

The Krumlov Intro to Trinity RNA-seq



Brian Haas
Broad Institute

Workshop on Genomics, Cesky Krumlov, May 2023

Intro to Brian Haas



Education and Career History



BS,MS Molecular Bio
DNA Repair
SUNY Albany
1991-1999



The Institute for Genomic Research
Rockville, Maryland, USA
(1999-2007)

Bioinformatics Analyst & Engineer

MS. Computer Science / Johns Hopkins



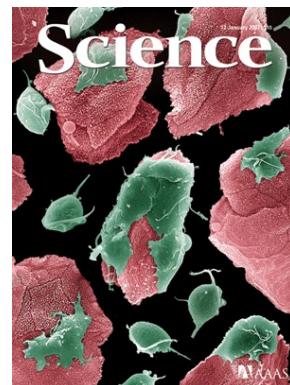
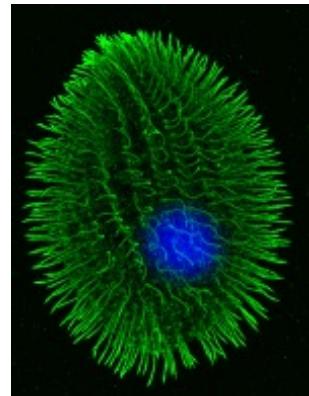
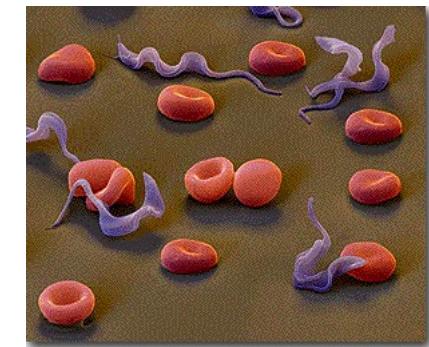
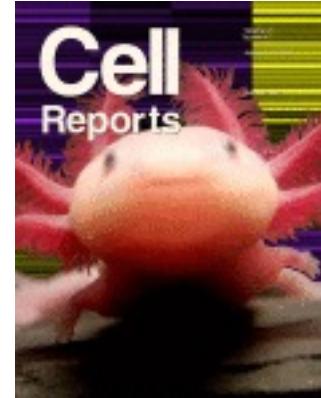
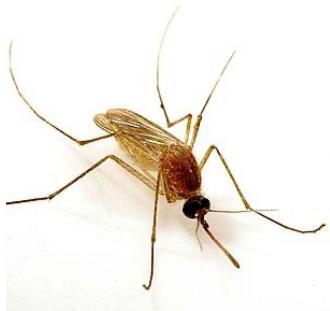
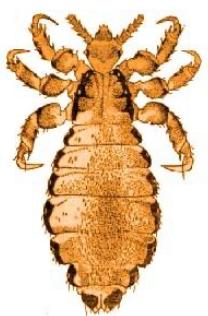
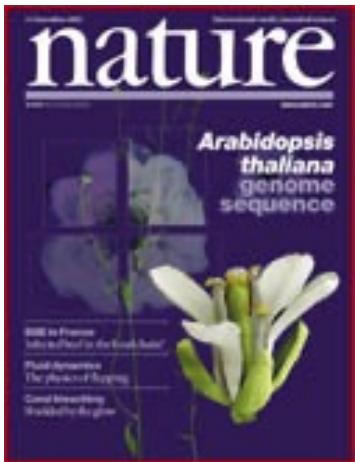
Cambridge, Massachusetts, USA

2007-current

Computational Biologist / Manager / PI

Ph.D. Bioinformatics / Boston University

Annotation and Analysis for Diverse Genomes and Transcriptomes



My Favorite Activity – Bioinformatics Tool Development and Application



NAR, 2003



Bioinformatics, 2004



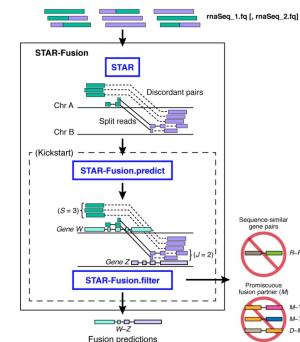
EvidenceModeler
Genome Biology, 2008



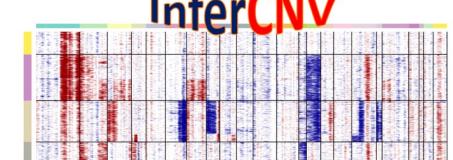
Chimera Slayer
Genome Research, 2011



Nature Biotech, 2011
Nature Protocols, 2013



STAR-Fusion
Genome Biology, 2019



FusionInspector
Cell Reports Methods, 2023

My Favorite Activity – Bioinformatics Tool

Development and Application



NAR, 2003



Bioinformatics, 2004



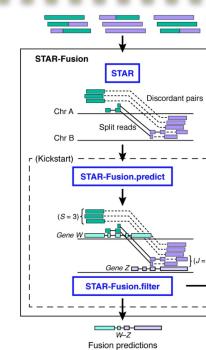
EvidenceModeler
Genome Biology, 2008



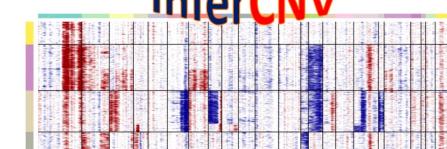
Chimera Slayer
Genome Research, 2011



Nature Biotech, 2011
Nature Protocols, 2013

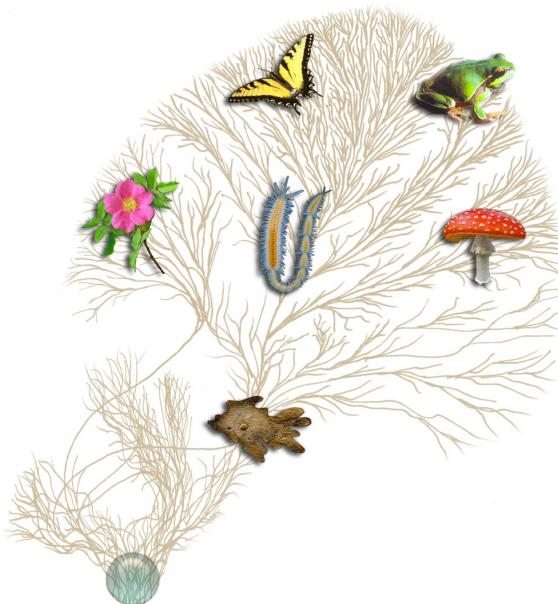


STAR-Fusion
Genome Biology, 2019

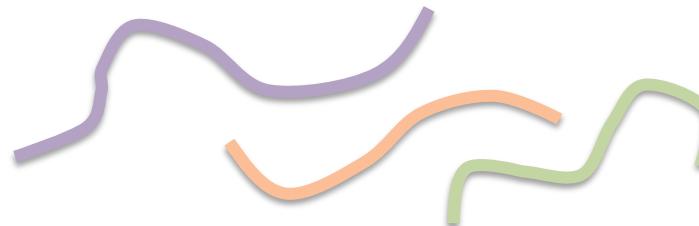


FusionInspector
Cell Reports Methods, 2023

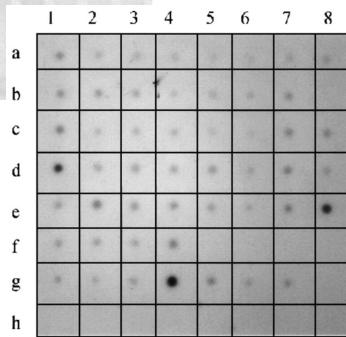
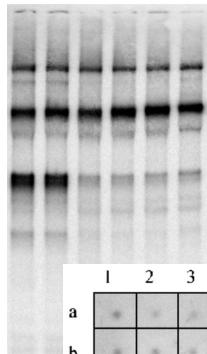
Biological Investigations Empowered by Transcriptomics



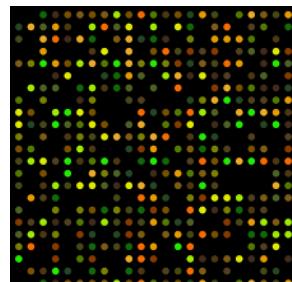
Extract RNA,
... some protocol for processing, ...



Northern

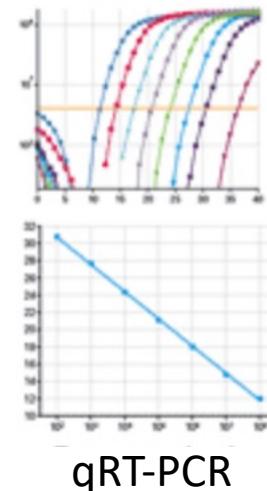


Dot Blot

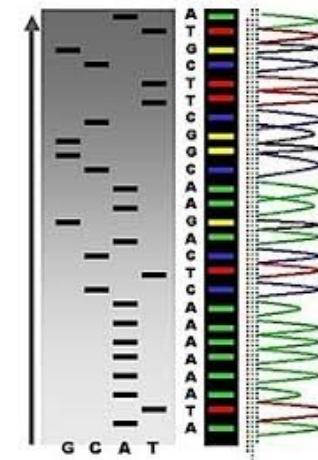


Microarray

Analysis Method
(pick your favorite)



qRT-PCR



Sanger Sequencing



Minion



Other...



MinION MkI: portable, real time biological analyses

MinION

Historical Timeline to Modern Transcriptomics (from 1970)

Reverse Transcription (1970)

Northern Blot
Sanger Sequencing
(1977)

Expressed Sequence Tags (1992)

cDNA microarrays (1995)

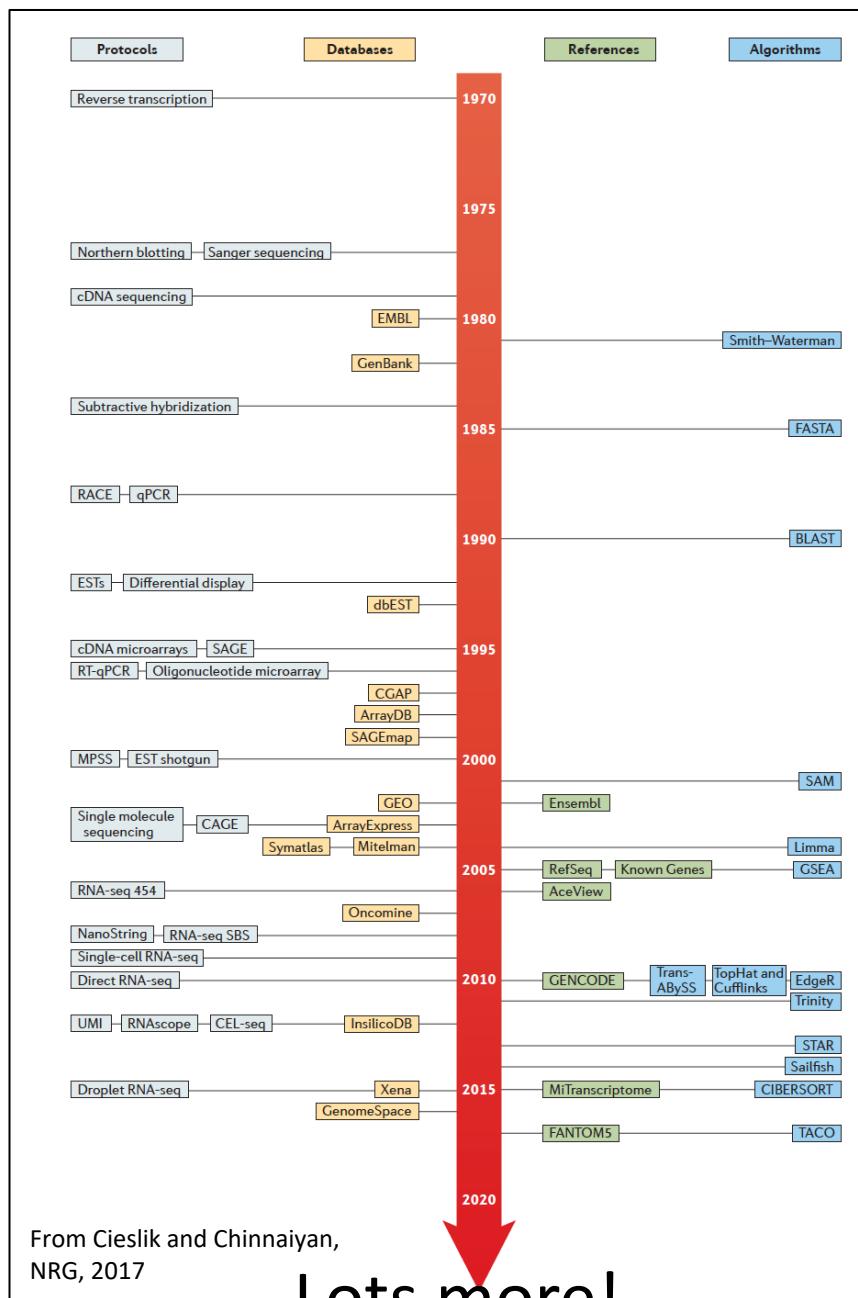
RNA-Seq (2006-2008)

PacBio IsoSeq (2014)

Droplet single cell RNA-Seq (2015)

Direct RNA Seq Nanopore (2018)

SlideSeq-v2 (2021)



Note: Just a small sampling of what's available.

Smith Waterman (1981)

BLAST (1990)

SAMtools (2009)
Tophat/Cufflinks (2010)

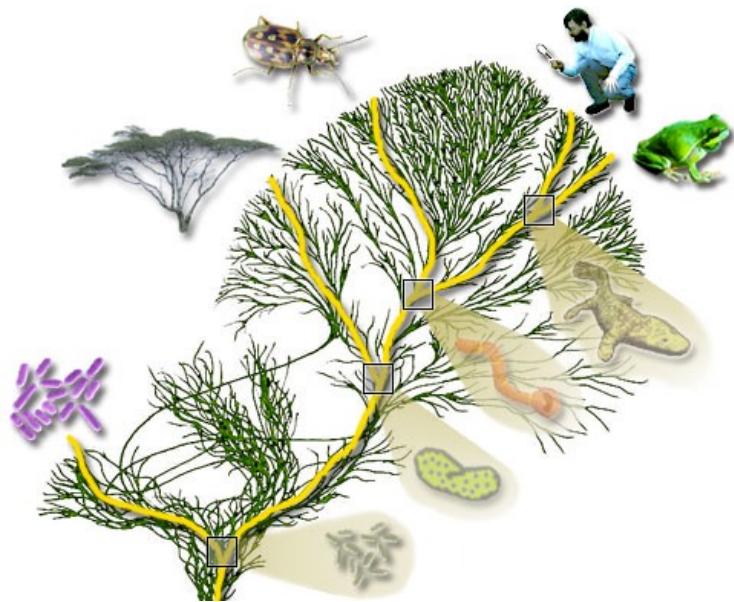


RSEM

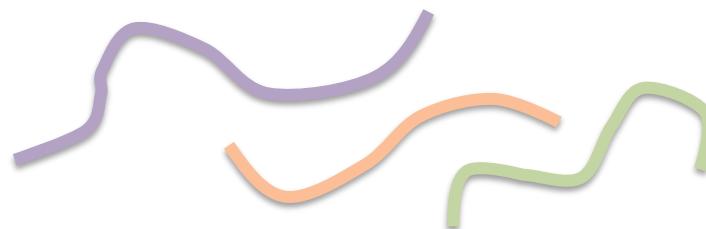
(2011)

STAR (2013)
StringTie (2015)
Kallisto (2016)
Salmon (2017)
minimap2 (2018)
Seurat-v2 (2021)

Modern Transcriptome Studies Empowered by RNA-seq



Extract RNA, convert to cDNA

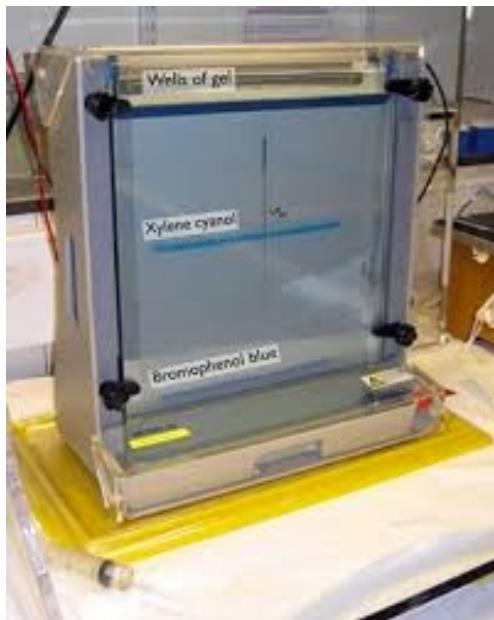
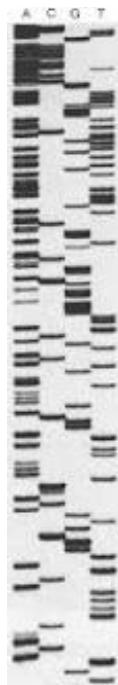


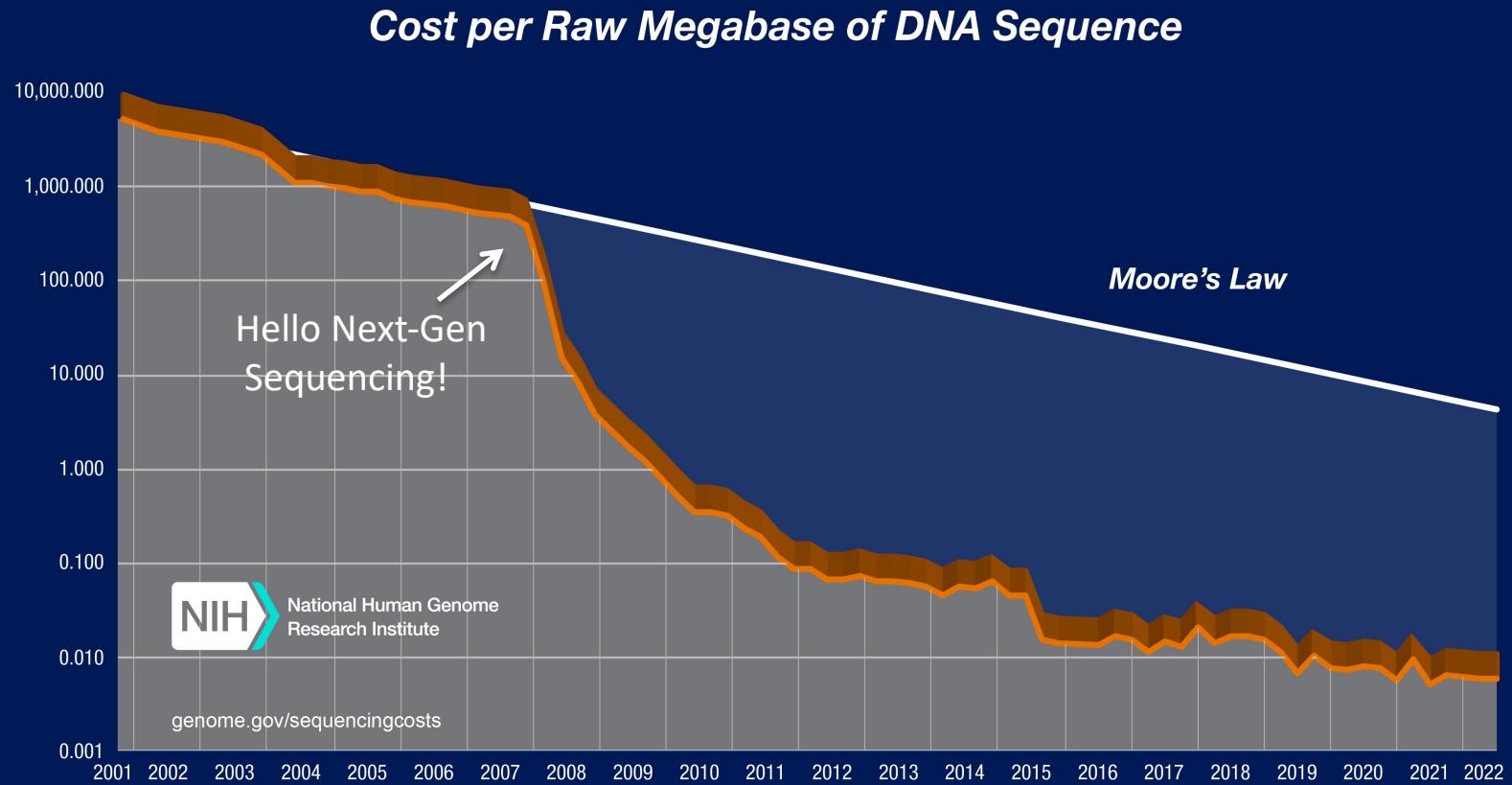
Next-gen Sequencer
(pick your favorite)

Millions to Billions of Reads

Personal Reflections...

Circa 1995





From <https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data>

Generating RNA-Seq: How to Choose?

Platform	iSeq Project Firefly 2018	MinSeq	MiSeq	Next Seq 550	HiSeq 2500 RR	Hiseq 2500 V3	HiSeq 2500 V4	HiSeq 4000	HiSeq X	Nova Seq S1 2018	Nova Seq S2	Nova Seq S4	5500 XL	318 HiQ 520	Ion 530	Ion Proton P1	PGM HiQ 540	RS P6-C4	Sequel	R&D end 2018	Smidg ION RnD	Mini ION R9.5	Grid ION X5	PromethION RnD	PromethION theoretical	QiaGen Gene Reader	BGI SEQ 500	BGI SEQ 50	#
Reads: (M)	4	25	25	400	600	3000	4000	5000	6000	3300	6600	20000	1400	3-5	15-20	165	60-80	5.5	38.5	--	--	--	--	--	--	400	1600	1600	--
Read length: (paired-end*)	150*	150*	300*	150*	100*	100*	125*	150*	150*	150*	150*	150*	60	200 400	200 400	200	200	15K	12K	32K	--	--	--	--	--	--	100*	50	--
Run time: (d)	0.54	1	2	1.2	1.125	11	6	3.5	3	1.66	1.66	1.66	7	0.37	0.16	--	0.16	4.3	--	--	--	2	2	2	--	--	1	0.4	--
Yield: (Gb)	1	7.5	15	120	120	600	1000	1500	1800	1000	2000	6000	180	1.5	7	10	12	12	5	150	4	8	40	2400	11000	80	200	8	--
Rate: (Gb/d)	1.85	7.5	7.5	100	106.6	55	166	400	600	600	1200	3600	30	5.5	50	--	93.75	2.8	--	--	--	4	20	1200	5500	--	200	20	--
Reagents: (\$K)	0.1	1.75	1	5	6.145	23.47	29.9	--	--	--	--	--	10.5	0.6	--	1	1.2	2.4	--	1	--	0.5	1.5	--	--	0.5	--	--	--
per-Gb: (\$)	100	233	66	50	51.2	39.1	31.7	20.5	7.08	18	15	5.8	58.33	--	--	100	--	200	80	6.6	--	62.5	37.5	20	4.3	--	--	--	--
hg-30x: (\$)	12000	28000	8000	5000	6144	4692	3804	2460	849.6	1800	1564	700	7000	--	--	12000	--	24000	9600	1000	--	7500	4500	2400	500	--	600	--	
Machine: (\$)	30K	49.5K	99K	250K	740K	690K	690K	900K	1M	999K	999K	999K	595K	50K	65K	243K	242K	695K	350K	350K	--	--	125K	75K	75K	--	200K	--	

#Page maintained by http://twitter.com/albertvilella http://tinyurl.com/ngslytics #Editable version: http://tinyurl.com/ngsspecsshared

#curl "https://docs.google.com/spreadsheets/d/1GMMfhLyLK0-q8Xklo3YxlWaZA5vVMuhU1kg41g4xLkXc/export?gid=4&format=csv" | grep -v '^#' | grep -v '\"' | column -t -s\| less -S

Stats circa 2018

For current, see: <https://tinyurl.com/wbgcs65>



*Not all shown at scale

Generating RNA-Seq: How to Choose?

Platform	Project Firefly 2018	MiniSeq	MiSeq	Next Seq 550	HiSeq 2500 RR	Hiseq 2500 V
Reads: (M)	4	25	25	400	600	300
Read length: (paired-end*)	150*	150*	300*	150*	100*	100
Run time: (d)	0.54	1	2	1.2	1.125	1
Yield: (Gb)	1	7.5	15	120	120	60
Rate: (Gb/d)	1.85	7.5	7.5	100	106.6	5
Reagents: (\$K)	0.1	1.75	1	5	6.145	23.4
per-Gb: (\$)	100	233	66	50	51.2	39.
hg-30x: (\$)	12000	28000	8000	5000	6144	469
Machine: (\$)	30K	49.5K	99K	250K	740K	690K

#Page maintained by http://twitter.com/albertvilella http://t.co/...
#curl "https://docs.google.com/spreadsheets/d/1GMMfhylK0-q8...



Platform	Mini ION R9.5	Grid ION X5	PromethION RnD	PromethION theoretical	QiaGen Gene Reader	BGI SEQ 500	BGI SEQ 50	#
--	--	--	--	--	400	1600	1600	--
--	--	--	--	--	100*	50	--	--
--	2	2	2	--	--	1	0.4	--
4	8	40	2400	11000	80	200	8	--
--	4	20	1200	5500	--	200	20	--
--	0.5	1.5	--	--	0.5	--	--	--
--	62.5	37.5	20	4.3	--	--	--	--
--	7500	4500	2400	500	--	600	--	--
--	--	125K	75K	75K	--	200K	--	--



Thx Joshua Levin, for the cartoon. ☺

Maybe something more portable?



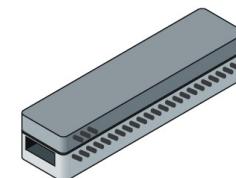
Today's Most Popular Sequencing Technologies



Illumina



Pacific Biosciences



Oxford Nanopore

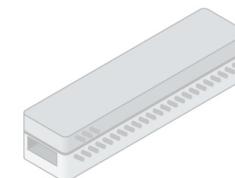
Today's Most Popular Sequencing Technologies



Illumina

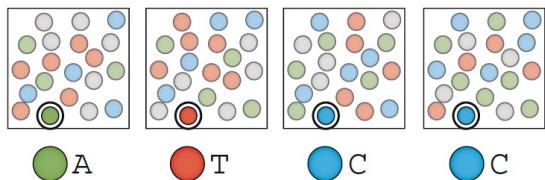
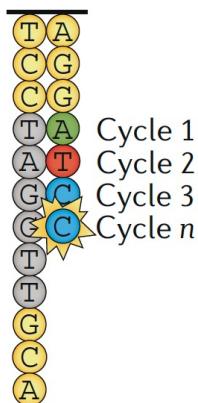


Pacific Biosciences



Oxford Nanopore

Flowcell



Hundreds of millions to billions of
highly accurate but shorter reads.

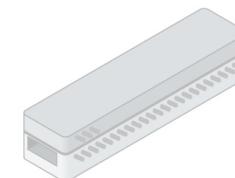
Today's Most Popular Sequencing Technologies



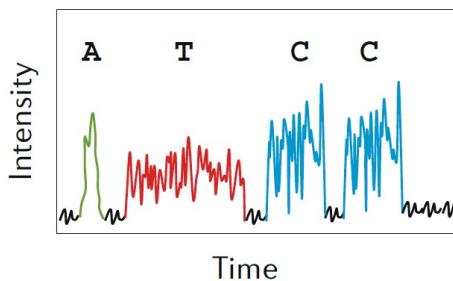
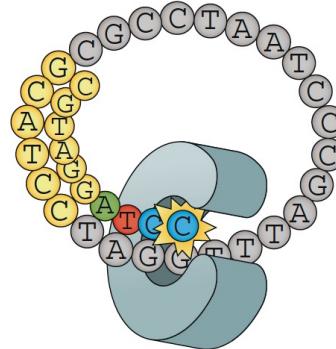
Illumina



Pacific Biosciences



Oxford Nanopore



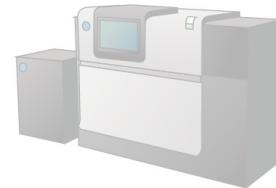
Limited sequencing depth, but
highly accurate full-length single
molecule reads.

Video at: https://www.youtube.com/watch?v=_ID8JyAbwEo

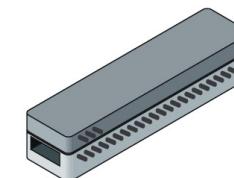
Today's Most Popular Sequencing Technologies



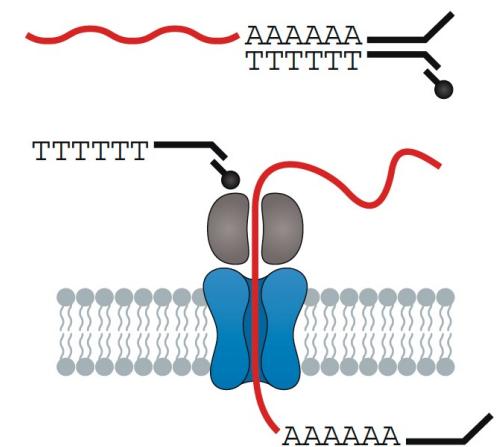
Illumina



Pacific Biosciences



Oxford Nanopore



Video:

<https://nanoporetech.com/how-it-works#fullVideo&modal=fullVideo>

Limited sequencing depth, and moderate-to-highly accurate full-length single molecule reads.

Can do direct RNA sequencing! and find evidence for methylation

A Plethora of Biological Sequence Analyses Enabled by RNA-Seq

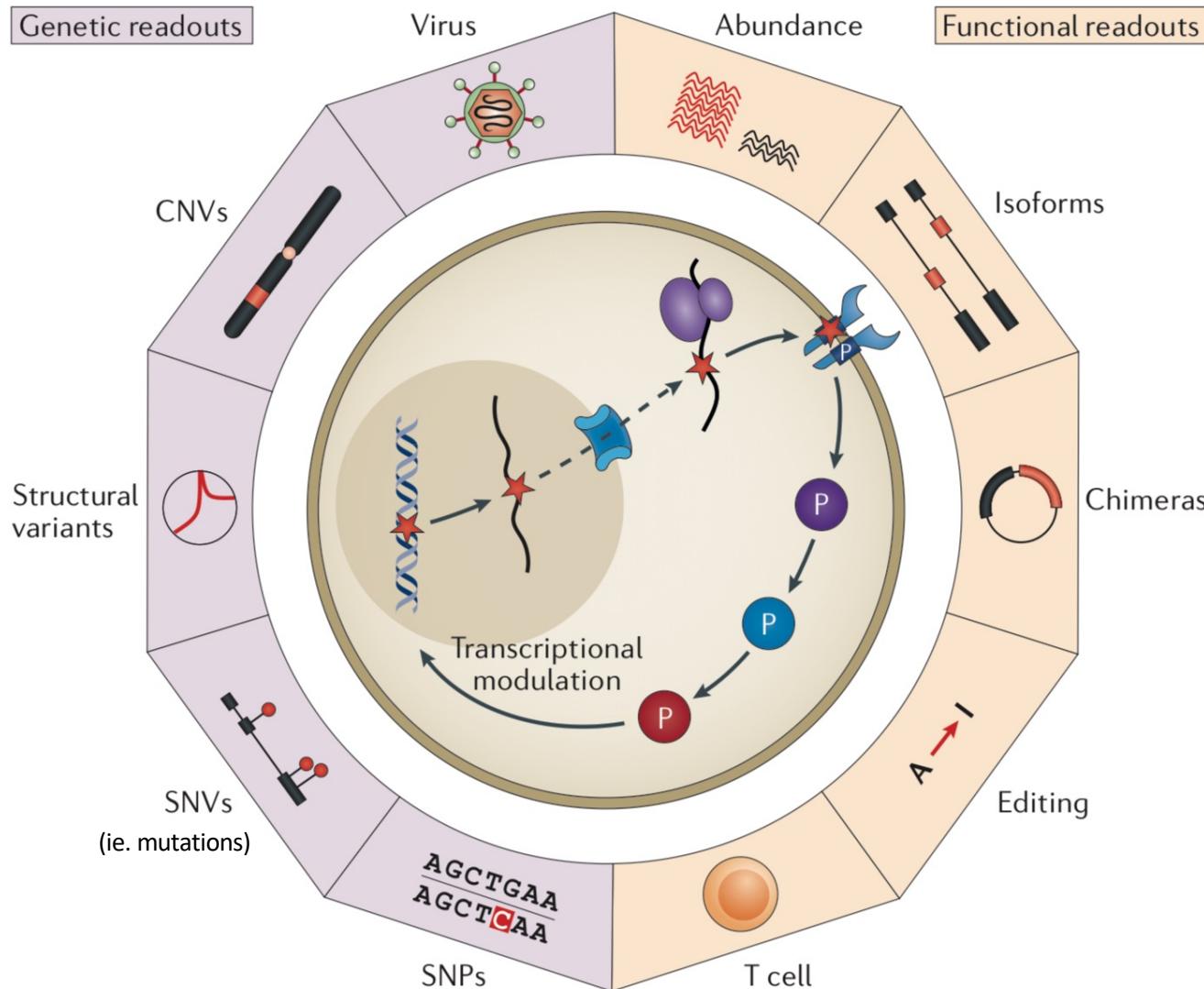
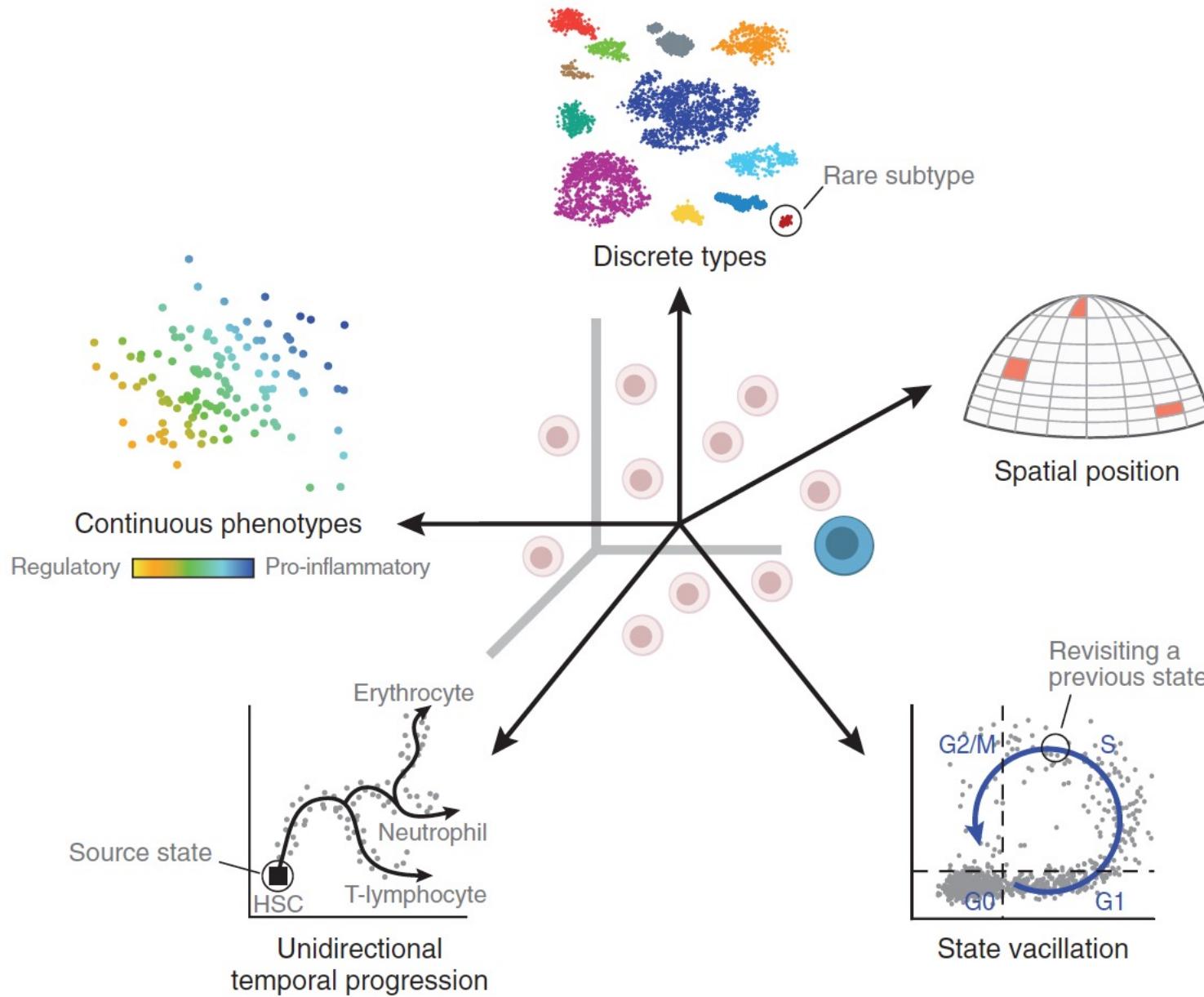


Figure 2 | Transcriptome profiling for genetic causes and functional phenotypic readouts.

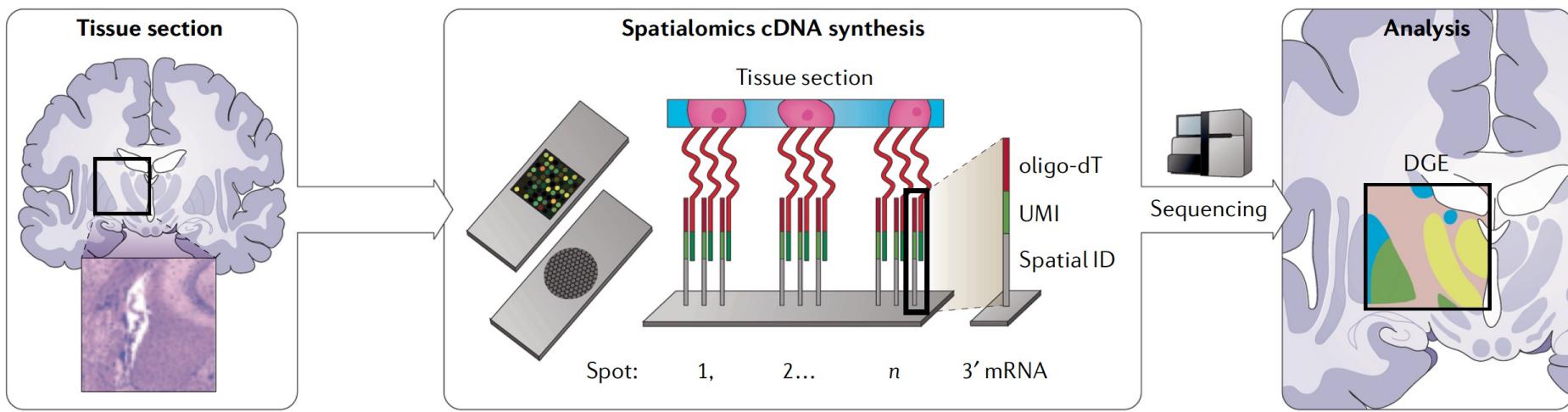
From Cieslik and Chinnaiyan, NRG, 2017

RNA-Seq is Empowering Discovery at Single Cell Resolution



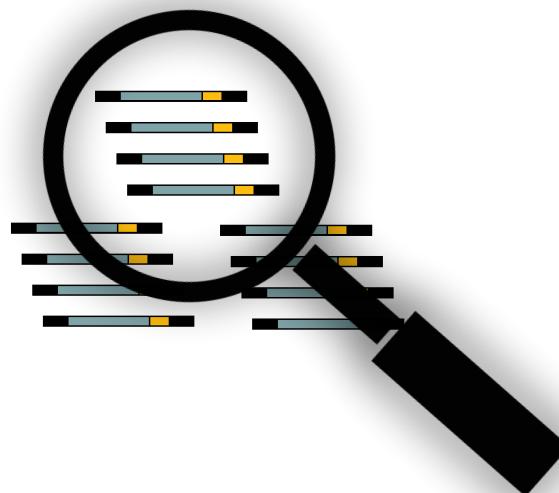
Spatial Transcriptomics

Spatial Encoding



A Myriad of Other Specialized RNA-seq -based Applications

RNA-Sequencing as your lens towards biological discovery



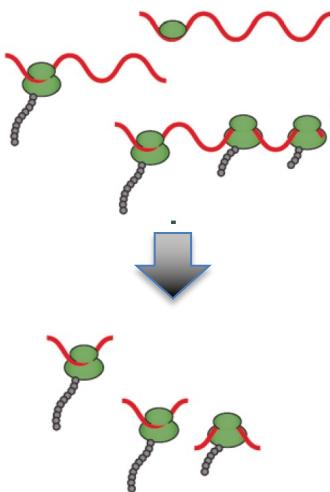
UV crosslink Biotin

RNase V1
(digests dsRNA) RNase S1
(digests ssRNA)

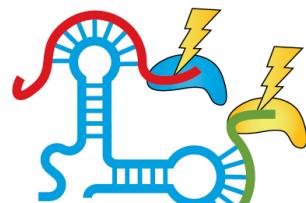
Adapted from "RNA sequencing: the teenage years"
Rory Stark, Marta Grzelak & James Hadfield
Nature Reviews Genetics volume 20, pages631–656(2019)

A Myriad of Other Specialized RNA-seq -based Applications

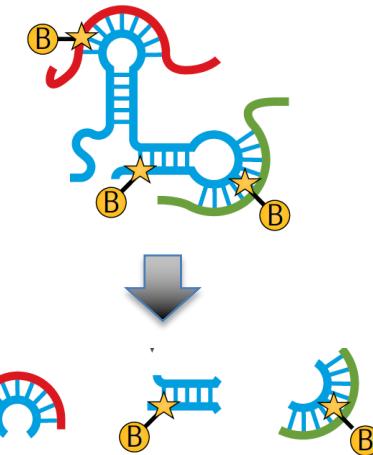
Ribosomal profiling



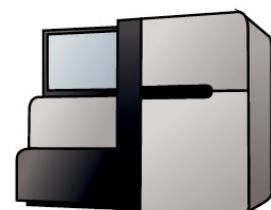
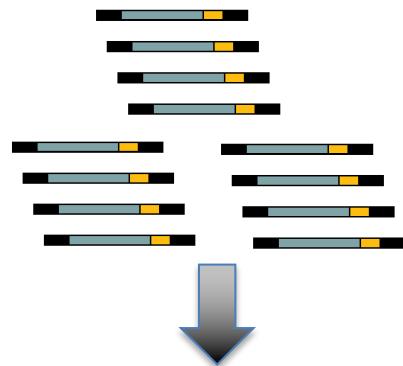
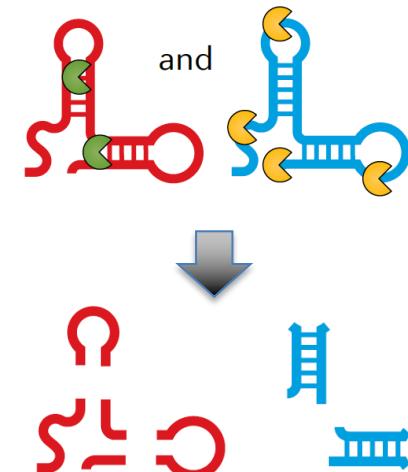
RNA-Protein Interactions



RNA-RNA interactions



RNA Structuromics



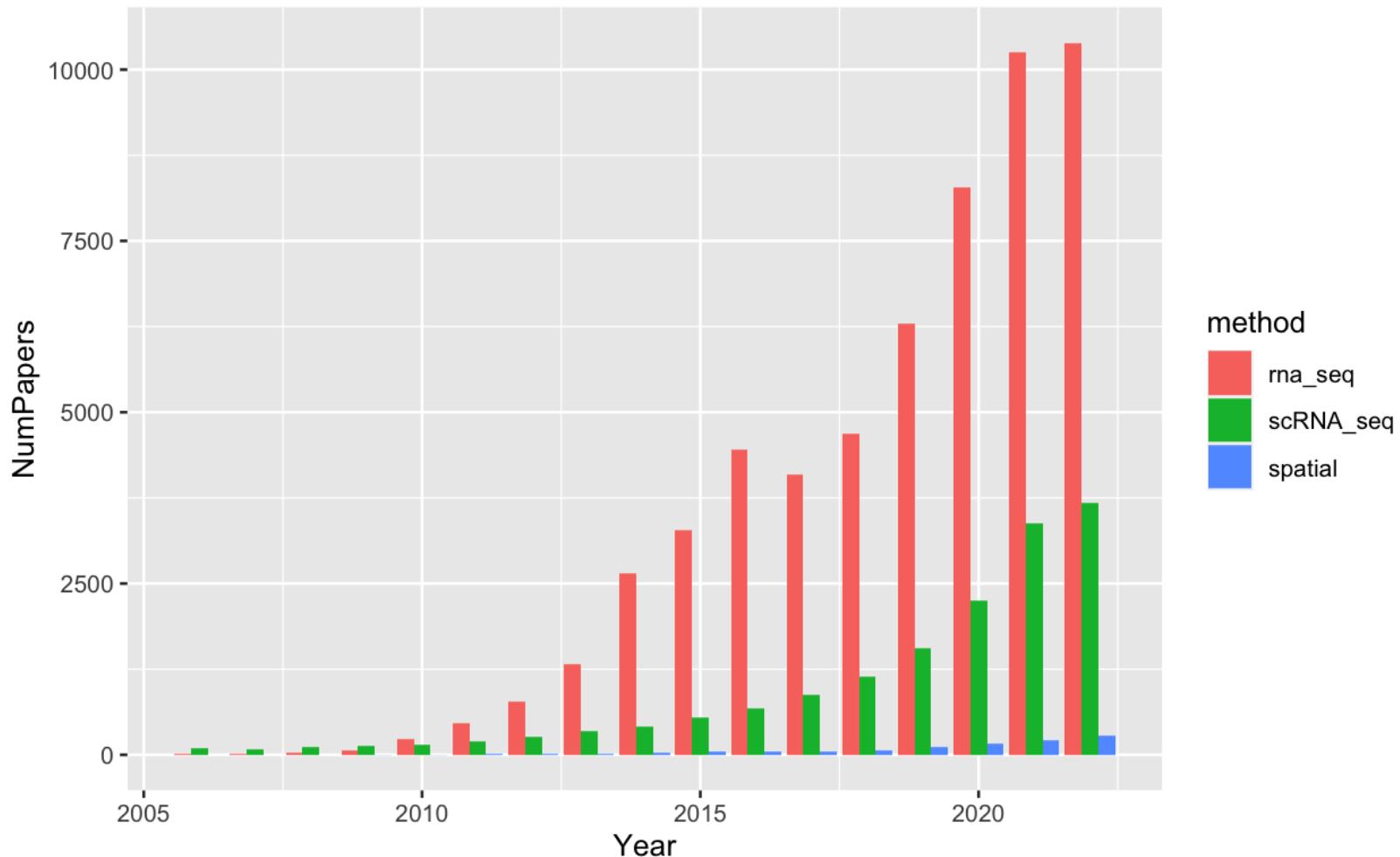
UV crosslink Biotin

RNase V1
(digests
dsRNA) RNase S1
(digests
ssRNA)

Adapted from "RNA sequencing: the teenage years"
Rory Stark, Marta Grzelak & James Hadfield
Nature Reviews Genetics volume 20, pages631–656(2019)

RNA-seq Publication Trend

Paper Counts from PubMed



Transcriptomics Lecture Overview

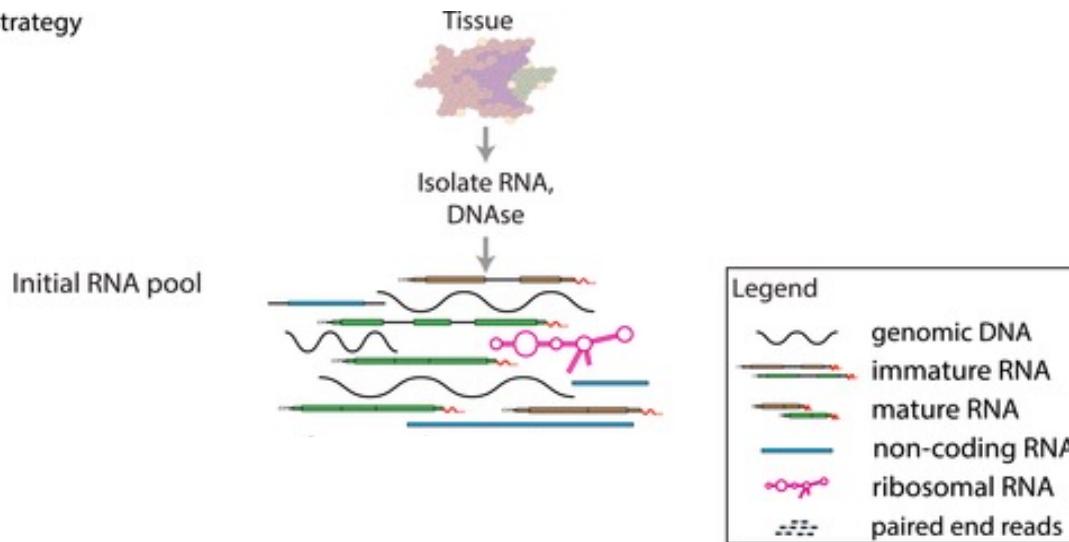
1. Overview of RNA-Seq
2. Transcript reconstruction methods
3. Trinity de novo assembly
4. Transcriptome quality assessment
5. Latest advances for RNA-seq
6. Short lab activity – running Trinity

Part 1. Overview of RNA-Seq

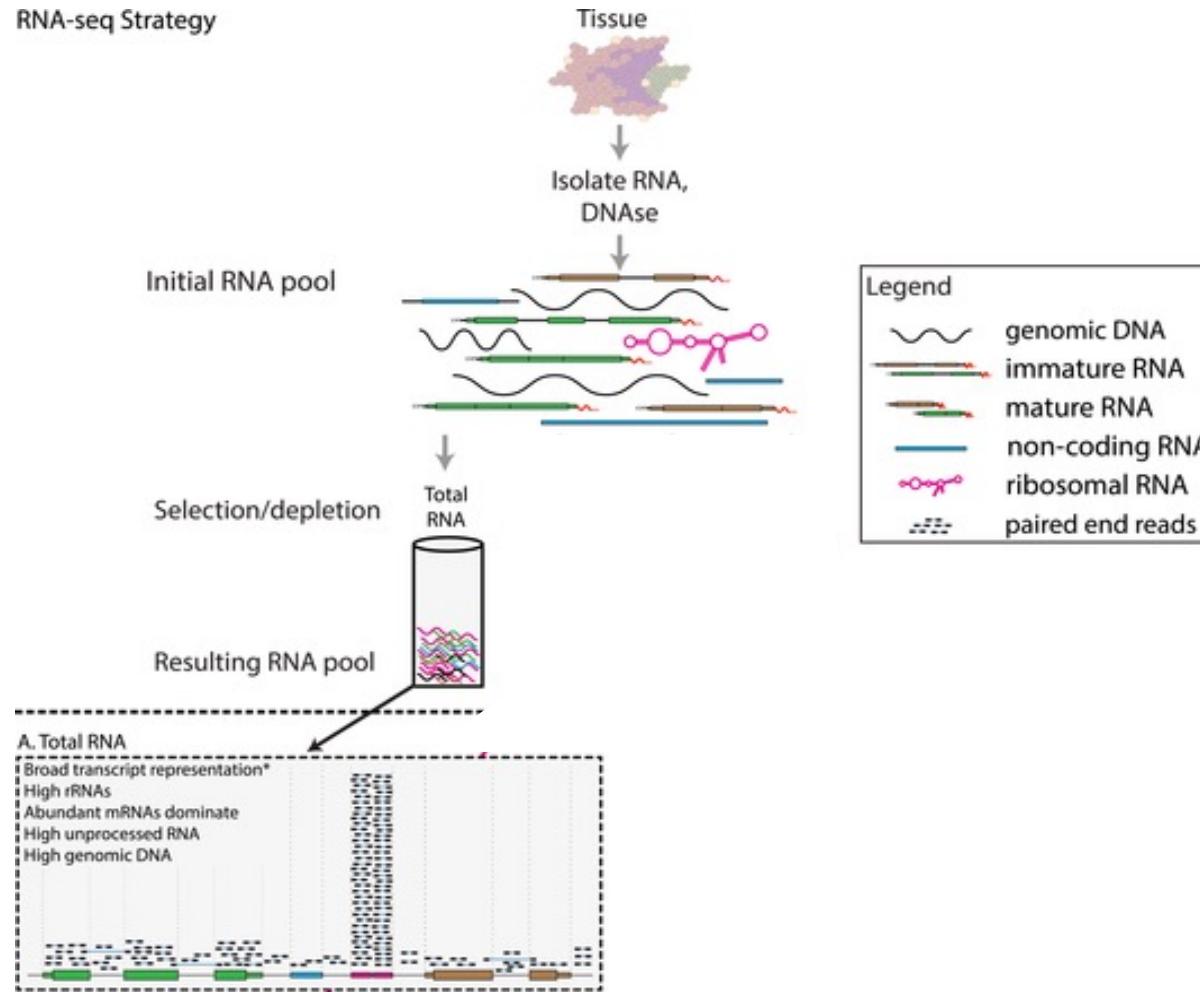


RNA-seq library enrichment strategies that influence interpretation and analysis.

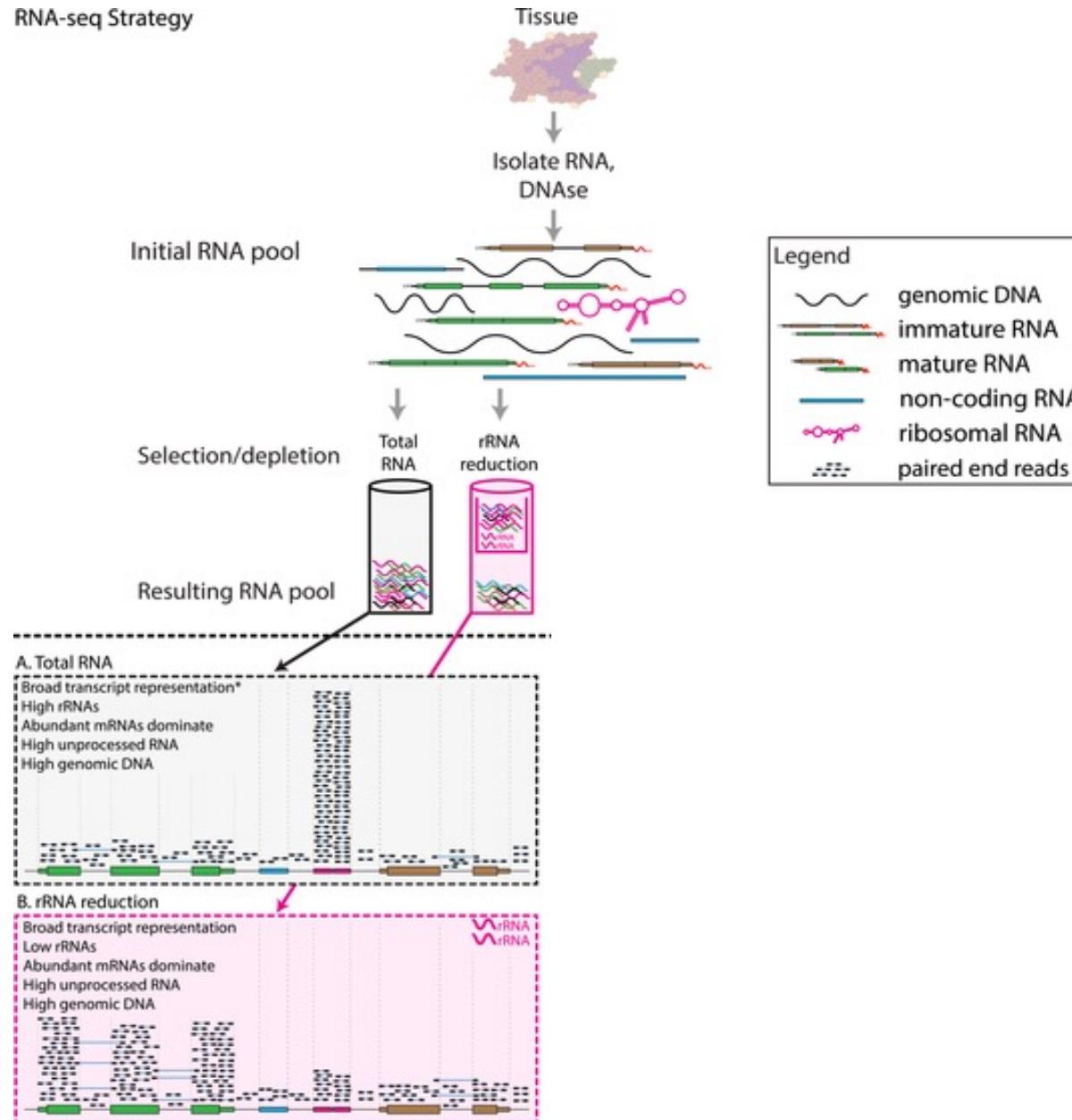
RNA-seq Strategy



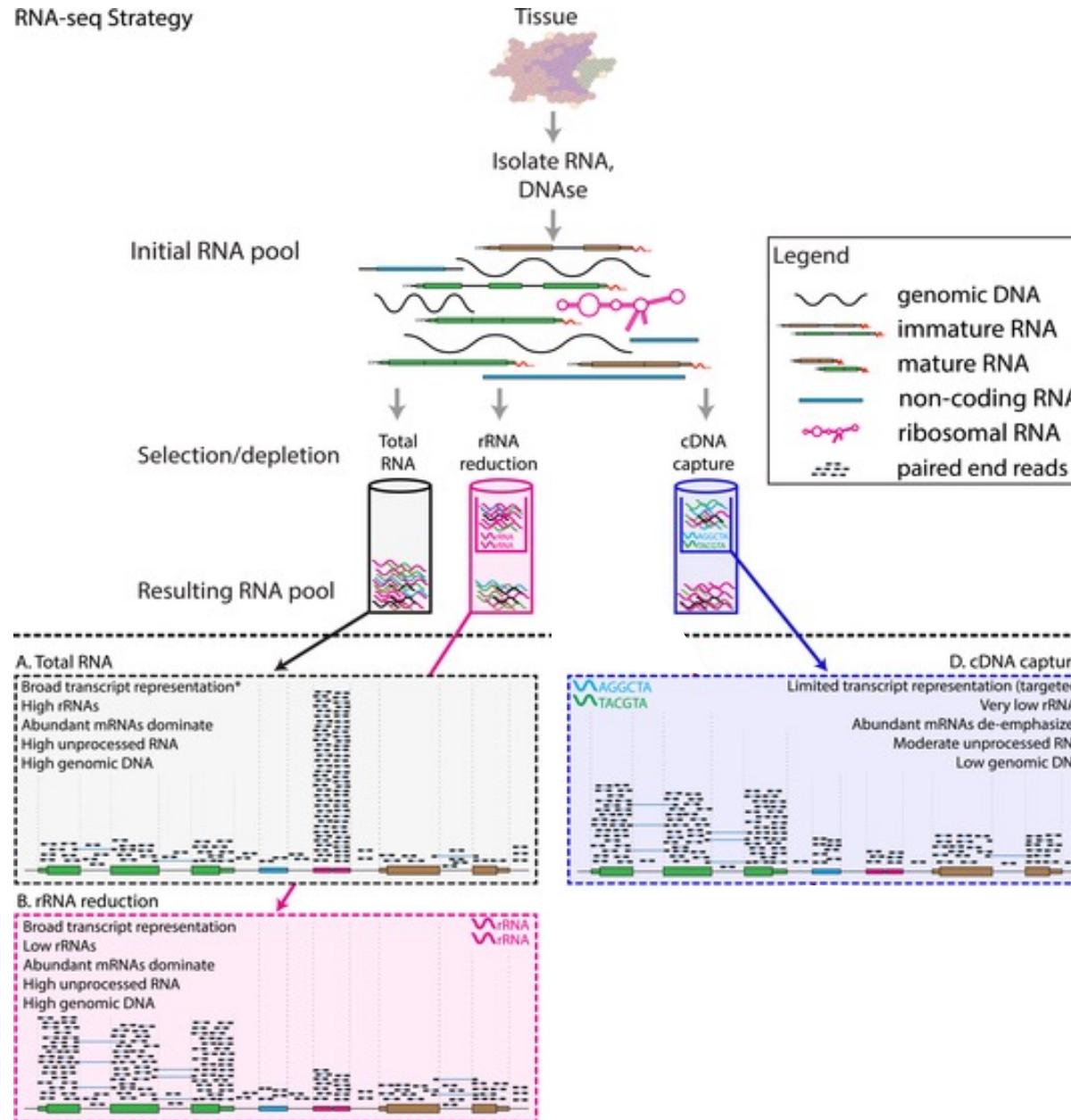
RNA-seq library enrichment strategies that influence interpretation and analysis.



RNA-seq library enrichment strategies that influence interpretation and analysis.



RNA-seq library enrichment strategies that influence interpretation and analysis.

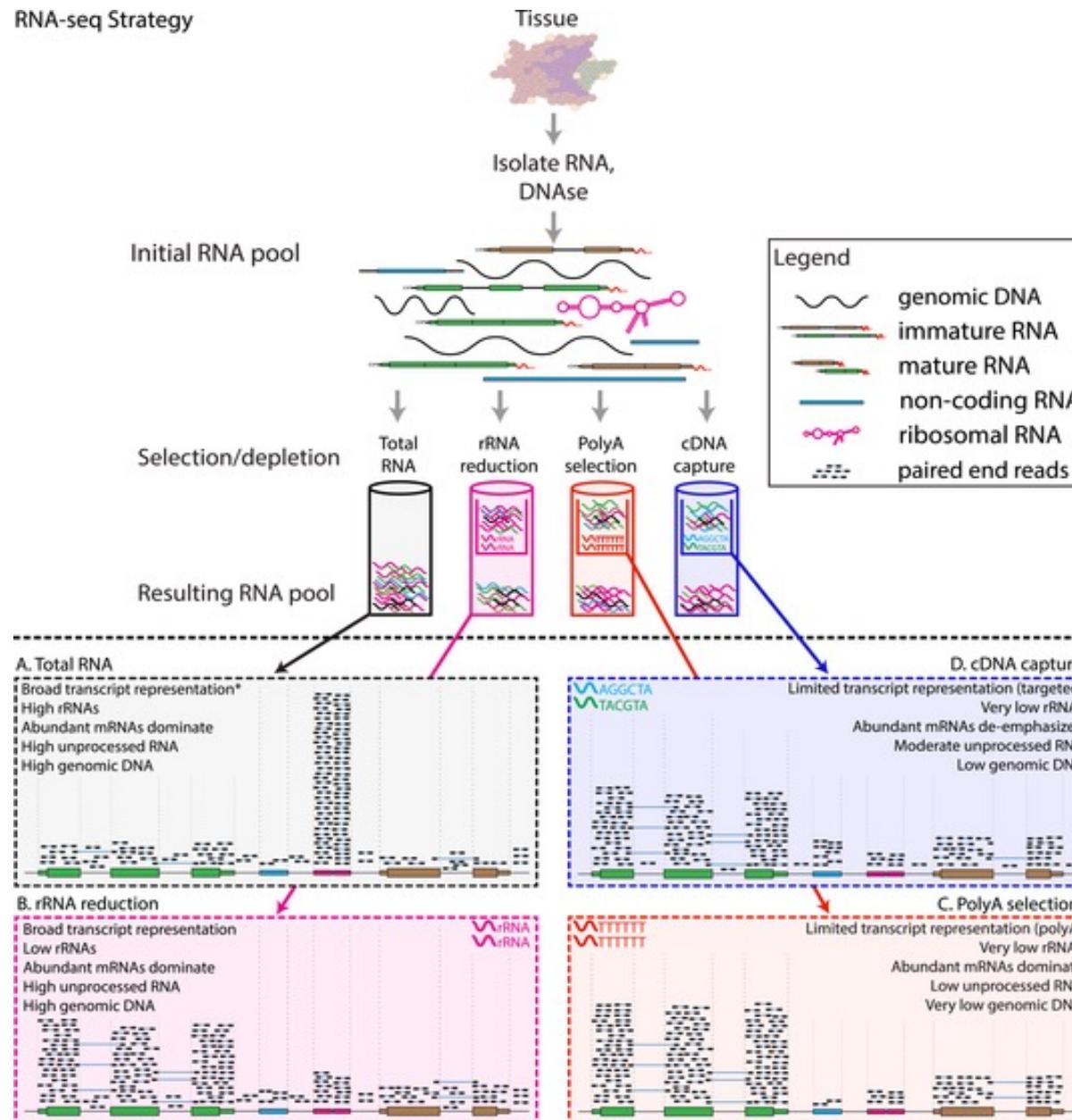


Expected Alignments

<http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004393>

RNA-seq library enrichment strategies that influence interpretation and analysis.

RNA-seq Strategy



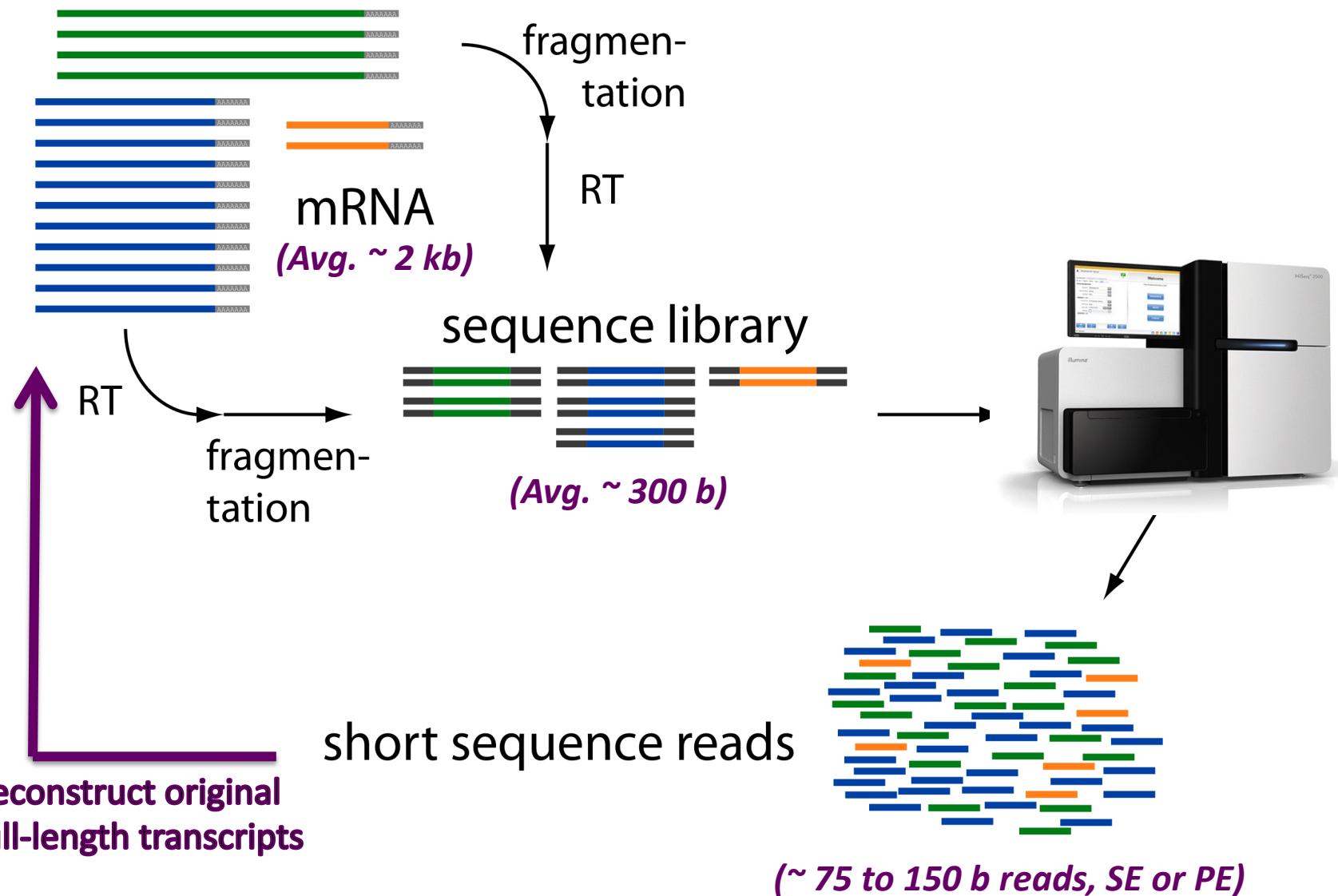
Expected Alignments

<http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004393>

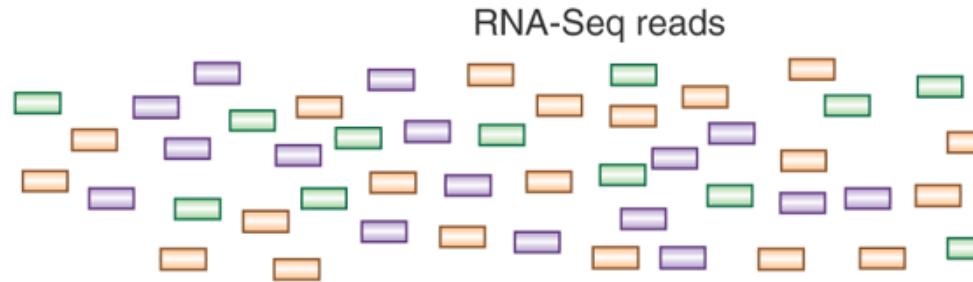
Part 2. Transcript Reconstruction Methods



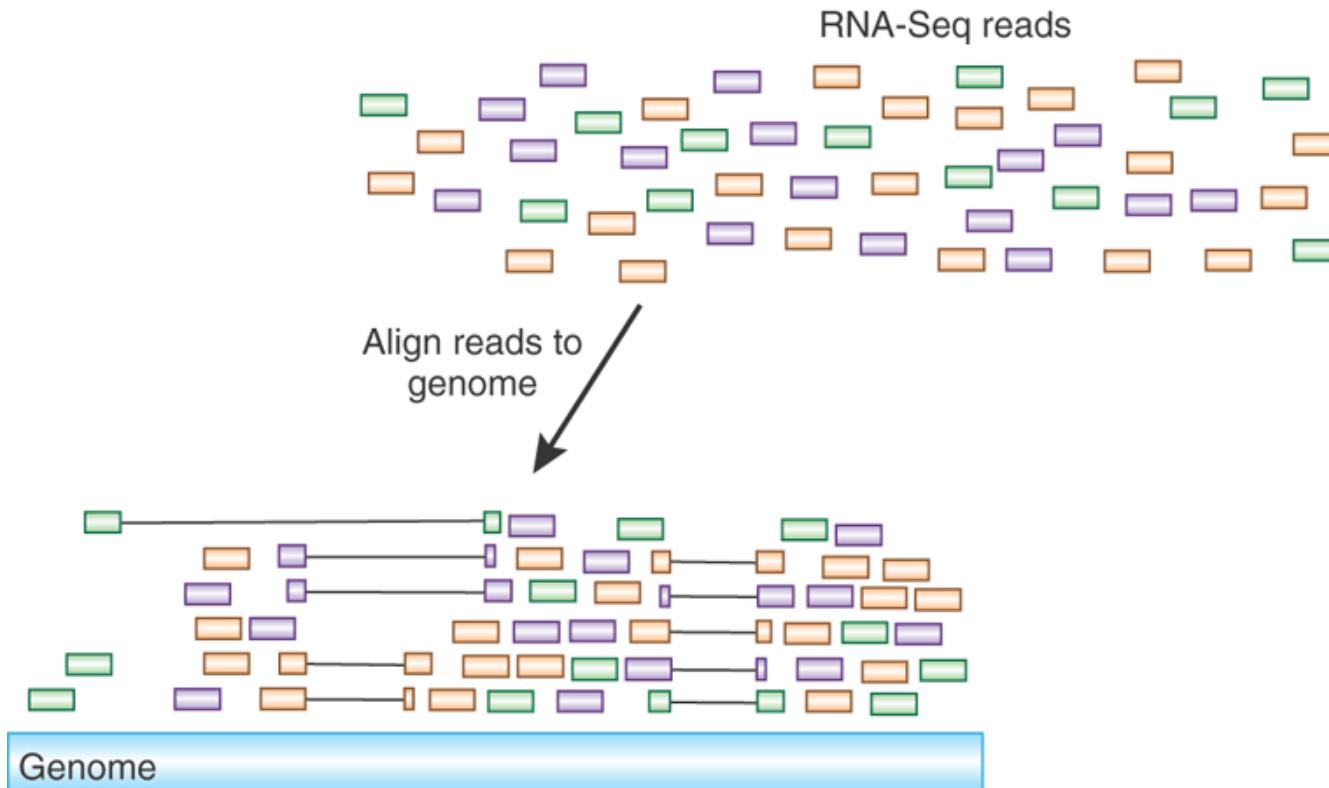
RNA-Seq Challenge: Transcript Reconstruction



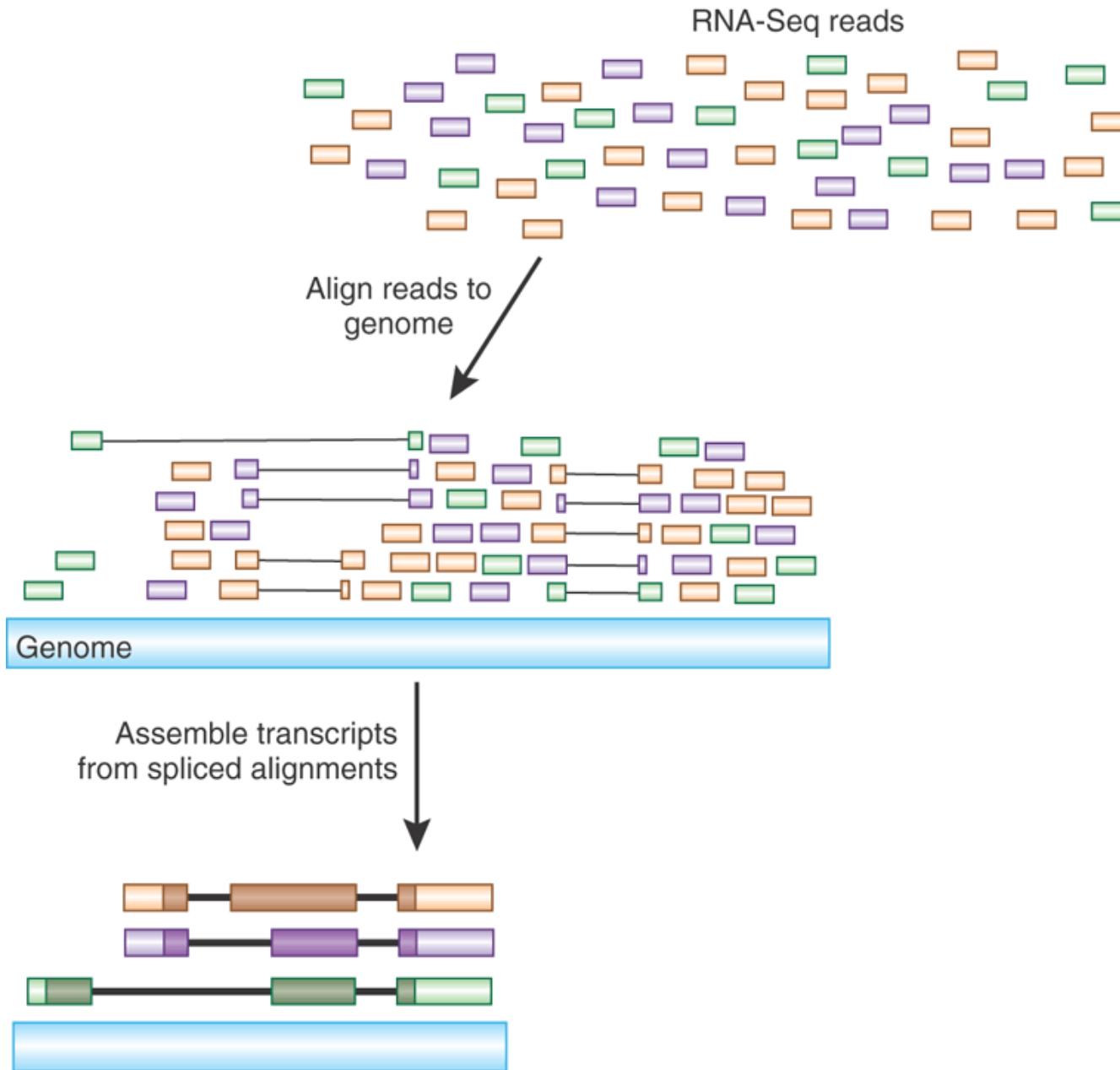
Transcript Reconstruction from RNA-Seq Reads



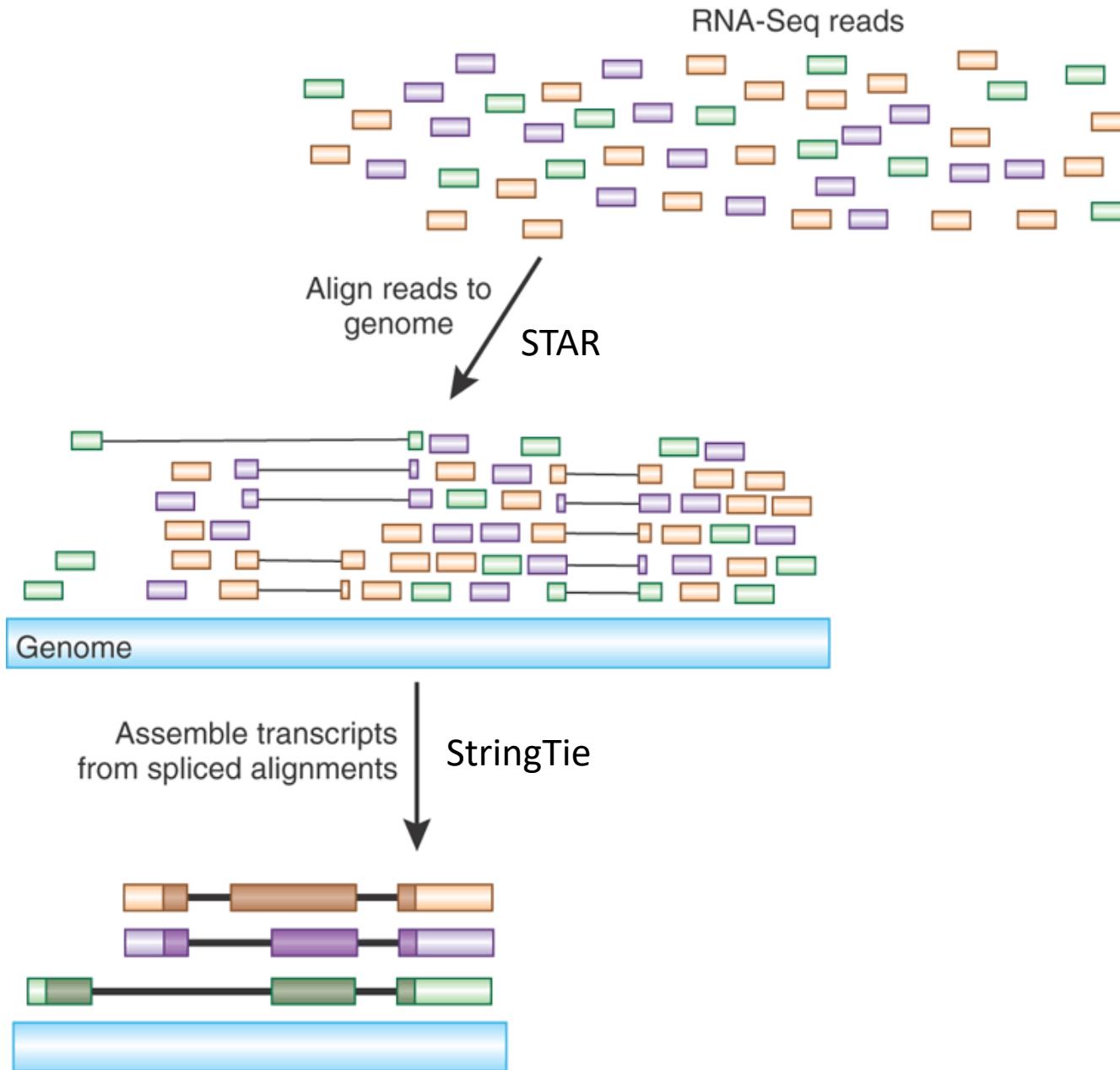
Transcript Reconstruction from RNA-Seq Reads



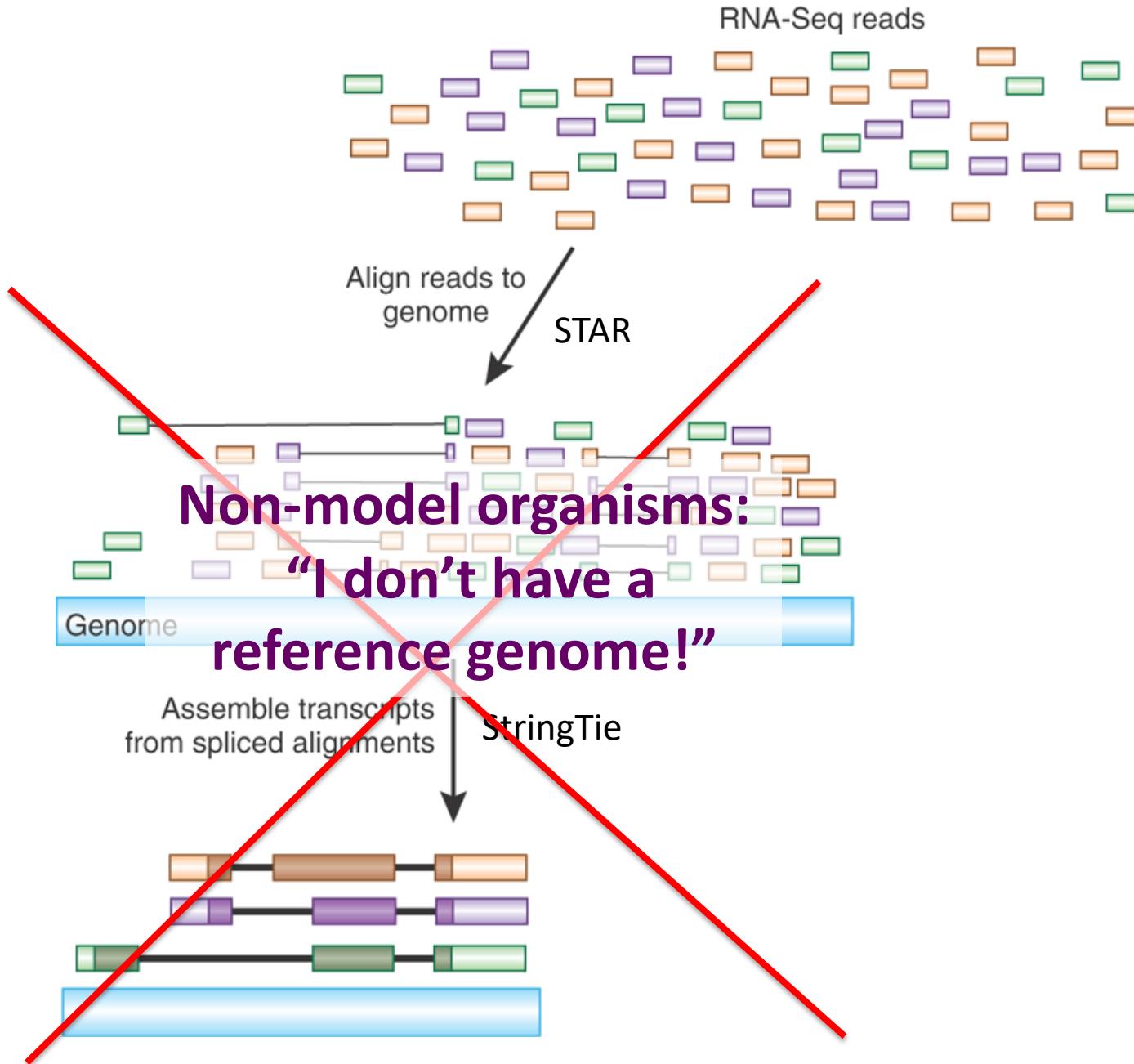
Transcript Reconstruction from RNA-Seq Reads



Transcript Reconstruction from RNA-Seq Reads



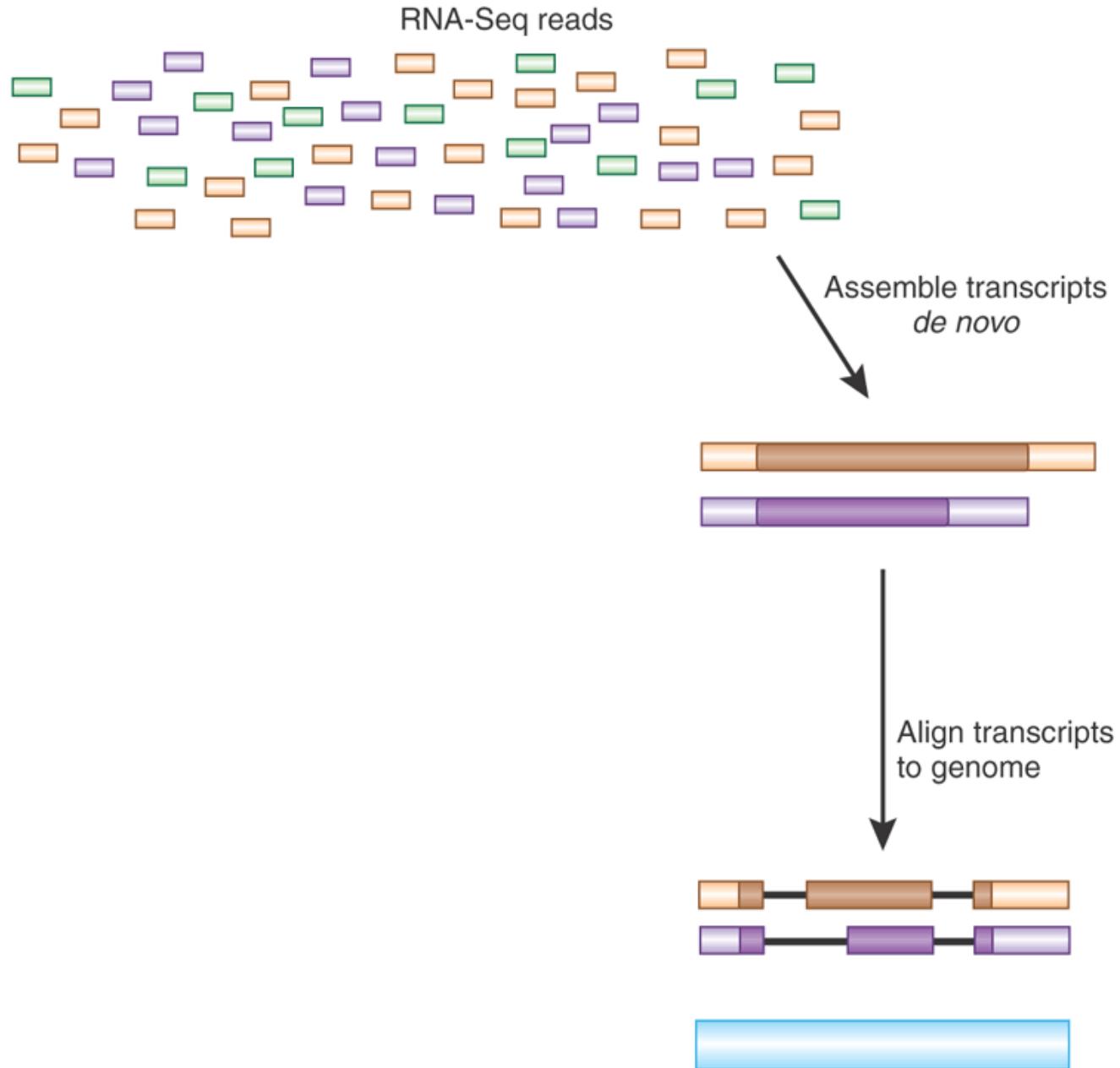
Transcript Reconstruction from RNA-Seq Reads



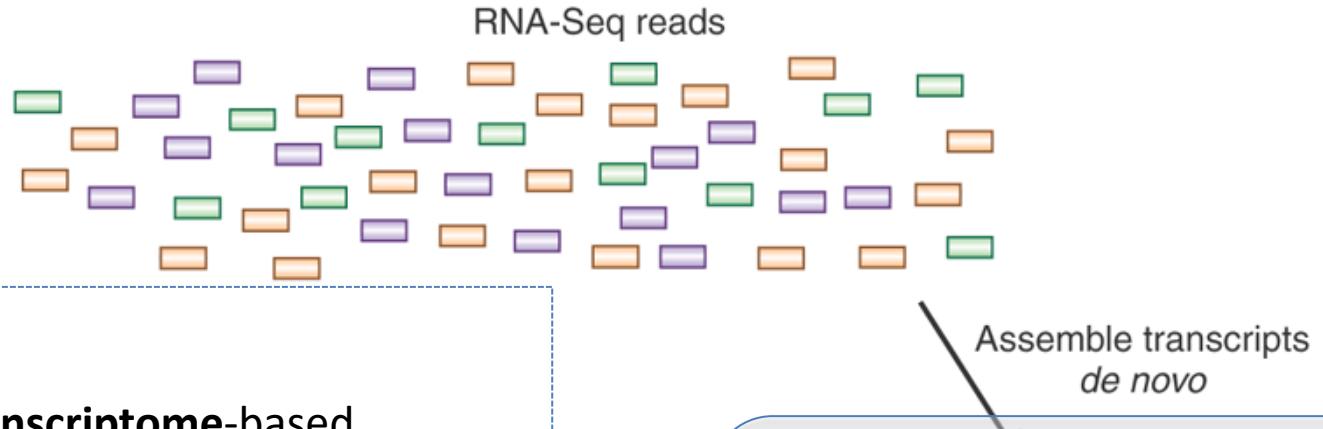
Transcript Reconstruction from RNA-Seq Reads



Transcript Reconstruction from RNA-Seq Reads



Transcript Reconstruction from RNA-Seq Reads



End-to-end Transcriptome-based
RNA-Seq Analysis
Software Package

NATURE PROTOCOLS | PROTOCOL

De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis

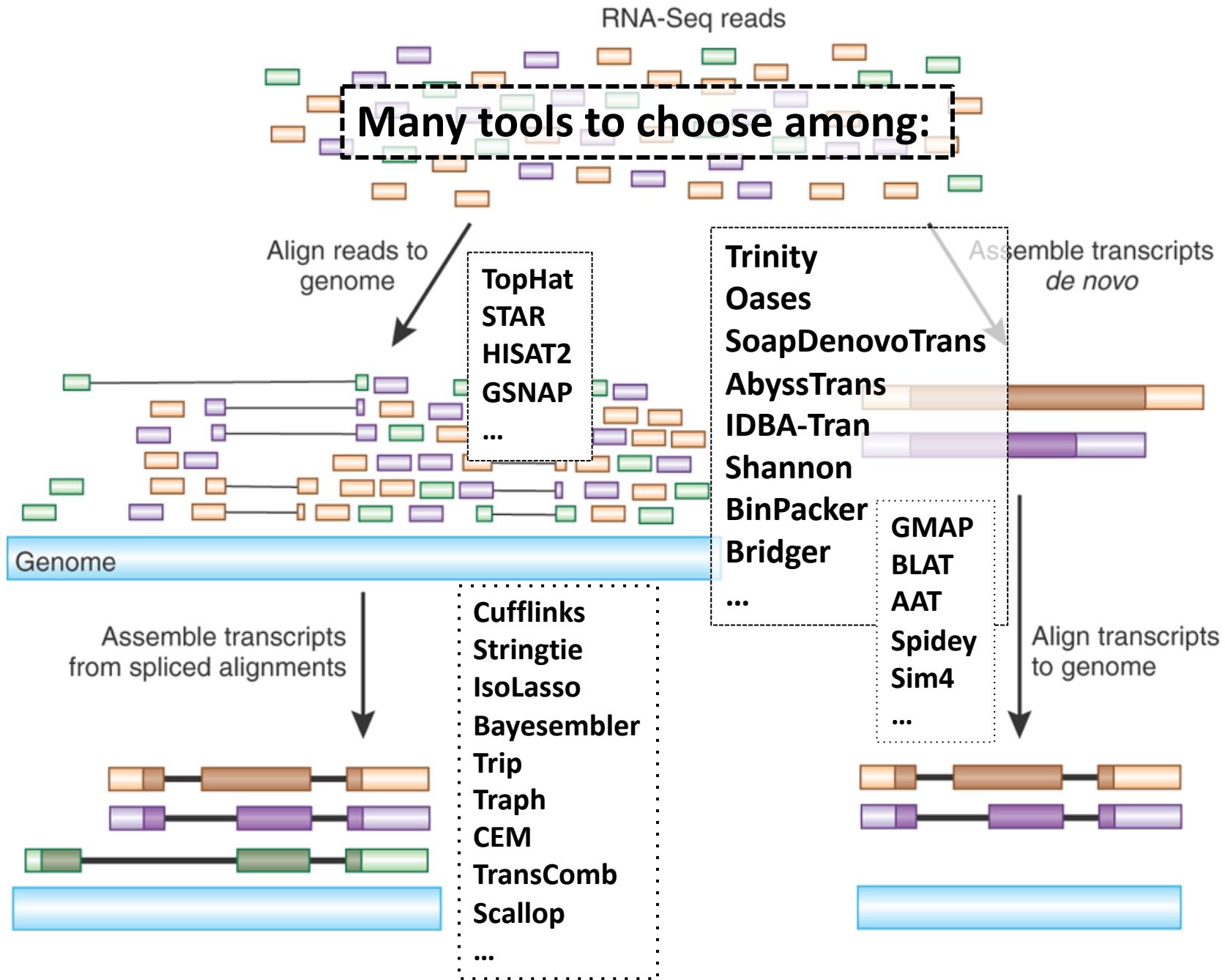
Brian J Haas, Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, Philip D Blood, Joshua Bowden, Matthew Brian Couger, David Eccles, Bo Li, Matthias Lieber, Matthew D MacManes, Michael Ott, Joshua Orvis, Nathalie Pochet, Francesco Strozzi, Nathan Weeks, Rick Westerman, Thomas William, Colin N Dewey, Robert Henschel, Richard D LeDuc, Nir Friedman & Aviv Regev

[Affiliations](#) | [Contributions](#) | [Corresponding authors](#)

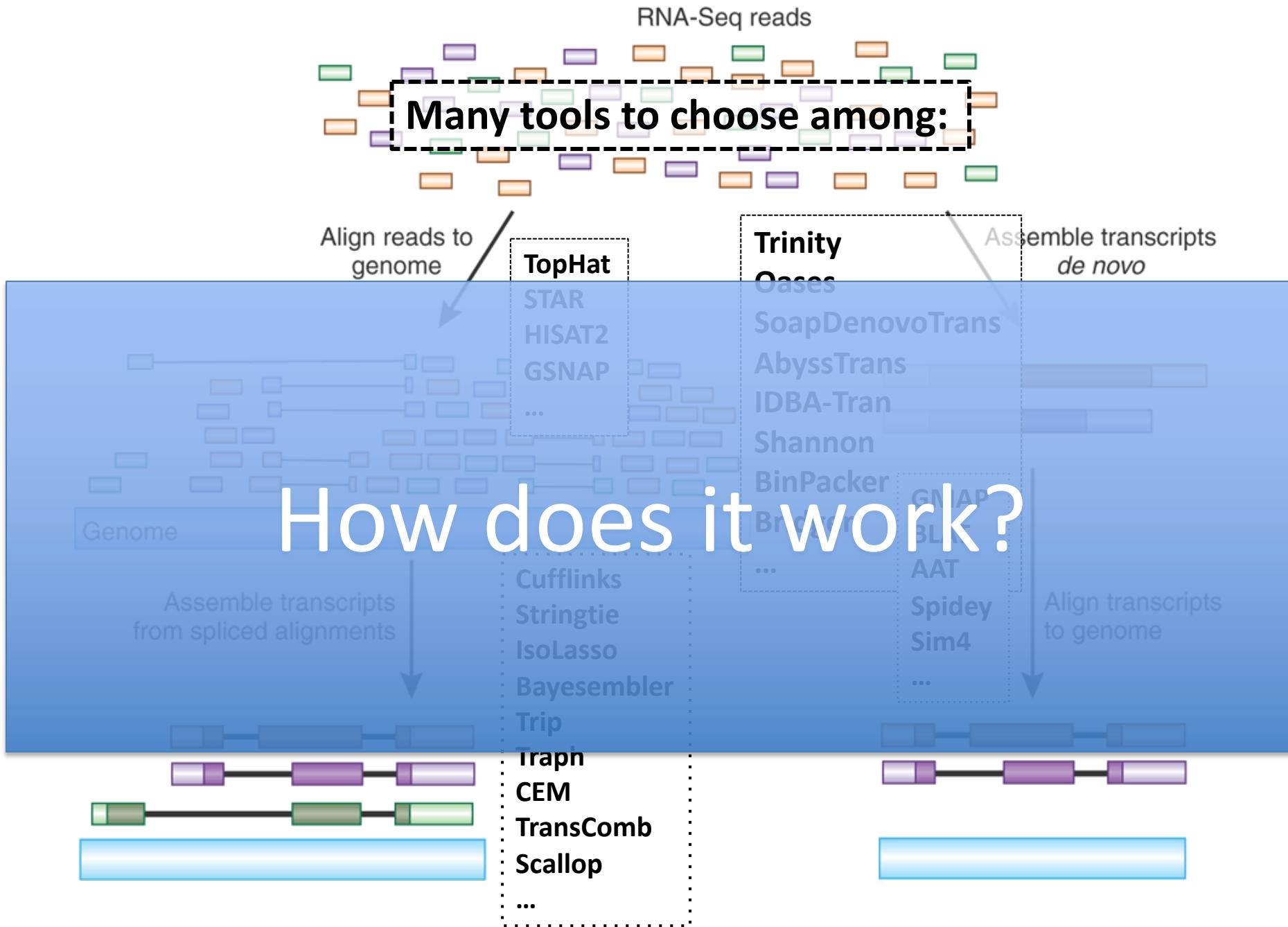
Nature Protocols 8, 1494–1512 (2013) | doi:10.1038/nprot.2013.084

Published online 11 July 2013

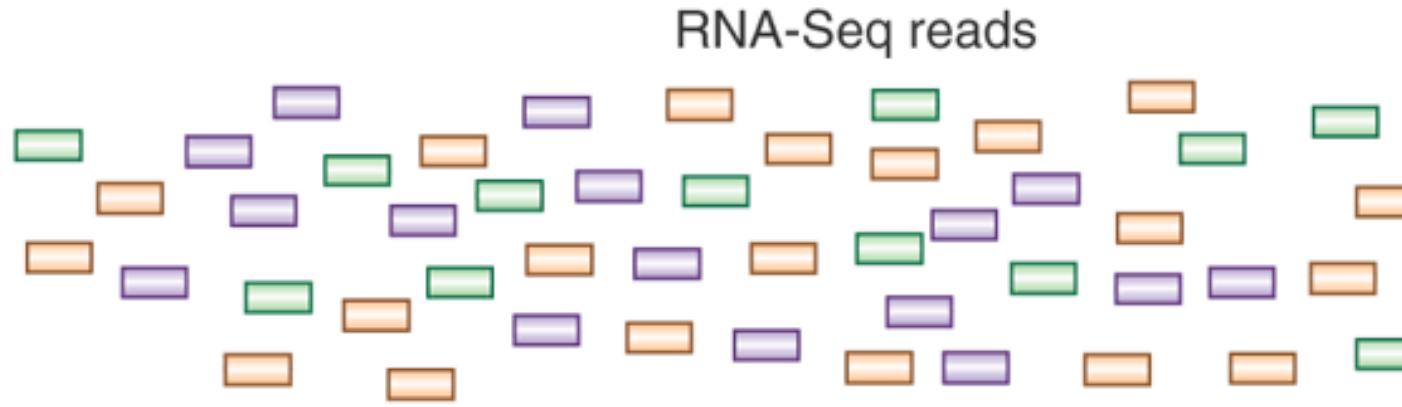
Transcript Reconstruction from RNA-Seq Reads



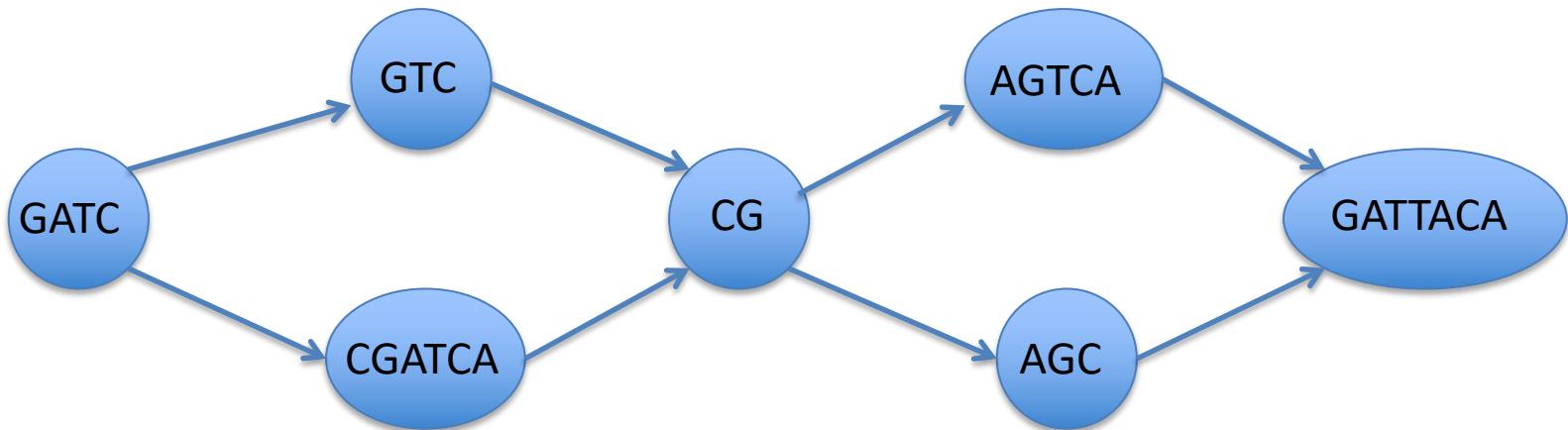
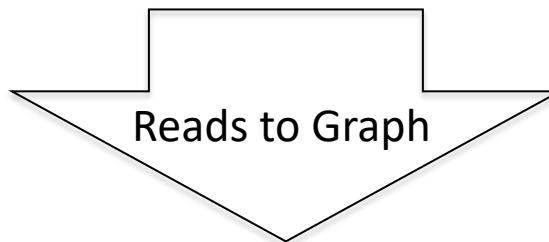
Transcript Reconstruction from RNA-Seq Reads



Graph Data Structures Commonly Used For Assembly

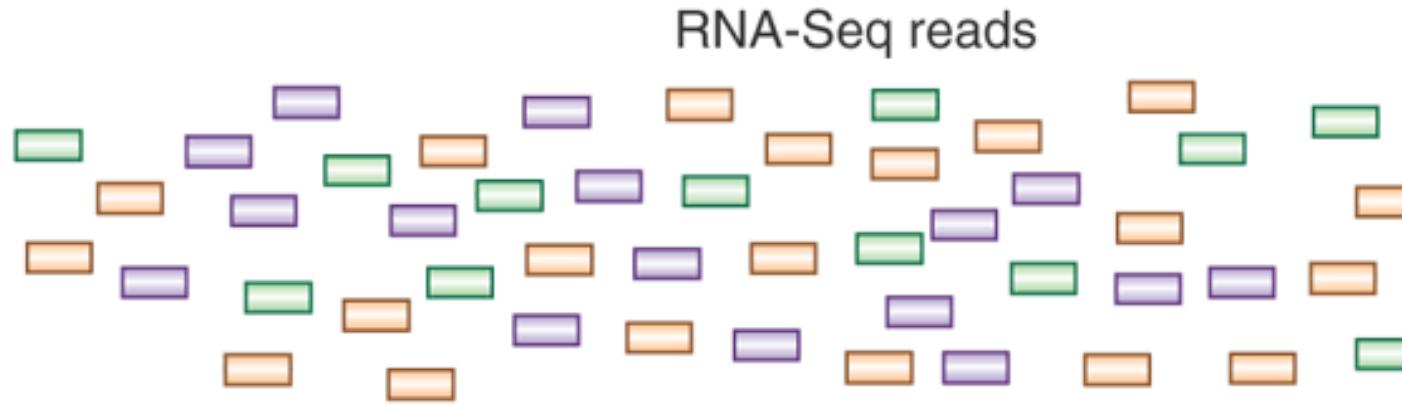


- Sequence
- Order
- Orientation (+, -)
- Overlap

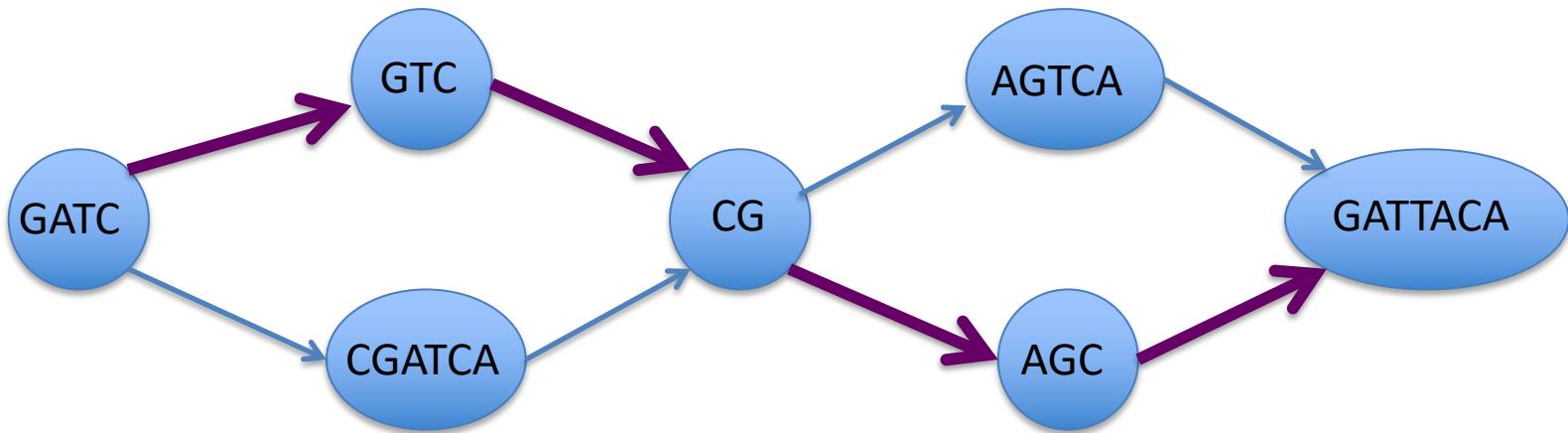
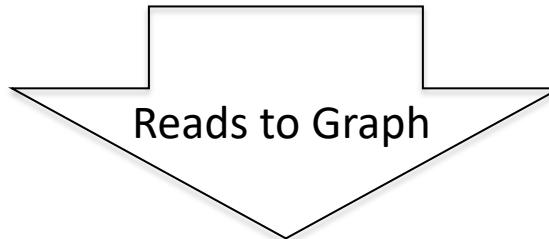


Nodes = sequence (+/-)
Edges = order, overlap

Graph Data Structures Commonly Used For Assembly



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

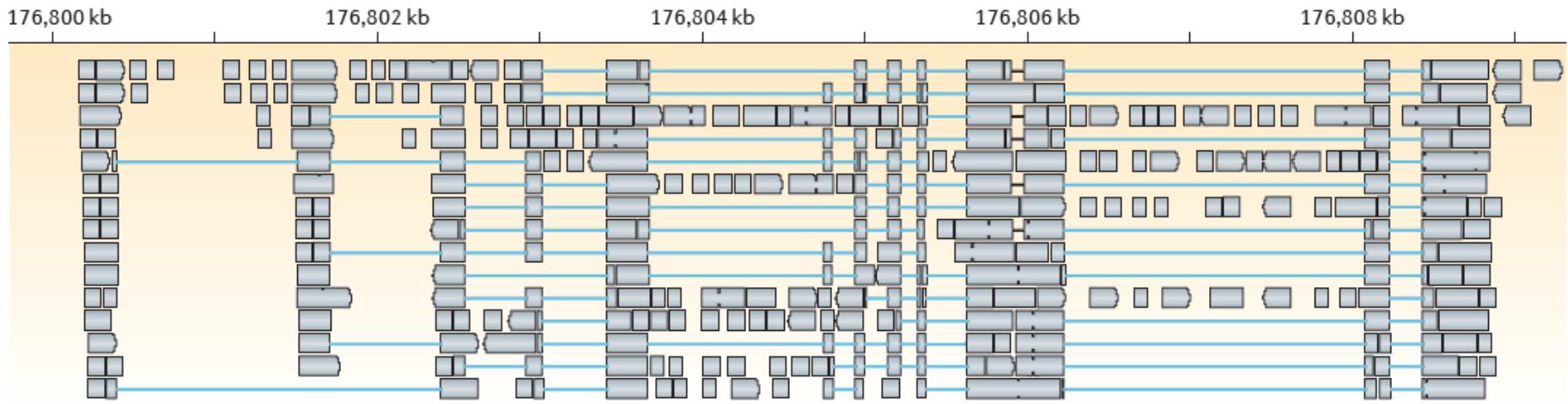


GATCGTCCGAGCGATTACA

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

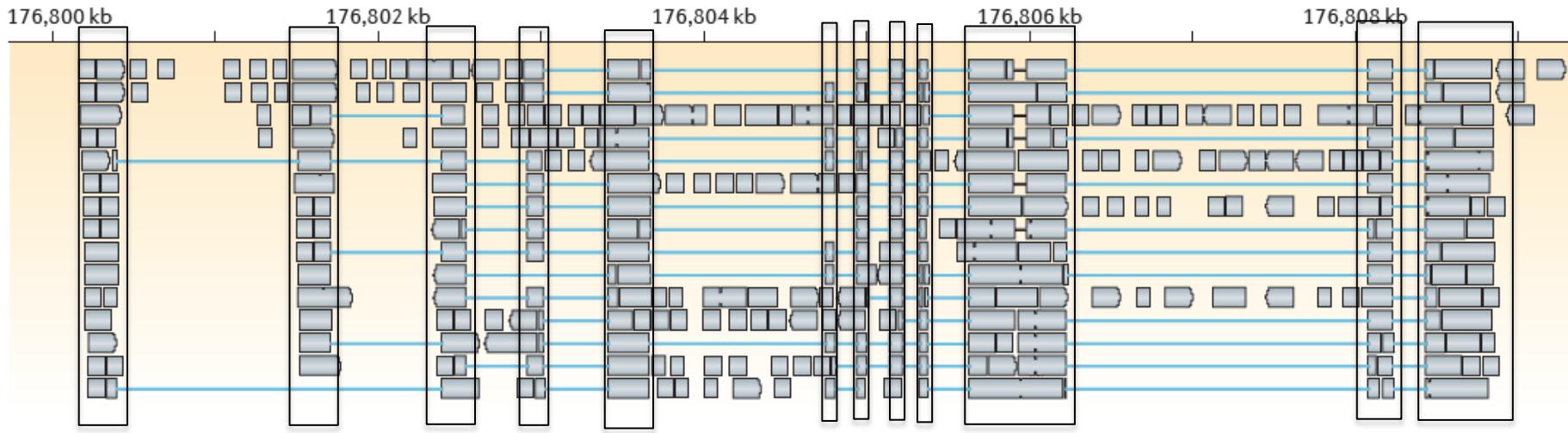
Genome-Guided Transcript Reconstruction

Splice-align reads to the genome



Genome-Guided Transcript Reconstruction

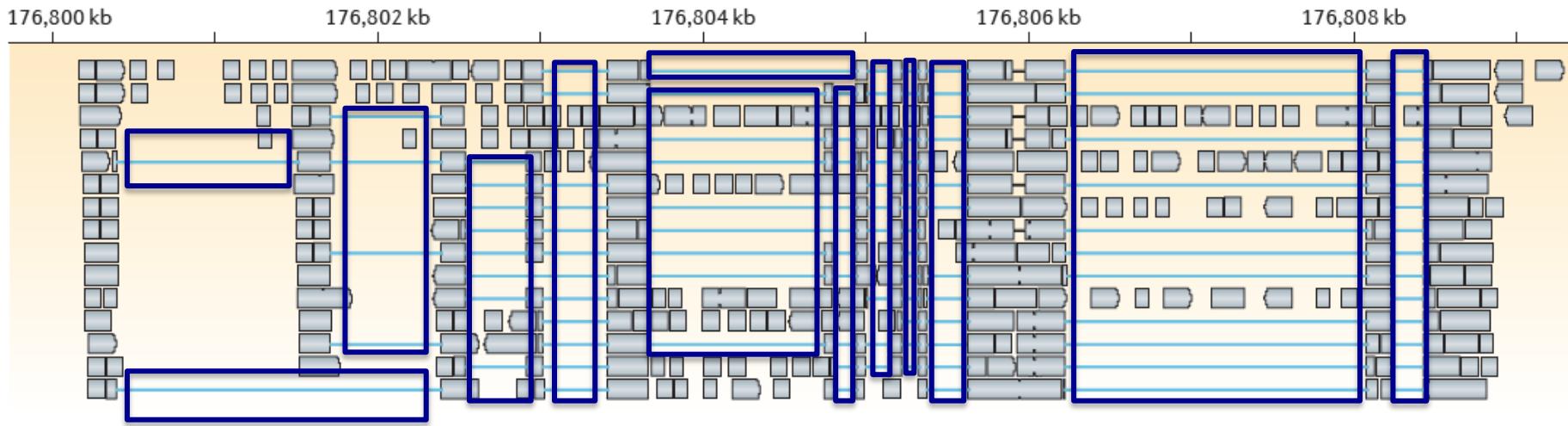
Splice-align reads to the genome



Alignment segment piles => exon regions

Genome-Guided Transcript Reconstruction

Splice-align reads to the genome



Large alignment gaps => introns

Genome-Guided Transcript Reconstruction

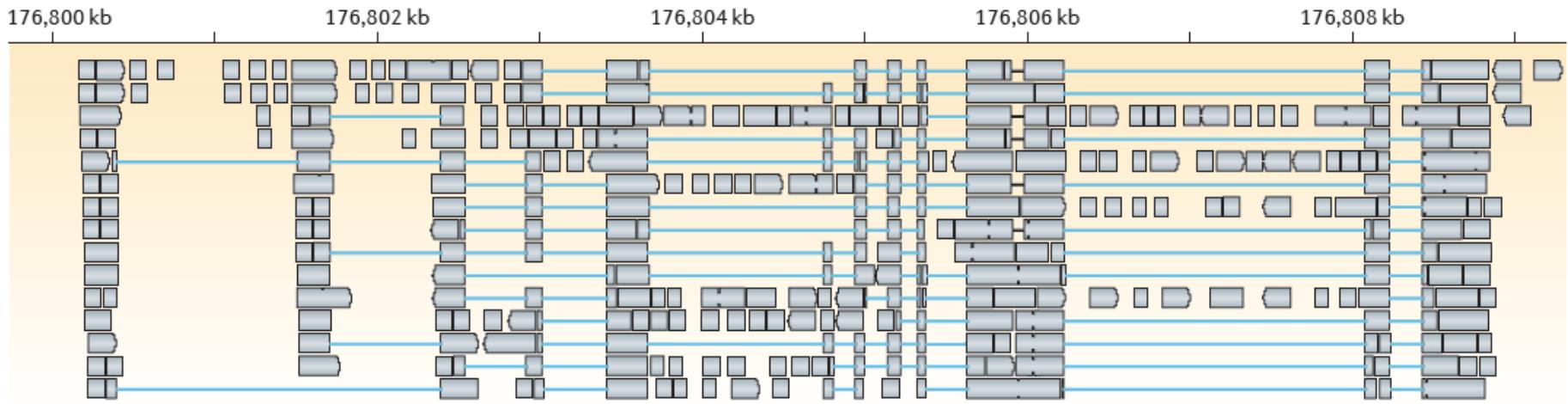
Splice-align reads to the genome



Overlapping but different introns = evidence of alternative splicing

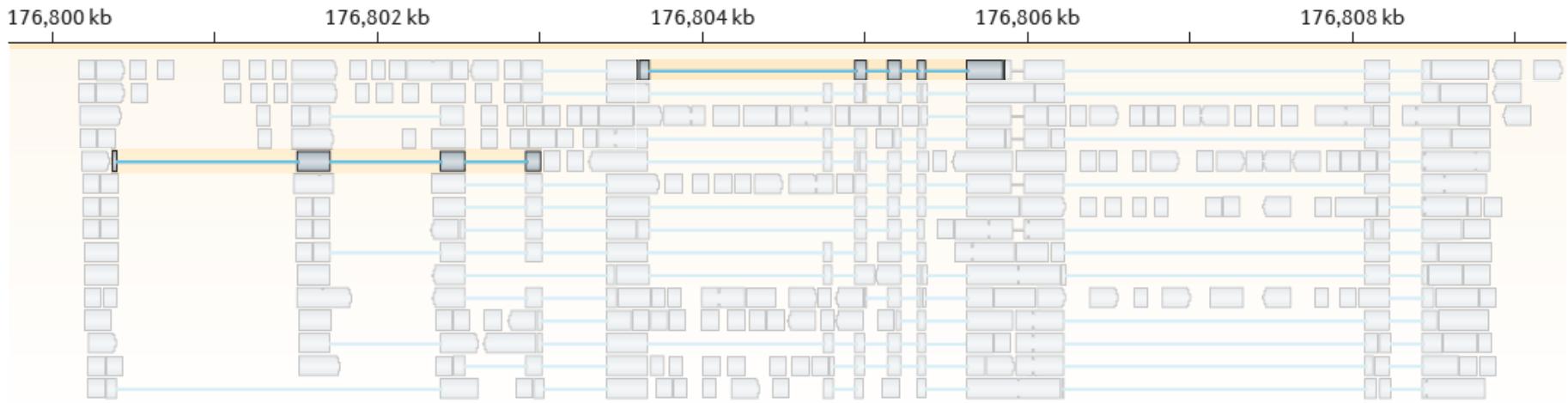
Genome-Guided Transcript Reconstruction

Splice-align reads to the genome



Genome-Guided Transcript Reconstruction

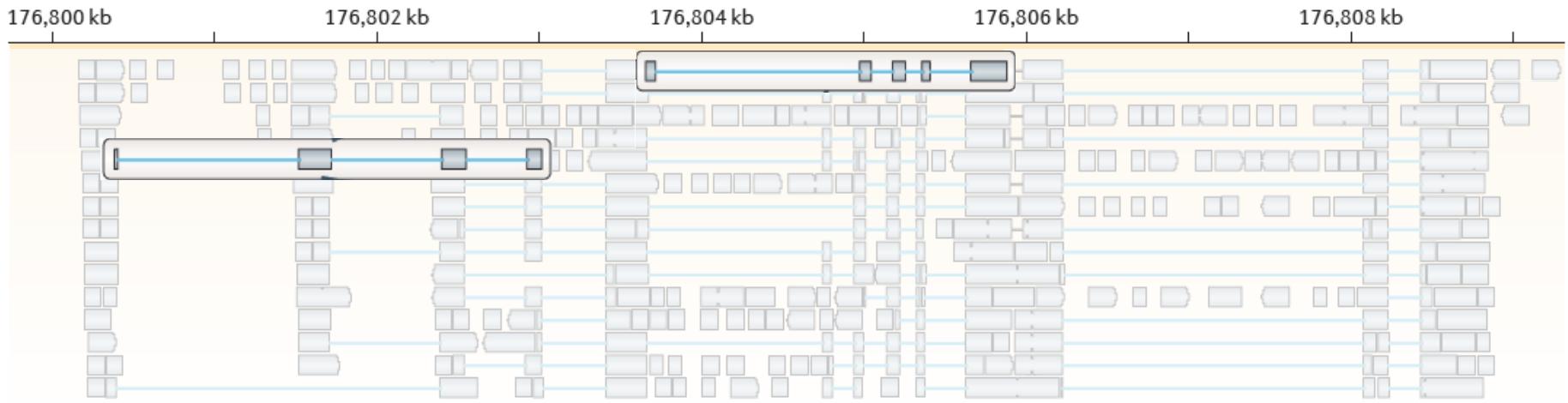
Splice-align reads to the genome



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

Genome-Guided Transcript Reconstruction

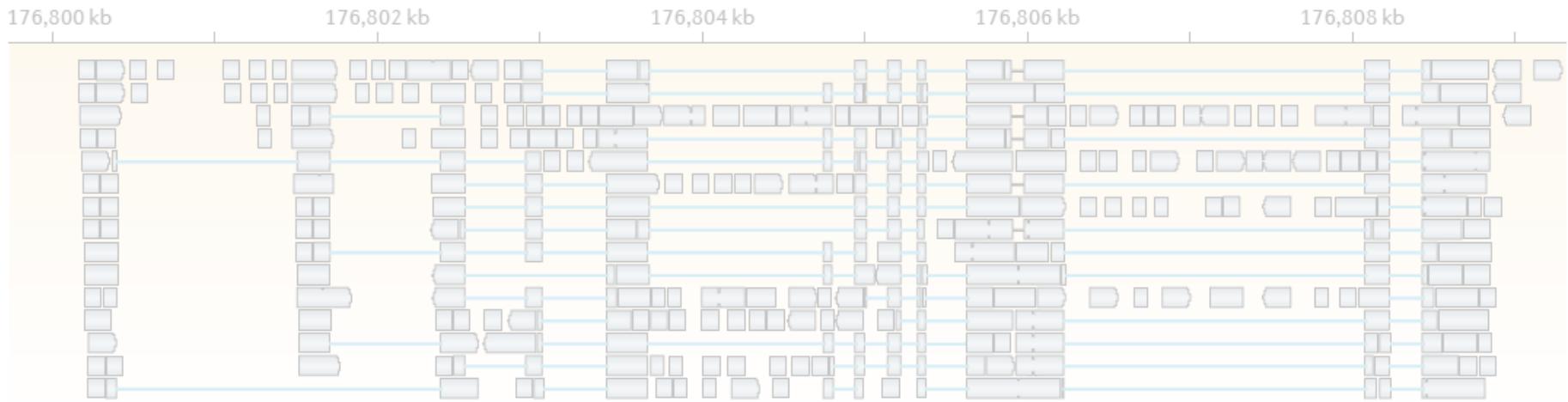
Splice-align reads to the genome



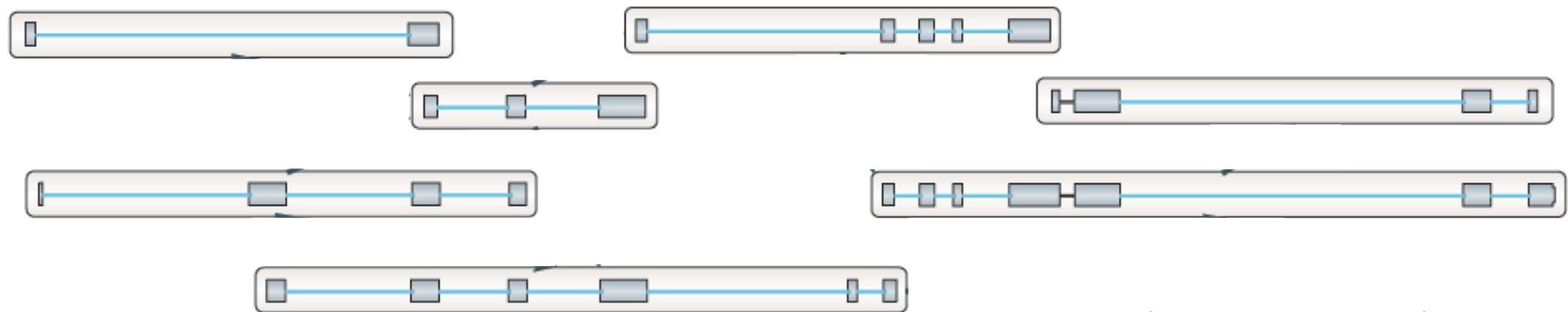
Nodes = unique splice patterns

Genome-Guided Transcript Reconstruction

Splice-align reads to the genome



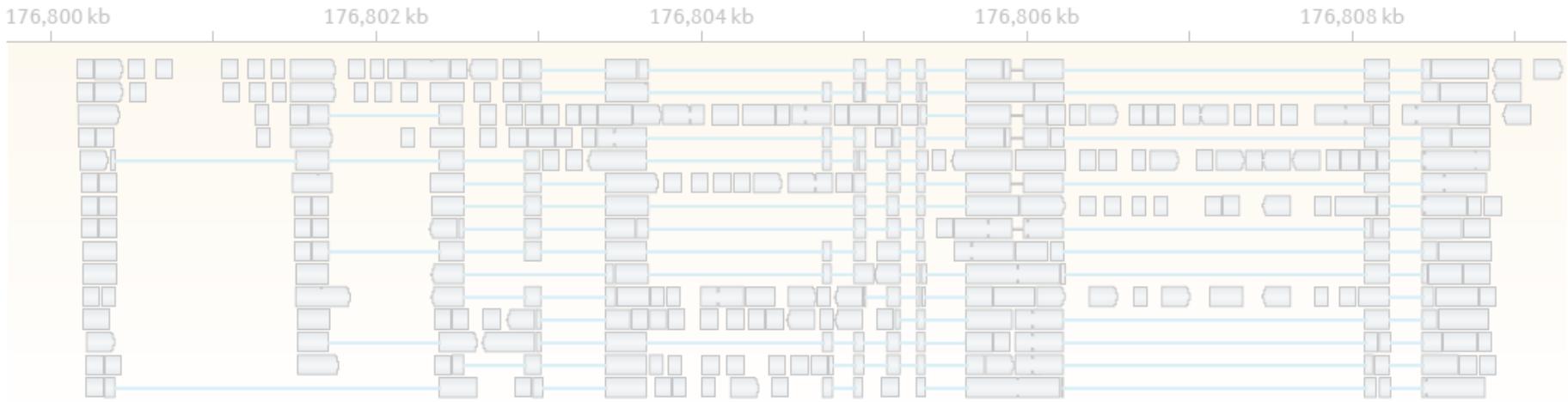
Construct graph from unique splice patterns of aligned reads.



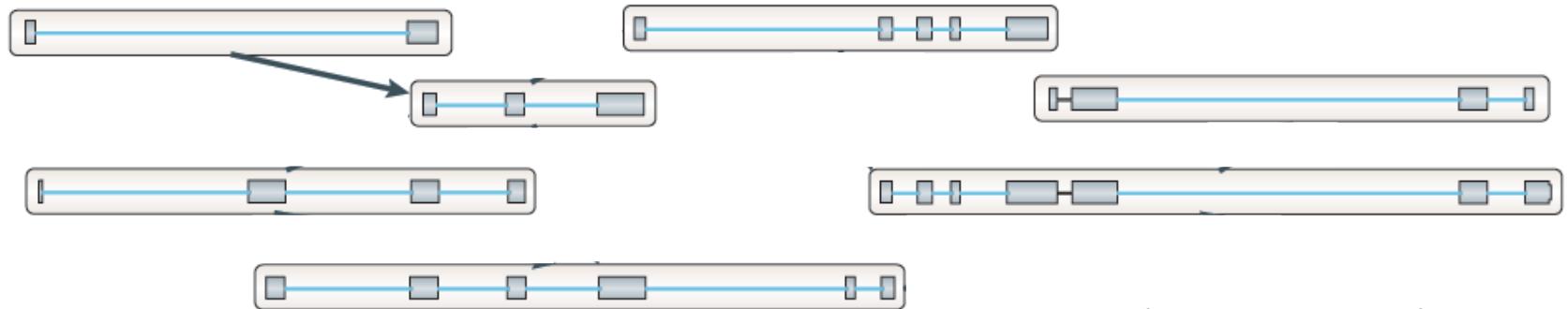
Nodes = unique splice patterns

Genome-Guided Transcript Reconstruction

Splice-align reads to the genome



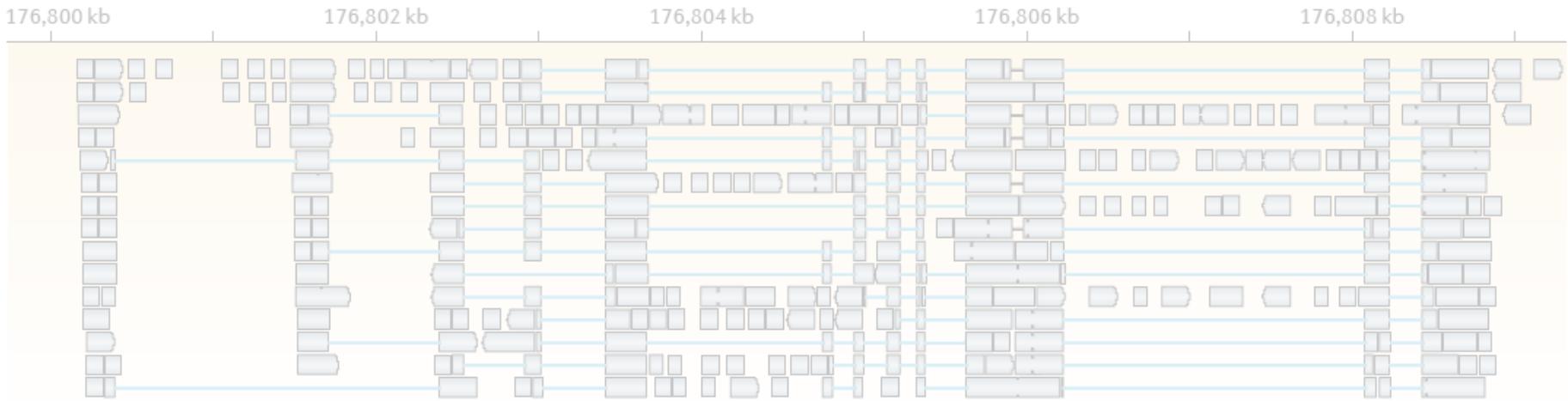
Construct graph from unique splice patterns of aligned reads.



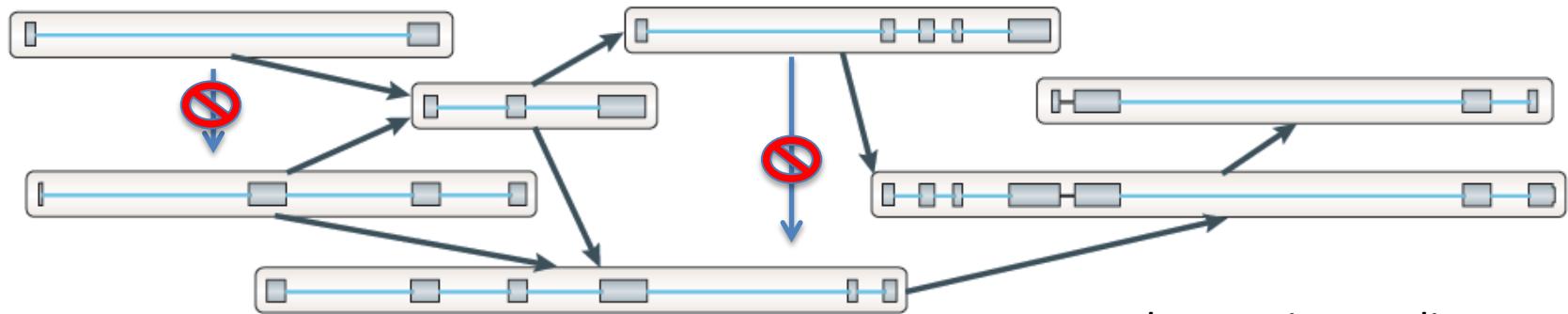
Nodes = unique splice patterns
Edges = compatible patterns

Genome-Guided Transcript Reconstruction

Splice-align reads to the genome

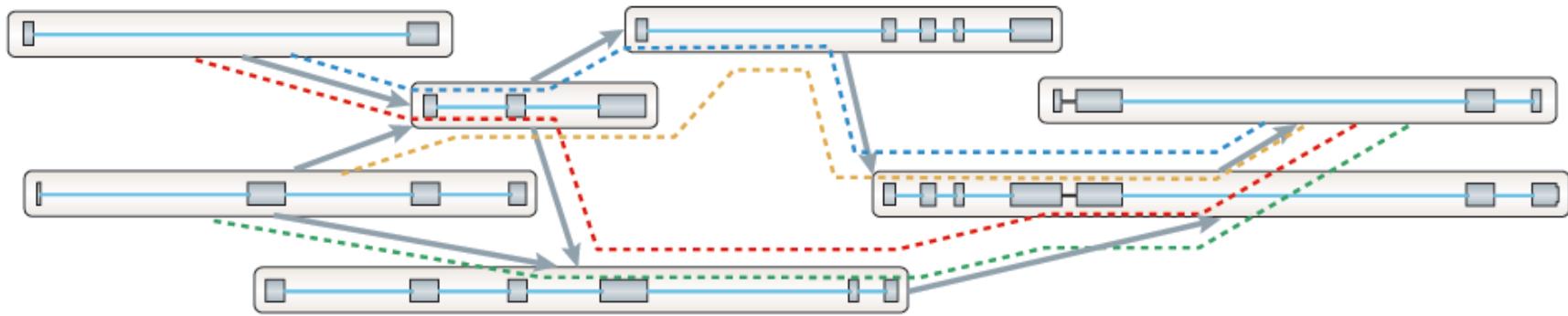


Construct graph from unique splice patterns of aligned reads.



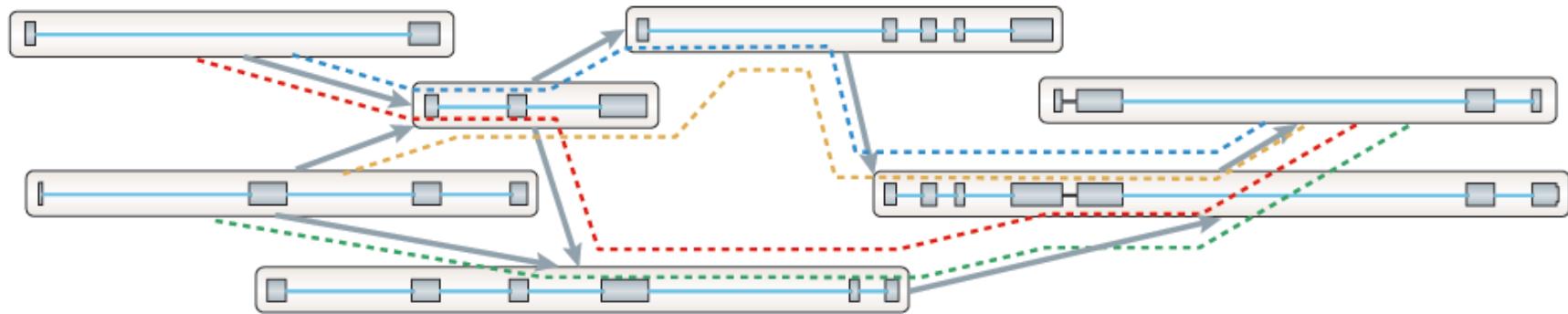
Genome-Guided Transcript Reconstruction

Traverse paths through the graph to assemble transcript isoforms

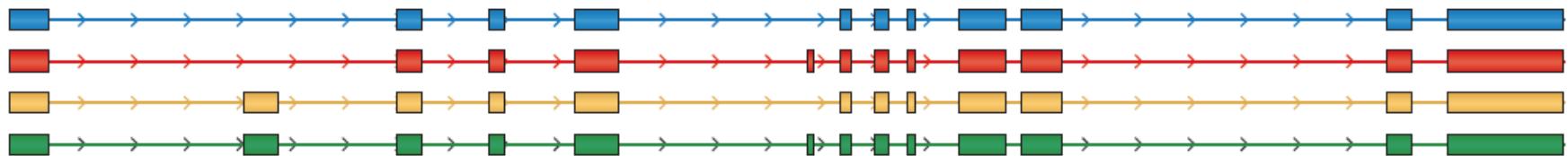


Genome-Guided Transcript Reconstruction

Traverse paths through the graph to assemble transcript isoforms



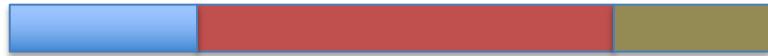
Reconstructed isoforms



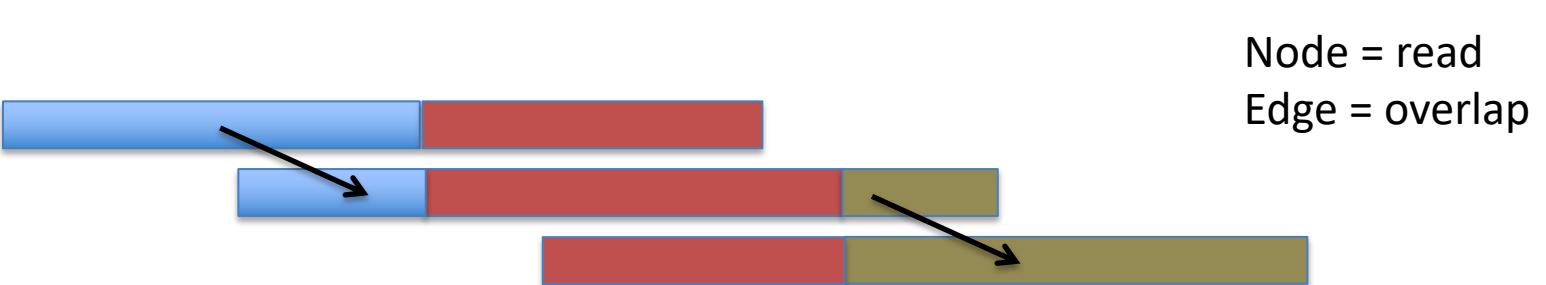
What if you don't have a high quality reference genome sequence?

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



Read Overlap Graph: Reads as nodes, overlaps as edges

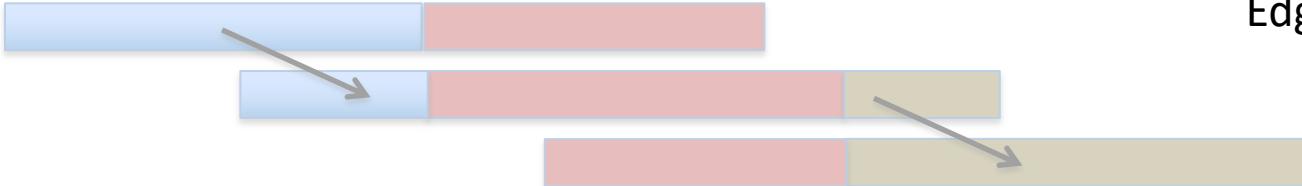


Transcript A



Generate consensus sequence where reads overlap

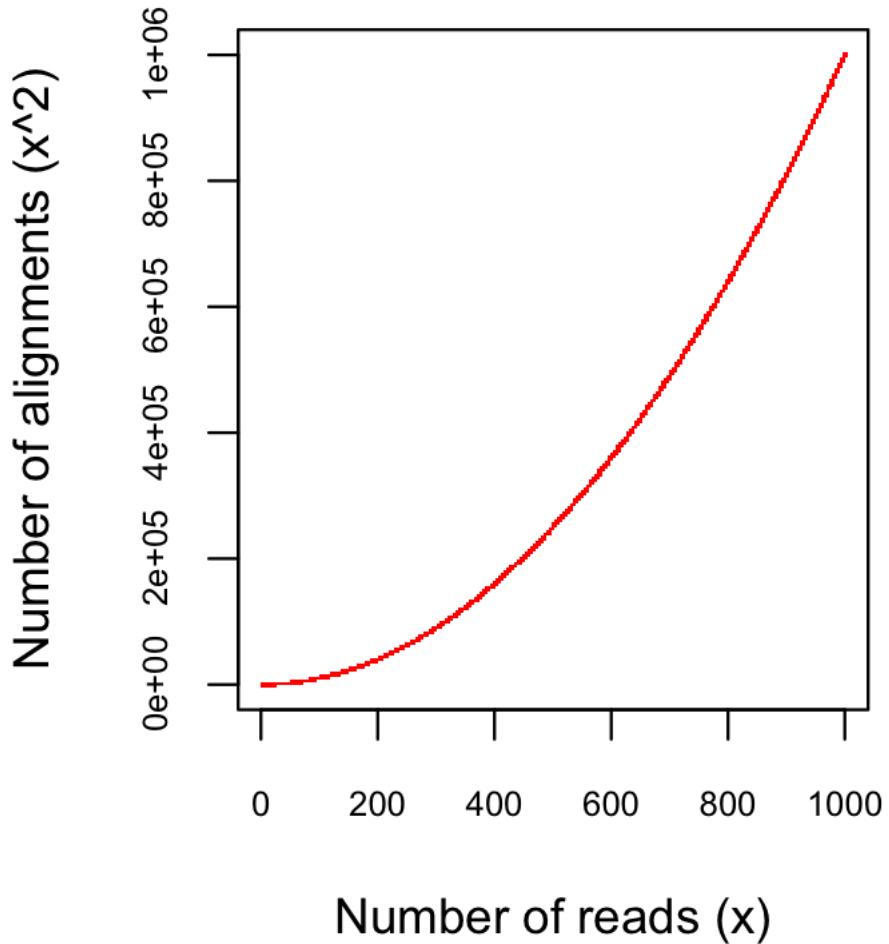
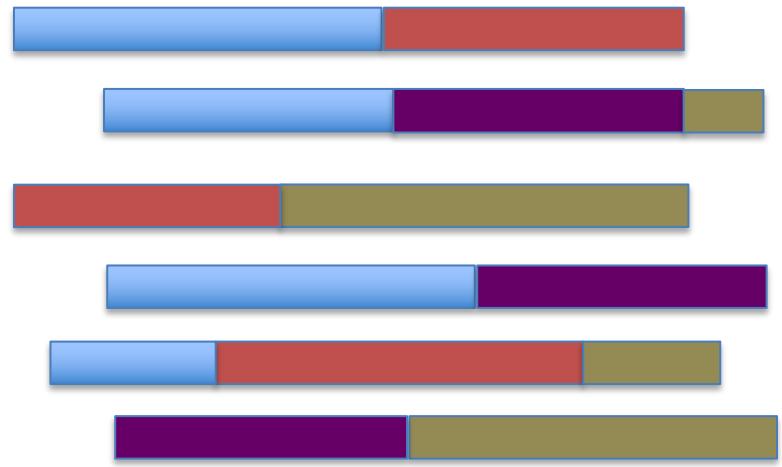
Node = read
Edge = overlap



Transcript B



Finding pairwise overlaps between n reads involves $\sim n^2$ 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

Have you learned about the de Bruijn graph already?

Sequence Assembly via de Bruijn Graphs

Generate all substrings of length k from the reads



Sequence Assembly via De Bruijn Graphs

Generate all substrings of length k from the reads



Sequence Assembly via De Bruijn Graphs

Generate all substrings of length k from the reads



Construct the de Bruijn graph



Nodes = unique k-mers

Sequence Assembly via De Bruijn Graphs

Generate all substrings of length k from the reads



Construct the de Bruijn graph



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



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



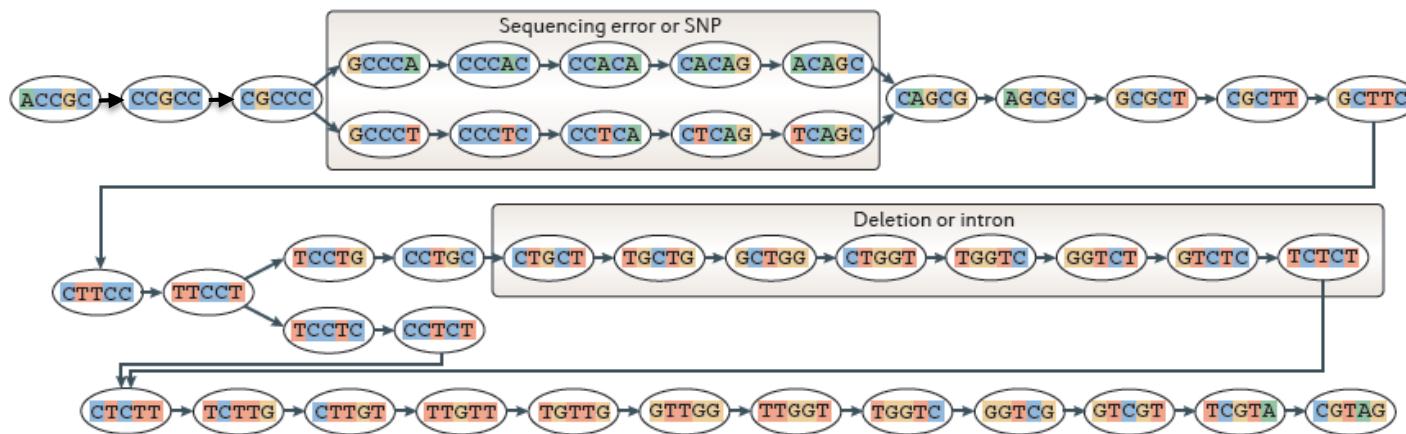
Nodes = unique k-mers
Edges = overlap by (k-1)

Sequence Assembly via De Bruijn Graphs

Generate all substrings of length k from the reads

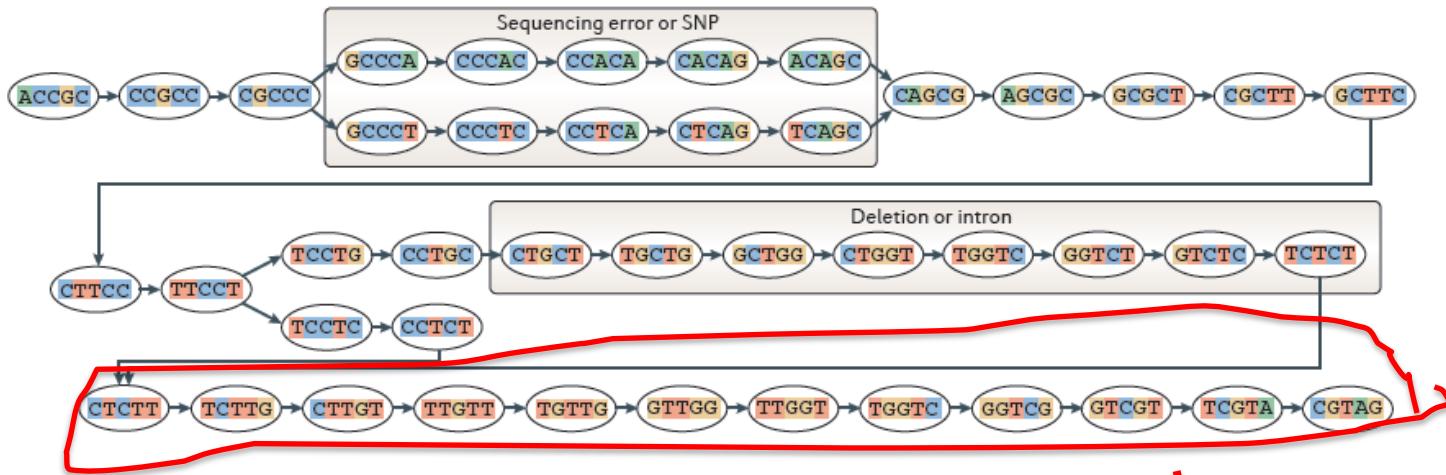
ACAGC	TCCTG	GTCTC		AGCGC	CTCTT	GGTCG	k-mers (k=5)
CACAG	TTCCT	GGTCT		CAGCG	CCTCT	TGGTC	
CCACA	CTTCC	TGGTC	TGTTG	TCAGC	TCCTC	TTGGT	
CCCAC	GCTTC	CTGGT	TTGTT	CTCAG	TTCCT	GTTGG	
GCCCA	CGCTT	GCTGG	CTTGT	CCTCA	CTTCC	TGTTG	
CGCCC	GCGCT	TGCTG	TCTTG	CCCTC	GCTTC	TTGTT	
CCGCC	AGCGC	CTGCT	CTCTT	GCCCT	CGCTT	CTTGT	
ACCGC	CAGCG	CCTGC	TCTCT	CGCCC	GCGCT	TCTTG	
ACCGCCCCACAGCGCTTCCTGCTGGTCTCTTGTG				CGCCCTCAGCGCTTCCTCTTGTGGTCGTAG			
							Reads

Construct the de Bruijn graph

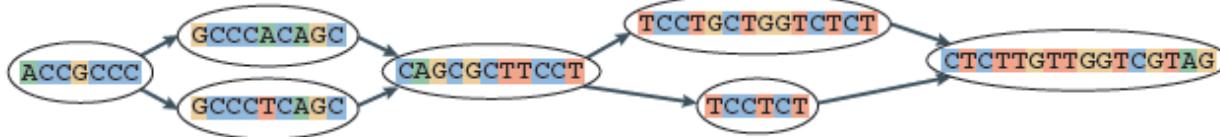


Nodes = unique k-mers
Edges = overlap by (k-1)

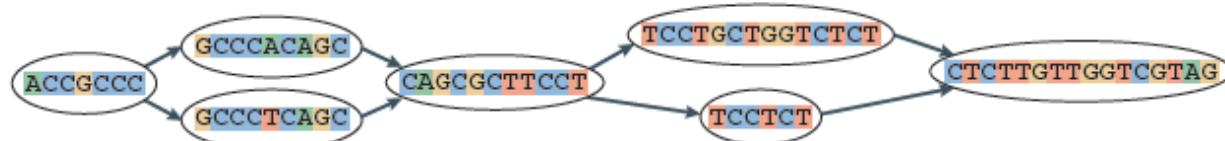
Construct the de Bruijn graph



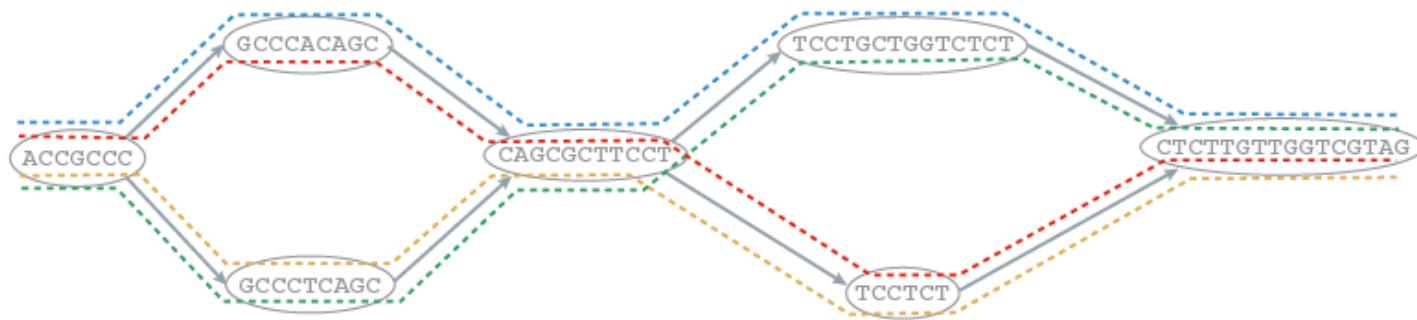
Collapse the de Bruijn graph



Collapse the de Bruijn graph



Traverse the graph



Assemble Transcript Isoforms

Transcript Isoforms:

- Blue dashed line: ACCGGCCACAGCGCTTCCTGCTGGTCTCTTGGTGGT CGTAG
- Red dashed line: ACCGGCCACAGCGCTTCCT-----CTTGGTGGT CGTAG
- Yellow dashed line: ACCGGCCCTCAGCGCTTCCT-----CTTGGTGGT CGTAG
- Green dashed line: ACCGGCCCTCAGCGCTTCCTGCTGGTCTCTTGGTGGT CGTAG

Part 3. Trinity De novo Assembly



Contrasting Genome and Transcriptome Assembly

Genome Assembly

- Uniform coverage
- Single contig per locus
- Double-stranded

Transcriptome Assembly

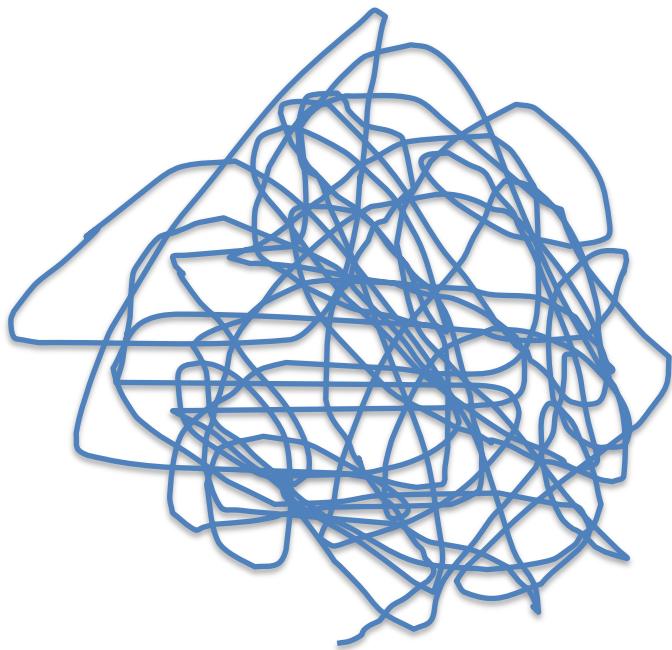
- Exponentially distributed coverage levels
- Multiple contigs per locus (alt splicing)
- Strand-specific



Trinity Aggregates Isolated Transcript Graphs

Genome Assembly

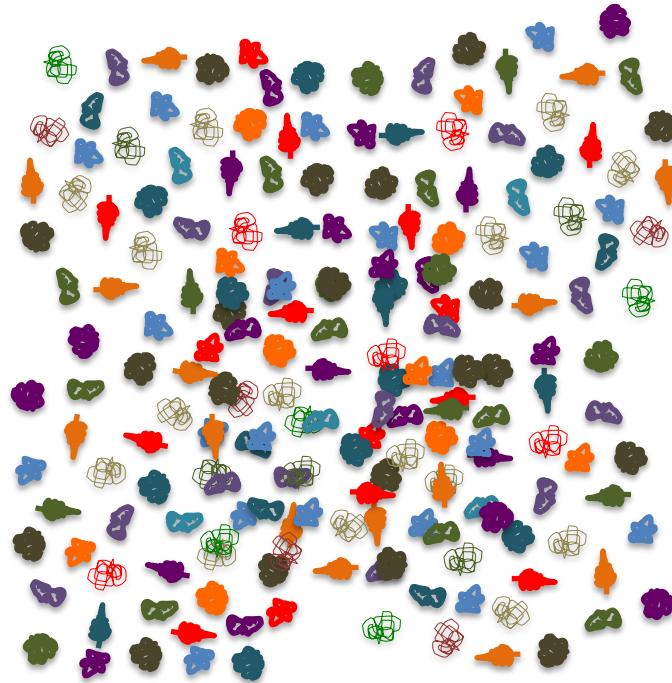
Single Massive Graph



Entire chromosomes represented.

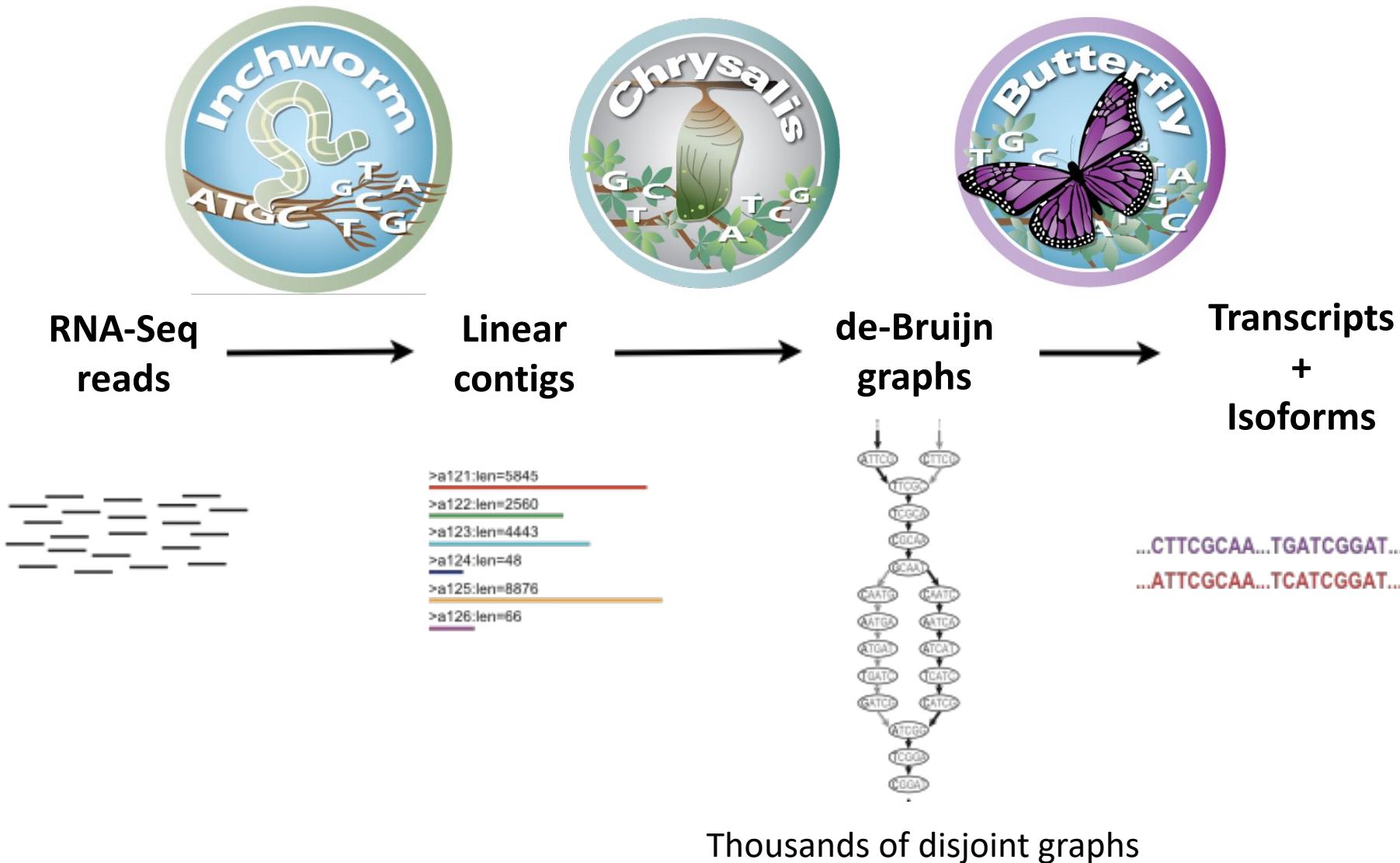
Trinity Transcriptome Assembly

Many Thousands of Small Graphs



Ideally, one graph per expressed gene.

Trinity – How it works:



Trinity – How it works:

Younger
me



Manfred
Grabherr



Moran
Yassour

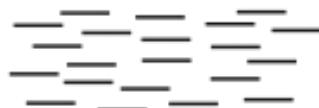


RNA-Seq
reads

Linear
contigs

de-Bruijn
graphs

Transcripts
+
Isoforms



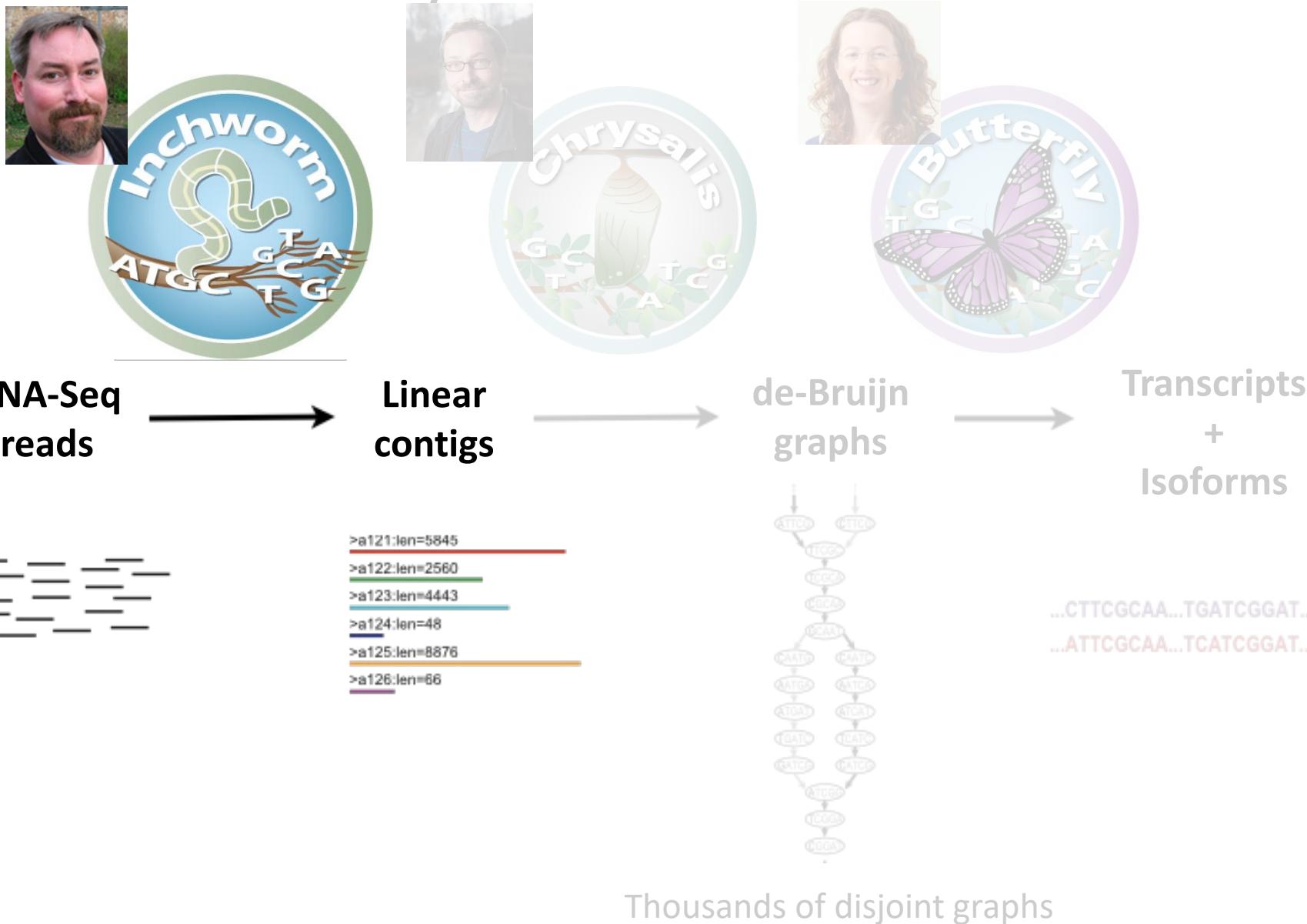
>a121:len=5845
>a122:len=2560
>a123:len=4443
>a124:len=48
>a125:len=8876
>a126:len=66



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

Thousands of disjoint graphs

Trinity – How it works:





Inchworm Algorithm

- Decompose all reads into overlapping Kmers => hashtable(kmer, count)

Read: **AATGTGAAACTGGATTACATGCTGGTATGTC...**

AATGTGA

ATGTGAA

Overlapping kmers of length (k)

TGTGAAA

...

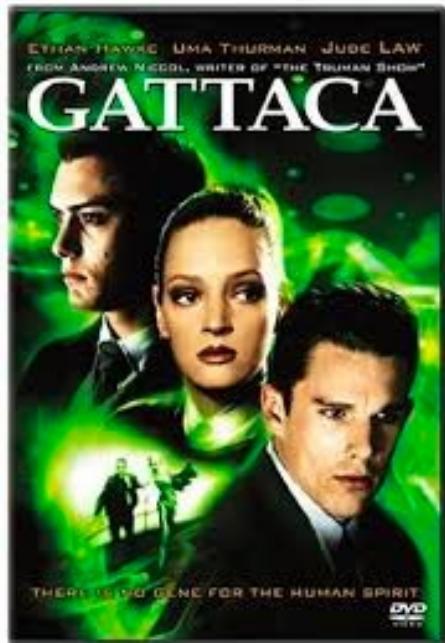
Kmer Catalog (hashtable)

Kmer	Count among all reads
AATGTGA	4
ATGTGAA	2
TGTGAAA	1
GATTACA	9



Inchworm Algorithm

- Decompose all reads into overlapping Kmers => hashtable(kmer, count)
- Identify seed kmer as most abundant Kmer, ignoring low-complexity kmers.



<https://en.wikipedia.org/wiki/Gattaca>

GATTACA
9

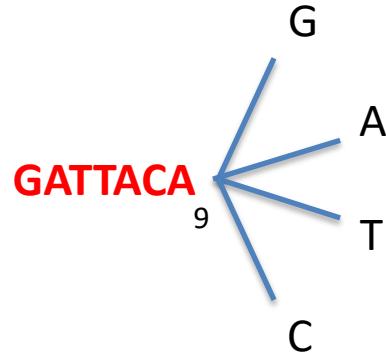
Kmer Catalog (hashtable)

Kmer	Count among all reads
AATGTGA	4
ATGTGAA	2
TGTGAAA	1
GATTACA	9



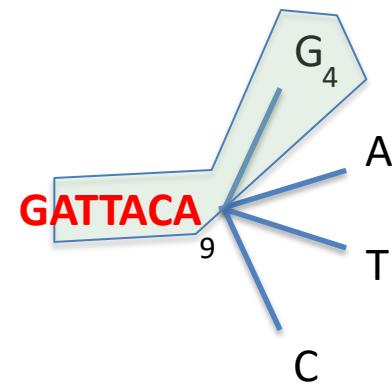
Inchworm Algorithm

- Decompose all reads into overlapping Kmers => hashtable(kmer, count)
- Identify seed kmer as most abundant Kmer, ignoring low-complexity kmers.
- Extend kmer at 3' end, guided by coverage.



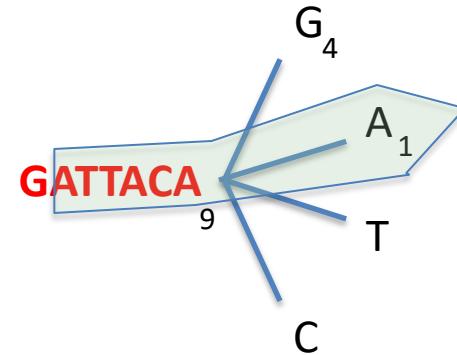


Inchworm Algorithm



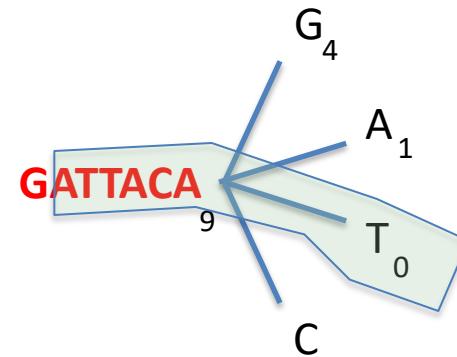


Inchworm Algorithm



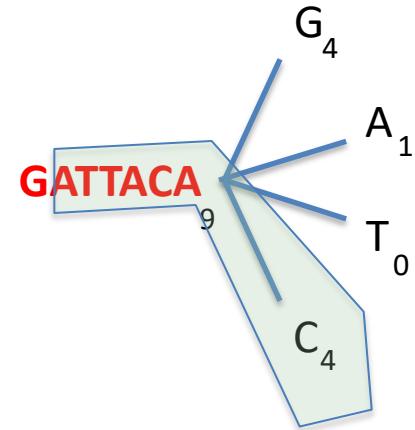


Inchworm Algorithm



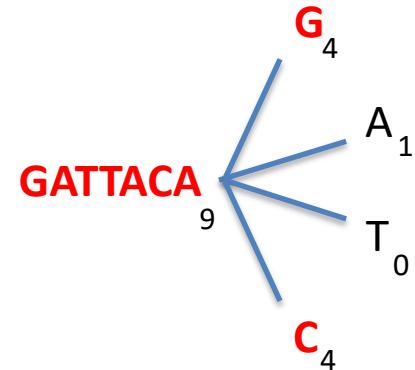


Inchworm Algorithm



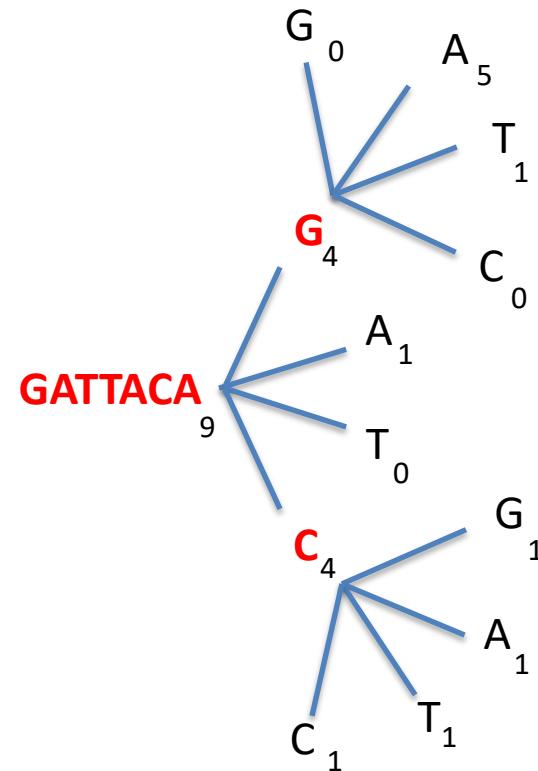


Inchworm Algorithm



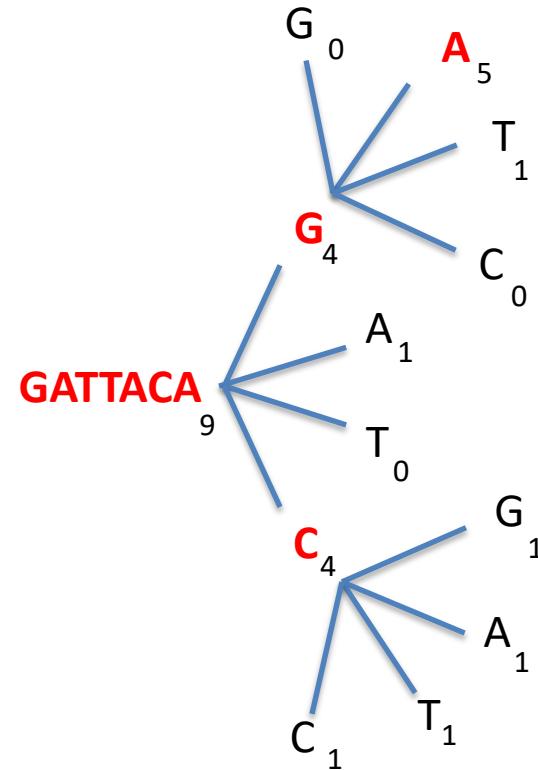


Inchworm Algorithm



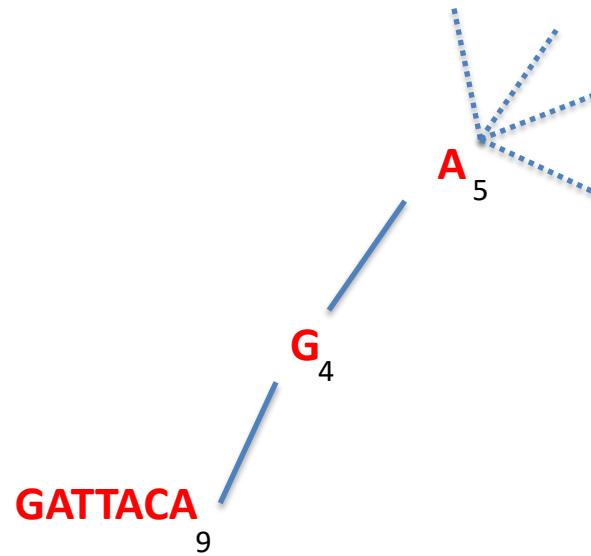


Inchworm Algorithm



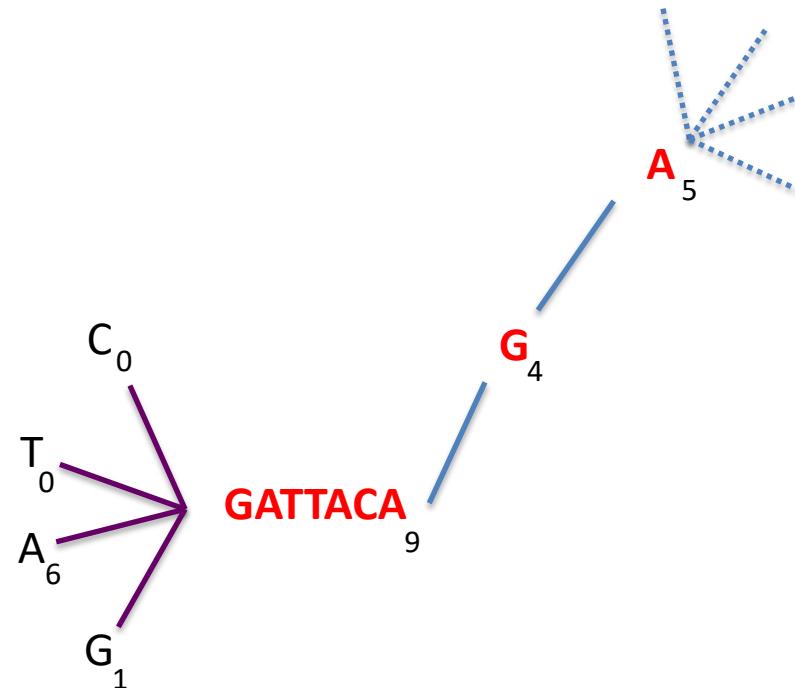


Inchworm Algorithm



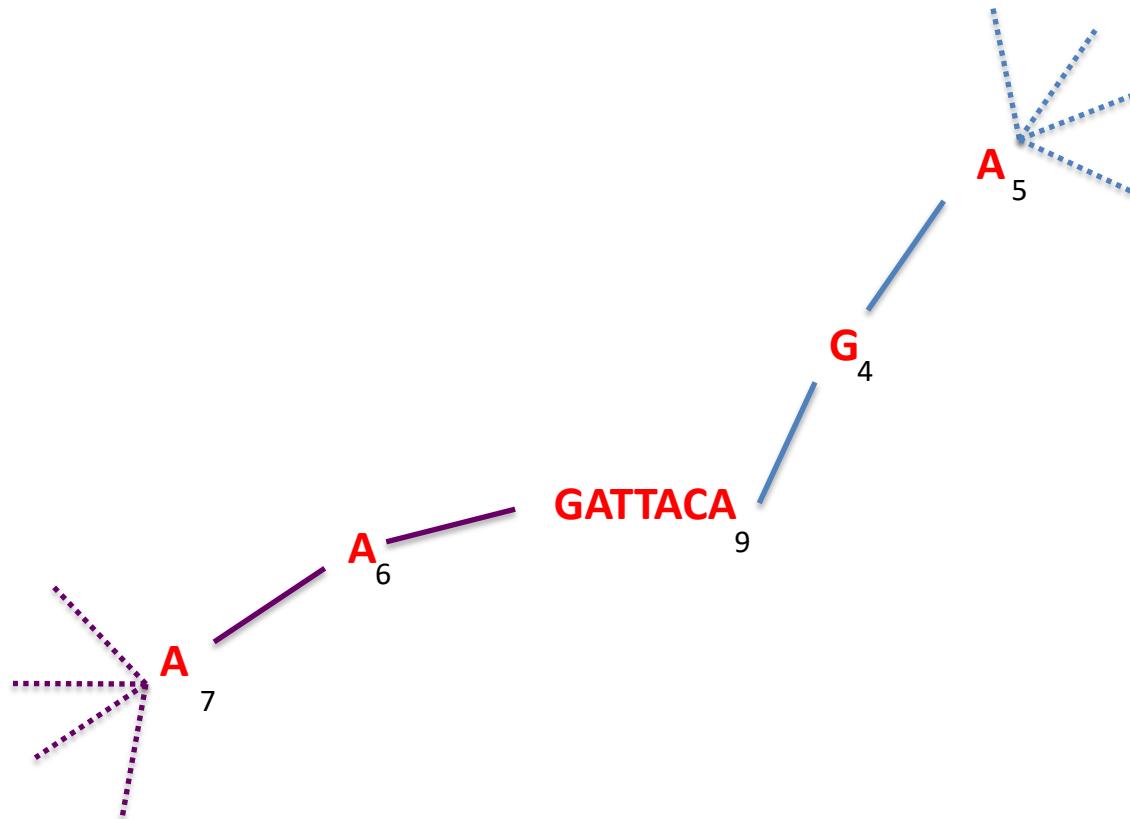


Inchworm Algorithm





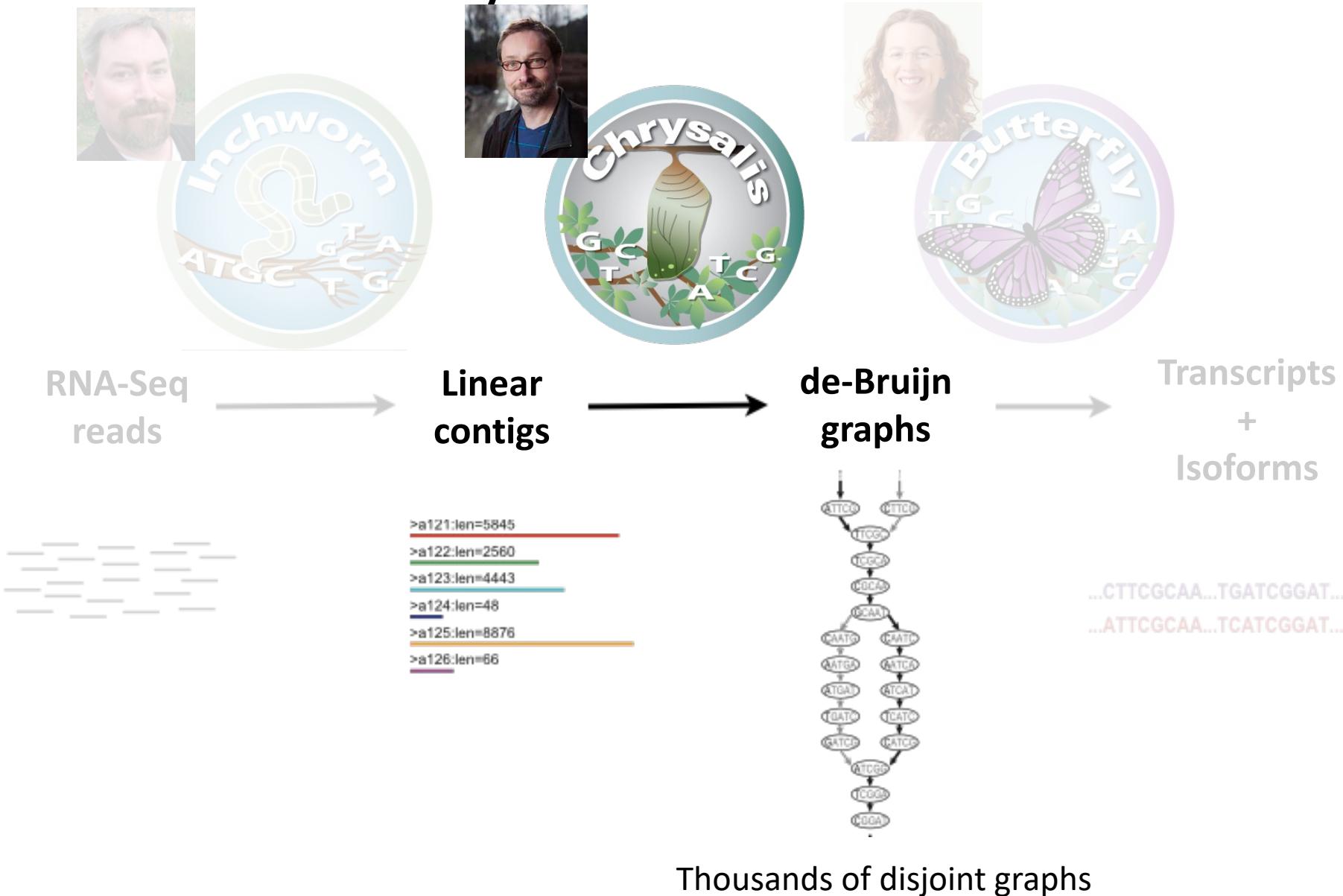
Inchworm Algorithm



Report contig:**AAGATTACAGA**....

Remove assembled kmers from catalog, then repeat the entire process.

Trinity – How it works:



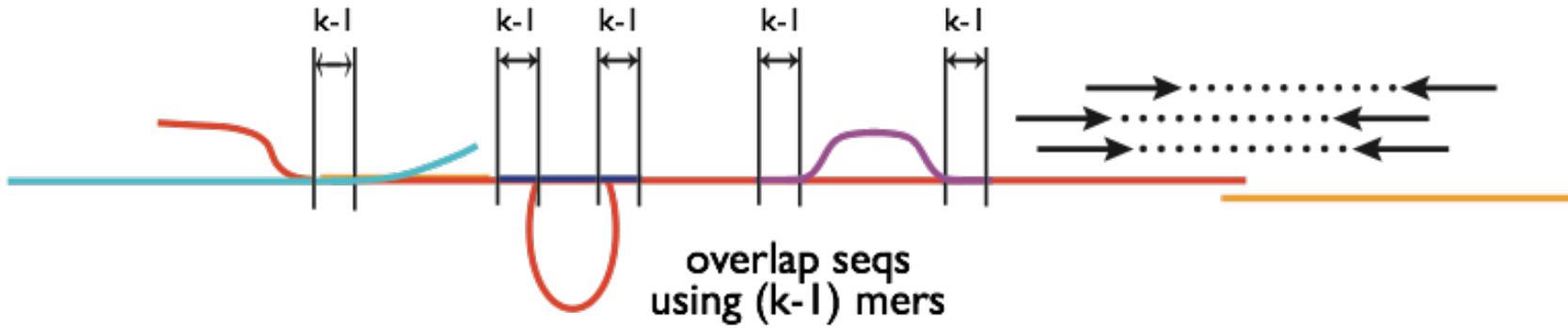
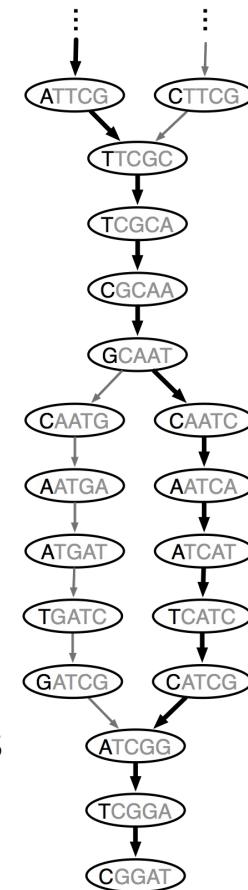
Chrysalis

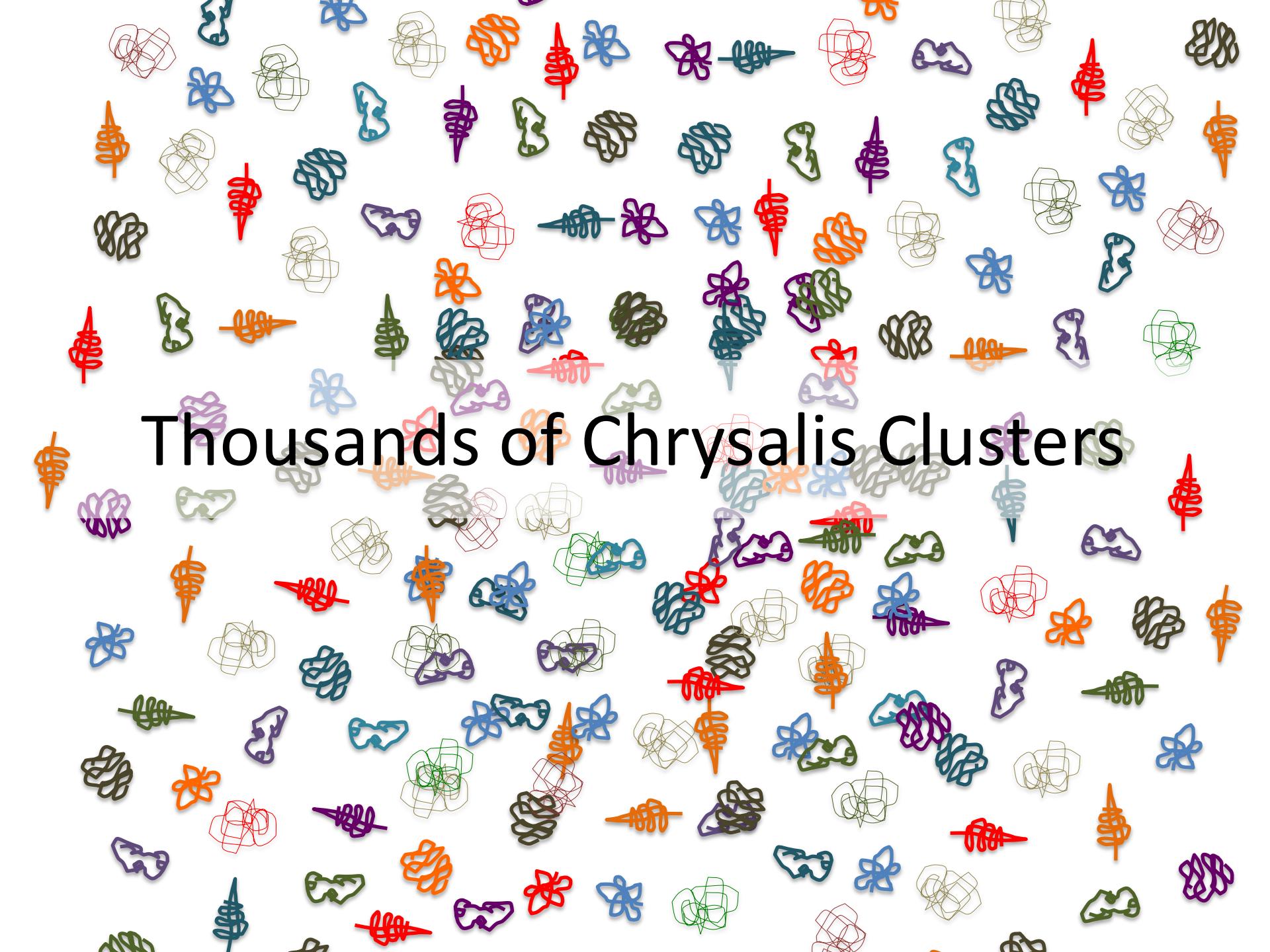
>a121:len=5845
 |
>a122:len=2560
 |
>a123:len=4443
 |
>a124:len=48
 |
>a125:len=8876
 |
>a126:len=66

Integrate isoforms via k-1 overlaps



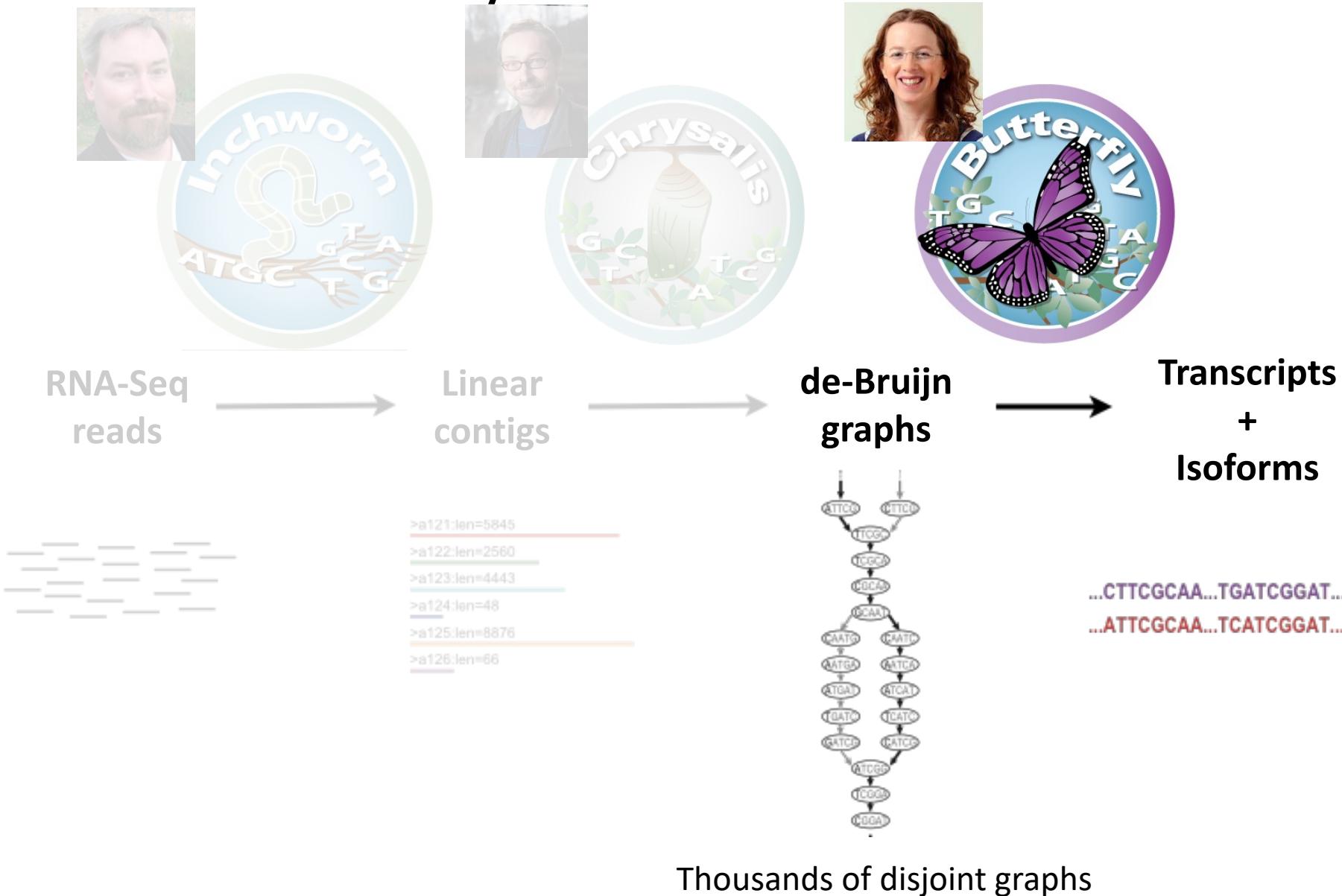
Build de Bruijn Graphs (ideally, one per gene)

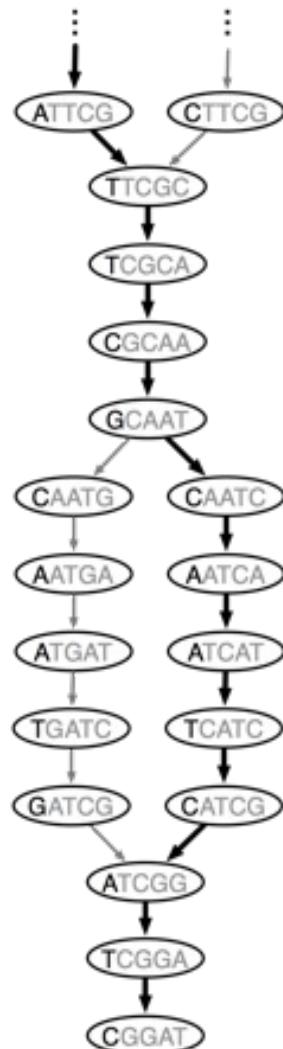




Thousands of Chrysalis Clusters

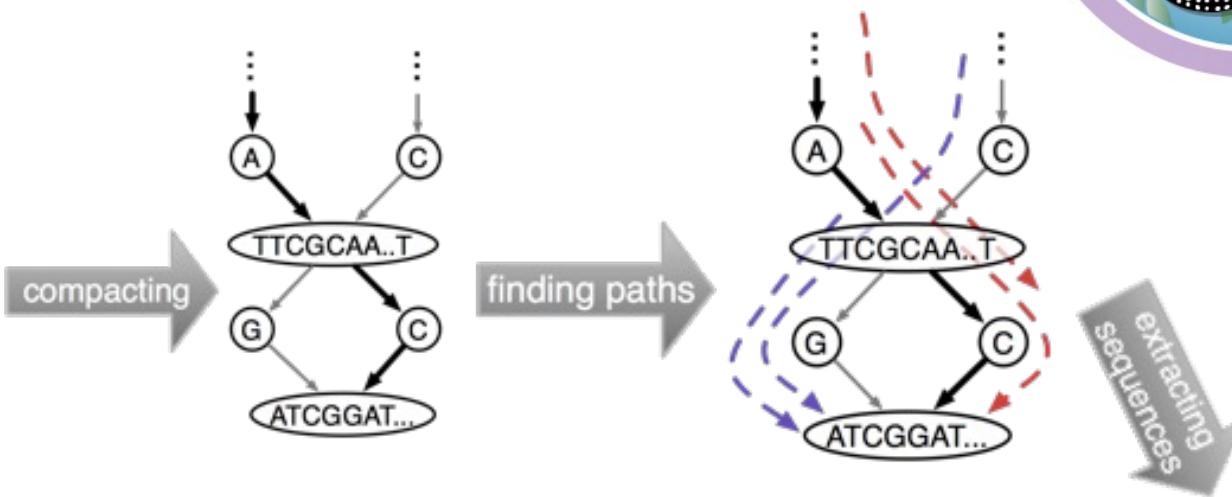
Trinity – How it works:





de Bruijn
graph

Butterfly



compact
graph

compact
graph with
reads

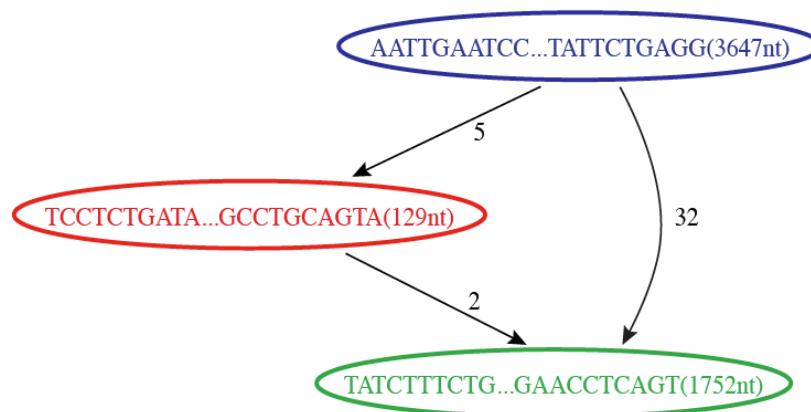
sequences
(isoforms and paralogs)



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

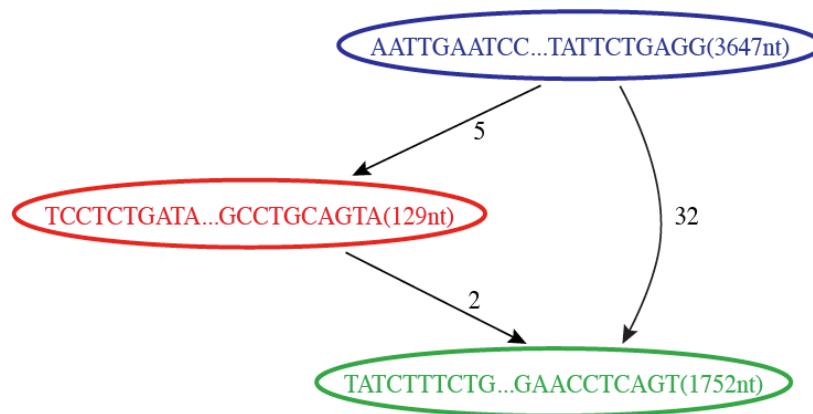
Butterfly Example 1: Reconstruction of Alternatively Spliced Transcripts

Butterfly's Compacted
Sequence Graph



Reconstruction of Alternatively Spliced Transcripts

Butterfly's Compacted
Sequence Graph

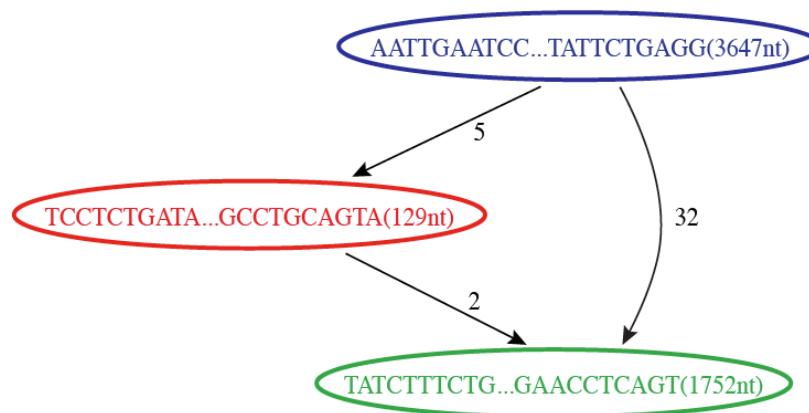


Reconstructed Transcripts



Reconstruction of Alternatively Spliced Transcripts

Butterfly's Compacted
Sequence Graph

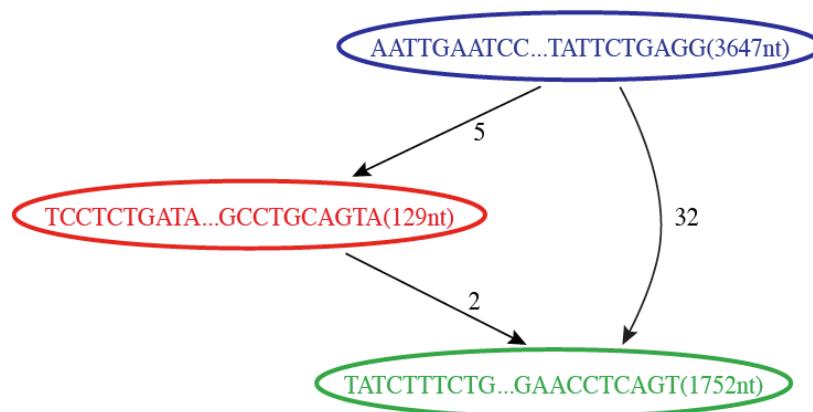


Reconstructed Transcripts



Reconstruction of Alternatively Spliced Transcripts

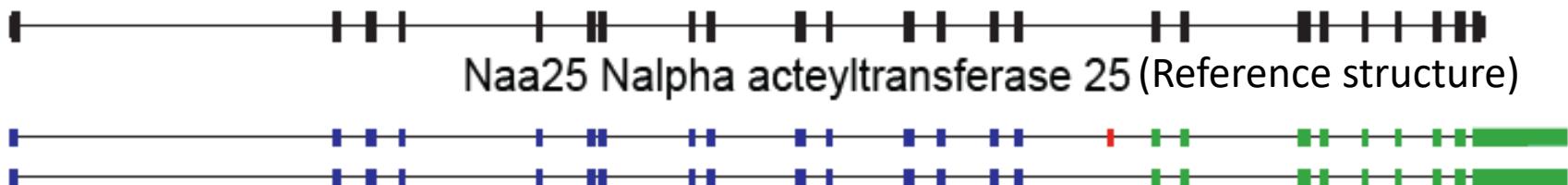
Butterfly's Compacted Sequence Graph



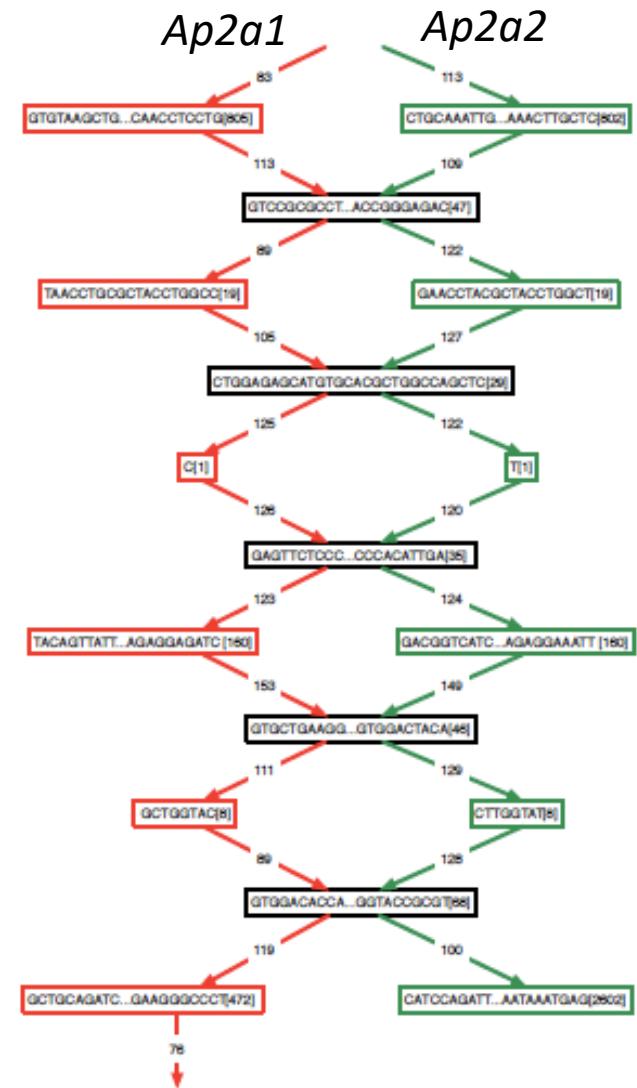
Reconstructed Transcripts



Aligned to Mouse Genome



Butterfly Example 2: Teasing Apart Transcripts of Paralogous Genes



Teasing Apart Transcripts of Paralogous Genes

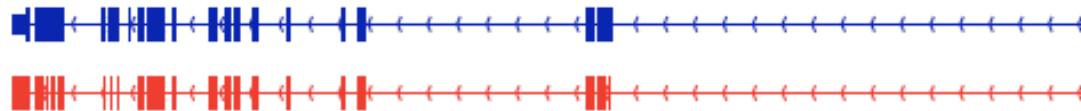
chr7:148,744,197-148,821,437

NM_007459; Ap2a2 adaptor protein complex AP-2, alpha 2 subunit



chr7:52,150,889-52,189,508

NM_001077264; Ap2a1 adaptor protein complex AP-2, alpha 1 subunit



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

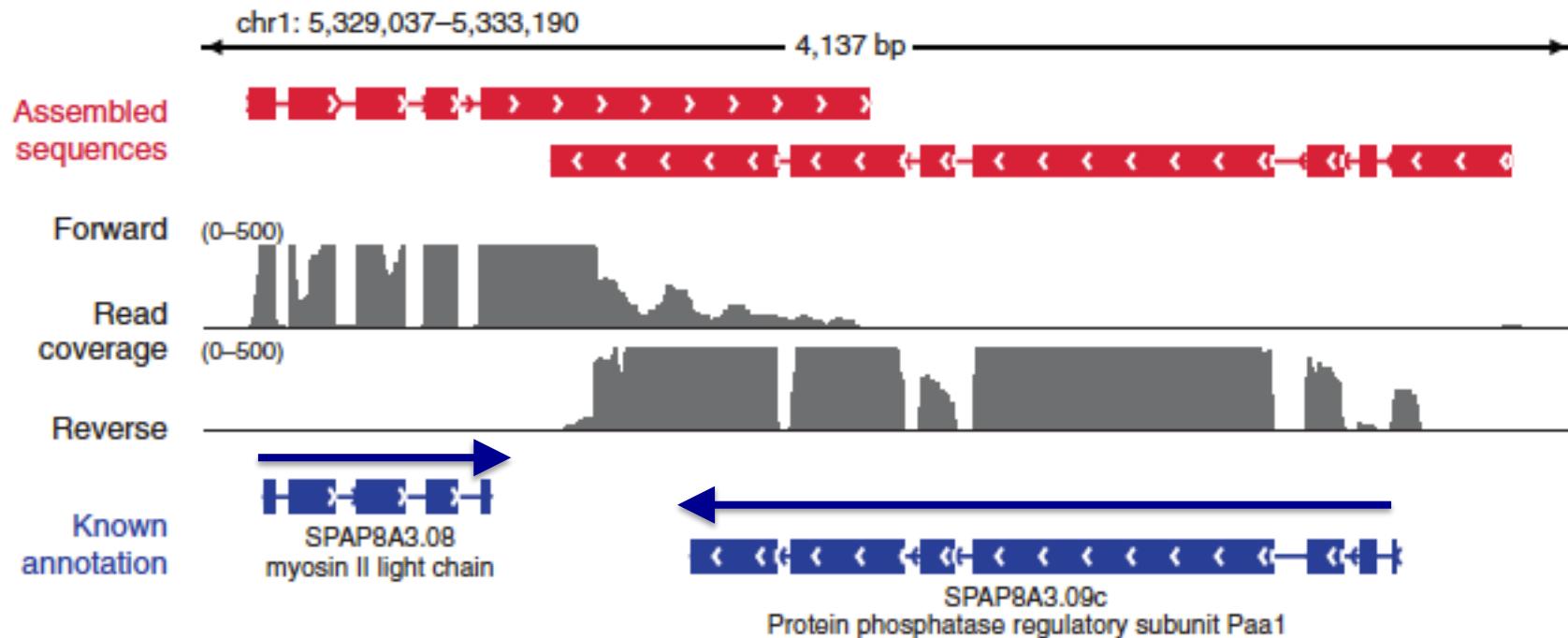
Illumina TruSeq Stranded mRNA Kit:



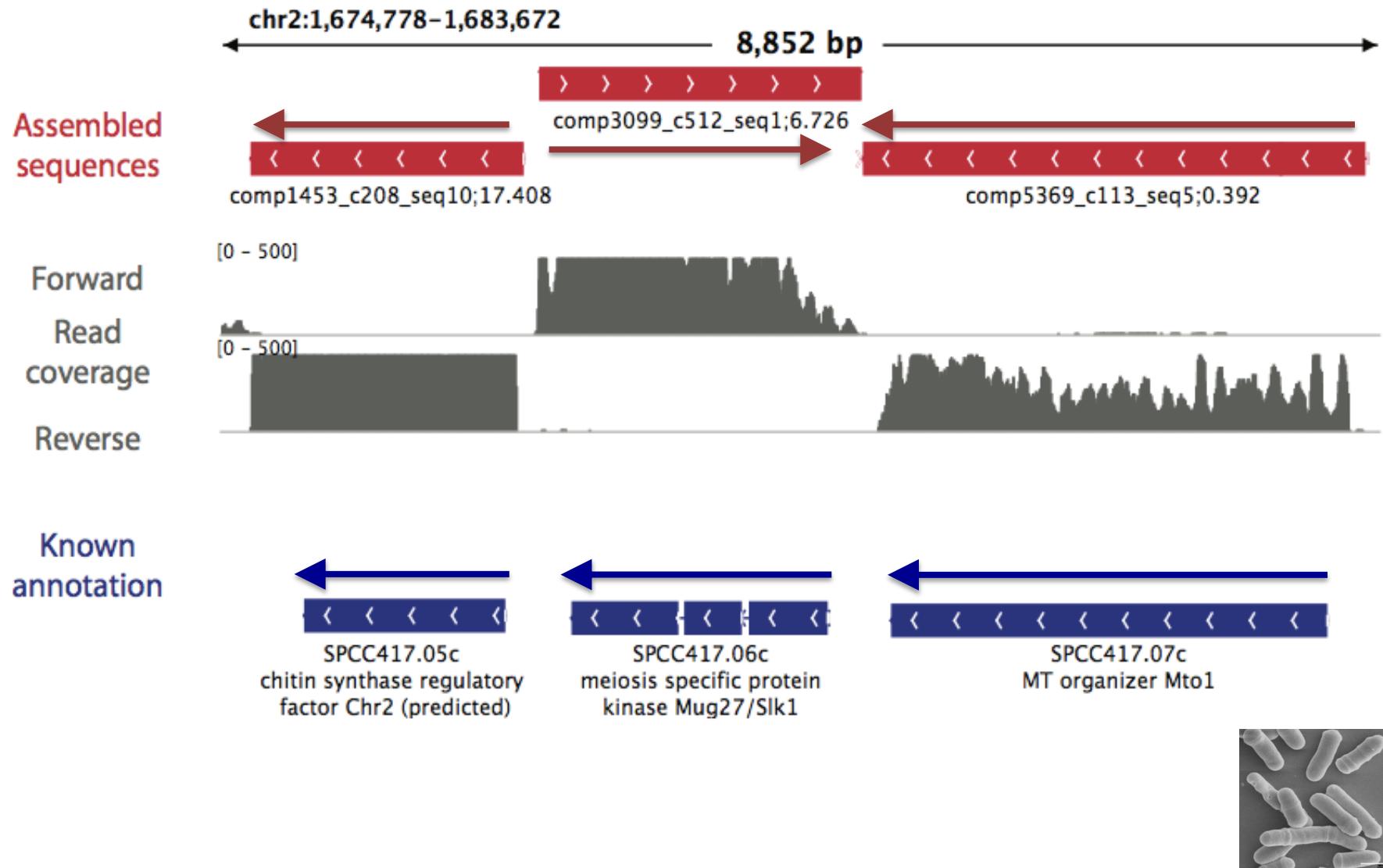
Overlapping UTRs from Opposite Strands



Schizosaccharomyces pombe
(fission yeast)



Antisense-dominated Transcription



Trinity is a Highly Effective and Popular RNA-Seq Assembler



Thousands of routine users.

>15k literature citations

Freely Available, Well-supported,
Open Source Software

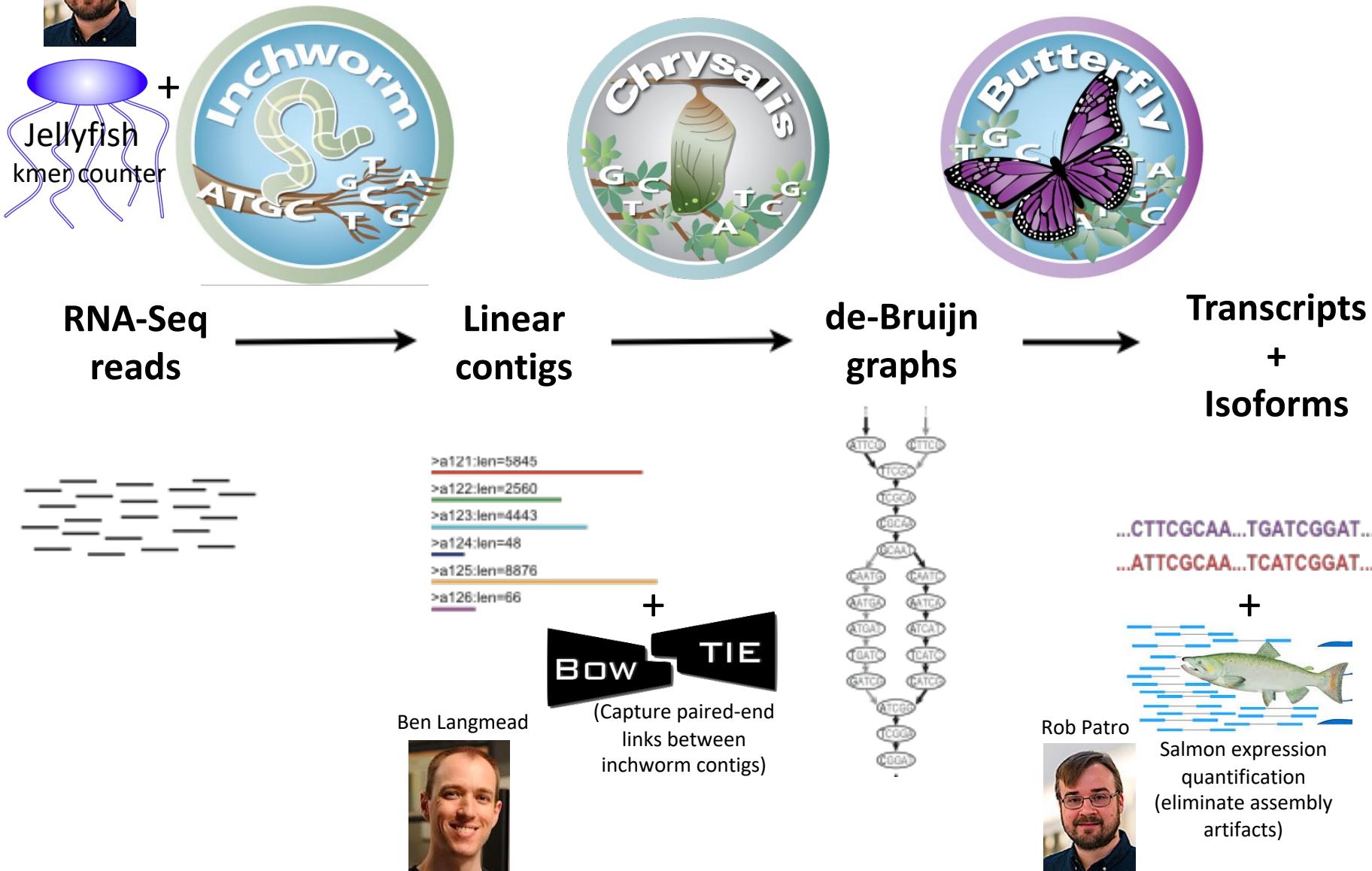


<http://trinityrnaseq.github.io>

Nature Biotechnology, 2011



Trinity – Today, Many More Components (off-the-shelf and into the Trinity ecosystem)



Transcriptome Assembly is Just the End of the Beginning...

NATURE PROTOCOLS | PROTOCOL

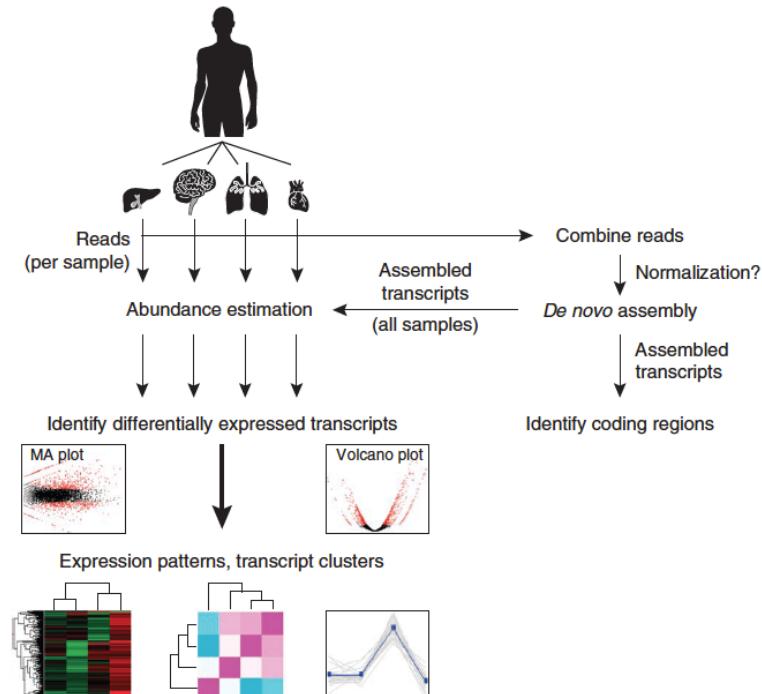
De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis

Brian J Haas, Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, Philip D Blood, Joshua Bowden, Matthew Brian Couger, David Eccles, Bo Li, Matthias Lieber, Matthew D MacManes, Michael Ott, Joshua Orvis, Nathalie Pochet, Francesco Strozzi, Nathan Weeks, Rick Westerman, Thomas William, Colin N Dewey, Robert Henschel, Richard D LeDuc, Nir Friedman & Aviv Regev

[Affiliations](#) | [Contributions](#) | [Corresponding authors](#)

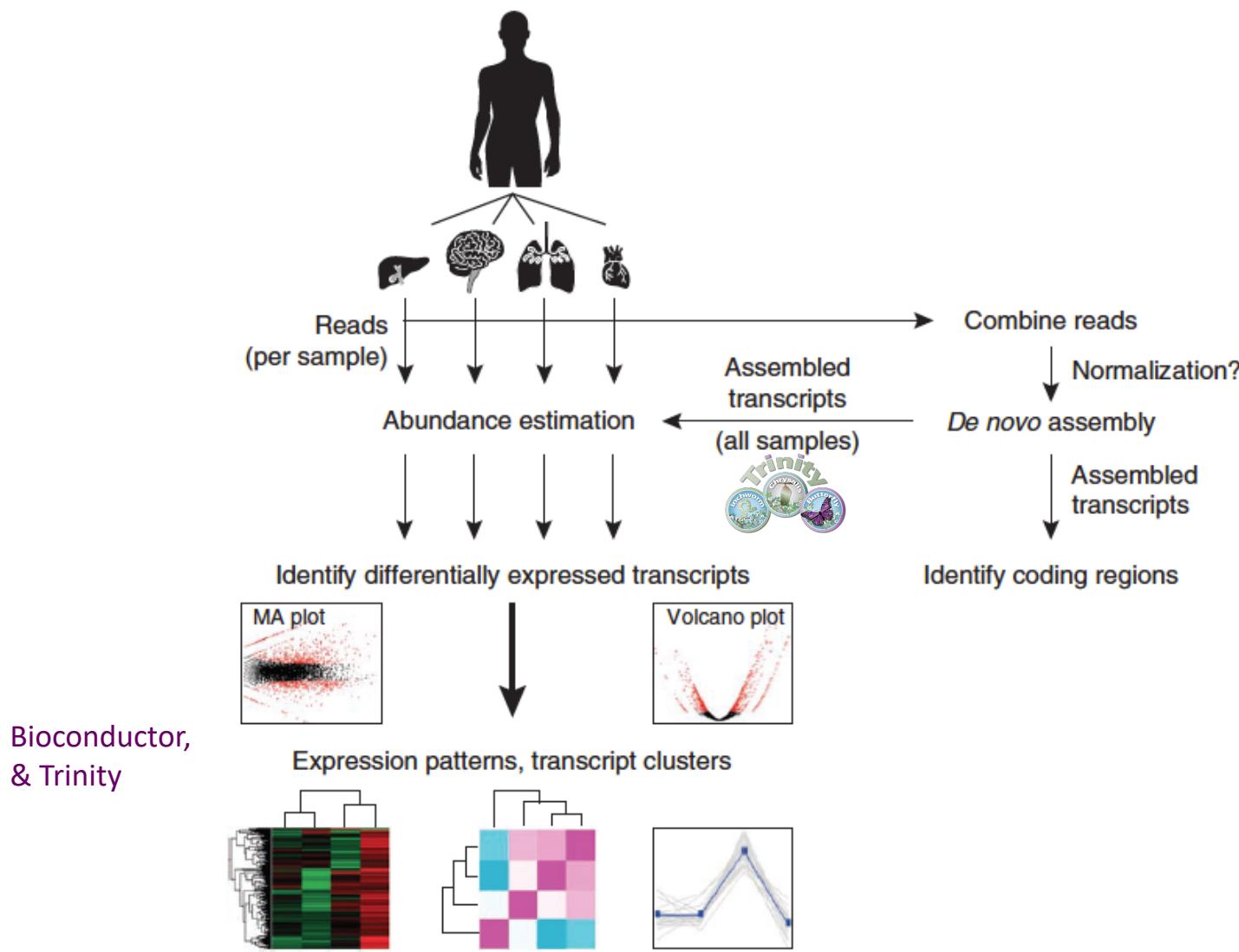
Nature Protocols 8, 1494–1512 (2013) | doi:10.1038/nprot.2013.084

Published online 11 July 2013



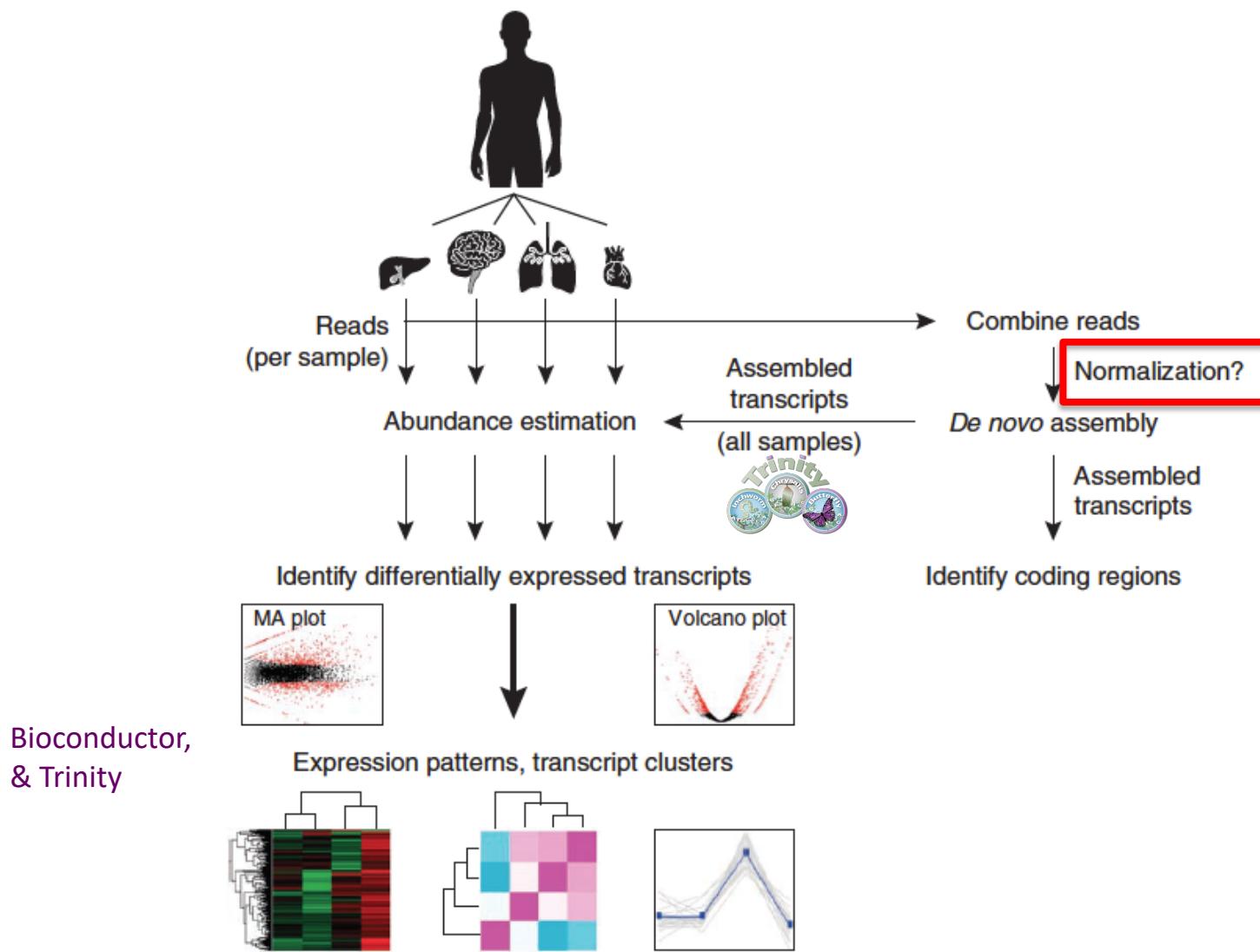
Trinity Framework for De novo Transcriptome Assembly and Analysis

(focus of the transcriptomics lab)

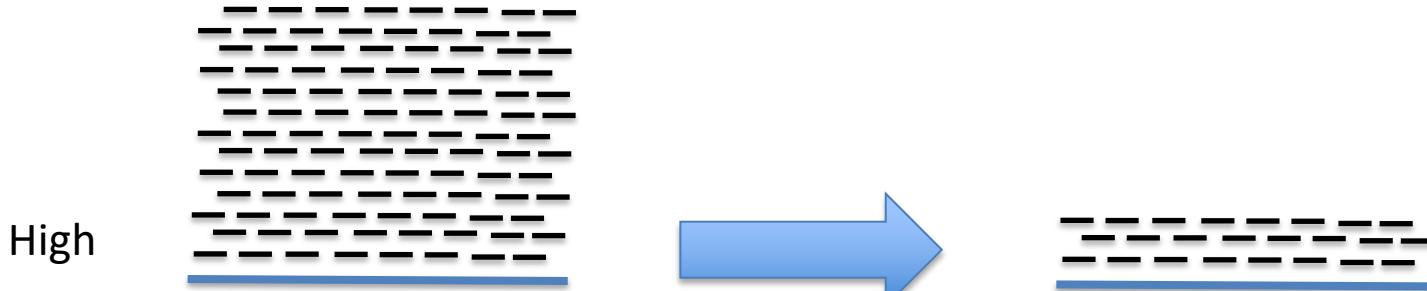


Trinity Framework for De novo Transcriptome Assembly and Analysis

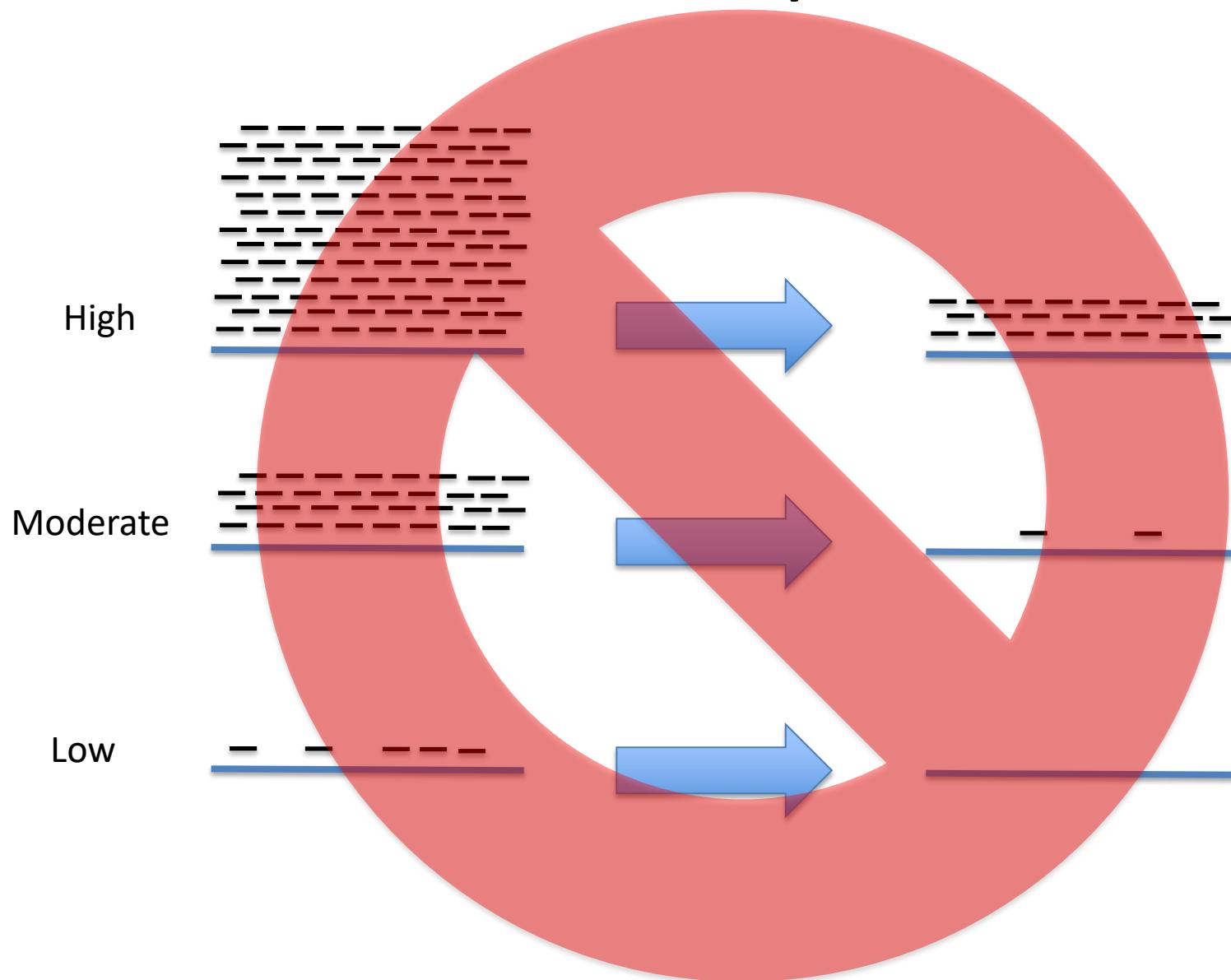
(focus of the transcriptomics lab)



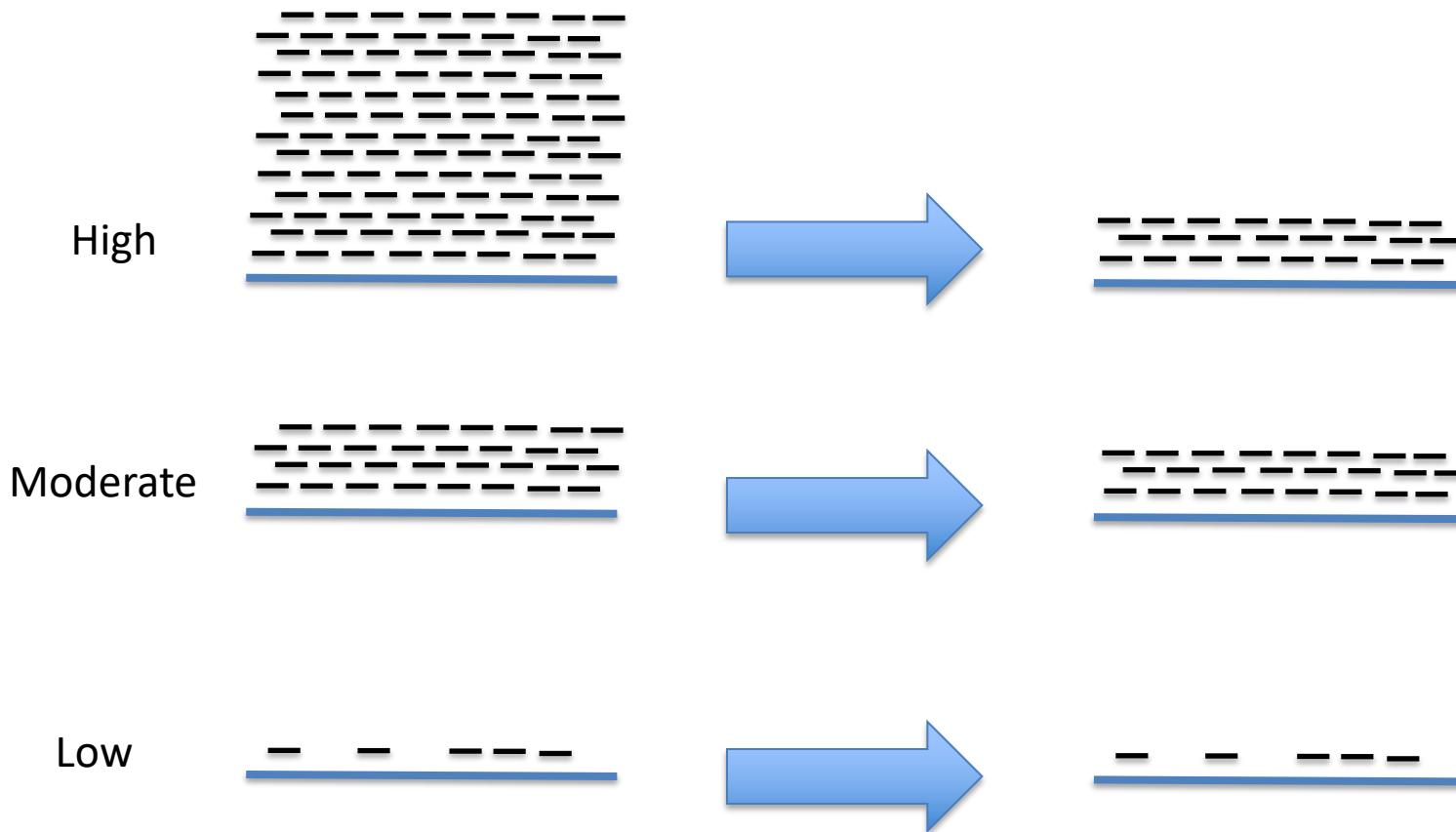
Could sub-sample the reads



Could sub-sample the reads



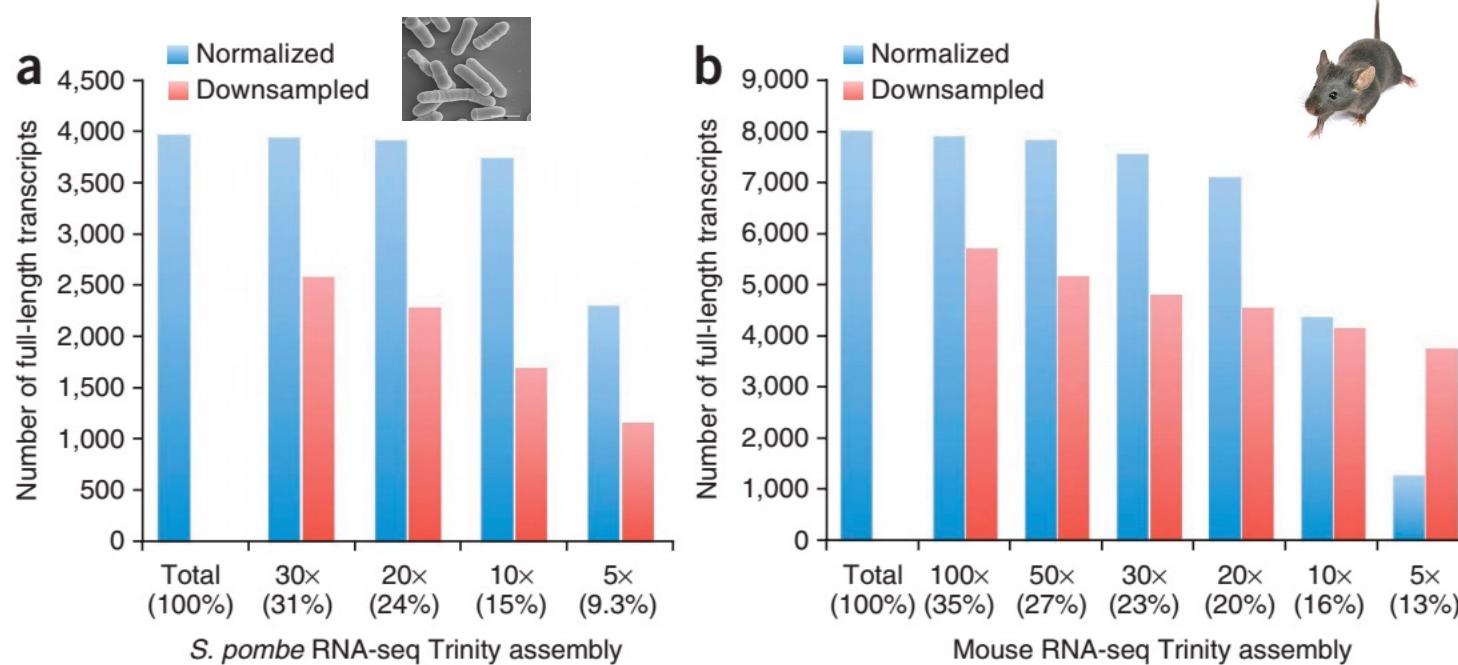
In silico normalization of reads



Select reads according to the probability:

$$P(\text{select read}) = \text{Min}\left(\frac{\text{target_coverage(read)}}{\text{observed_coverage(read)}}, 1\right)$$

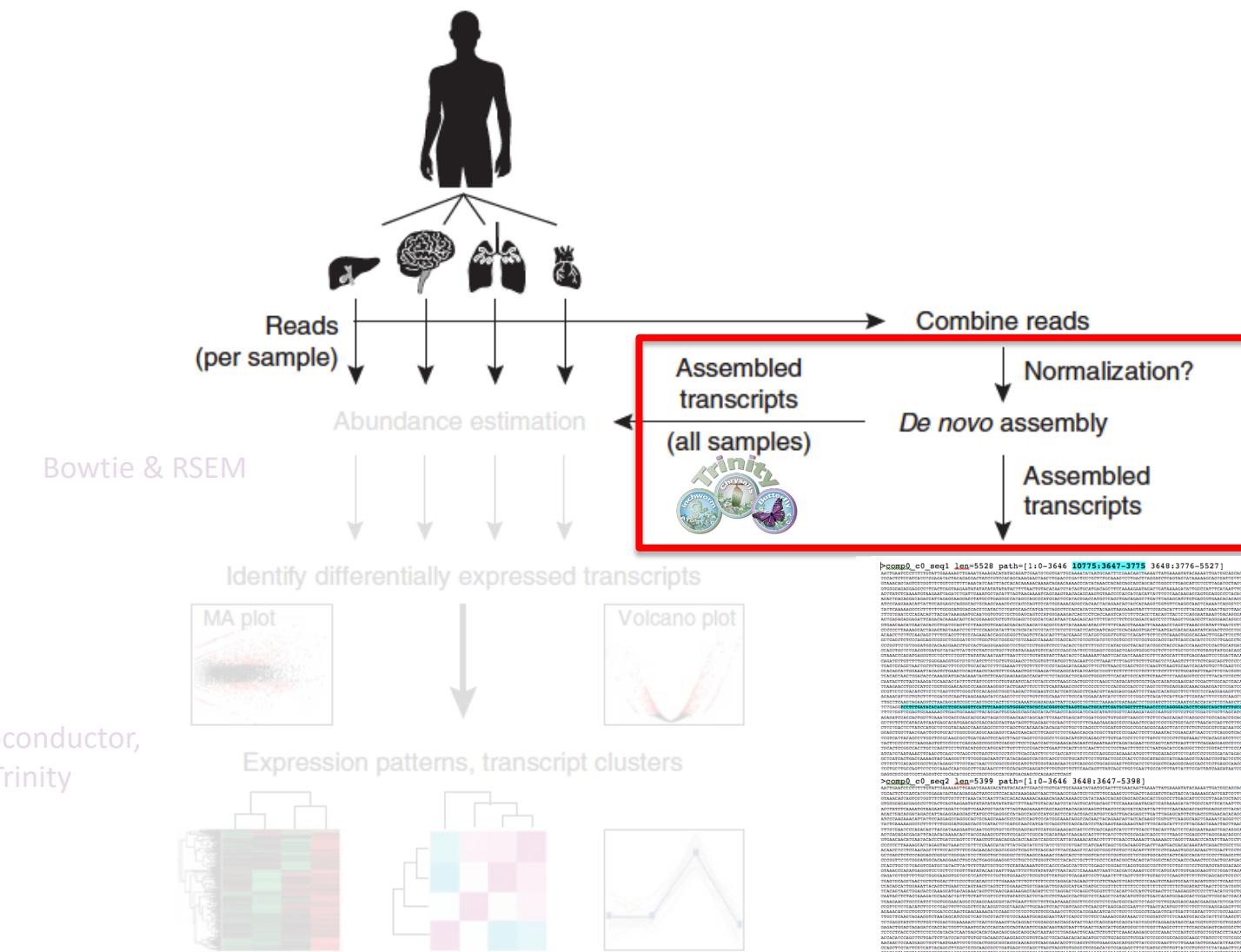
Impact of Normalization on *De novo* Full-length Transcript Reconstruction



Largely retain full-length reconstruction, but use less RAM and assemble much faster.

Can go from >1 billion reads down to < 100 M reads used in assembly.

The product of Trinity: a Fasta file of assembled transcripts



Trinity output: A multi-fasta file

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

De Brujin graph information

- Nodes: 279
- Edges: 332
- Total length: 4,685,914

Graph drawing

- Scope: Entire graph
- Style: Single Double
- Draw graph

Graph display

- Zoom: 44.4%
- Node width: 8.5
- Random colours

Node labels

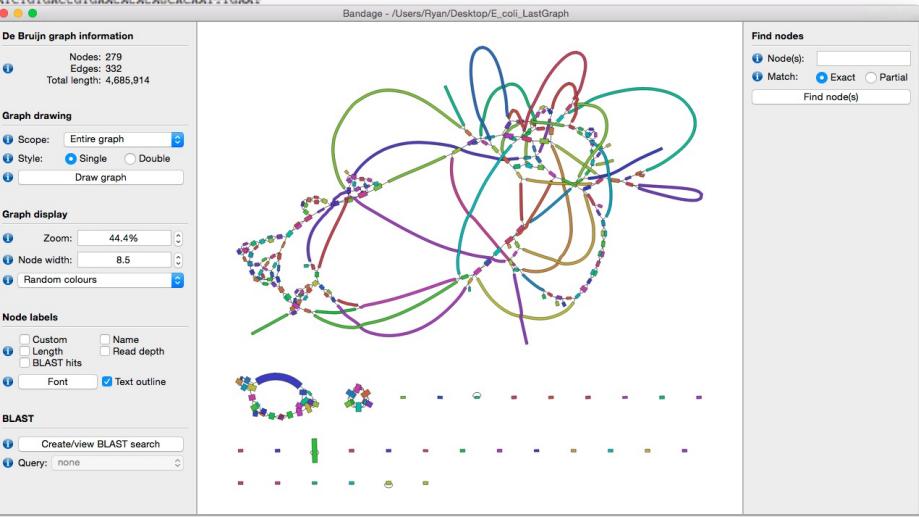
- Custom
- Length
- Name
- Read depth
- BLAST hits
- Font
- Outline outline

BLAST

- Create/view BLAST search
- Query: none

Can visualize using Bandage

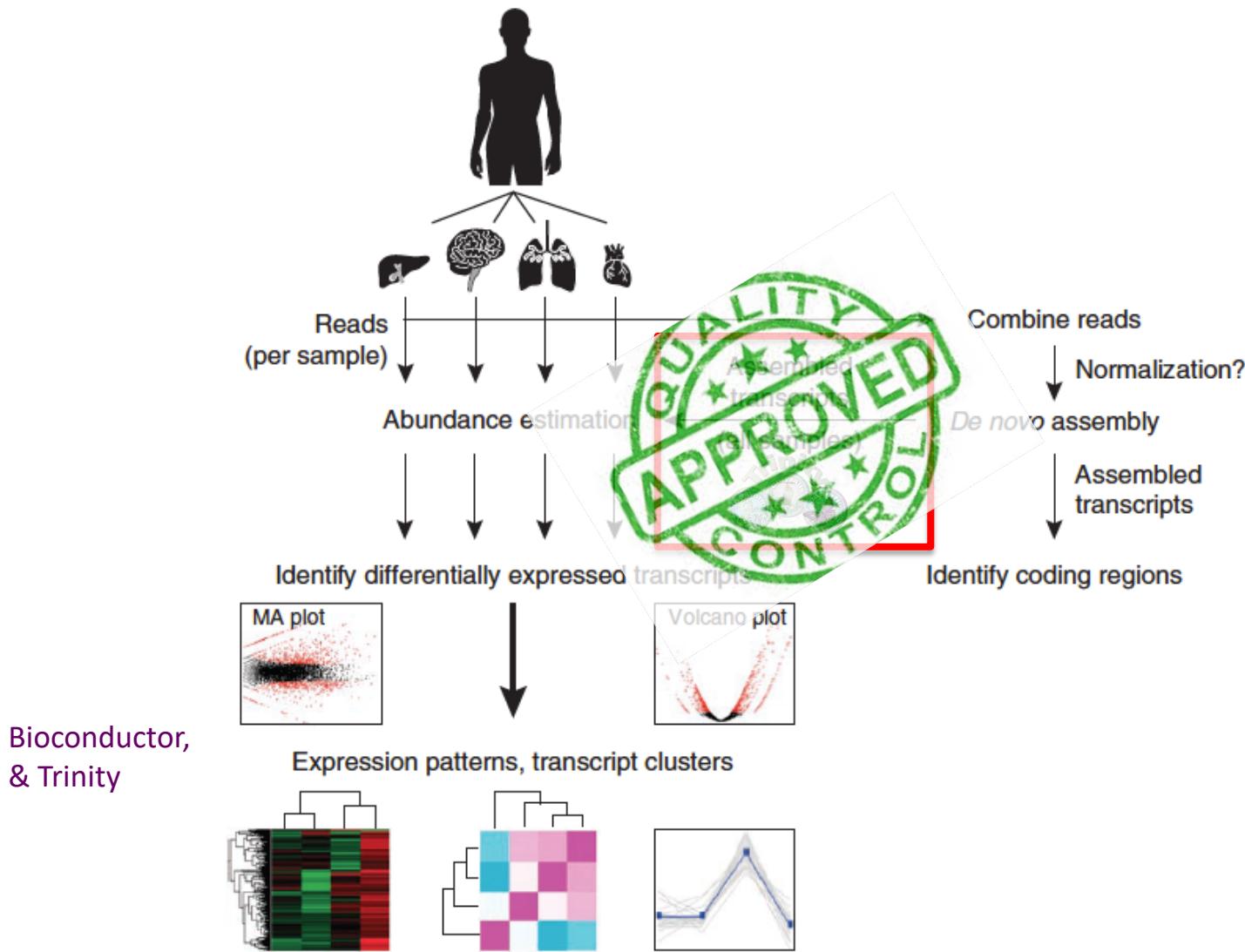
<https://rrwick.github.io/Bandage/>



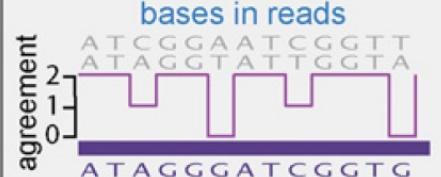
Part 4. Transcriptome Quality Assessment



Evaluating the quality of your transcriptome assembly

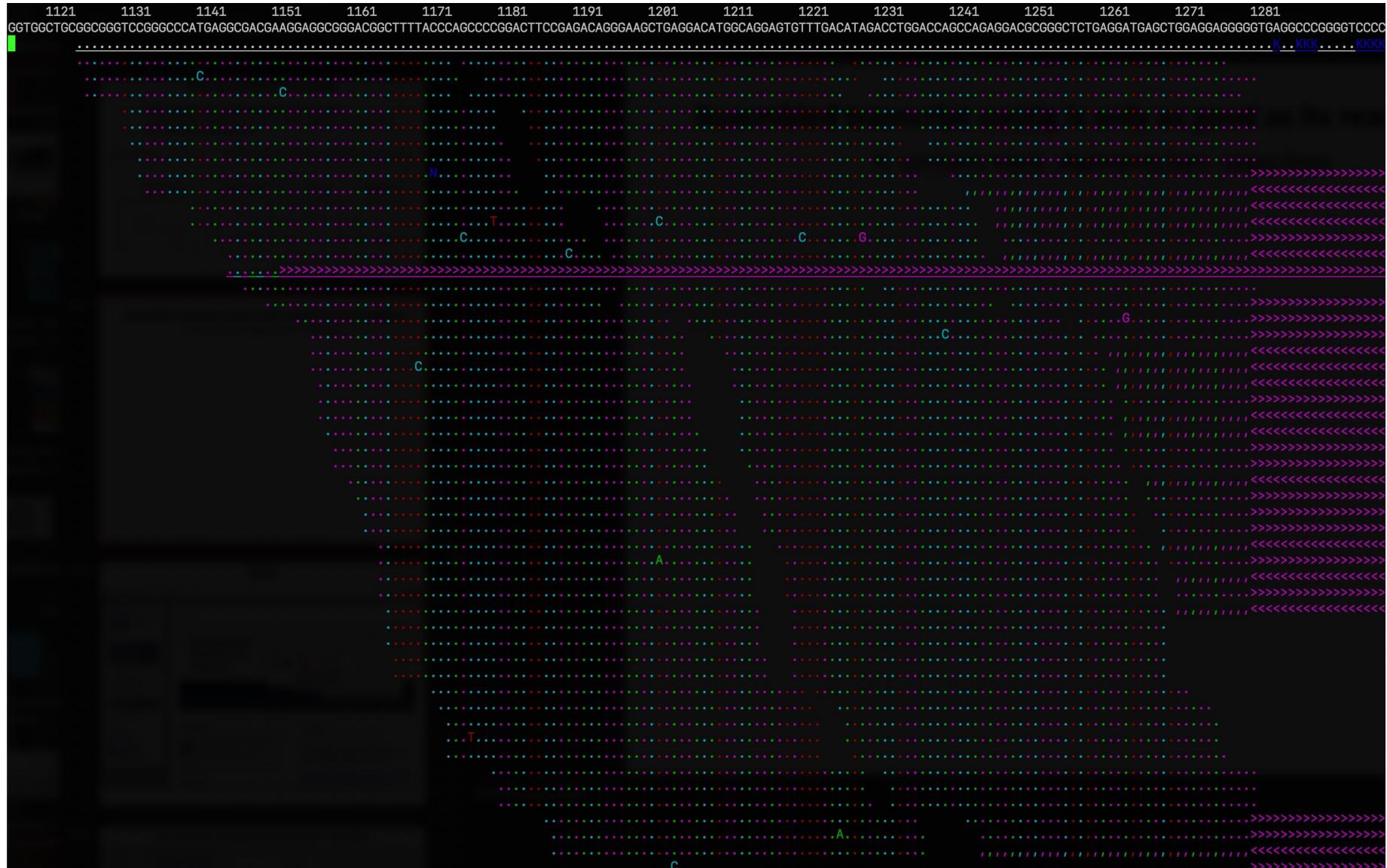


De novo Transcriptome Assembly is Prone to Certain Types of Errors

Error type	Transcripts	Assembly	Read evidence
Family collapse	geneAA geneAB geneAC n=3	n=1	
Chimerism	geneC geneB n=2	n=1	
Unsupported insertion	n=1	n=1	no reads align to insertion
Incompleteness	n=1	n=1	read pairs align off end of contig
Fragmentation	n=1	n=4	bridging read pairs
Local misassembly	n=1	n=1	read pairs in wrong orientation
Redundancy	n=1	n=3	all reads assign to best contig

Assembled transcript contig is only as good as its read support.

```
% samtools tview alignments.bam target.fasta
```



IGV

www.broadinstitute.org/igv/

igv Integrative Genomics Viewer ALABSL

- Home
- Downloads
- Documents
 - Hosted Genomes
 - FAQ
 - IGV User Guide
 - File Formats
 - Release Notes
 - Credits
- Contact

Search website

search

[Broad Home](#)

[Cancer Program](#)

BROAD INSTITUTE
© 2012 Broad Institute

Home

Integrative Genomics Viewer



What's New

July 3, 2012. Soybean (*Glycine max*) and Rat (rn5) genomes have been updated.

April 20, 2012. IGV 2.1 has been released.
See the [release notes](#) for more details.

April 19, 2012. See our new [IGV paper](#) in *Briefings in Bioinformatics*.

Overview

Citing IGV

To cite your use of IGV in your publication:

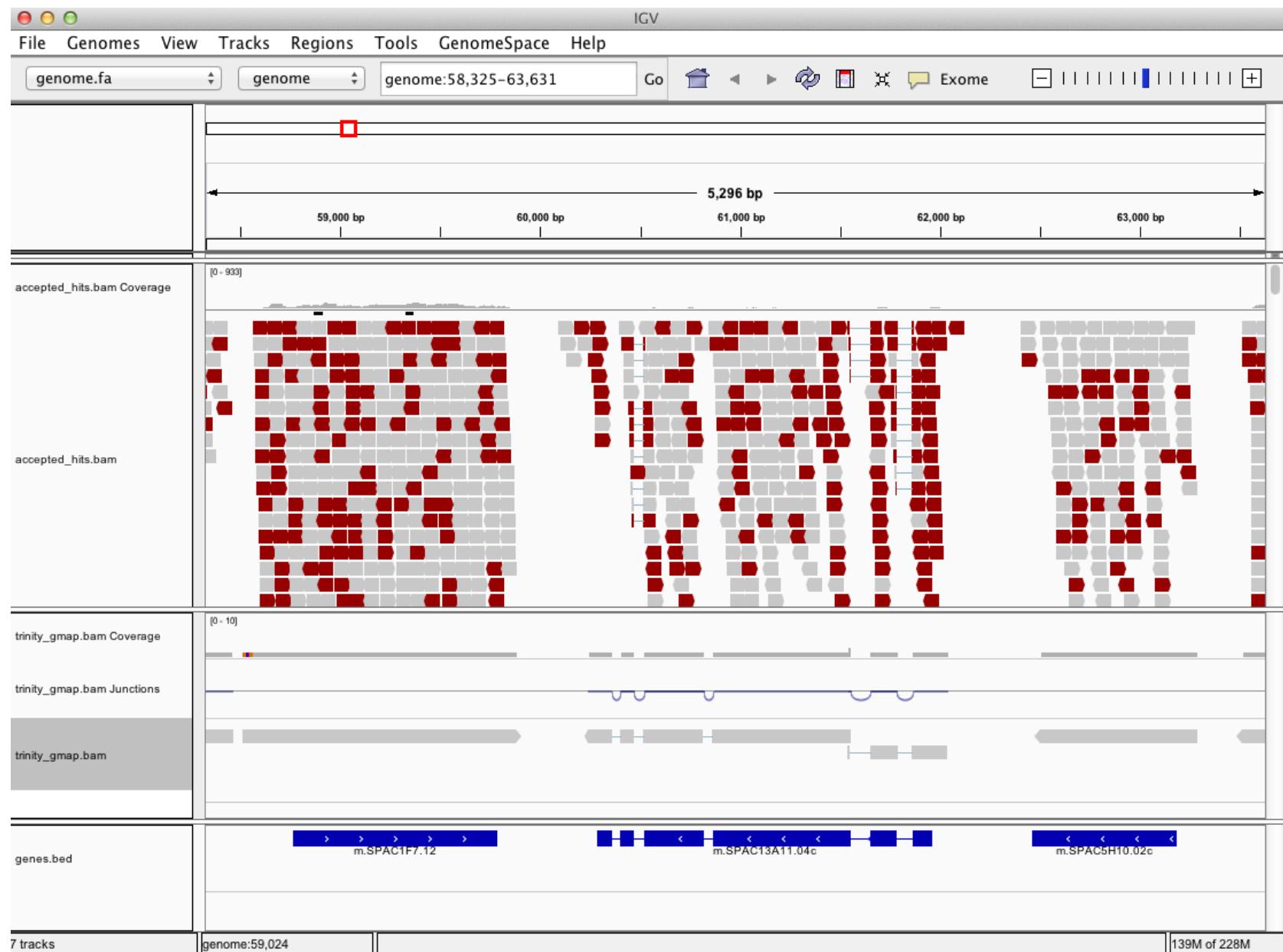
James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, Jill P. Mesirov. [Integrative Genomics Viewer](#). *Nature Biotechnology* 29, 24–26 (2011), or
Helga Thorvaldsdottir, James T. Robinson, Jill P. Mesirov. [Integrative Genomics Viewer \(IGV\): high-performance genomics data visualization and exploration](#).

Can Examine Transcript Read Support Using IGV



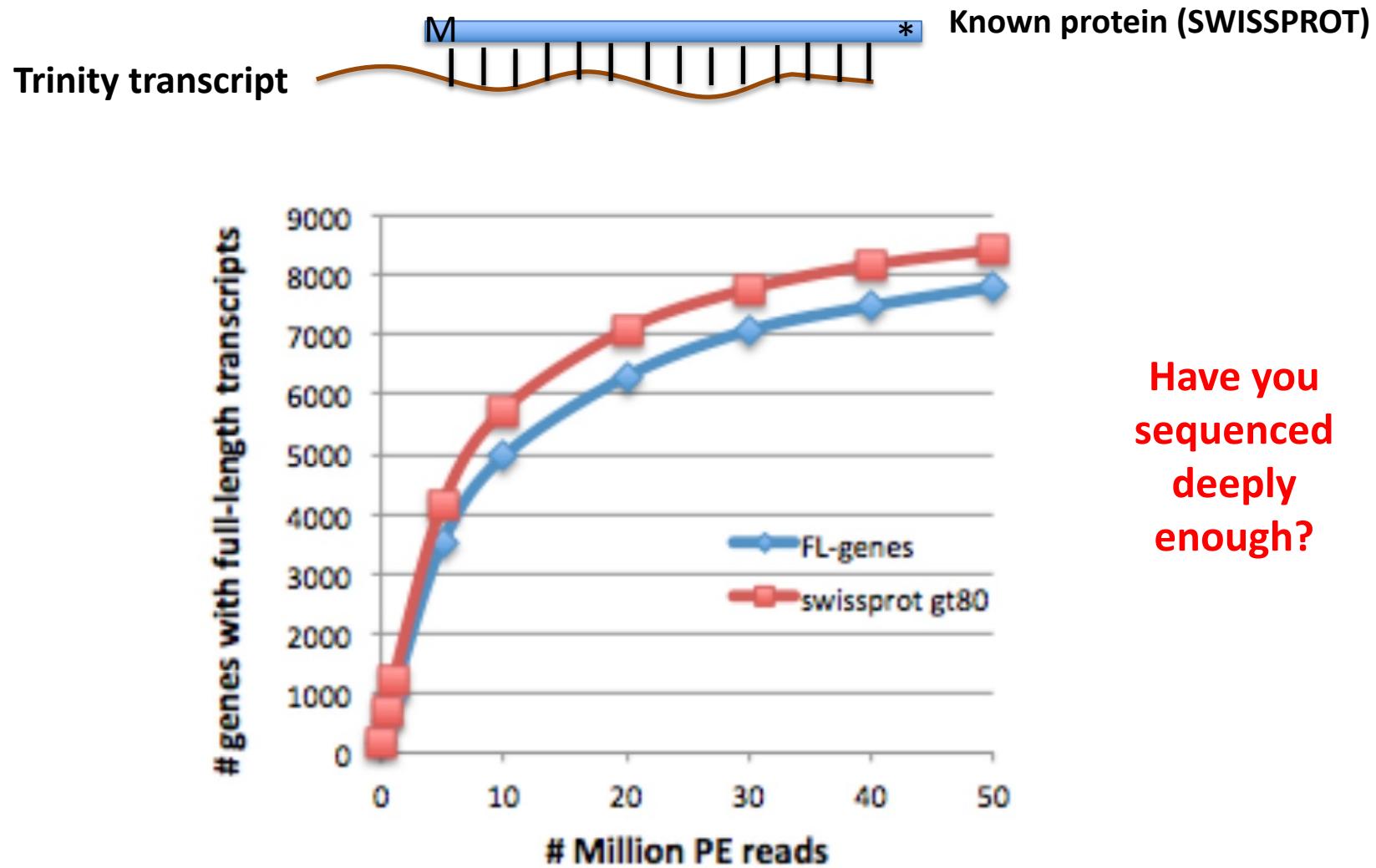
Can align Trinity transcripts to genome scaffolds to examine intron/exon structures

(Trinity transcripts aligned to the genome using GMAP)



Evaluating the quality of your transcriptome assembly

Full-length Transcript Detection via BLASTX





UNIVERSITÉ
DE GENÈVE

FACULTÉ DE MÉDECINE

Zdobnov's Computational Evolutionary Genomics
group

CEGG Home | OrthoDB v9 | BUSCO v2

Latest is v5.4.7

BUSCO v2

Assessing genome assembly and
annotation completeness with
Benchmarking Universal Single-
Copy Orthologs

About BUSCO

BUSCO v2 provides quantitative measures for the assessment of genome assembly, gene set, and transcriptome completeness, based on evolutionarily-informed expectations of gene content from near-universal single-copy orthologs selected from [OrthoDB v9](#).

BUSCO assessments are implemented in open-source software, with a large selection of lineage-specific sets of Benchmarking Universal Single-Copy Orthologs. These conserved orthologs are ideal candidates for large-scale phylogenomics studies, and the annotated BUSCO gene models built during genome assessments provide a comprehensive gene predictor training set for use as part of genome annotation pipelines.



UNIVERSITÉ
DE GENÈVE

FACULTÉ DE MÉDECINE

Zdobnov's Computational Evolutionary Genomics
group

CEGG Home | OrthoDB v9 | BUSCO v2

Latest is v5.4.7

BUSCO v2

Assessing genome assembly and
annotation completeness with
Benchmarking Universal Single-
Copy Orthologs

```
#Summarized BUSCO benchmarking for file: Trinity.fasta
#BUSCO was run in mode: trans
```

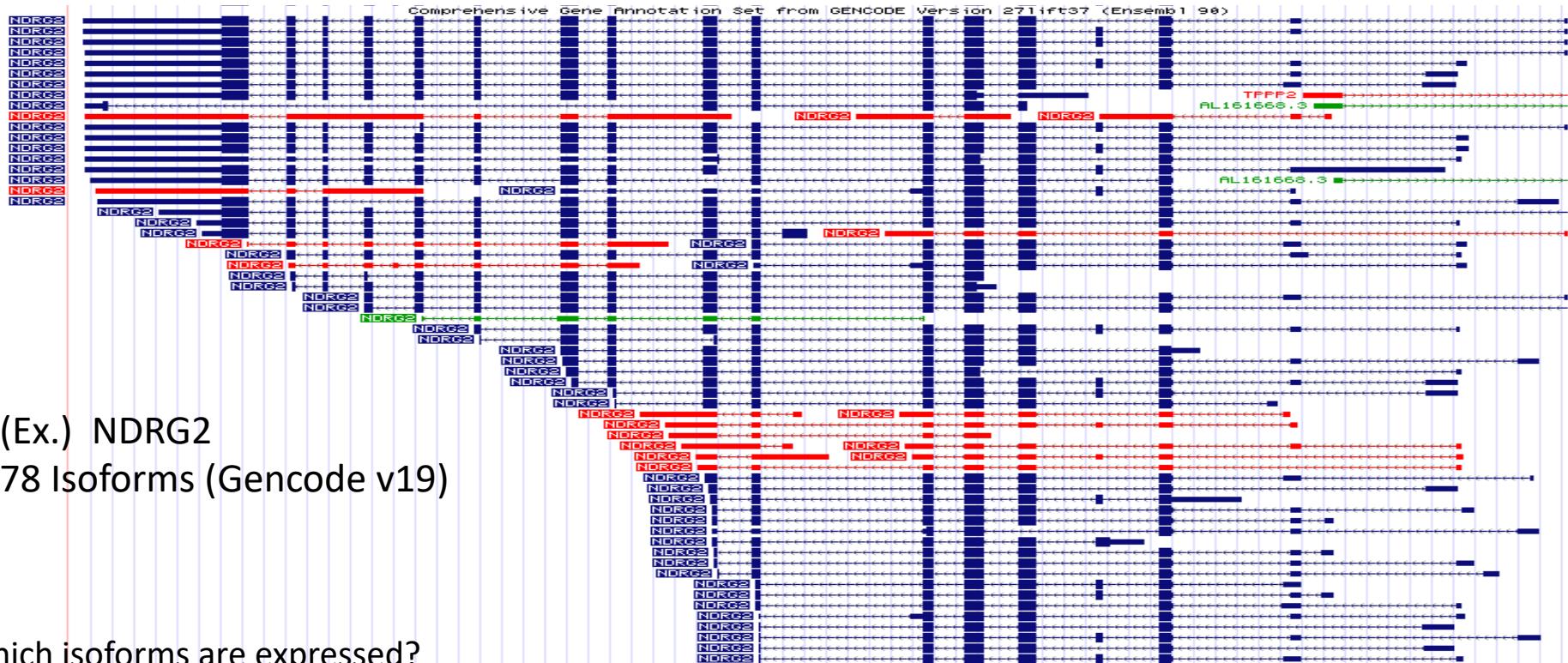
Summarized benchmarks in BUSCO notation:

C:88%[D:53%],F:4.5%,M:7.3%,n:3023

Representing:

1045	Complete Single-copy BUSCOs
1617	Complete Duplicated BUSCOs
139	Fragmented BUSCOs
222	Missing BUSCOs
3023	Total BUSCO groups searched

Transcript Reconstruction or Expression Analysis can be Quite Difficult at Complex Loci

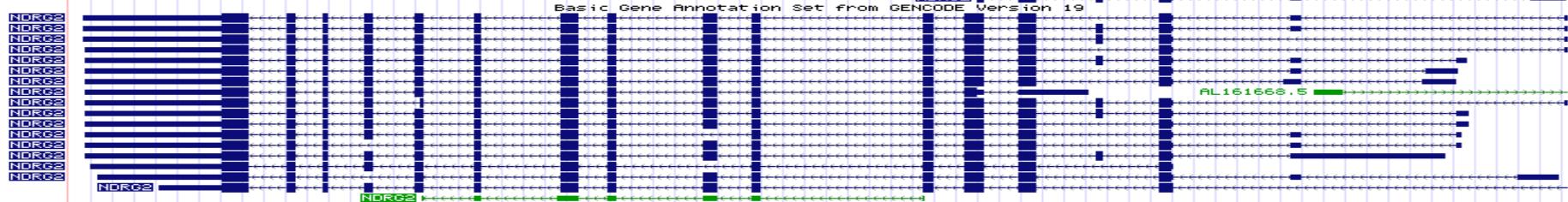


(Ex.) *NDRG2*

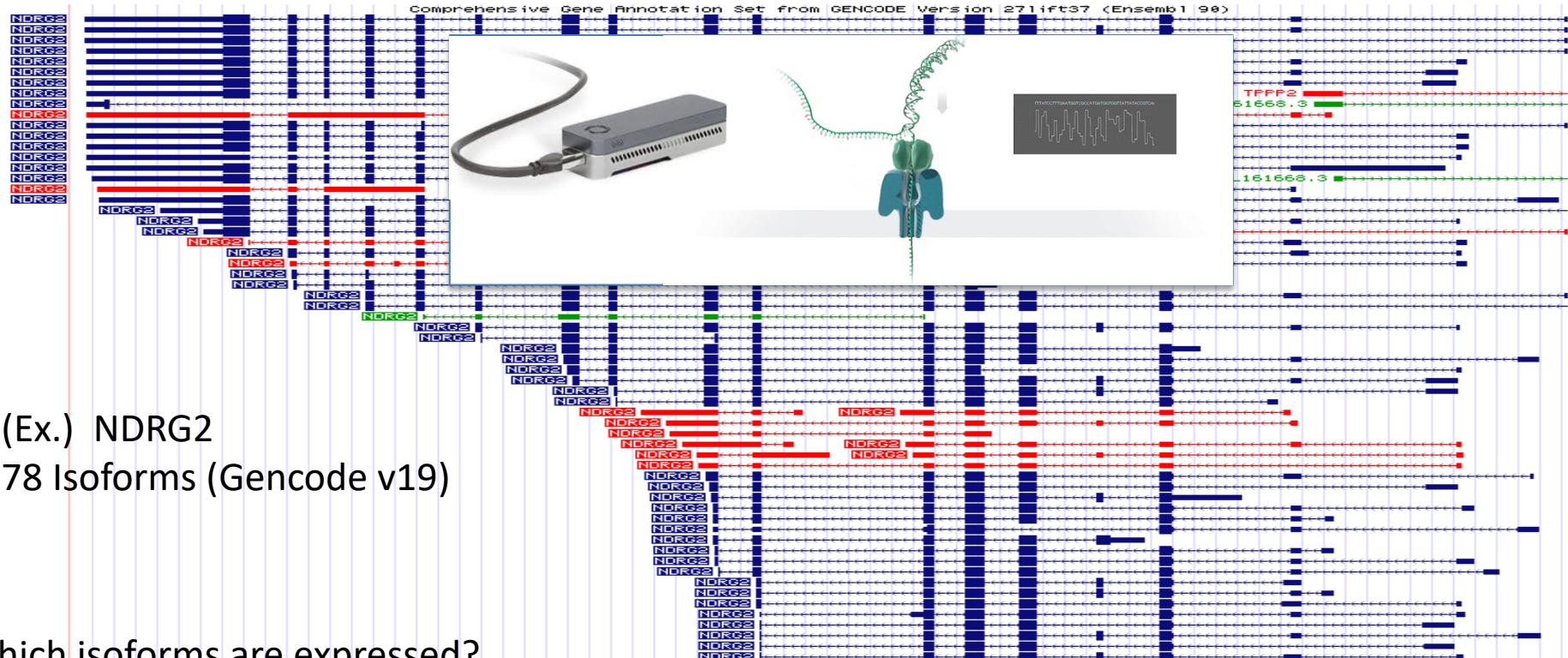
78 Isoforms (Gencode v19)

Which isoforms are expressed?

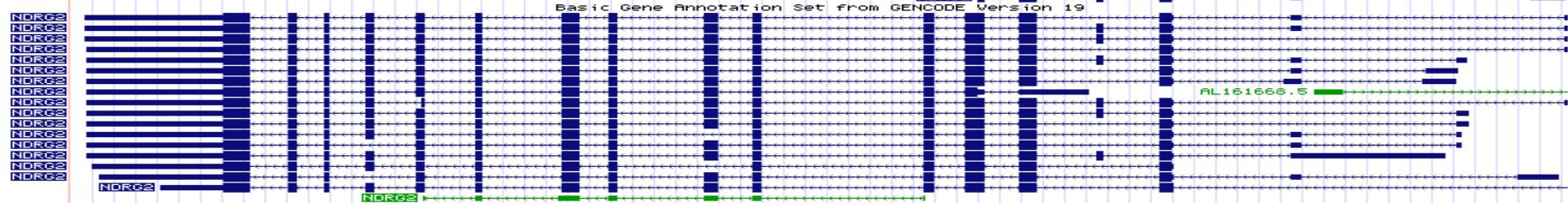
Which can be confidently reconstructed from short reads?



Too complex... don't guess from short reads, use long reads.

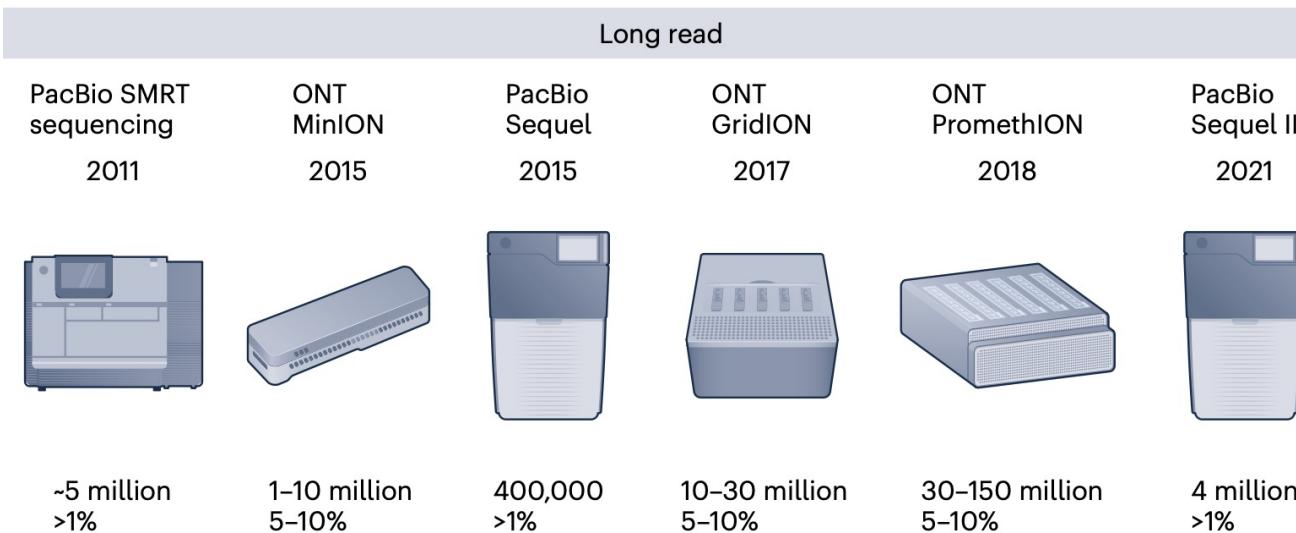


Which isoforms are expressed?
Is there evidence of differential transcript usage?



Method of the Year 2022: long-read sequencing

The variables on RNA molecules: concert or cacophony? Answers in long-read sequencing



Inflection point for LR transcriptomics



MAS-seq → 40-120 million cDNA reads

Info on error rates for long reads – impressive!!

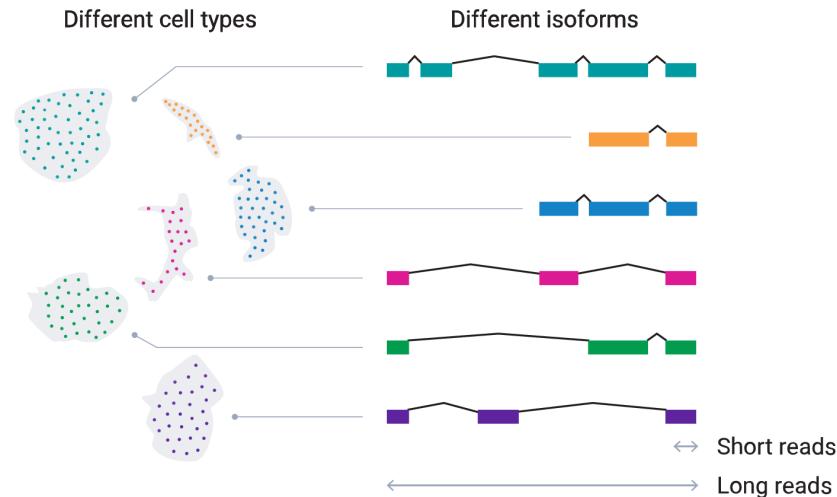
<https://nanoporetech.com/accuracy>

<https://www.pacb.com/technology/hifi-sequencing/>

99% 99.9%

Q30 Q40

Long reads for Single Cell Transcriptomes!!



Key Points

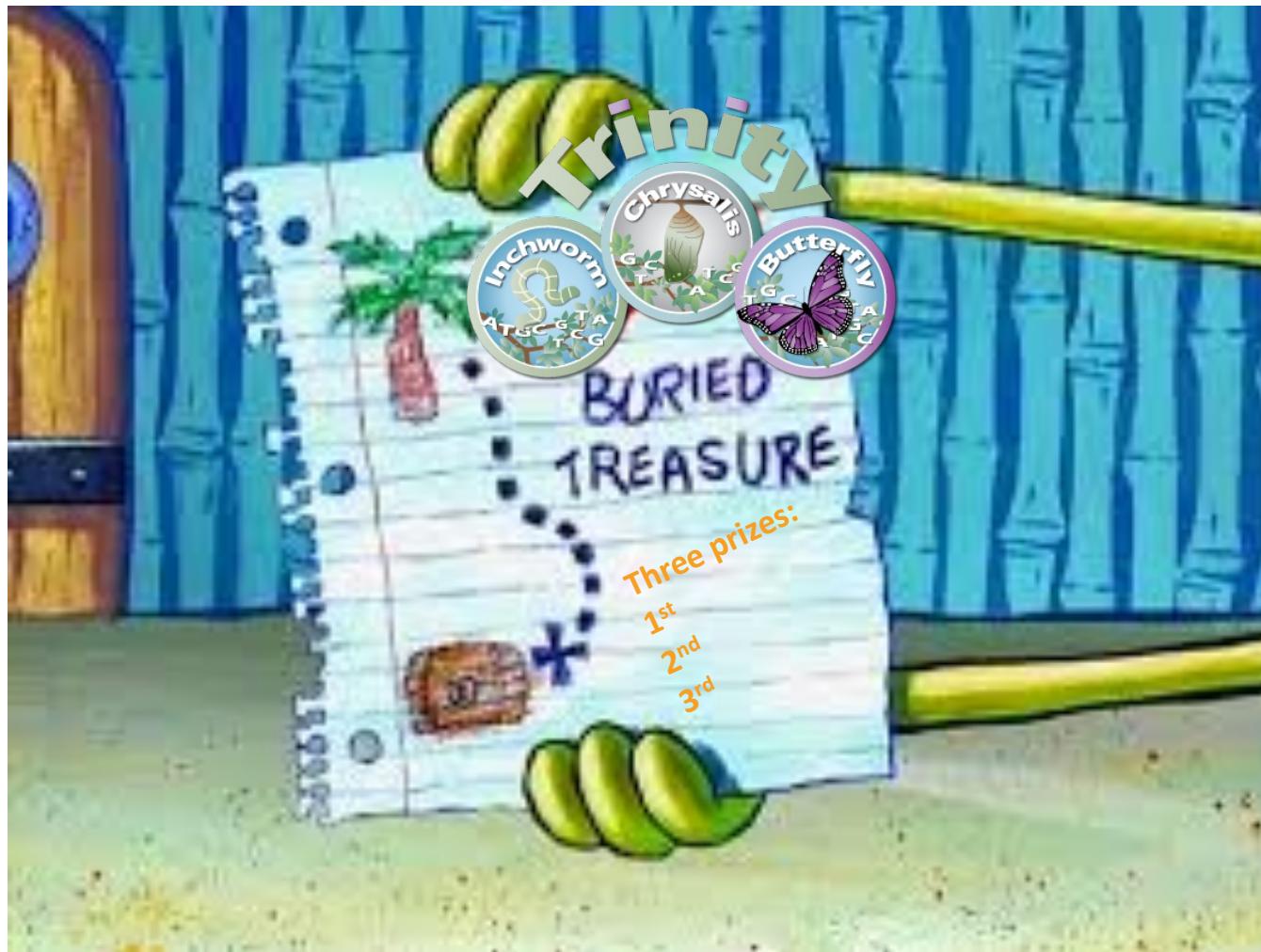
- RNA-seq enables many aspects of biology to be studied at single base & single cell resolution plus spatial context.
- Different isolation/capture methods available
- Reconstruction typically involves graph reconstruction from reads or alignments and path traversal.
- Do strand-specific sequencing whenever possible (eg. TruSeq)
- For QC – examine read support and full-length reconstruction stats.
- Latest advancements: long read transcriptome sequencing yields isoform structure info at single cell resolution (eg. MAS-seq).

Running Trinity

(on small sets of reads)

```
Trinity --left reads.left.fa \
    --right reads.right.fa \
    --seqType fa \
    --max_memory 1G \
    --CPU 1 \
    --output trinity_outdir \
    --no_normalize_reads
```

Trinity Treasure Hunt!!! 😊



Will provide link to the challenge via slack – stay tuned, will start ~ 8pm

Slack channel: #transcriptomicslab