

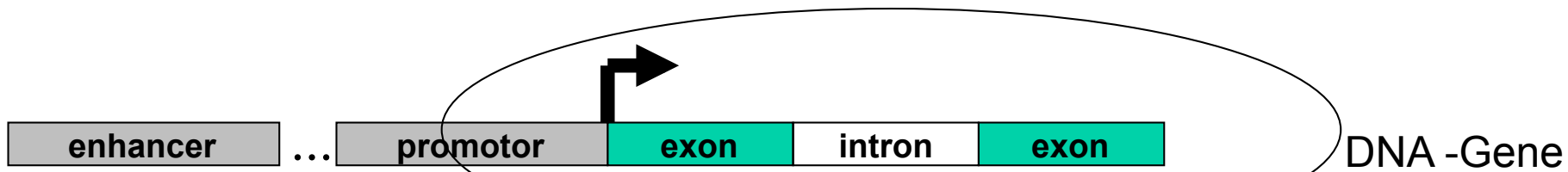
RNA-seq 2

transcriptome assembly

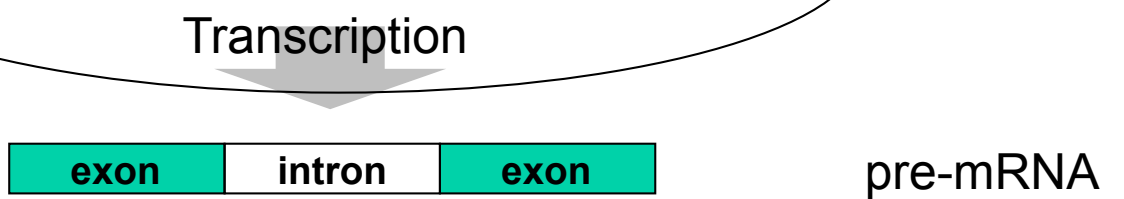
Roland Krause

Acknowledgments

- Ho-Ryun Chung
- Marcel Schulz



Where are the transcripts?



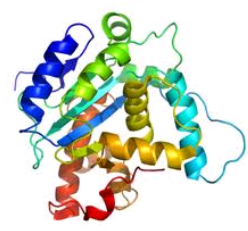
Capping
Splicing
Polyadenylation



Nuclear export

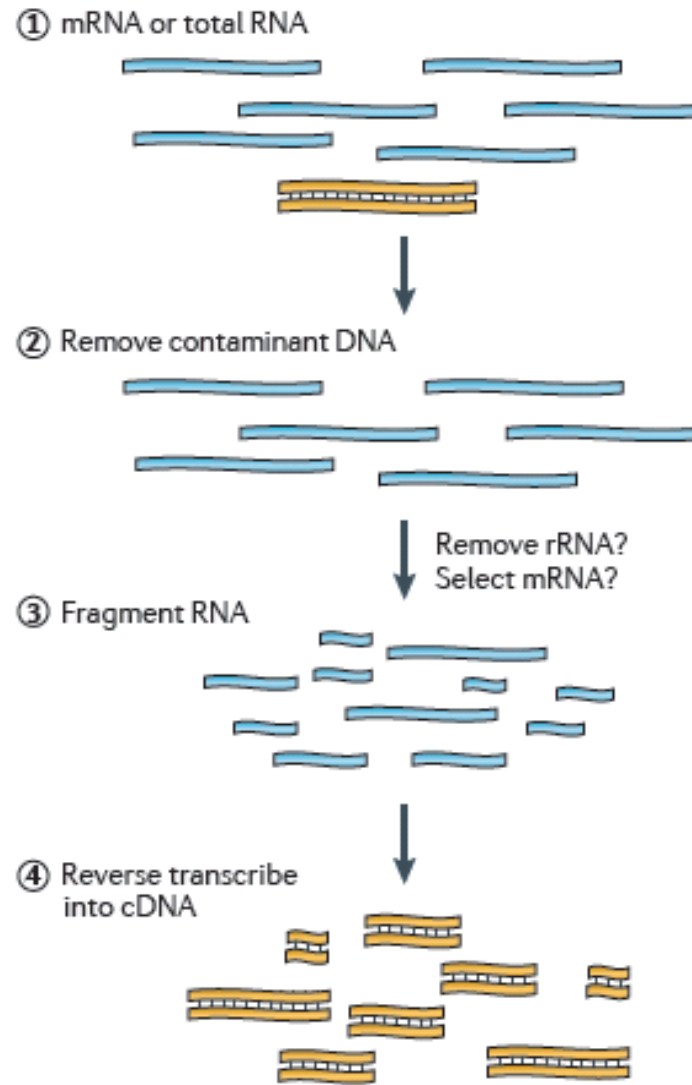


Translation

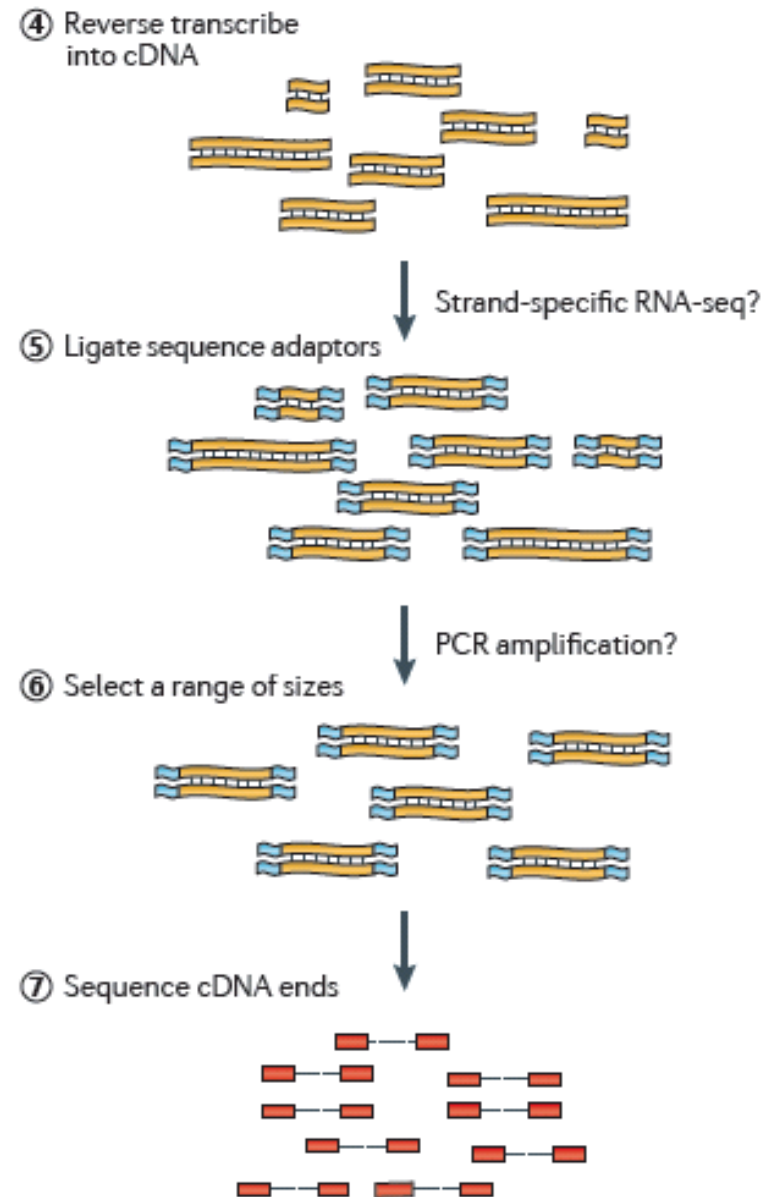


Protein

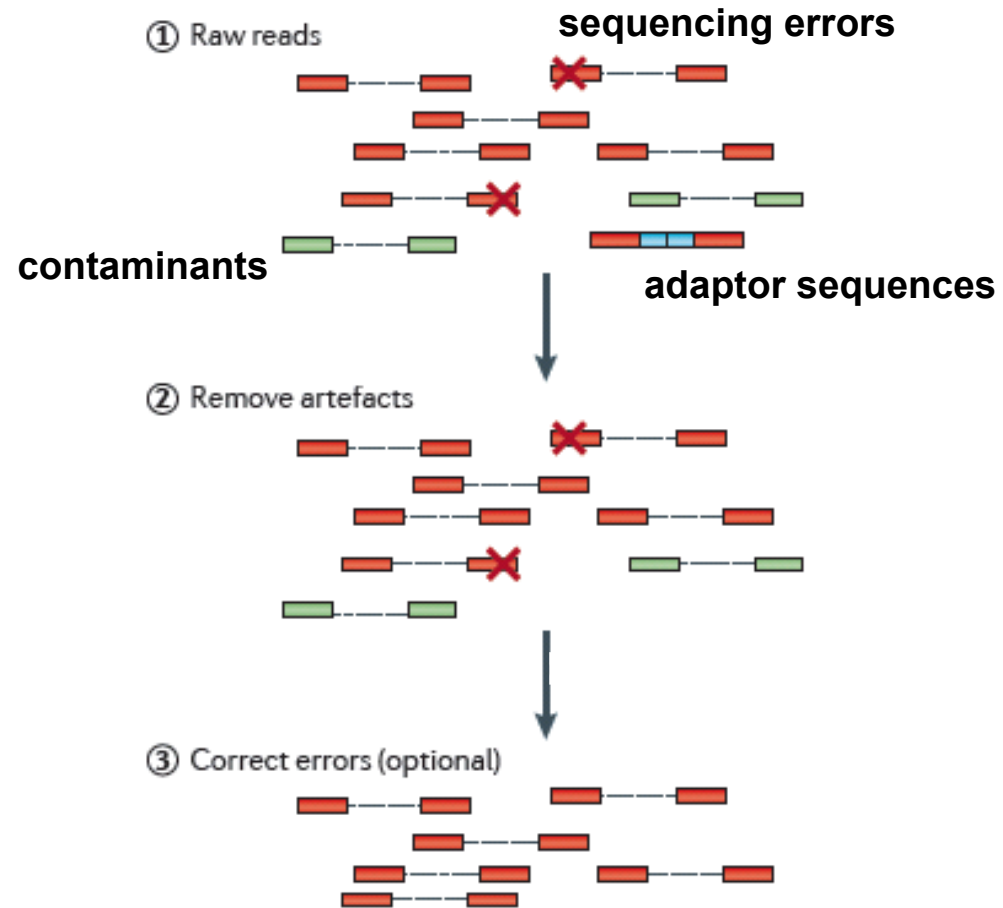
... data generation



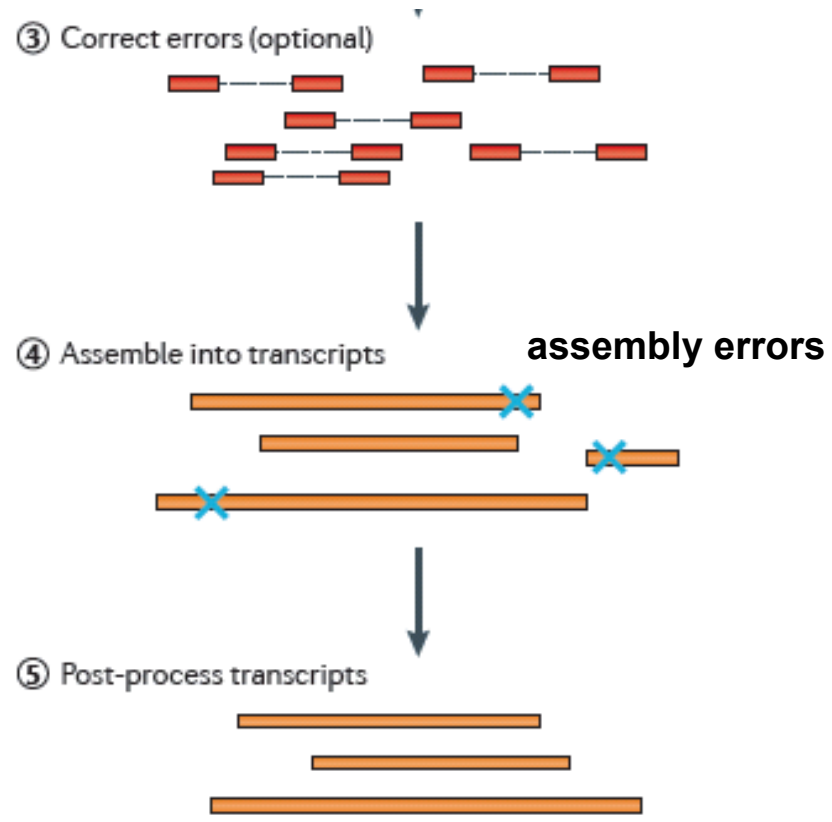
... data generation



... data analysis



... data analysis

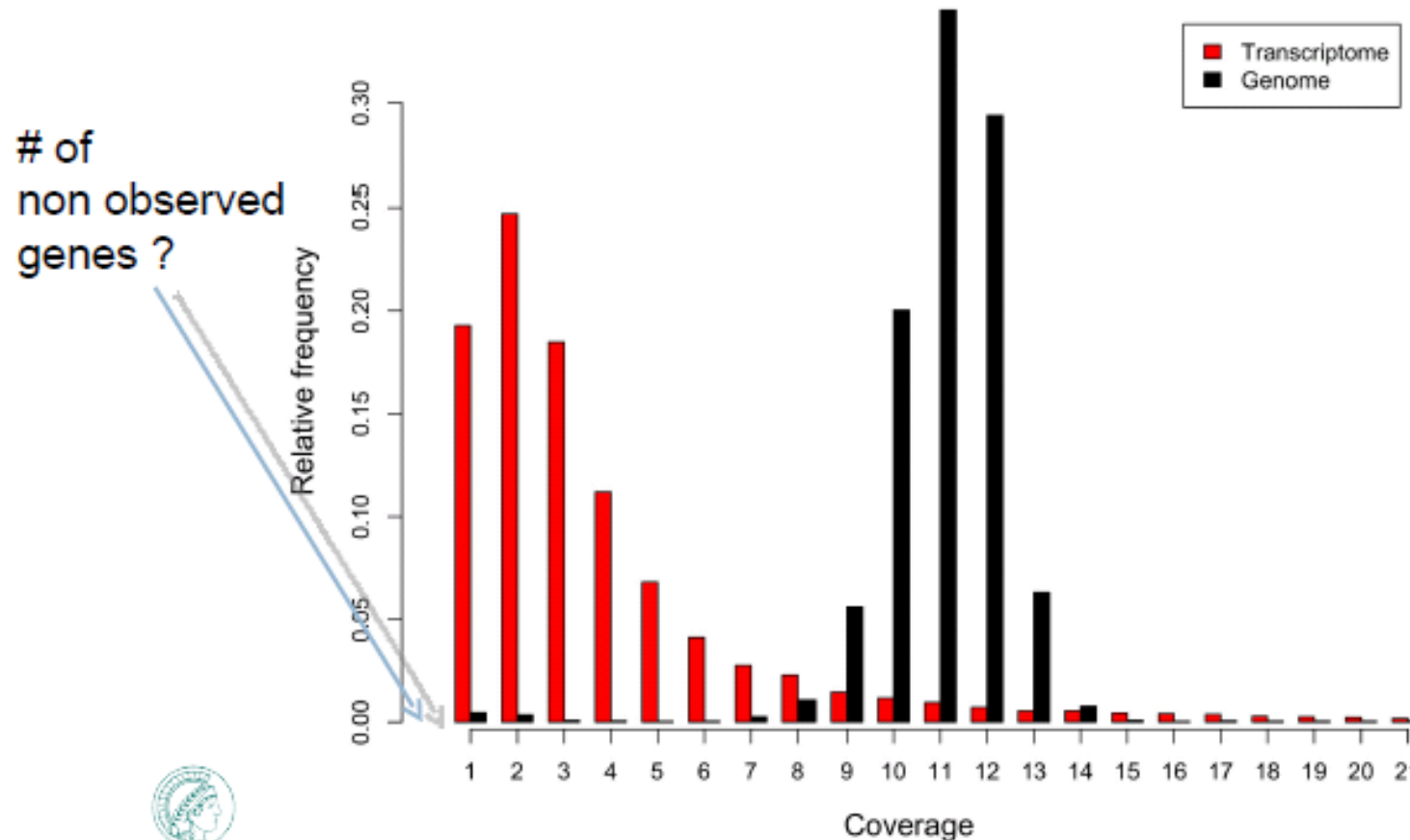


... challenges for transcriptome assembly

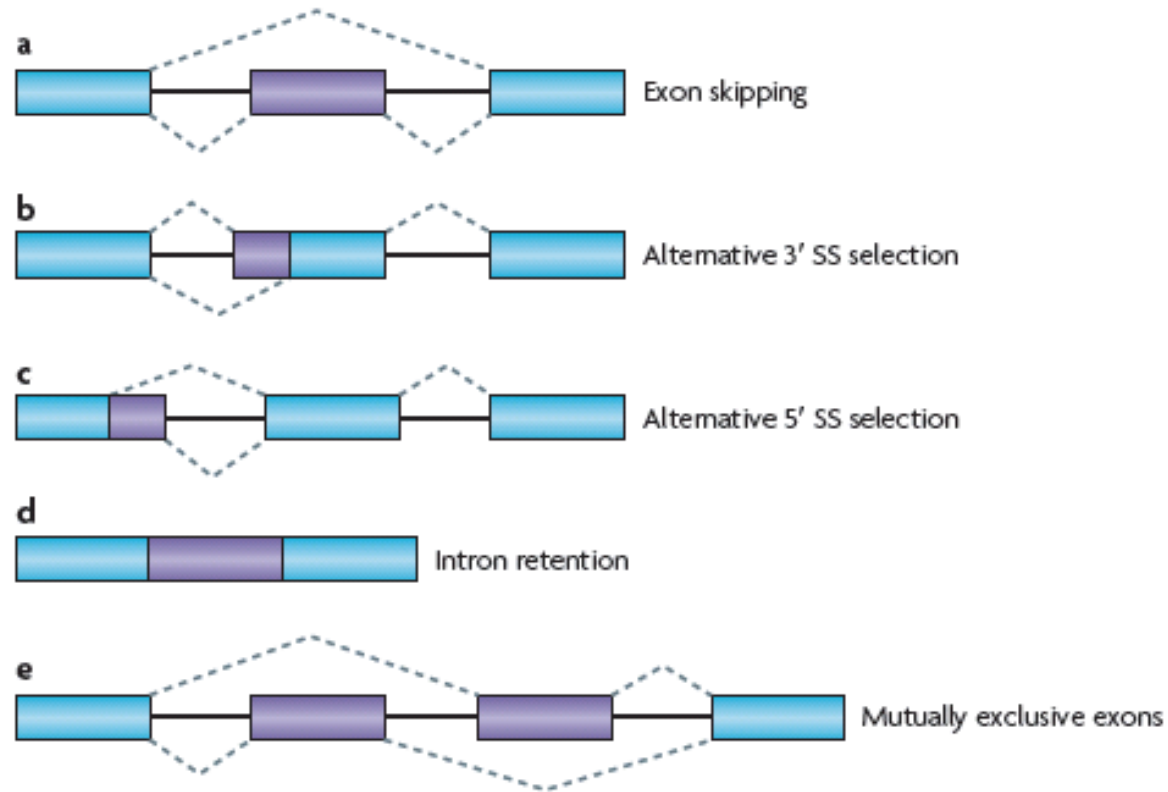
- highly non-uniform coverage
- alternative splicing
- alternative promoter usage
- alternative poly(A)

... highly non-uniform coverage

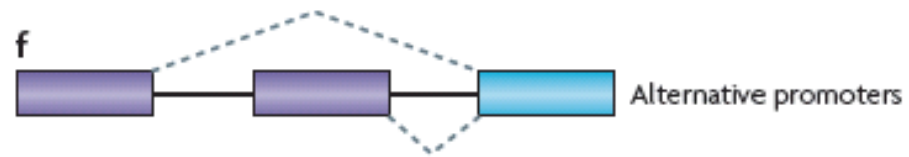
- RNA-Seq reads are distributed according to transcript expression levels.



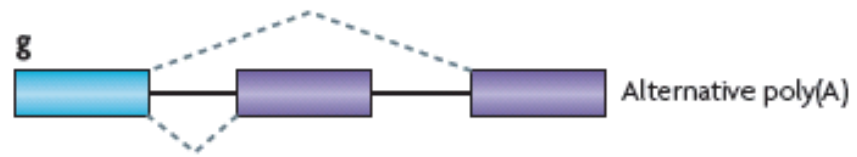
... alternative splicing



... alternative promoter



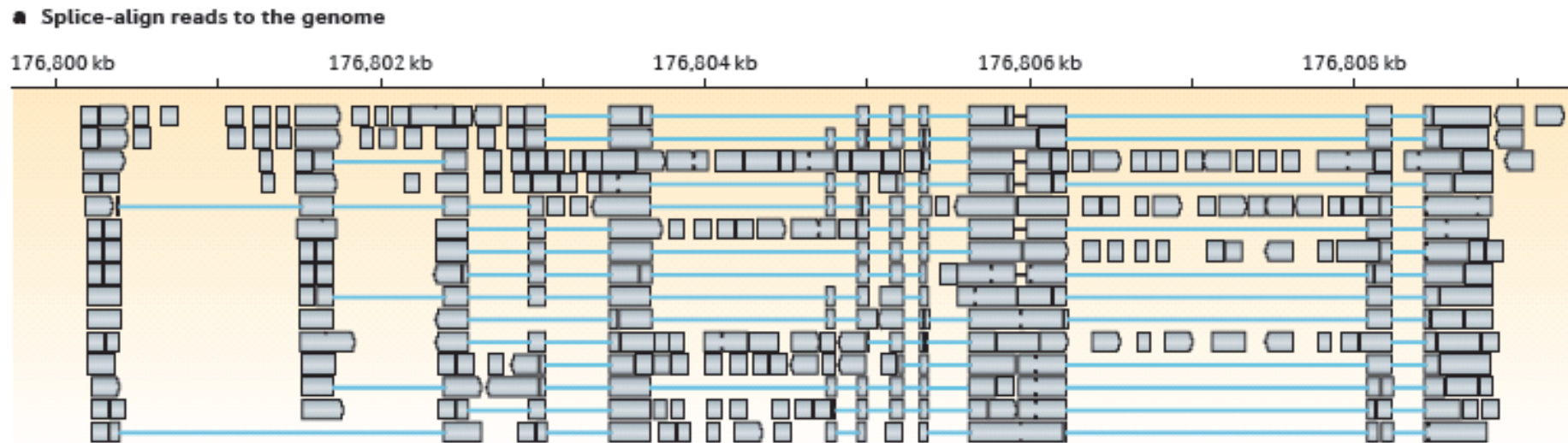
... alternative poly(A)



... transcriptome assembly strategies

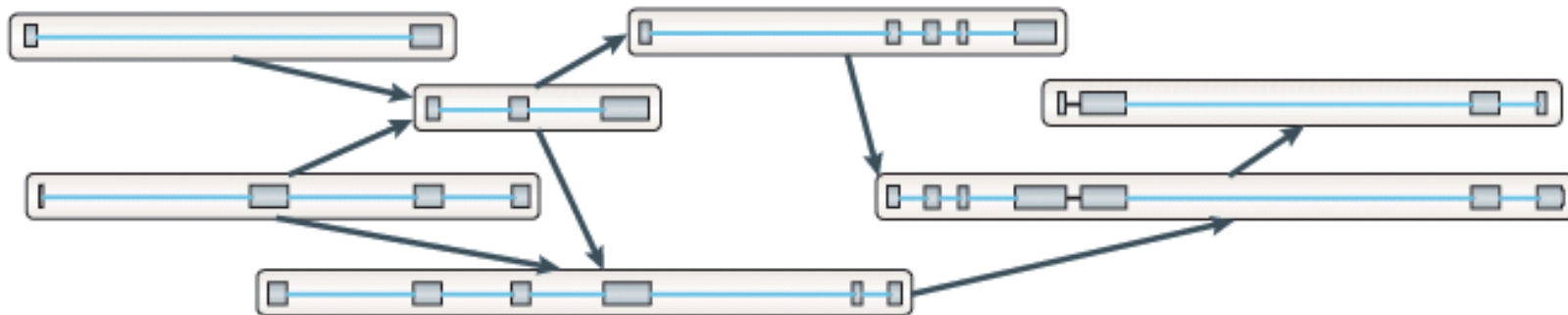
- reference-based
- *de novo*
- combination

... reference-based transcriptome assembly



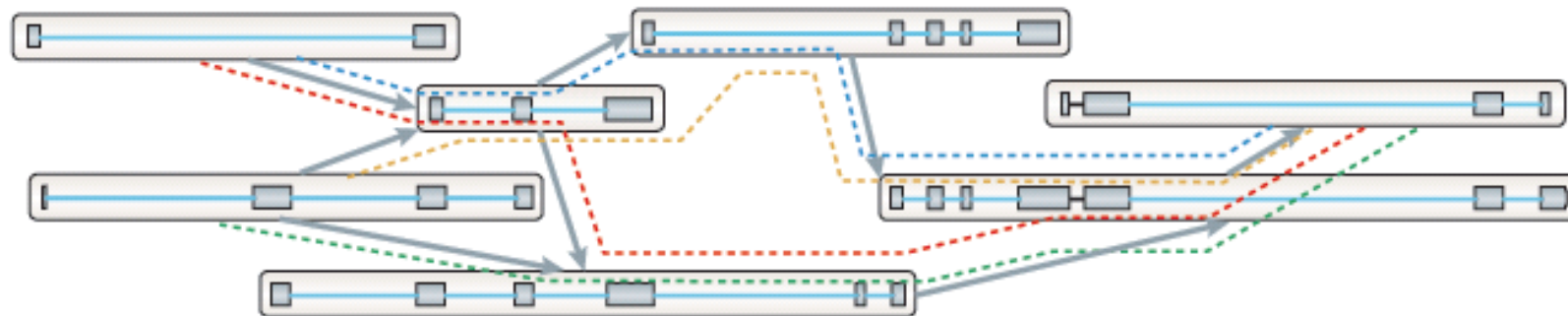
... reference-based transcriptome assembly

b Build a graph representing alternative splicing events



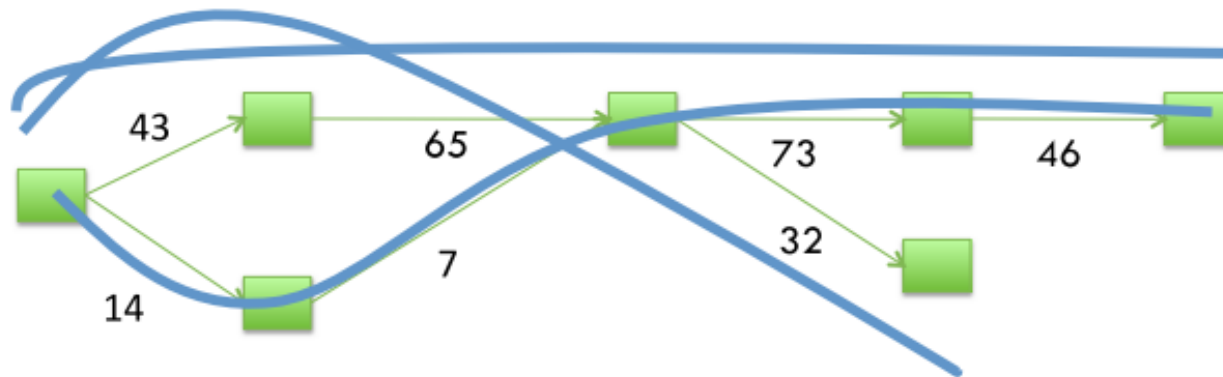
... reference-based transcriptome assembly

c Traverse the graph to assemble variants



... transcript prediction from splicing graphs

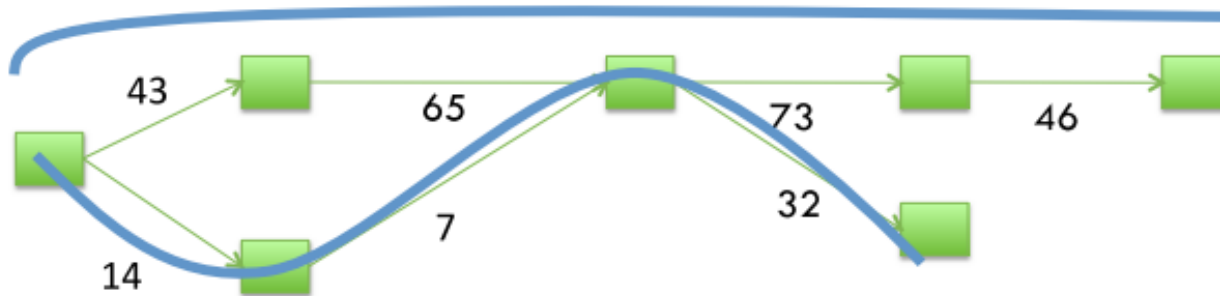
- maximum likelihood (heaviest path) approach



- assumptions:
 - AS events are independent
 - most exons are constitutive

... transcript prediction from splicing graphs

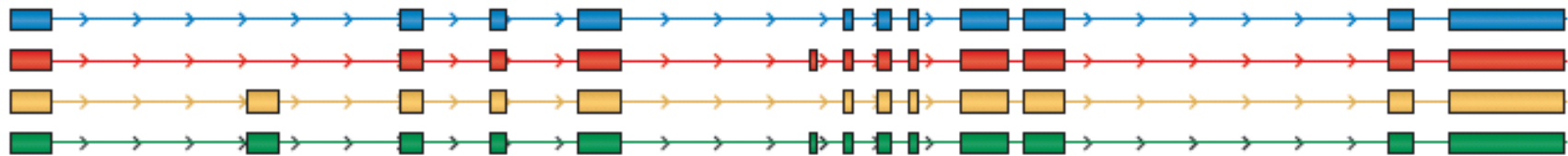
- maximum parsimony approach



- assumptions:
 - resources are sparse -> the cell is economical in the number of transcripts

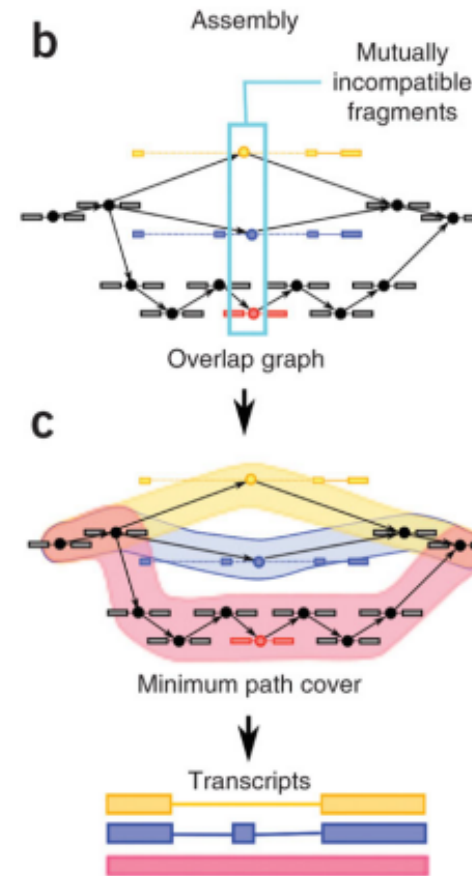
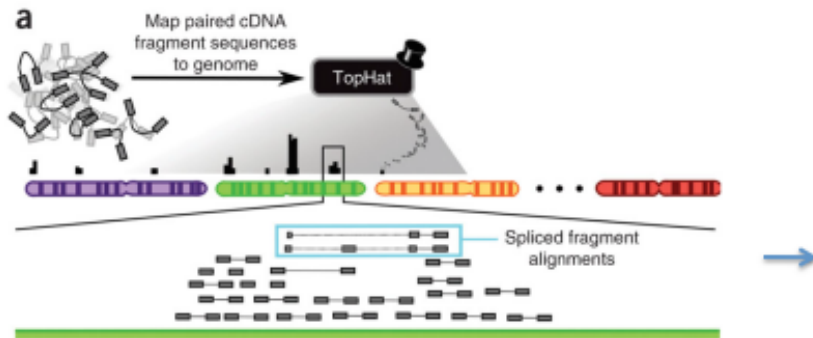
... reference-based transcriptome assembly

d Assembled isoforms



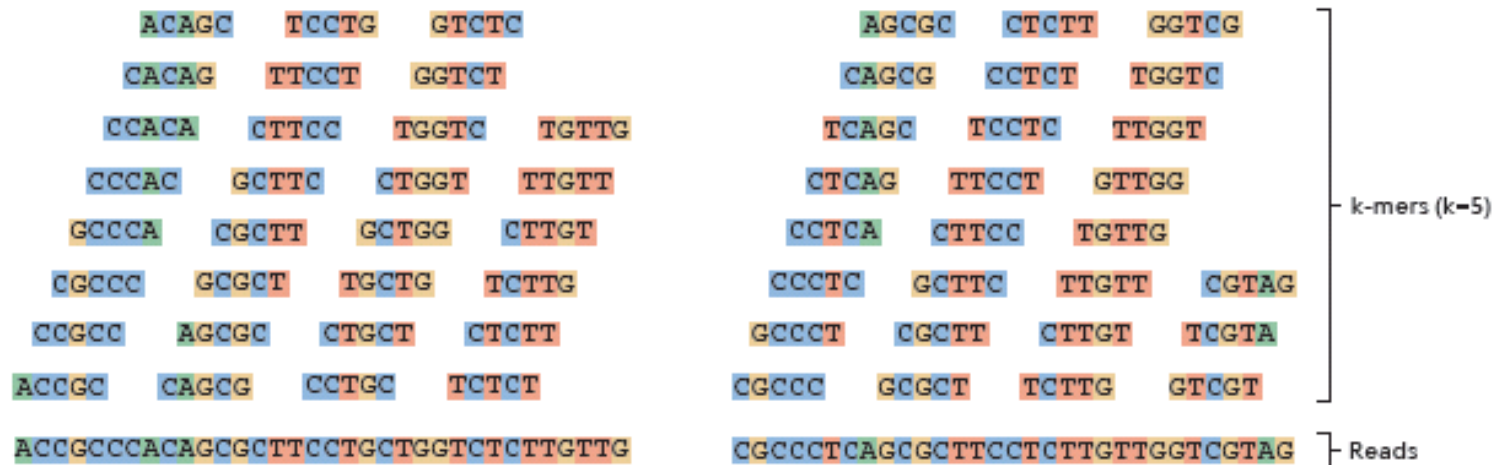
... for example: *cufflinks*

spliced alignment of paired-end reads to genome



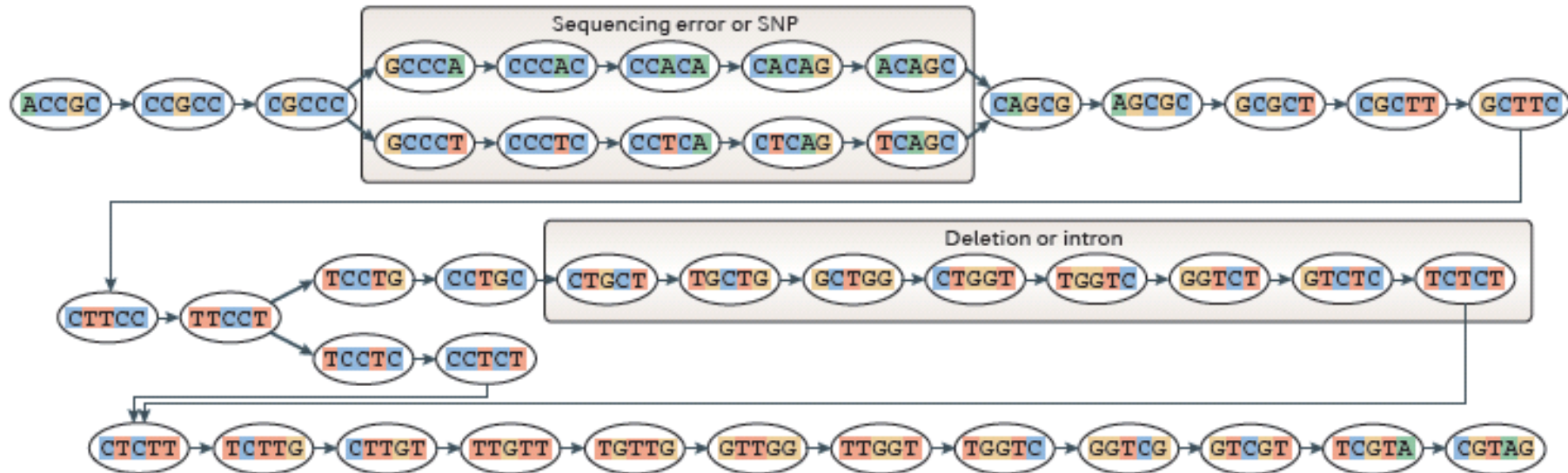
... *de novo* transcriptome assembly

- Generate all substrings of length k from the reads



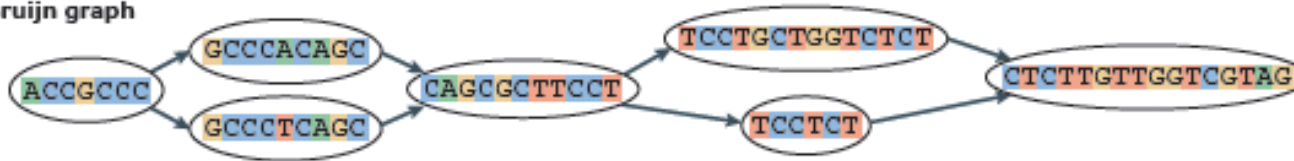
... *de novo* transcriptome assembly

b Generate the De Bruijn graph

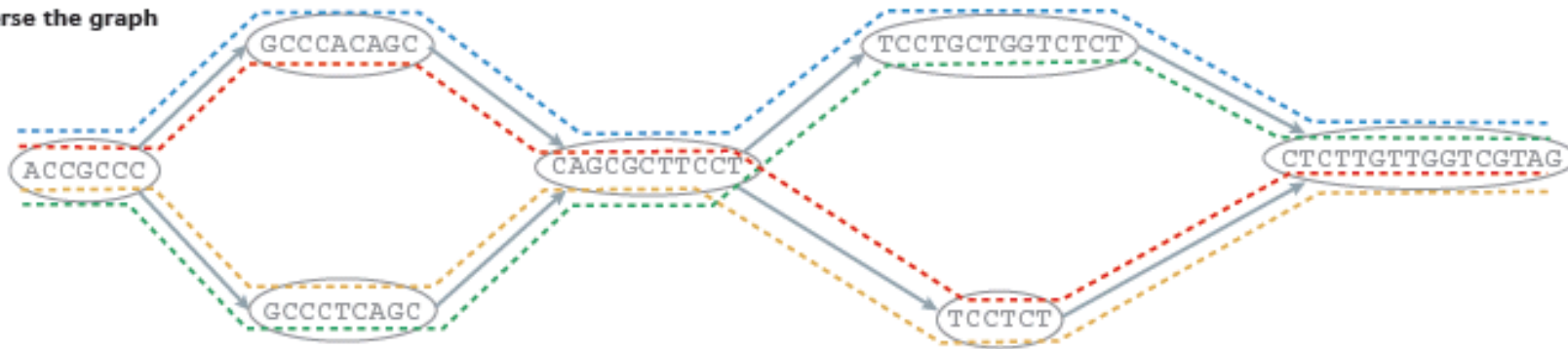


... *de novo* transcriptome assembly

c Collapse the De Bruijn graph



d Traverse the graph

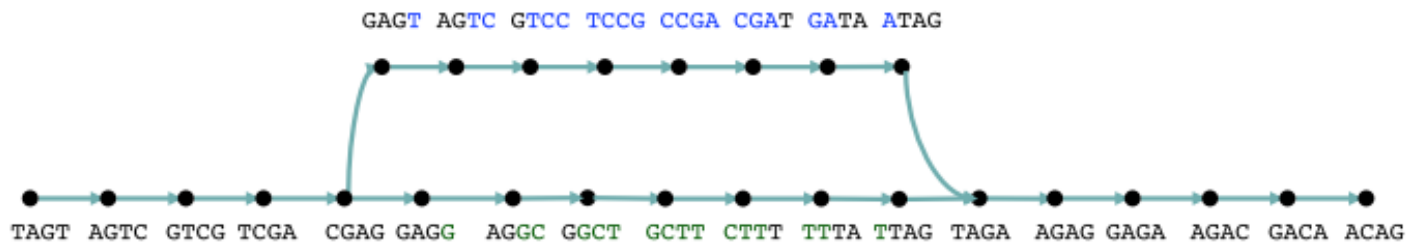


Assembled isoforms

```

----- ACCGCCCACAGCGCTTCCTGCTGGTCTCTTGGTTGGTCGTAG
----- ACCGCCCACAGCGCTTCCT-----CTTGGTTGGTCGTAG
----- ACCGCCCTCAGCGCTTCCT-----CTTGGTTGGTCGTAG
----- ACCGCCCTCAGCGCTTCCTGCTGGTCTCTTGGTTGGTCGTAG
  
```

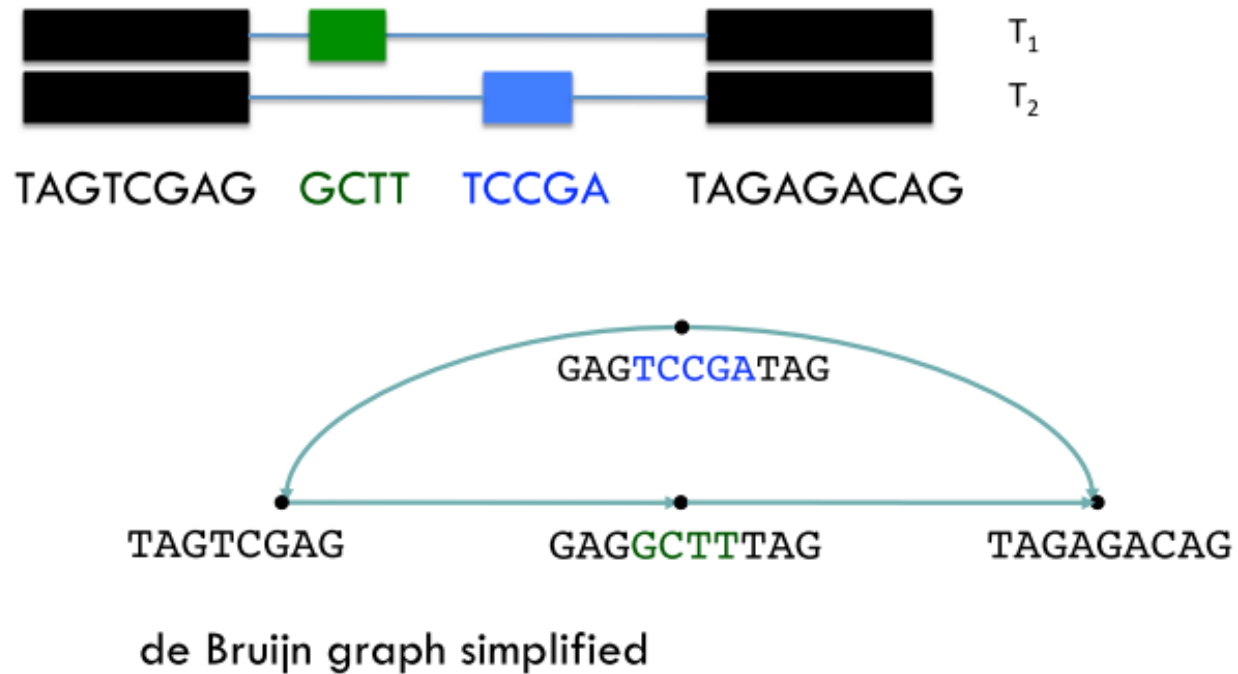
... from de Bruijn graph to splicing graph



de Bruijn graph of T_1 and T_2 with $k = 4$

Heber et al. Bioinformatics, 2001

... from de Bruijn graph to splicing graph

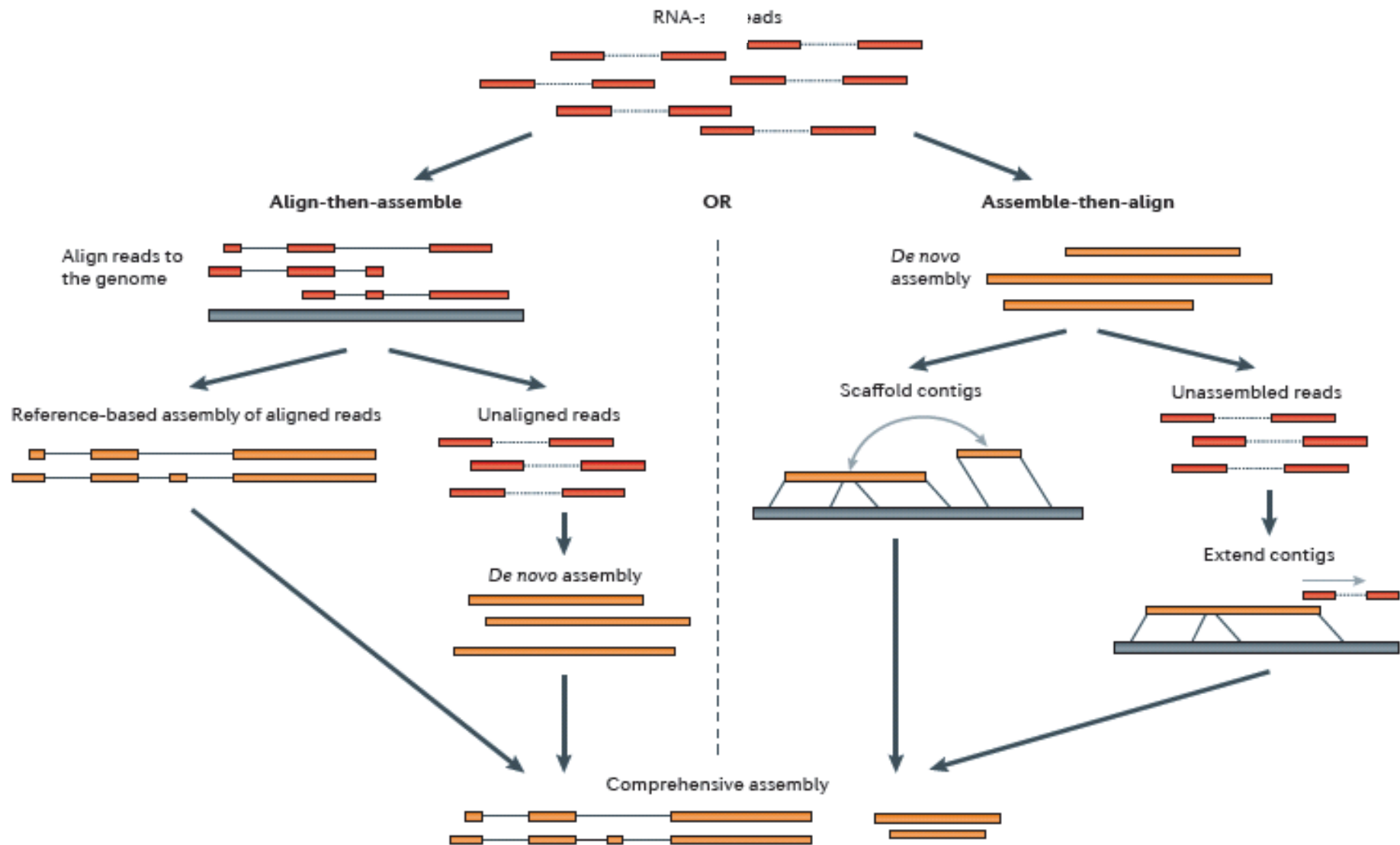


Heber et al. Bioinformatics, 2001

... for example Oases

- Assemble reads using Velvet
- Create clusters of contigs:
 - Connecting reads
 - Connecting read pairs
- Re-implement traditional algorithms to run on graphs, instead of a reference genome
 - Greedy transitive reduction (Myers, 2005)
 - Motif searches
 - Dynamic assembly of transcripts (Lee, 2003)

... combination



... reference-based vs *de novo*

Approach	Advantages	Disadvantages
<i>ab initio</i> reference based	<ul style="list-style-type: none">-alignment tolerates seq. errors-repeats are detected through alignment-grouping by genomic proximity	<ul style="list-style-type: none">-reference seq. needed-assumes transcripts are collinear with the genome
<i>de novo</i>	<ul style="list-style-type: none">-no reference needed-detection of non-collinear transcripts (cancer, trans-splicing)-handling of micro-exons (~ 25bp)	<ul style="list-style-type: none">-lowly expressed genes indistinguishable from seq. errors-missassemblies due to repeats

... *transcriptome assembly tools*

Assembler	De novo?	Parallelism	Support for paired-end reads?	Support for stranded reads?	Support for multiple insert sizes?	Outputs transcript counts?	Software availability	Refs
G-Mo.R-Se	No	None	No	No	No	No	http://www.genoscope.cns.fr/externe/gmorse/	17
Cufflinks	No	MP	Yes	Yes	Yes	Yes	http://cufflinks.cbc.umd.edu/	20
Scripture	No	None	Yes	Yes	Yes	Yes	http://www.broadinstitute.org/software/scripture/	16
ERANGE	No	None	Yes	Yes	Yes	Yes	http://woldlab.caltech.edu/rnaseq	50
Multiple-k	Yes	None	Yes	Yes	Yes	No	http://www.surget-groba.ch/downloads/	19
Rnnotator	Yes	MP	Yes	Yes	Yes	Yes	Contact David Gilbert (DEGilbert@lbl.gov)	15
Trans-ABYSS	Yes	MPI	Yes	No	Yes	Yes	http://www.bcgsc.ca/platform/bioinfo/software/trans-abyss	18
Oases	Yes	MP	Yes	Yes	Yes	No	http://www.ebi.ac.uk/~zerbino/oases/	-
Trinity	Yes	MP	Yes	Yes	No	Yes	http://trinityrnaseq.sourceforge.net/	59

... *transcriptome assembly tools*

Assembler	De novo?	Parallelism	Support for paired-end reads?	Support for stranded reads?	Support for multiple insert sizes?	Outputs transcript counts?	Software availability	Refs
G-Mo.R-Se	No	None	No	No	No	No	http://www.genoscope.cns.fr/externe/gmorse/	17
Cufflinks	No	MP	Yes	Yes	Yes	Yes	http://cufflinks.cbc.umd.edu/	20
Scripture	No	None	Yes	Yes	Yes	Yes	http://www.broadinstitute.org/software/scripture/	16
ERANGE	No	None	Yes	Yes	Yes	Yes	http://woldlab.caltech.edu/rnaseq	50
Multiple-k	Yes	None	Yes	Yes	Yes	No	http://www.surget-groba.ch/downloads/	19
Rnnotator	Yes	MP	Yes	Yes	Yes	Yes	Contact David Gilbert (DEGilbert@lbl.gov)	15
Trans-ABYSS	Yes	MPI	Yes	No	Yes	Yes	http://www.bcgsc.ca/platform/bioinfo/software/trans-abyss	18
Oases	Yes	MP	Yes	Yes	Yes	No	http://www.ebi.ac.uk/~zerbino/oases/	-
Trinity	Yes	MP	Yes	Yes	No	Yes	http://trinityrnaseq.sourceforge.net/	59

... reference-based vs *de novo*

Method	transfrag > 100 bps	N50	total in mb	Nucleotide Sensitivity			Nucleotide Specificity
				all genes	lowly exp. genes RPKM < 1	highly exp. genes RPKM > 20	
Cufflinks*	72,745	2,613	~ 73	45.3	24.1	71.4	67
Oases [§]	73,357	1,287	~ 64	28.3	0.9	64.8	85.3

*Trapnell *et al.* *Nat. Biotech.* 2010

[§]Schulz, Zerbino *et. al.*, submitted

Results for Ensembl 57 annotation

Success stories

- New genes in low coverage genomes
 - Cufflinks: 649 new rice genes
- New splice forms
 - Alzheimer related splicing
- Comparison by the authors of *Trinity*
 - For mice: assembly-based outperform *de novo*
 - In yeast: vice versa

Accuracy

The accuracy metric is defined as the percentage of the correctly assembled bases estimated using the set of expressed reference transcripts (N). If reference transcripts are not available, then the reference genome can be used as an alternative. Accuracy can be formally written as:

$$\text{Accuracy} = 100 \times \frac{\sum_{i=1}^M A_i}{\sum_{i=1}^M L_i} \quad (1)$$

where L_i is the length of alignment between a reference transcript and an assembled transcript T_i , A_i is the correct bases in transcript T_i , and M represents the number of best alignments between assembled transcripts and reference.

Completeness

The completeness metric is defined as the percentage of expressed reference transcripts covered by all the assembled transcripts and is written as:

$$\text{Completeness} = 100 \times \frac{\sum_{i=1}^N I(C_i \geq \delta)}{N} \quad (2)$$

where the indicator function, I , represents whether (1) or not (0) C_i (the percentage of a reference transcript, i , that is covered by assembled transcripts) is greater than some arbitrary threshold, δ : for example, 80%.

Contiguity

The contiguity metric is defined as the percentage of expressed reference transcripts covered by a single, longest-assembled transcript and is similarly written as:

$$\text{Contiguity} = 100 \times \frac{\sum_{i=1}^N I(C_i \geq \delta)}{N} \quad (3)$$

where the indicator function, I , represents whether (1) or not (0) C_i (the percentage of a reference transcript, i , that is covered by a single, longest-assembled transcript) is greater than some arbitrary threshold, δ : for example, 80%.

Chimerism

The percentage of chimaeras that occur owing to misassemblies among all of the assembled transcripts. A chimeric transcript is one that contains non-repetitive parts from two or more different reference genes. They can arise from biological sources (gene fusions or *trans*-splicing), experimental sources (intermolecular ligation) or informatics sources (misassemblies). Misassembled chimeric transcripts can be distinguished from true chimaeras by determining whether the number of reads spanning the chimeric junction is significant when compared to the number of reads spanning other segments of the transcript.

Variant resolution

The percentage of transcript variants assembled. This can be calculated by the average of the percentage of assembled variants within the reference set as:

$$\text{Variants} = 100 \times \frac{\sum_{i=1}^N \frac{\max((C_i - E_i), 0)}{V_i}}{N} \quad (4)$$

where C_i and E_i are the number of correctly and incorrectly assembled variants for reference gene i , respectively, and V_i is the total number of variants for i .