bioinformatics.ca

Canadian Bioinformatics Workshops

www.bioinformatics.ca

bioinformaticsdotca.github.io

Supported by

OICR
Ontario Institute for Cancer Research

Creative Commons

This page is available in the following languages:
Afrikaans български Català Dansk Deutsch Ελληνικά English English (CA) English (GB) English (US) Esperanto Castellano Castellano (AR) Español (CL) Castellano (CO) Español (Ecuador) Castellano (MX) Castellano (PE) Euskara Suomeksi français français (CA) Galego עברית hrvatski Magyar Italiano 日本語 한국어 Macedonian Melayu Nederlands Norsk Sesotho sa Leboa polski Português română slovenski jezik српски srpski (latinica) Sotho svenska 中文 華語 (台灣) isiZulu

creative commons

Attribution-Share Alike 2.5 Canada

You are free:

- **to Share** — to copy, distribute and transmit the work
- **to Remix** — to adapt the work

Free Cultural Works APPROVED FOR

Under the following conditions:

- **Attribution**. You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).
- **Share Alike**. If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar licence to this one.

- For any reuse or distribution, you must make clear to others the licence terms of this work.
- Any of the above conditions can be waived if you get permission from the copyright holder.
- The author's moral rights are retained in this licence.

Disclaimer

Your fair dealing and other rights are in no way affected by the above.
This is a human-readable summary of the Legal Code (the full licence) available in the following languages:
English French

Learn how to distribute your work using this licence

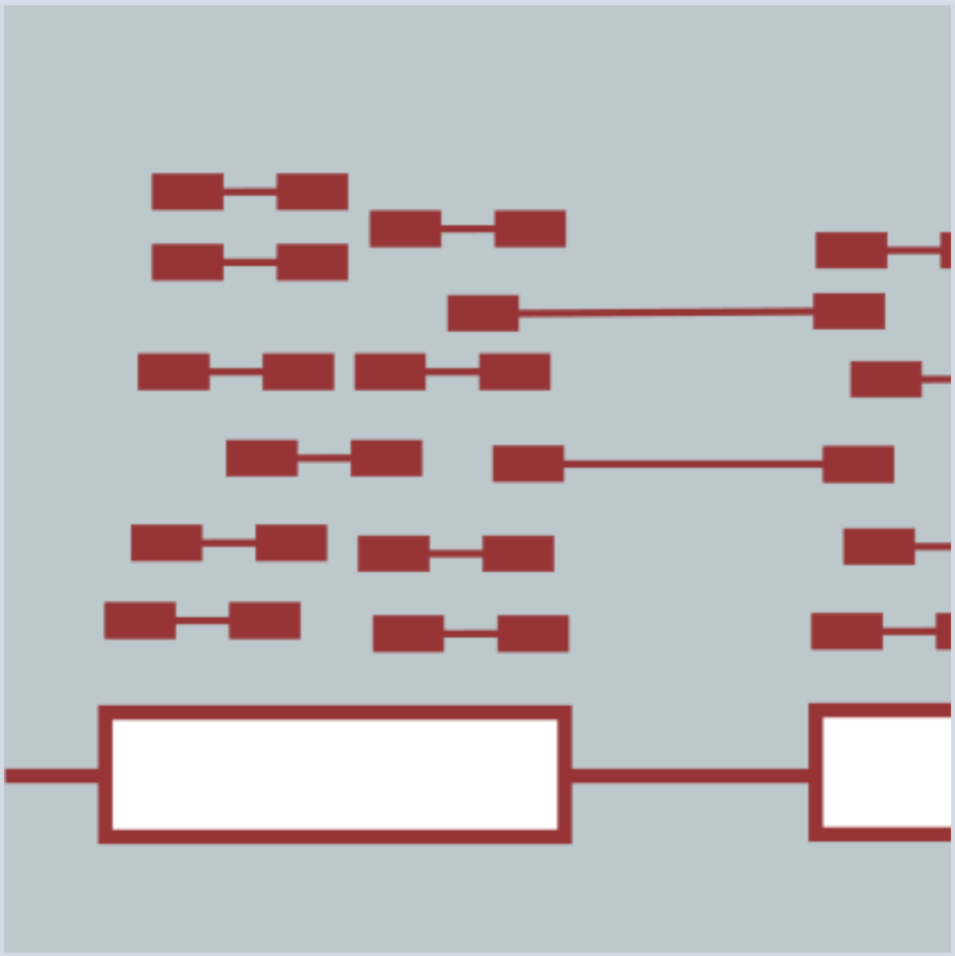
Genome-Guided and Genome-Free Transcriptome Assembly

Brian Haas

Informatics for RNA-Seq Analysis

June 11-13, 2019

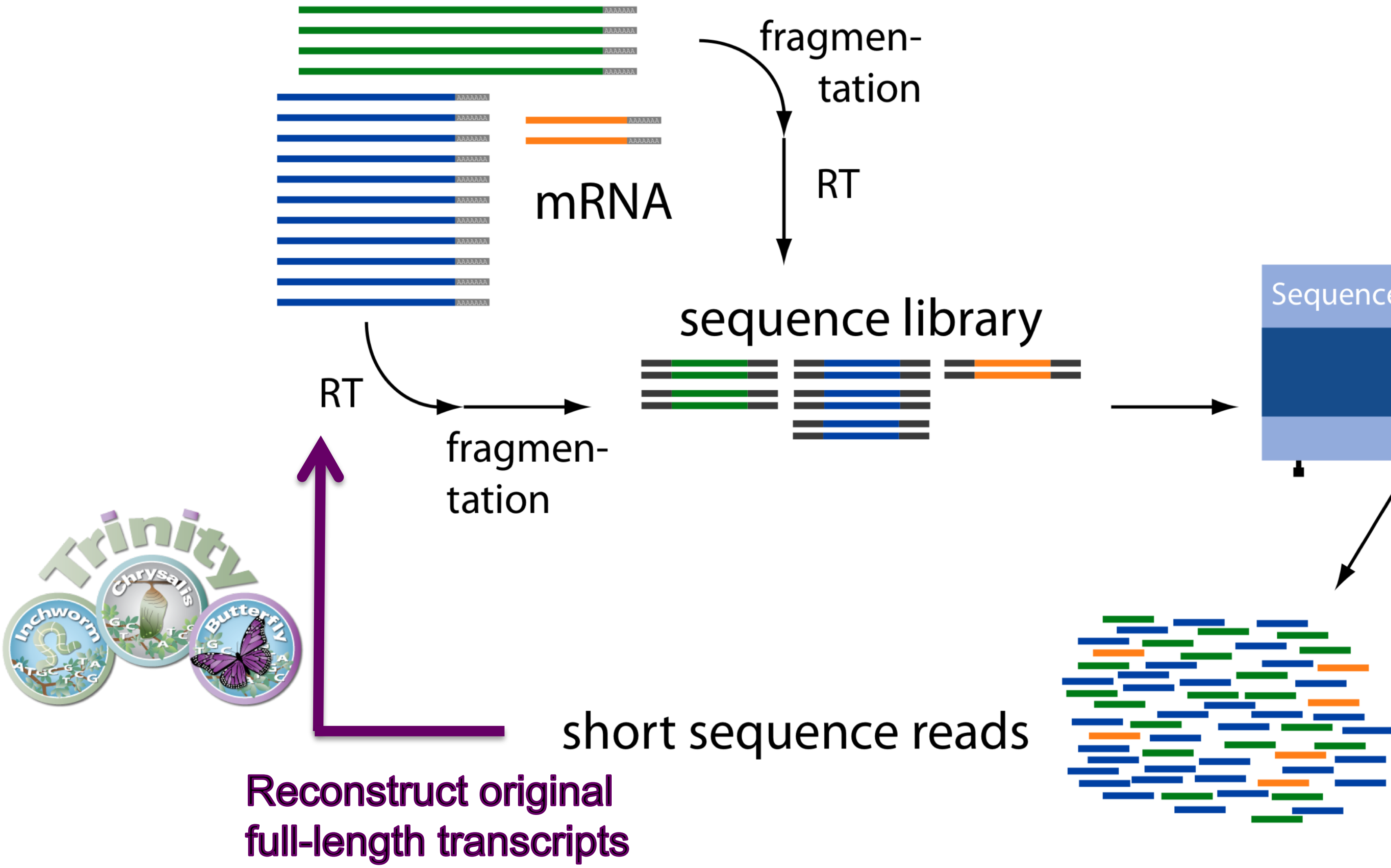
bioinformatics.ca



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

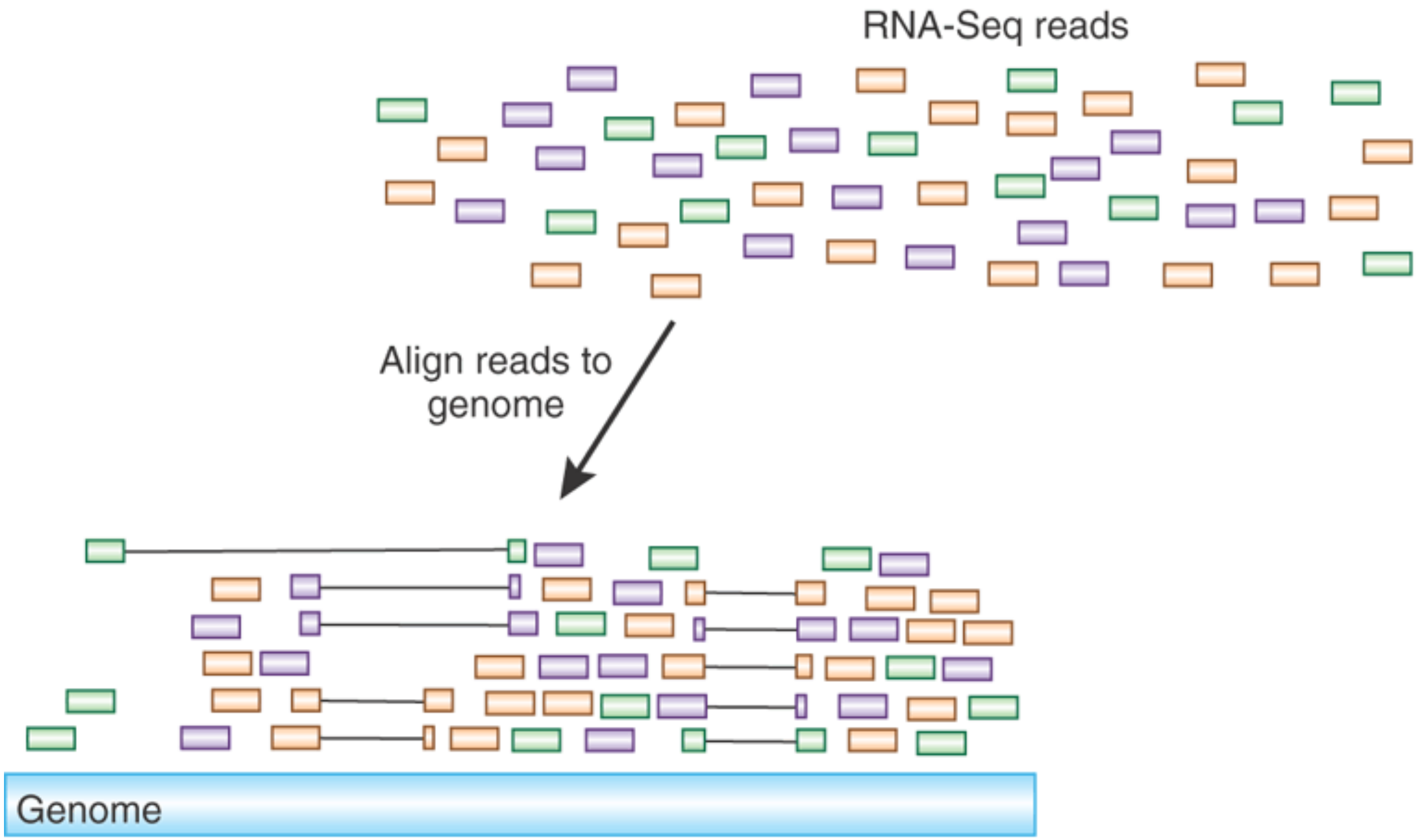
Advancing RNA-Seq analysis

Brian J Haas & Michael C Zody

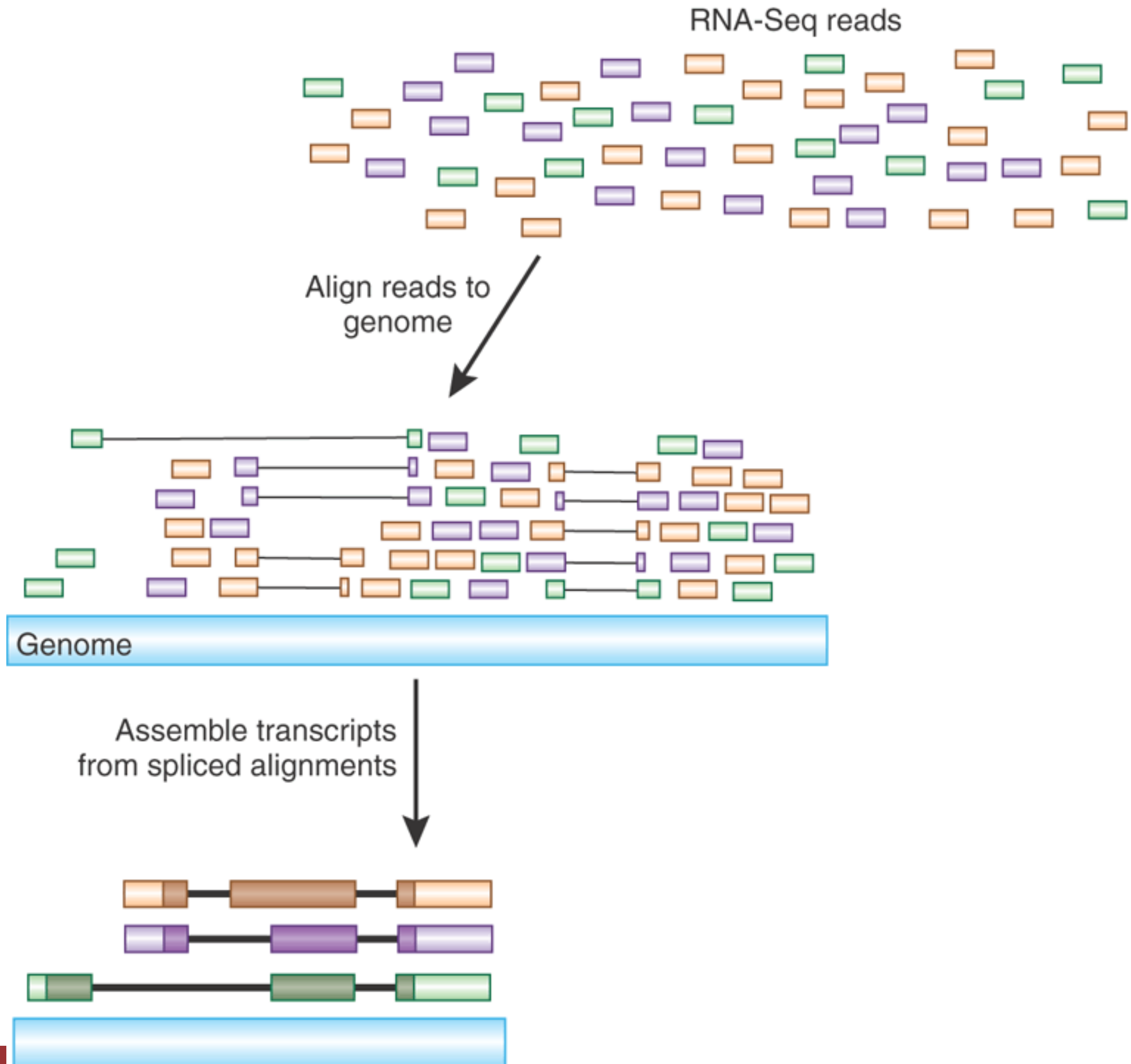
Nature Biotech, 2010

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

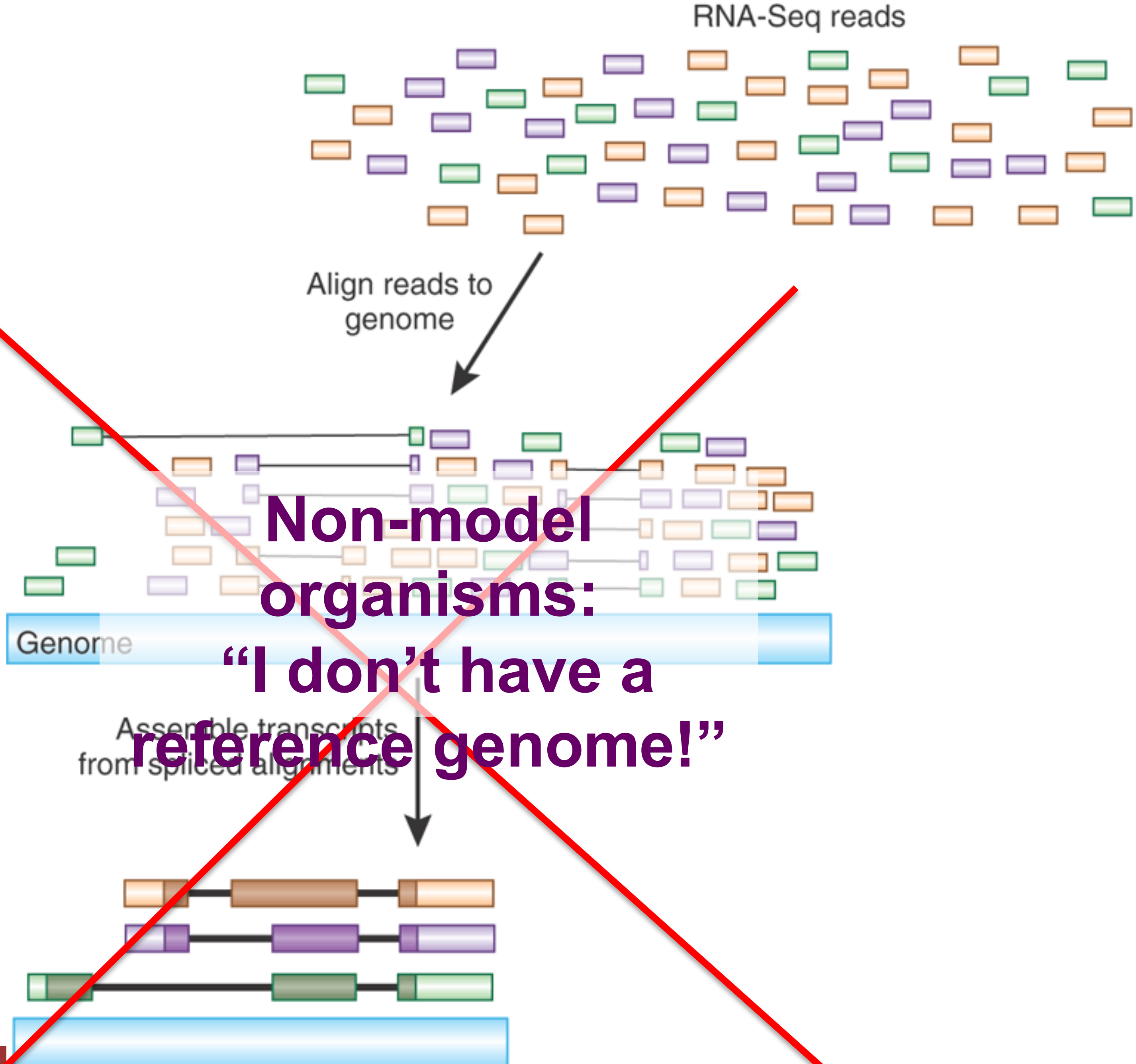
Transcript Reconstruction from RNA-Seq Reads



Transcript Reconstruction from RNA-Seq Reads



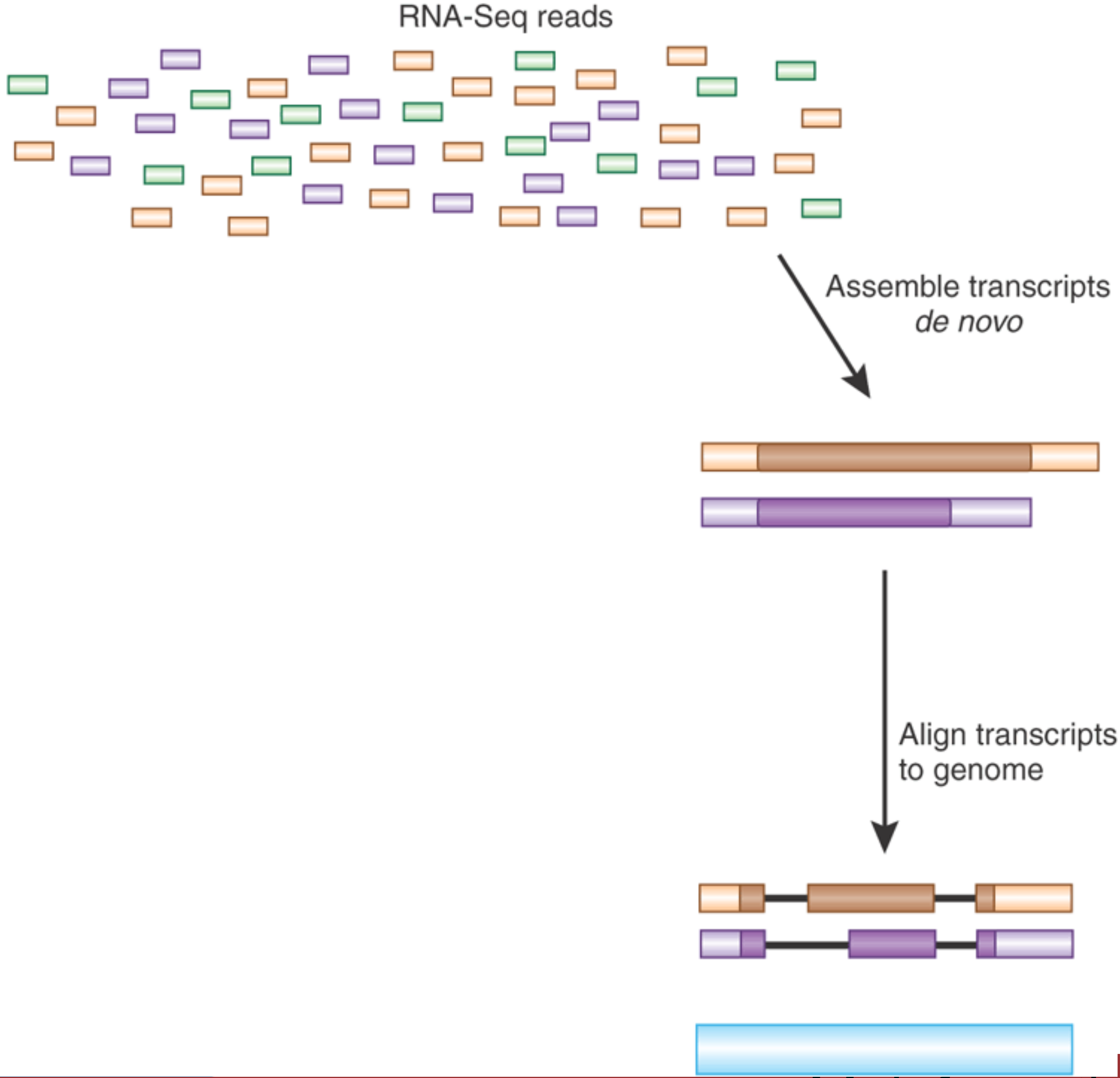
Transcript Reconstruction from RNA-Seq Reads



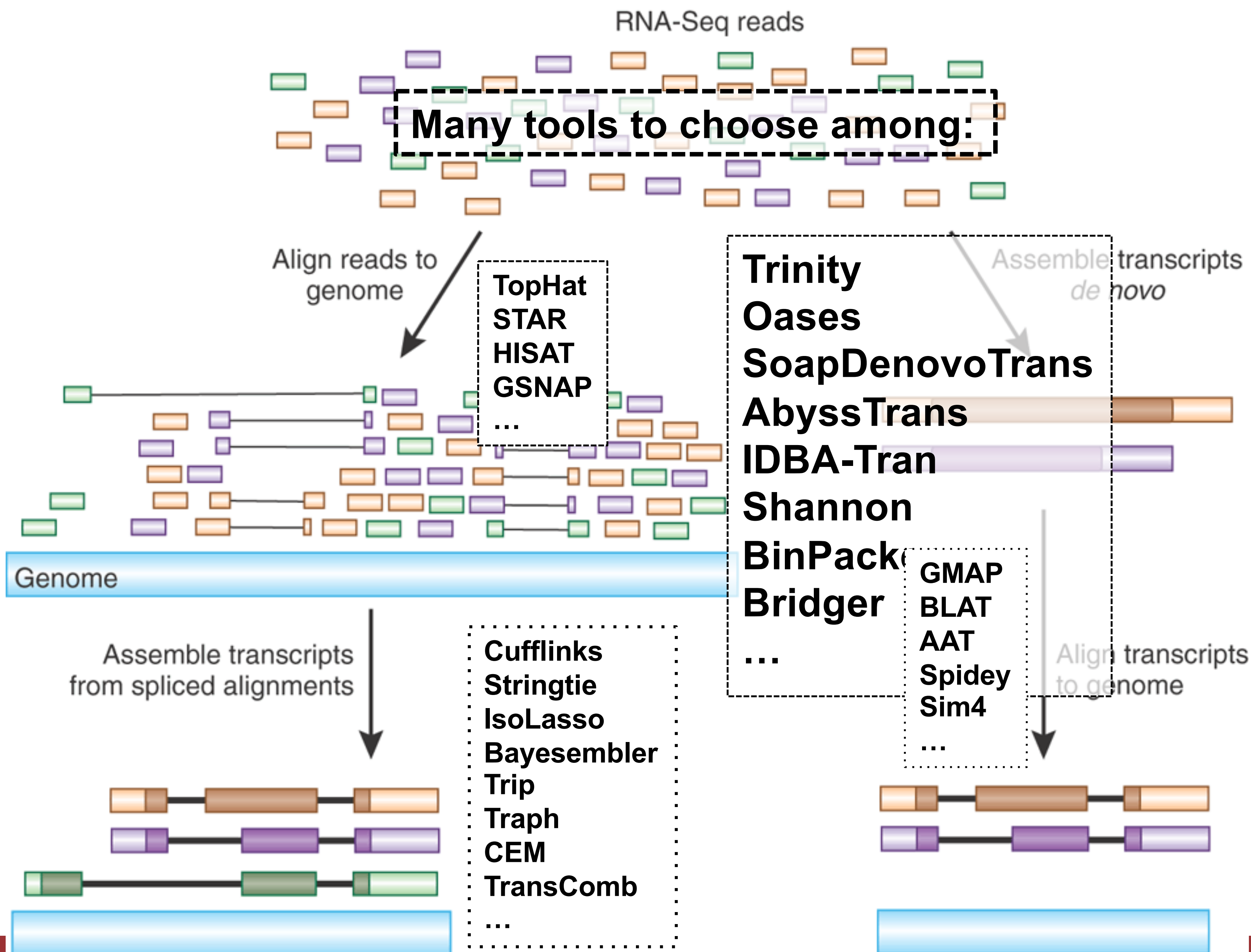
Non-model organisms: "I don't have a reference genome!"

Transcript Reconstruction from RNA-Seq Reads

Transcript Reconstruction from RNA-Seq Reads

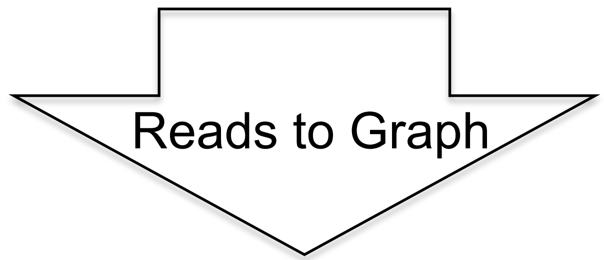


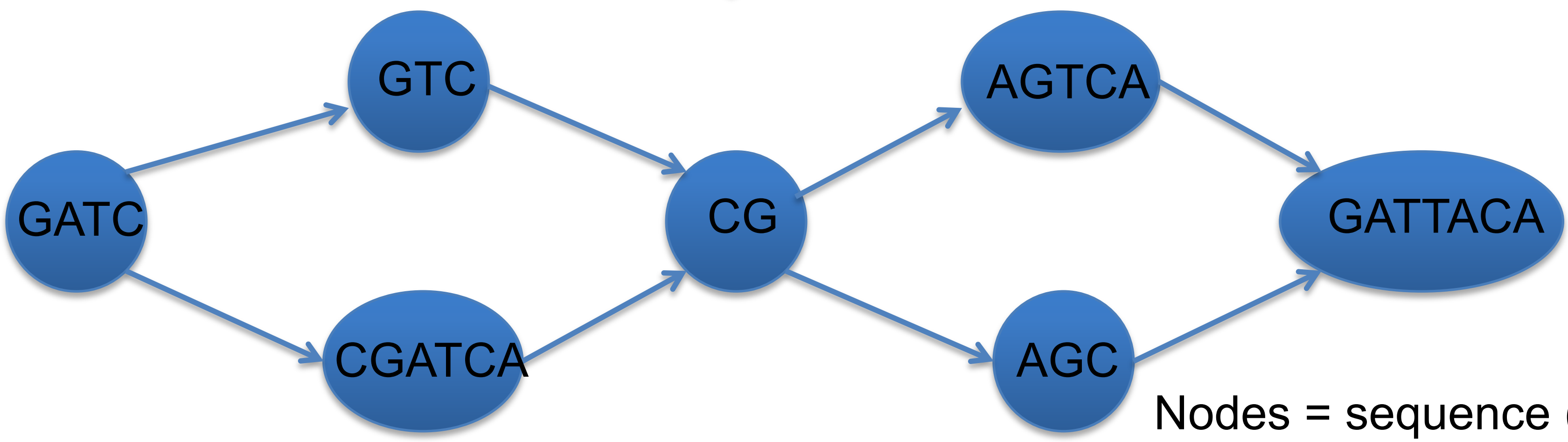
Transcript Reconstruction from RNA-Seq Reads



Graph Data Structures Commonly Used For Assembly

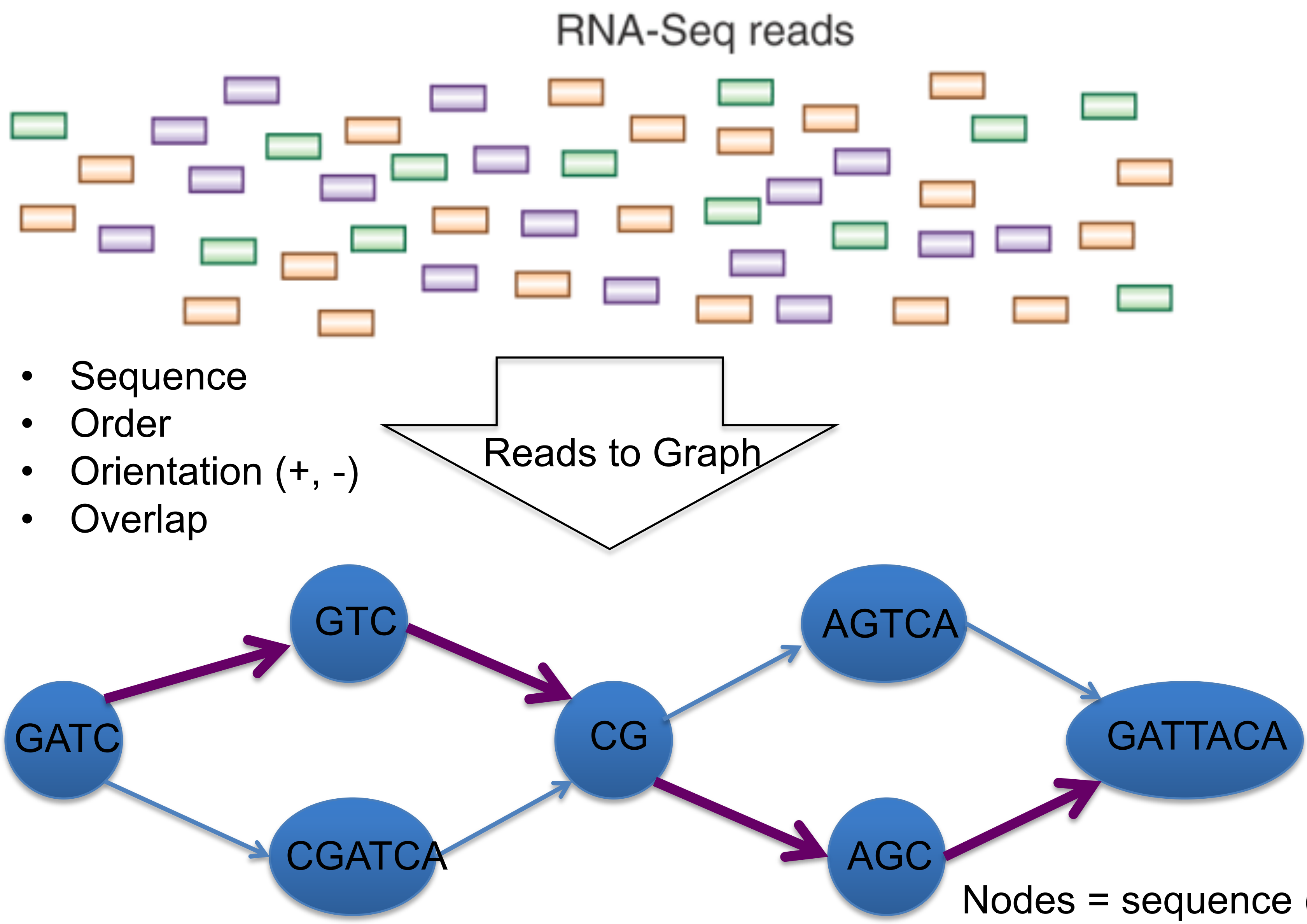
- Sequence
- Order
- Orientation (+, -)
- Overlap





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

Graph Data Structures Commonly Used For Assembly



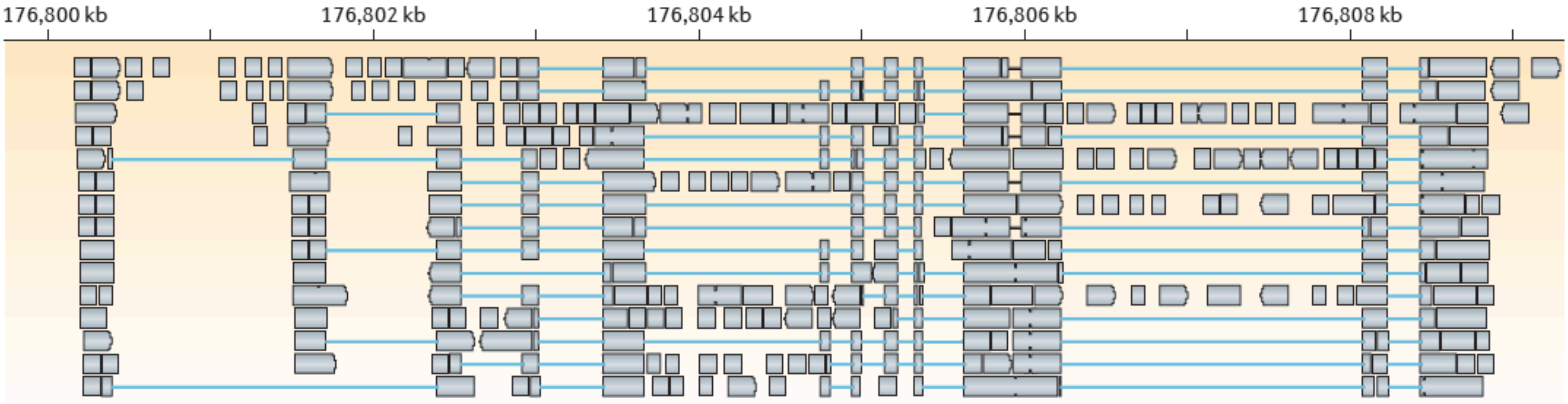
- Sequence
- Order
- Orientation (+, -)
- Overlap

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

GATCGTCCGAGCGATTACA

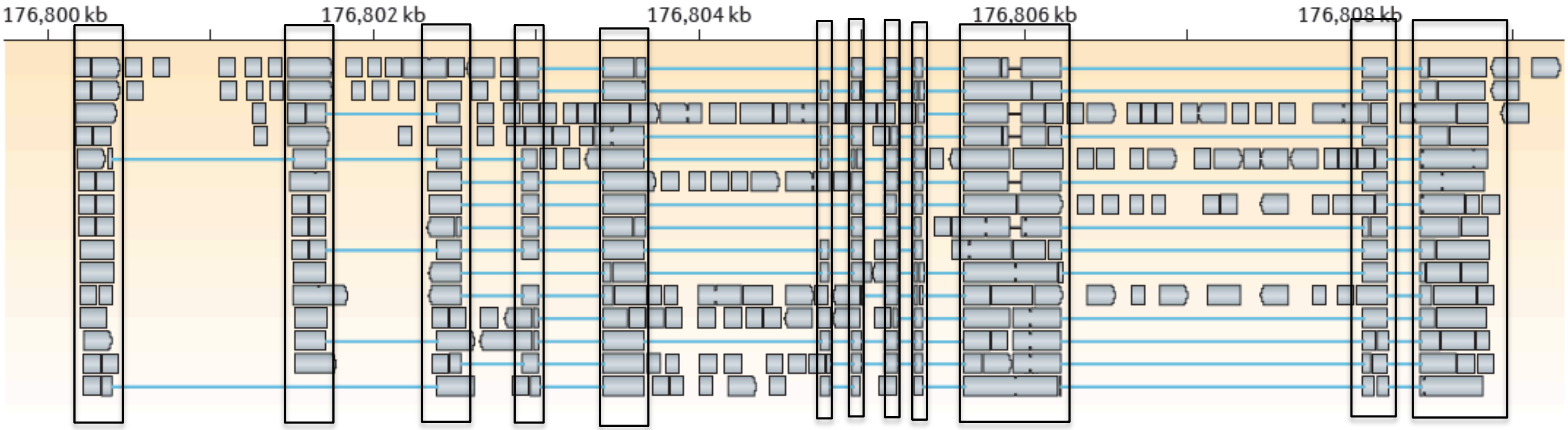
Genome-Guided Transcript Reconstruction

Splice-align reads to the genome



Genome-Guided Transcript Reconstruction

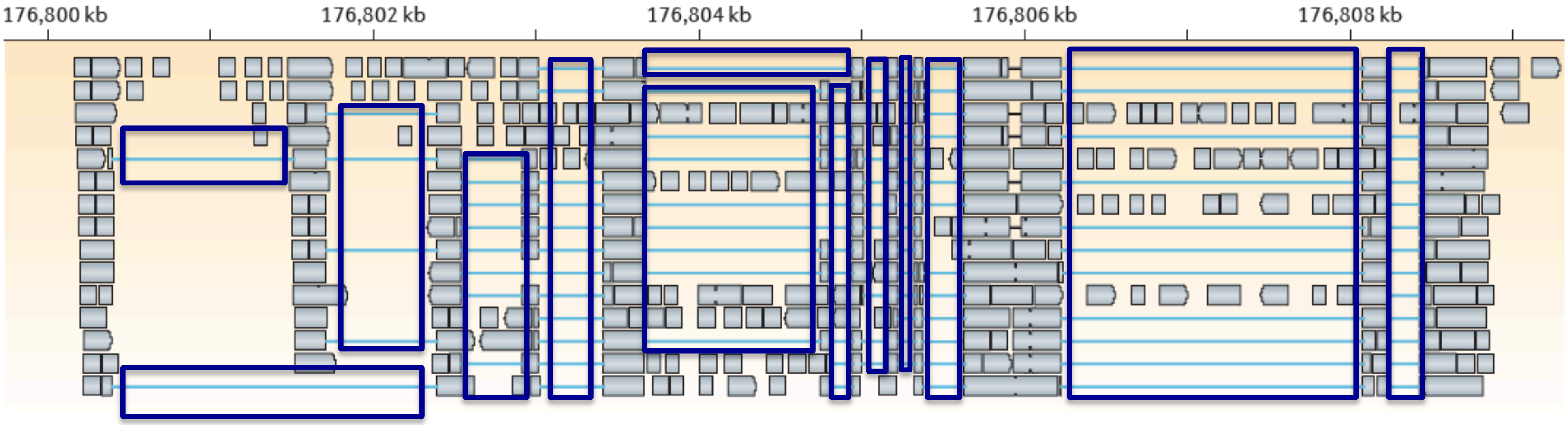
Splice-align reads to the genome



Alignment segment piles => exon regions

Genome-Guided Transcript Reconstruction

Splice-align reads to the genome

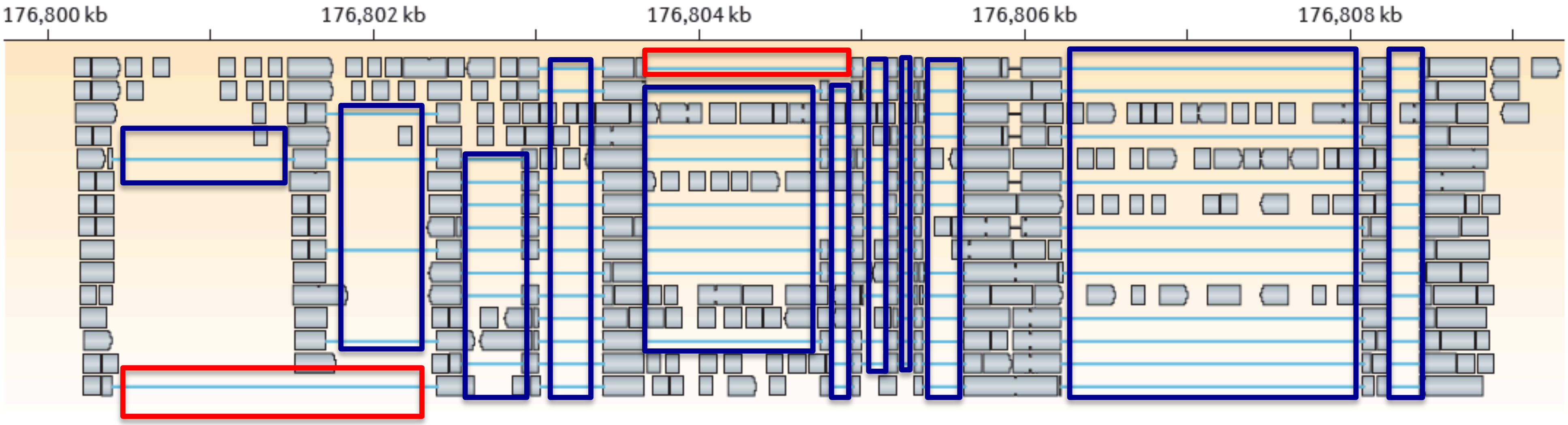


Large alignment gaps => introns

From Martin & Wang. Nature Reviews in Genetics. 2011

Genome-Guided Transcript Reconstruction

Splice-align reads to the genome

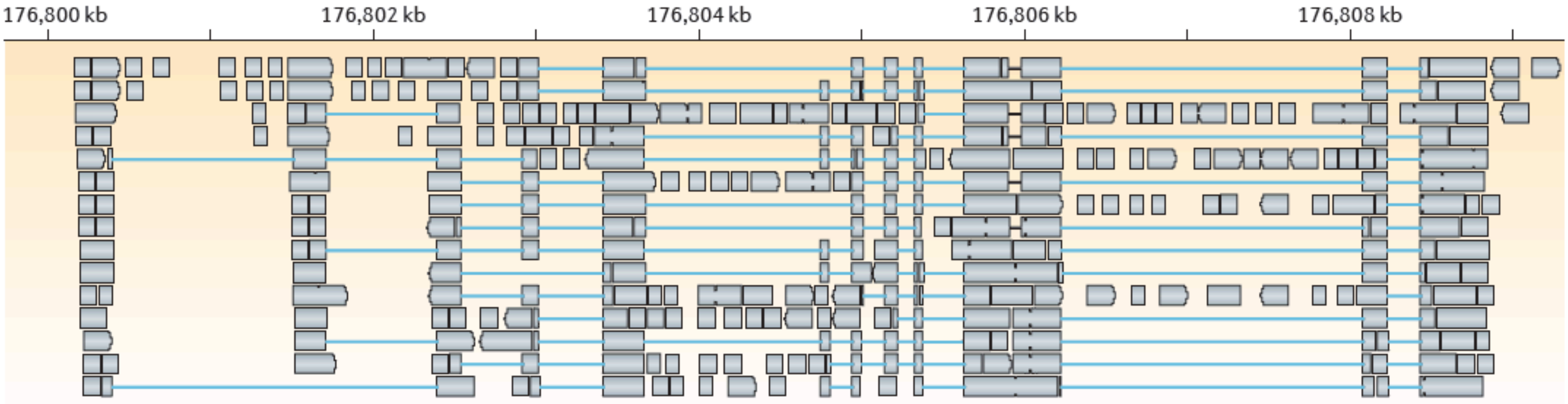


Overlapping but different introns = evidence of alternative splicing

From Martin & Wang. Nature Reviews in Genetics. 2011

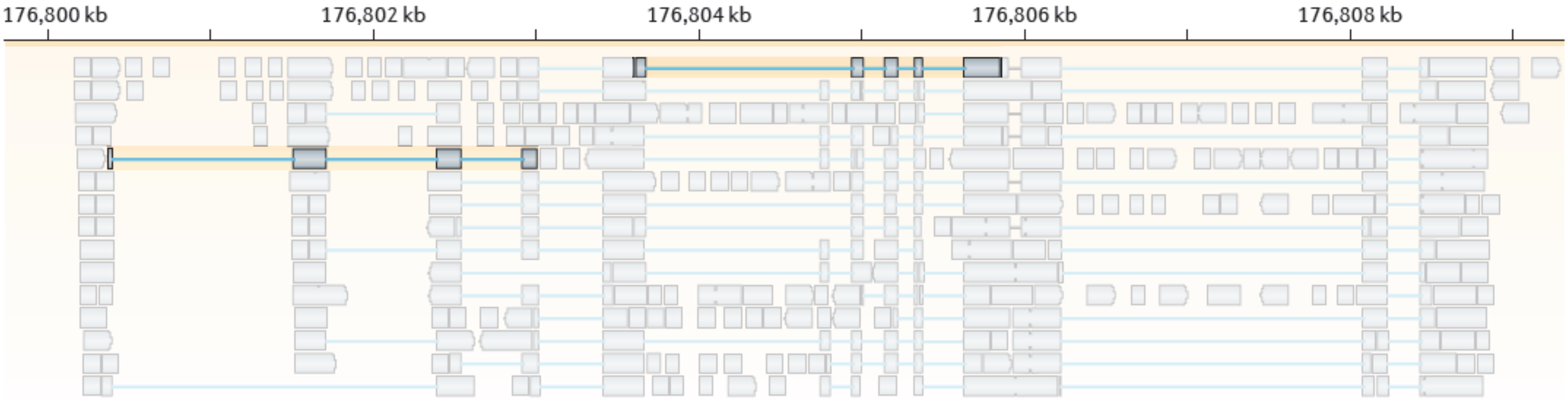
Genome-Guided Transcript Reconstruction

Splice-align reads to the genome



Genome-Guided Transcript Reconstruction

Splice-align reads to the genome

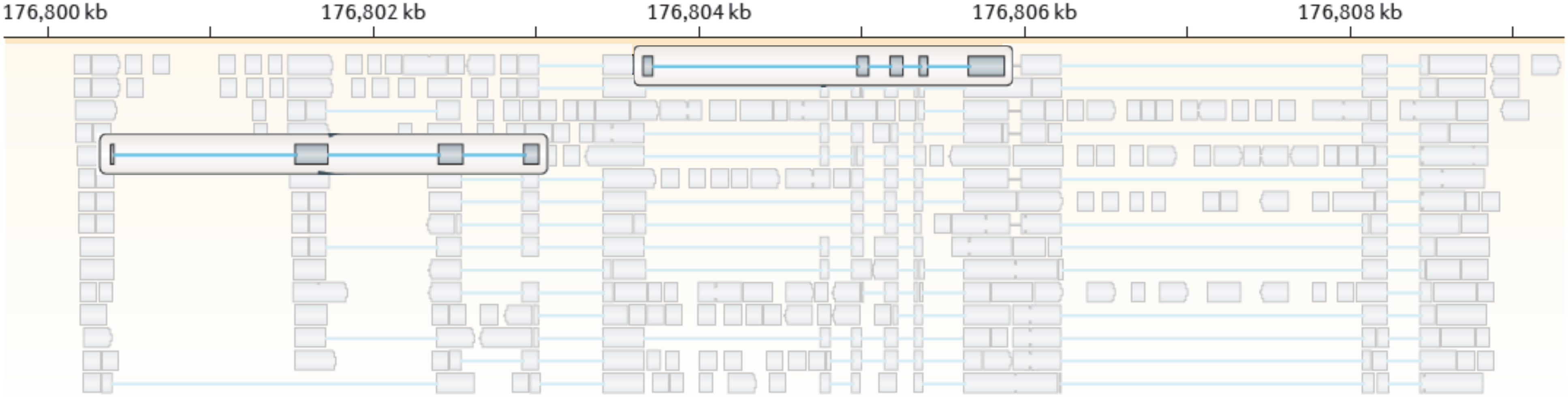


Individual reads can yield multiple exon and intron segments (splice patterns)

From Martin & Wang. Nature Reviews in Genetics. 2011

Genome-Guided Transcript Reconstruction

Splice-align reads to the genome

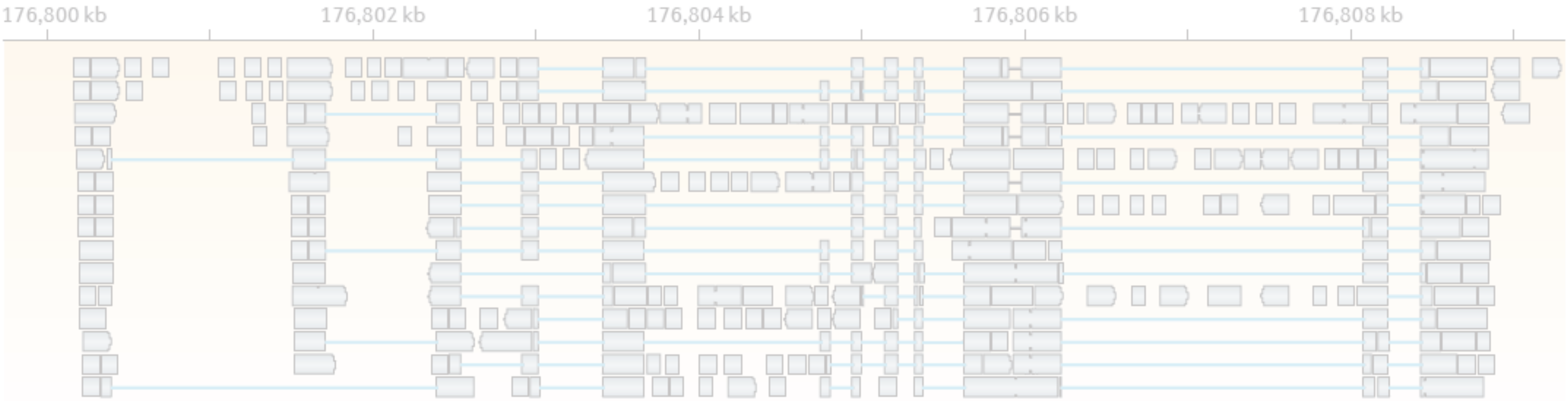


Nodes = unique splice patterns

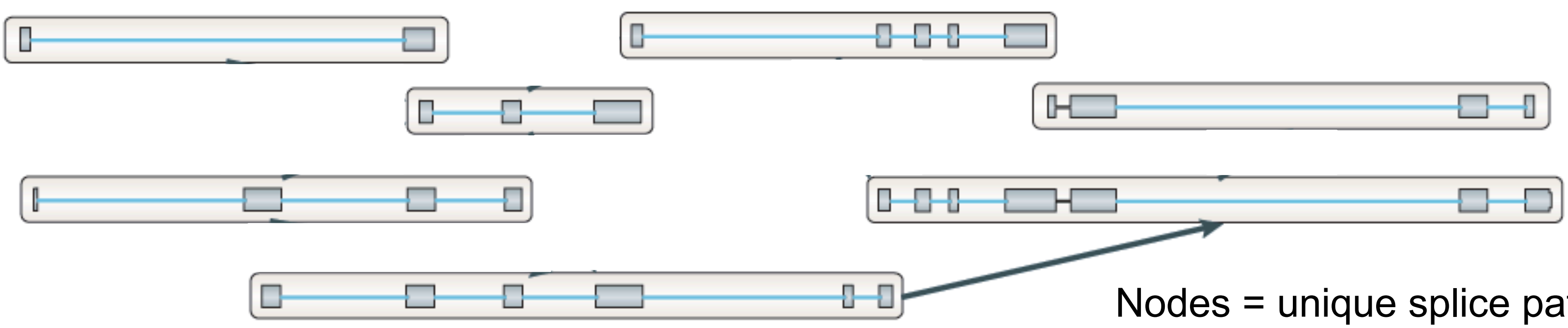
From Martin & Wang. Nature Reviews in Genetics. 2011

Genome-Guided Transcript Reconstruction

Splice-align reads to the genome



Construct graph from unique splice patterns of aligned reads.

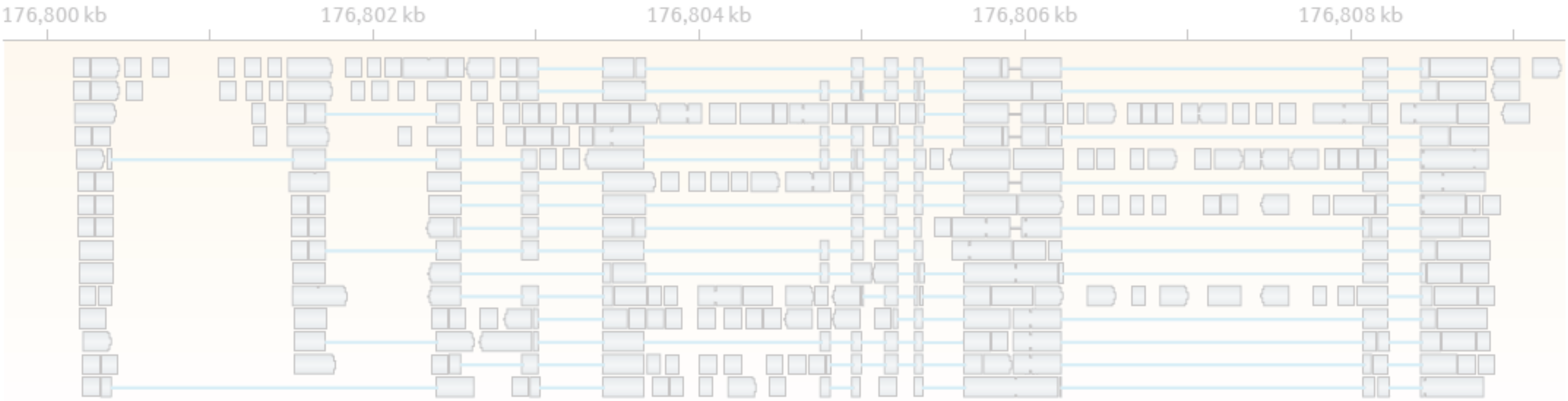


Nodes = unique splice patterns
Edges = compatible patterns

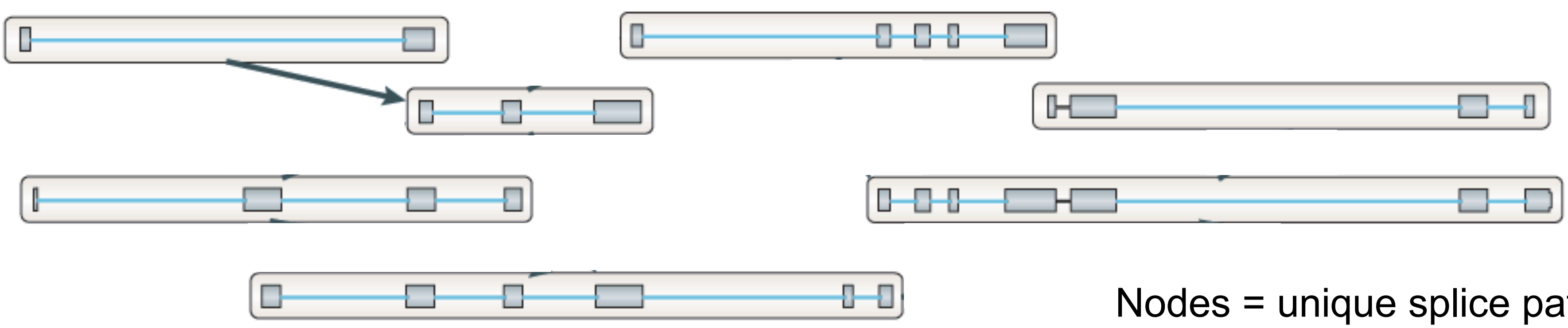
From Martin & Wang. Nature Reviews in Genetics. 2011

Genome-Guided Transcript Reconstruction

Splice-align reads to the genome



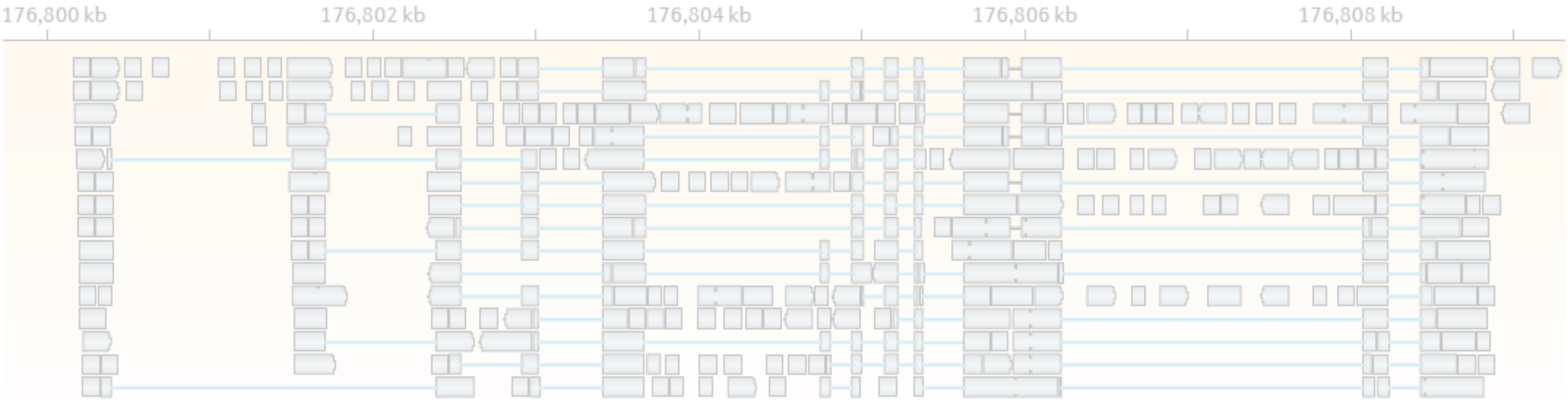
Construct graph from unique splice patterns of aligned reads.



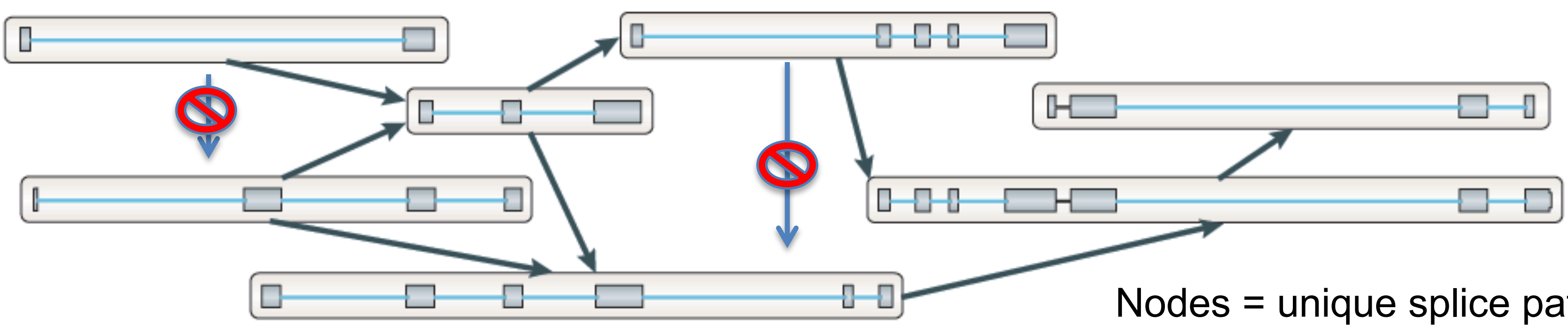
Nodes = unique splice patterns
Edges = compatible patterns

Genome-Guided Transcript Reconstruction

Splice-align reads to the genome



Construct graph from unique splice patterns of aligned reads.

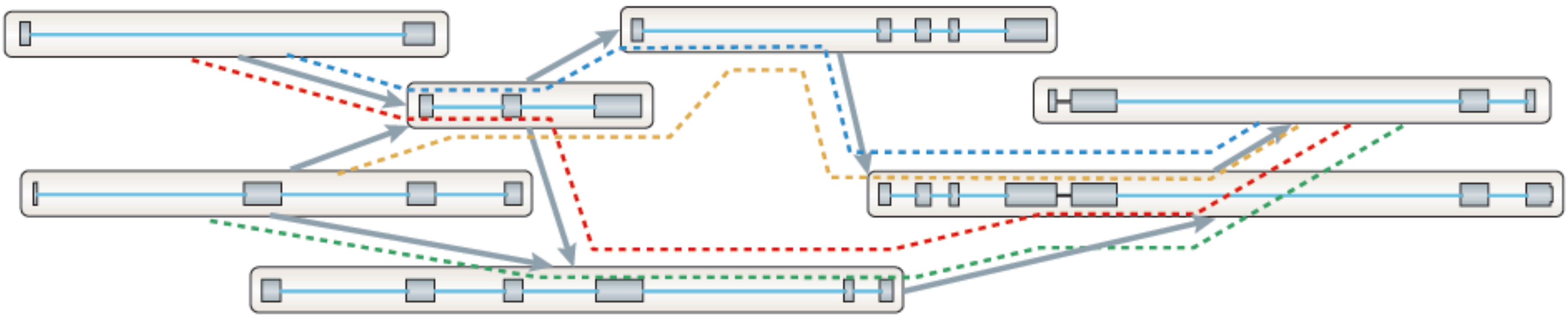


Nodes = unique splice patterns
Edges = compatible patterns

From Martin & Wang. Nature Reviews in Genetics. 2011

Genome-Guided Transcript Reconstruction

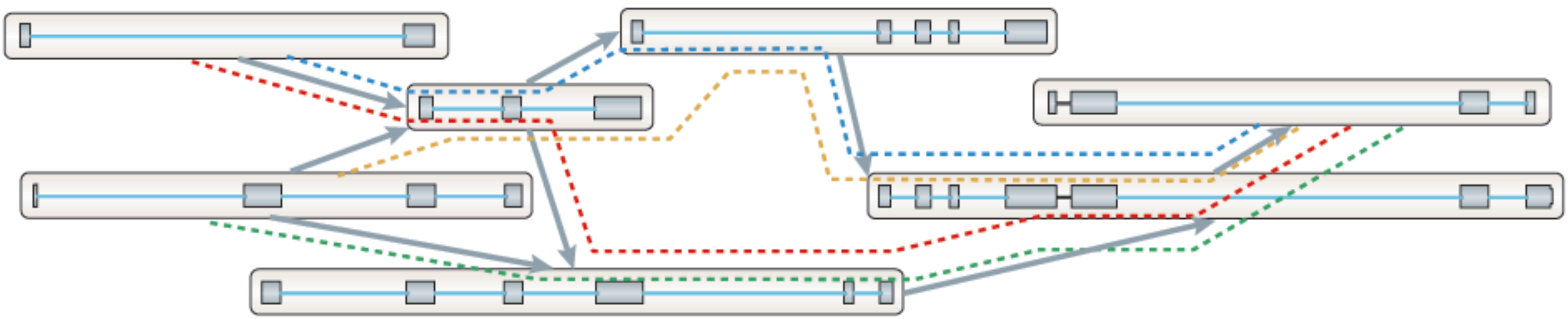
Traverse paths through the graph to assemble transcript isoforms



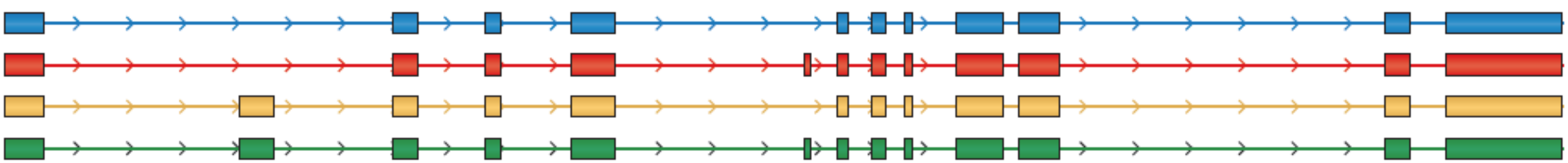
From Martin & Wang. Nature Reviews in Genetics. 2011

Genome-Guided Transcript Reconstruction

Traverse paths through the graph to assemble transcript isoforms



Reconstructed isoforms

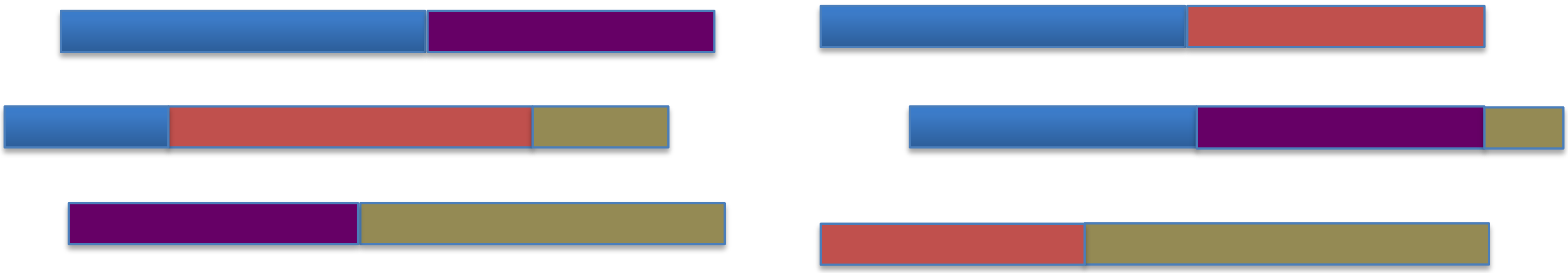


From Martin & Wang. Nature Reviews in Genetics. 2011

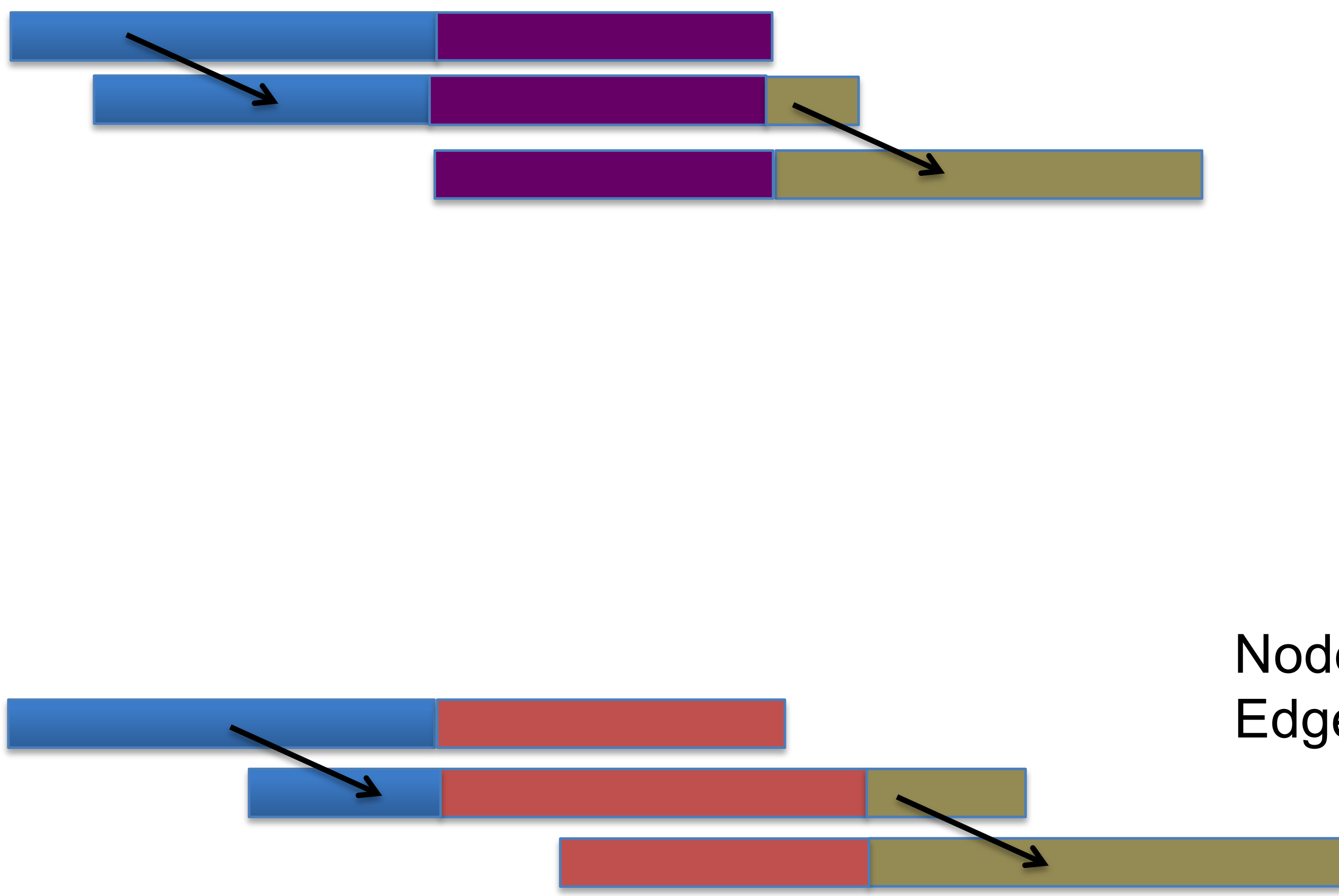
What if you don't have a high quality reference genome sequence?

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



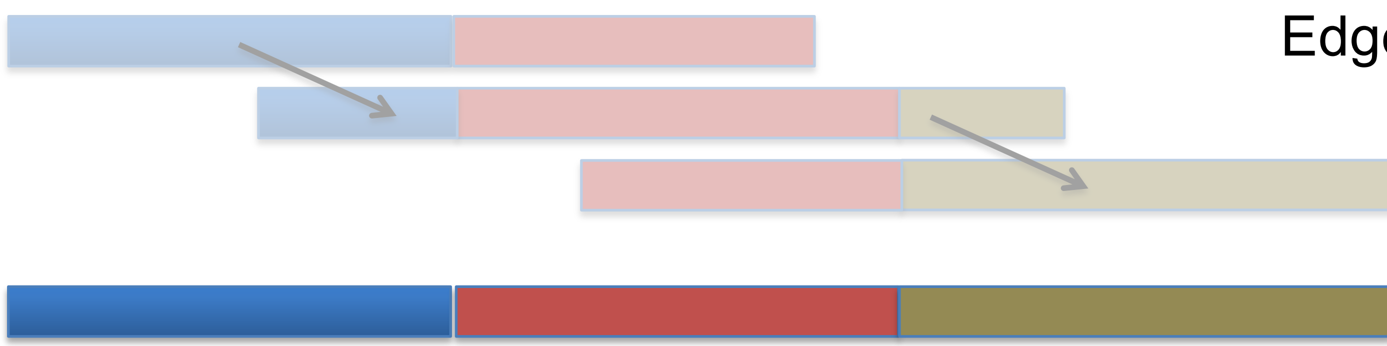
Node = read
Edge = overlap

Read Overlap Graph: Reads as nodes, overlaps as edges

Transcript A

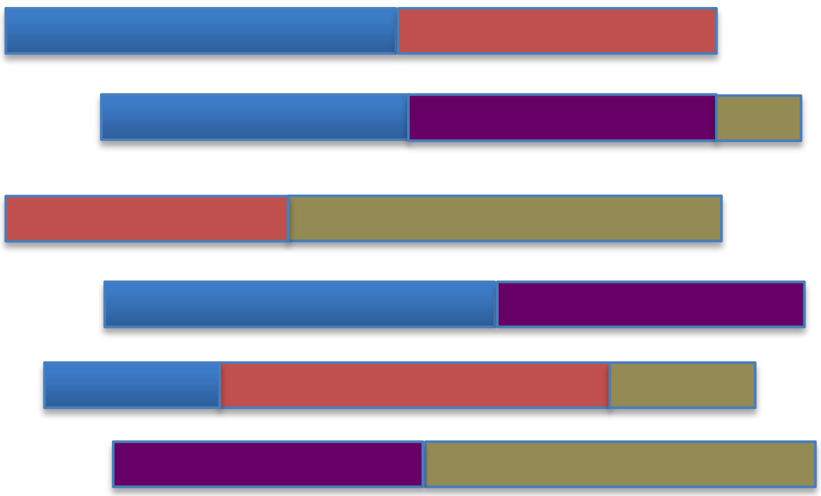
Generate consensus sequence where reads overlap

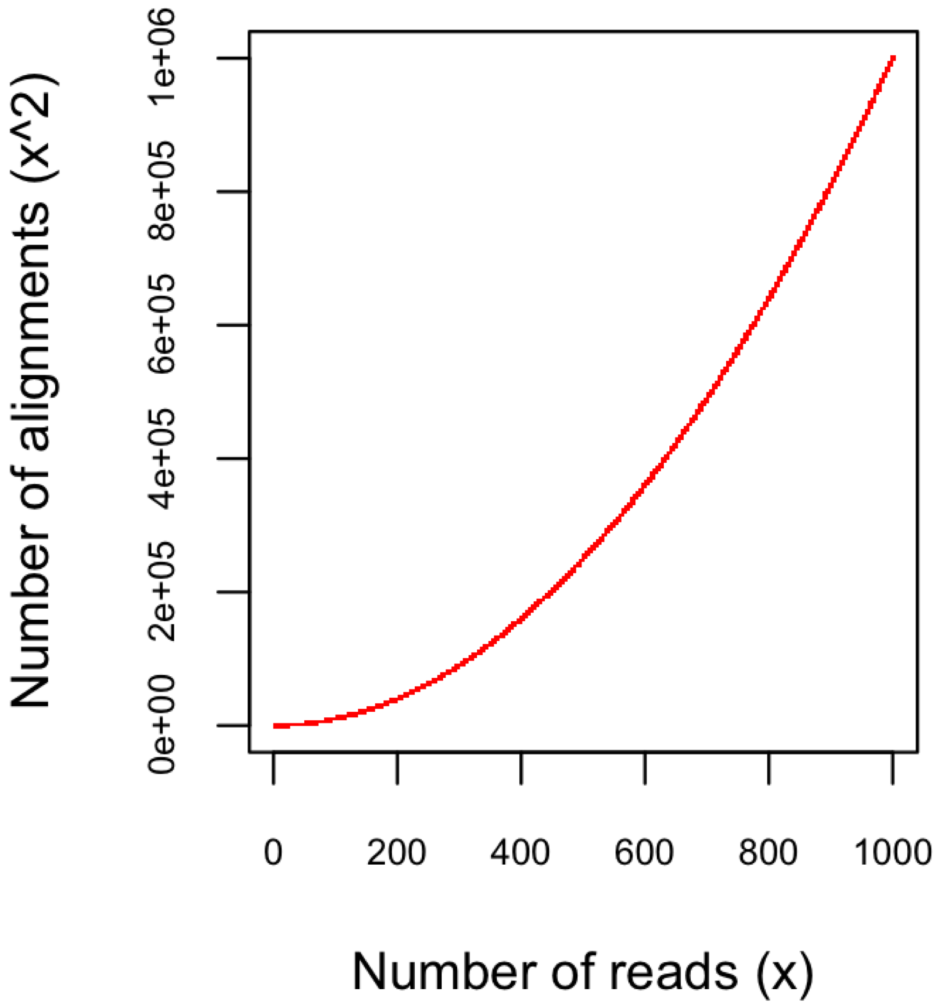
Node = read
Edge = overlap



Transcript B

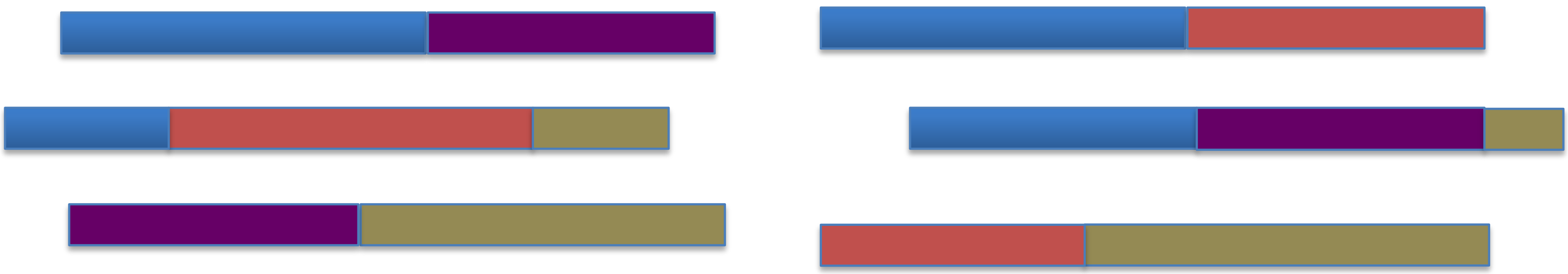
Finding pairwise overlaps between *n* reads involves ~ n^2 comparisons.





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

No genome to align to… De novo assembly required

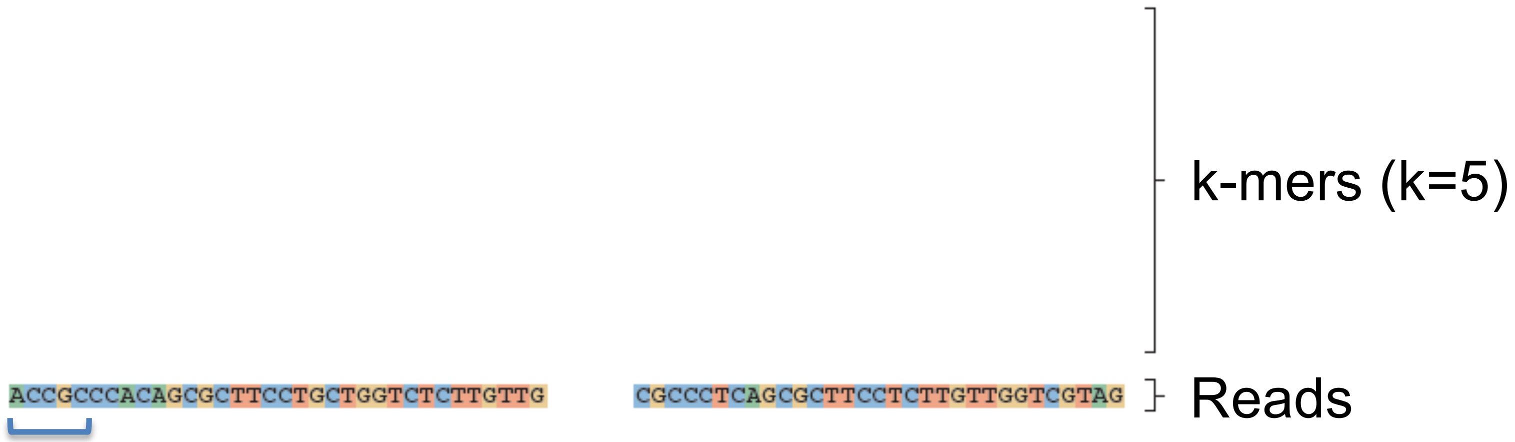


Want to avoid n^2 read alignments to define overlaps

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

k-mers (k=5)

```
ACCGC
ACCGCCCACAGCGCTTCCTGCTGGTCTCTTGTTG      CGCCCTCAGCGCTTCCTCTTGTTGGTCGTAG
```

Reads

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

Sequence Assembly via De Bruijn Graphs

Generate all substrings of length k from the reads

ACCGC

k-mers (k=5)

ACCGCCCACAGCGCTTCCTGCTGGTCTCTTGTTG   CGCCCTCAGCGCTTCCTCTTGTTGGTCGTAG

Reads

Construct the de Bruijn graph

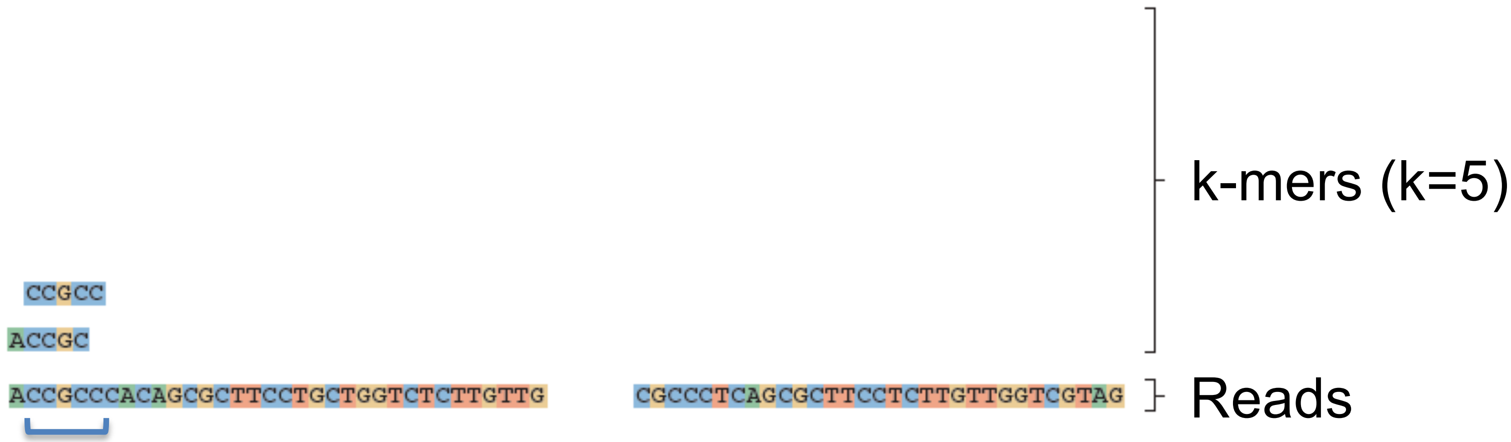
ACCGC

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



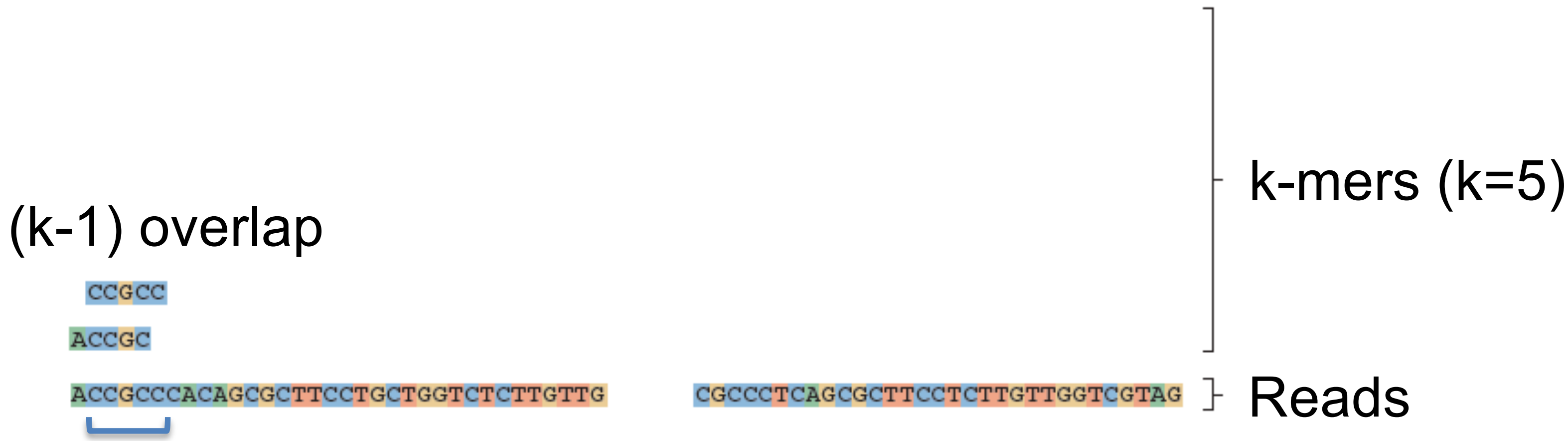
Construct the de Bruijn graph

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



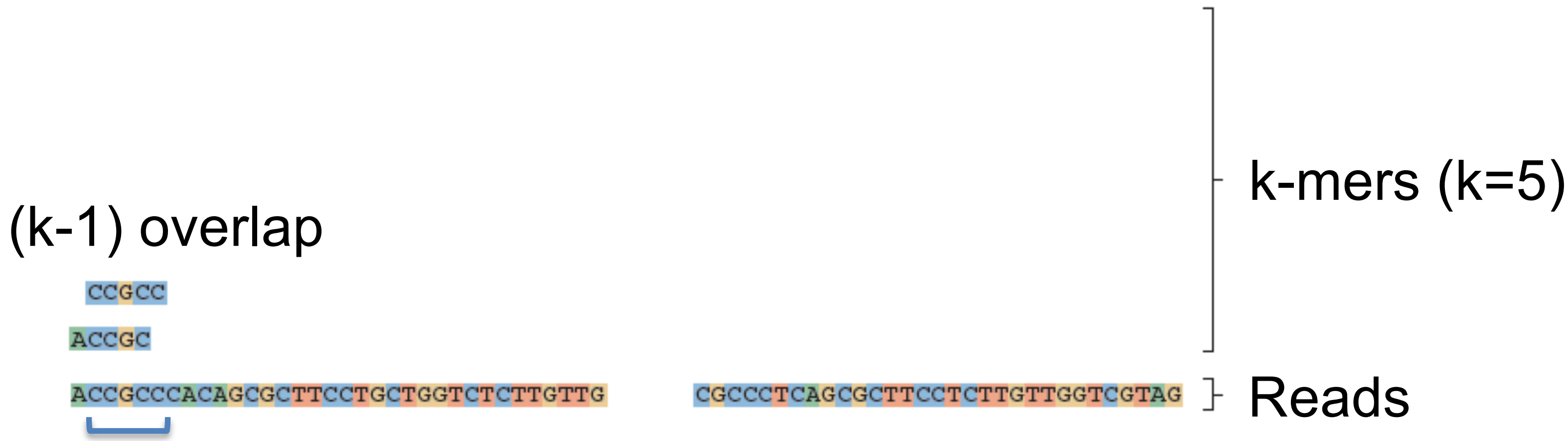
Construct the de Bruijn graph

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

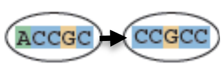
Nodes = unique k-mers, Edges = overlap by (k-1)

Sequence Assembly via De Bruijn Graphs

Generate all substrings of length k from the reads



Construct the de Bruijn graph

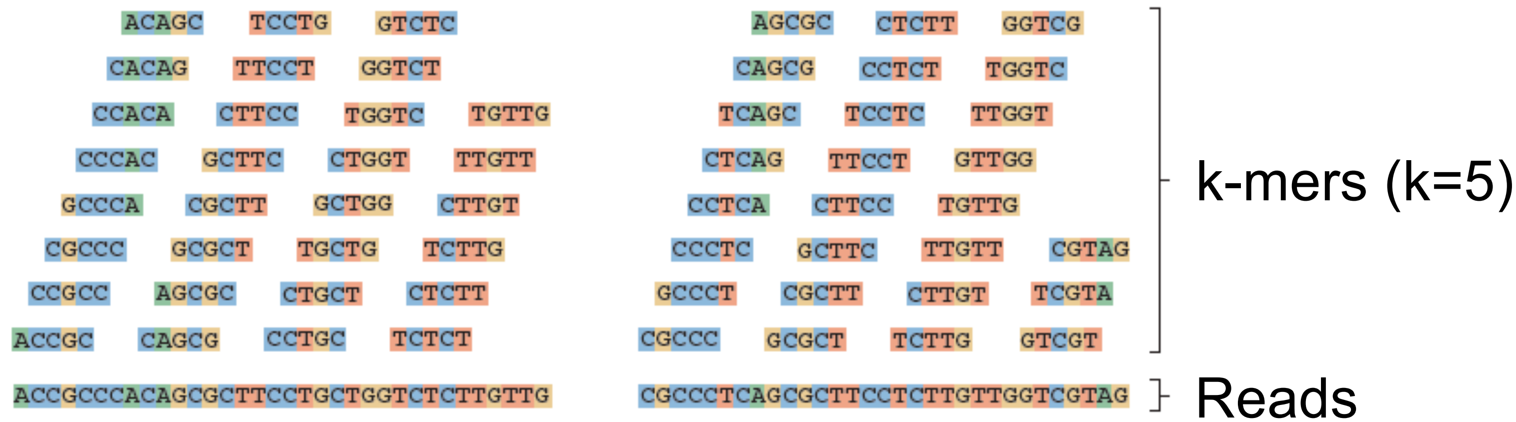


From Martin & Wang, Nat. Rev. Genet. 2011

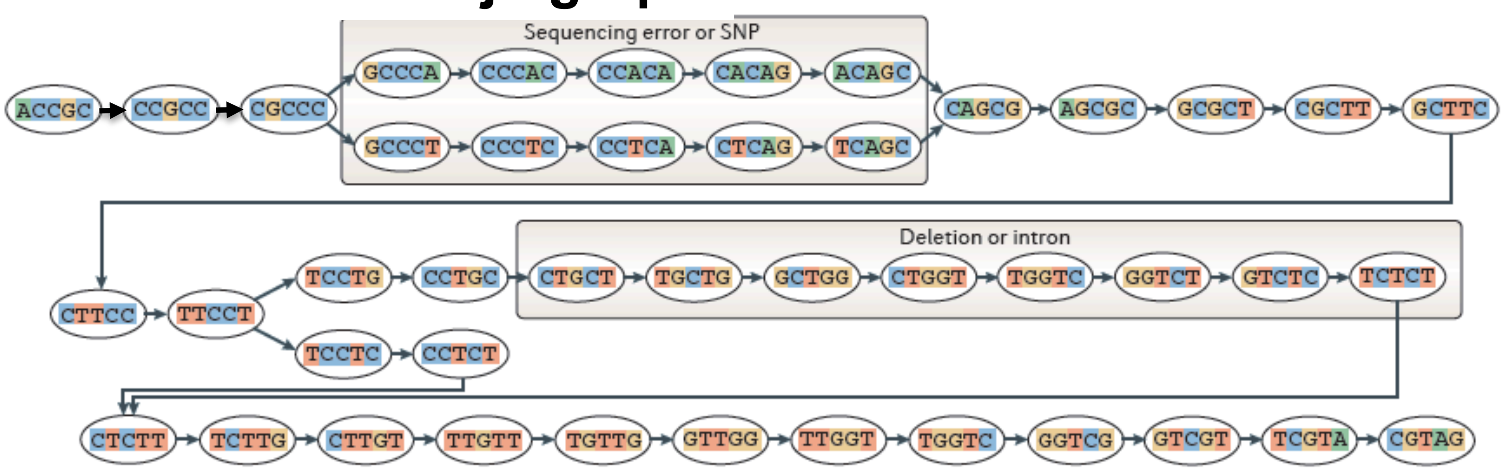
Nodes = unique k-mers, Edges = overlap by (k-1)

Sequence Assembly via De Bruijn Graphs

Generate all substrings of length k from the reads



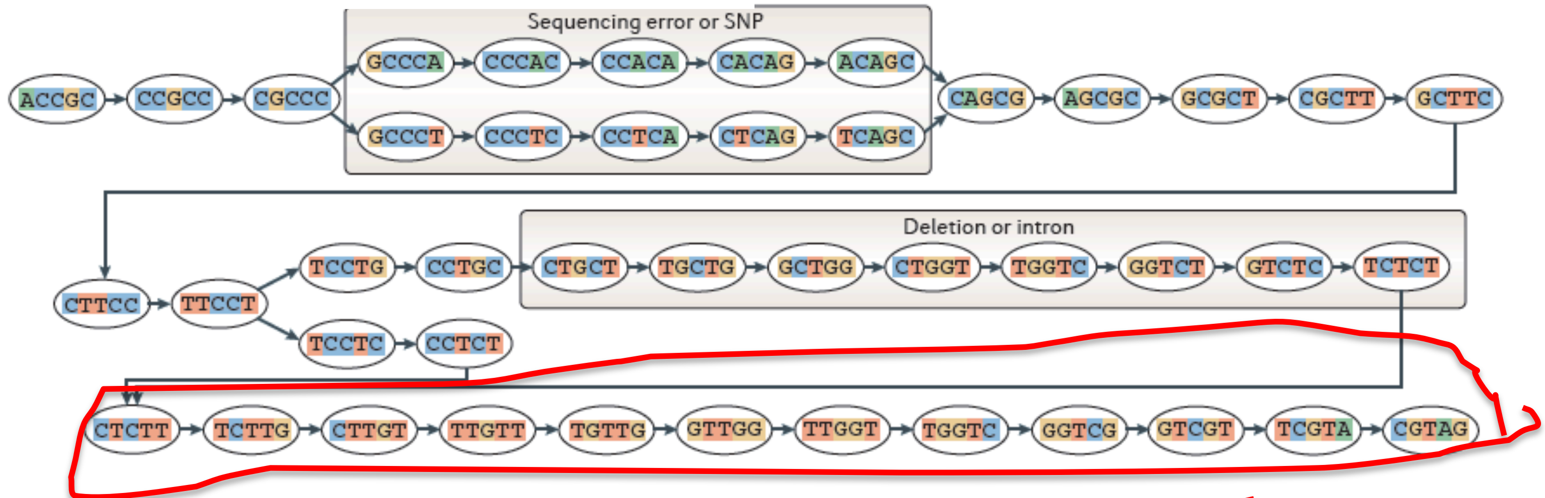
Construct the de Bruijn graph



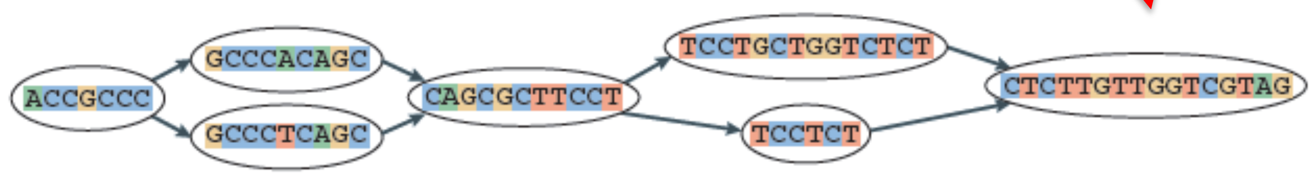
From Martin & Wang, Nat. Rev. Genet. 2011

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

Construct the de Bruijn graph

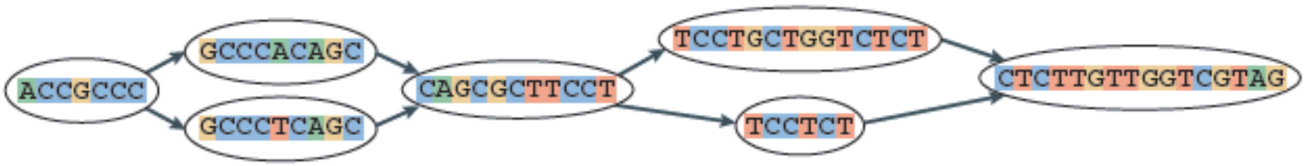


Collapse the de Bruijn graph

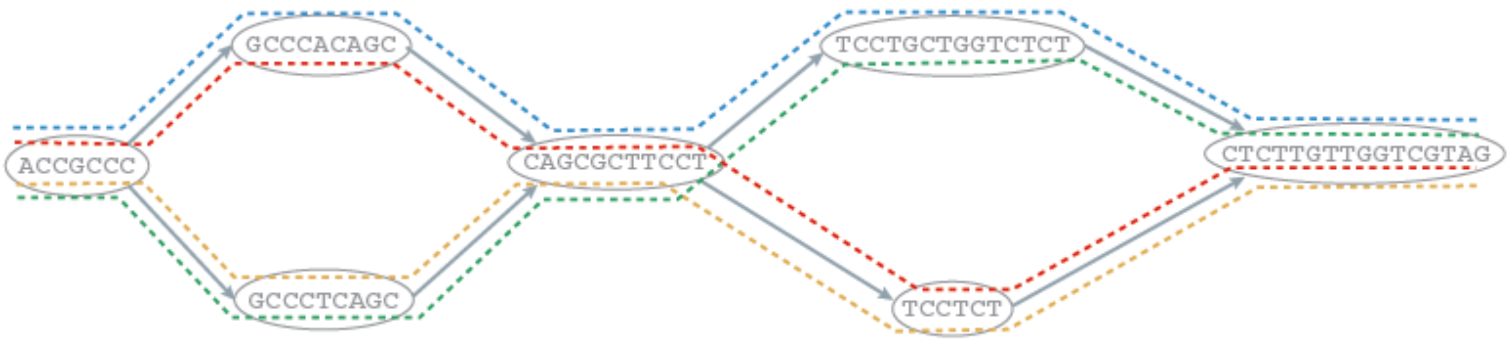


From Martin & Wang, Nat. Rev. Genet. 2011

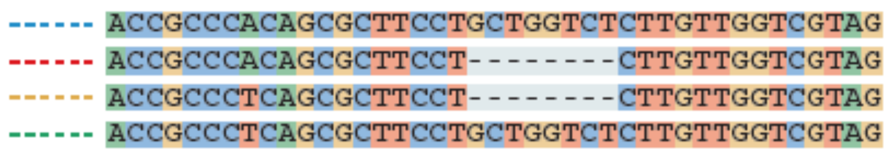
Collapse the de Bruijn graph



Traverse the graph



Assemble Transcript Isoforms



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

Contrasting Genome and Transcriptome *De novo* Assembly

Genome Assembly

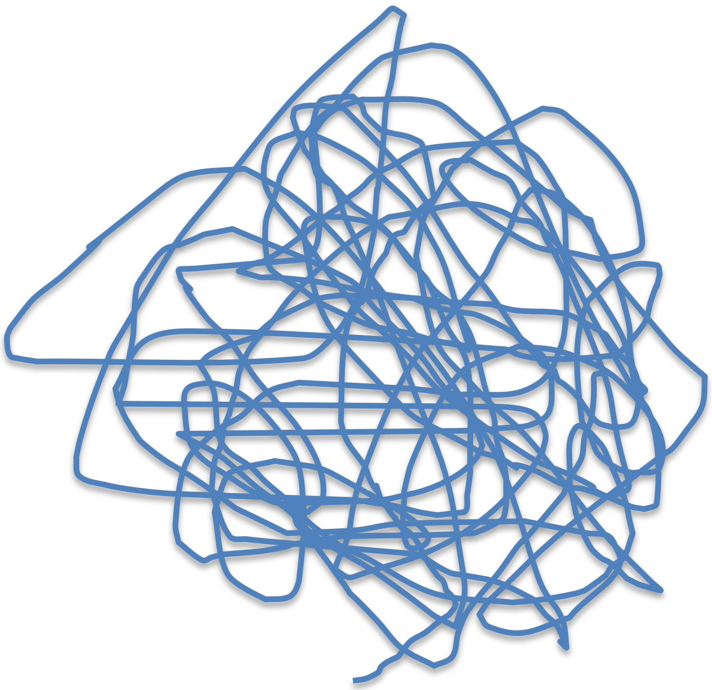
- Uniform coverage
- Single contig per locus
- Assemble small numbers of large Mb-length chromosomes
- Double-stranded data

Transcriptome Assembly

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

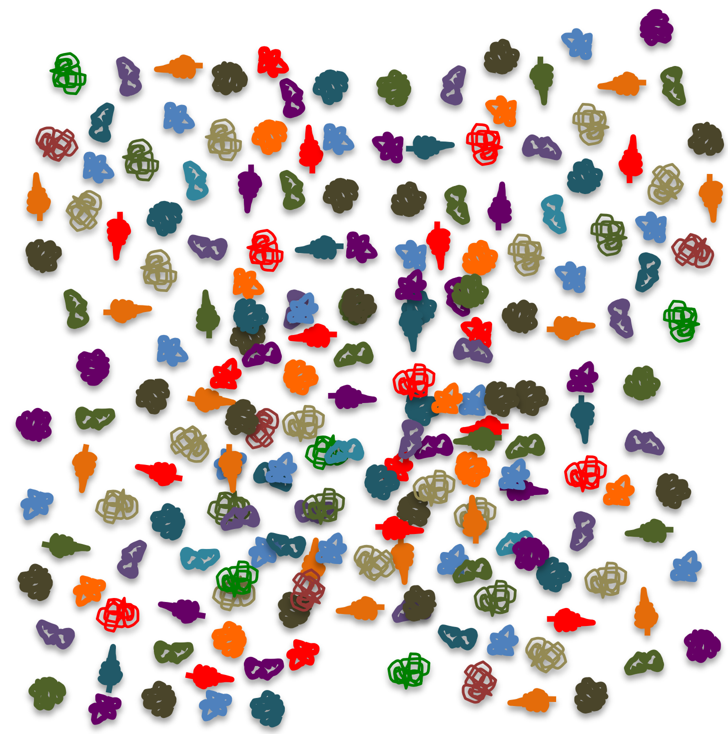
Trinity Aggregates Isolated Transcript Graphs

Genome Assembly
Single Massive Graph



Entire chromosomes represented.

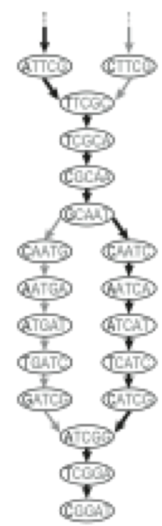
Trinity Transcriptome Assembly
Many Thousands of Small Graphs



Ideally, one graph per expressed gene.

Trinity – How it works:

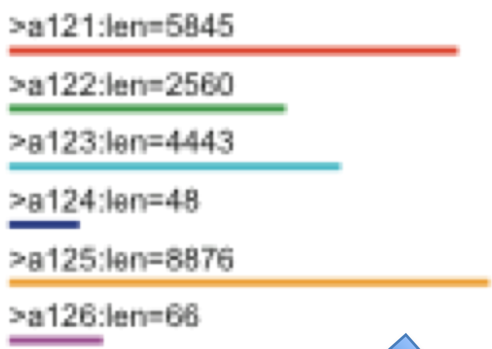
RNA-Seq reads → **Linear contigs** → **de-Bruijn graphs** → **Transcripts + Isoforms**



Thousands of disjoint graphs

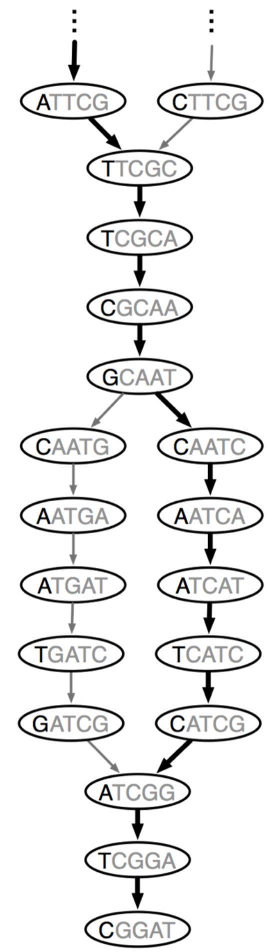
...CTTCGCAA...TGATCGGAT...
...ATTCGCAA...TCATCGGAT...

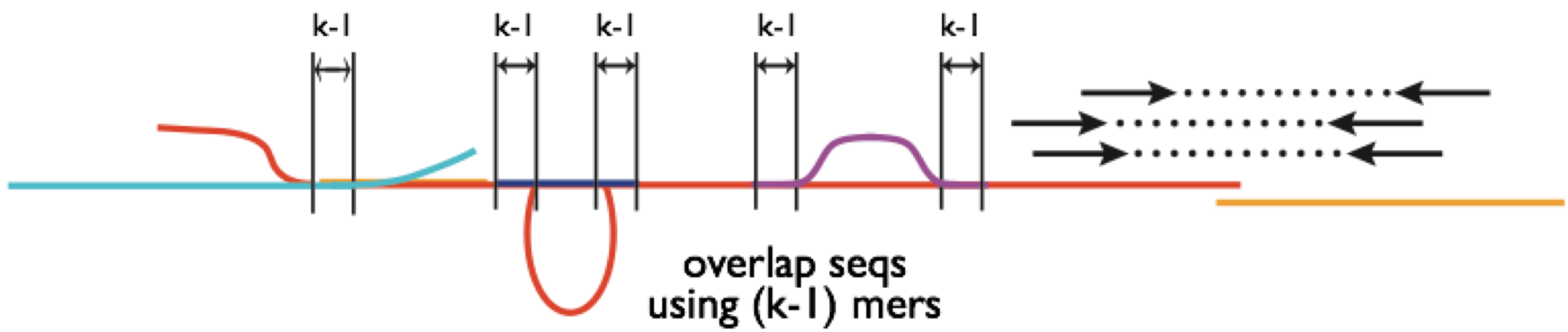
Chrysalis



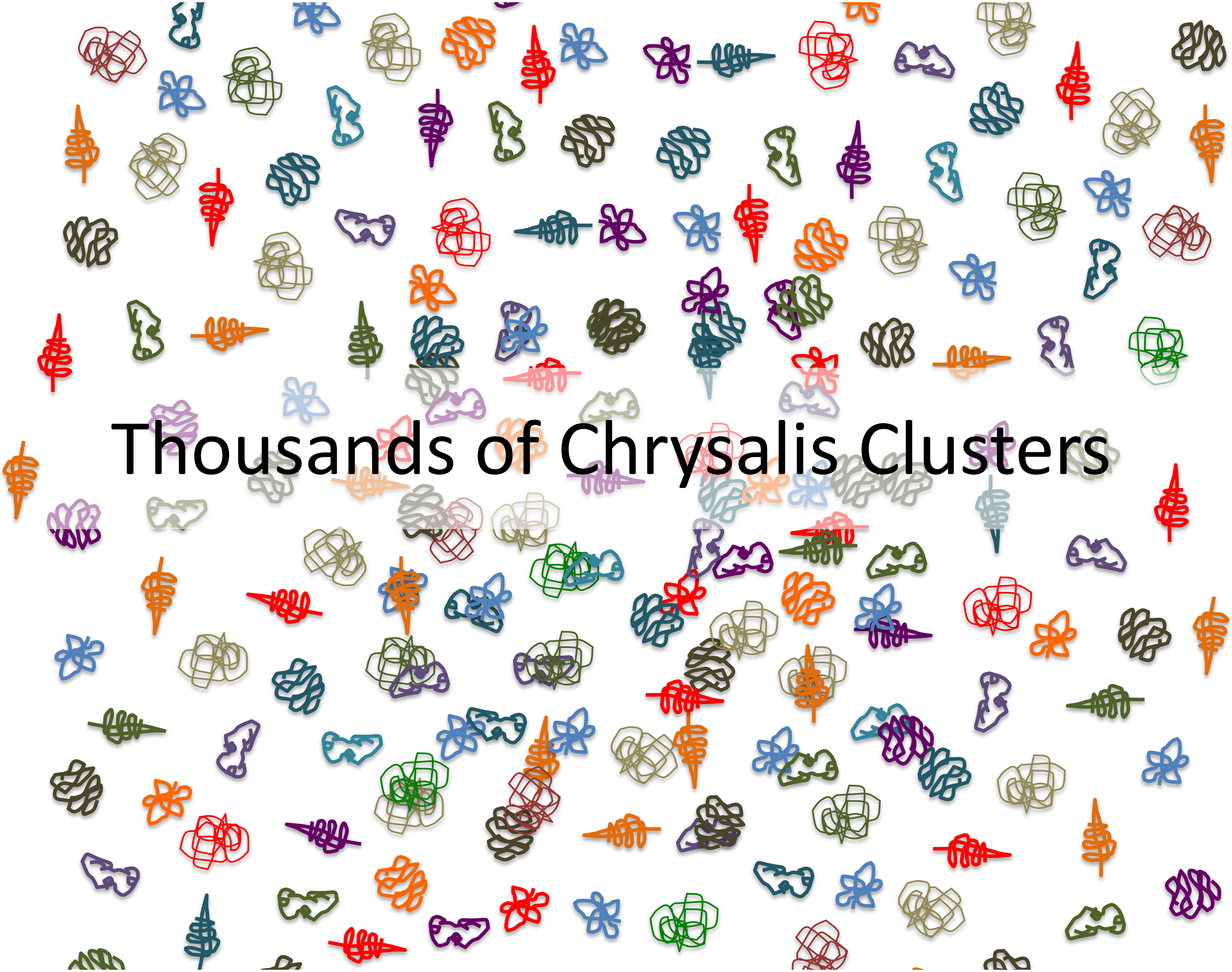
Integrate isoforms
via k-1 overlaps

Build de Bruijn Graphs
(ideally, one per gene)

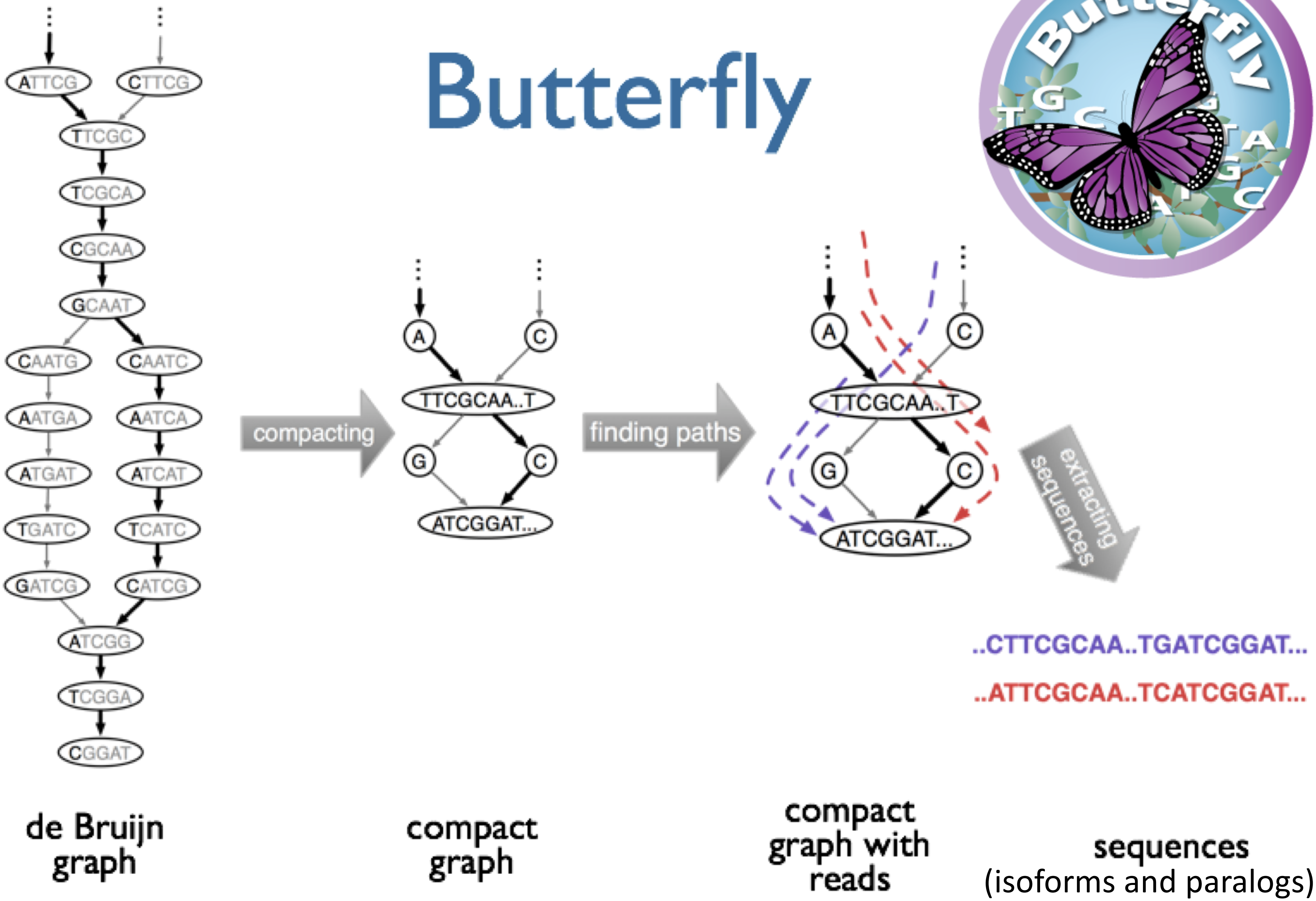




Thousands of Chrysalis Clusters

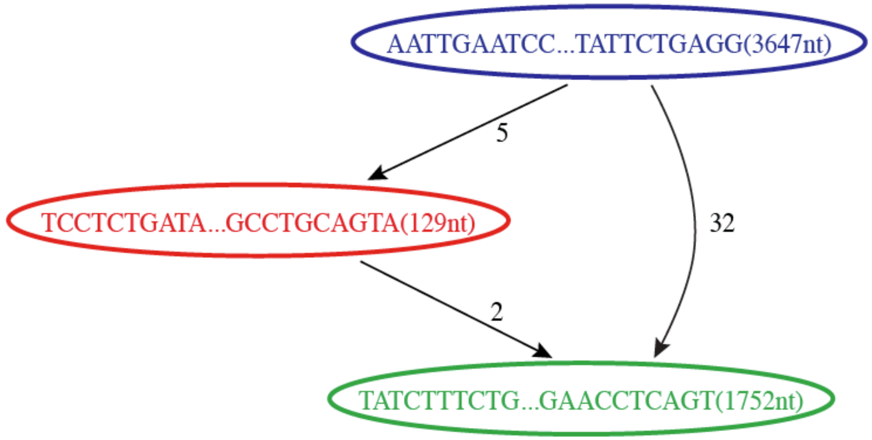


Butterfly


ATTCG
CTTCG
TTCGC
TCGCA
CGCAA
GCAAT
CAATG
CAATC
AATGA
AATCA
ATGAT
ATCAT
TGATC
TCATC
GATCG
CATCG
ATCGG
TCGGA
CGGAT
compacting
A
C
TTCGCAA..T
G
C
ATCGGAT...
finding paths
A
C
TTCGCAA..T
G
C
ATCGGAT...
extracting sequences
..CTTCGCAA..TGATCGGAT...
..ATTCGCAA..TCATCGGAT...
de Bruijn graph
compact graph
compact graph with reads
sequences
(isoforms and paralogs)

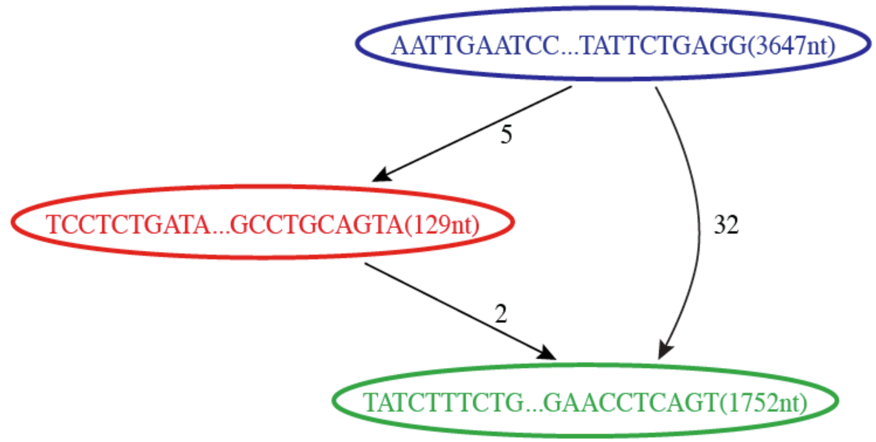
Butterfly Example 1:
Reconstruction of Alternatively Spliced Transcripts

Butterfly's Compacted Sequence Graph



Reconstruction of Alternatively Spliced Transcripts

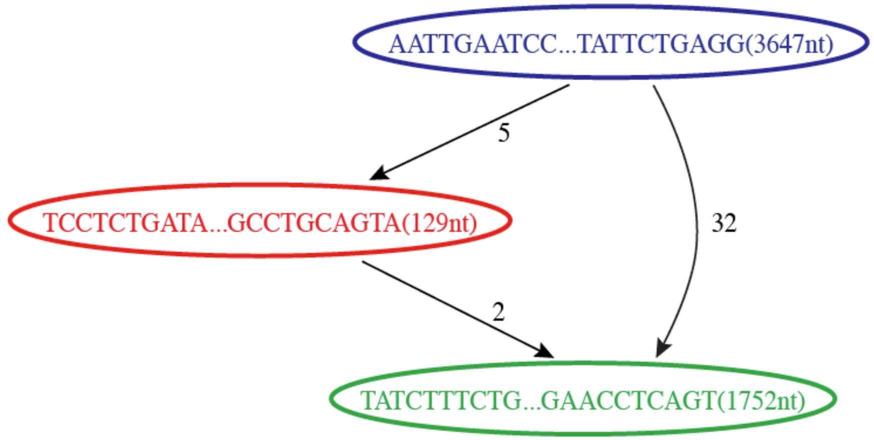
Butterfly's Compacted Sequence Graph



Reconstructed Transcripts

Reconstruction of Alternatively Spliced Transcripts

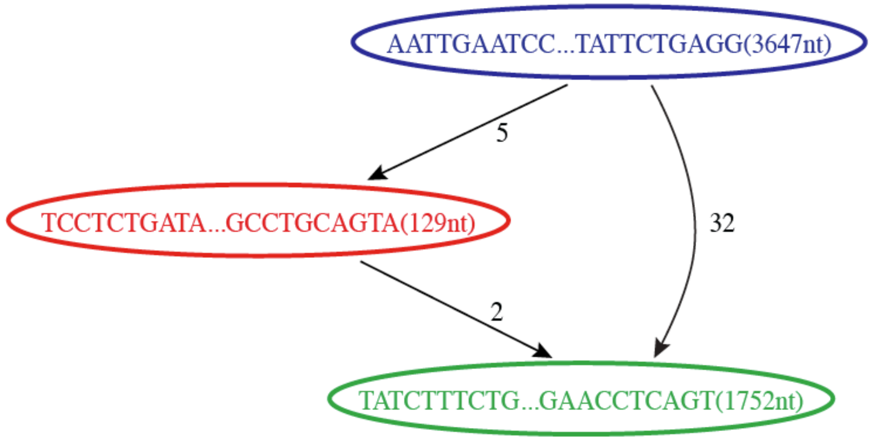
Butterfly's Compacted Sequence Graph



Reconstructed Transcripts

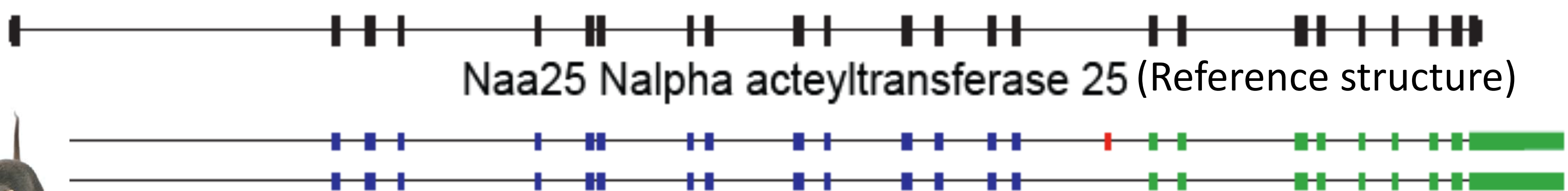
Reconstruction of Alternatively Spliced Transcripts

Butterfly's Compacted Sequence Graph



Reconstructed Transcripts

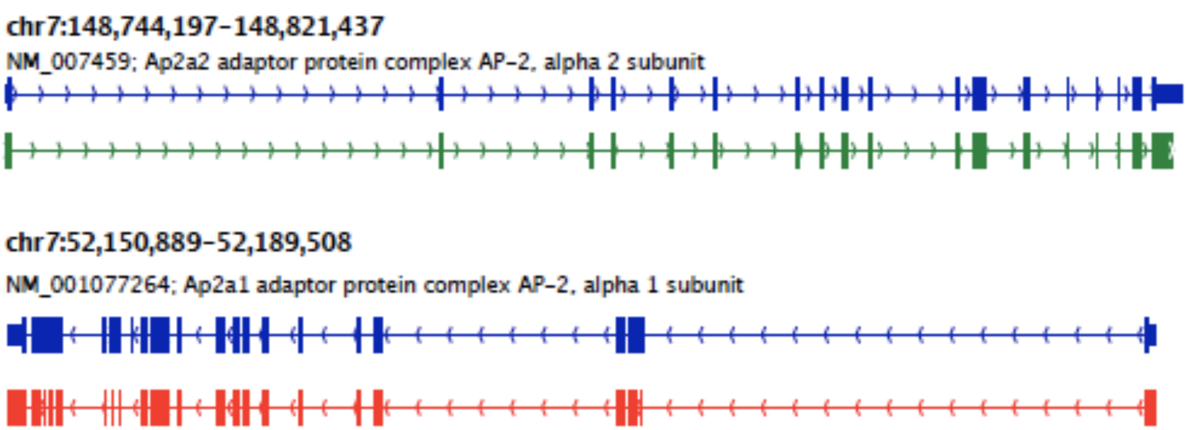
Aligned to Mouse Genome



(Reference structure)

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

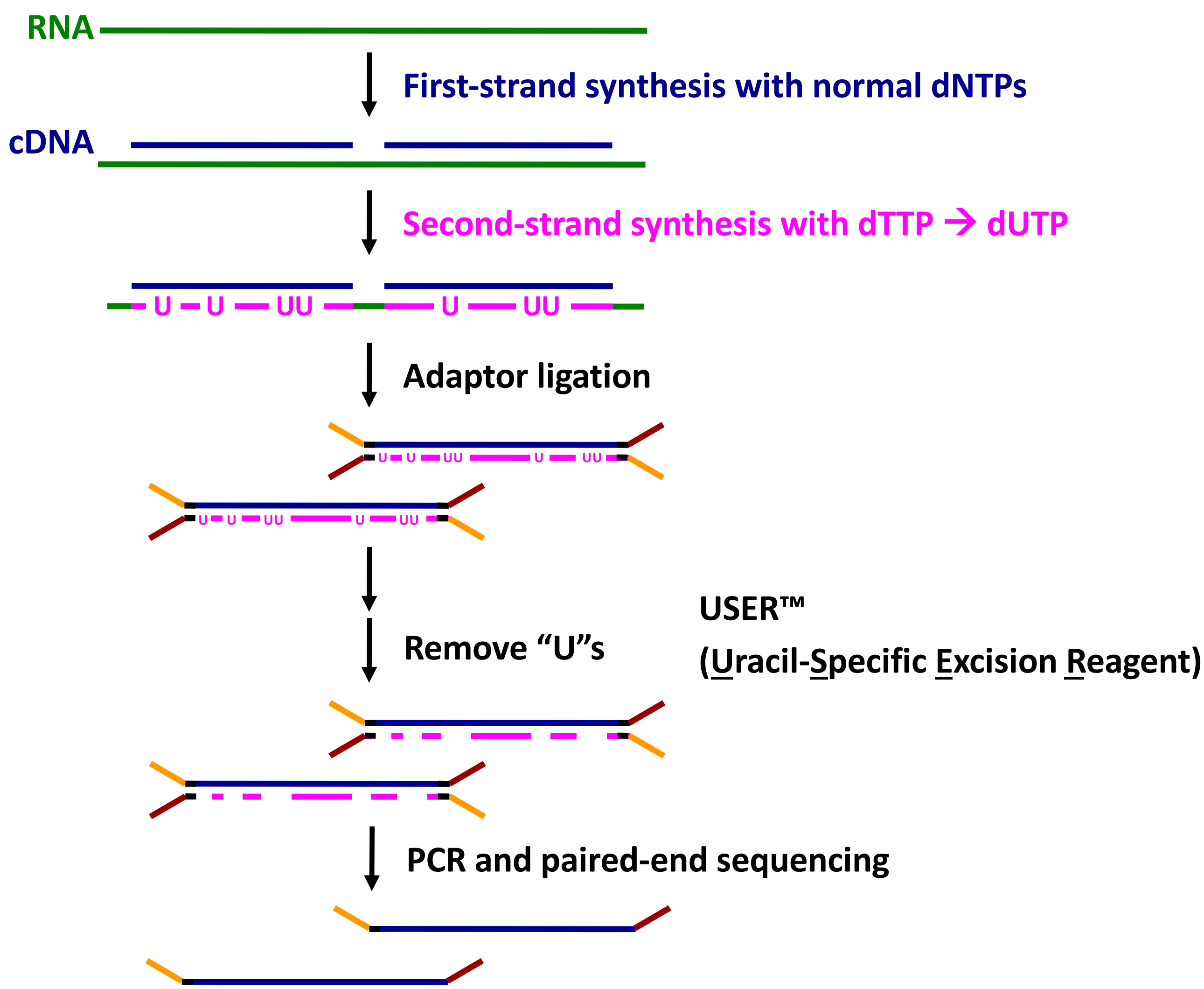
computational pipeline to compare library quality metrics from any RNA-seq method. Using the well-annotated *Saccharomyces cerevisiae* transcriptome as a benchmark, we compared seven library-construction protocols, including both published and

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-

boundaries of adjacent genes transcribed on opposite strands and resolve the correct expression levels of coding or noncoding overlapping transcripts. These tasks are particularly challenging in small microbial genomes, prokaryotic and eukaryotic, in which

'dUTP second strand marking' identified as the leading protocol

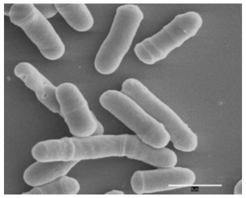
dUTP 2nd Strand Method: Our Favorite



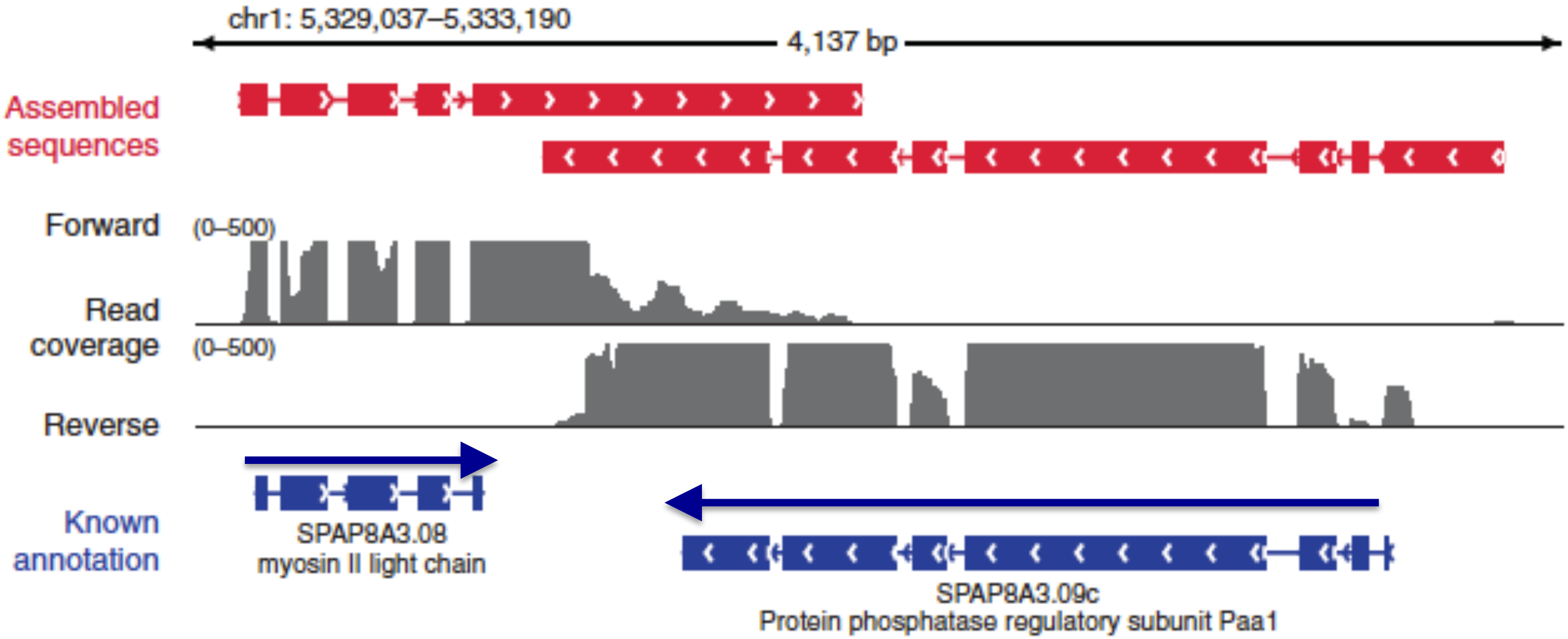
Modified from Parkhomchuk *et al.* (2009) *Nucleic Acids Res.* 37:e123

Slide from J. Levin

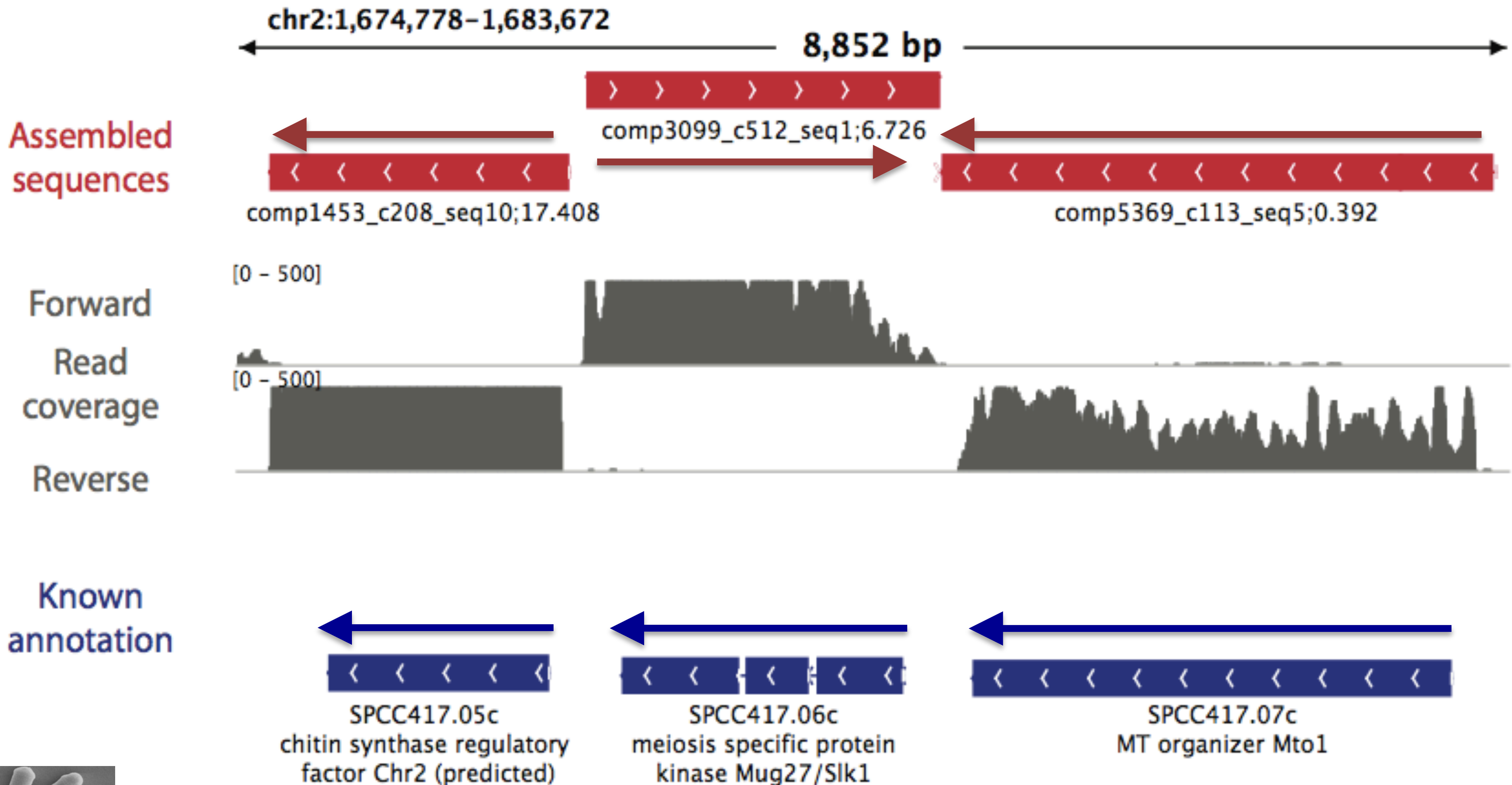
Overlapping UTRs from Opposite Strands

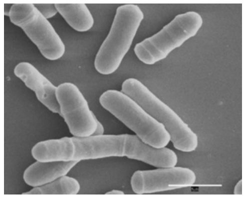


Schizosaccharomyces pombe
(fission yeast)

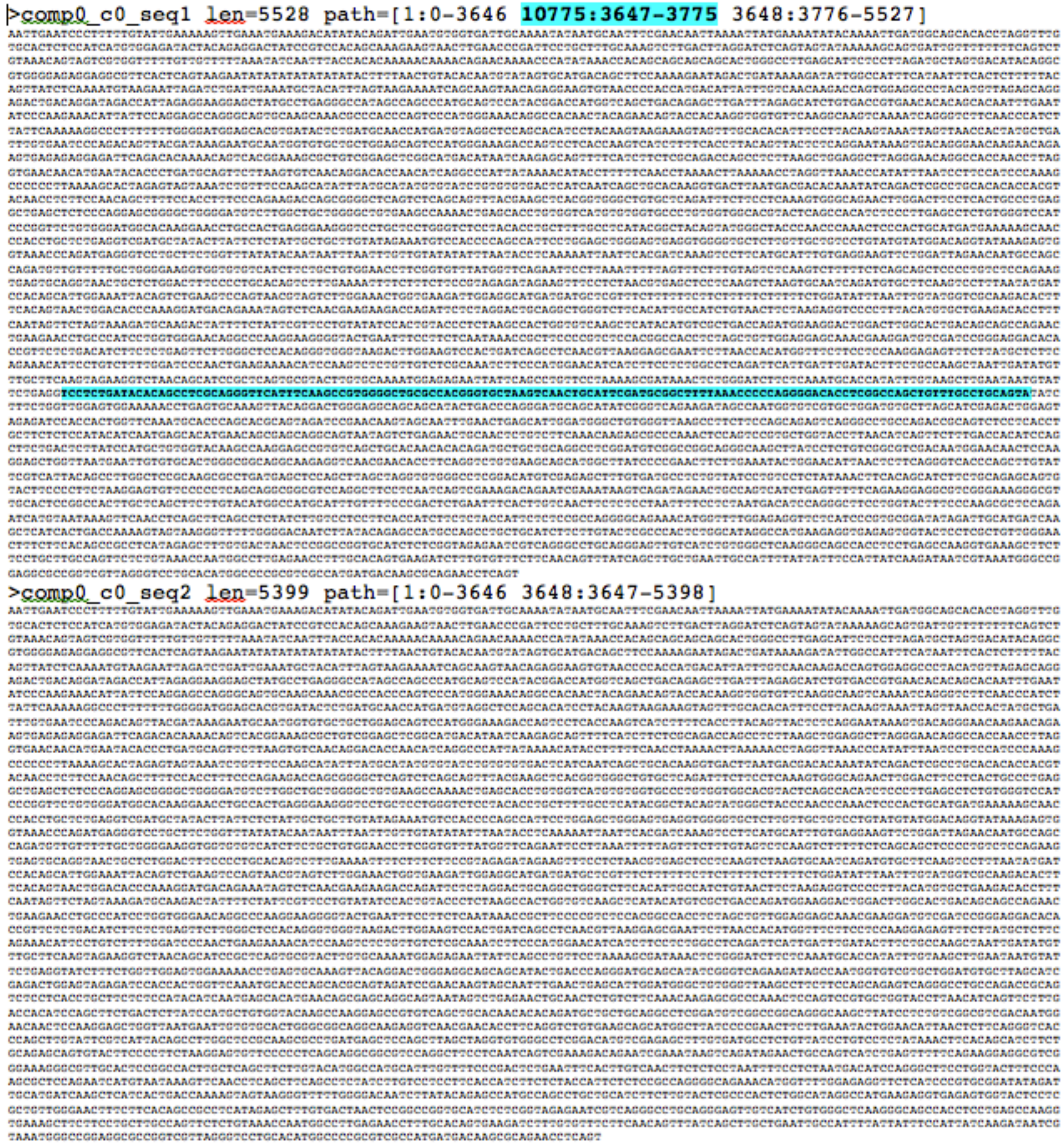


Antisense-dominated Transcription





Trinity output: a multi-fasta file



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

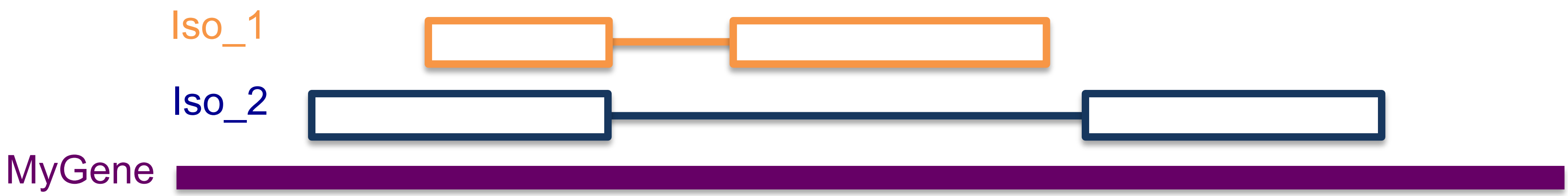
```
>comp0_c0_seq2 len=5399 path=[1:0-3646 3648:3647-5398]
```

Flavors of Differential Expression Analyses

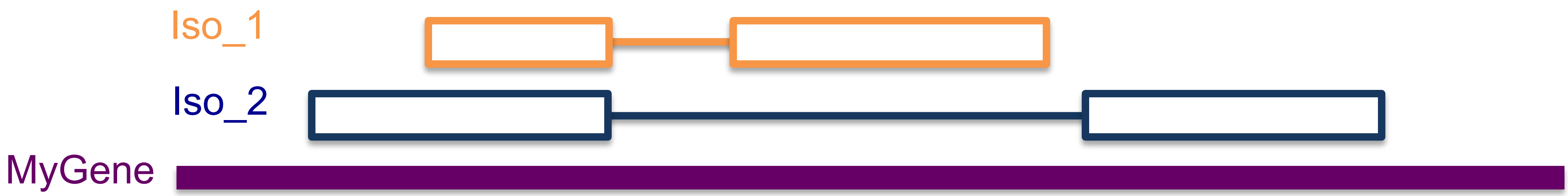
- Transcripts:
  - Differential Transcript Expression (DTE)
  - Differential Transcript Usage (DTU)
  - Differential Exon Usage (DEU)
- Gene:
  - Differential Gene Expression (DGE)
  - Gene Differential Expression (GDE)

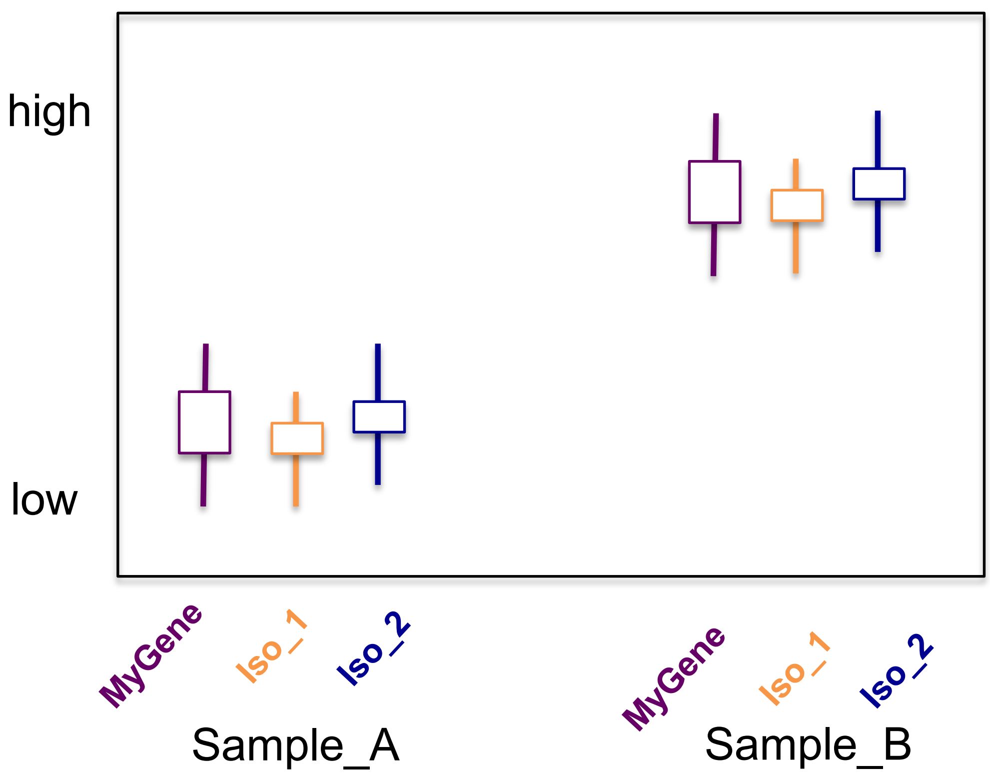
?

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



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

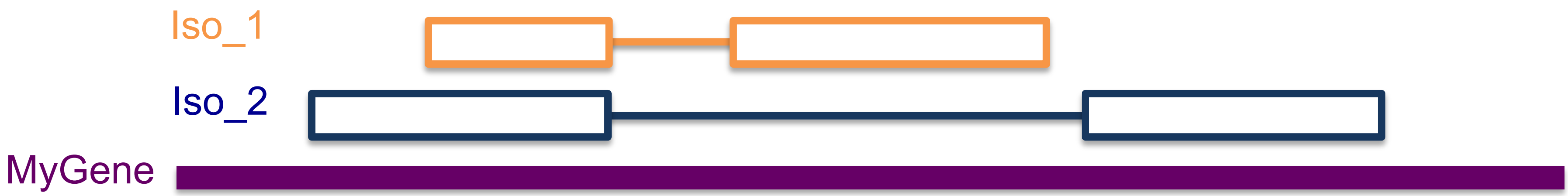


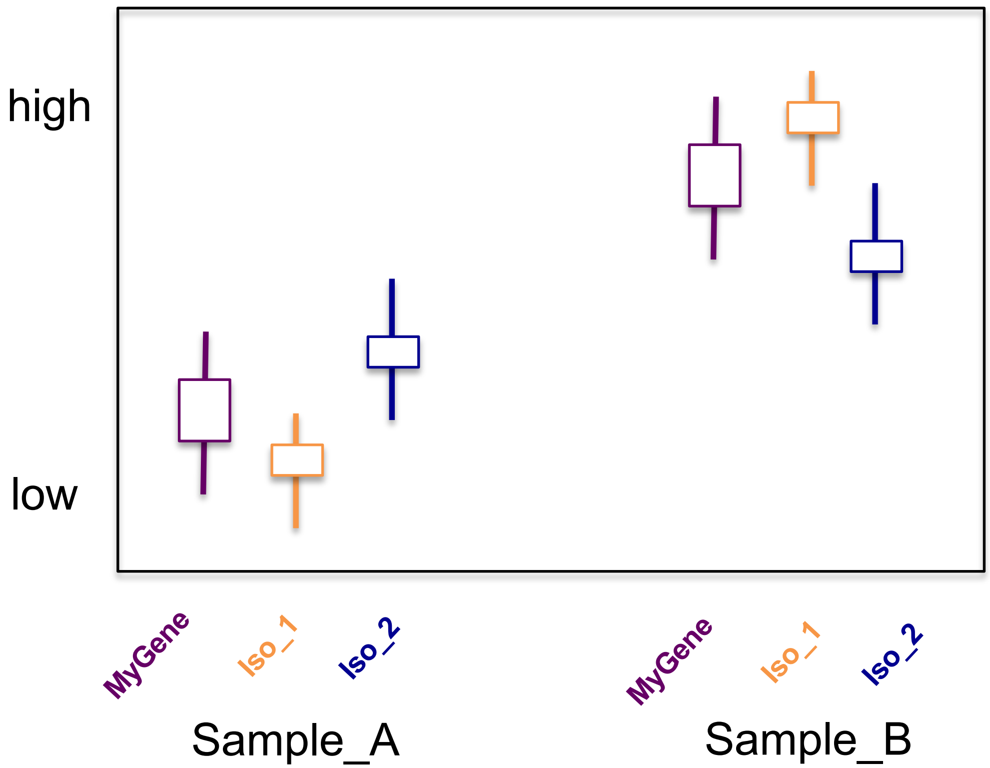


Feature	Diff Expressed?
MyGene	Yes
Iso_1	Yes
Iso_2	Yes

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

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

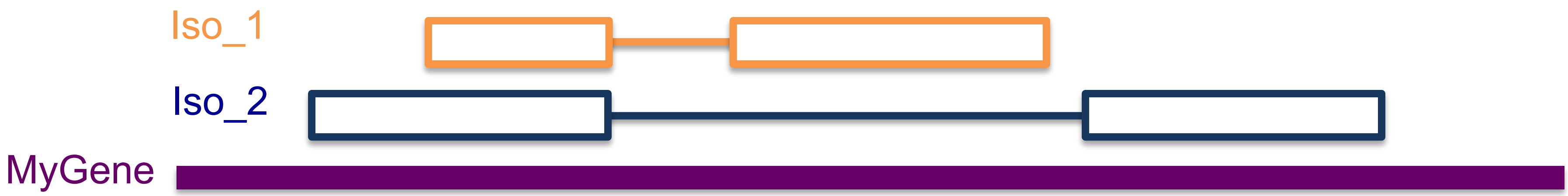


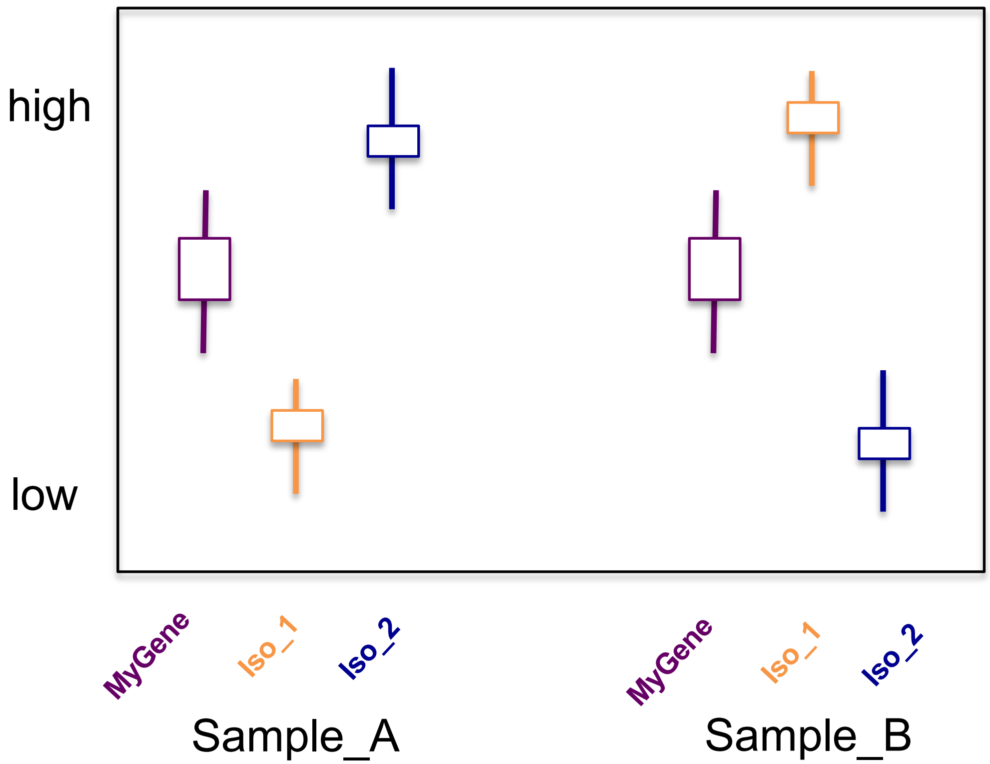


Feature	Diff Expressed?
MyGene	Yes
Iso_1	Yes
Iso_2	Yes

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

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





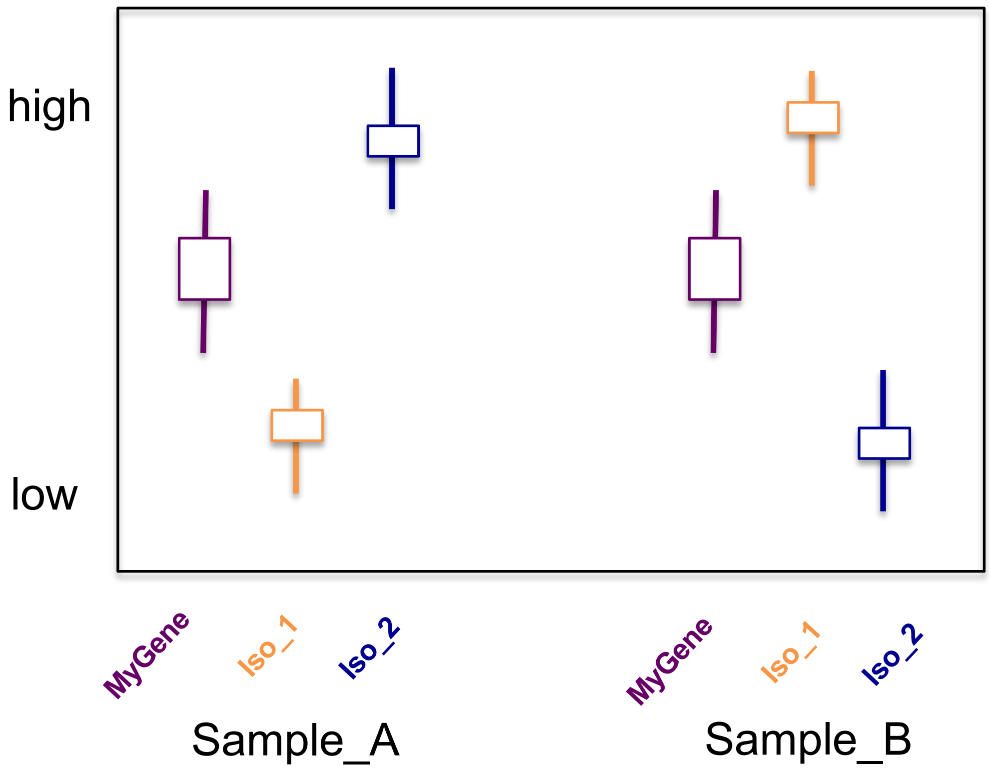
Feature	Diff Expressed?
MyGene	No
Iso_1	Yes
Iso_2	Yes

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

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

From Gene-level view (DGE): not apparent
From Transcript-level view (GDE): Yes, gene should be acknowledged as having changed.

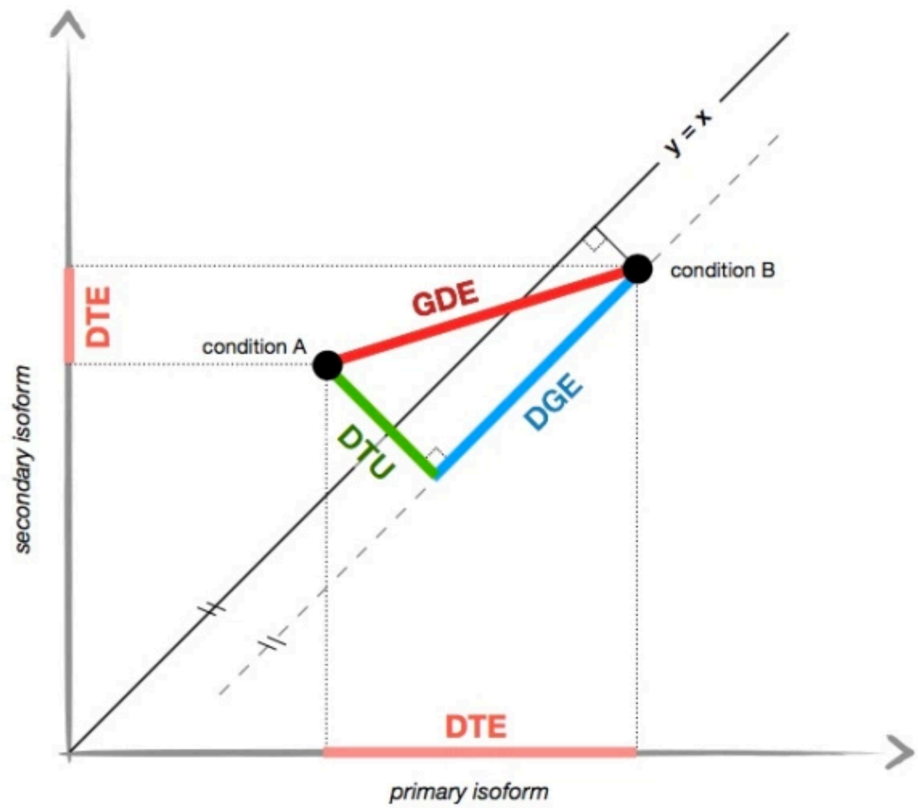
Prevailing viewpoint:
DTE or DTU -> Gene Diff Expressed (GDE)



Feature	Diff Expressed?
MyGene	No?
Iso_1	Yes
Iso_2	Yes

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

Clarifying view: (DTE or DTU or DGE) as special cases of Ge

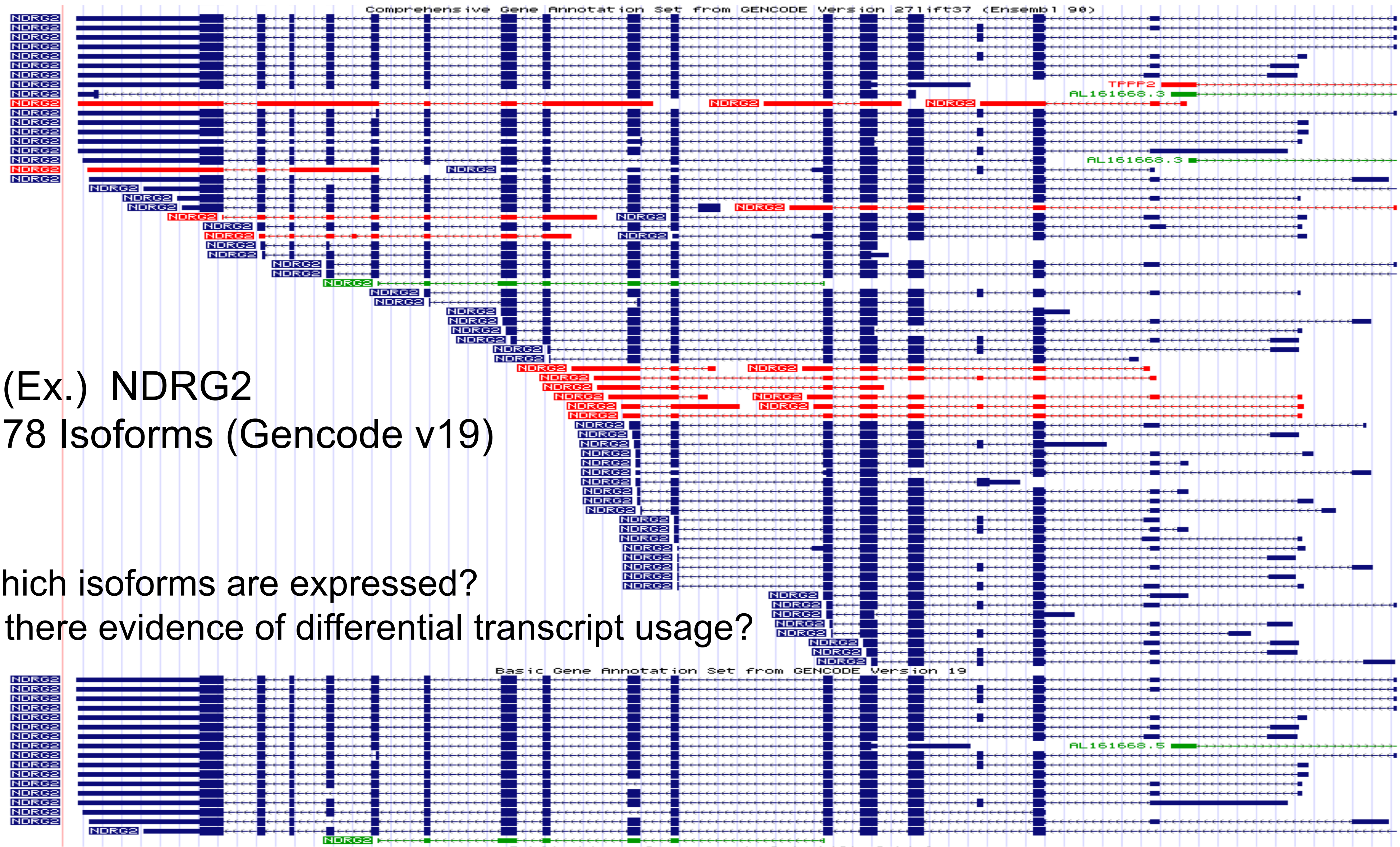


DTE: differential transcript expression
DTU: differential transcript usage
DGE: differential gene expression (gene-level analysis)
GDE: gene differential expression (transcript-level analysis)

Ntranos, Yi, et al., 2018 – see supp.

[See Lior Pachter's blog post: https://liorpachter.wordpress.com/2019/01/07/fast-and-accurate-gene-differential-expression-by-testing-transcript-compatibility-counts/](https://liorpachter.wordpress.com/2019/01/07/fast-and-accurate-gene-differential-expression-by-testing-transcript-compatibility-counts/)

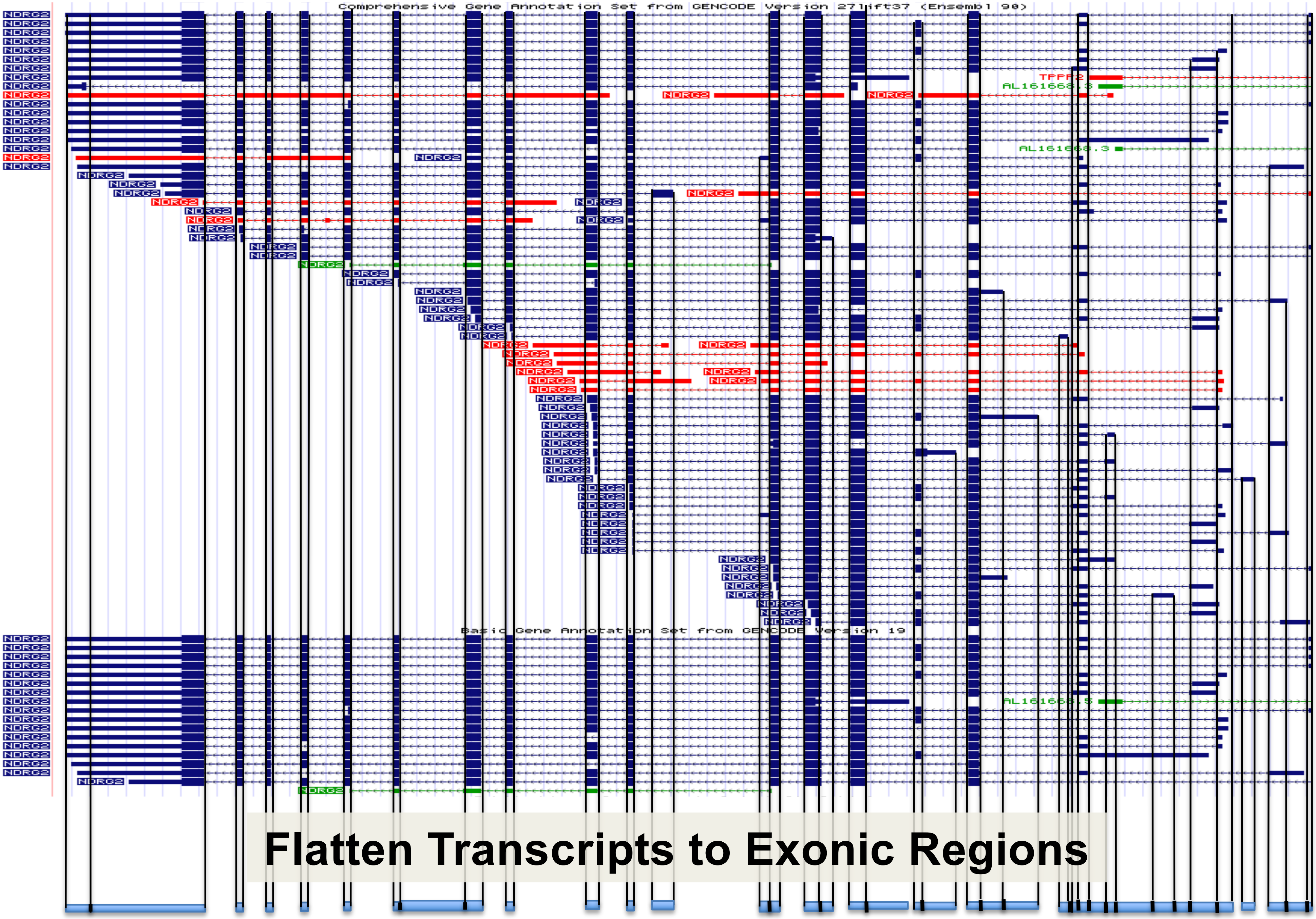
High Confidence Differential Transcript Expression is Difficult to Attain With Many Candidate Isoforms



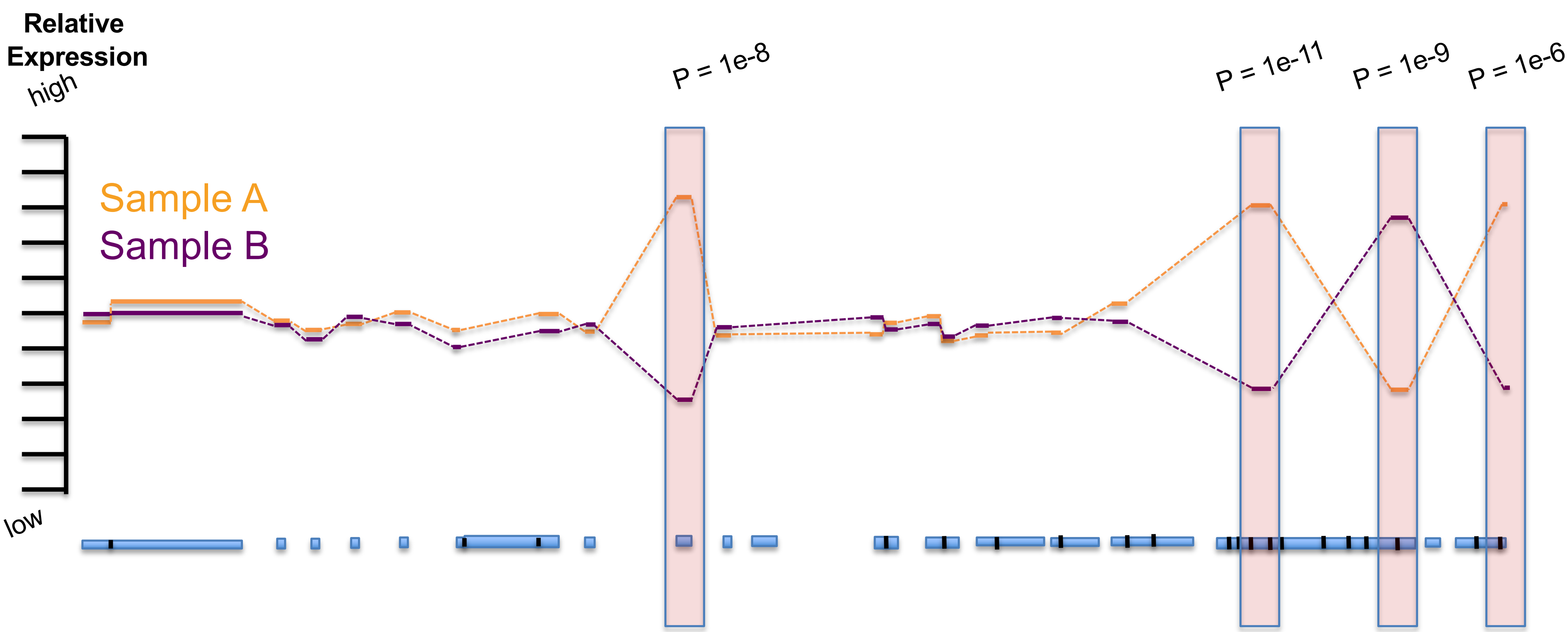
(Ex.) NDRG2
78 Isoforms (Gencode v19)

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

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



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:

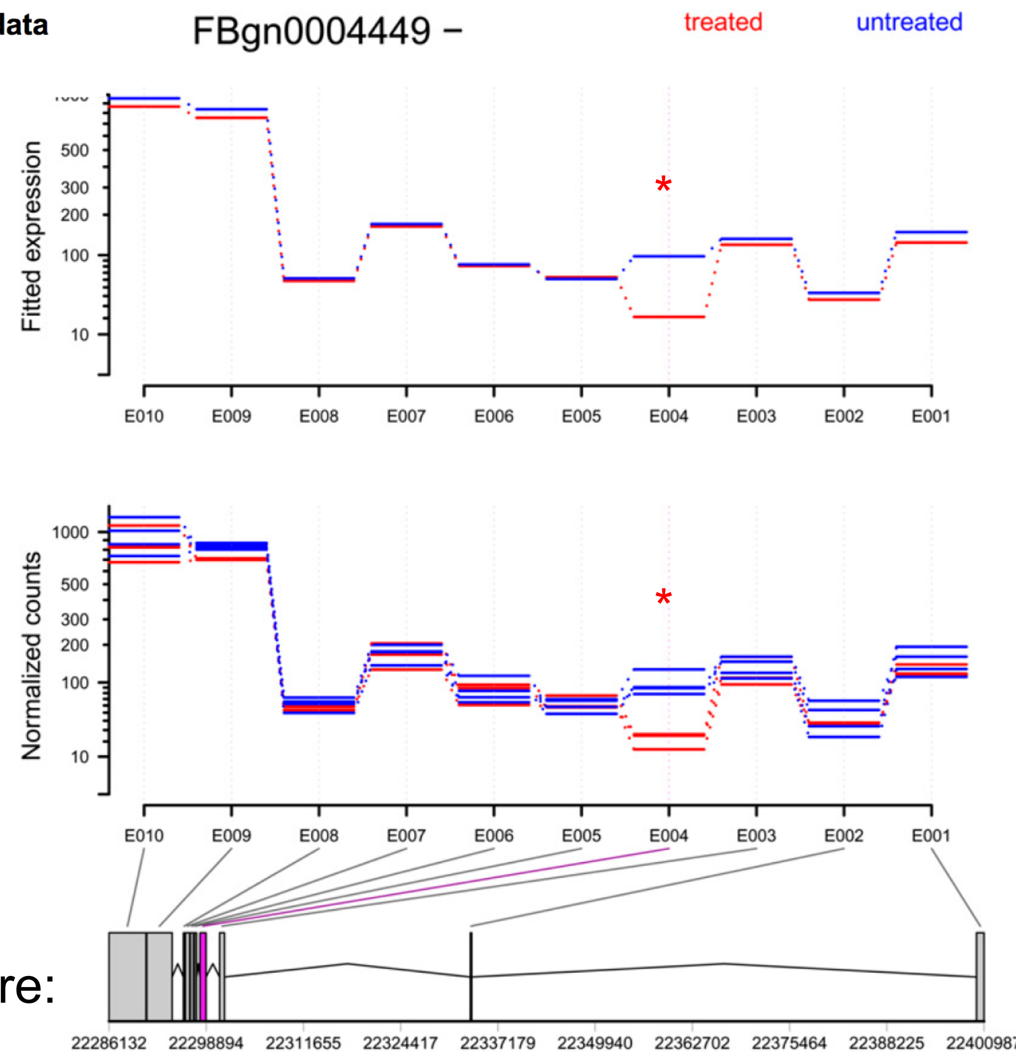


Figure 3. The treatment of knocking down the splicing factor *pasilla* affects the fourth exon (counting bin E004) of the gene *Ten-m* (CG5723). (*Top* panel) Fitted values according to the linear model; (*middle* panel) normalized counts for each sample; (*bottom* panel) flattened gene model. (Red)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

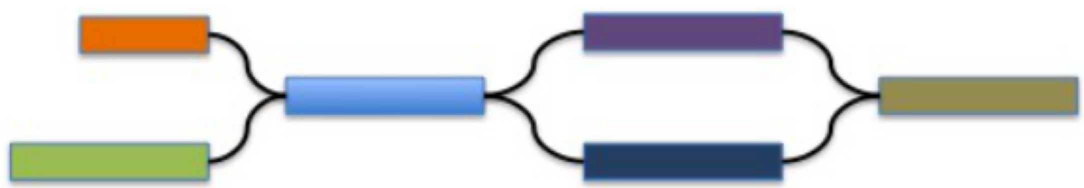
Genome Biology

METHOD Open Access

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

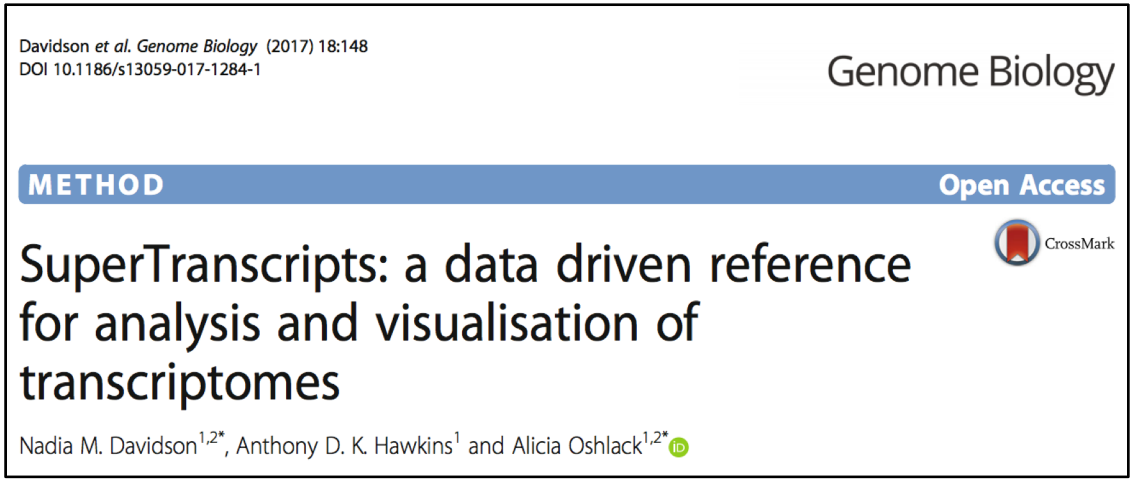
Nadia M. Davidson[1,2*], Anthony D. K. Hawkins[1] and Alicia Oshlack[1,2*]

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

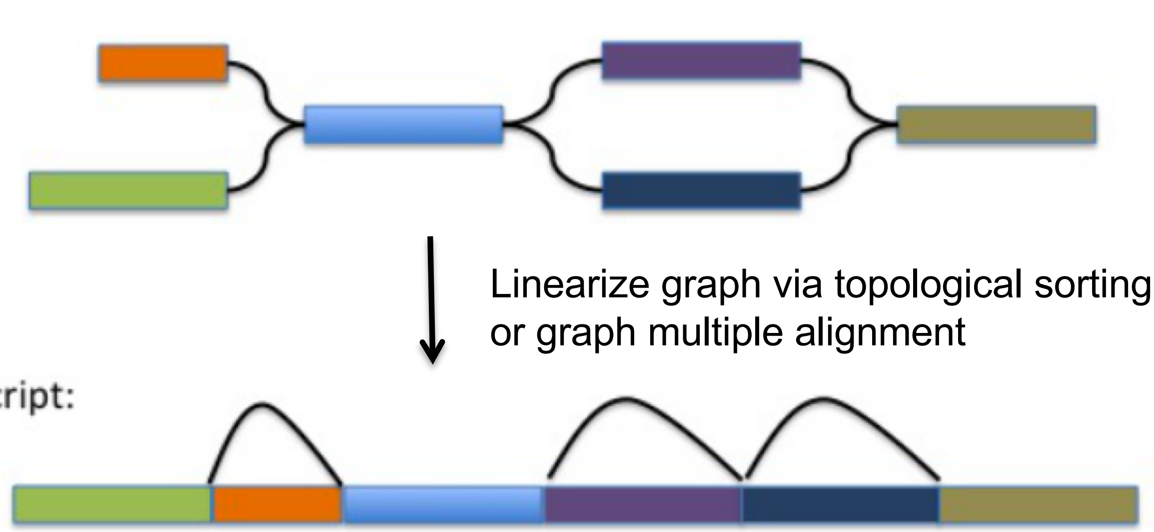
METHOD Open Access

CrossMark

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

We are on a Coffee Break & Networking Session

Workshop Sponsors:

Ontario Institute for Cancer Research

Canadian Centre for Computational Genomics

McGill initiative in Computational Medicine