bioinformatics.ca

# Canadian Bioinformatics Workshops

www.bioinformatics.ca

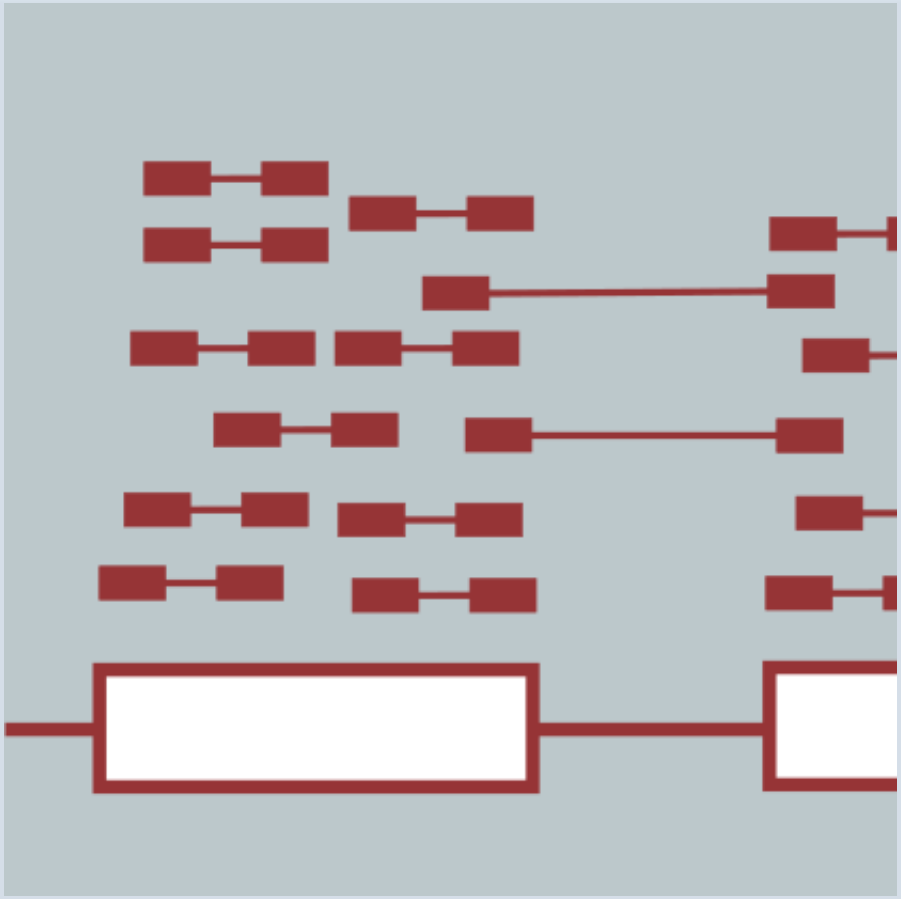
# Genome-Guided and Genome-Free Transcriptome Assembly

Brian Haas

Informatics for RNA-Seq Analysis

May 28-30, 2018

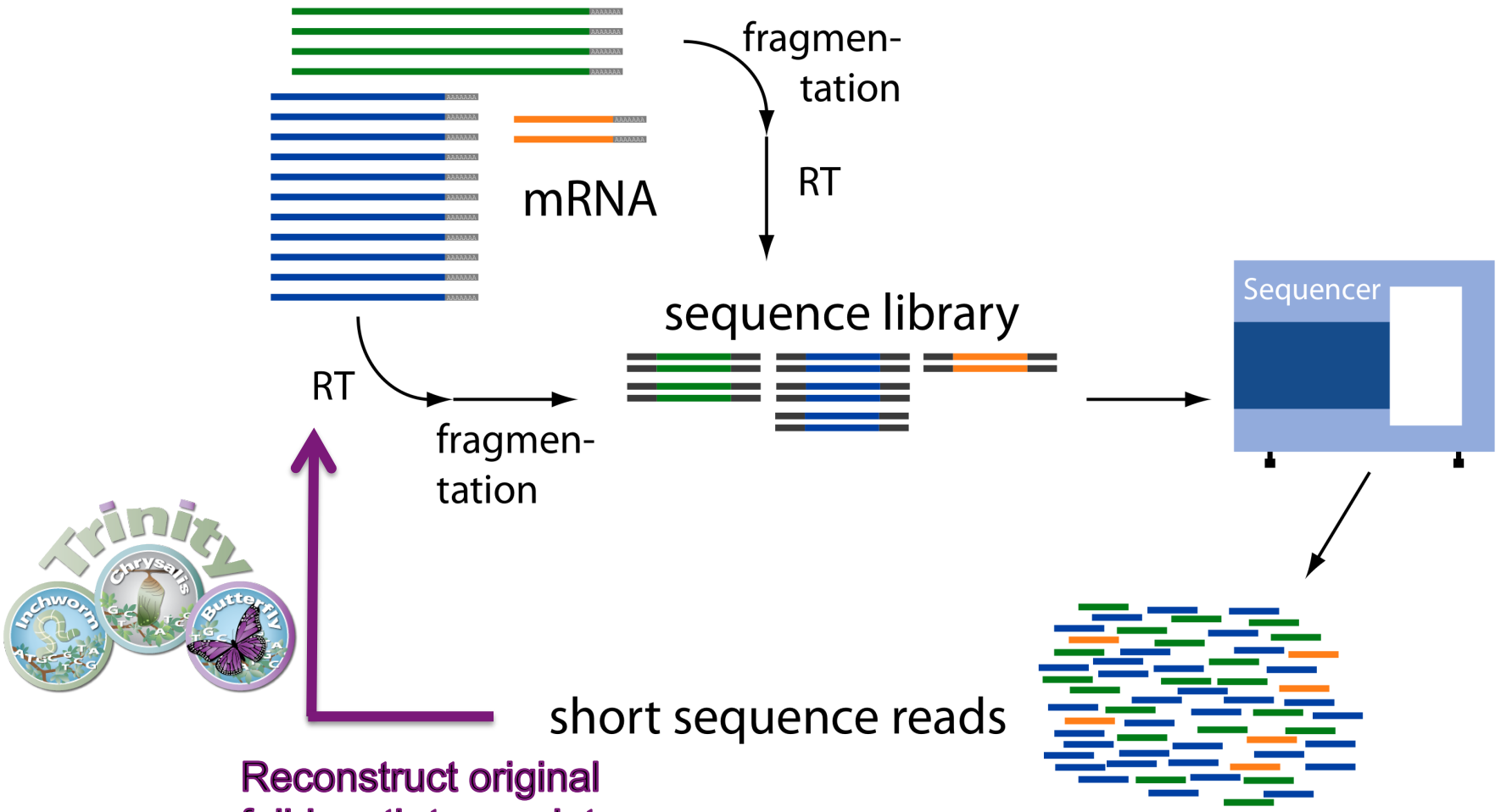
bioinformatics.ca



# Learning Objectives of Module

- Understand the challenges involved in reconstructing transcripts from RNA-Seq data
- Become familiar with computational algorithms and data structures leveraged for transcript assembly
- Appreciate the importance of strand-specific RNA-Seq data for transcript reconstruction
- Differentiate between differential gene expression and differential transcript usage.

# Assembly Required




Adapted from G. Raetsch

# Transcript Reconstruction from RNA-Seq Reads


## Advancing RNA-Seq analysis

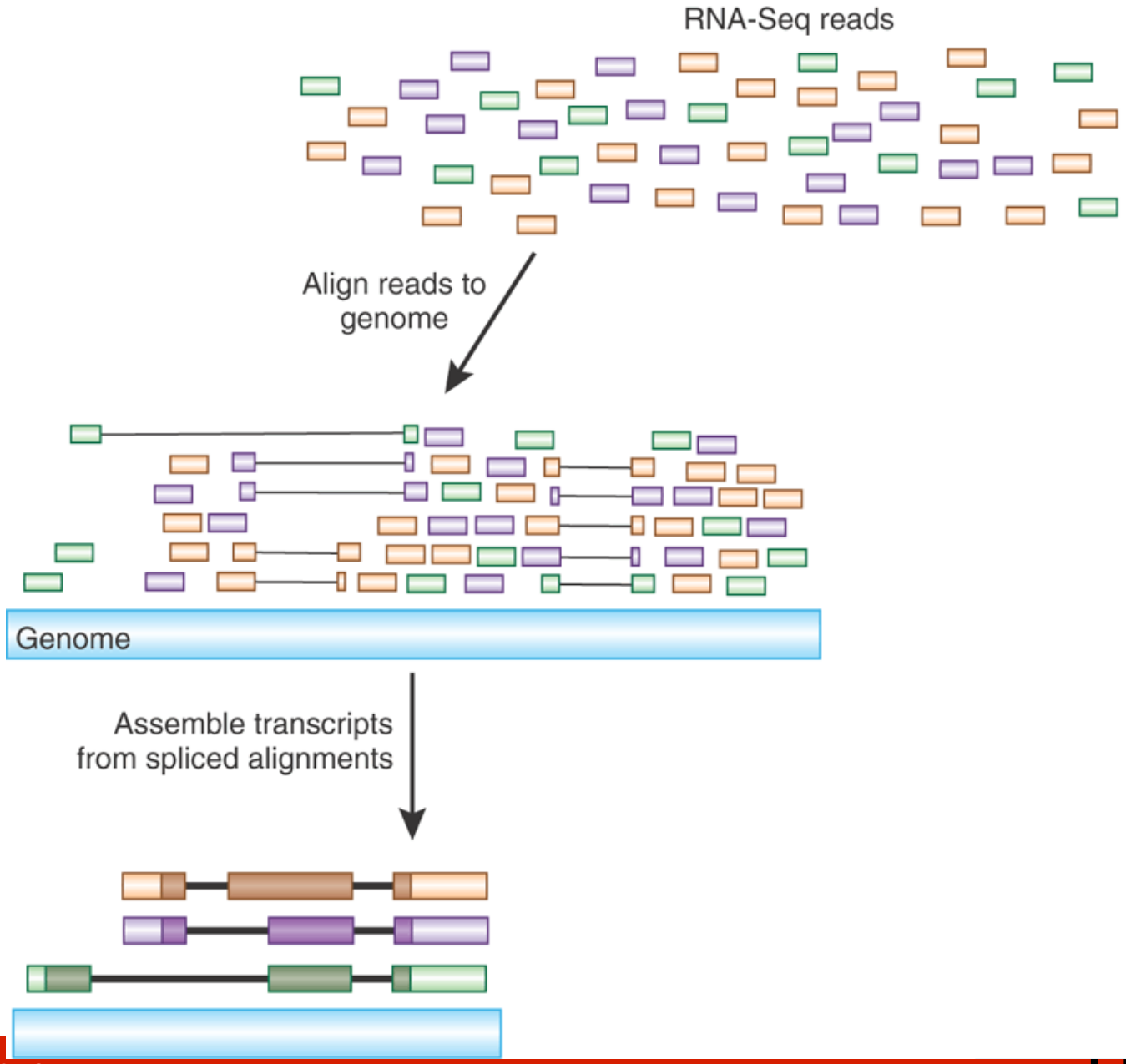
Brian J Haas & Michael C Zody

Nature Biotech, 2010

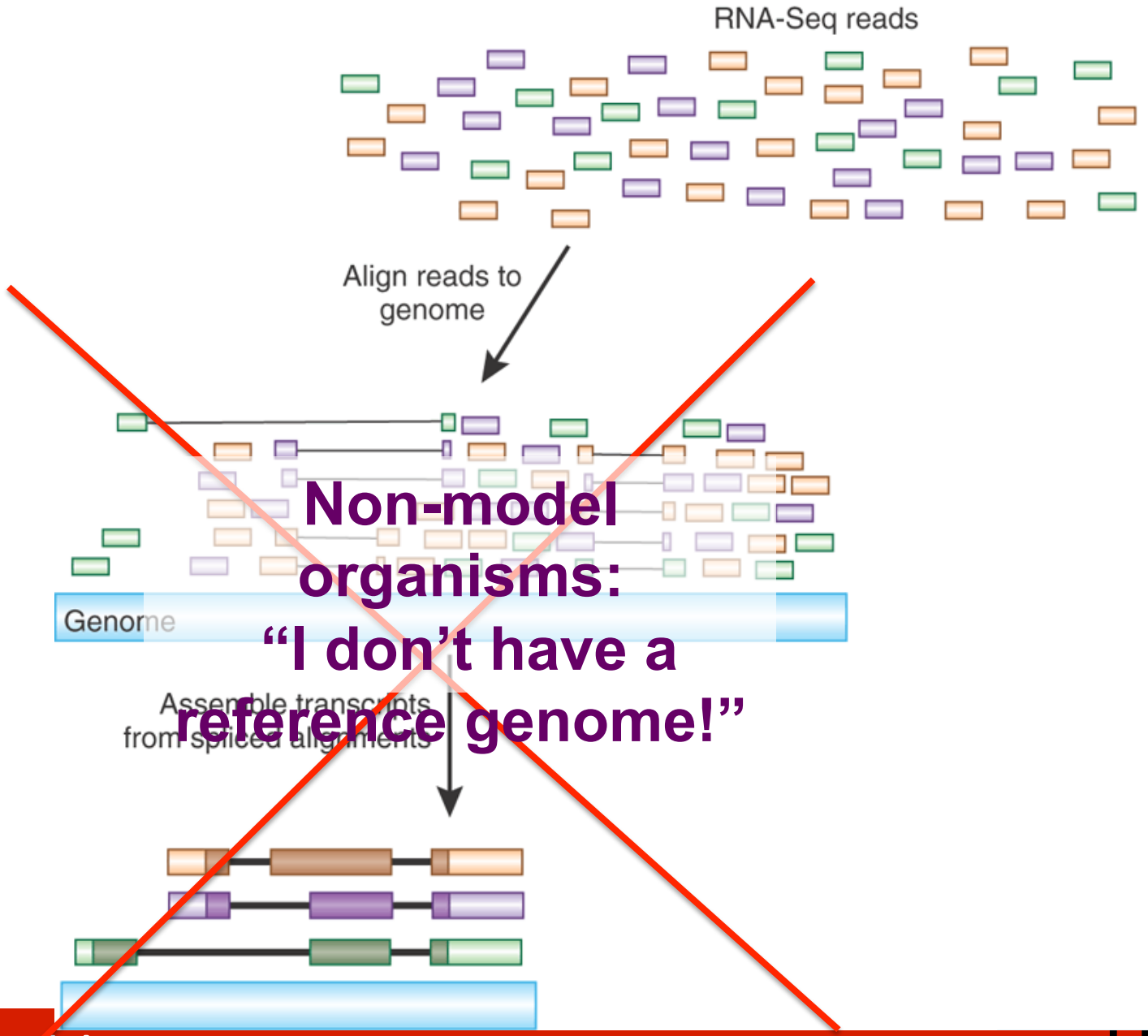
**New methods for analyzing RNA-Seq data enable *de novo* reconstruction of the transcriptome.**

# Transcript Reconstruction from RNA-Seq Reads

# Transcript Reconstruction from RNA-Seq Reads



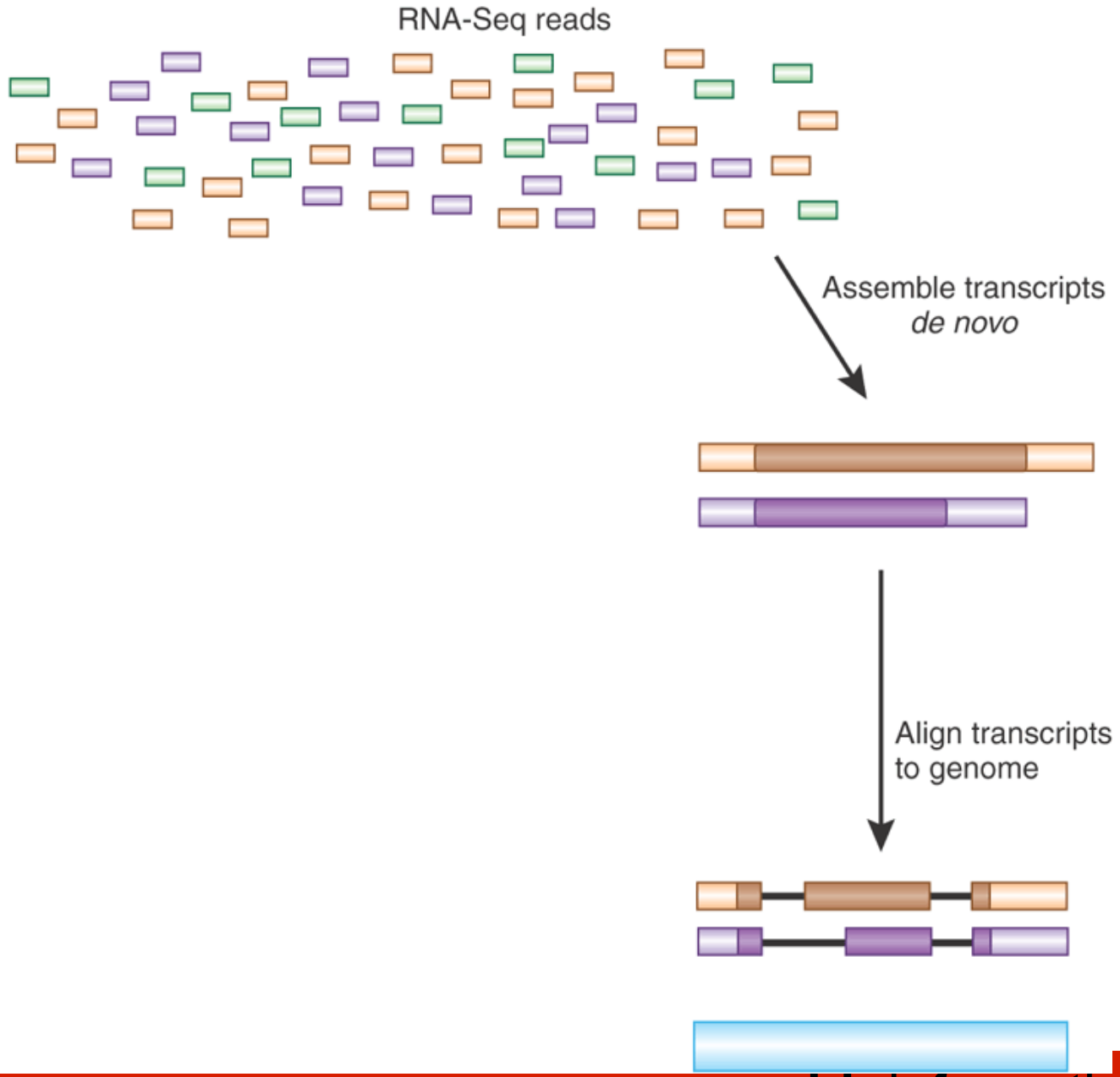
# Transcript Reconstruction from RNA-Seq Reads



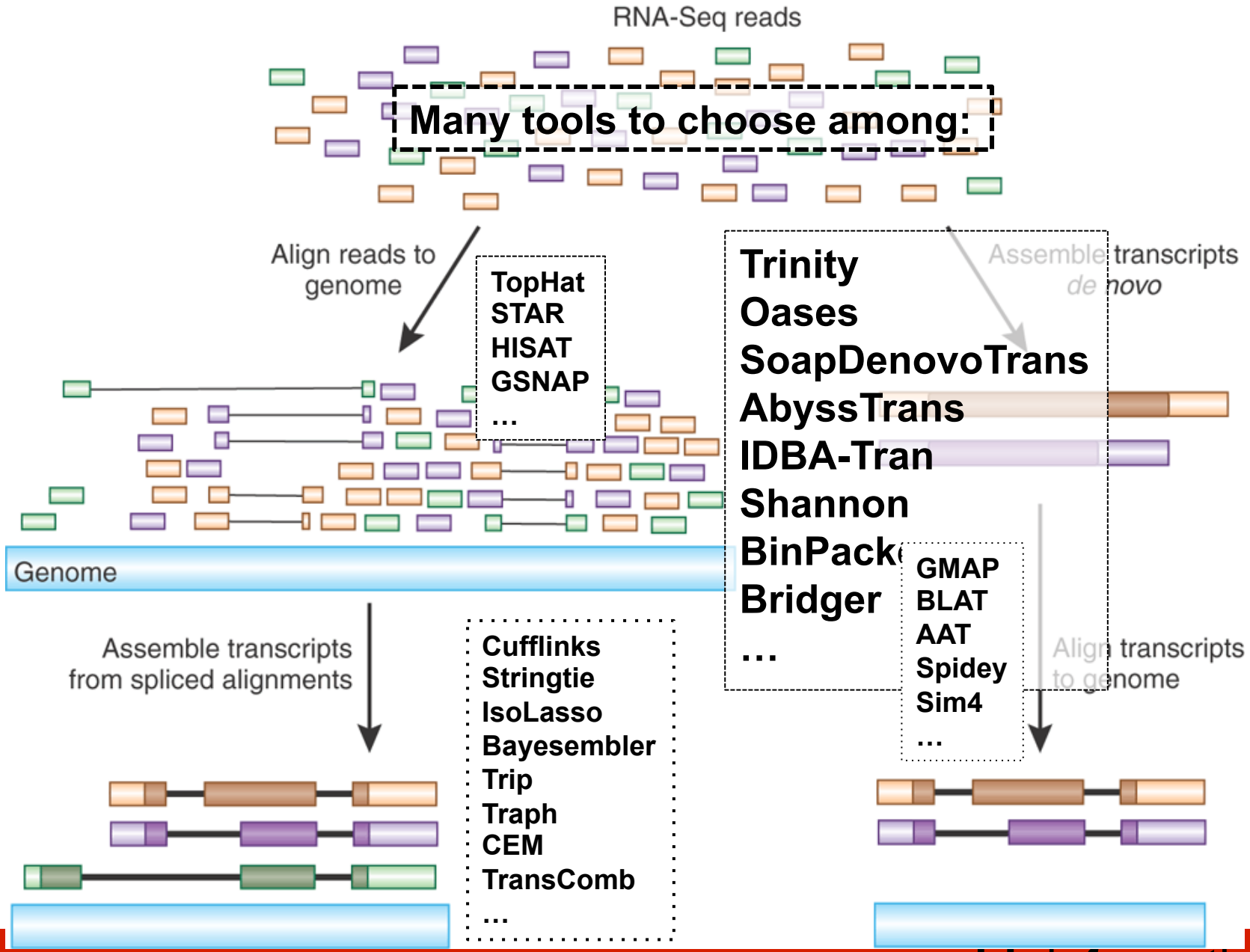

**Non-model organisms: "I don't have a reference genome!"**

# Transcript Reconstruction from RNA-Seq Reads

# Transcript Reconstruction from RNA-Seq Reads



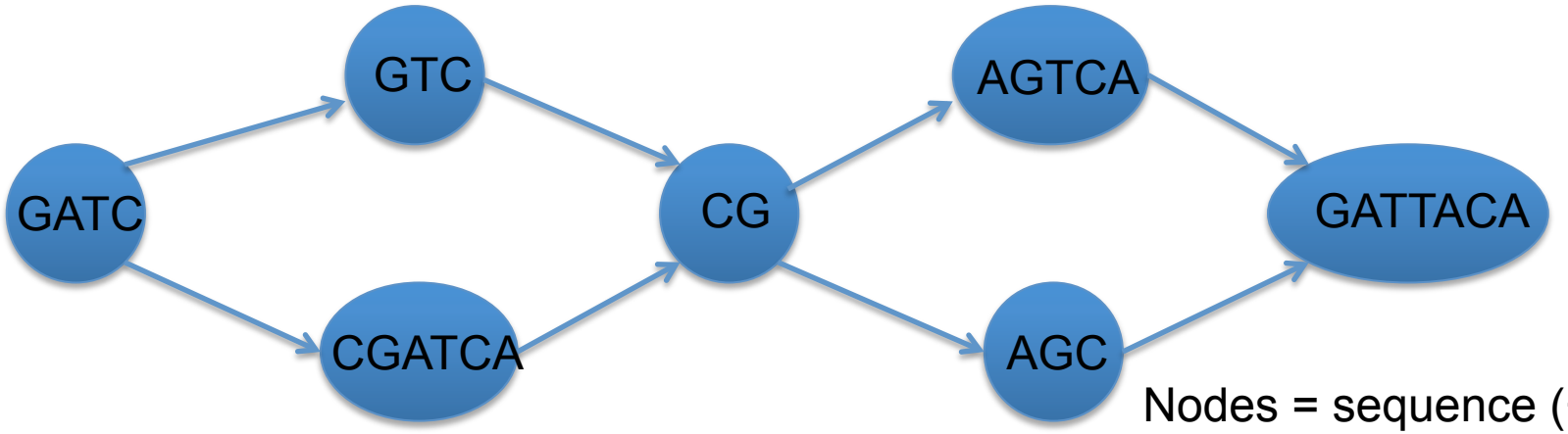
# Transcript Reconstruction from RNA-Seq Reads



# Graph Data Structures Commonly Used For Assembly


- Sequence
- Order
- Orientation (+, -)
- Overlap

Reads to Graph

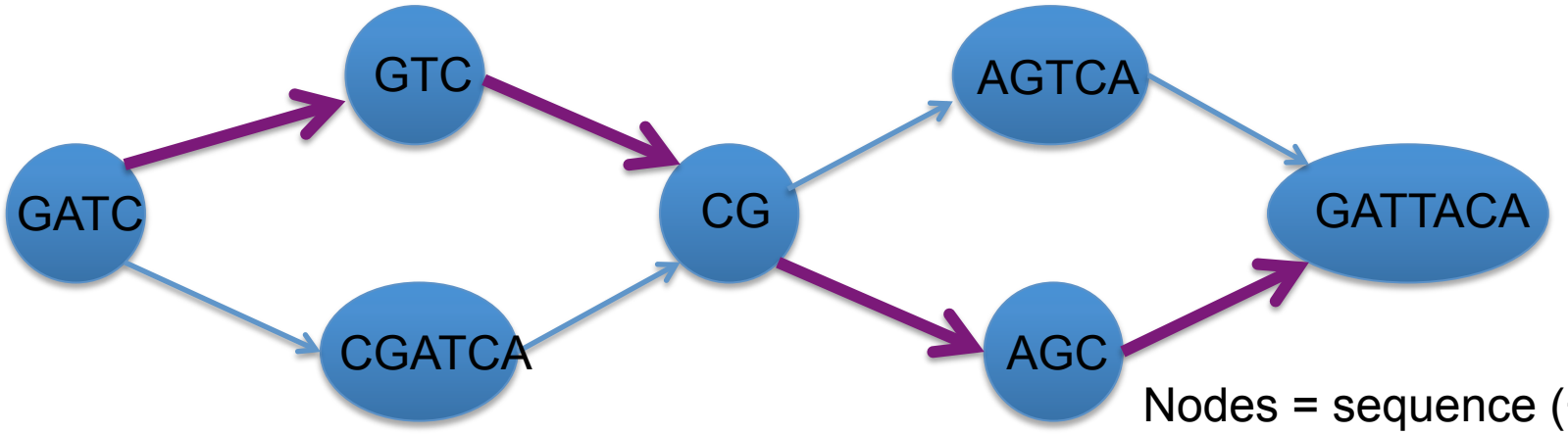



Nodes = sequence (+/-)
Edges = order, overlap

# Graph Data Structures Commonly Used For Assembly


- Sequence
- Order
- Orientation (+, -)
- Overlap

Reads to Graph

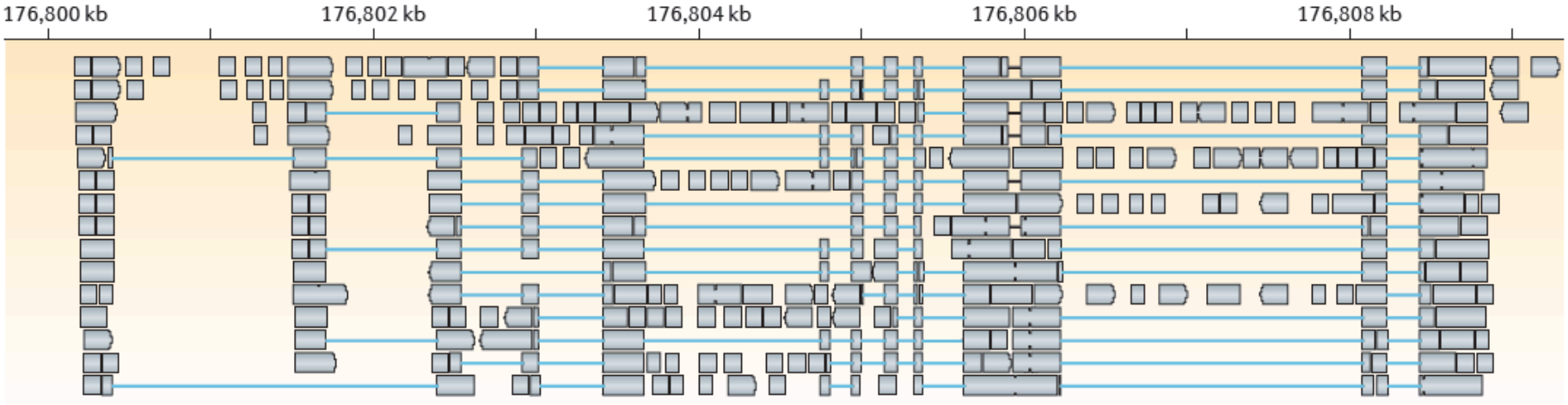



Nodes = sequence (+/-)
Edges = order, overlap

**GATCGTCCGAGCGATTACA**

# Genome-Guided Transcript Reconstruction

## Splice-align reads to the genome

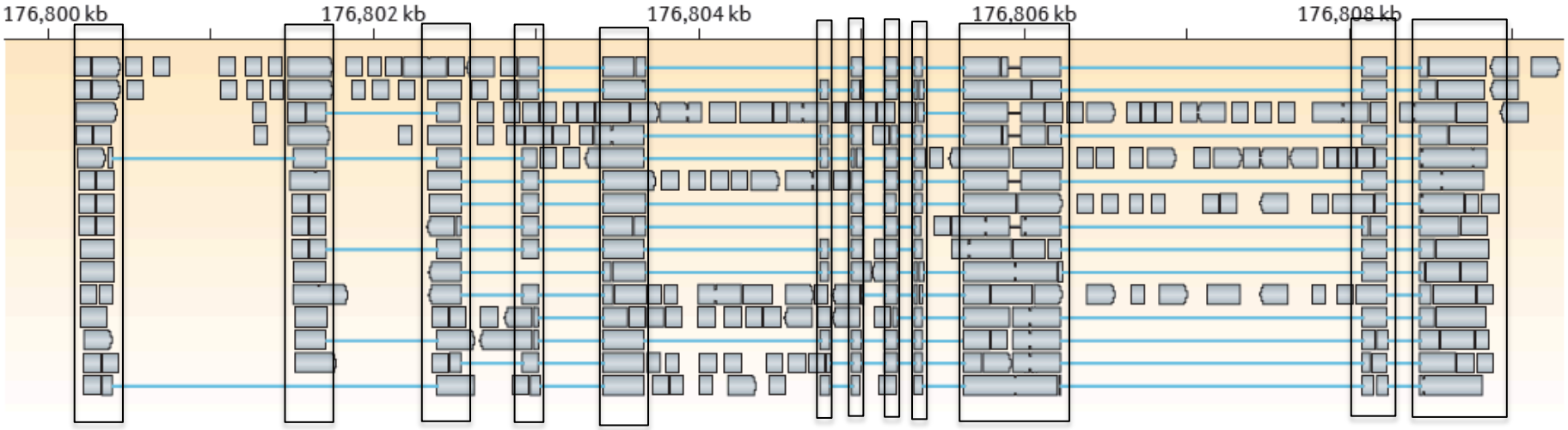





From Martin & Wang. Nature Reviews in Genetics. 2011

# Genome-Guided Transcript Reconstruction

## Splice-align reads to the genome



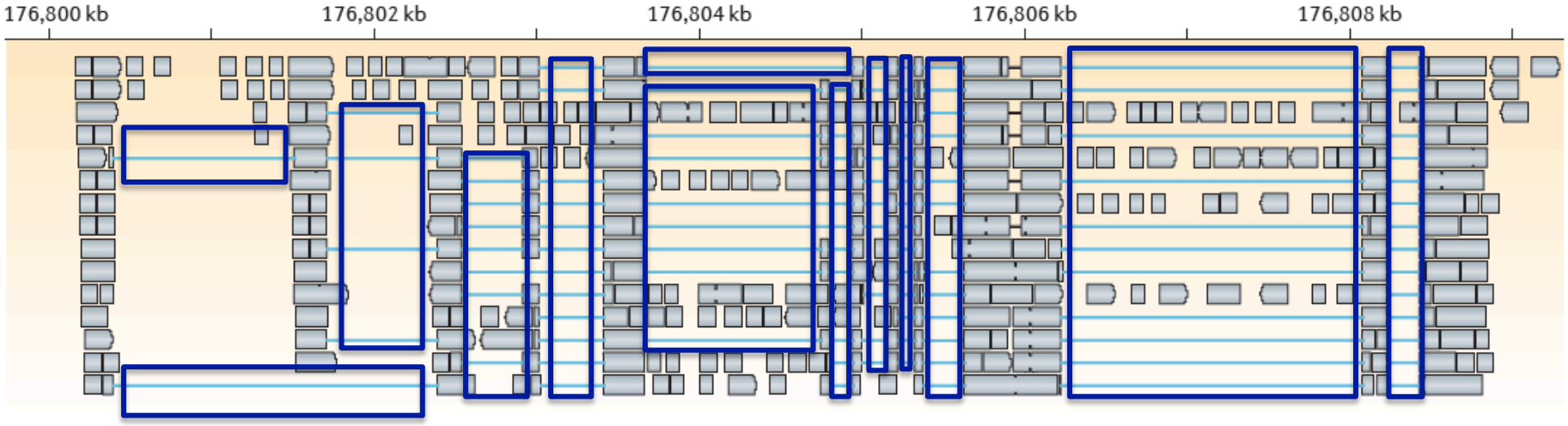

Alignment segment piles => exon regions



From Martin & Wang. Nature Reviews in Genetics. 2011

# Genome-Guided Transcript Reconstruction

## Splice-align reads to the genome



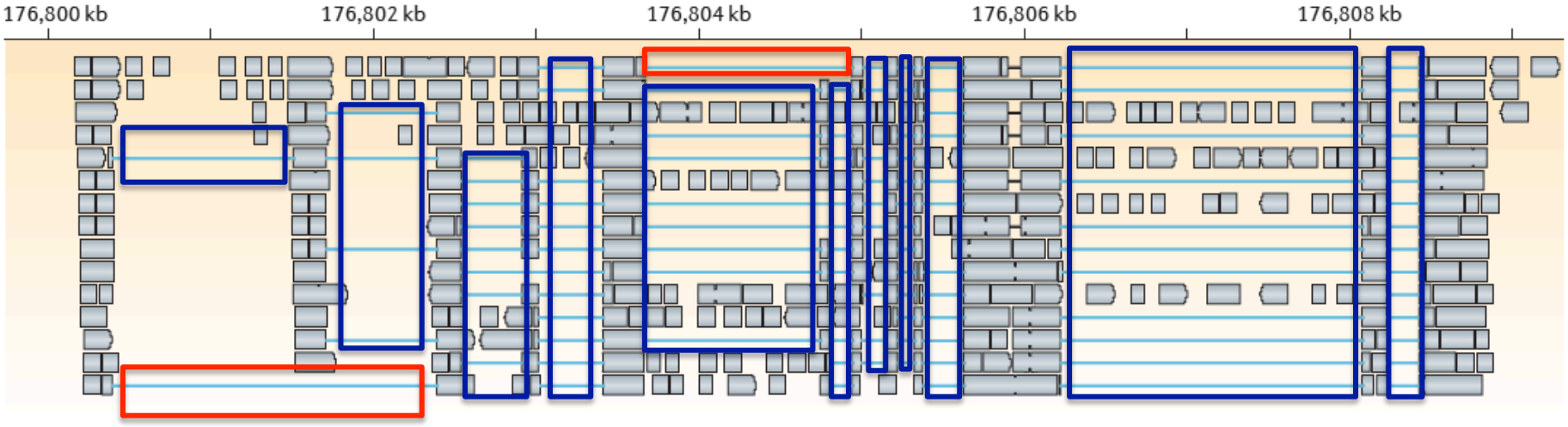

Large alignment gaps => introns



From Martin & Wang. Nature Reviews in Genetics. 2011

# Genome-Guided Transcript Reconstruction

## Splice-align reads to the genome



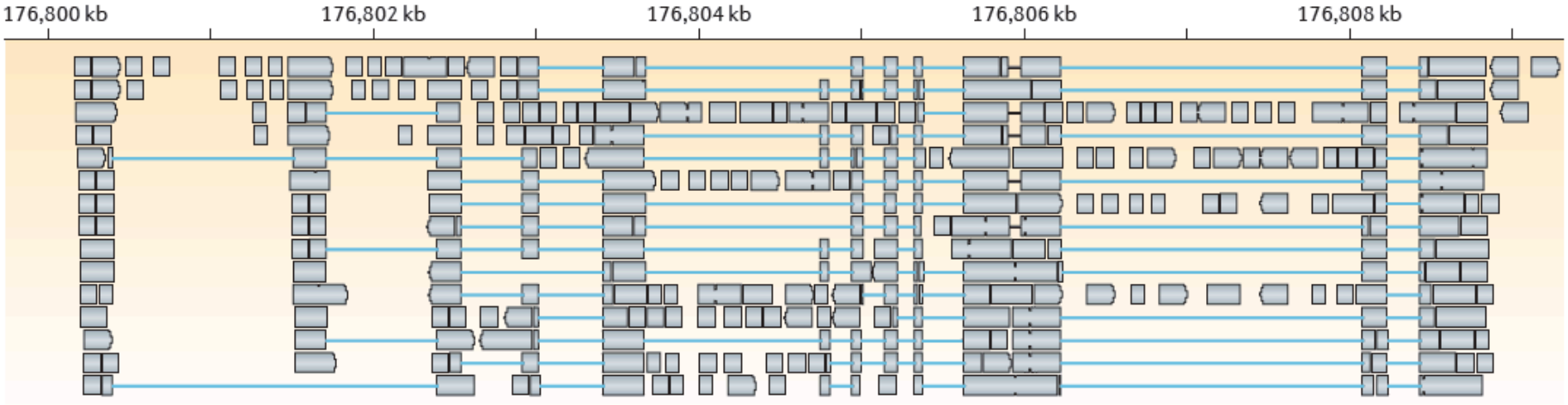

Overlapping but different introns = evidence of alternative splicing



From Martin & Wang. Nature Reviews in Genetics. 2011

# Genome-Guided Transcript Reconstruction

## Splice-align reads to the genome

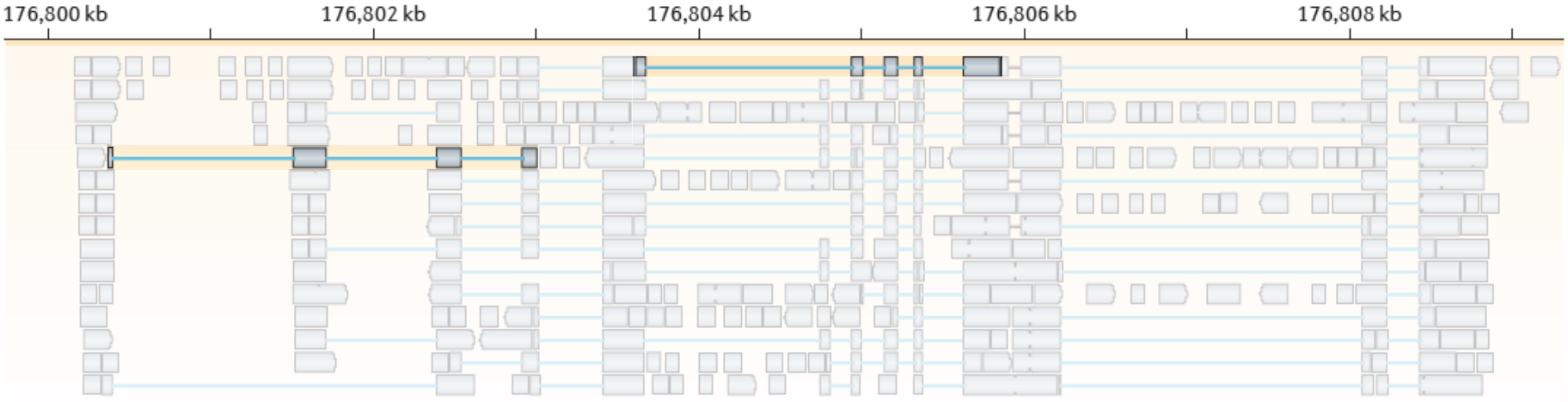





From Martin & Wang. Nature Reviews in Genetics. 2011

# Genome-Guided Transcript Reconstruction

## Splice-align reads to the genome



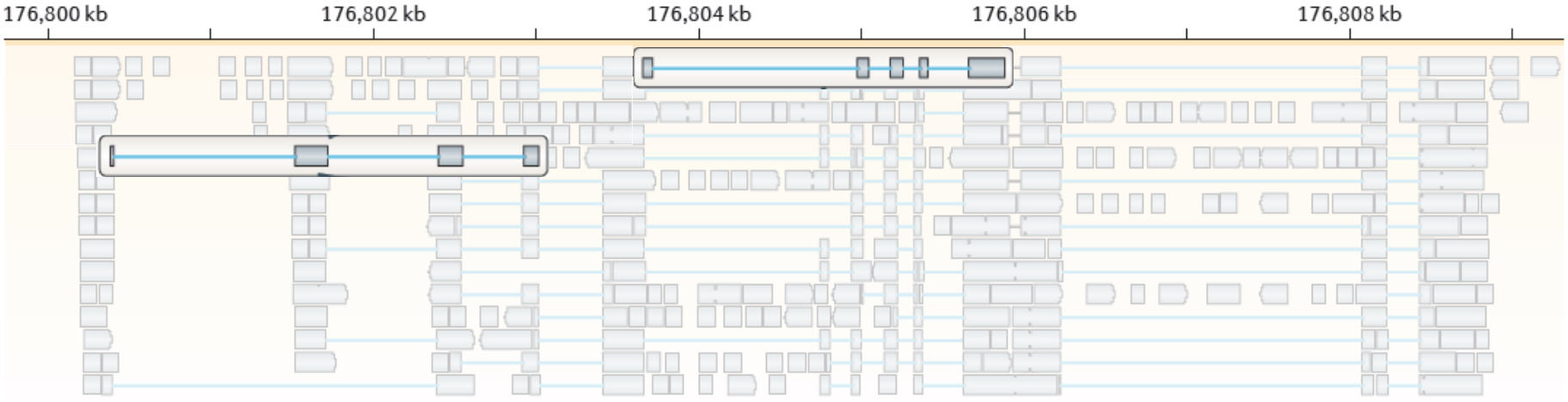

Individual reads can yield multiple exon and intron segments (splice patterns)



From Martin & Wang. Nature Reviews in Genetics. 2011

# Genome-Guided Transcript Reconstruction

## Splice-align reads to the genome



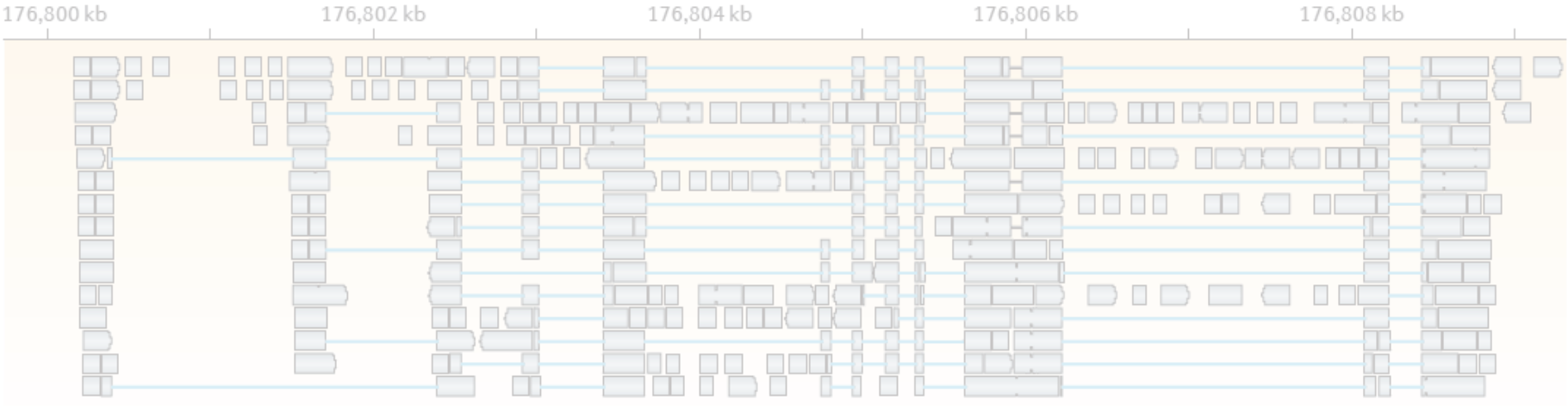

Nodes = unique splice patterns



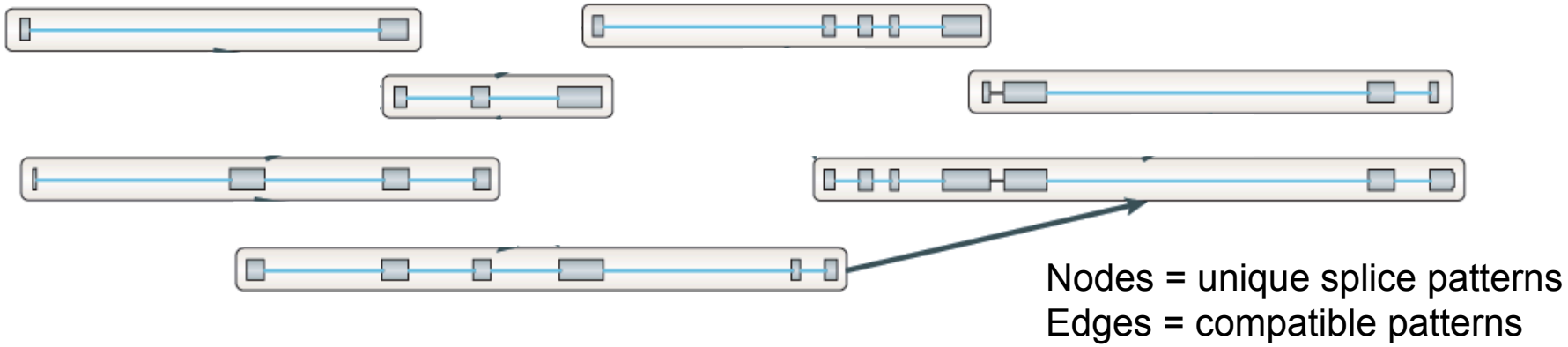
From Martin & Wang. Nature Reviews in Genetics. 2011

# Genome-Guided Transcript Reconstruction

## Splice-align reads to the genome




**Construct graph from unique splice patterns of aligned reads.**

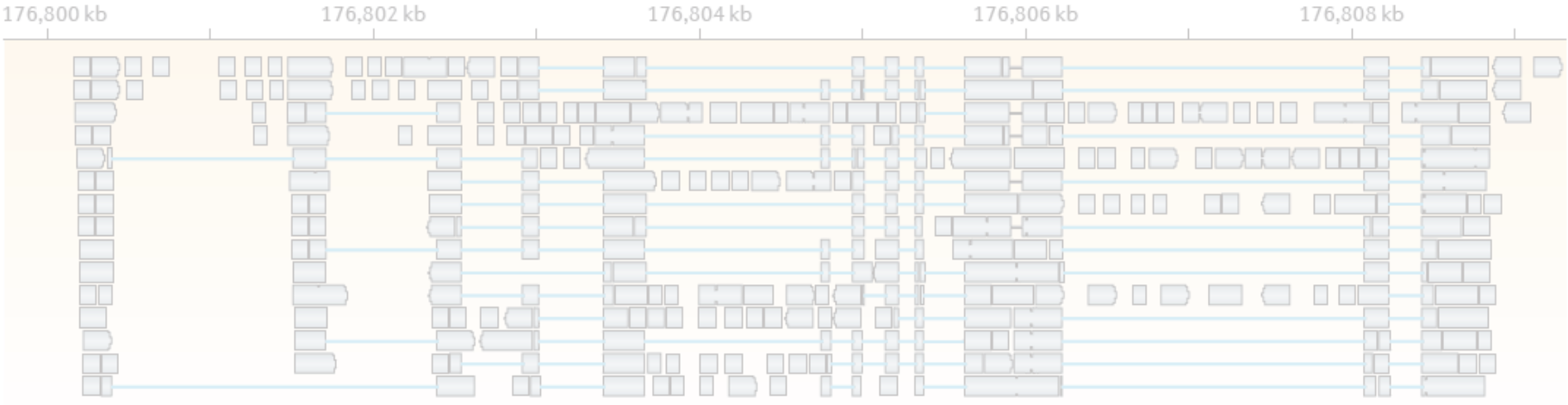




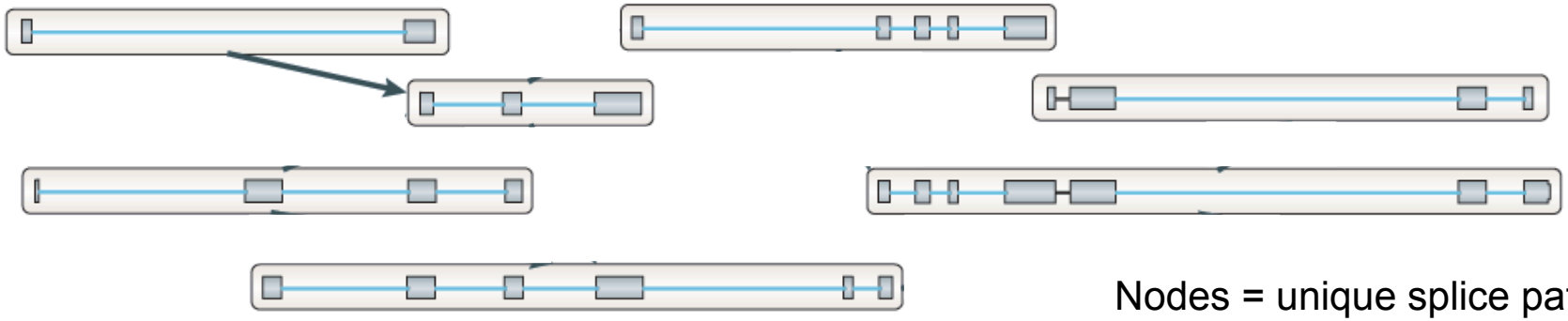

From Martin & Wang. Nature Reviews in Genetics. 2011

# Genome-Guided Transcript Reconstruction

## Splice-align reads to the genome




**Construct graph from unique splice patterns of aligned reads.**



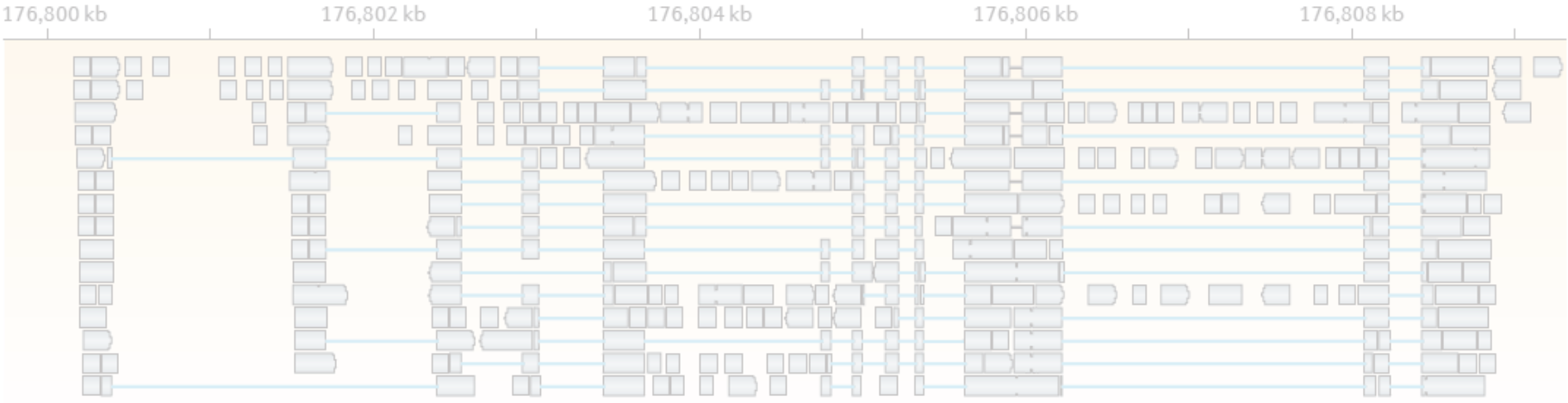
Nodes = unique splice patterns
Edges = compatible patterns



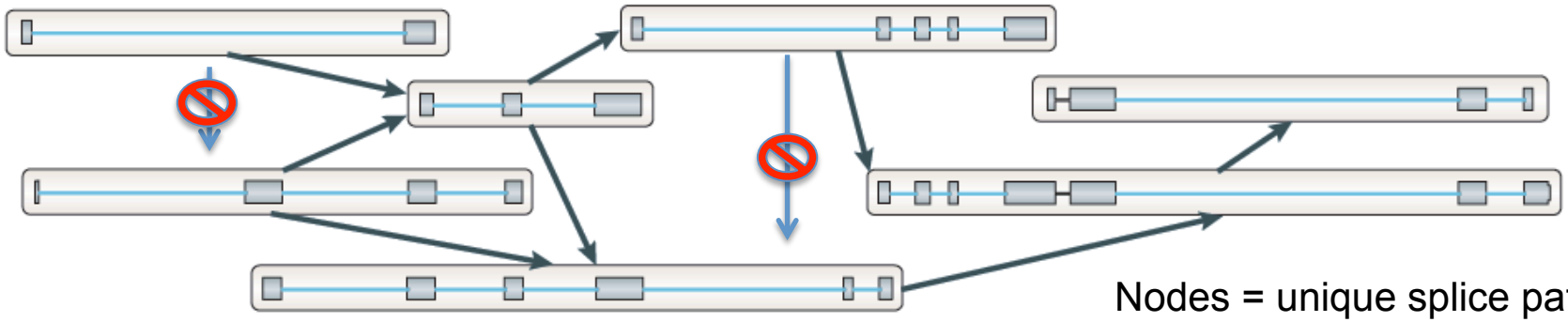
From Martin & Wang. Nature Reviews in Genetics. 2011

# Genome-Guided Transcript Reconstruction

## Splice-align reads to the genome




**Construct graph from unique splice patterns of aligned reads.**



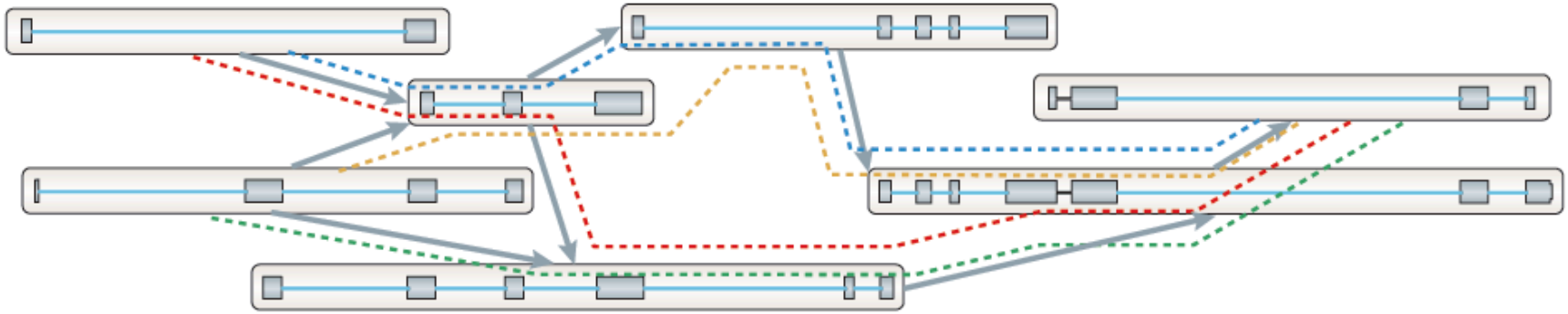
Nodes = unique splice patterns
Edges = compatible patterns



From Martin & Wang. Nature Reviews in Genetics. 2011

# Genome-Guided Transcript Reconstruction

**Traverse paths through the graph to assemble transcript isoforms**

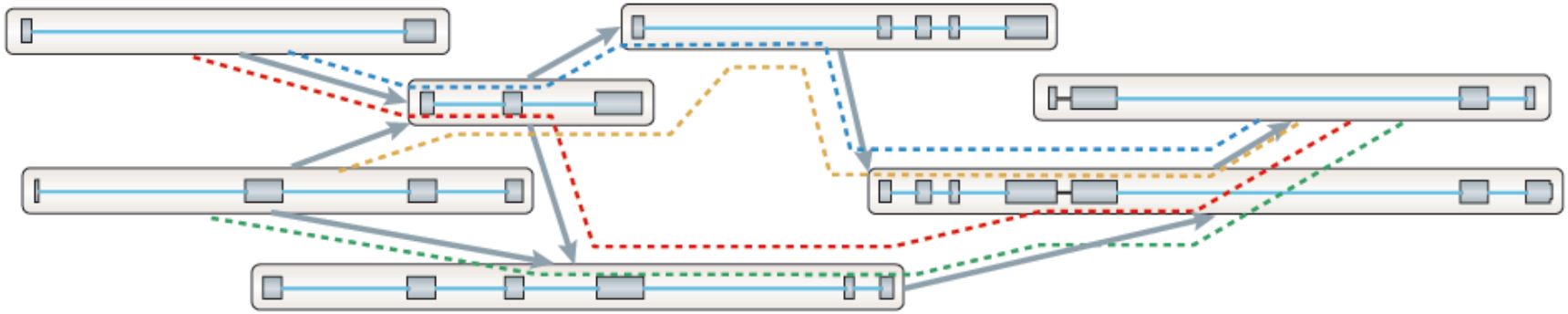




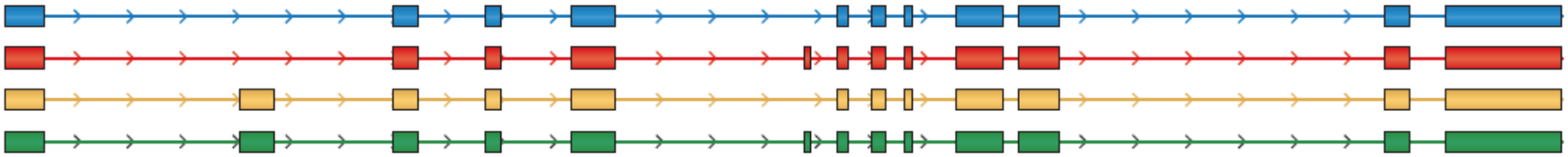
From Martin & Wang. Nature Reviews in Genetics. 2011

# Genome-Guided Transcript Reconstruction

**Traverse paths through the graph to assemble transcript isoforms**



**Reconstructed isoforms**



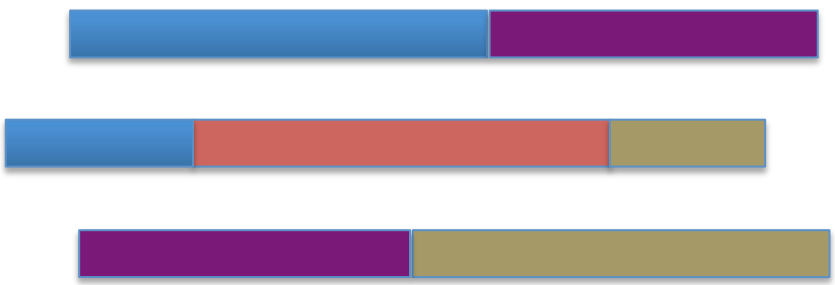


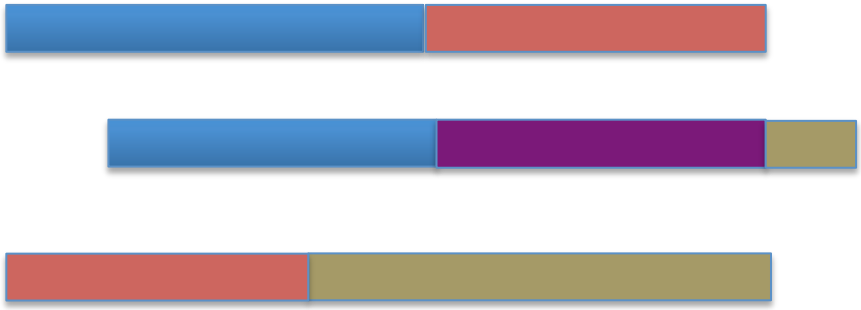
From Martin & Wang. Nature Reviews in Genetics. 2011

What if you don’t have a high quality reference genome sequence?

**Genome-free de novo transcript reconstruction to the rescue.**

# Read Overlap Graph: Reads as nodes, overlaps as edges

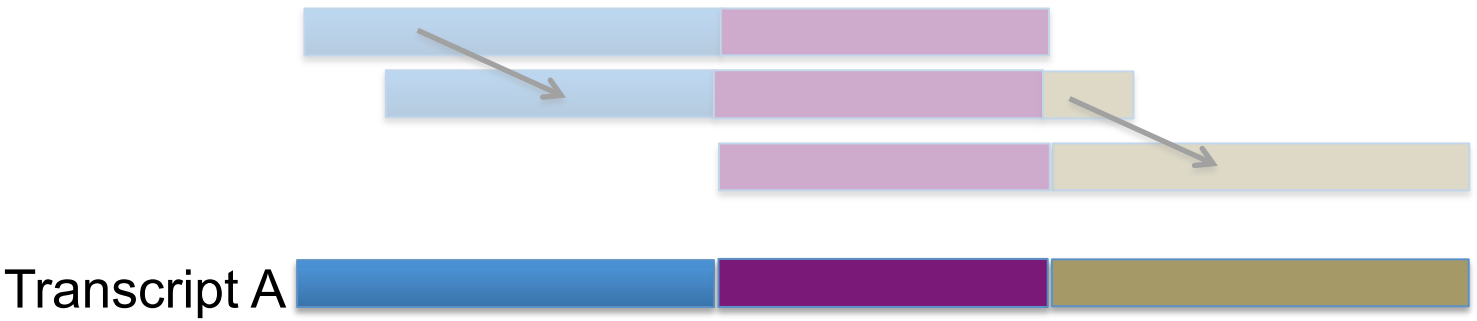




# Read Overlap Graph: Reads as nodes, overlaps as edges

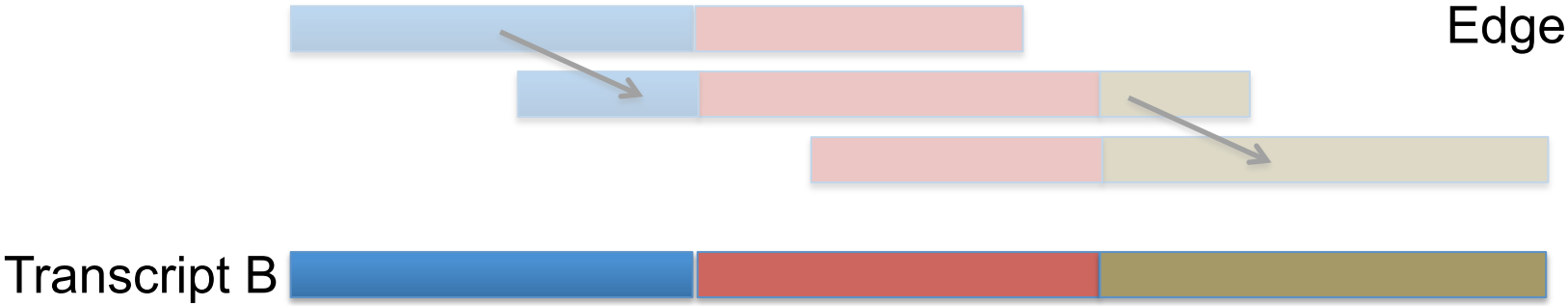
Node = read
Edge = overlap

## Read Overlap Graph: Reads as nodes, overlaps as edges

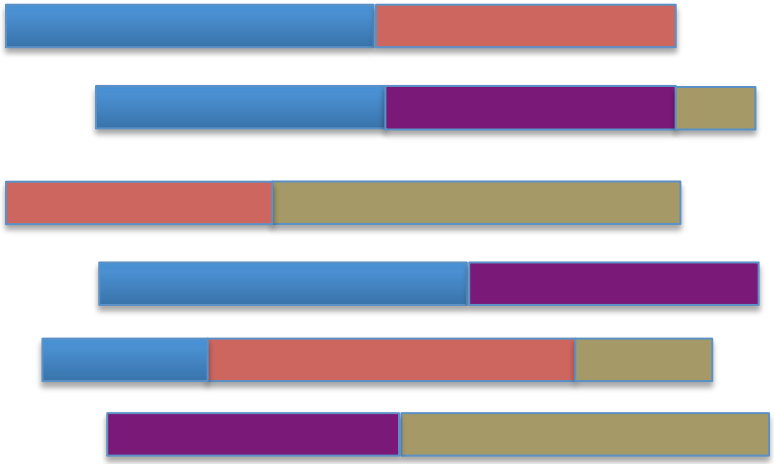


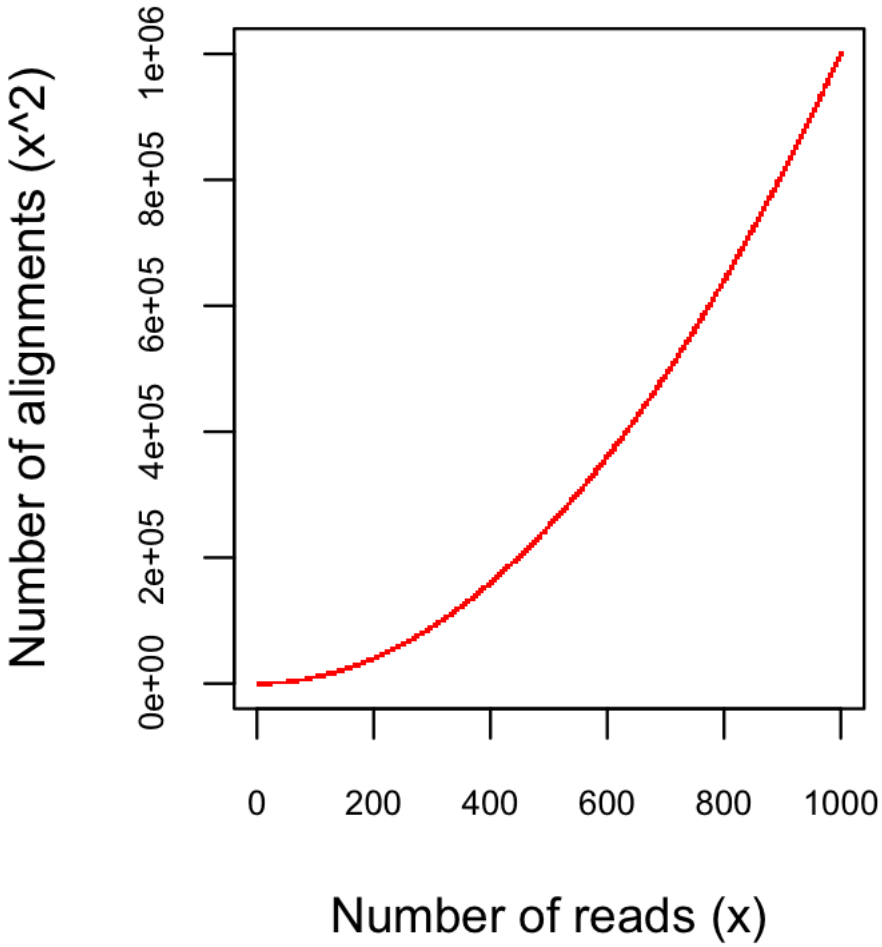

Generate consensus sequence where reads overlap

Node = read
Edge = overlap



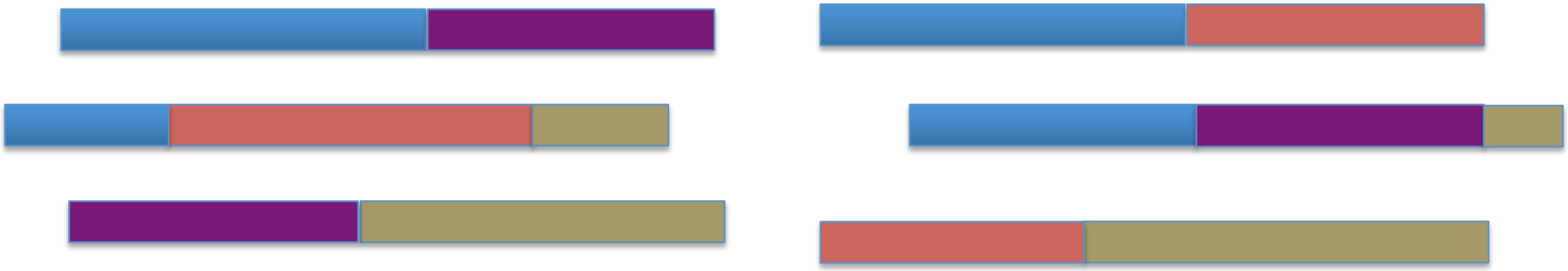
# Finding pairwise overlaps between *n* reads involves ~ $n^2$ comparisons.






*Impractical for typical RNA-Seq data (50M reads)*

## No genome to align to…  De novo assembly required

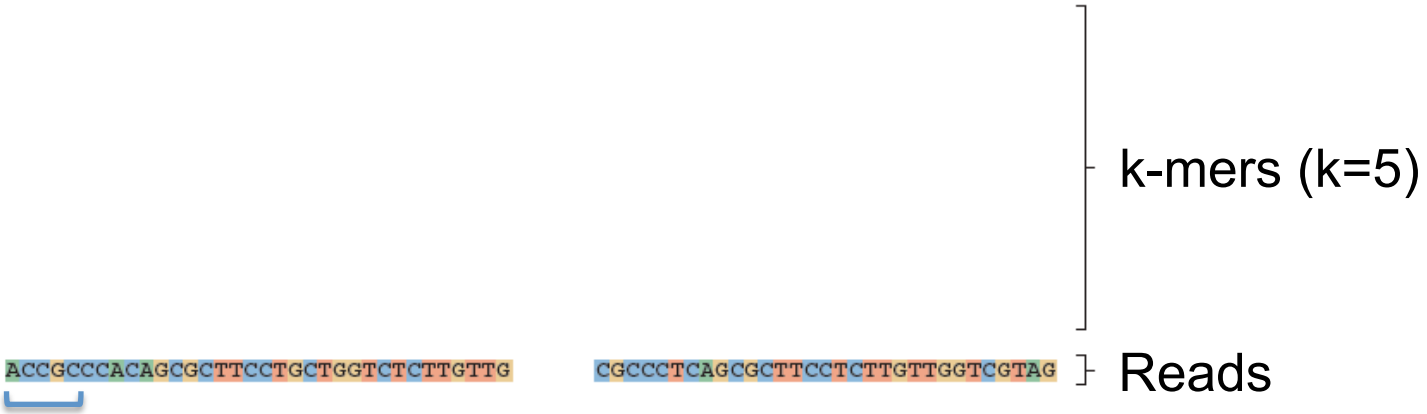


Want to avoid $n^2$ read alignments to define overlaps

**Use a de Bruijn graph**

# Sequence Assembly via De Bruijn Graphs

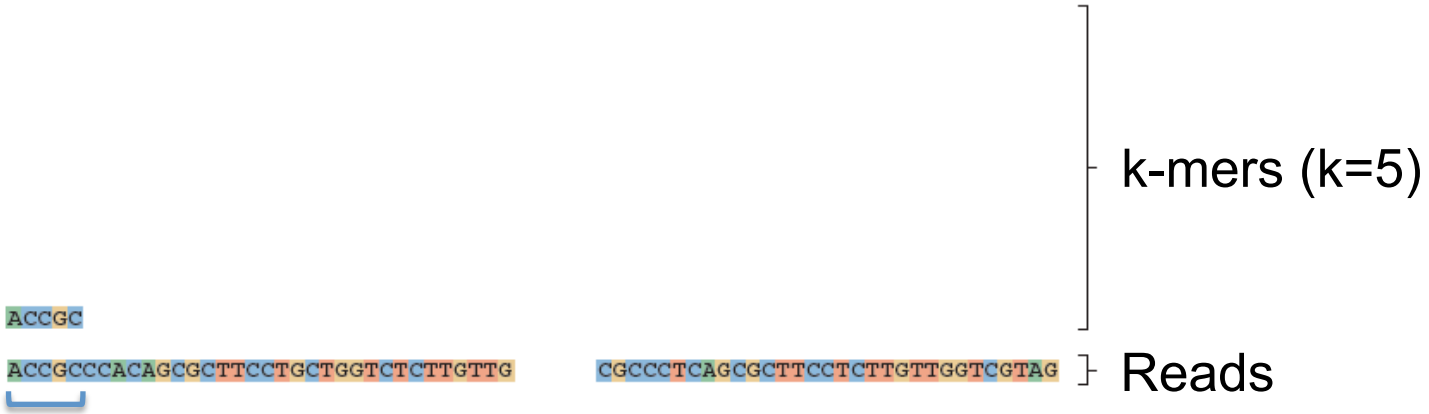
## Generate all substrings of length k from the reads




From Martin & Wang, Nat. Rev. Genet. 2011

# Sequence Assembly via De Bruijn Graphs

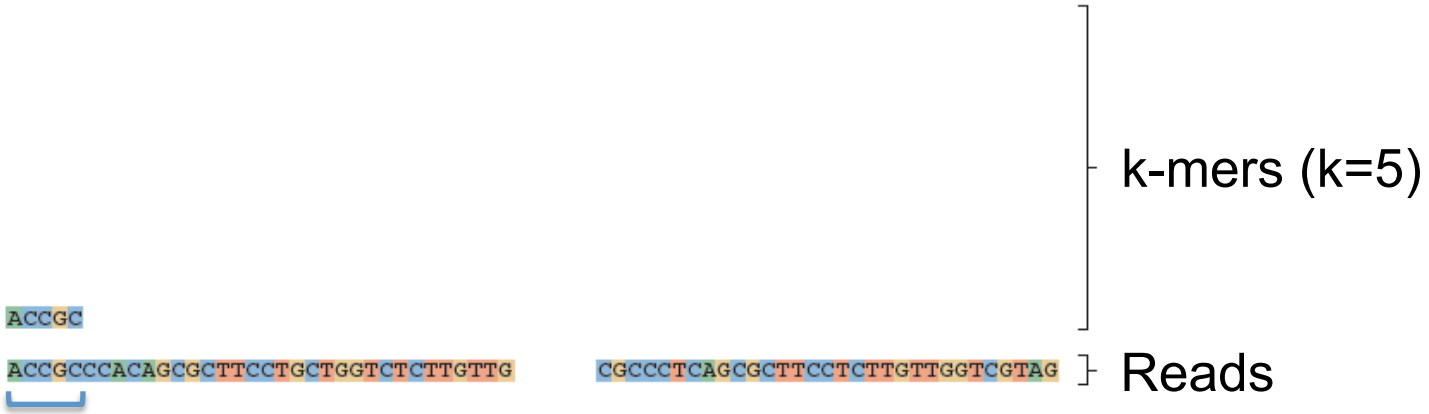
## Generate all substrings of length k from the reads




From Martin & Wang, Nat. Rev. Genet. 2011

# Sequence Assembly via De Bruijn Graphs

## Generate all substrings of length k from the reads



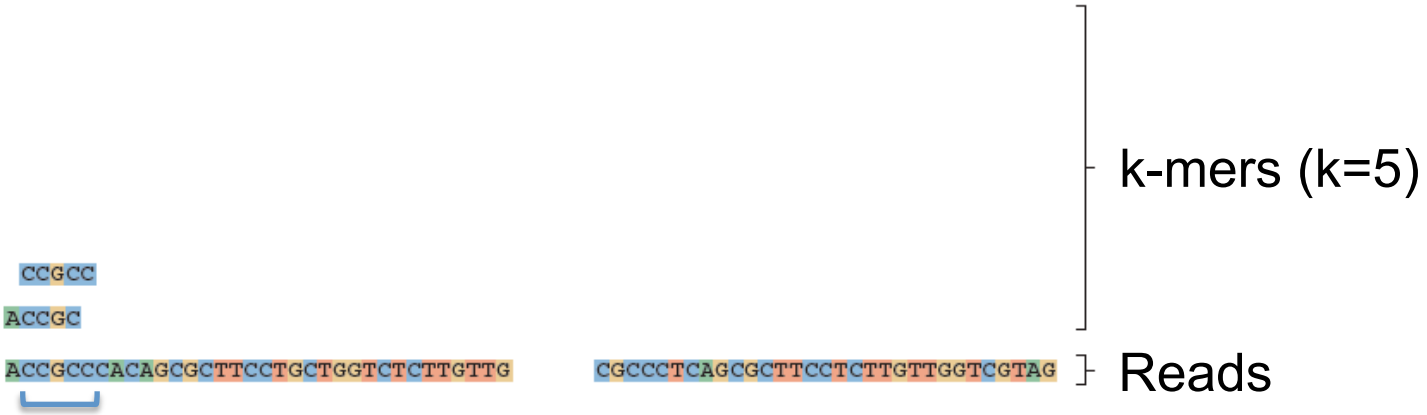

## Construct the de Bruijn graph


From Martin & Wang, Nat. Rev. Genet. 2011

Nodes = unique k-mers, Edges = overlap by (k-1)

# Sequence Assembly via De Bruijn Graphs

## Generate all substrings of length k from the reads



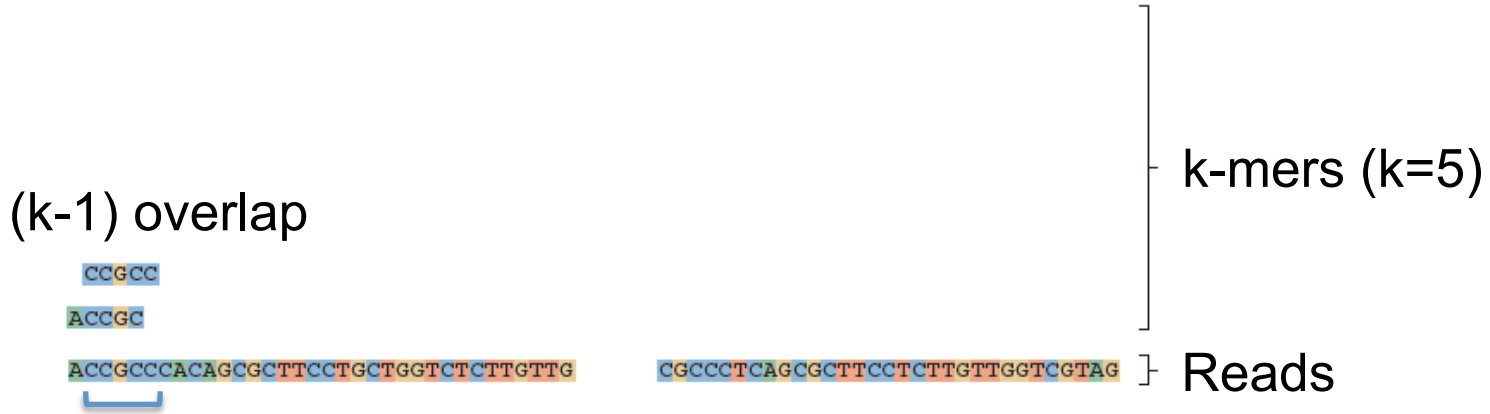

## Construct the de Bruijn graph


From Martin & Wang, Nat. Rev. Genet. 2011

Nodes = unique k-mers, Edges = overlap by (k-1)

# Sequence Assembly via De Bruijn Graphs

## Generate all substrings of length k from the reads



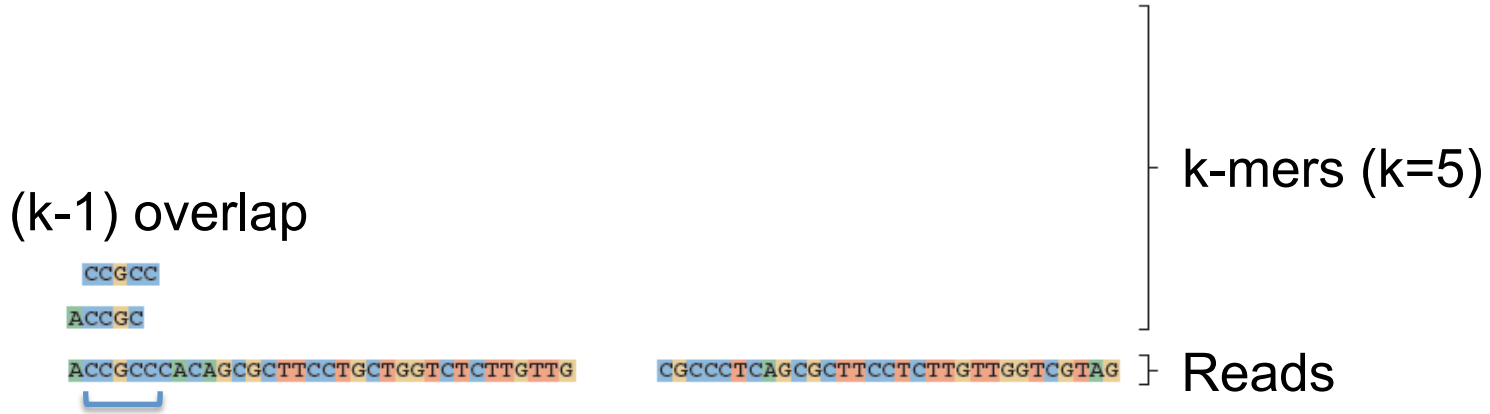

## Construct the de Bruijn graph


From Martin & Wang, Nat. Rev. Genet. 2011

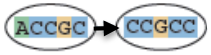
Nodes = unique k-mers, Edges = overlap by (k-1)

# Sequence Assembly via De Bruijn Graphs

## Generate all substrings of length k from the reads




## Construct the de Bruijn graph

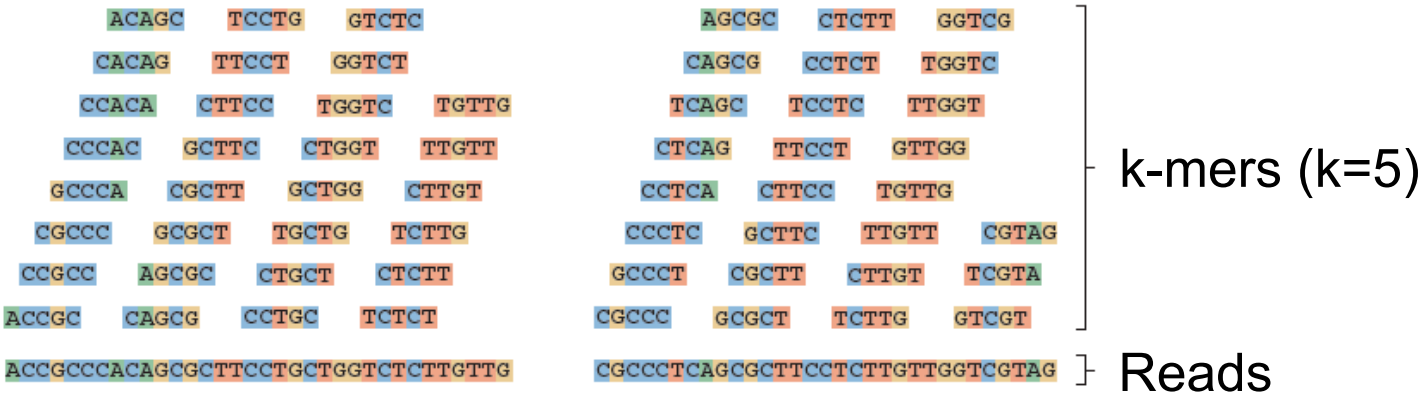



From Martin & Wang, Nat. Rev. Genet. 2011

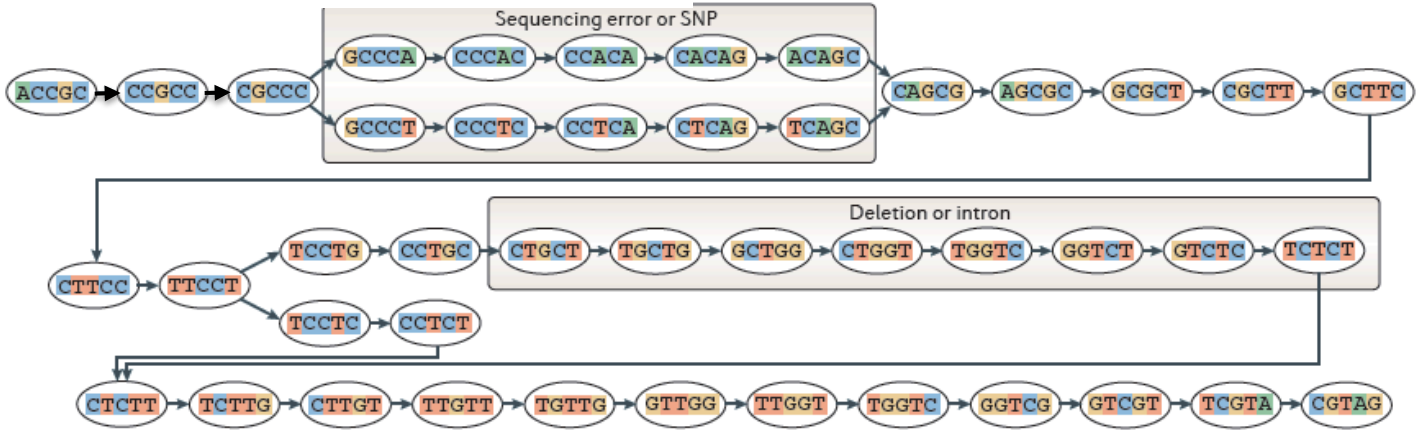
Nodes = unique k-mers, Edges = overlap by (k-1)

# Sequence Assembly via De Bruijn Graphs

## Generate all substrings of length k from the reads



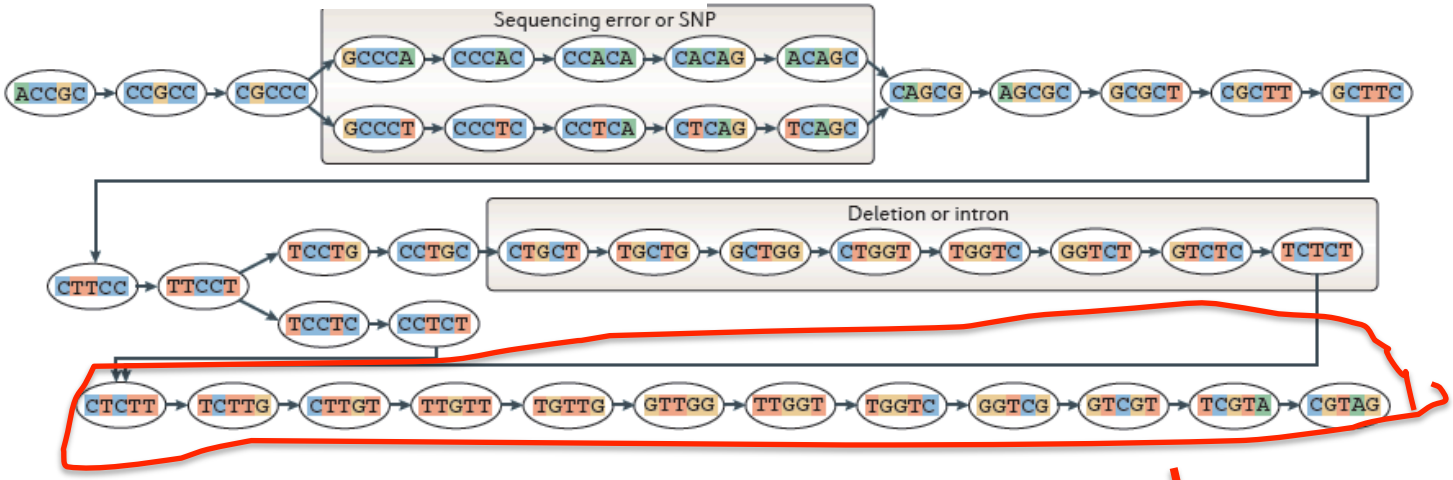

## Construct the de Bruijn graph



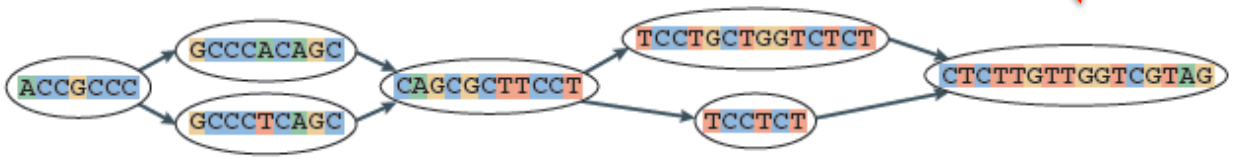

From Martin & Wang, Nat. Rev. Genet. 2011

Nodes = unique k-mers, Edges = overlap by (k-1)

# Construct the de Bruijn graph

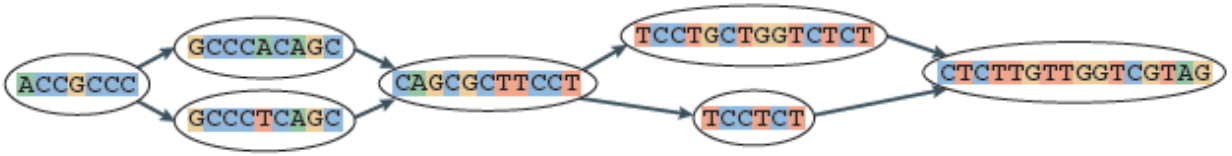



# Collapse the de Bruijn graph

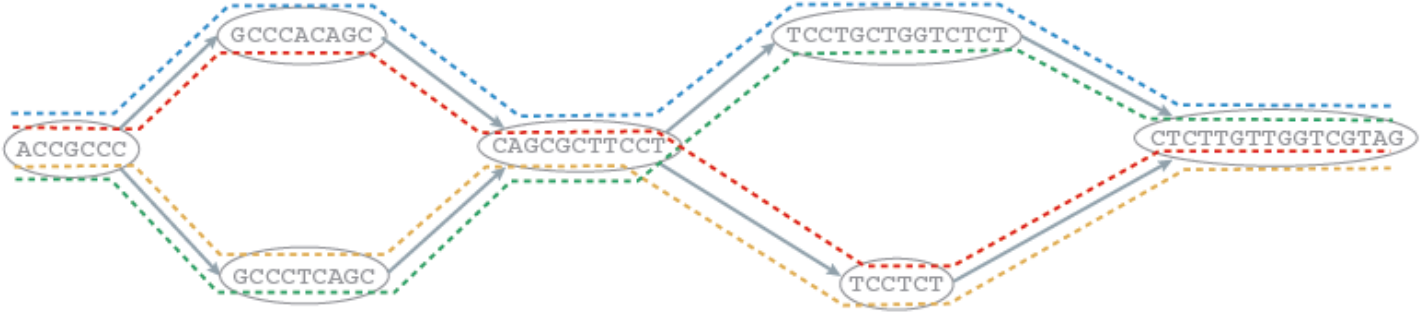



From Martin & Wang, Nat. Rev. Genet. 2011

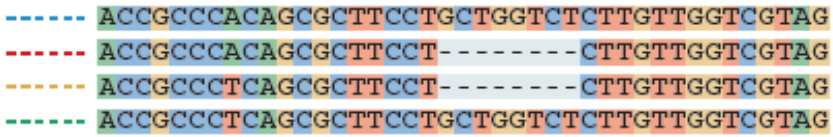
## Collapse the de Bruijn graph




## Traverse the graph




## Assemble Transcript Isoforms




From Martin & Wang, Nat. Rev. Genet. 2011

## Contrasting Genome and Transcriptome *De novo* Assembly

### Genome Assembly

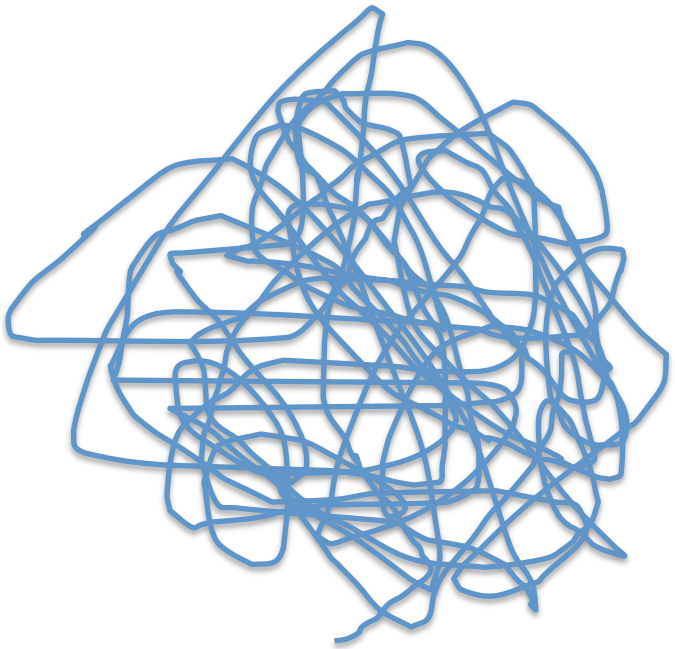
- Uniform coverage
- Single contig per locus
- Assemble small numbers of large Mb-length chromosomes
- Double-stranded data

### Transcriptome Assembly

- Exponentially distributed coverage levels
- Multiple contigs per locus (alt splicing)
- Assemble many thousands of Kb-length transcripts
- Strand-specific data available

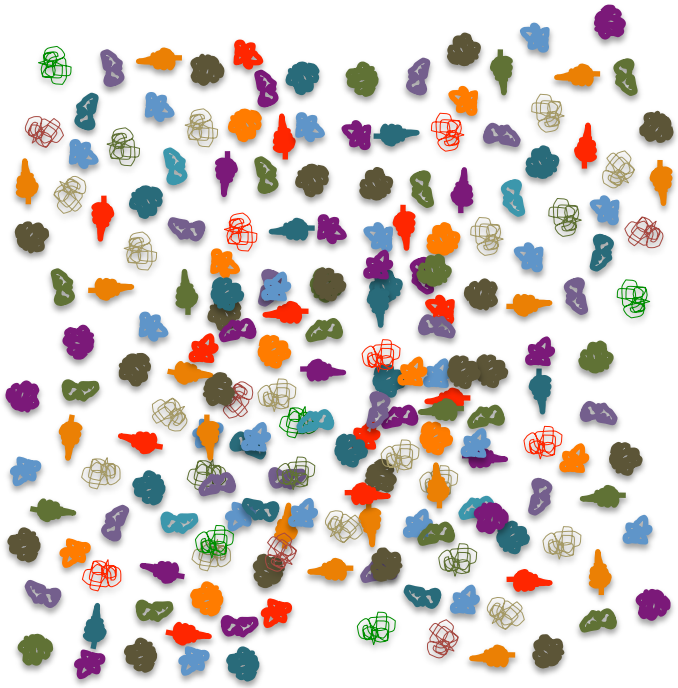
# Trinity Aggregates Isolated Transcript Graphs

**Genome Assembly**
Single Massive Graph



Entire chromosomes represented.

**Trinity Transcriptome Assembly**
Many Thousands of Small Graphs



Ideally, one graph per expressed gene.

# Trinity – How it works:




**RNA-Seq reads** → **Linear contigs** → **de-Bruijn graphs** → **Transcripts + Isoforms**

>a121:len=5845
>a122:len=2560
>a123:len=4443
>a124:len=48
>a125:len=8876
>a126:len=66

...CTTCGCAA...TGATCGGAT...
...ATTCGCAA...TCATCGGAT...

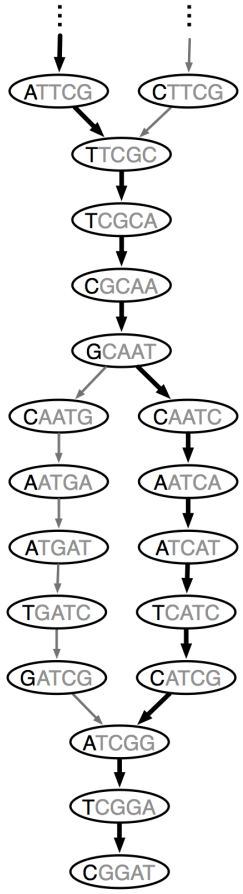
Thousands of disjoint graphs

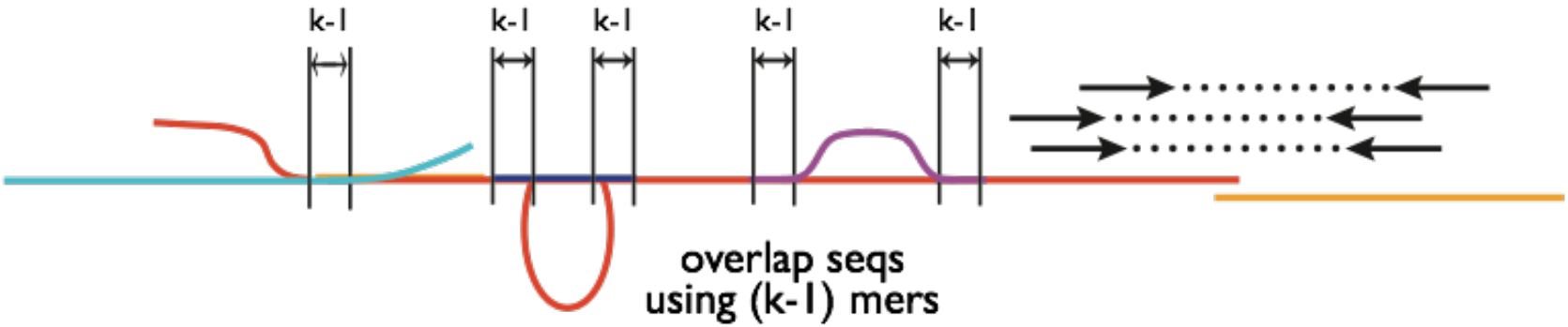
# Chrysalis

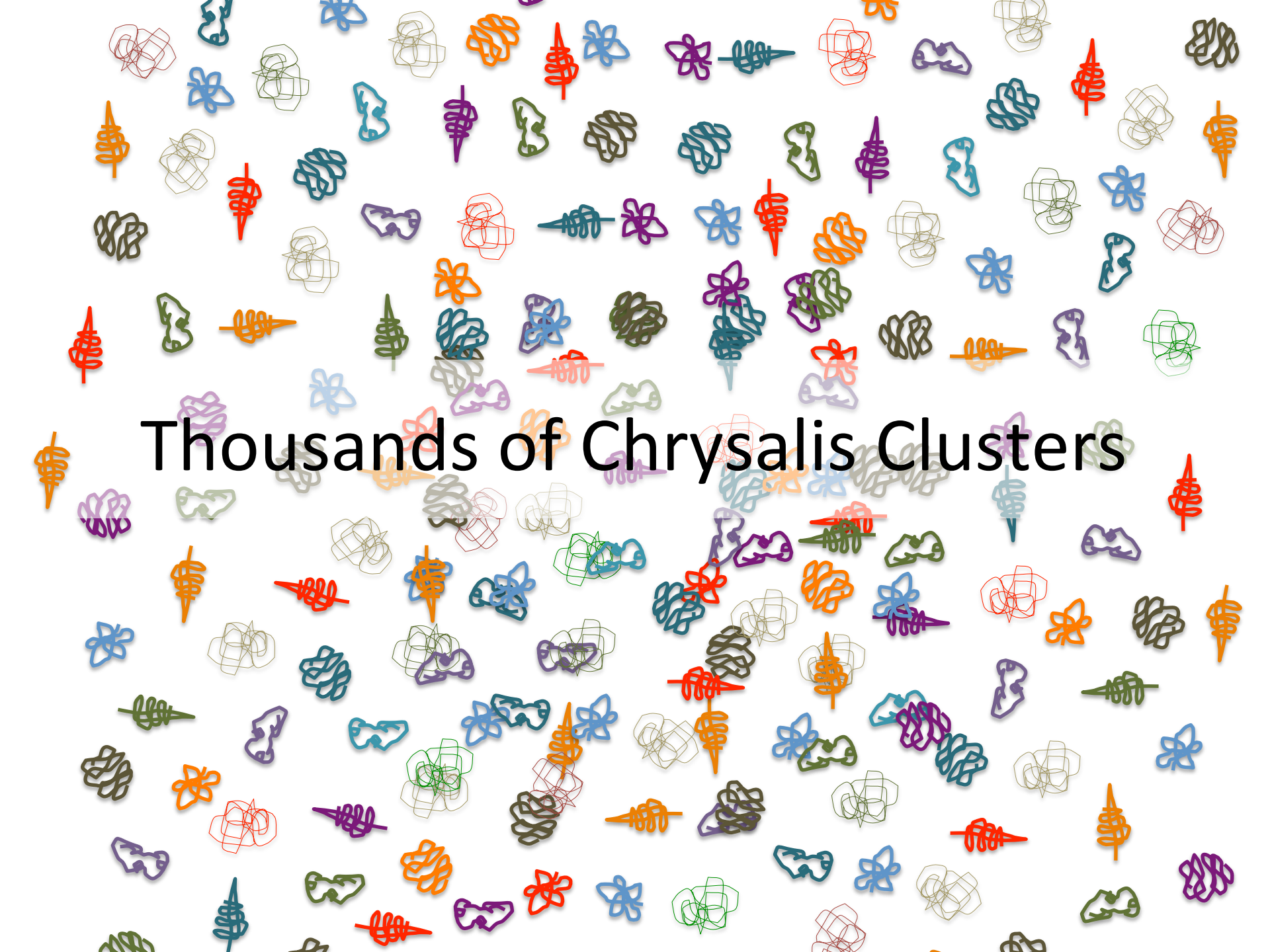
>a121:len=5845
>a122:len=2560
>a123:len=4443
>a124:len=48
>a125:len=8876
>a126:len=66

Integrate isoforms
via k-1 overlaps

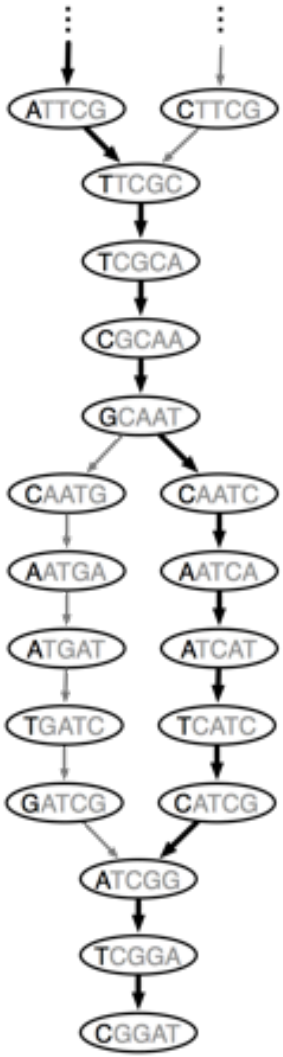
Build de Bruijn Graphs
(ideally, one per gene)



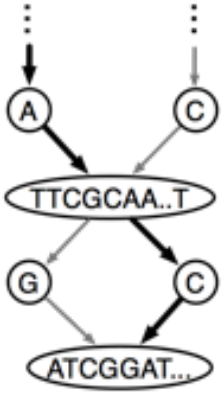

Thousands of Chrysalis Clusters

# Butterfly



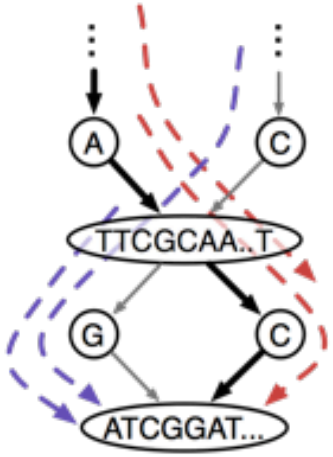

de Bruijn graph

compacting




compact graph

finding paths




compact graph with reads

extracting sequences

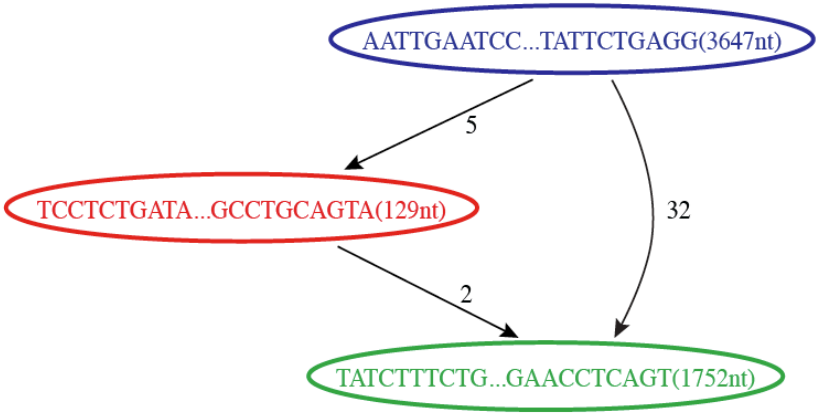
..CTTCGCAA..TGATCGGAT...

..ATTCGCAA..TCATCGGAT...

sequences (isoforms and paralogs)

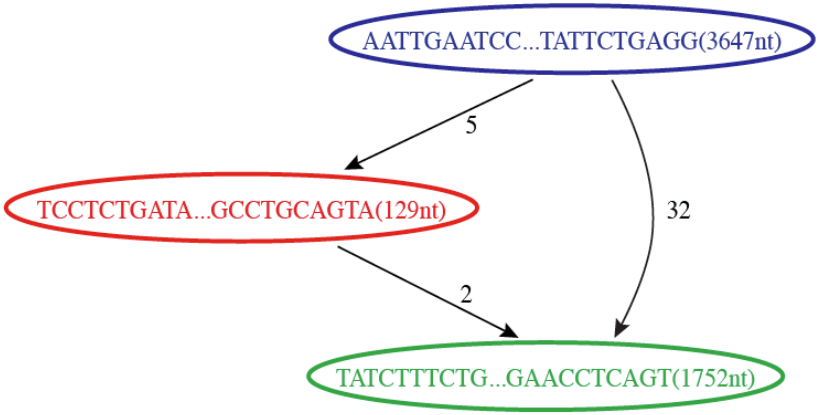
# Butterfly Example 1:
# Reconstruction of Alternatively Spliced Transcripts

Butterfly's Compacted Sequence Graph




# Reconstruction of Alternatively Spliced Transcripts

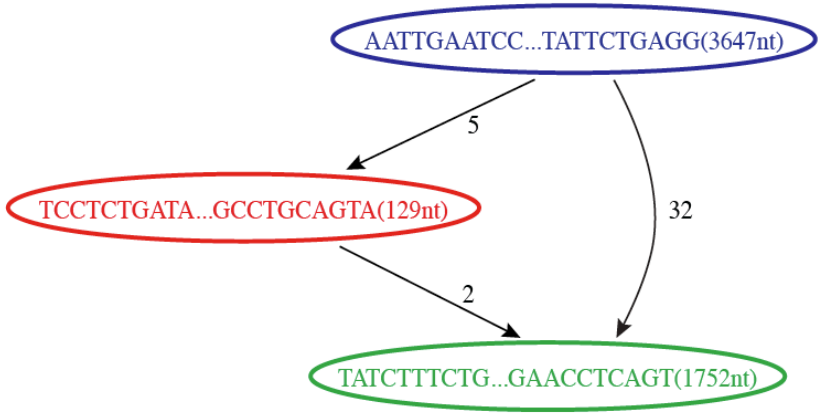
Butterfly's Compacted Sequence Graph




Reconstructed Transcripts

# Reconstruction of Alternatively Spliced Transcripts

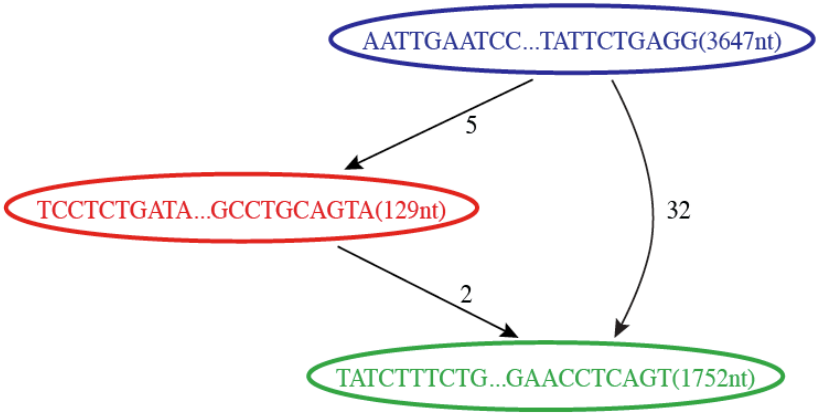
Butterfly's Compacted Sequence Graph




Reconstructed Transcripts

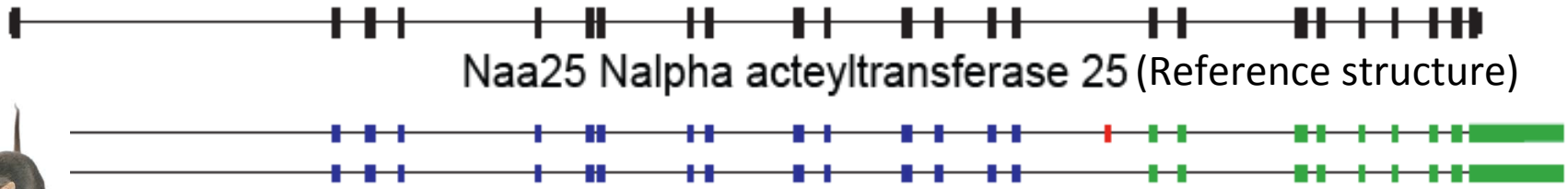
# Reconstruction of Alternatively Spliced Transcripts

Butterfly's Compacted Sequence Graph



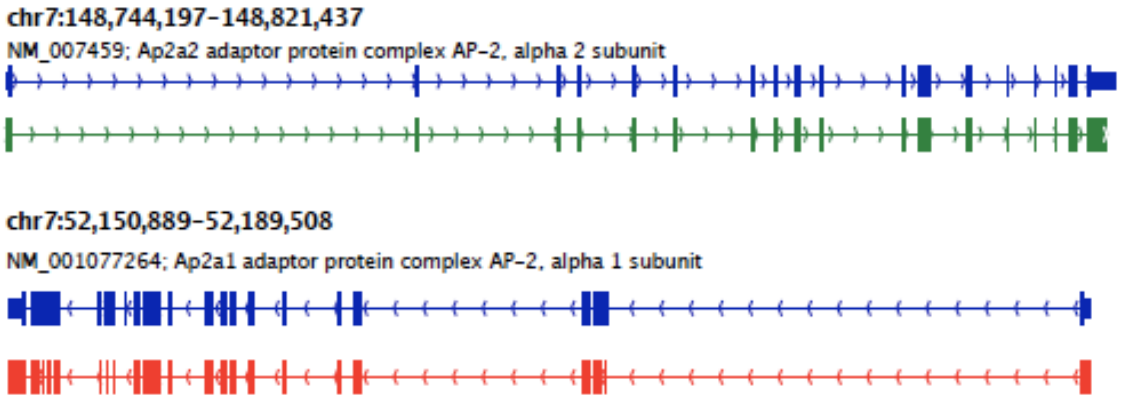

Reconstructed Transcripts

Aligned to Mouse Genome




# Butterfly Example 2:
# Teasing Apart Transcripts of Paralogous Genes

# Butterfly Example 2:
# Teasing Apart Transcripts of Paralogous Genes





# Strand-specific RNA-Seq is Preferred

Computationally: fewer confounding graph structures in de novo assembly:

ex. Forward != reverse complement

(GGAA != TTCC)

Biologically: separate sense vs. antisense transcription

NATURE METHODS | VOL.7 NO.9 | SEPTEMBER 2010 |

BROAD INSTITUTE

## Comprehensive comparative analysis of strand-specific RNA sequencing methods

Joshua Z Levin[1,6], Moran Yassour[1–3,6], Xian Adiconis[1], Chad Nusbaum[1], Dawn Anne Thompson[1], Nir Friedman[3,4], Andreas Gnirke[1] & Aviv Regev[1,2,5]

**Strand-specific, massively parallel cDNA sequencing (RNA-seq) is a powerful tool for transcript discovery, genome annotation and expression profiling. There are multiple published methods**

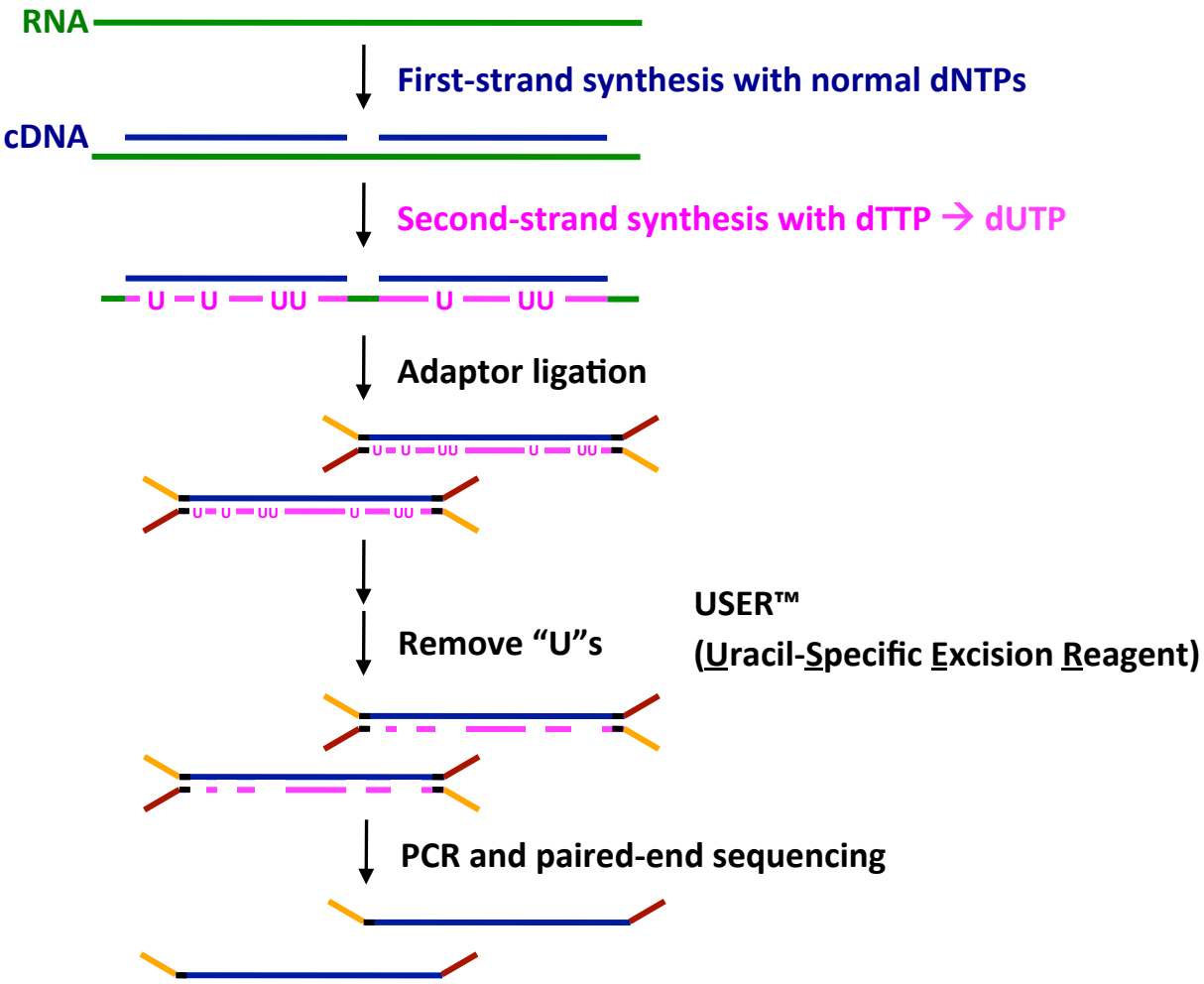
'dUTP second strand marking' identified as the leading protocol

**any RNA-seq method. Using the well-annotated *Saccharomyces cerevisiae* transcriptome as a benchmark, we compared seven library-construction protocols, including both published and**

Nevertheless, direct information on the originating strand can substantially enhance the value of an RNA-seq experiment. For example, such information would help to accurately identify anti-

resolve the correct expression levels of coding or noncoding over-lapping transcripts. These tasks are particularly challenging in small microbial genomes, prokaryotic and eukaryotic, in which

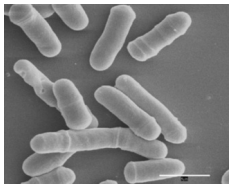
# dUTP 2nd Strand Method: Our Favorite



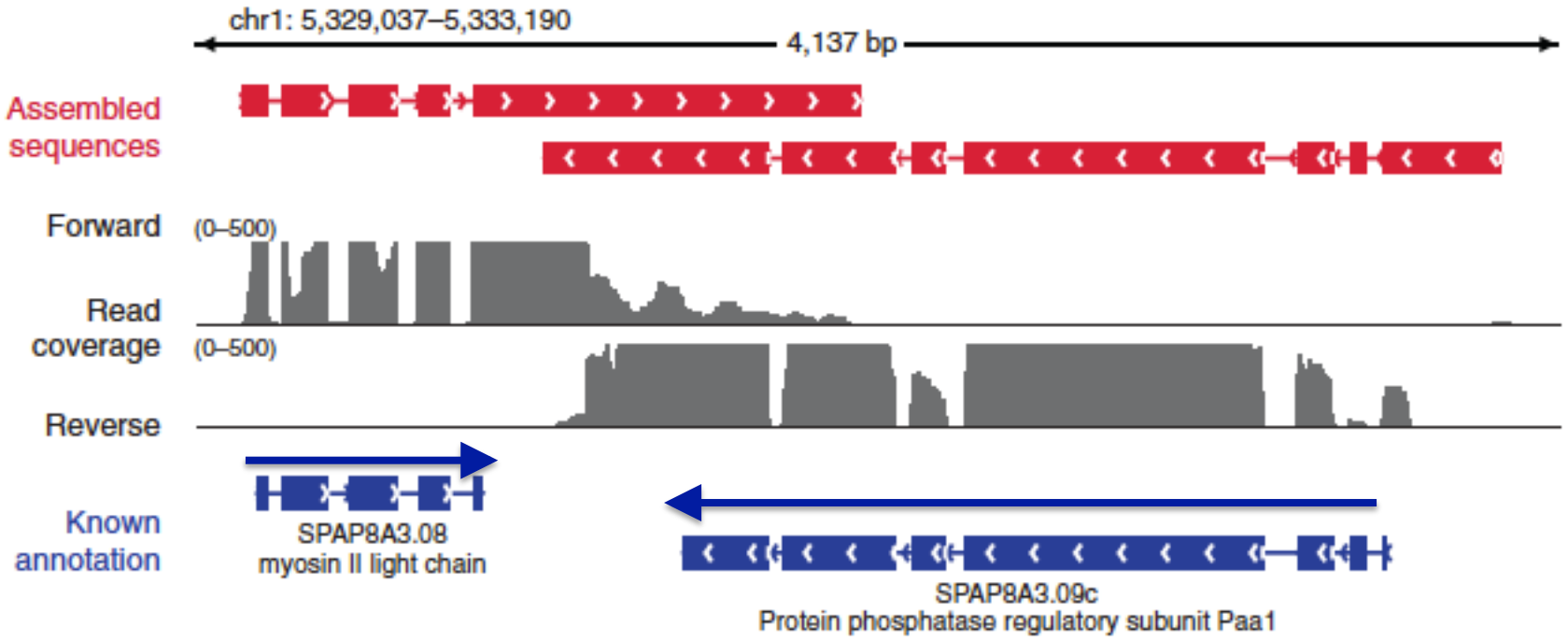

**Modified from Parkhomchuk *et al.* (2009) *Nucleic Acids Res.* 37:e123**

Slide from J. Levin

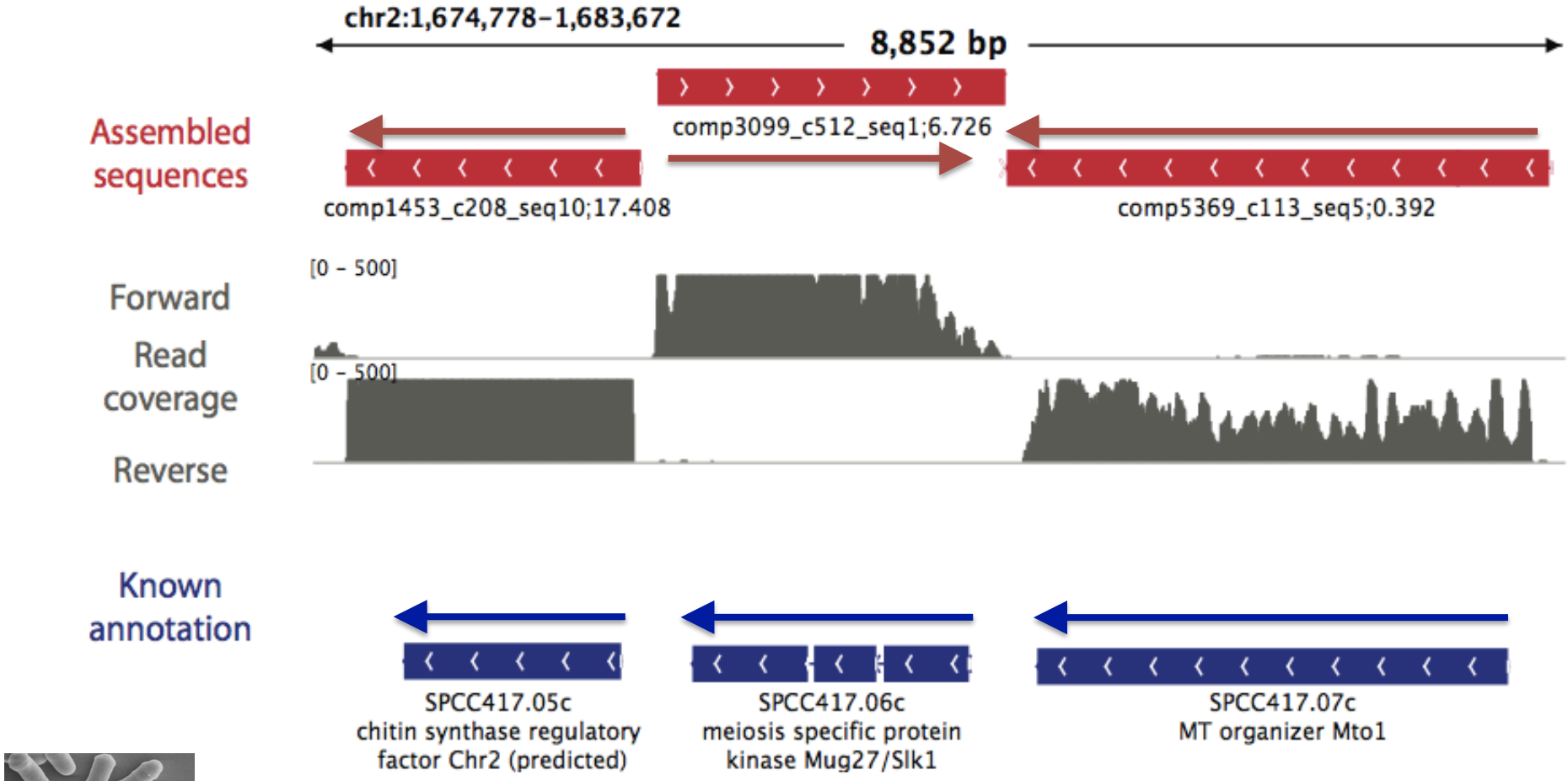
# Overlapping UTRs from Opposite Strands

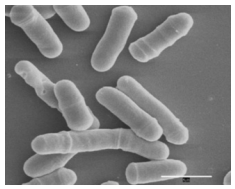


*Schizosacharomyces pombe*
(fission yeast)



# Antisense-dominated Transcription

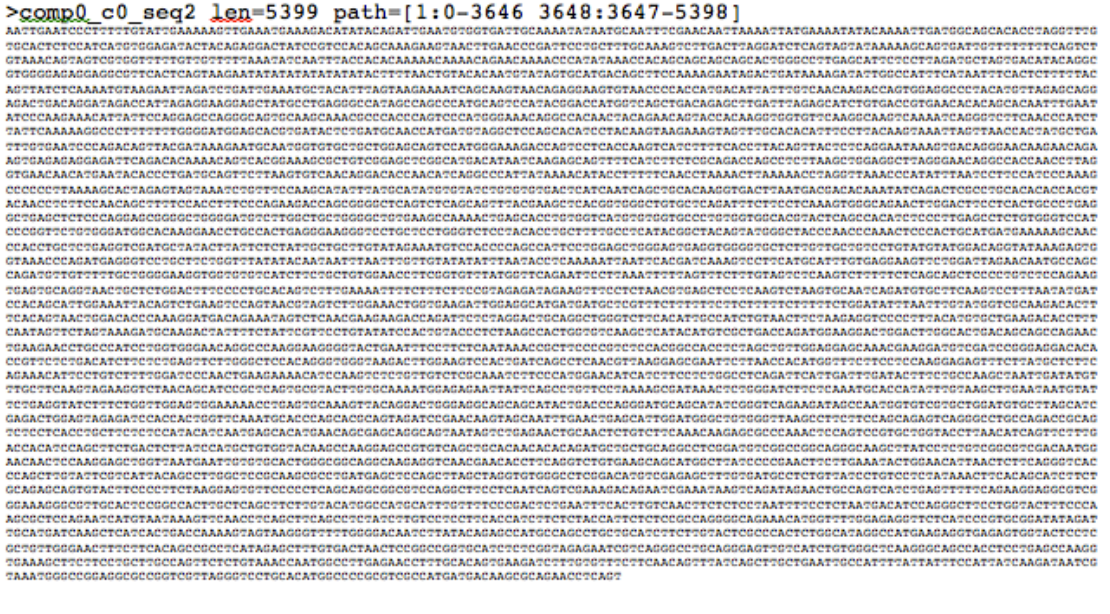


# Trinity output: a multi-fasta file

>comp0_c0_seq1 len=5528 path=[1:0-3646 10775:3647-3775 3648:3776-5527]

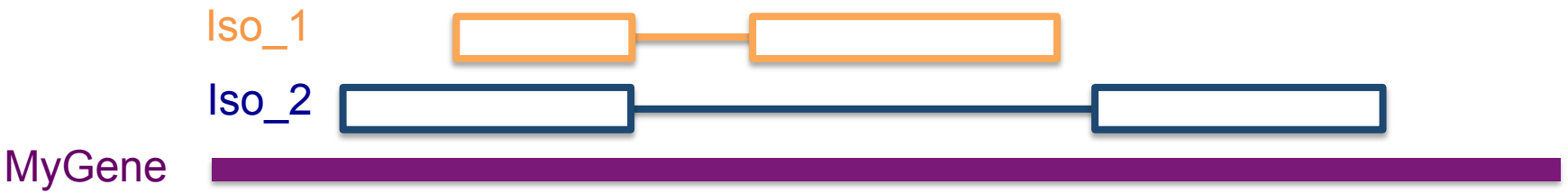
>comp0_c0_seq2 len=5399 path=[1:0-3646 3648:3647-5398]



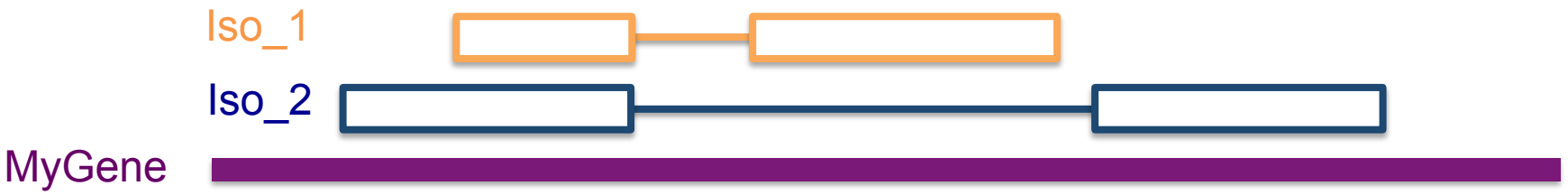
# Flavors of Differential Expression Analyses

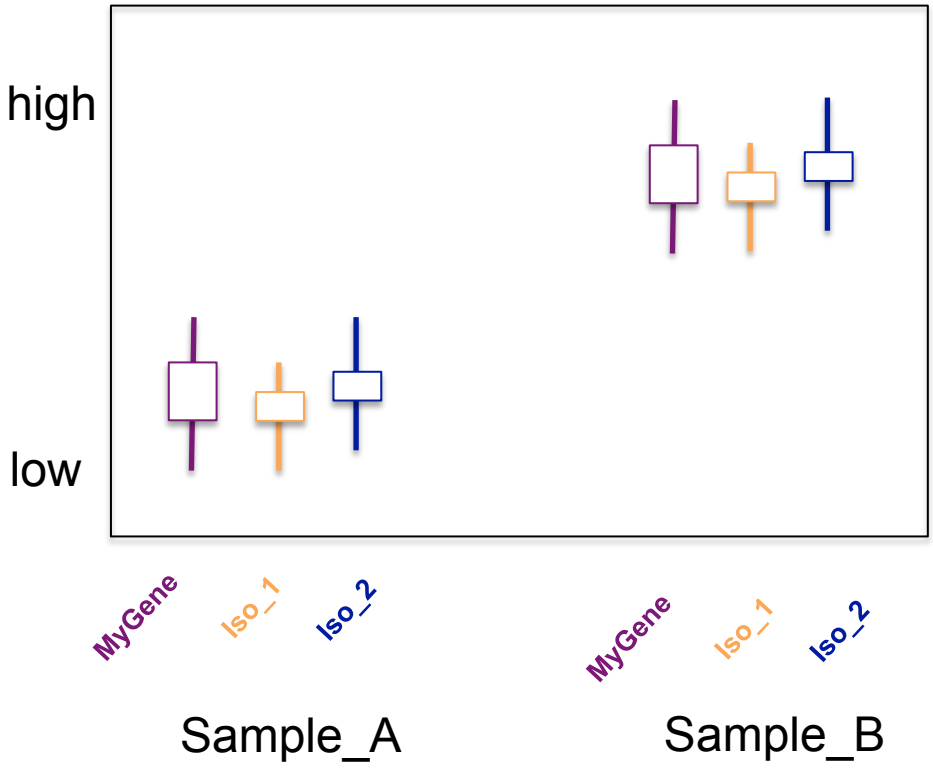
- Differential Gene Expression (DGE)
- Differential Transcript Expression (DTE)
- Differential Transcript Usage (DTU)
- Differential Exon Usage (DEU)

# Differential Gene Expression (DGE) and Differential Transcript Expression (DTE) (Example 1)



# Differential Gene Expression (DGE) and Differential Transcript Expression (DTE) (Example 1)

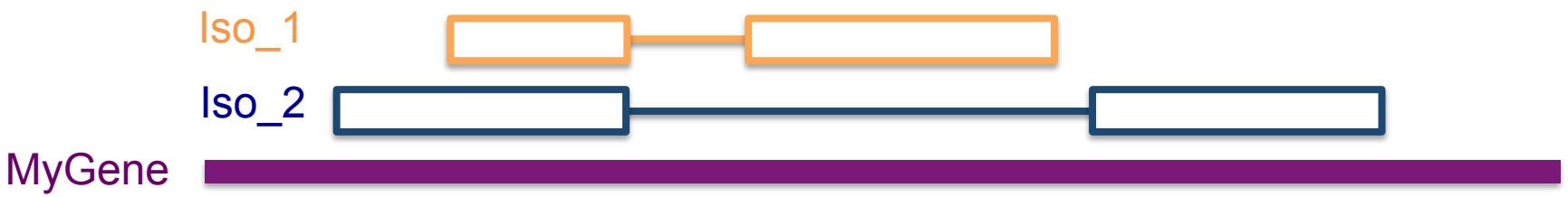


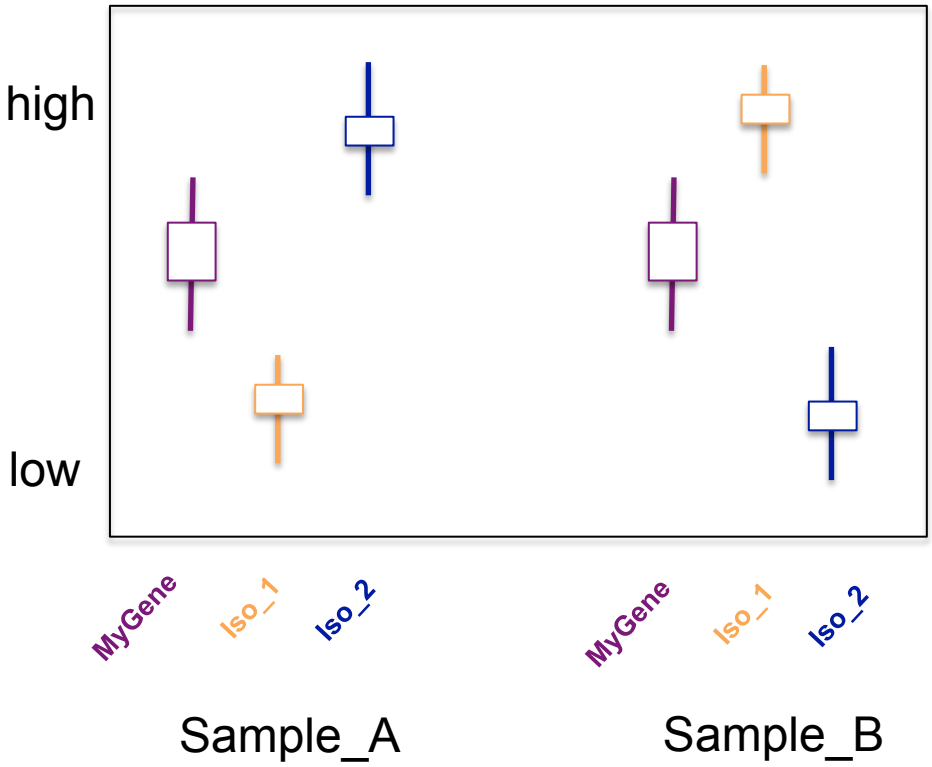




| Feature | Diff Expressed? |
|---|---|
| MyGene | Yes |
| Iso_1 | Yes |
| Iso_2 | Yes |

Diff. Transcript Usage ? (eg. Isoform switching) No

# Differential Gene Expression (DGE) and Differential Transcript Expression (DTE) (Example 2)

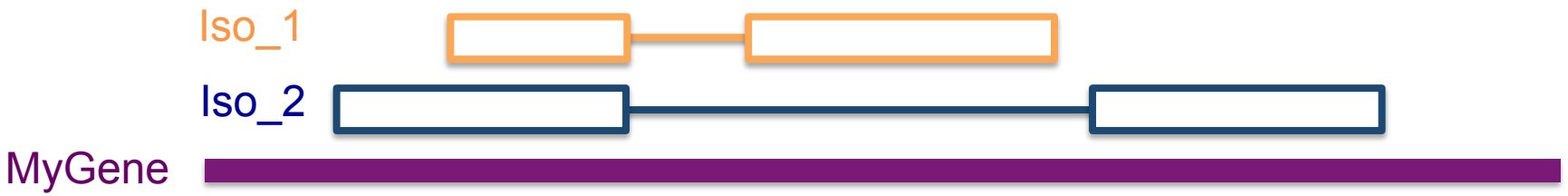


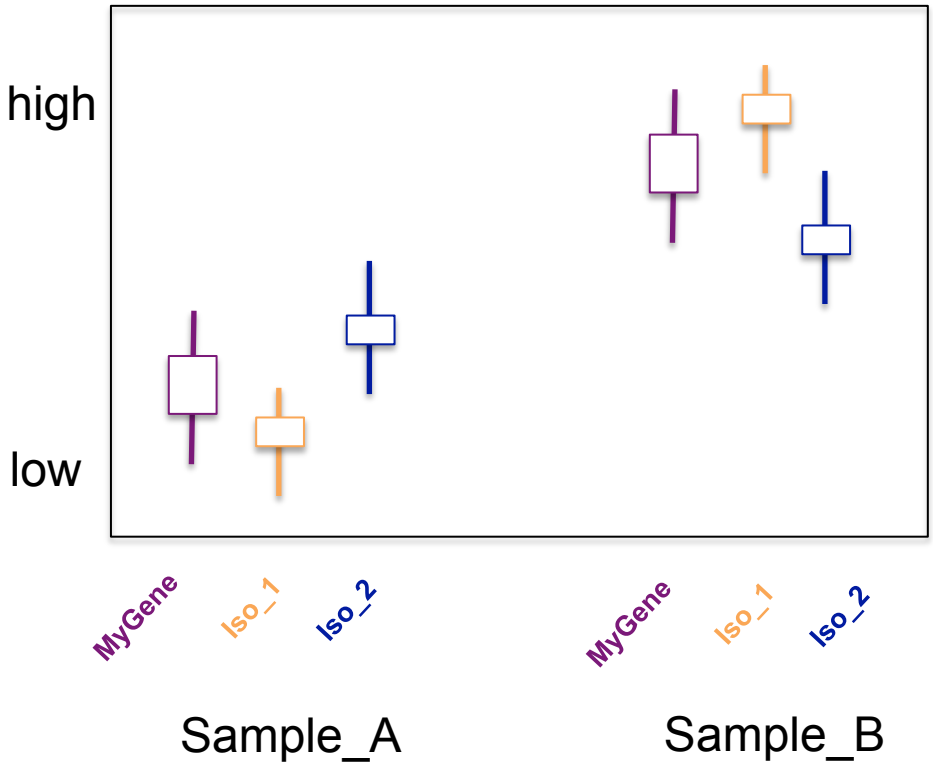




| Feature | Diff Expressed? |
|---|---|
| MyGene | No |
| Iso_1 | Yes |
| Iso_2 | Yes |

Diff. Transcript Usage ? (eg. Isoform switching) Yes

# Differential Gene Expression (DGE) and Differential Transcript Expression (DTE) (Example 3)

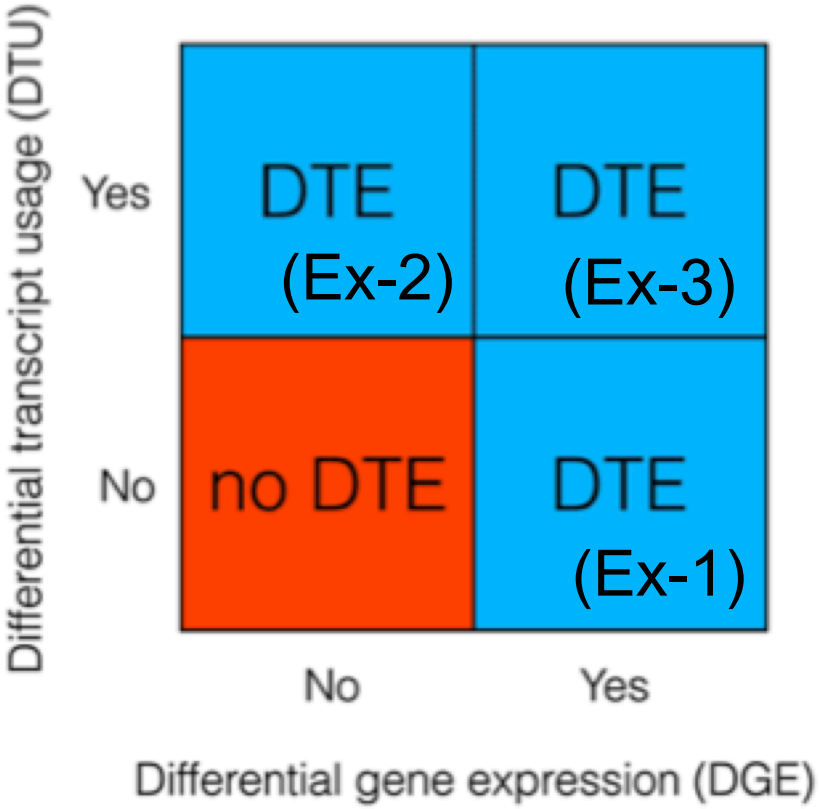



| Feature | Diff Expressed? |
|---|---|
| MyGene | Yes |
| Iso_1 | Yes |
| Iso_2 | Yes |

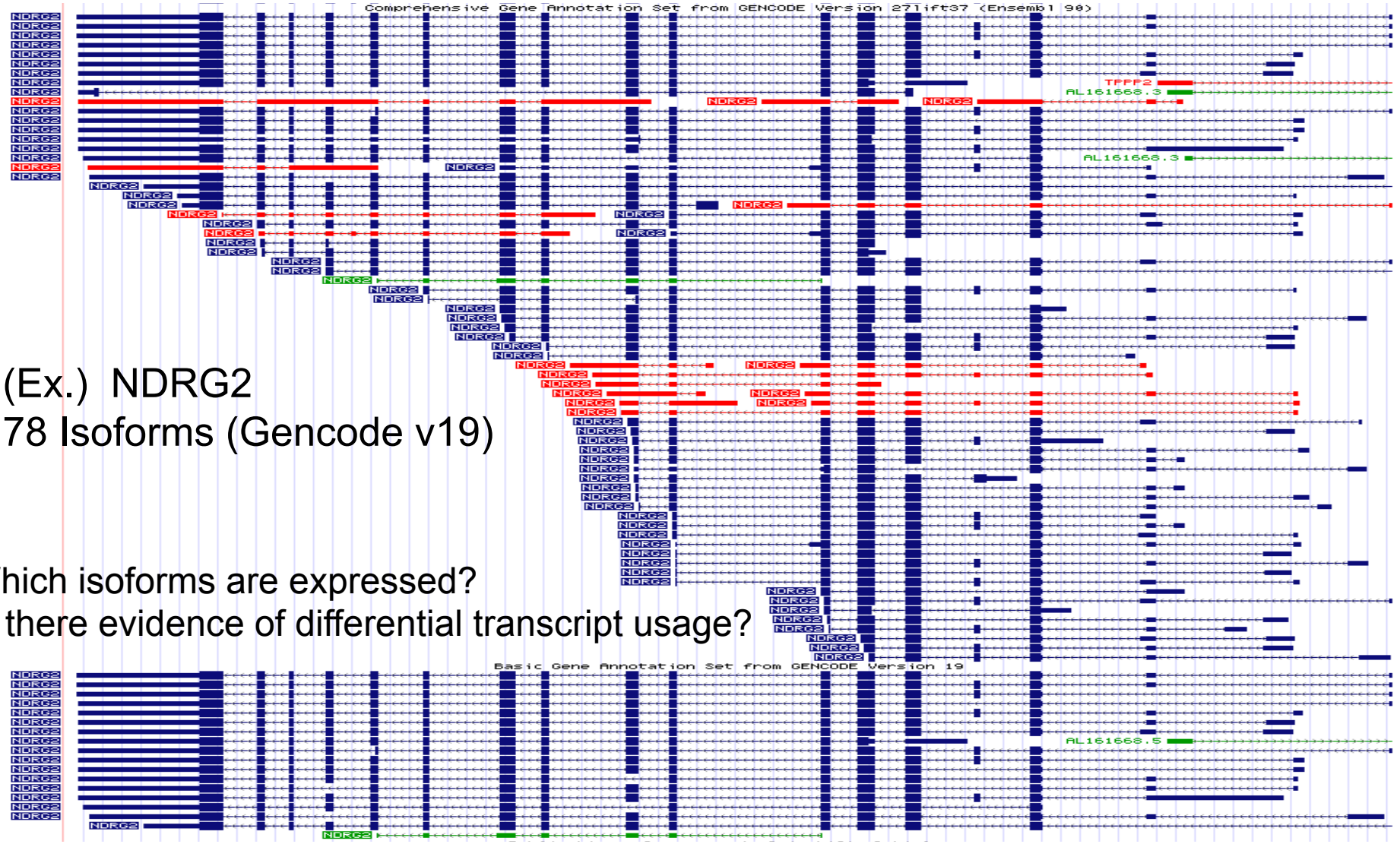
Diff. Transcript Usage ? (eg. Isoform switching) Yes

# Differential Transcript Usage(DTU)
# vs
# Differential Gene Expression (DGE)
# vs.
# Differential Transcript Expression (DTE)




Soneson C, Love MI and Robinson MD. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences [version 2]. F1000Research 2016, 4:1521 (doi: 10.12688/f1000research.7563.2) F1000Research

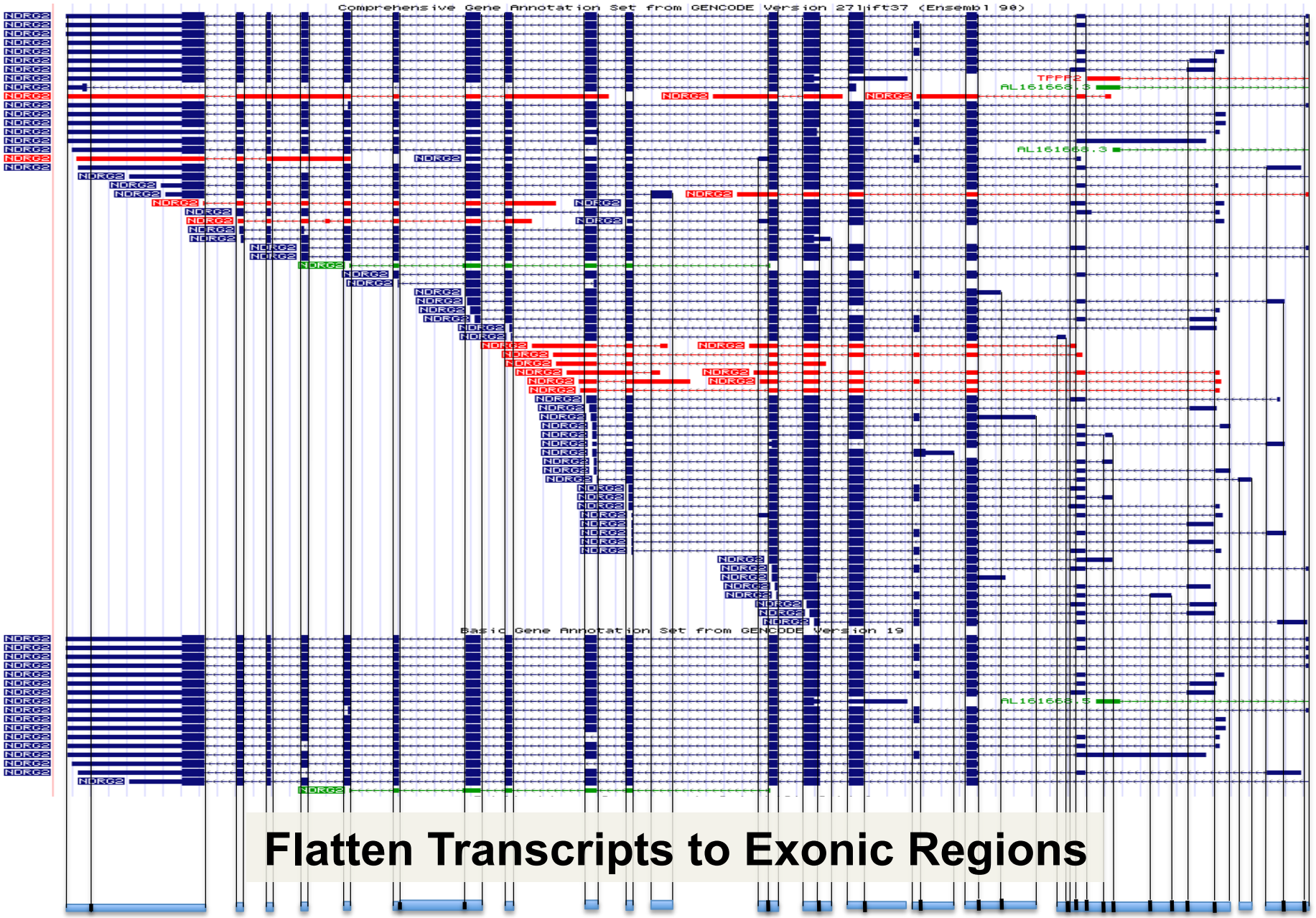
# High Confidence Differential Transcript Expression is Difficult to Attain With Many Candidate Isoforms




(Ex.) NDRG2
78 Isoforms (Gencode v19)

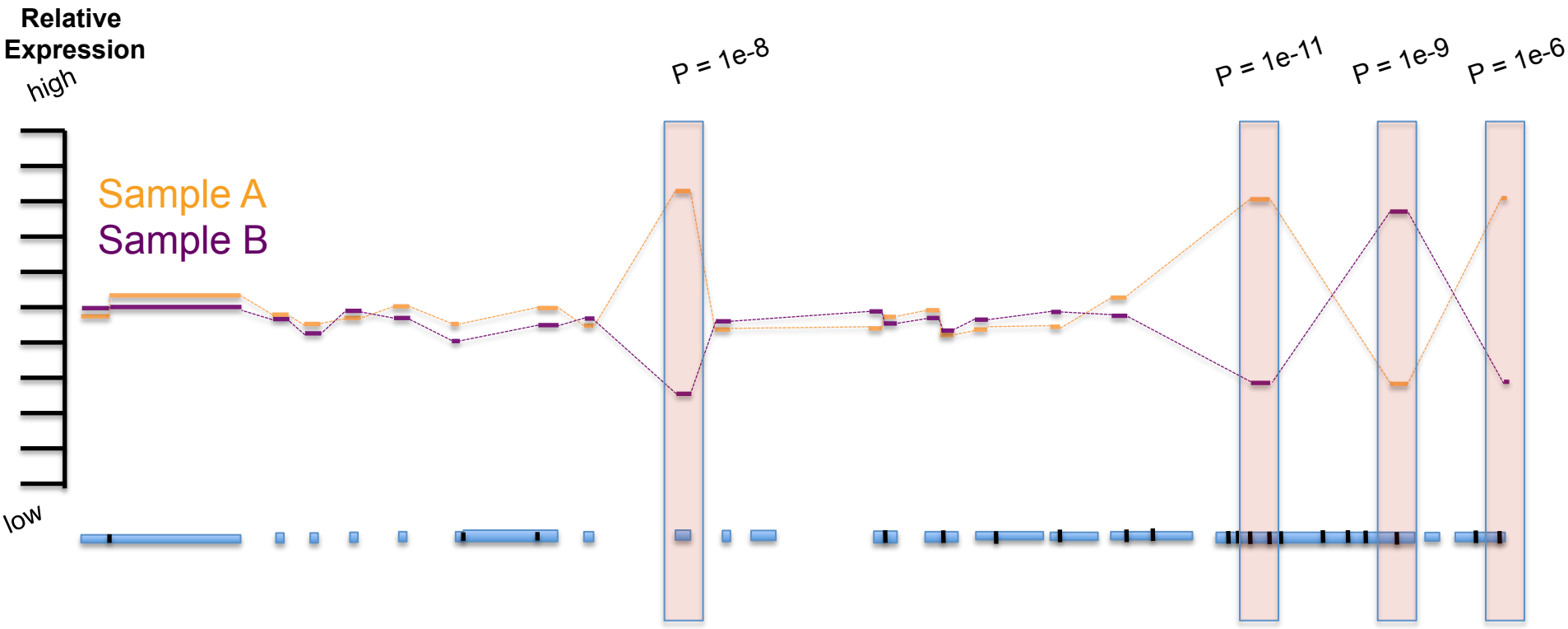
Which isoforms are expressed?
Is there evidence of differential transcript usage?

# Measure Differential Transcript Usage (DTU) via Differential Exon Usage (DEU)




## Flatten Transcripts to Exonic Regions

# Measure Differential Transcript Usage (DTU) via Differential Exon Usage (DEU)



GENOME RESEARCH
CSHL Press | Journal Home | Subscriptions | eTOC Alerts | BioSupplyNet

Genome Res. 2012 Oct; 22(10): 2008–2017. PMCID: PMC3460195
doi: 10.1101/gr.133744.111

**Detecting differential usage of exons from RNA-seq data**

Simon Anders,[1,2] Alejandro Reyes,[1] and Wolfgang Huber

Averaged Replicates

Each Replicate

Flattened gene structure:

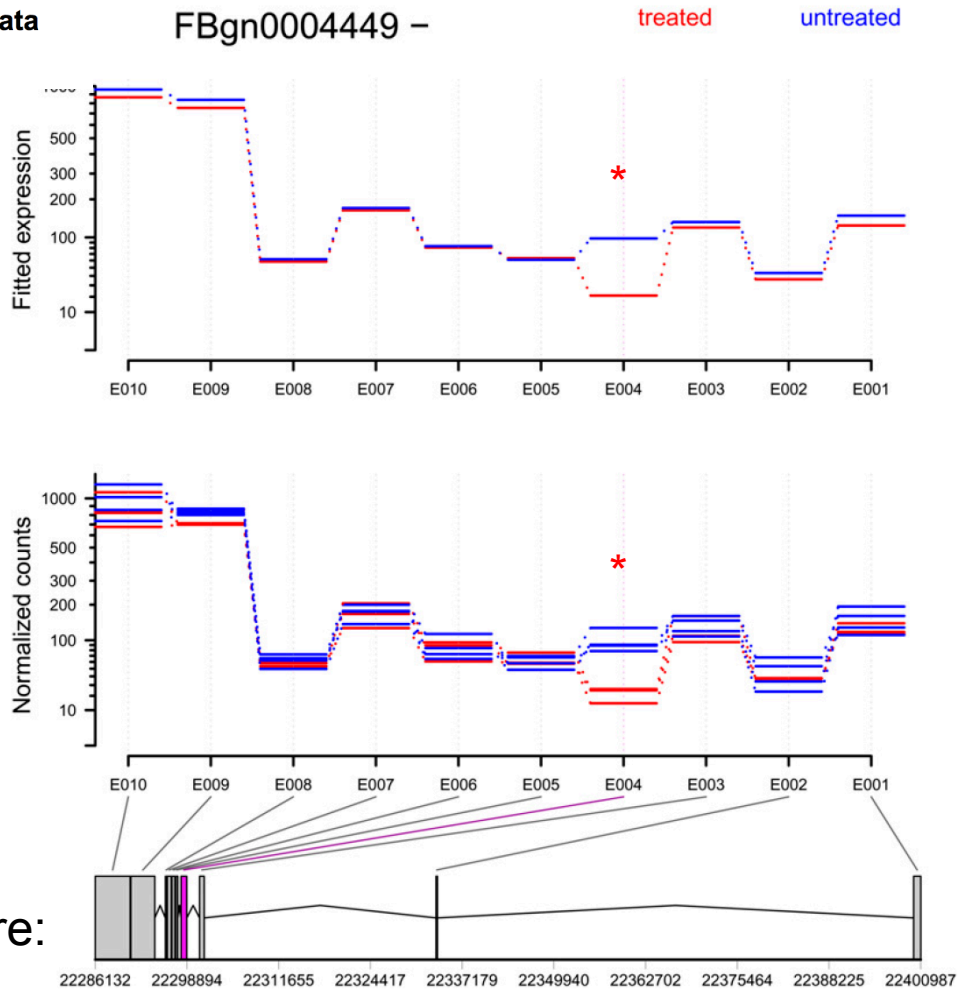



**Figure 3.** The treatment of knocking down the splicing factor *pasilla* affects the fourth exon (counting bin E004) of the gene *Ten-m* (CG5723). (*Top* panel) Fitted values according to the linear model; (*middle* panel) normalized counts for each sample; (*bottom* panel) flattened gene model. (Red) Data for knockdown samples; (blue) control.

# Enabling Differential Transcript Usage Analysis for De novo Transcriptome Assemblies

Davidson *et al. Genome Biology* (2017) 18:148
DOI 10.1186/s13059-017-1284-1

Genome Biology

METHOD Open Access

## SuperTranscripts: a data driven reference for analysis and visualisation of transcriptomes

Nadia M. Davidson[1,2*], Anthony D. K. Hawkins[1] and Alicia Oshlack[1,2*]

# Enabling Differential Transcript Usage Analysis for De novo Transcriptome Assemblies

Davidson *et al. Genome Biology* (2017) 18:148
DOI 10.1186/s13059-017-1284-1

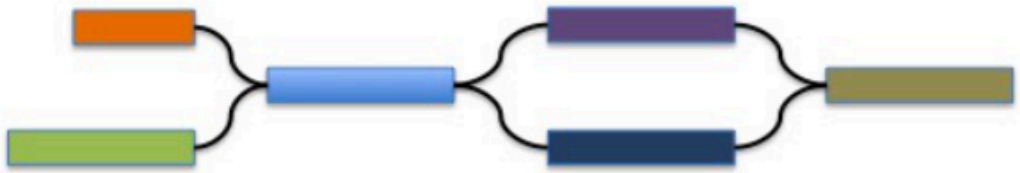
Genome Biology

METHOD Open Access

## SuperTranscripts: a data driven reference for analysis and visualisation of transcriptomes

Nadia M. Davidson[1,2*], Anthony D. K. Hawkins[1] and Alicia Oshlack[1,2*]

Transcript splice graph:



Similar method and protocols now integrated into Trinity:
https://github.com/trinityrnaseq/trinityrnaseq/wiki/SuperTranscripts

# Enabling Differential Transcript Usage Analysis for De novo Transcriptome Assemblies

Davidson *et al. Genome Biology* (2017) 18:148
DOI 10.1186/s13059-017-1284-1

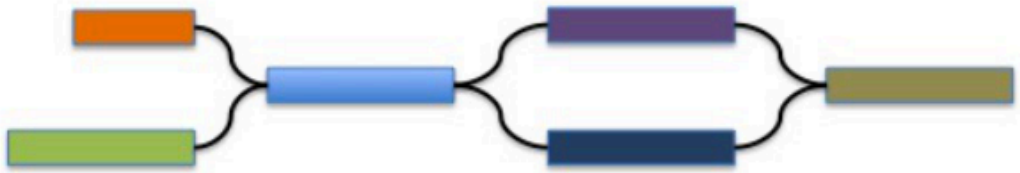
Genome Biology

METHOD Open Access

## SuperTranscripts: a data driven reference for analysis and visualisation of transcriptomes

Nadia M. Davidson[1,2*], Anthony D. K. Hawkins[1] and Alicia Oshlack[1,2*]

Transcript splice graph:



Linearize graph via topological sorting
or graph multiple alignment

SuperTranscript:

DEXseq for DTU,
GATK for Variant Detection

Similar method and protocols now integrated into Trinity:
https://github.com/trinityrnaseq/trinityrnaseq/wiki/SuperTranscripts

# Time for Transcript Reconstruction Lab