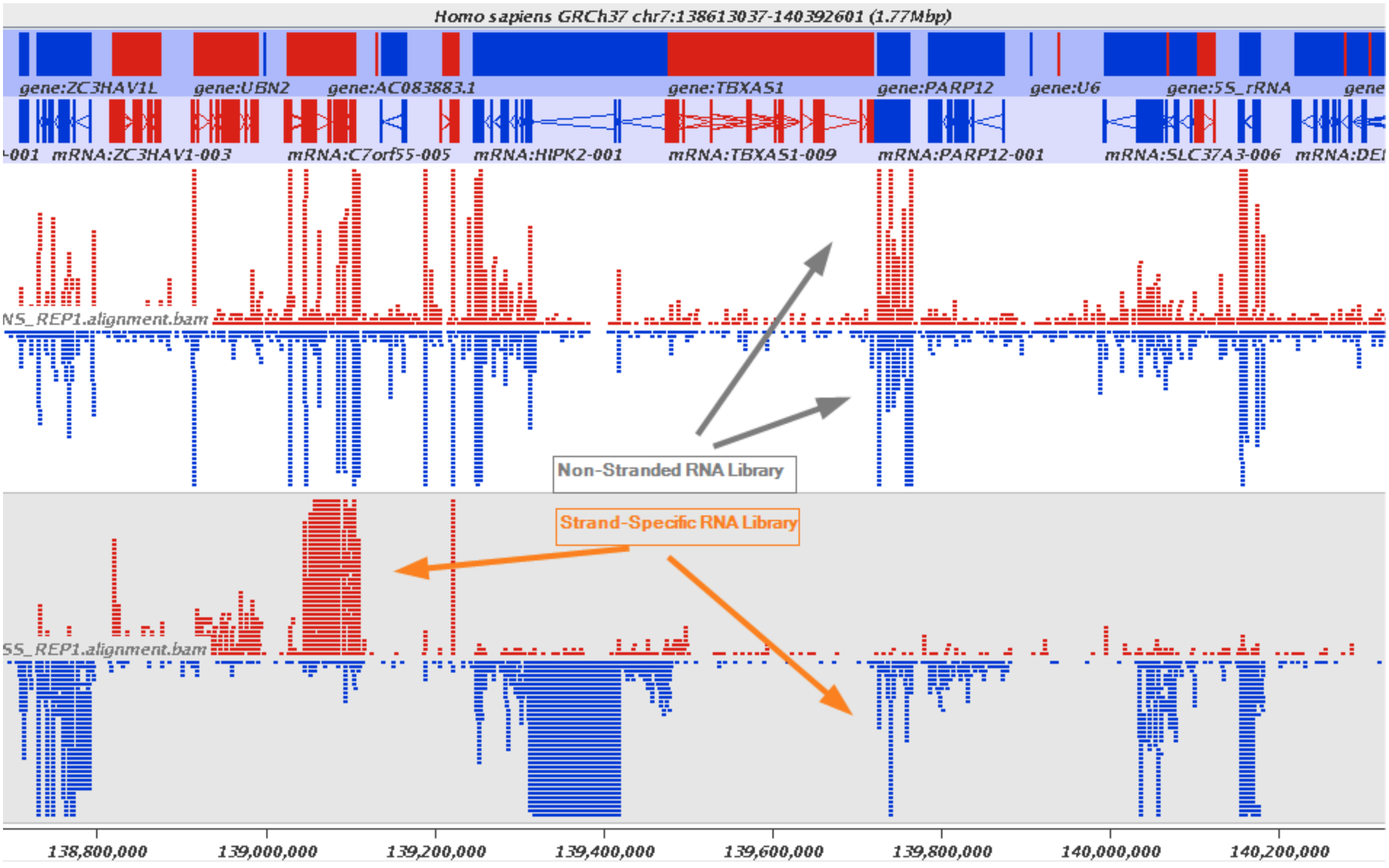
Estimation of the strandness

- Some library preparation protocols create so called stranded RNA-Seq libraries that preserve the strand information.
- In practice, with Illumina paired-end RNA-Seq protocols you are unlikely to encounter many of these possibilities.
- You will most likely deal with either:
  - Unstranded RNA-Seq data
  - Stranded RNA-Seq data generated by the use of specialized RNA isolation kits during sample preparation

Genomics109

Estimation of the strandness

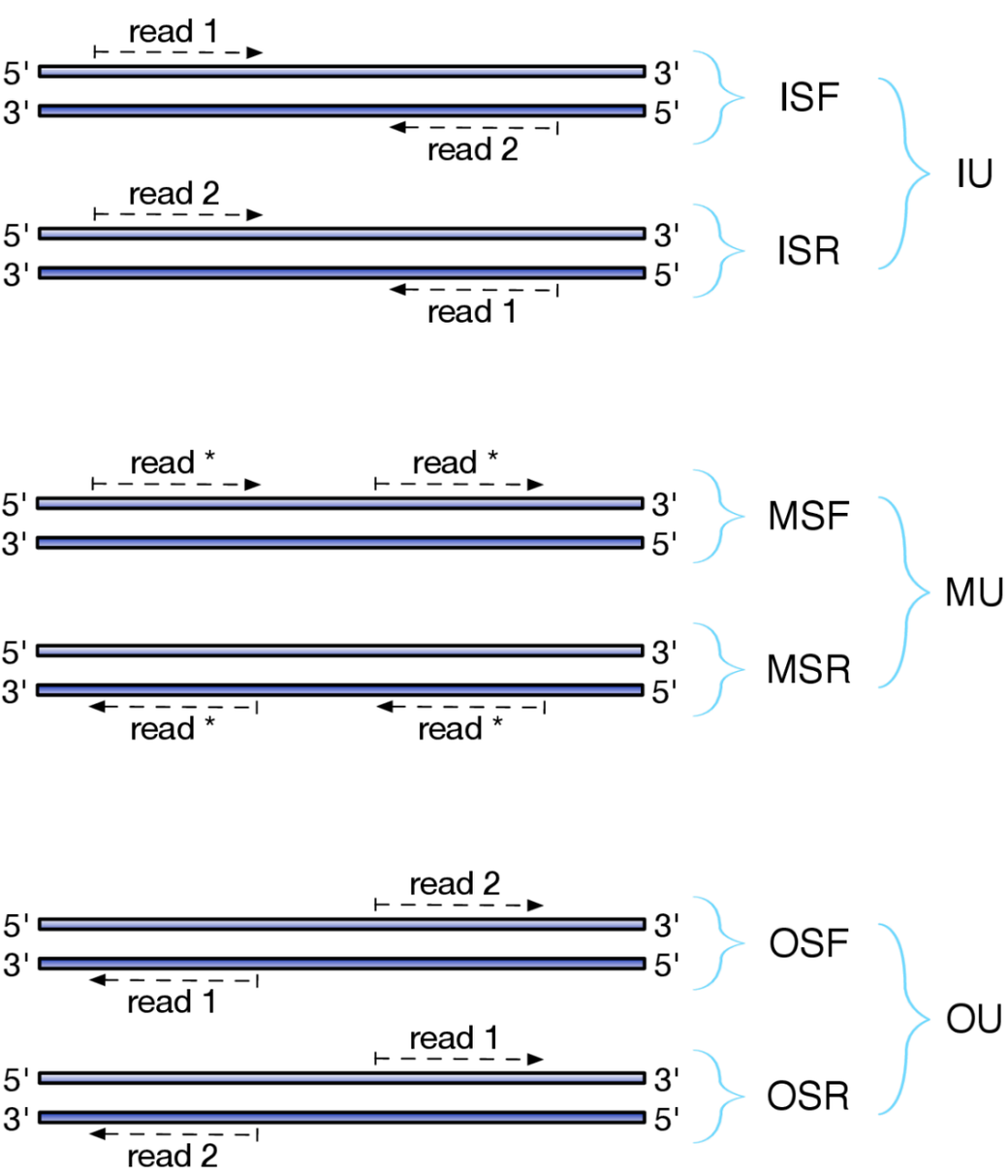
- **The implication of stranded RNA-Seq is that you can distinguish whether the reads are derived from forward or reverse-encoded transcripts:**



Genomics109

Estimation of the strandness

- **Depending on the approach, and whether one performs single-end or paired-end sequencing, there are multiple possibilities on how to interpret the results of the mapping of these reads to the genome:**



Relative orientation of the reads (paired-end)
I = inward
O = outward
M = matching (co-directional)

Library type
U = unstranded
S = stranded
Read origin
F = read1 (paired-end) or single-end read from forward strand
R = read1 (paired-end) or single-end read from reverse strand

Genomics109

Estimation of the strandness

- **Another option is to estimate these parameters with a tool called Infer Experiment from the RSeQC tool suite.**

BIOINFORMATICS APPLICATIONS NOTE

Vol. 28 no. 16 2012, pages 2184–2185
doi:10.1093/bioinformatics/bts356

Sequence analysis

Advance Access publication June 27, 2012

RSeQC: quality control of RNA-seq experiments

Liguo Wang[1,2], Shengqin Wang[3] and Wei Li[1,2,*]

[1]Division of Biostatistics, Dan L. Duncan Cancer Center and [2]Department of Molecular and Cellular Biology, Baylor College of Medicine, Houston, TX 77030, USA and [3]State Key Laboratory of Bioelectronics, School of Biological Science and Medical Engineering, Southeast University, Nanjing, China

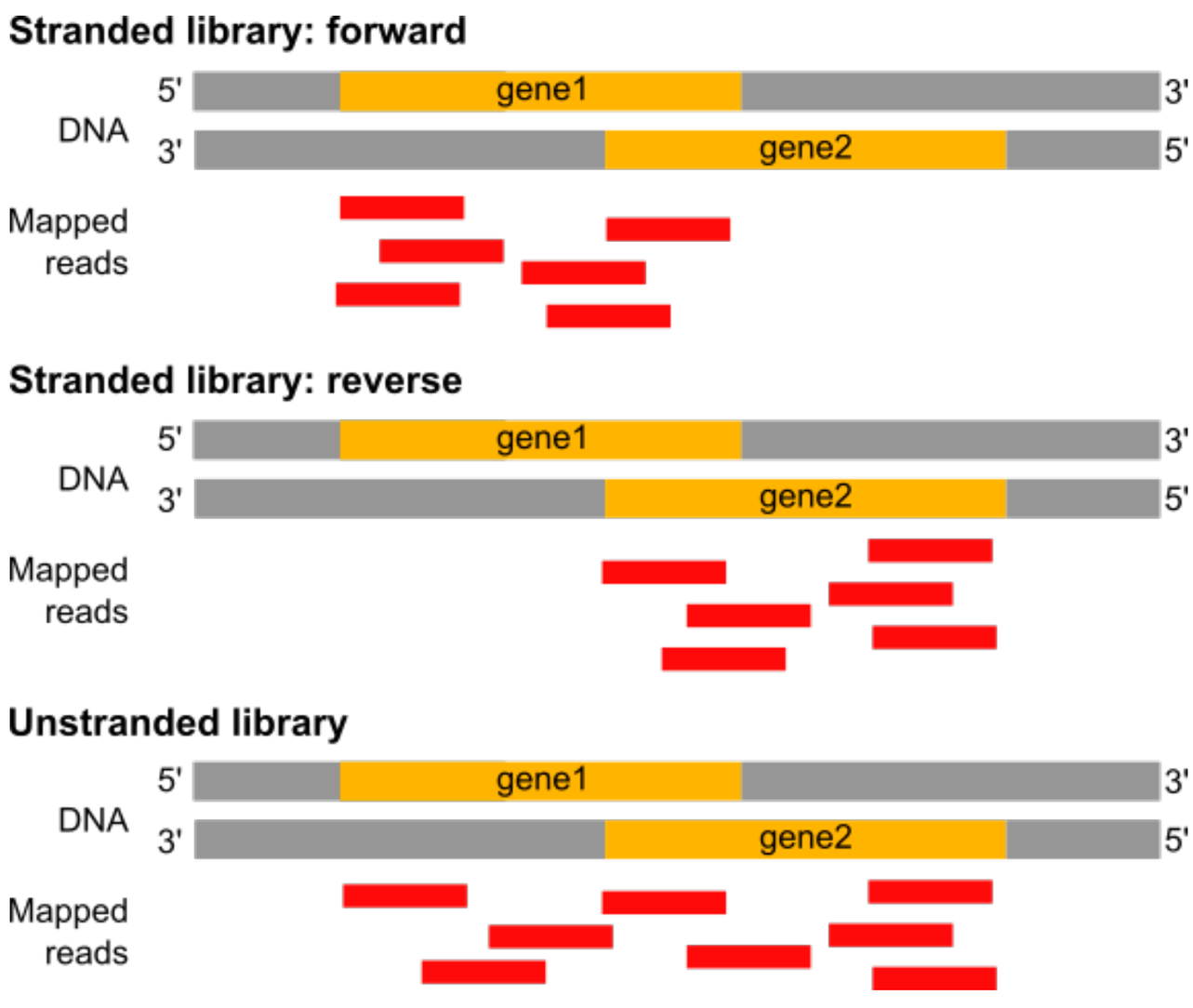
Associate Editor: Ivo Hofacker

Genomics109

Estimation of the strandness

- This tool takes the BAM files from the mapping, selects a subsample of the reads and compares their genome coordinates and strands with those of the reference gene model (from an annotation file).

- Based on the strand of the genes, it can gauge whether sequencing is strand-specific, and if so, how reads are stranded (forward or reverse):



Genomics109

Estimation of the strandness

- Infer Experiment tool generates one file with information on:
  - Paired-end or single-end library
  - Fraction of reads failed to determine
  - 2 lines
    - For single-end
      - Fraction of reads explained by “++,–”: the fraction of reads that assigned to forward strand
      - Fraction of reads explained by ”-,-”: the fraction of reads that assigned to reverse strand
    - For paired-end
      - Fraction of reads explained by “1++,1–,2+-,2-+”: the fraction of reads that assigned to forward strand
      - Fraction of reads explained by “1+-,1-+,2++,2–”: the fraction of reads that assigned to reverse strand

- If the two “Fraction of reads explained by” numbers are close to each other, we conclude that the library is not a strand-specific dataset (or unstranded).

- As it is sometimes quite difficult to find out which settings correspond to those of other programs, the following table might be helpful to identify the library type:

Genomics109

Estimation of the strandness

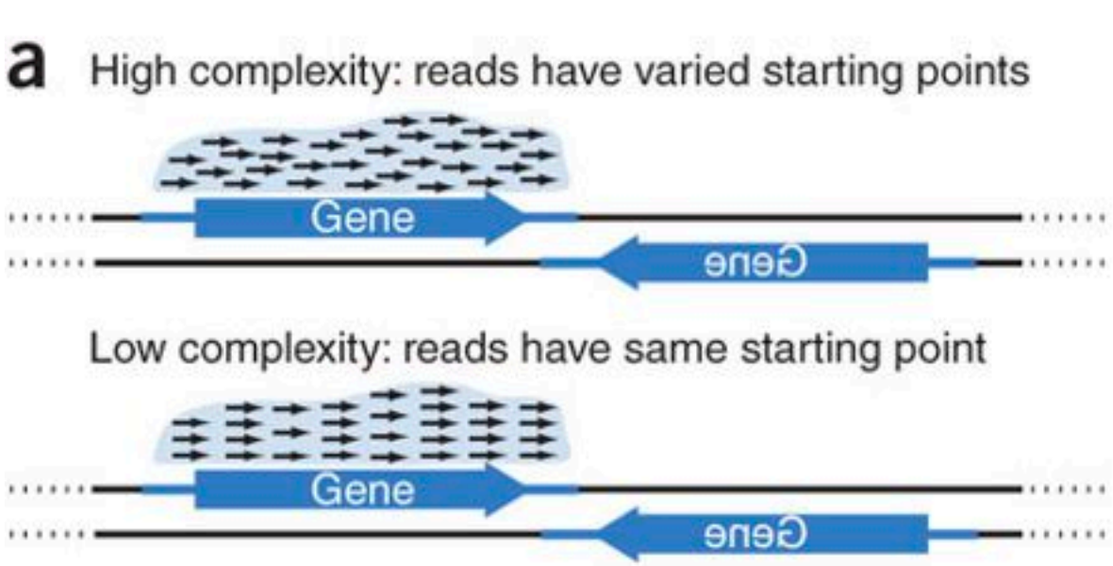
- **If the two “Fraction of reads explained by” numbers are close to each other, we conclude that the library is not a strand-specific dataset (or unstranded).**

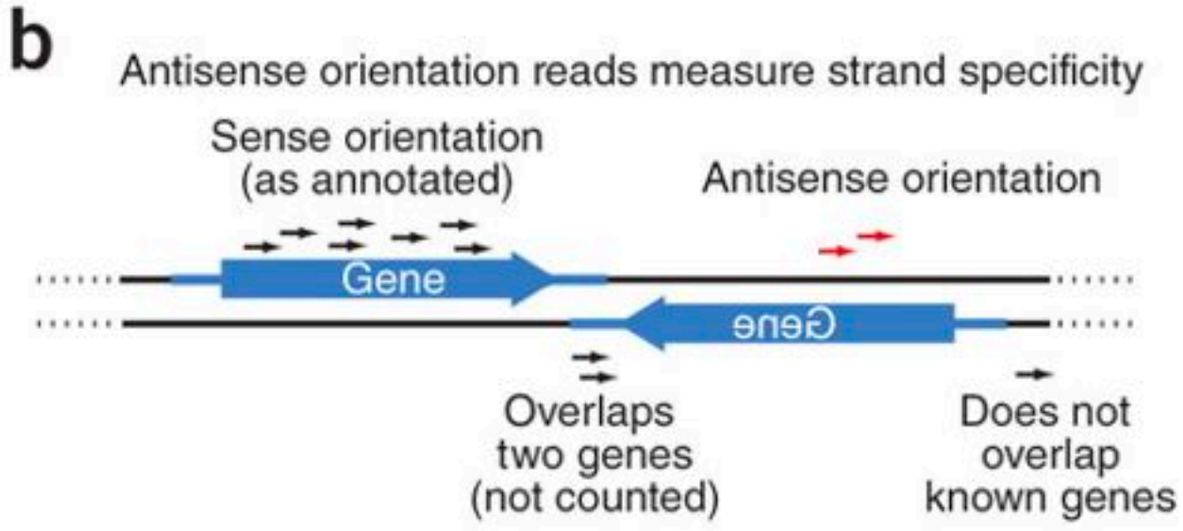
- **As it is sometimes quite difficult to find out which settings correspond to those of other programs, the following table might be helpful to identify the library type:**

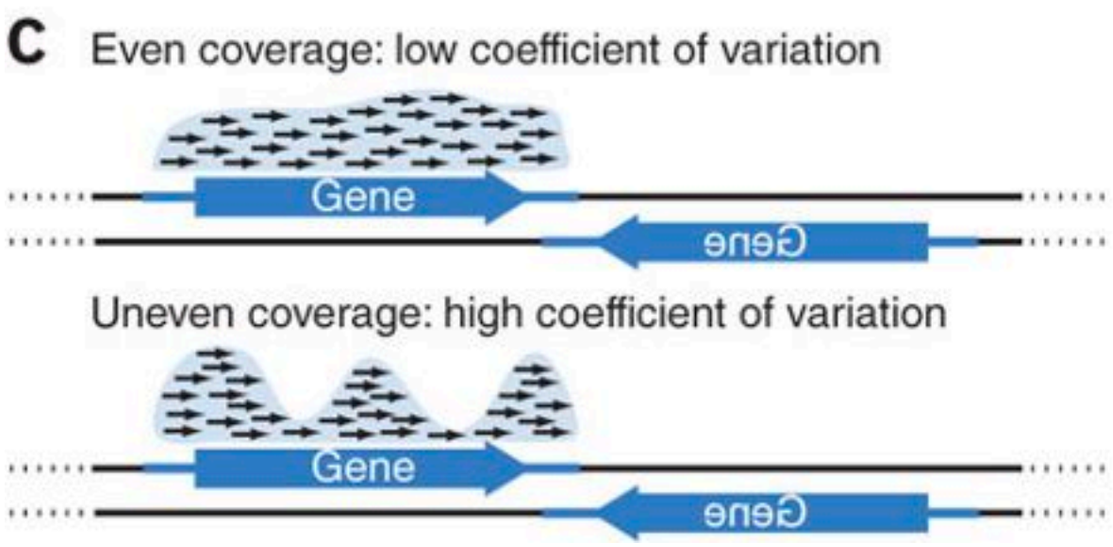
Library type	Infer Experiment	TopHat	HISAT2	HTSeq-count	featureCounts
Paired-End (PE) - SF	1 + + , 1 - - , 2 + - , 2 - +	FR Second Strand	Second Strand F/FR	yes	Forward (1)
PE - SR	1 + - , 1 - + , 2 + + , 2 - -	FR First Strand	First Strand R/RF	reverse	Reverse (2)
Single-End (SE) - SF	+ + , - -	FR Second Strand	Second Strand F/FR	yes	Forward (1)
SE - SR	+ - , - +	FR First Strand	First Strand R/RF	reverse	Reverse (2)
PE, SE - U	undecided	FR Unstranded	default	no	Unstranded (0)

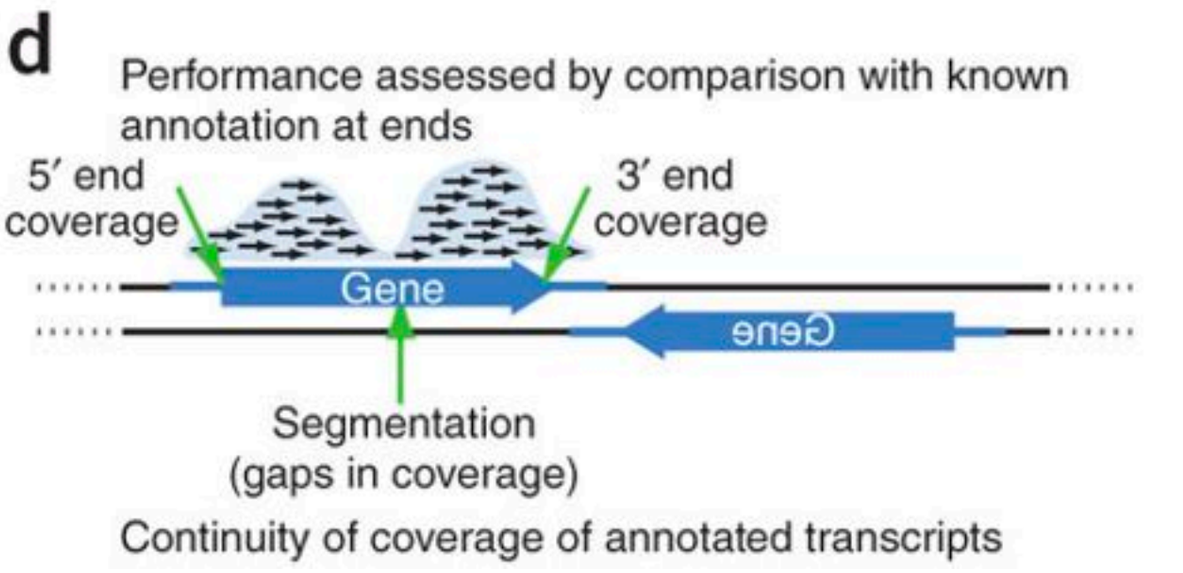
Genomics109

Mapping Complexity Considerations



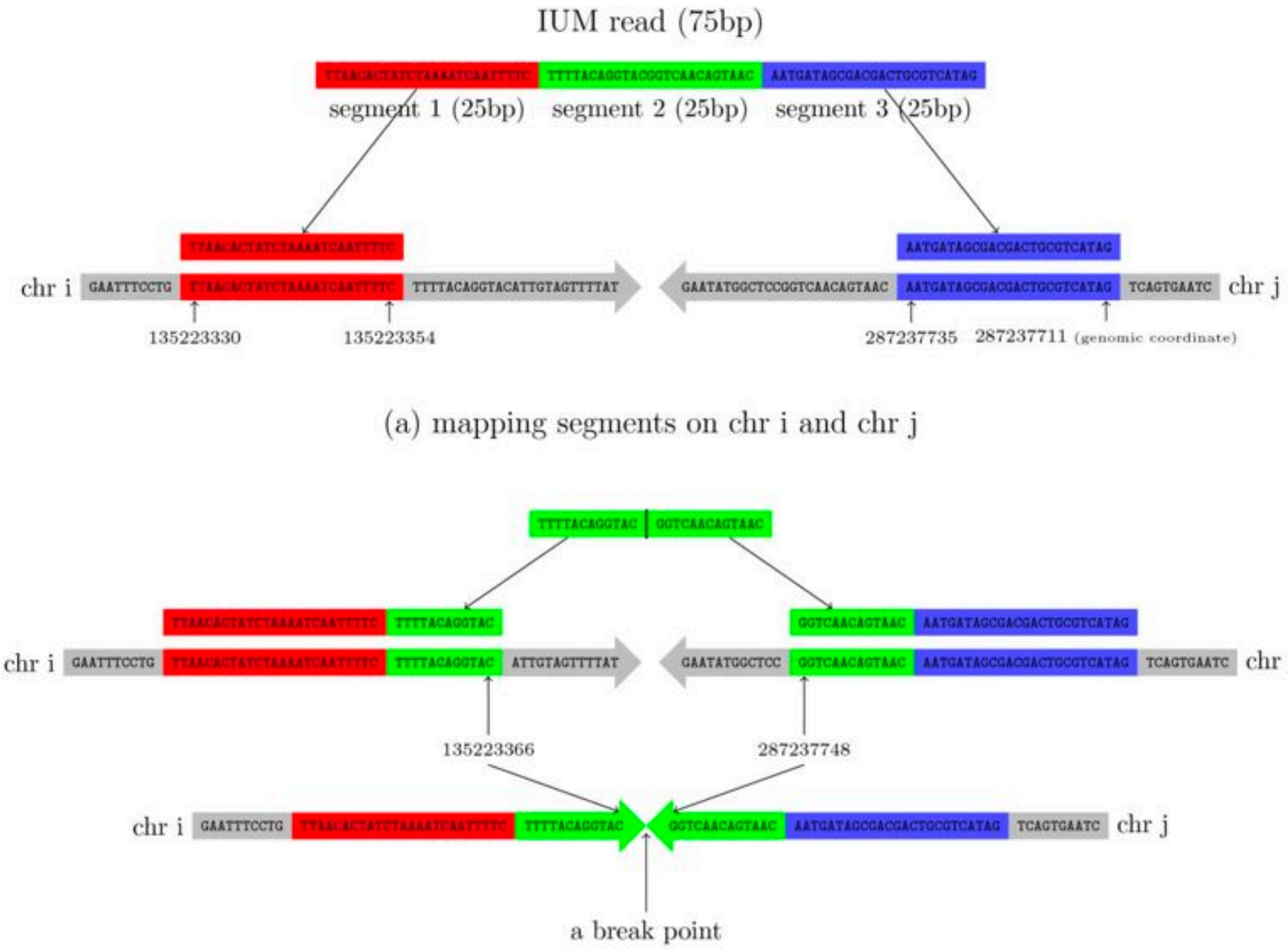






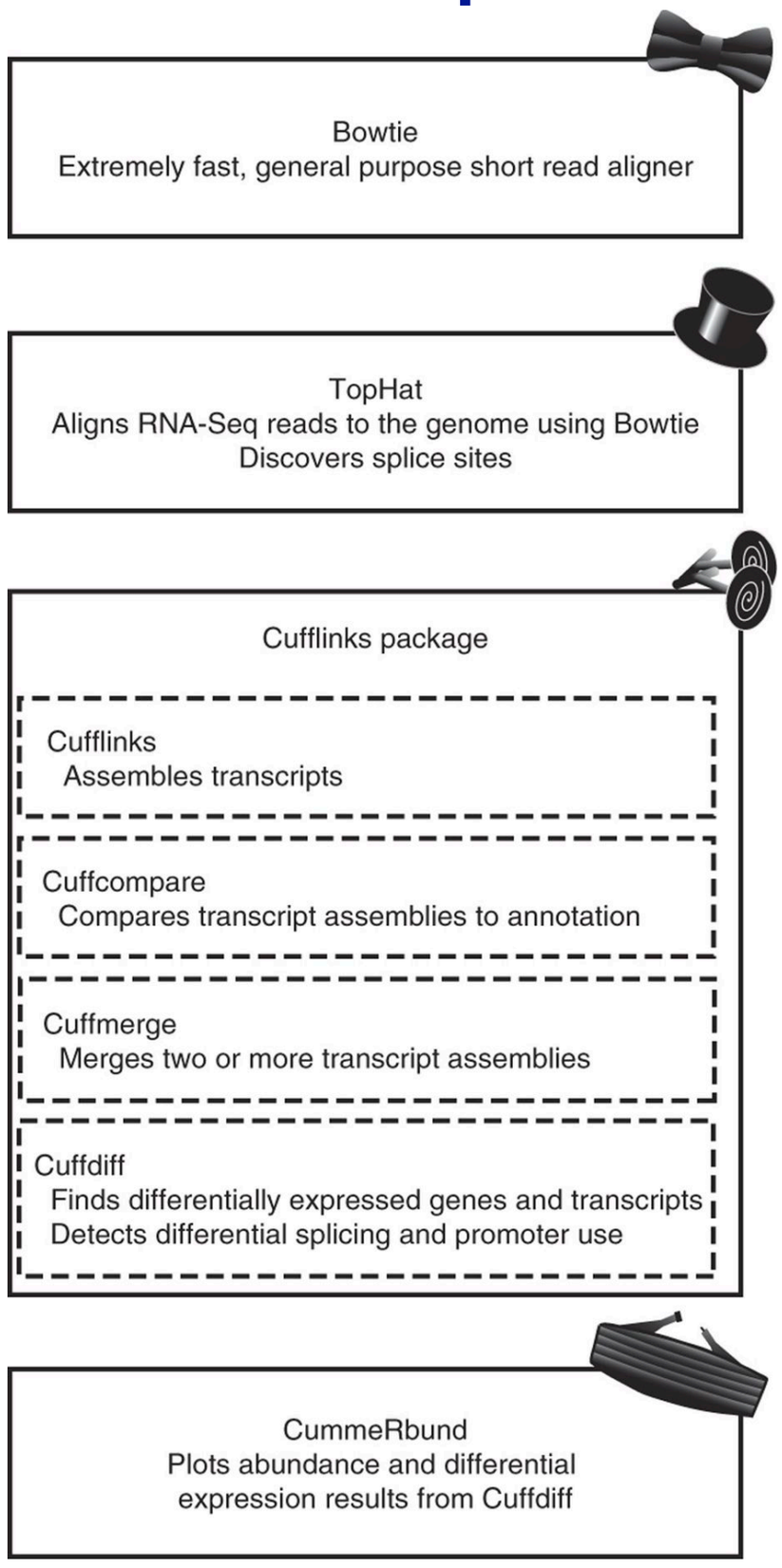
Genomics109

Identifying Fusion Events



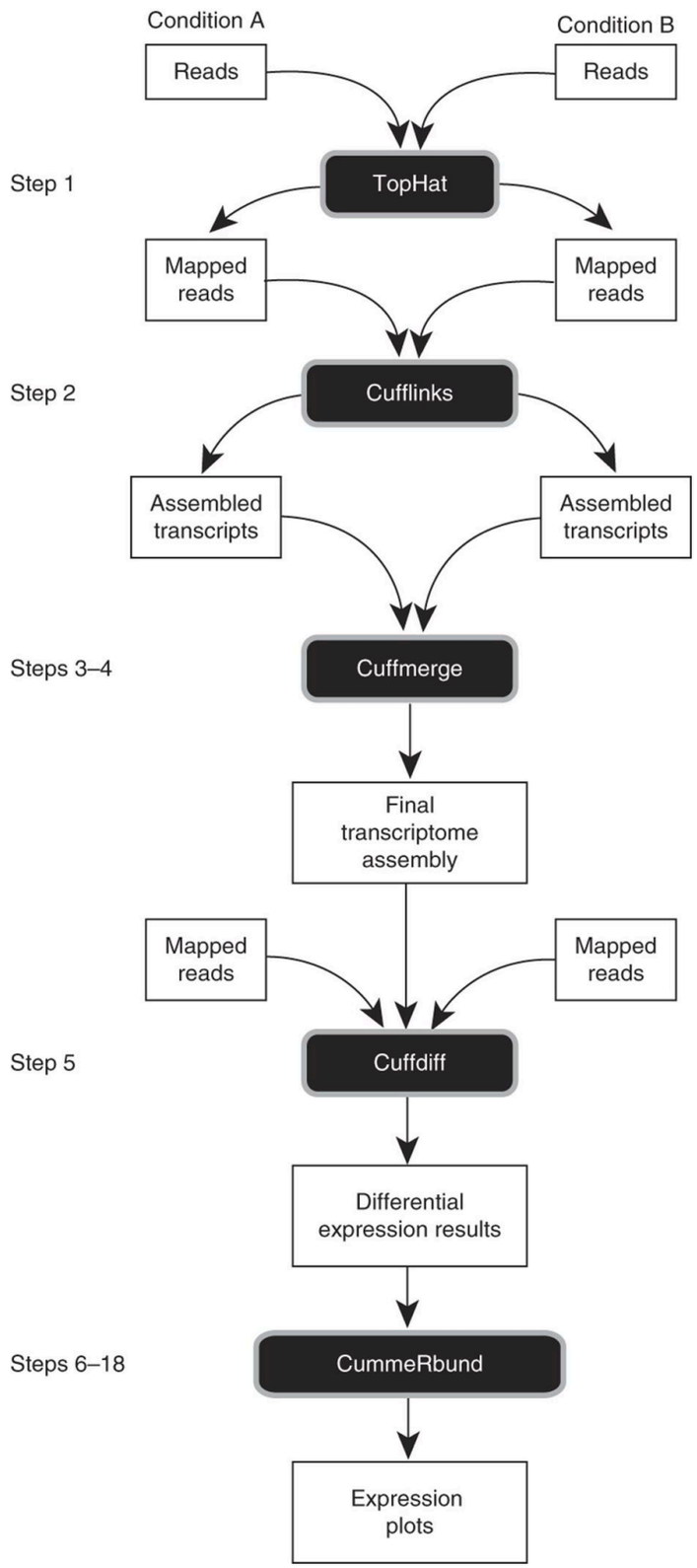
Genomics109

Tuxedo Pipeline



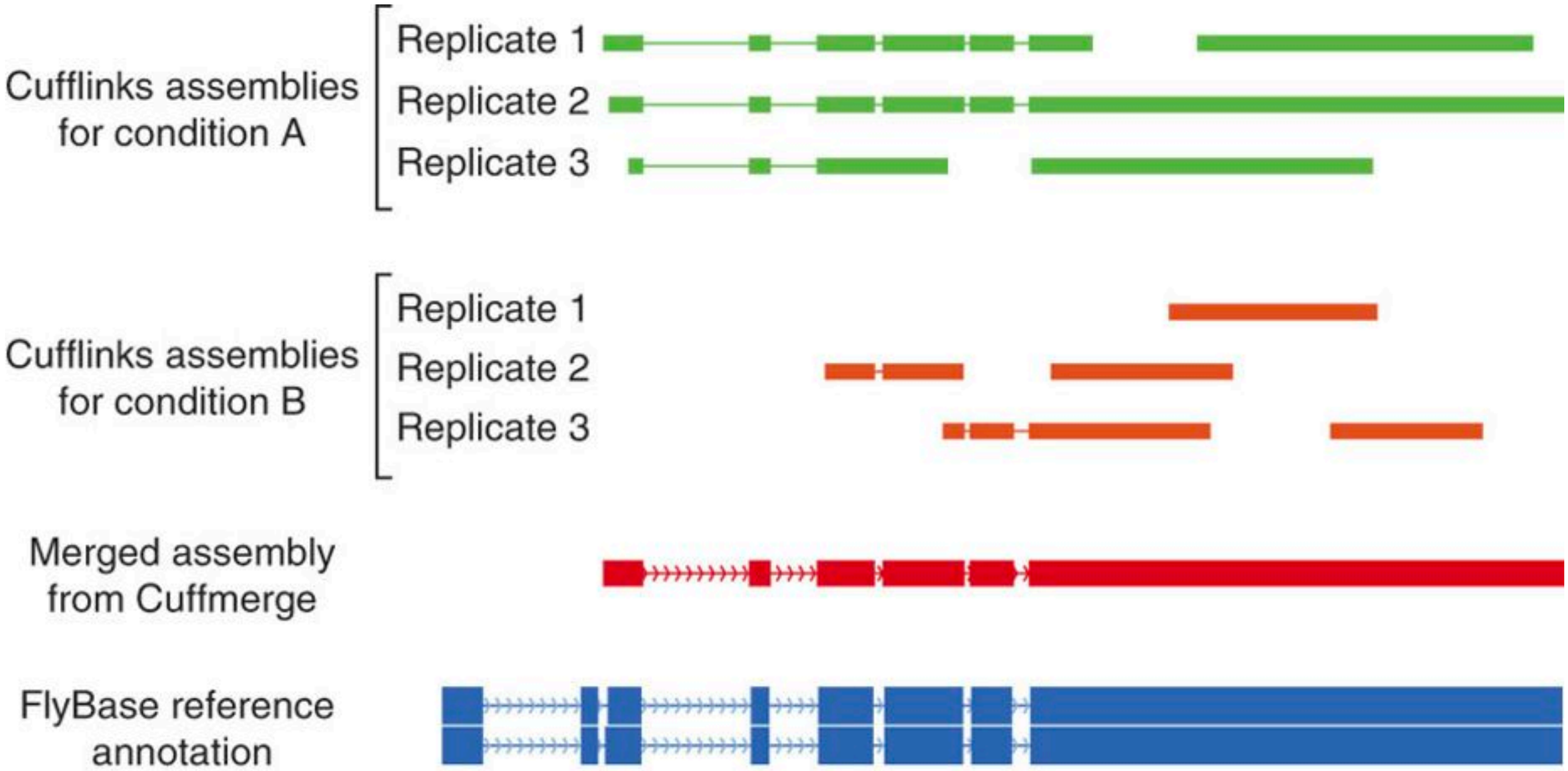
Genomics109

Comparing Conditions



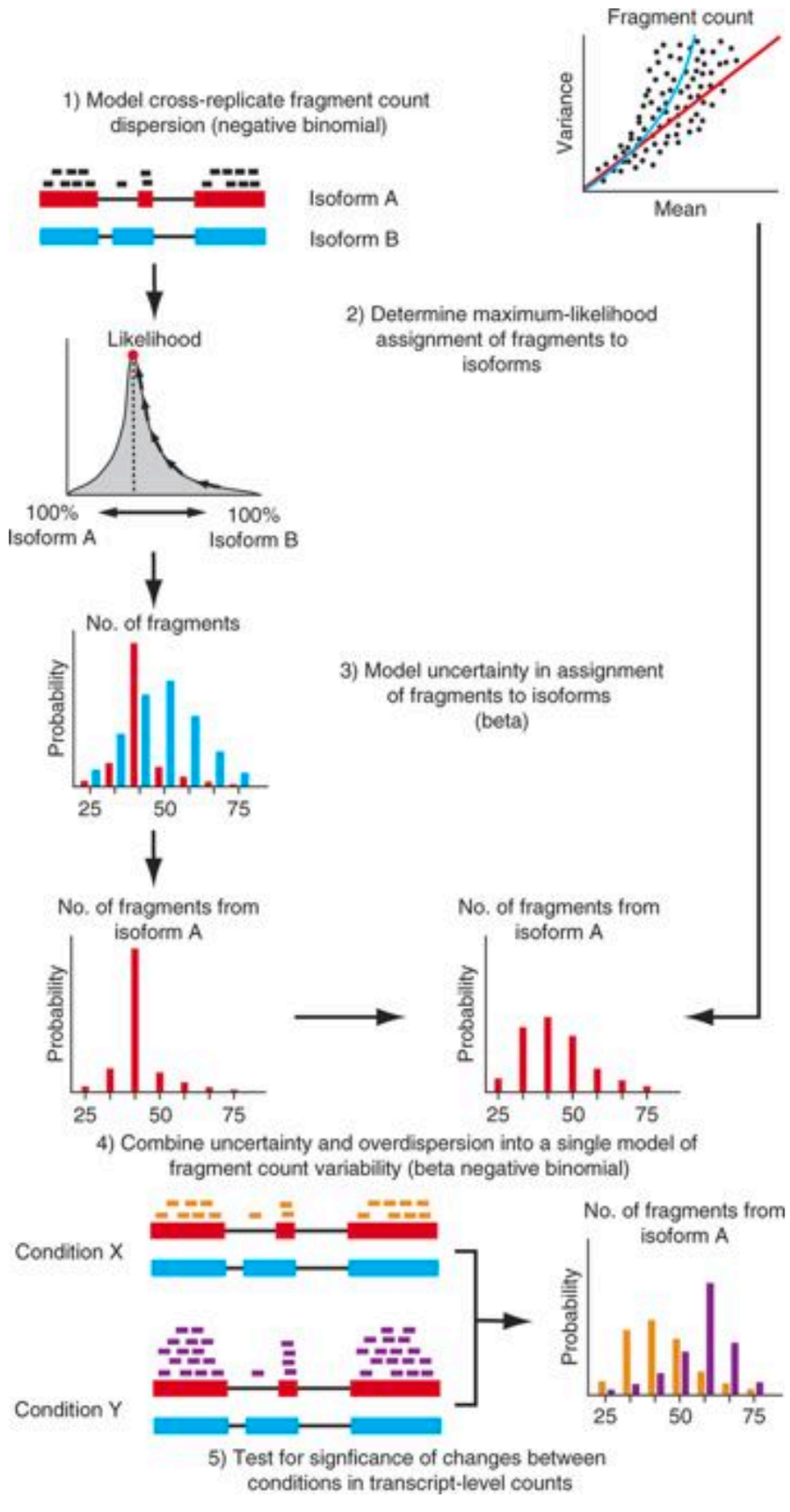
Genomics109

Analysis



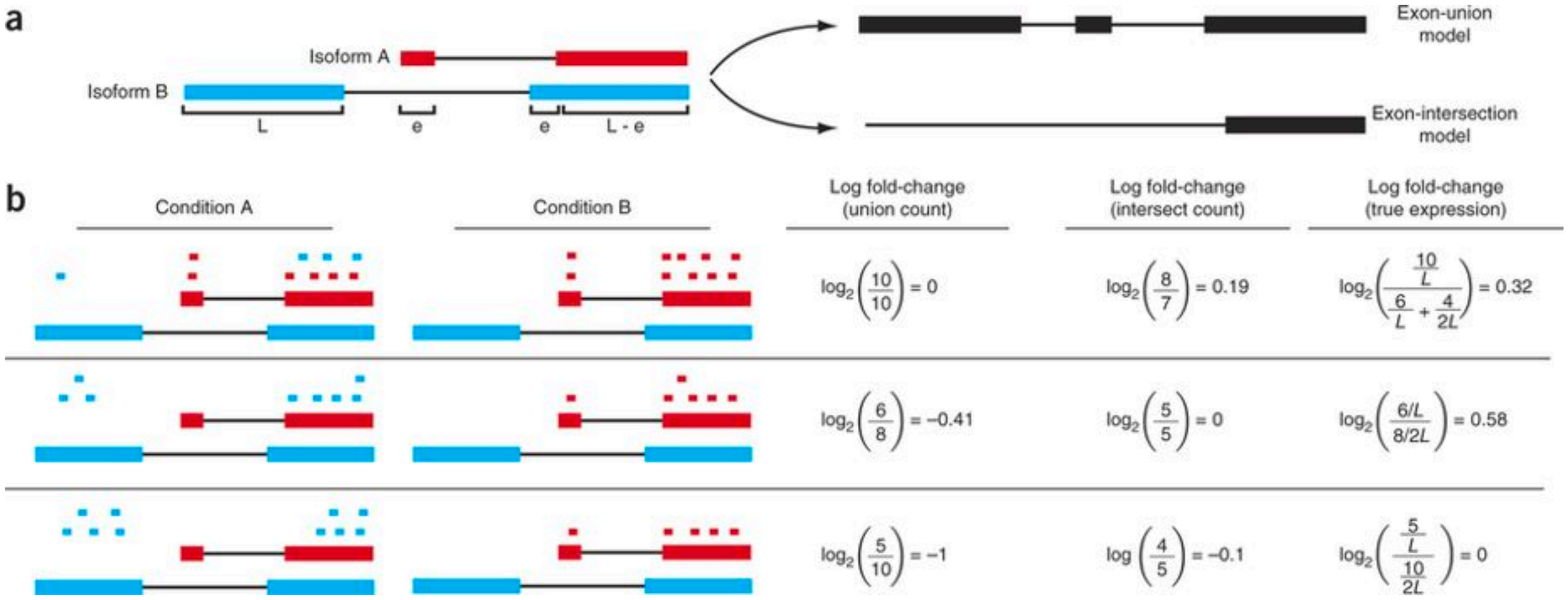
Genomics109

Analysis



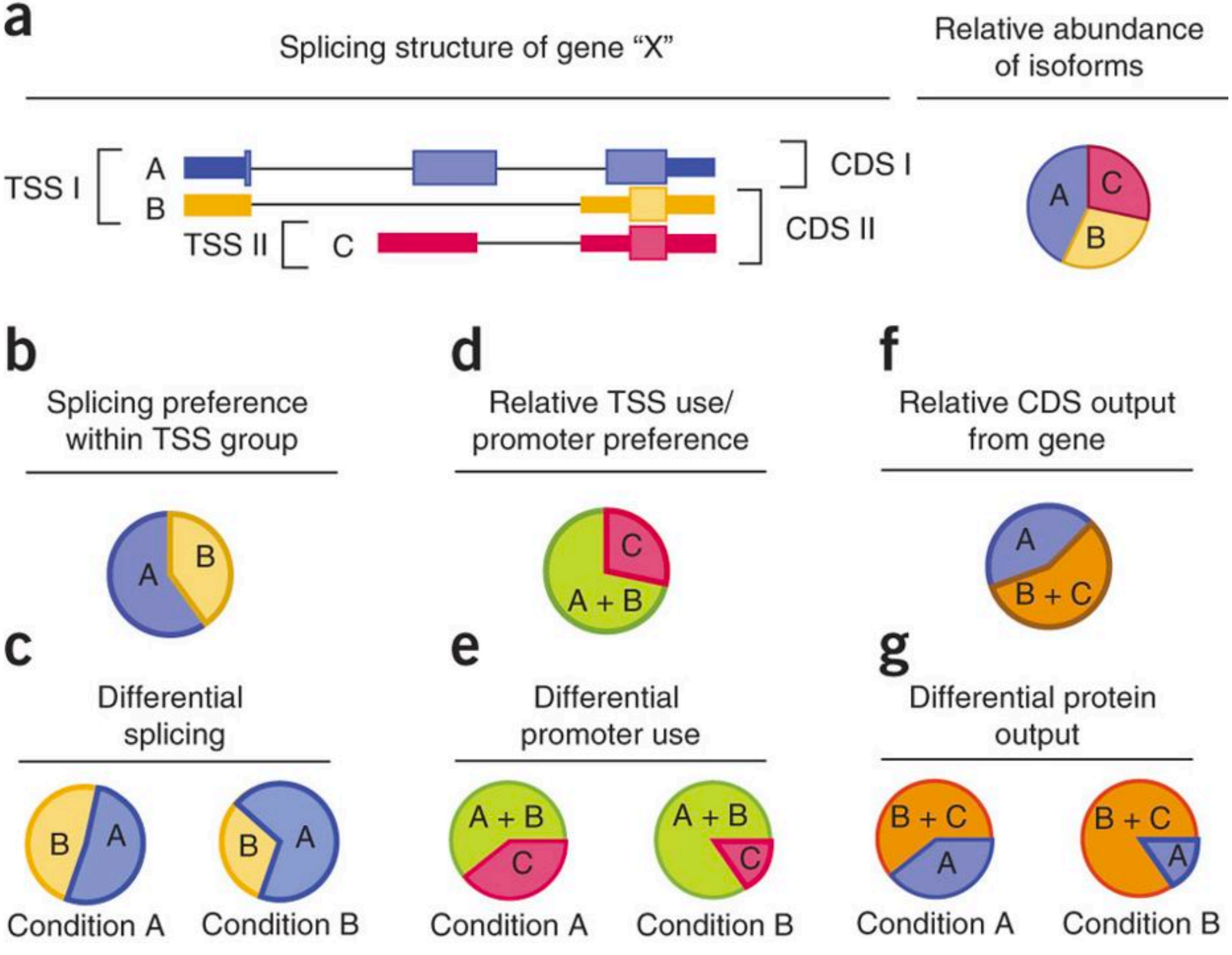
Genomics109

Analysis



Genomics109

Analysis

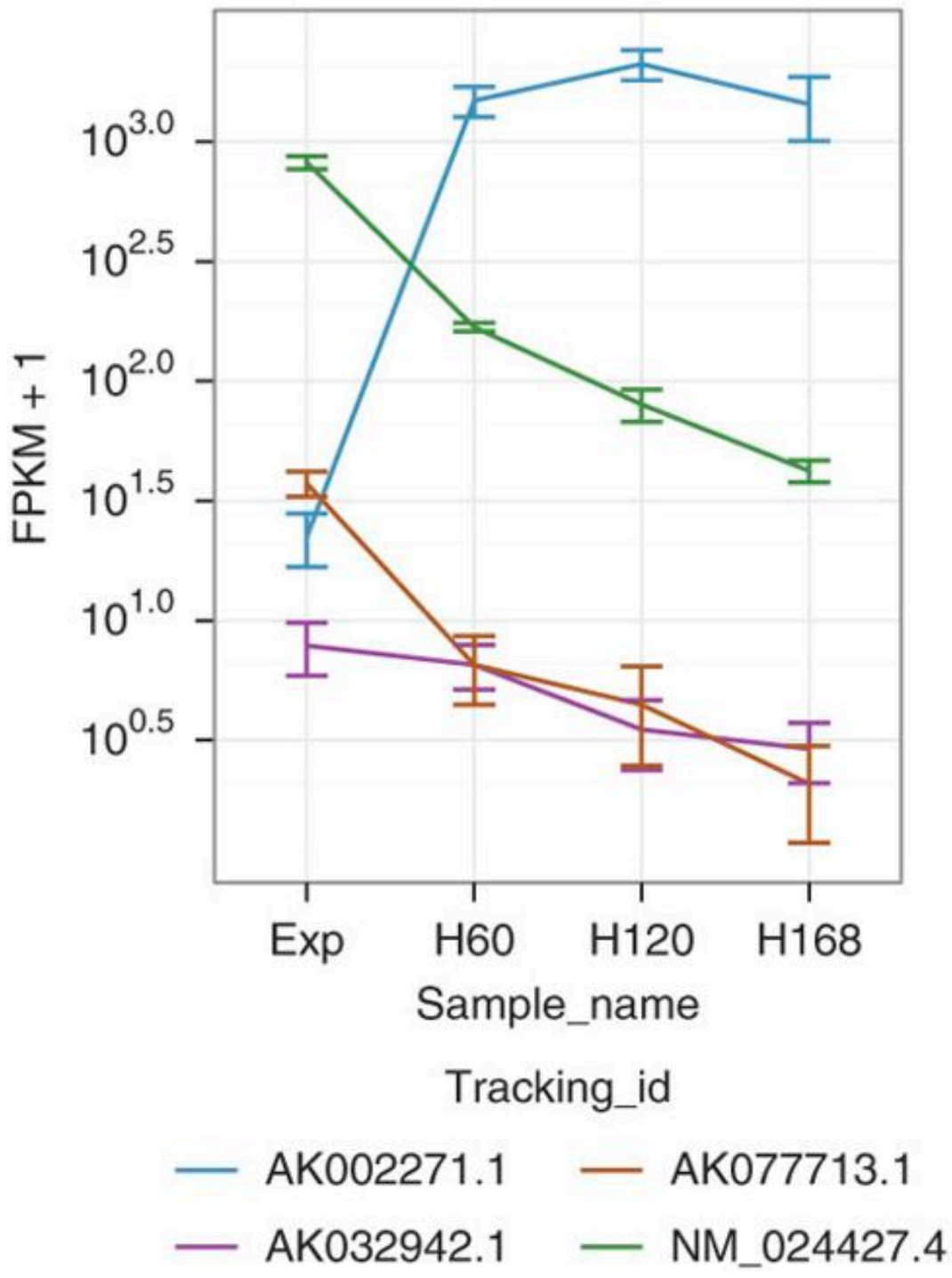


Genomics109

Analysis

a

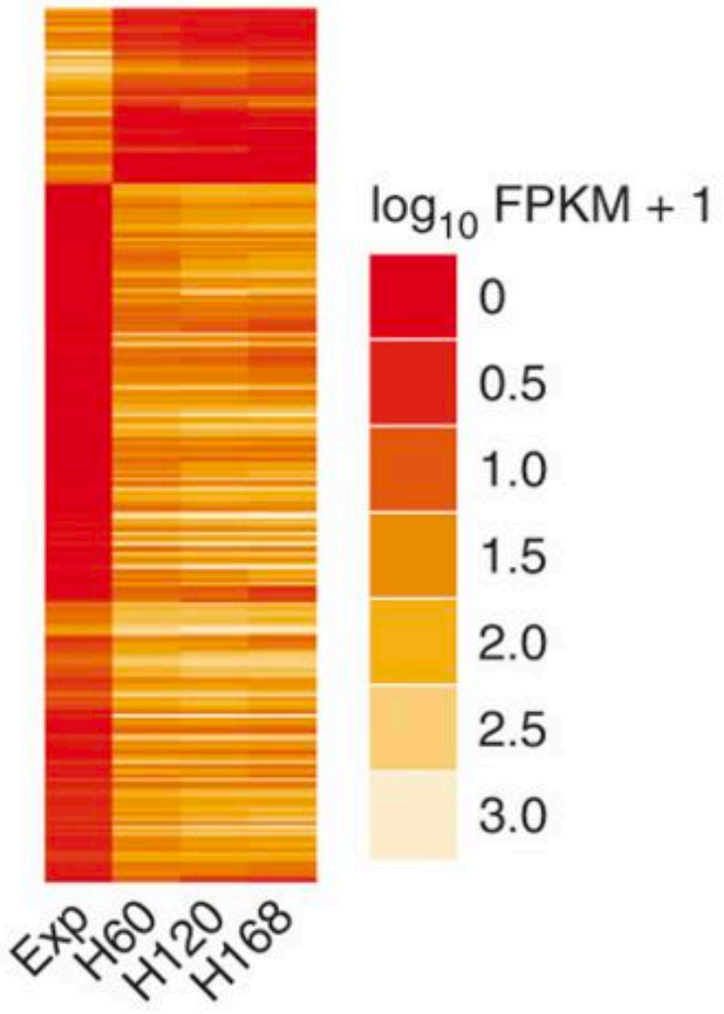
```
expressionPlot(isoforms(tpn1),logMode=T)
```



b

```
sig_genes <- getGenes(cd, geneIdList)

csHeatmap(sig_genes,
          clustering="row",
          labRow=F)
```



Genomics109

Tuxedo Pipeline Information

Program	Manual	Publications
TopHat	TopHat2_Manual	TopHat: discovering splice junctions with RNA-Seq
		TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions
TopHat-Fusion	TopHat-Fusion_Manual	TopHat-Fusion: an algorithm for discovery of novel fusion transcripts
Cufflinks	Cufflinks_Manual	Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform …
		Improving RNA-Seq expression estimates by correcting for fragment bias
		Identification of novel transcripts in annotated genomes using RNA-Seq
		Differential analysis of gene regulation at transcript resolution with RNA-seq

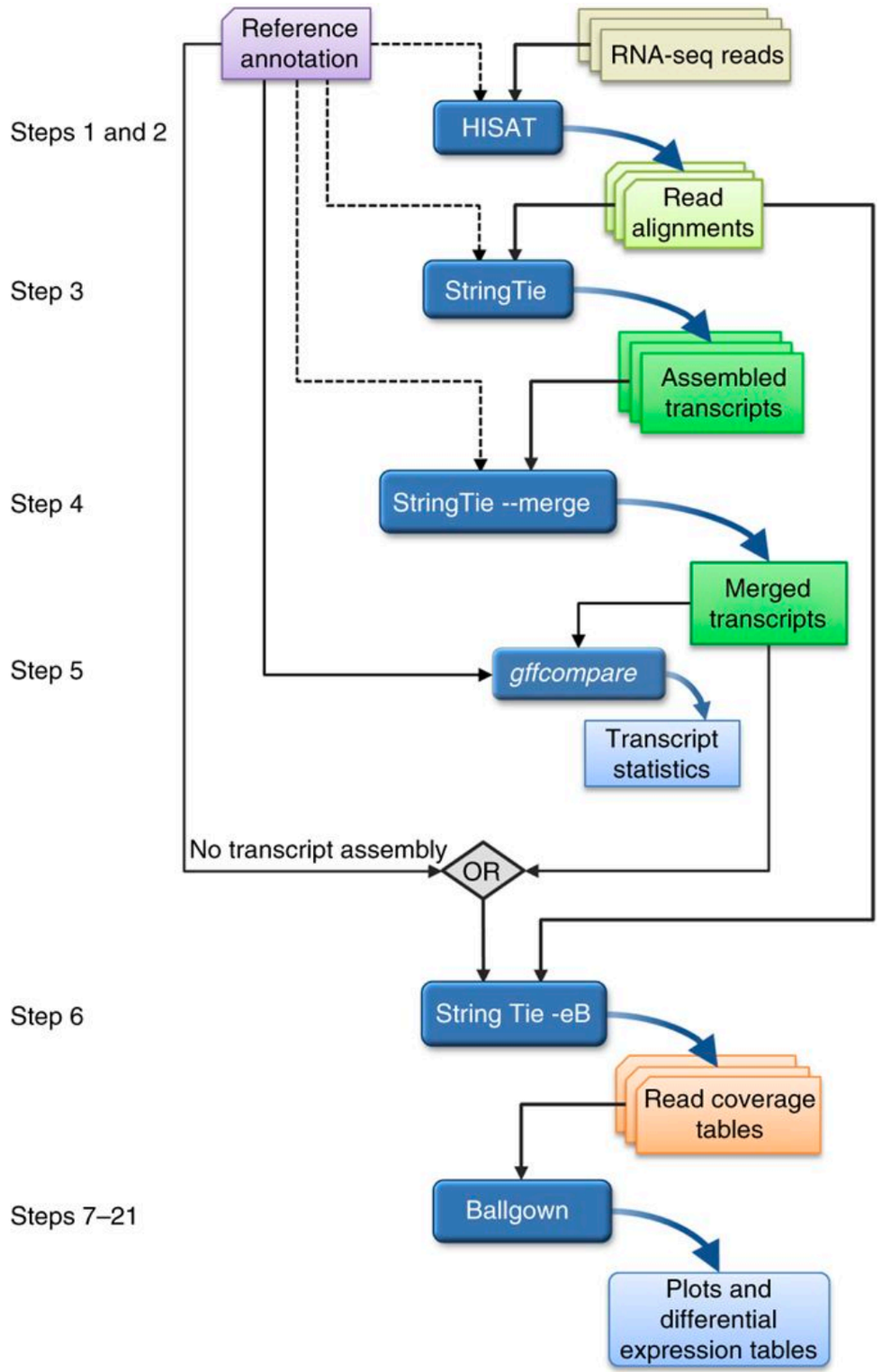
Genomics109

Tuxedo Pipeline Tutorials

Journal	Title
Nature Protocols:	**Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks**

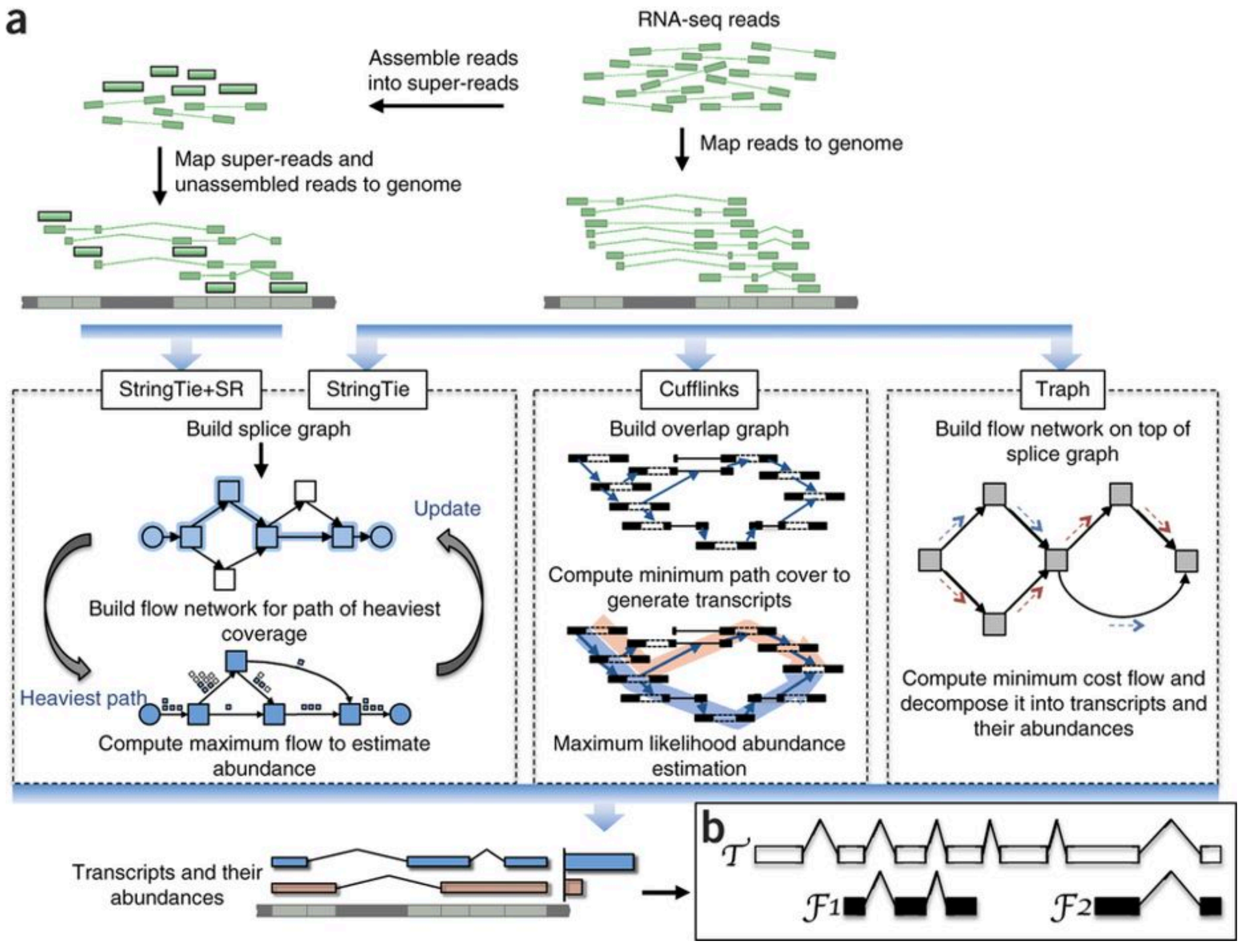
Genomics109

The New Pipelines
HISAT2, StringTie and Ballgown



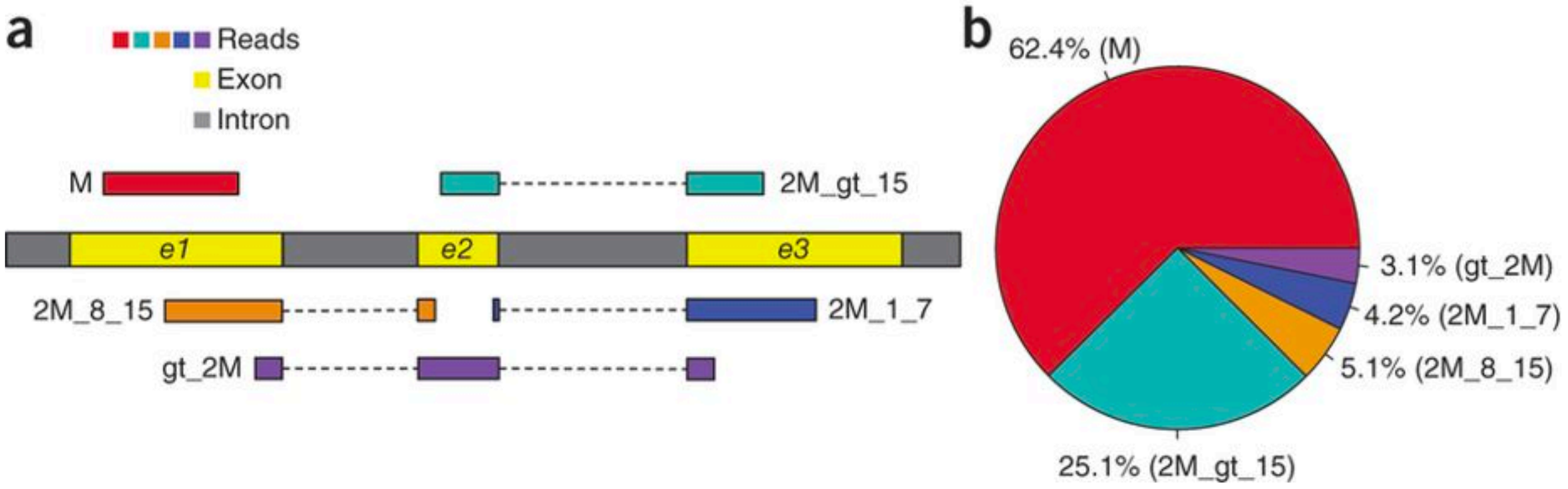
Genomics109

Logic



Genomics109

Classes of Reads



Genomics109

HISAT2, StringTie and Ballgown Information

PIPELINE:	HISAT2, StringTie and Ballgown
HISAT2:	Manual
	HISAT: a fast spliced aligner with low memory requirements
StringTie:	Manual
	StringTie enables improved reconstruction of a transcriptome from RNA-seq reads
	Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown
Ballgown:	Manual
	Documentation
	Ballgown bridges the gap between transcriptome assembly and expression analysis

Genomics109

STAR, StringTie and DESeq2 Information

PIPELINE:	STAR, StringTie and DESeq2
STAR:	Manual
	STAR: ultrafast universal RNA-seq aligner
StringTie:	Manual
	StringTie enables improved reconstruction of a transcriptome from RNA-seq reads
	Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown
DESeq2:	Manual
	Count-based differential expression analysis of RNA sequencing data using R and Bioconductor
	Analyzing RNA-seq data with DESeq2

Genomics109

Kallisto and Sleuth and Salmon Information

PIPELINE:	Kallisto and Sleuth and Salmon
Kallisto:	About
	Near-optimal probabilistic RNA-seq quantification
Sleuth:	About
	Differential analysis of RNA-seq incorporating quantification uncertainty
	A sleuth for RNA-Seq
	GITHUB
Salmon	About
	Manual
	GITHUB
	Salmon provides fast and bias-aware quantification of transcript expression

Genomics109

Blog Wars

Blog_Wars	How not to perform a differential expression analysis (or science)
	Response to the blog post about Salmon and kallisto
	Not-quite alignments: Salmon, Kallisto and Efficient Quantification of RNA-Seq data

Other Aligners

OTHER Aligners:	
GMAP:	Info
	GMAP: a genomic mapping and alignment program for mRNA and EST sequences
	Aligner tutorial: GMAP, STAR, BLAT, and BLASR

Genomics109

Evaluating Results

There's a new RNA-seq metric on the block…

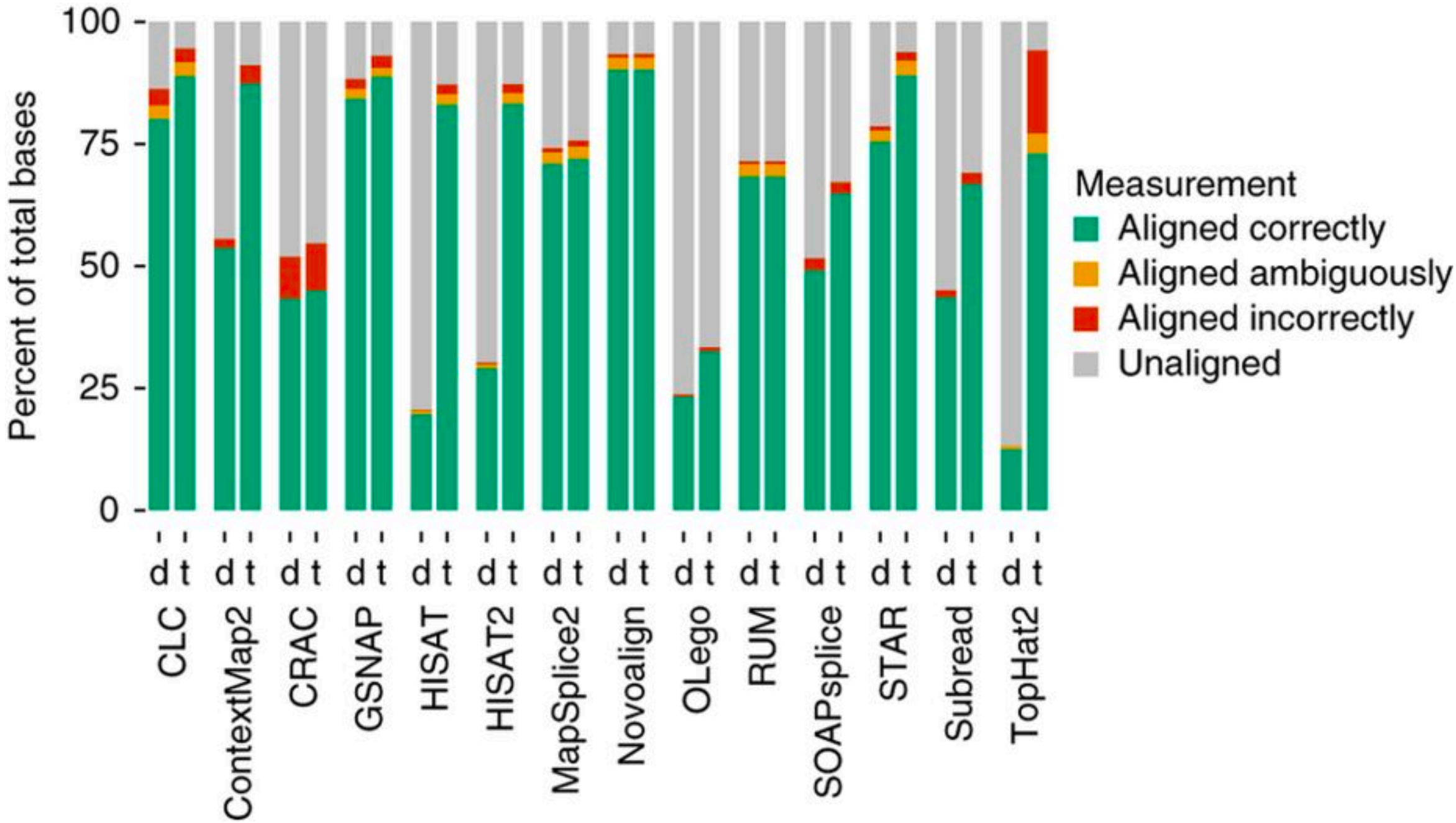
- We used to report RPKM (Reads Per Kilobase Million) or FPKM (Fragments Per Kilobase Million)
  - These normalized read counts for:
    - 1) The sequencing depth (that's the "Million" part)
      - Sequencing runs with more depth will have more reads mapping to each gene.
    - 2) The length of the gene (that's the "Kilobase" part)
      - Longer genes will have more reads mapping to them.
- Now they want us to use TPM – Transcripts per million

https://youtu.be/TTUrtCY2k-w

Genomics109

Final Considerations

Make sure to tune your parameters…Not just use defaults



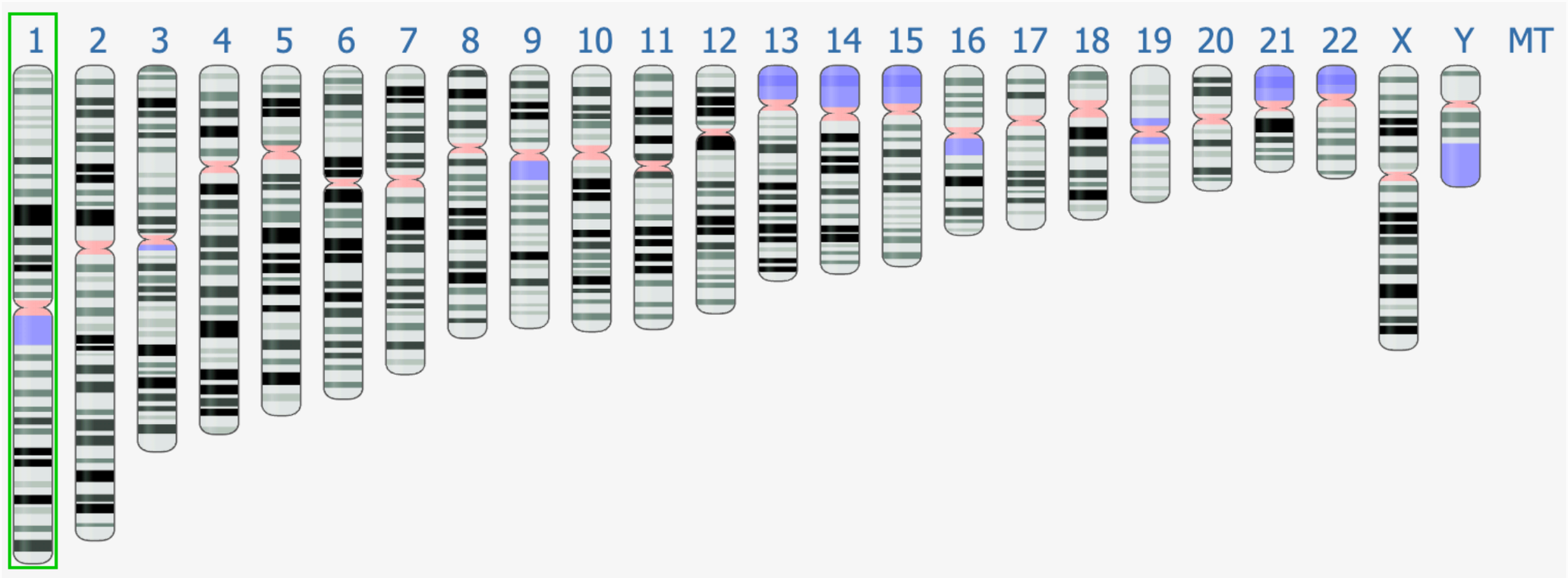
Genomics109

References

- *This lesson has been developed using materials from various sources, that include, but are not restricted to training tutorials developed by the Galaxy Project team. These are open access materials distributed under the terms of the Creative Commons Attribution license (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.*

- *Other Bibliographic References are:*

  - *RSeQC: quality control of RNA-seq experiments*
  - *Comprehensive comparative analysis of strand-specific RNA sequencing methods*
  - *Transcript assembly and quantification by RNA-seq..*
  - *Identification of novel transcripts in annotated genomes using…*
  - *Differential analysis of gene regulation at transcript…*
  - *TopHat: discovering splice junctions with RNA-Seq*
  - *TopHat-Fusion: an algorithm for discovery…*
  - *Differential gene and transcript expression analysis…*
  - *TopHat2: accurate alignment of transcriptomes…*
  - *HISAT: a fast spliced aligner with low memory…*
  - *Stringtie enables improved reconstruction…*
  - *Count-based differential expression analysis of RNA…*
  - *Transcript-level expression analysis of RNA-seq…*
  - *STAR: ultrafast universal RNA-seq aligner…*
  - *Near-optimal probabilistic RNA-Seq quantification…*
  - *Differential analysis of RNA-Seq…*
  - *Salmon provides fast…*
  - *GMAP: a genomic mapping and alignment…*
  - *Spatially resolved transcriptomics…*
  - *Simulation-Based omprehensive benchmarking…*
  - *A survey of best practices for RNA-seq data Analysis*



Genomics109

BIOL647
Digital Biology

Rodolfo Aramayo