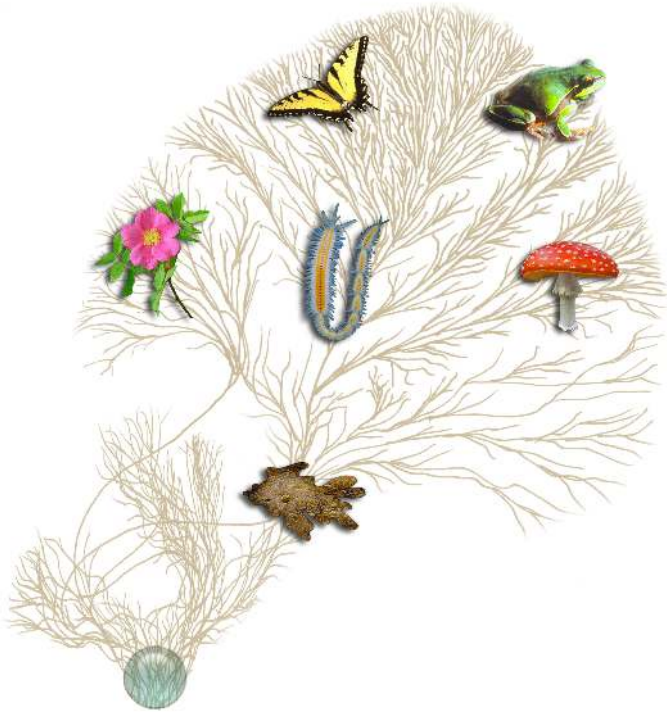
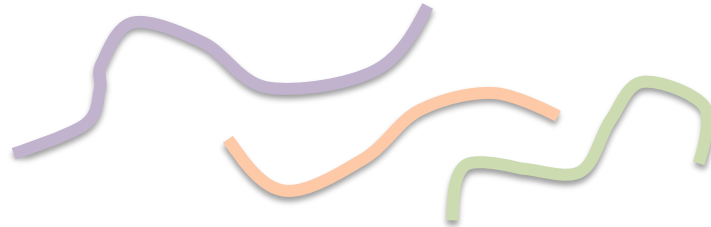


RNA-Seq Empowers Transcriptome Studies



Extract RNA, convert to cDNA



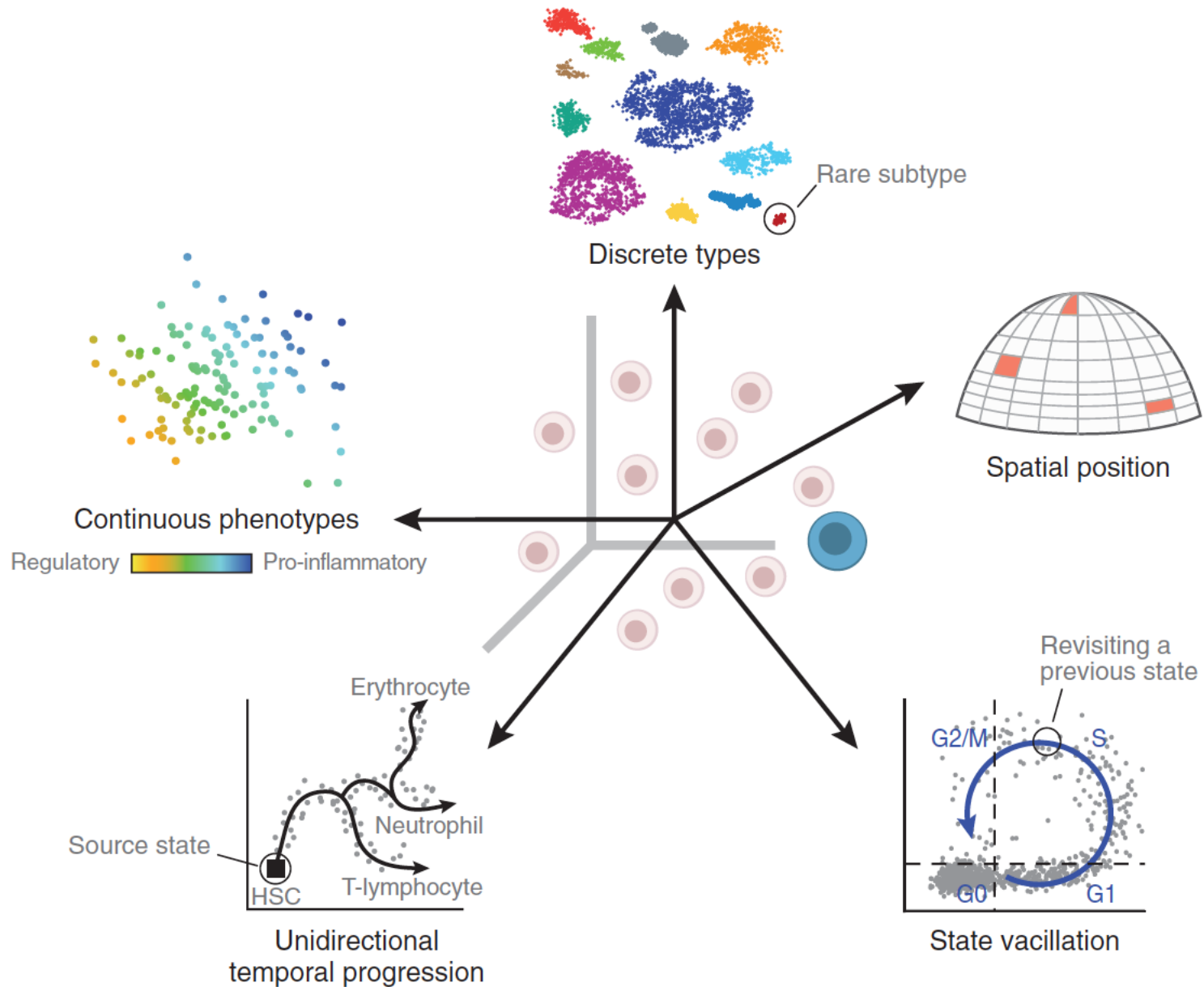
Next-gen Sequencer
(pick your favorite)



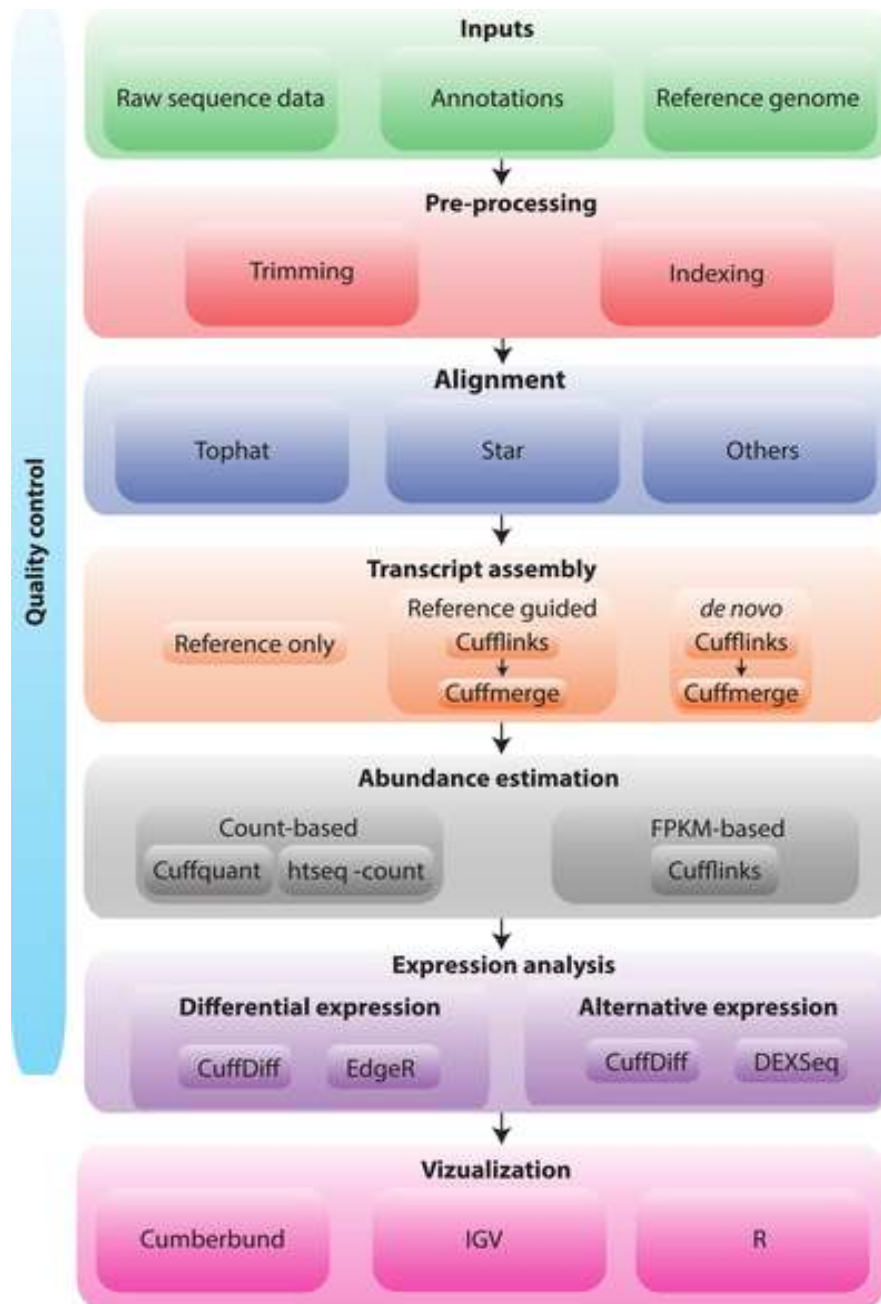
RNA-Seq Empowers Many Facets of Biological Investigations

- Transcript identification (ie. which genes active)
- Expression Levels
- Alternative splicing isoforms
- Allelic variants
- Mutations
- Fusion Transcripts
- RNA-editing

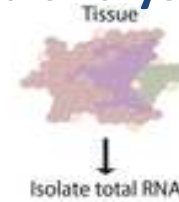
RNA-Seq is Empowering Discovery at Single Cell Resolution



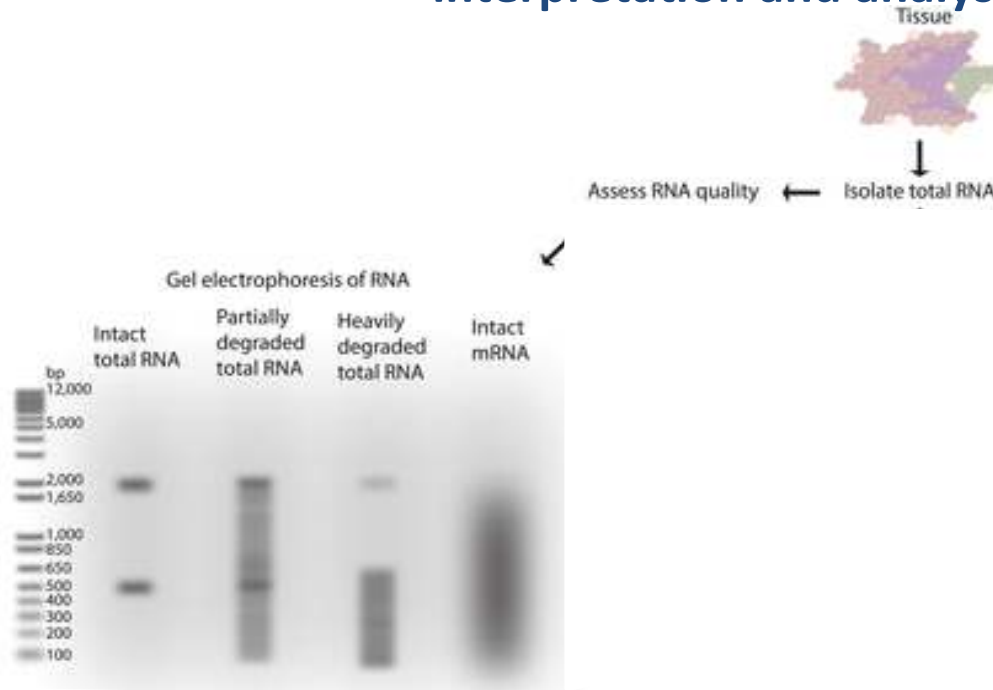
RNA-seq analysis flow chart.



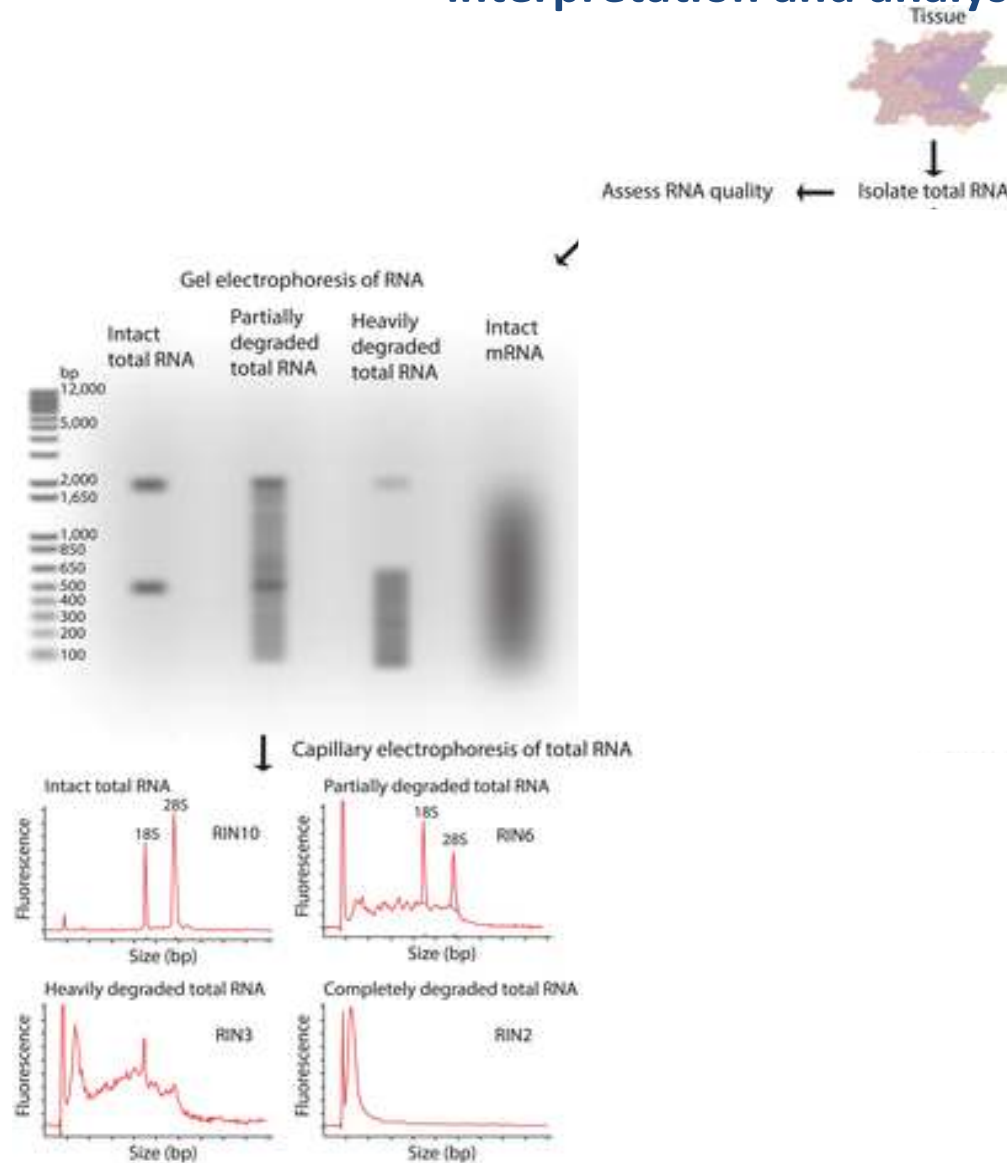
RNA-seq library fragmentation and size selection strategies that influence interpretation and analysis.



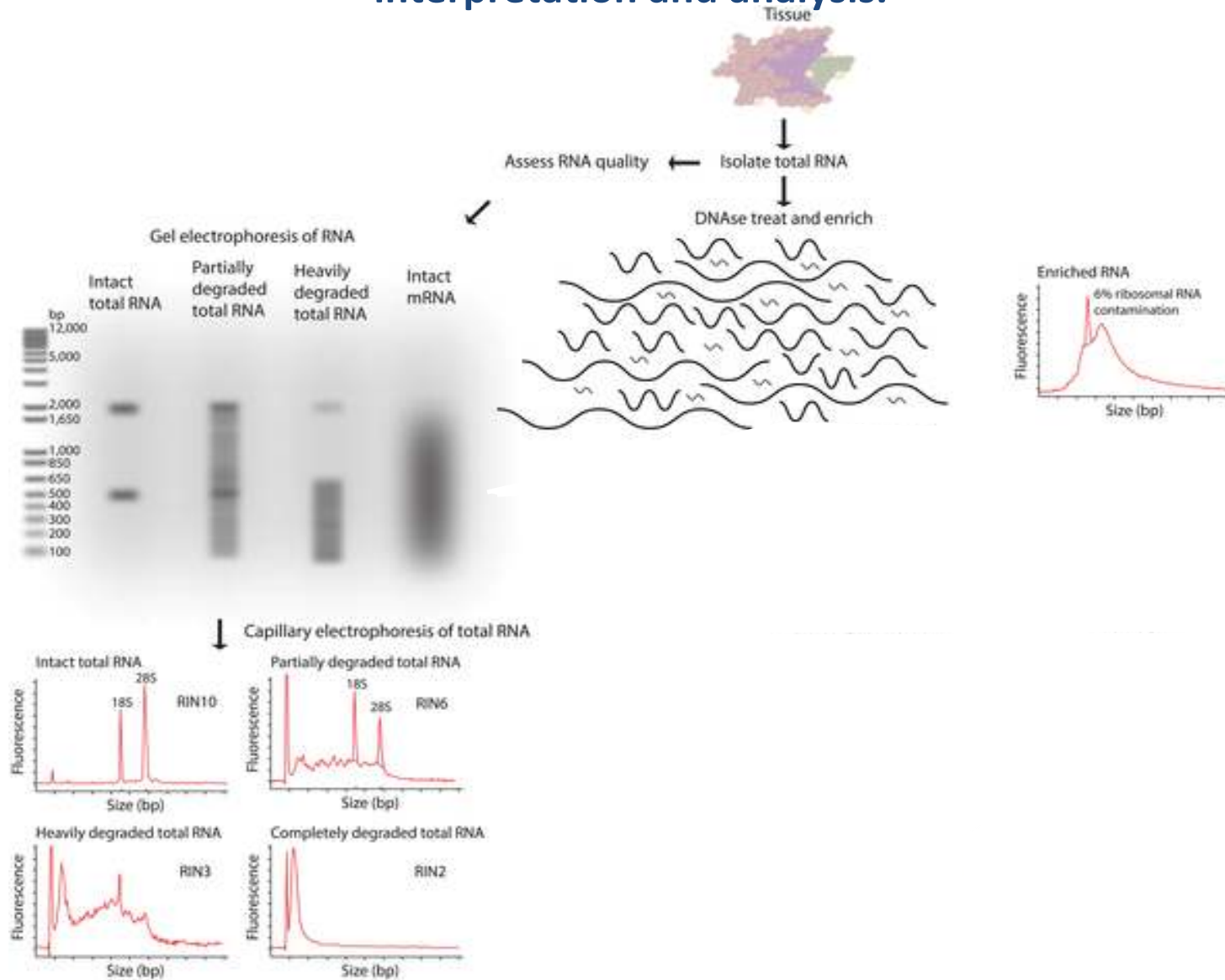
RNA-seq library fragmentation and size selection strategies that influence interpretation and analysis.



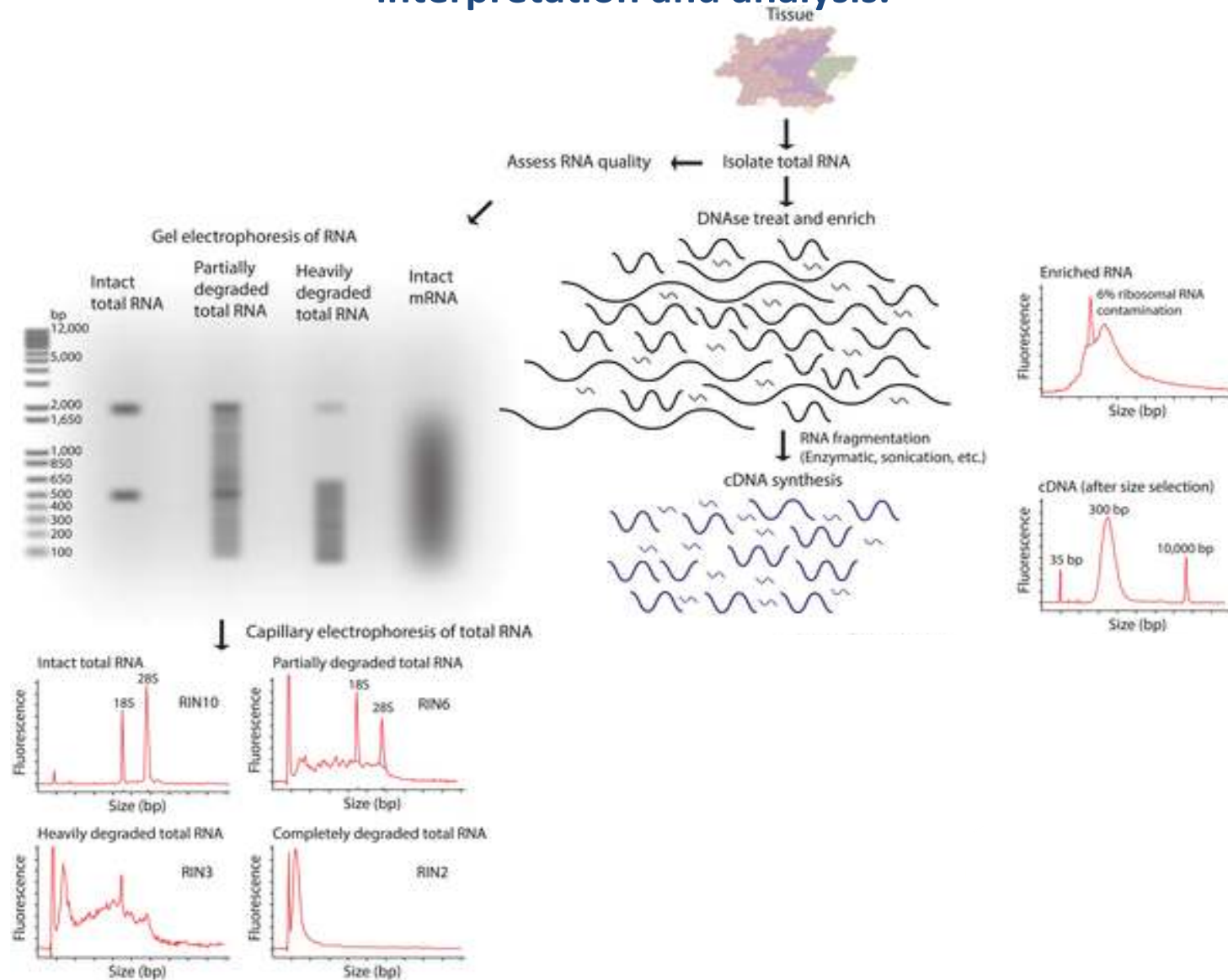
RNA-seq library fragmentation and size selection strategies that influence interpretation and analysis.



RNA-seq library fragmentation and size selection strategies that influence interpretation and analysis.

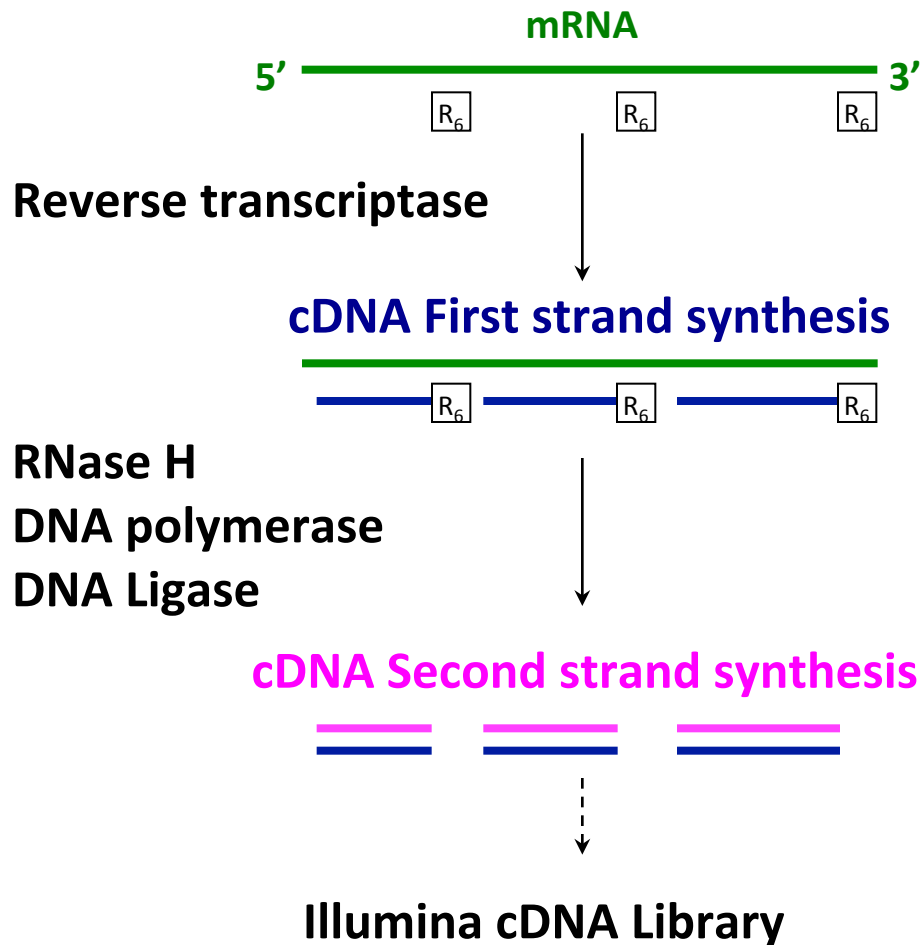


RNA-seq library fragmentation and size selection strategies that influence interpretation and analysis.

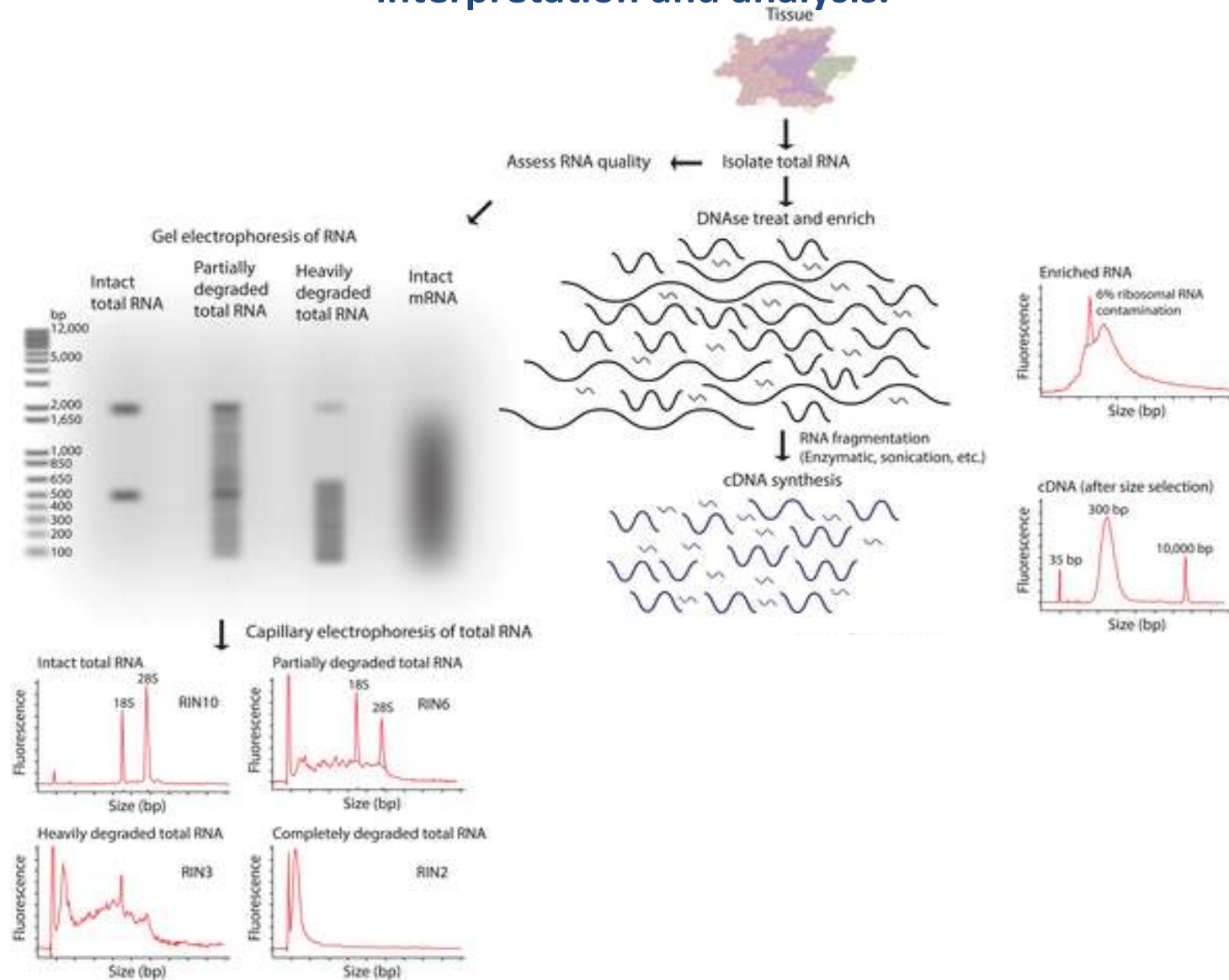


RNA-Seq: How do we make cDNA?

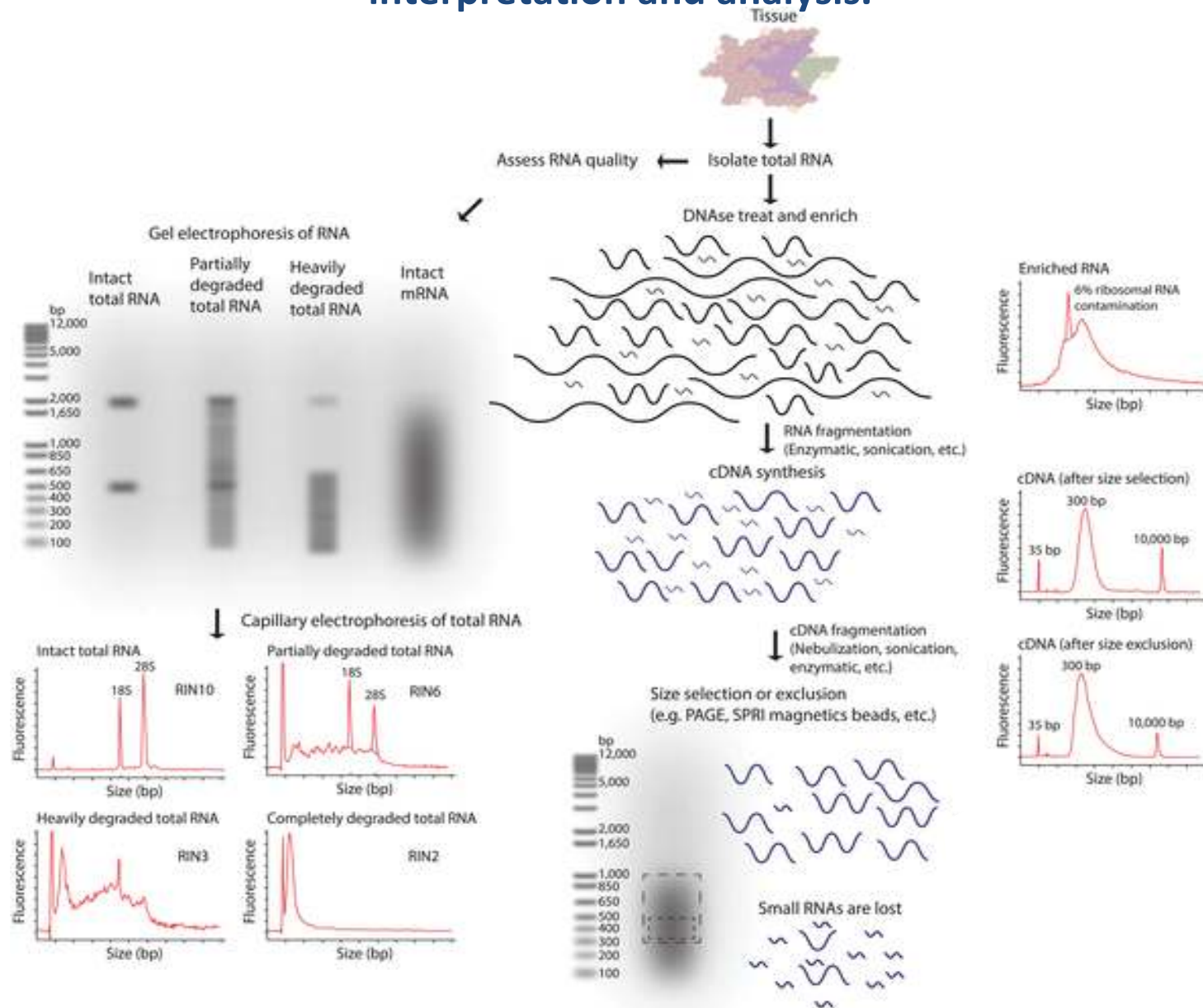
Prime with Random Hexamers (R6)



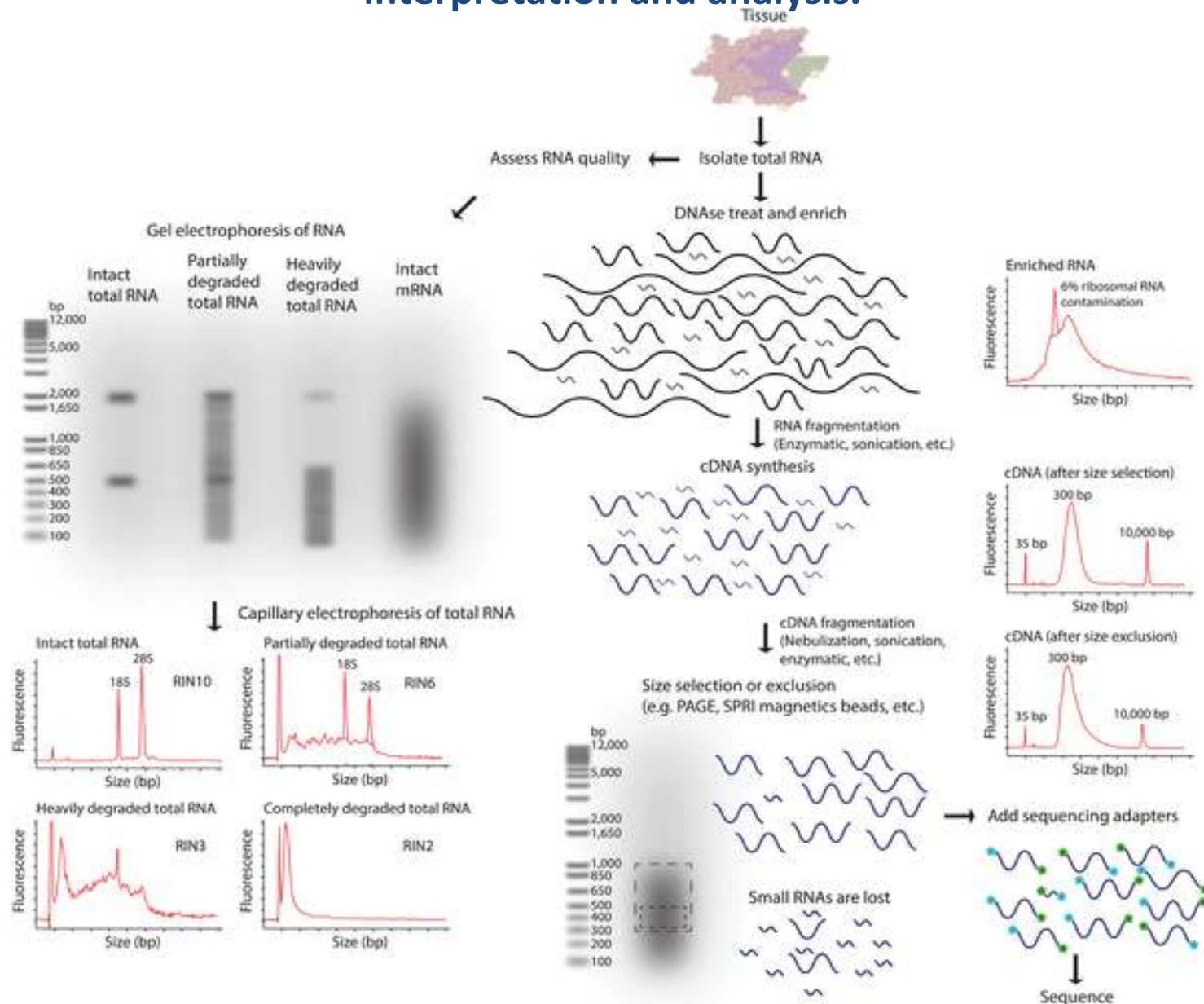
RNA-seq library fragmentation and size selection strategies that influence interpretation and analysis.



RNA-seq library fragmentation and size selection strategies that influence interpretation and analysis.

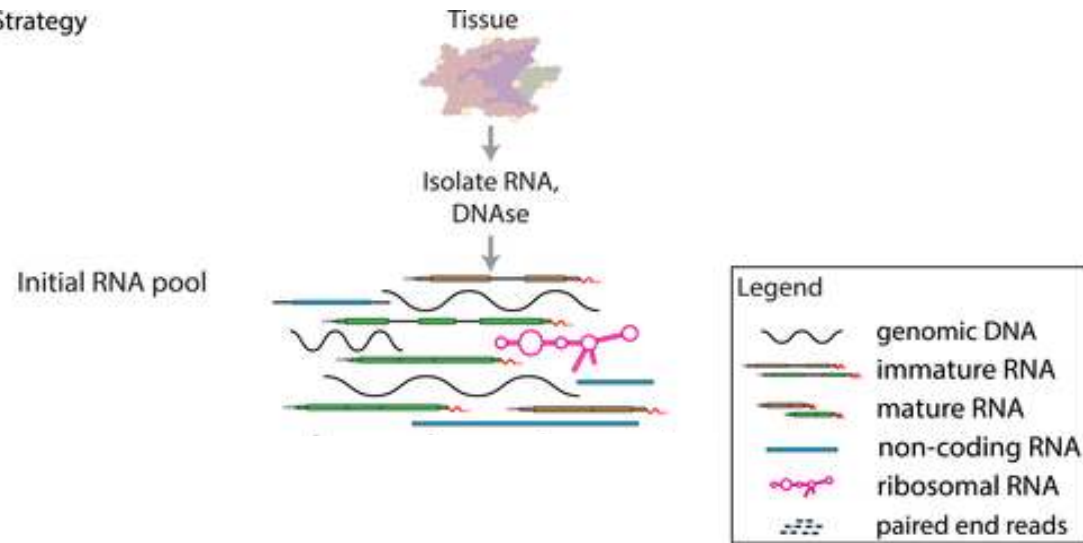


RNA-seq library fragmentation and size selection strategies that influence interpretation and analysis.



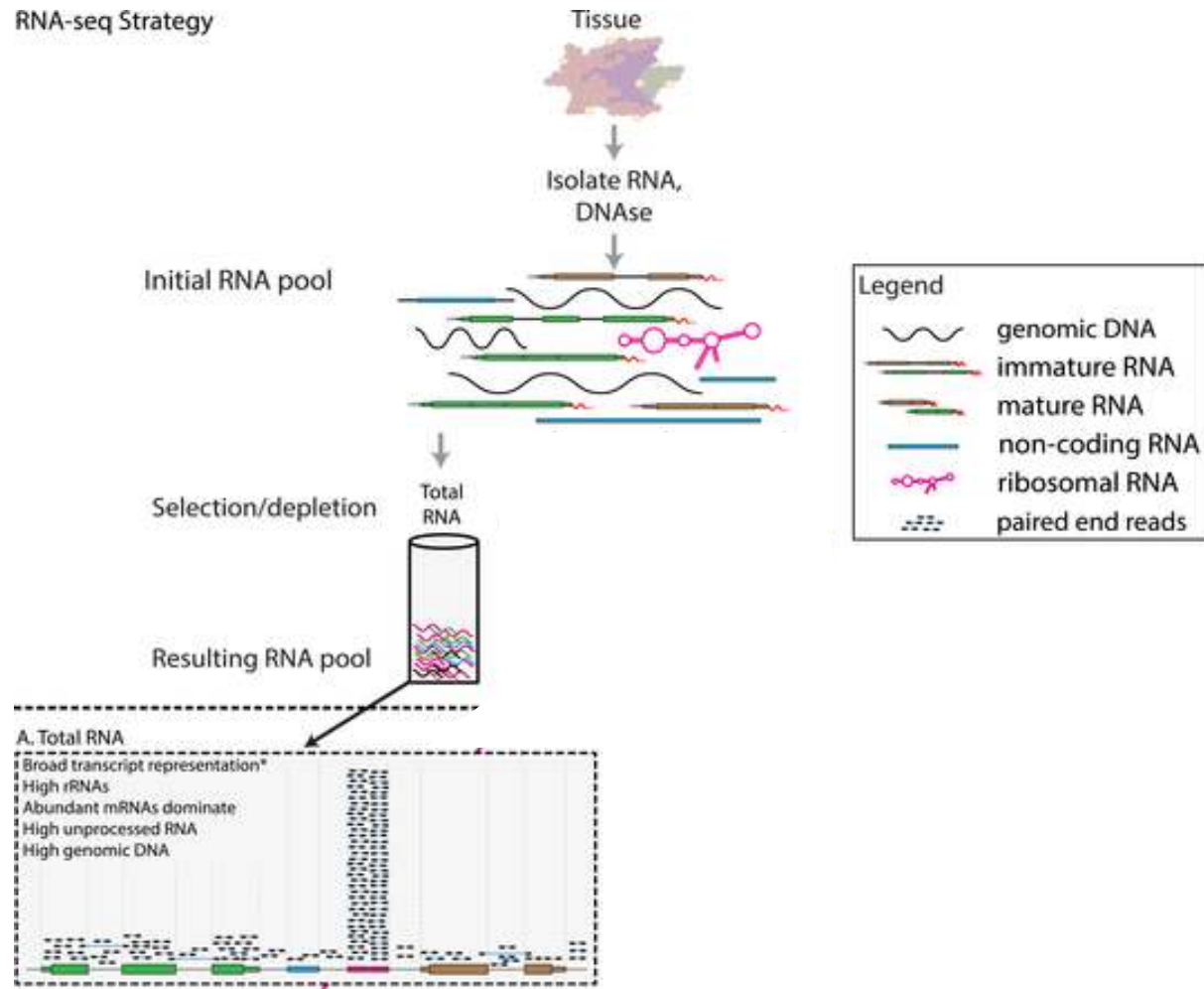
RNA-seq library enrichment strategies that influence interpretation and analysis.

RNA-seq Strategy



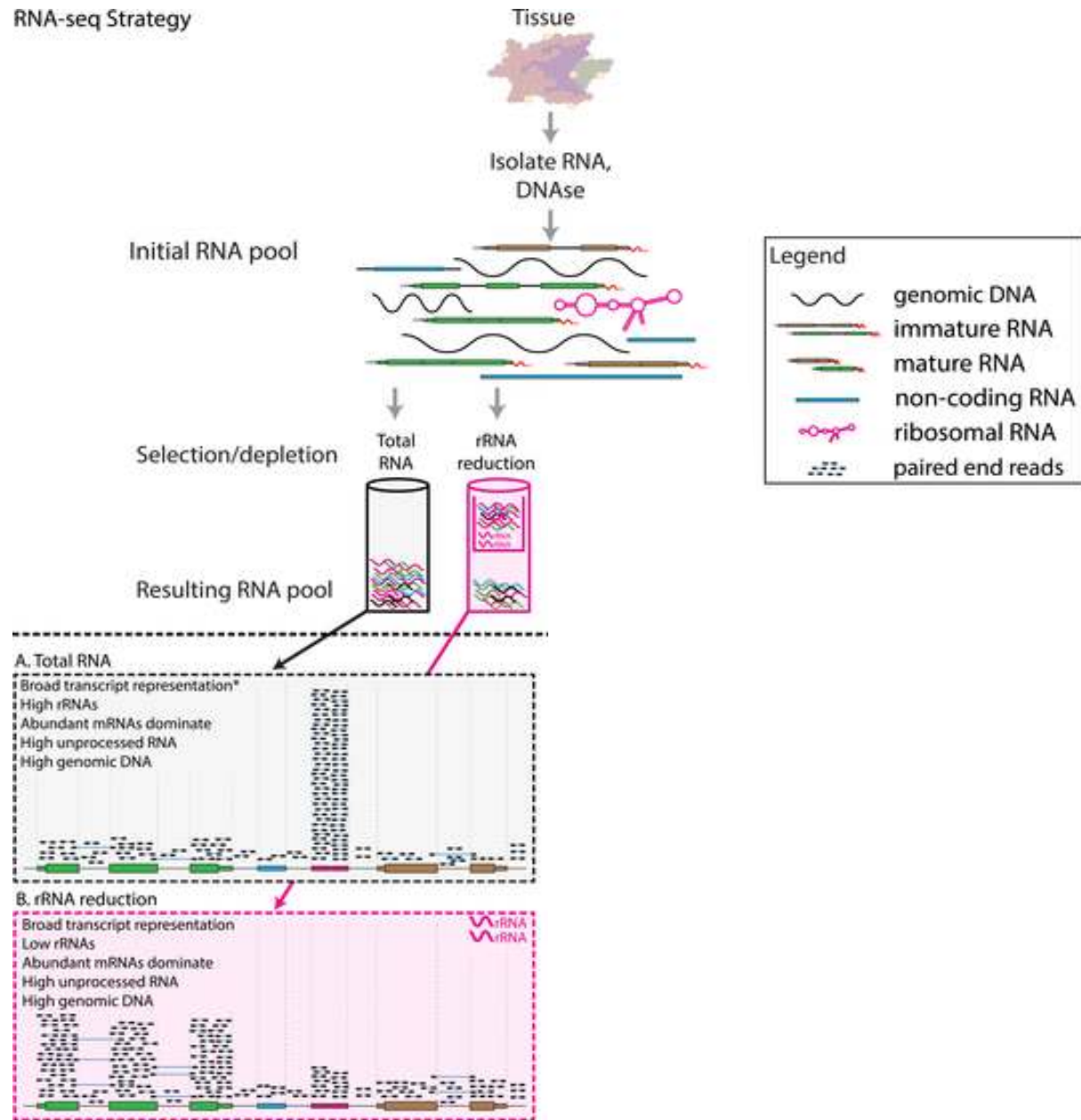
RNA-seq library enrichment strategies that influence interpretation and analysis.

RNA-seq Strategy



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

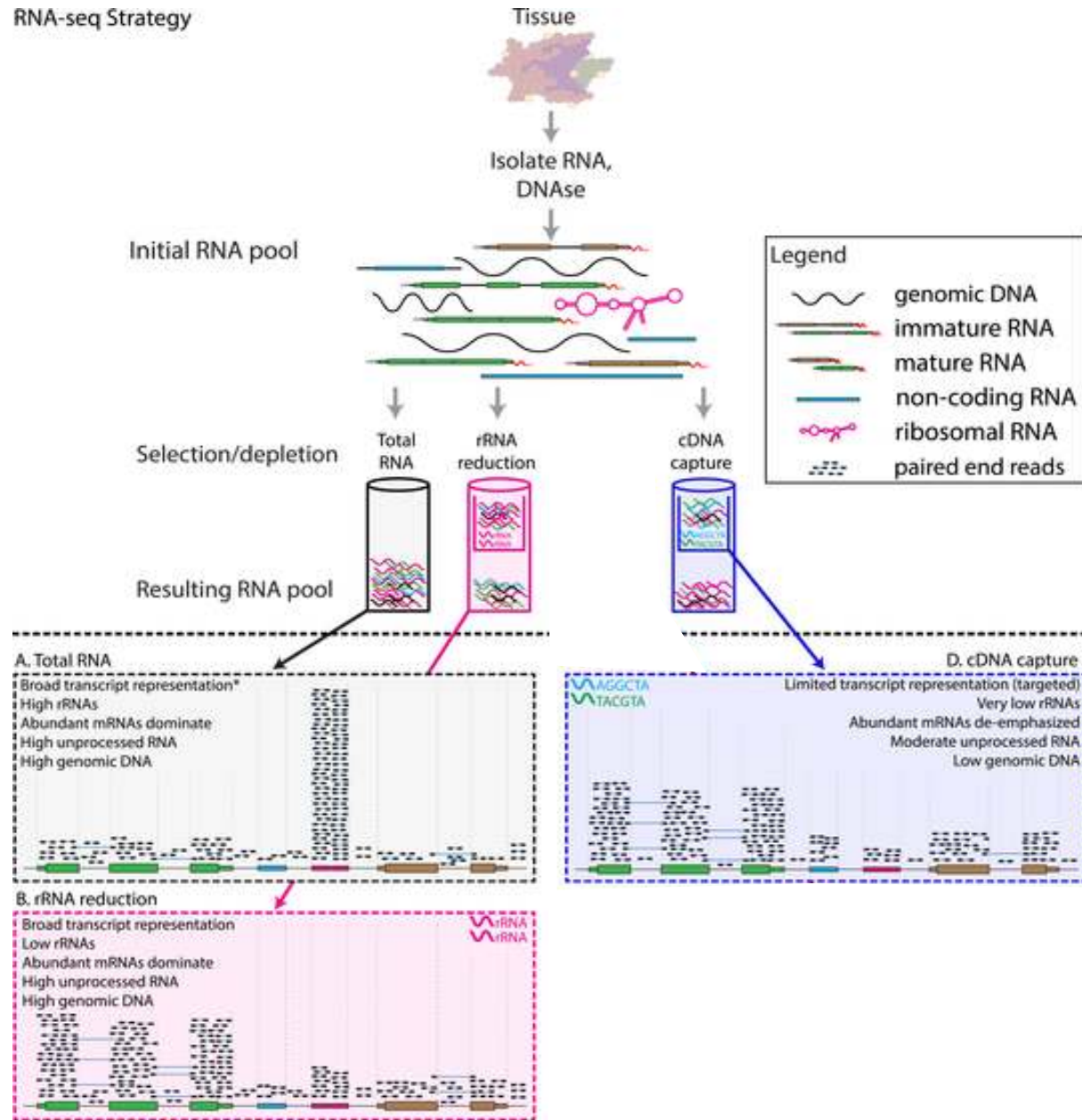
RNA-seq Strategy



Expected Alignments

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

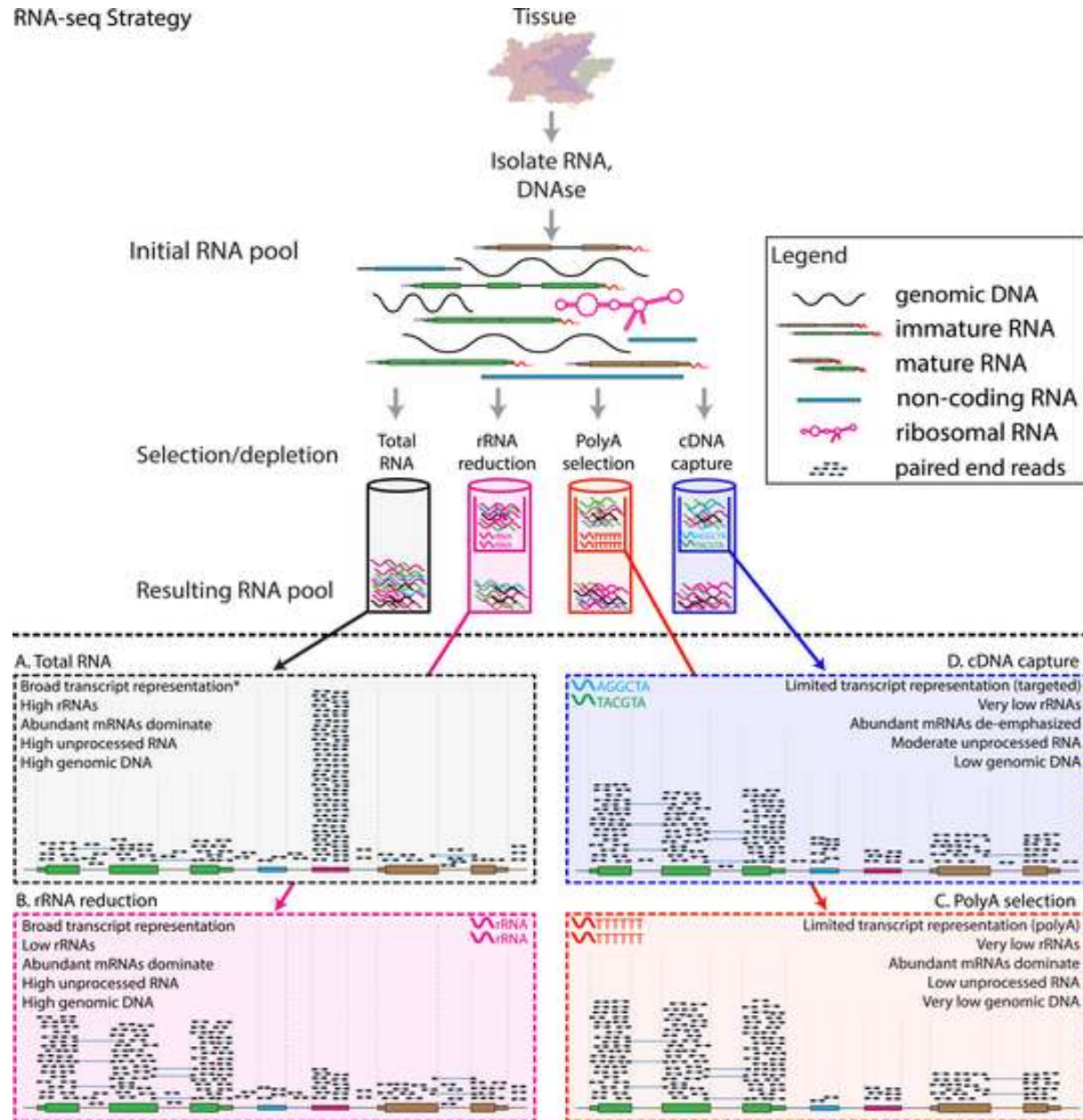
RNA-seq Strategy



Expected Alignments

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

RNA-seq Strategy



Expected Alignments

Generating RNA-Seq: *How to Choose?*

Many different instruments hit the scene in the last decade



Illumina



454



SOLiD



Helicos



Ion Torrent

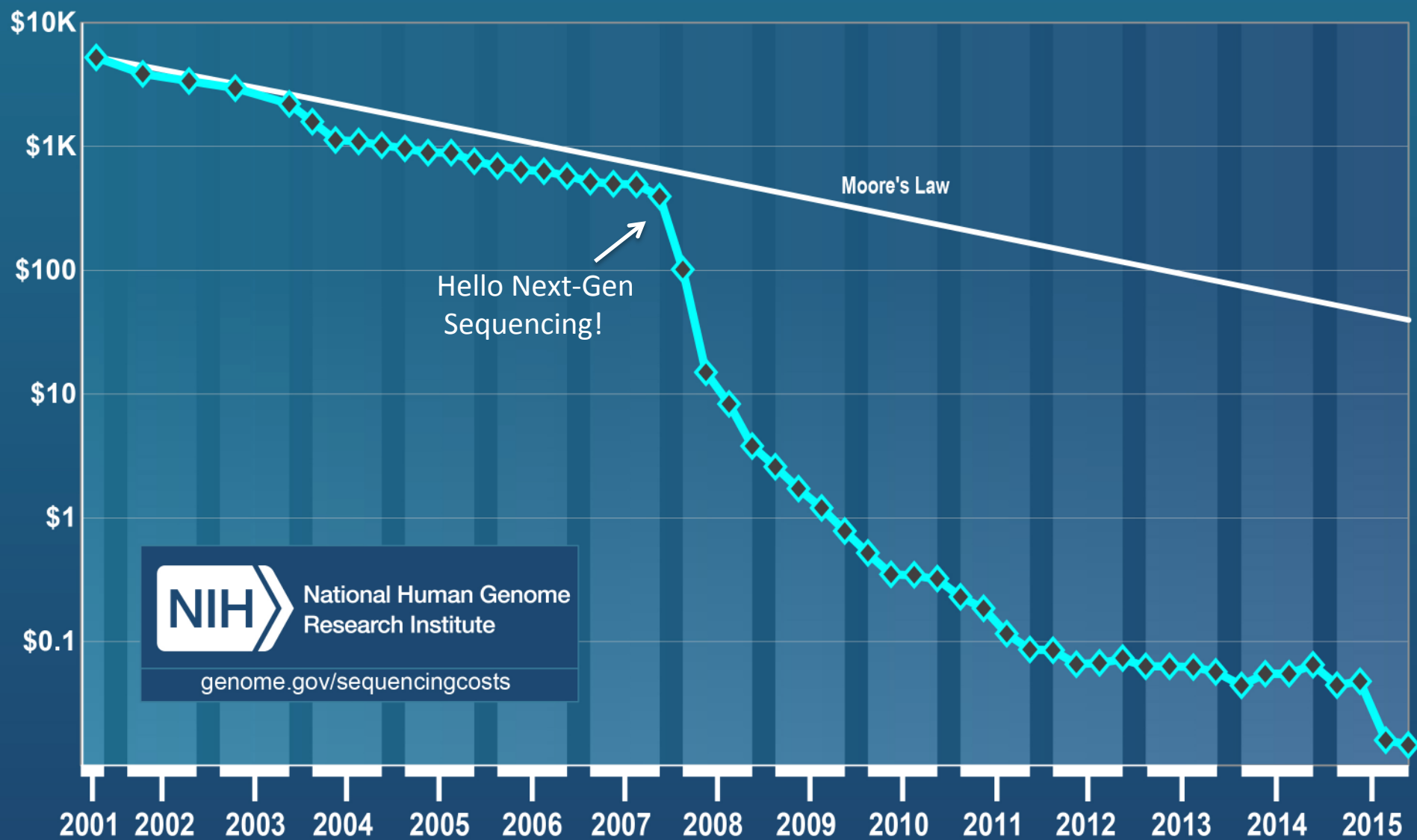


Pacific Biosciences



Oxford Nanopore

Cost per Raw Megabase of DNA Sequence



National Human Genome
Research Institute

genome.gov/sequencingcosts

From <https://www.genome.gov/sequencingcostsdata/>

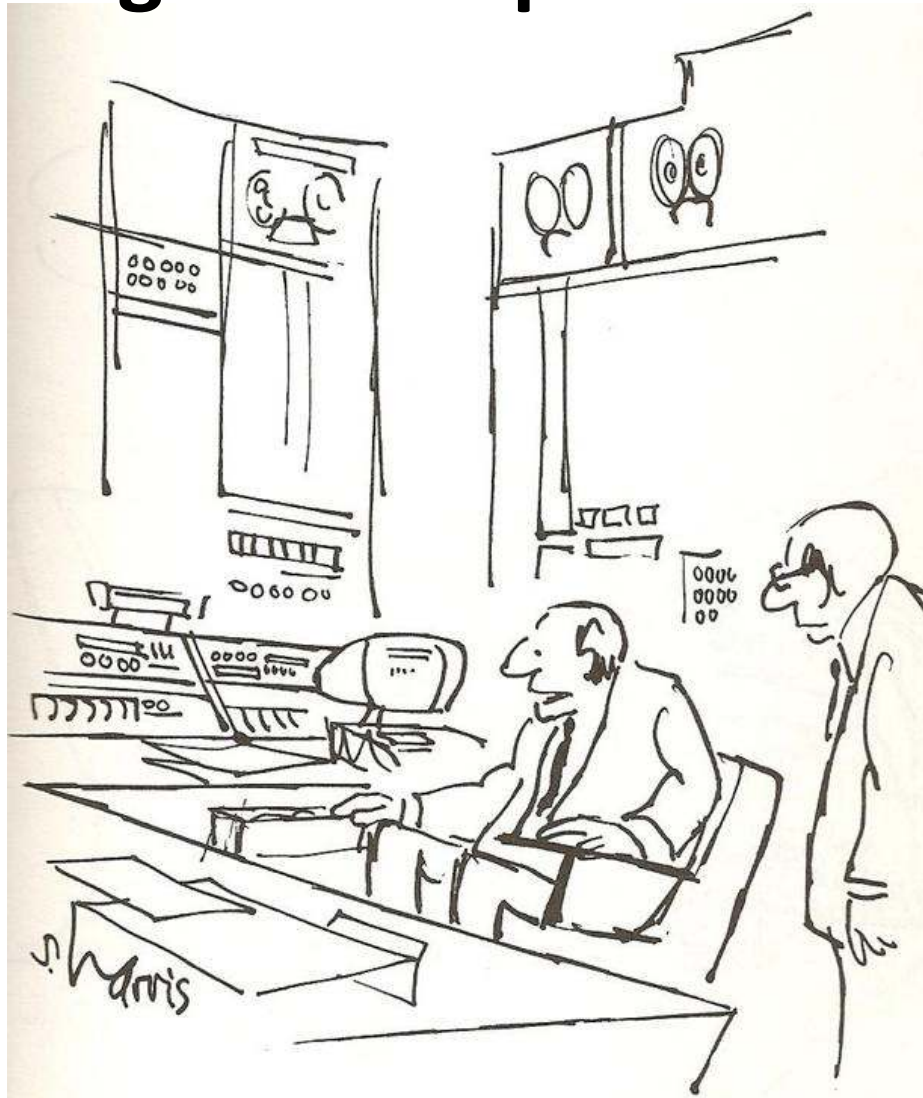
Generating RNA-Seq: *How to Choose?*



Illumina



Ion Torrent



“What I especially like about this baby is this little drawer where I can keep my lunch.”



Helicos



Oxford Nanopore

Generating RNA-Seq: *How to Choose?*

Popular choices for RNA-Seq today



Illumina



454



SOLiD



Helicos



Ion Torrent



Pacific Biosciences



Oxford Nanopore

Generating RNA-Seq: *How to Choose?*

Popular choices for RNA-Seq today

[Current RNA-Seq
workhorse]



Illumina



[Full-length single
molecule sequencing]



Pacific Biosciences

[Newly emerging
technology for full-length
single molecule sequencing]

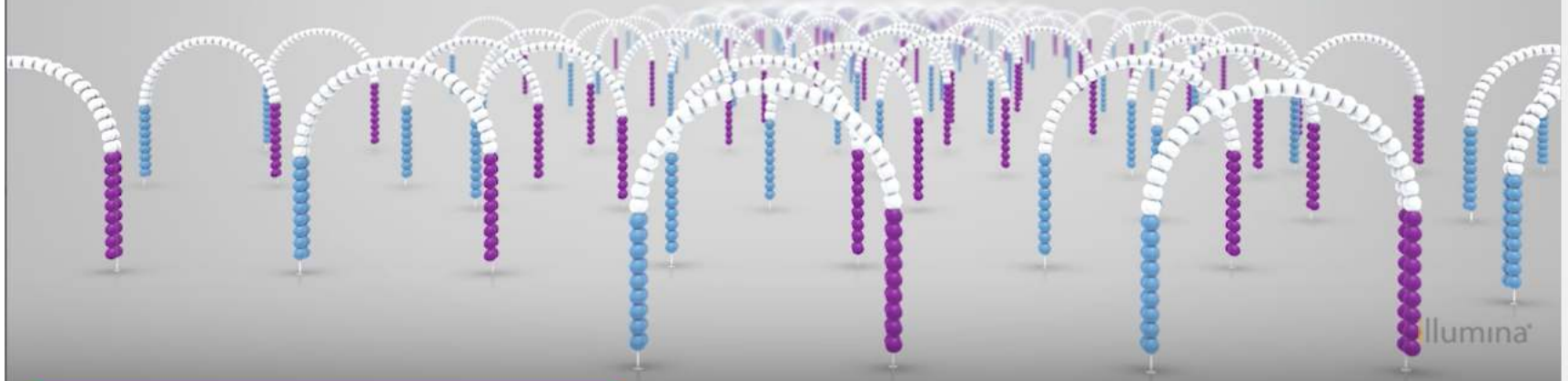


Oxford Nanopore



Ion Torrent

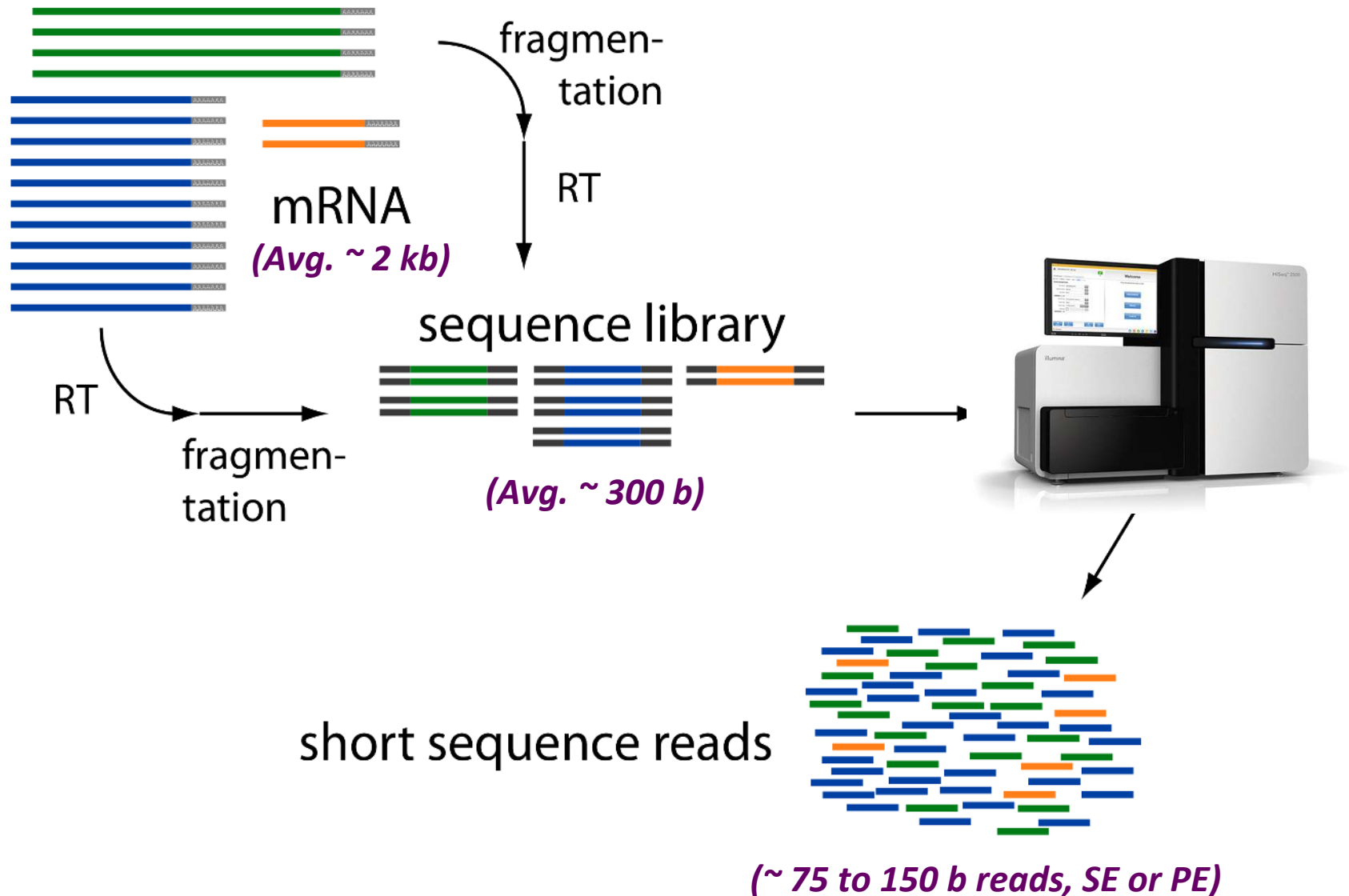
Cluster Generation



2:01 / 5:12



Millions to Billions of Reads



Common Data Formats for RNA-Seq

FASTA format:

```
>61DFRAAXX100204:1:100:10494:3070/1  
AAACAACAGGGGCACATTGTCACTCTTGTATTTGAAAAACACTTTCCGGCCAT
```

FASTQ format:

```
@61DFRAAXX100204:1:100:10494:3070/1  
AAACAACAGGGGCACATTGTCACTCTTGTATTTGAAAAACACTTTCCGGCCAT  
+  
ACCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCBC?CCCCCCCC@@@CACCCCCA
```

Read

Quality values

Interpreting Base Quality Values

<pre>@61DFRAAXX100204:1:100:10494:3070/1 AAACAACAGGGCACATTGTCACTCTTGTATTGAAAAACACTTTCCGGCCAT + ACCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCBC?CCCCCCCC@@@CACCCCCA</pre>	Read Quality values
	
AsciiEncodedQual ('B') = 63	

$$\text{Phred_Quality_Value} = \text{AsciiEncodedQual}('B') - 33 = 30$$

$$\text{Phred_Quality_Value} = -10 * \log_{10}(\text{Pwrong}('T'))$$

$$\text{Pwrong}('T') = 10^{(30/-10)} = 10^{-3} = 0.001$$

Paired-end Sequences



Two FastQ files, read name indicates left (/1) or right (/2) read of paired-end

```
@61DFRAAXX100204:1:100:10494:3070/1
AAACAACAGGGCACATTGTCACTCTTGTATTTGAAAAACACTTTCCGGCCAT
+
ACCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCBC?CCCCCCCC@@@CACCCCCA
```

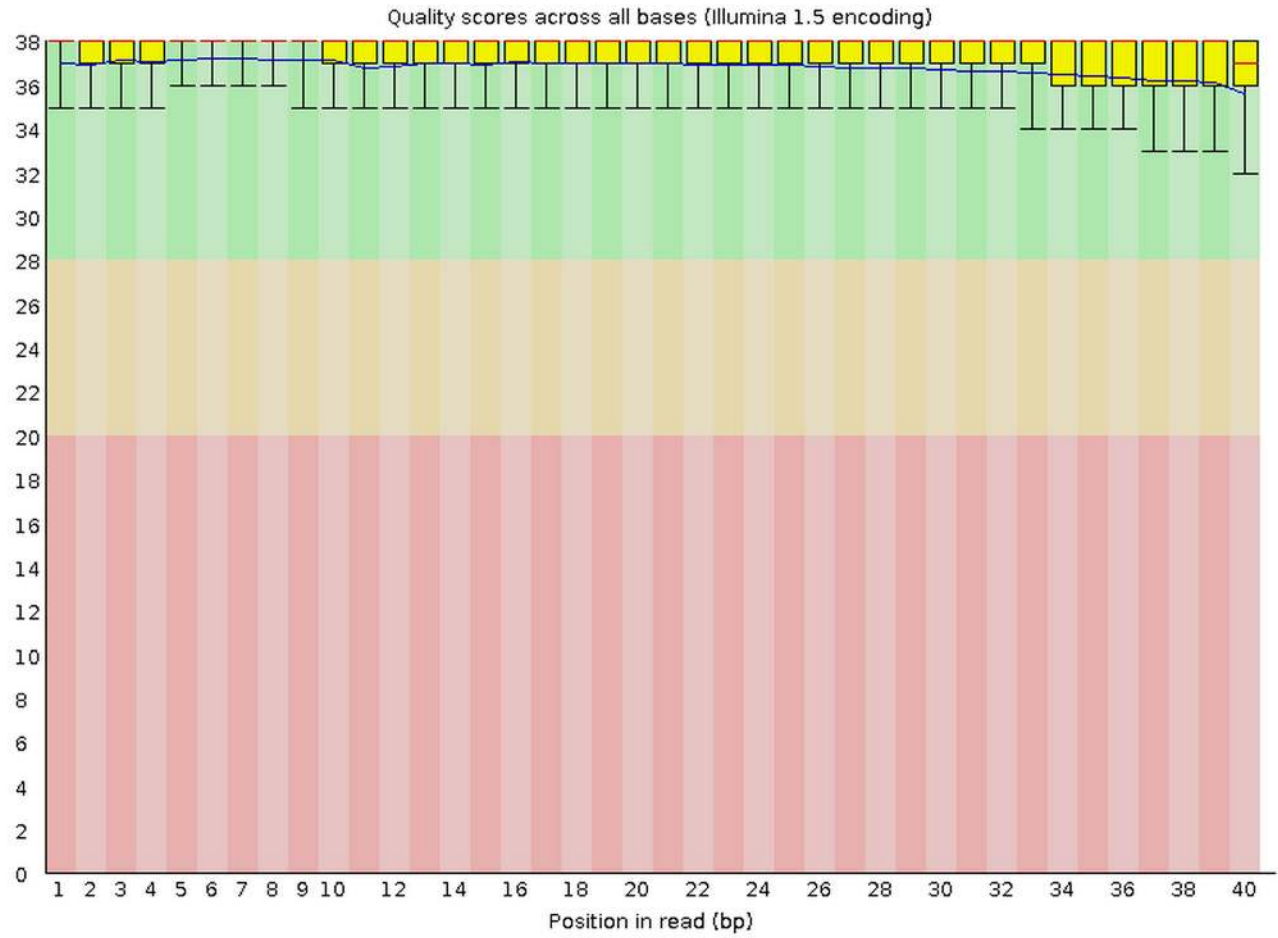
```
@61DFRAAXX100204:1:100:10494:3070/2
CTCAAATGGTTAATTCTCAGGCTGCAAATATTCGTTTCAGGATGGAAGAACA
+
C<CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCBCCCC
```

FastQC Report

Summary

- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per tile sequence quality
- ✓ Per sequence quality scores
- ! Per base sequence content
- ✓ Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ✓ Sequence Duplication Levels
- ✓ Overrepresented sequences
- ✓ Adapter Content
- ! Kmer Content

✓ Per base sequence quality



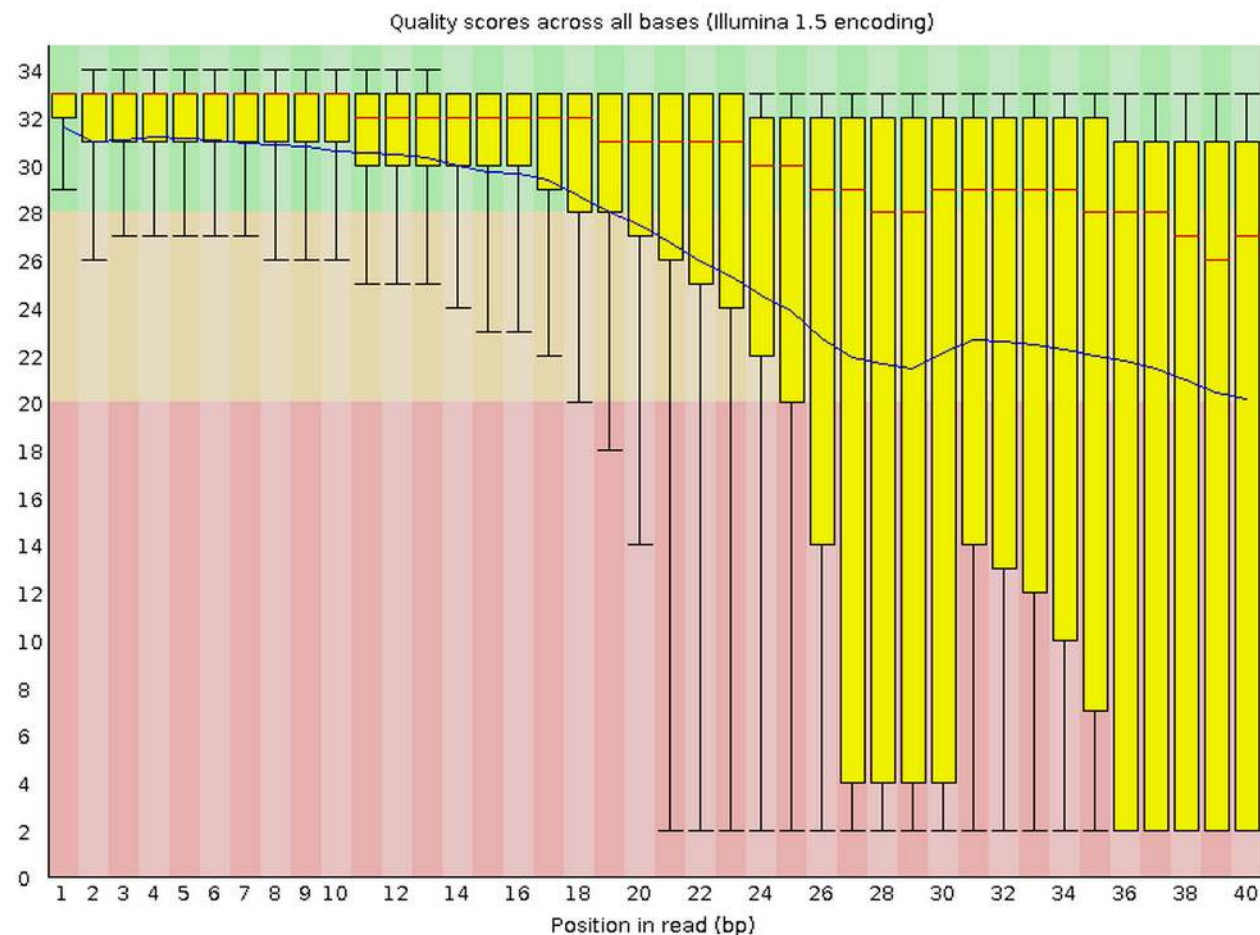
FastQC Report

Wed 25 Mar 2015
bad_sequence.txt

Summary

- ✓ Basic Statistics
- ✗ Per base sequence quality
- ✗ Per tile sequence quality
- ✓ Per sequence quality scores
- ! Per base sequence content
- ! Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ! Sequence Duplication Levels
- ! Overrepresented sequences
- ✓ Adapter Content
- ! Kmer Content

✗ Per base sequence quality



What to do?

- Trim the reads?
- Start over – try sequencing it again?

Trimming low quality regions of reads: Trimmomatic



www.usadellab.org/cms/?page=trimmomatic



USADELLAB.org

Home

Research

Education

Service & Software

Publications

Supporting Info

About Us

NGS, DE and other things

Trimmomatic: A flexible read trimming tool for Illumina NGS data

Citations

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*, btu170.

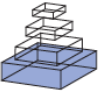
Downloading Trimmomatic

Version 0.36: [binary](#), [source](#) and [manual](#)

Quick start

Paired End:

```
java -jar trimmomatic-0.35.jar PE -phred33 input_forward.fq.gz input_reverse.fq.gz
output_forward_paired.fq.gz output_forward_unpaired.fq.gz
output_reverse_paired.fq.gz output_reverse_unpaired.fq.gz ILLUMINACLIP:TruSeq3-
PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36
```



On the optimal trimming of high-throughput mRNA sequence data

Matthew D. MacManes^{1,2*}

¹ Department of Molecular, Cellular and Biomedical Sciences, University of New Hampshire, Durham, NH, USA

² Hubbard Center for Genome Studies, Durham, NH, USA

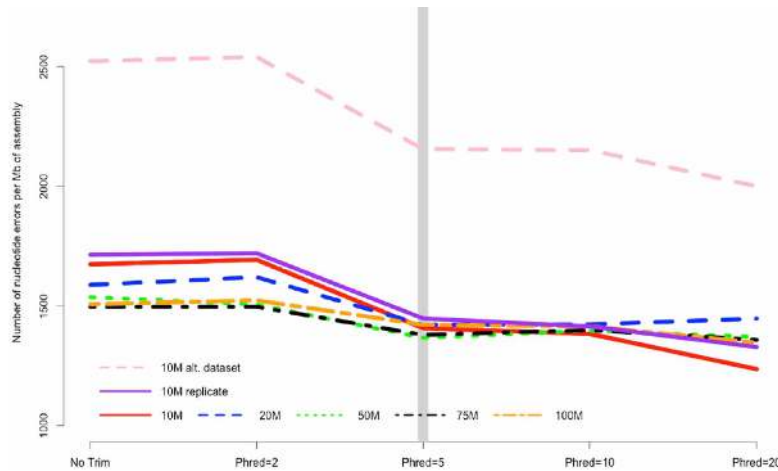
“... researchers interested in assembling transcriptomes *de novo* should elect for a much gentler quality trimming, or no trimming at all.”

“... trimming at PHRED=2 or PHRED=5 optimizes assembly quality.”

Aggressive Trimming may be harmful, whereas light trimming could be beneficial

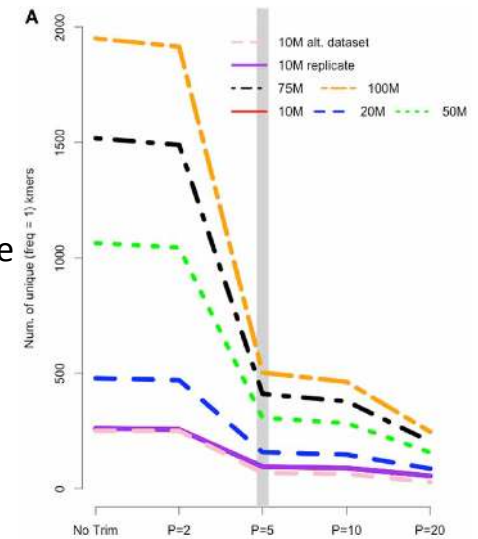
Fewer errors in the assembly

Nucleotide errors / Mb



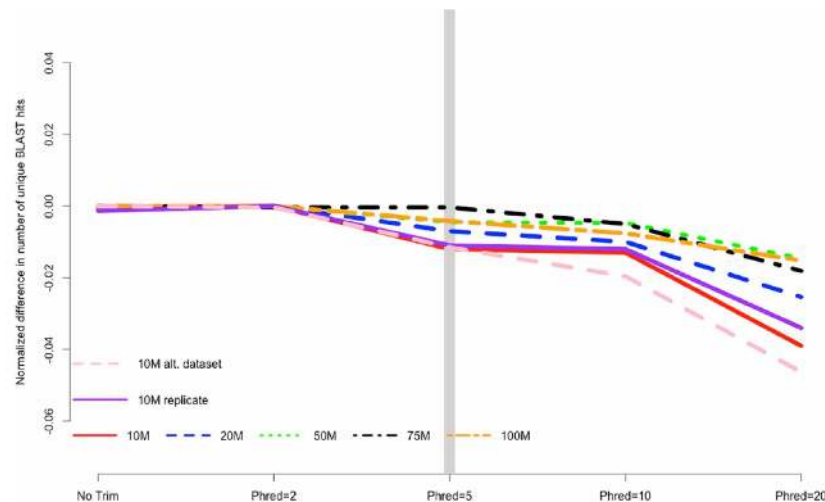
Fewer unique kmers

unique kmers



Light trimming doesn't reduce number of blast matches w/ higher sequencing depths.

Normalized # of blast matches



MultiQC - aggregation across all QC on all samples



Aggregate results from bioinformatics analyses across many samples into a single report

MultiQC searches a given directory for analysis logs and compiles a HTML report. It's a general use tool, perfect for summarising the output from numerous bioinformatics tools.



Phil Ewels
phil.ewels@scilifelab.se

- Introduction to MultiQC (1:19)
- Installing MultiQC (4:33)
- Running MultiQC (5:21)
- Using MultiQC Reports (6:06)

Current version: v1.2

[Home](#) [Docs](#) [Plugins](#) [Logo](#) [Example Reports](#)

[GitHub](#)

[Python Package Index](#)

[Documentation](#)

[51 supported tools](#)

[Publication / Citation](#)

[Get help on Gitter](#)

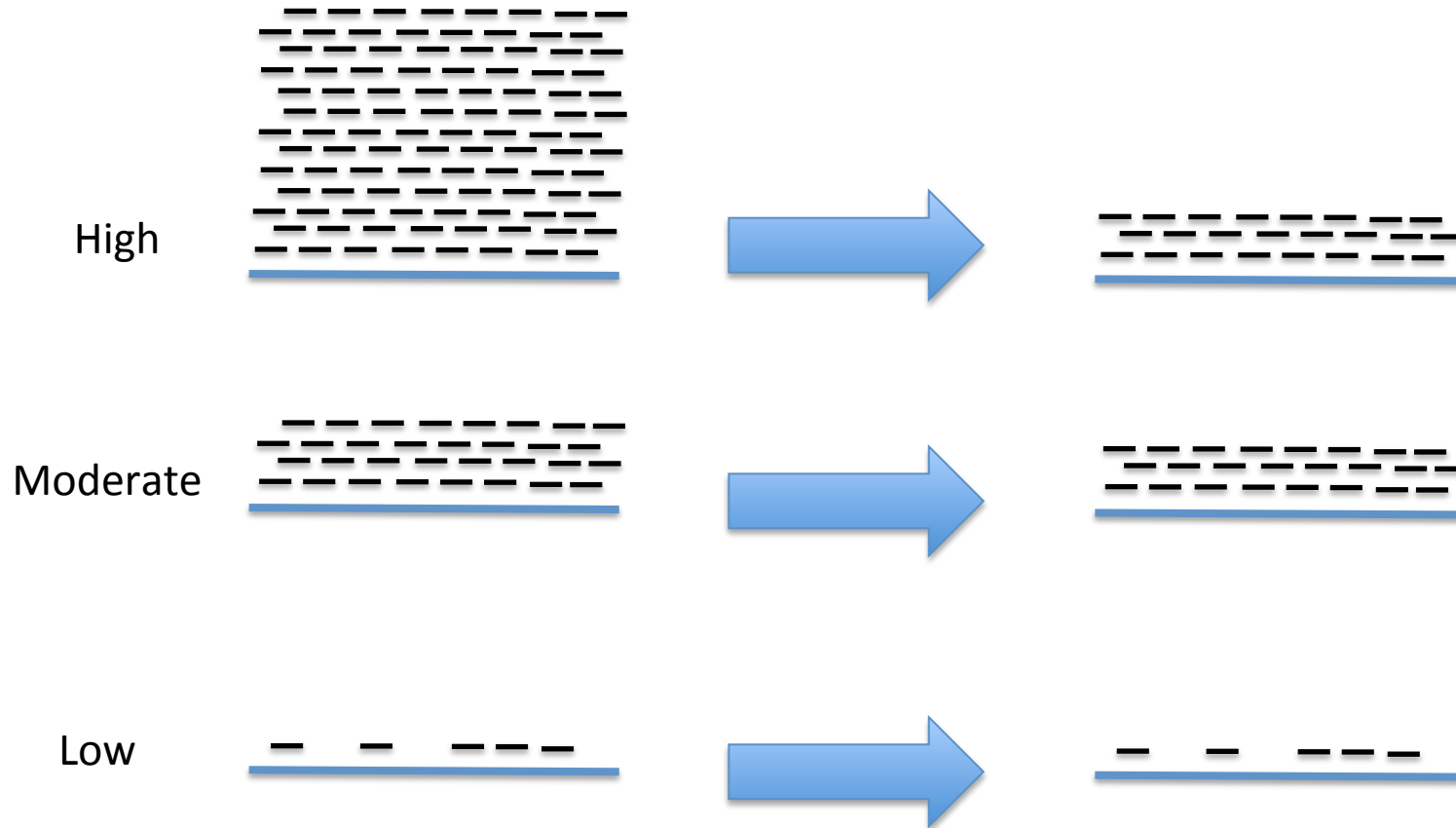
[Quick Install](#)

```
pip install multiqc      # Install
multiqc .                # Run
```

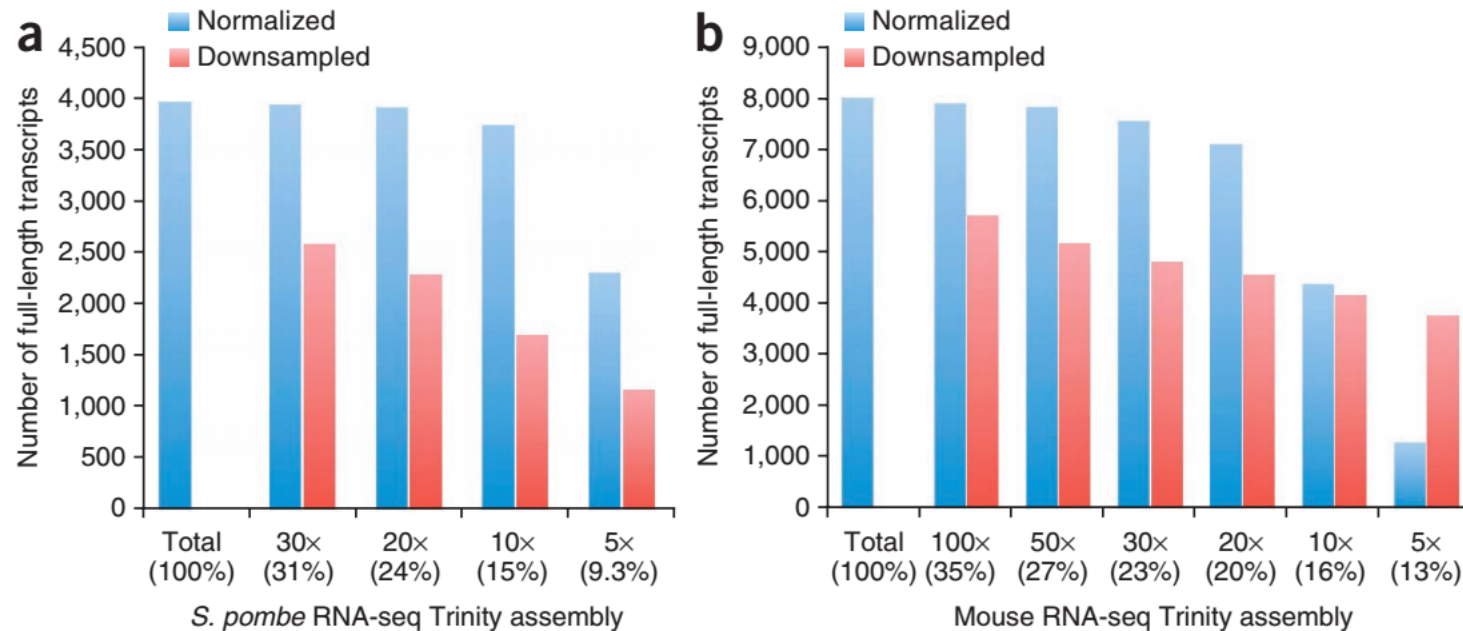
[pip](#) [conda](#) [manual](#)

Need a little more help? [See the full installation instructions.](#)

In silico normalization of reads

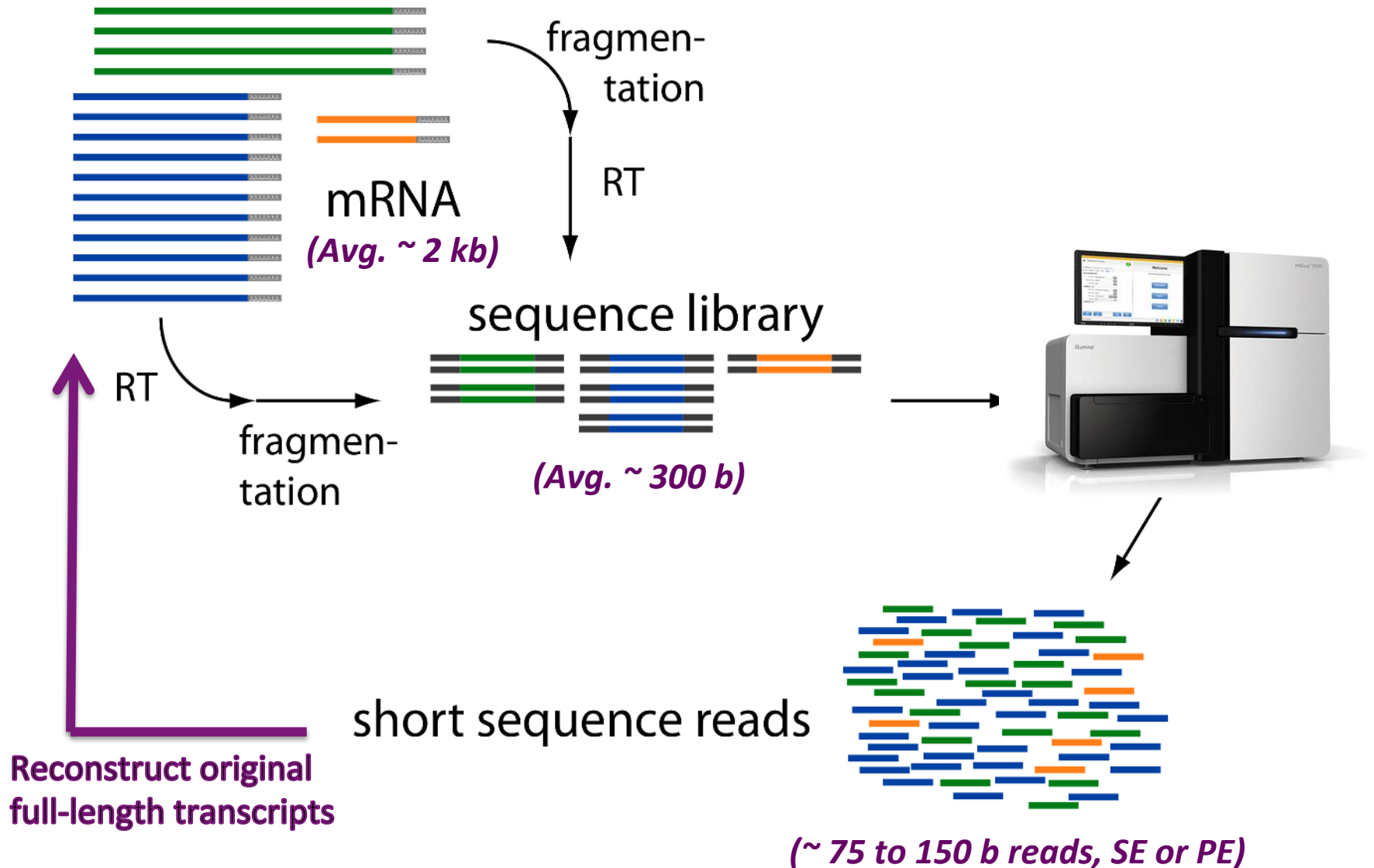


Impact of Normalization on *De novo* Full-length Transcript Reconstruction

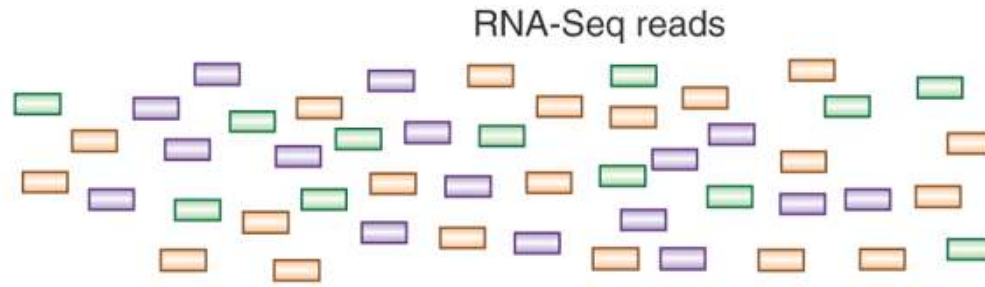


Largely retain full-length reconstruction, but use less RAM and assemble much faster.

RNA-Seq Challenge: Transcript Reconstruction



Transcript Reconstruction from RNA-Seq Reads



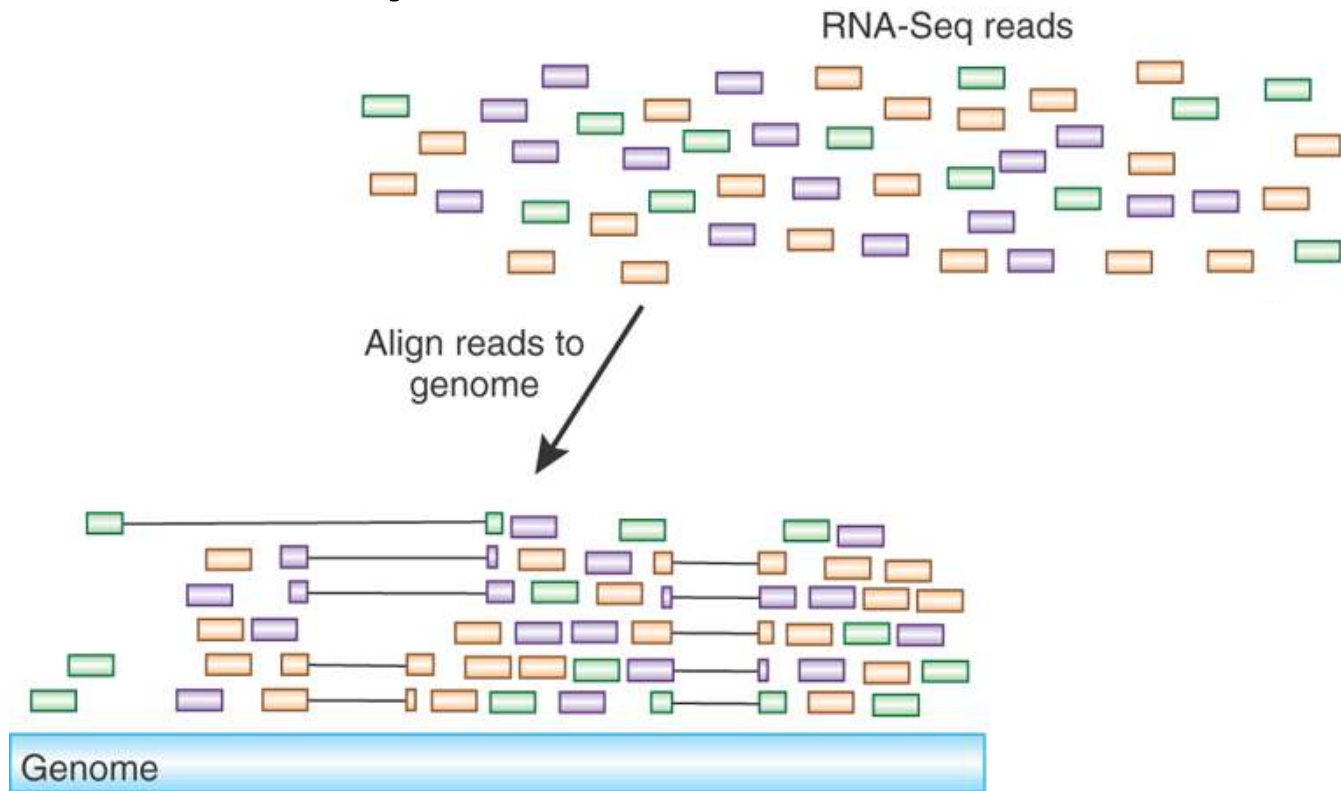
Advancing RNA-Seq analysis

Brian J Haas & Michael C Zody

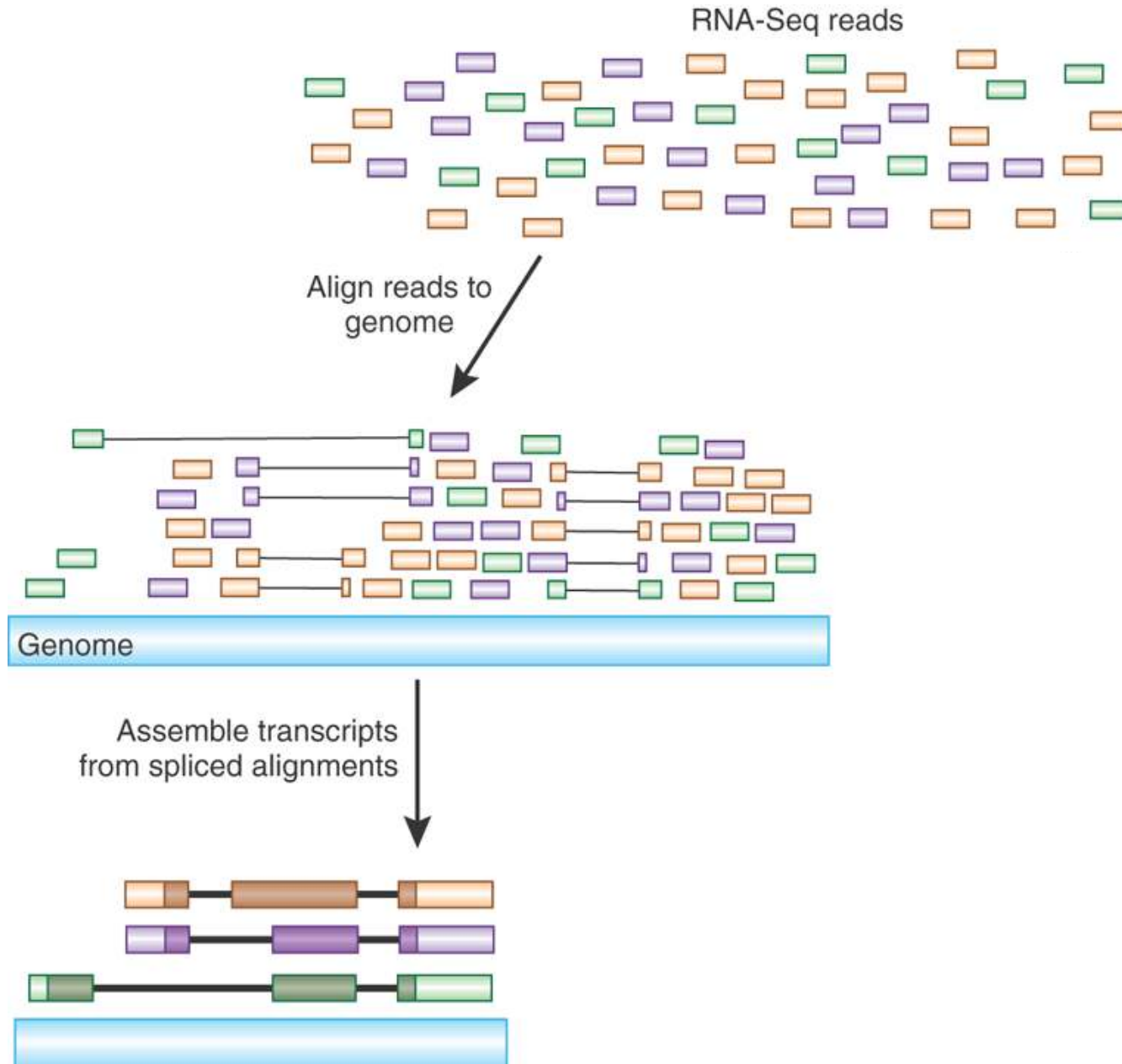
Nature Biotech, 2010

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

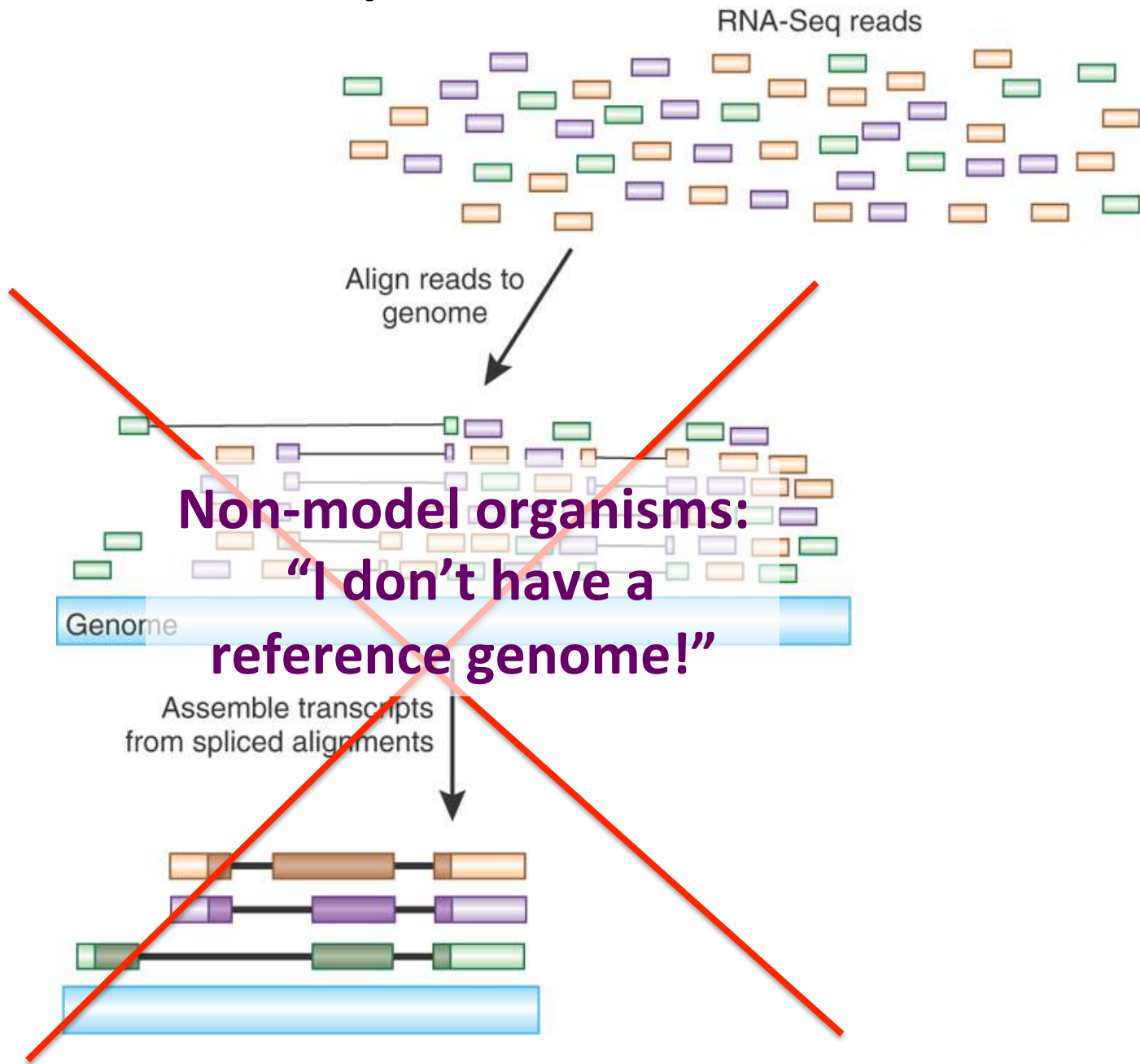
Transcript Reconstruction from RNA-Seq Reads



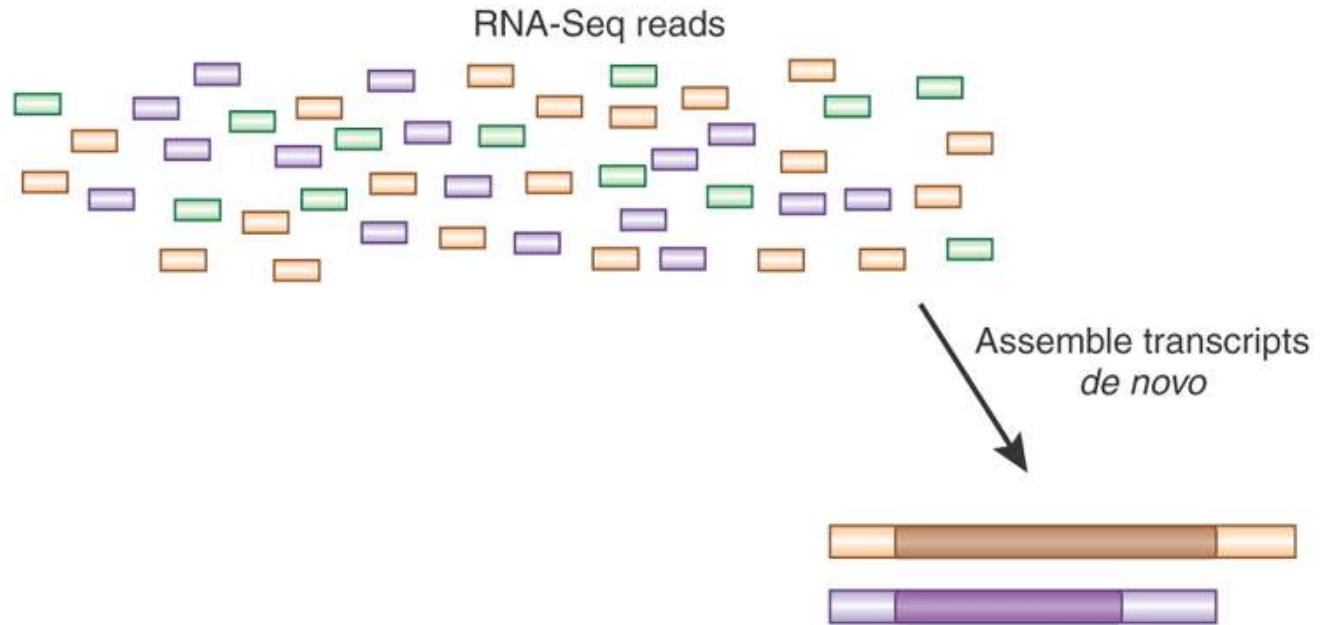
Transcript Reconstruction from RNA-Seq Reads



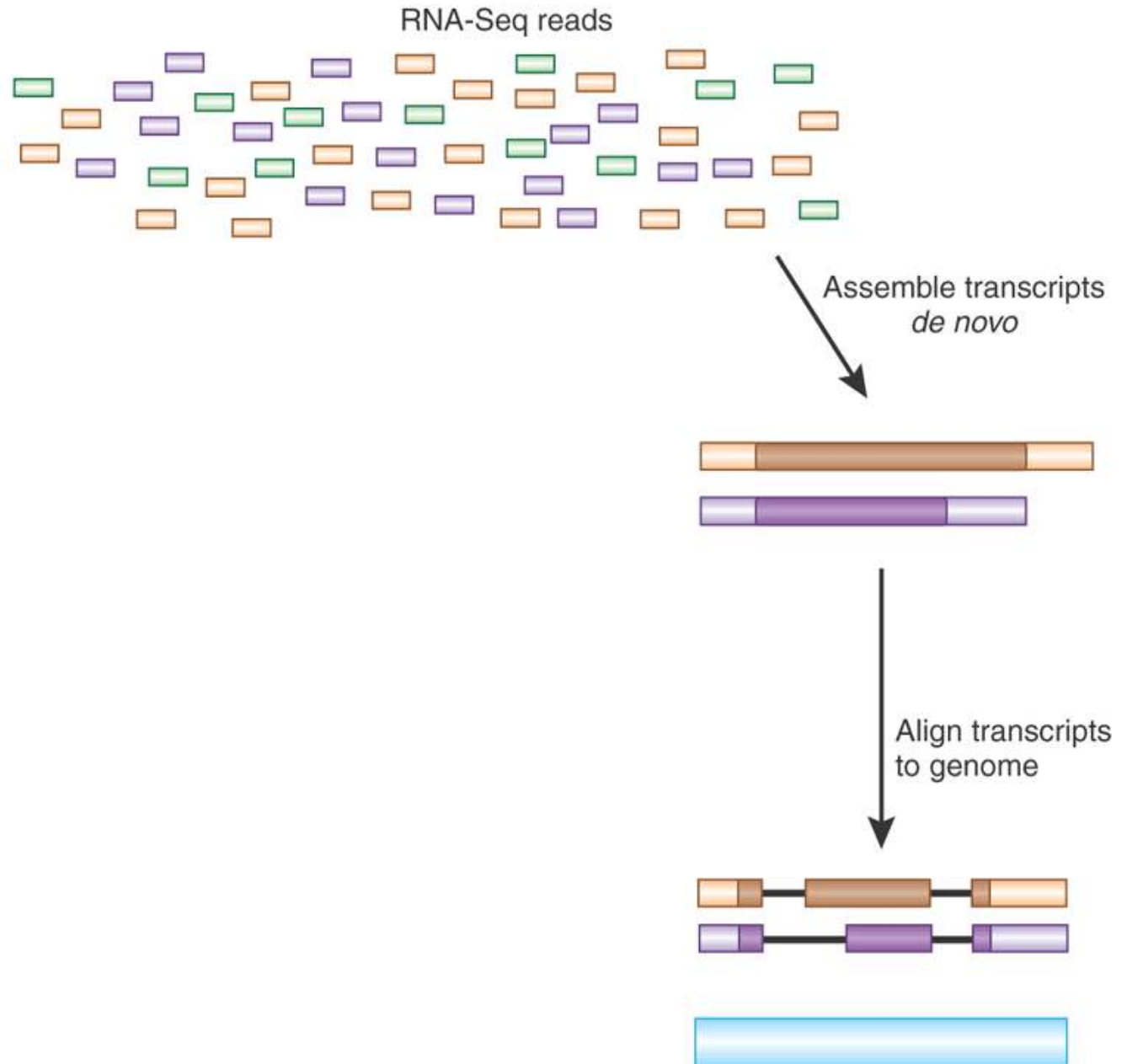
Transcript Reconstruction from RNA-Seq Reads



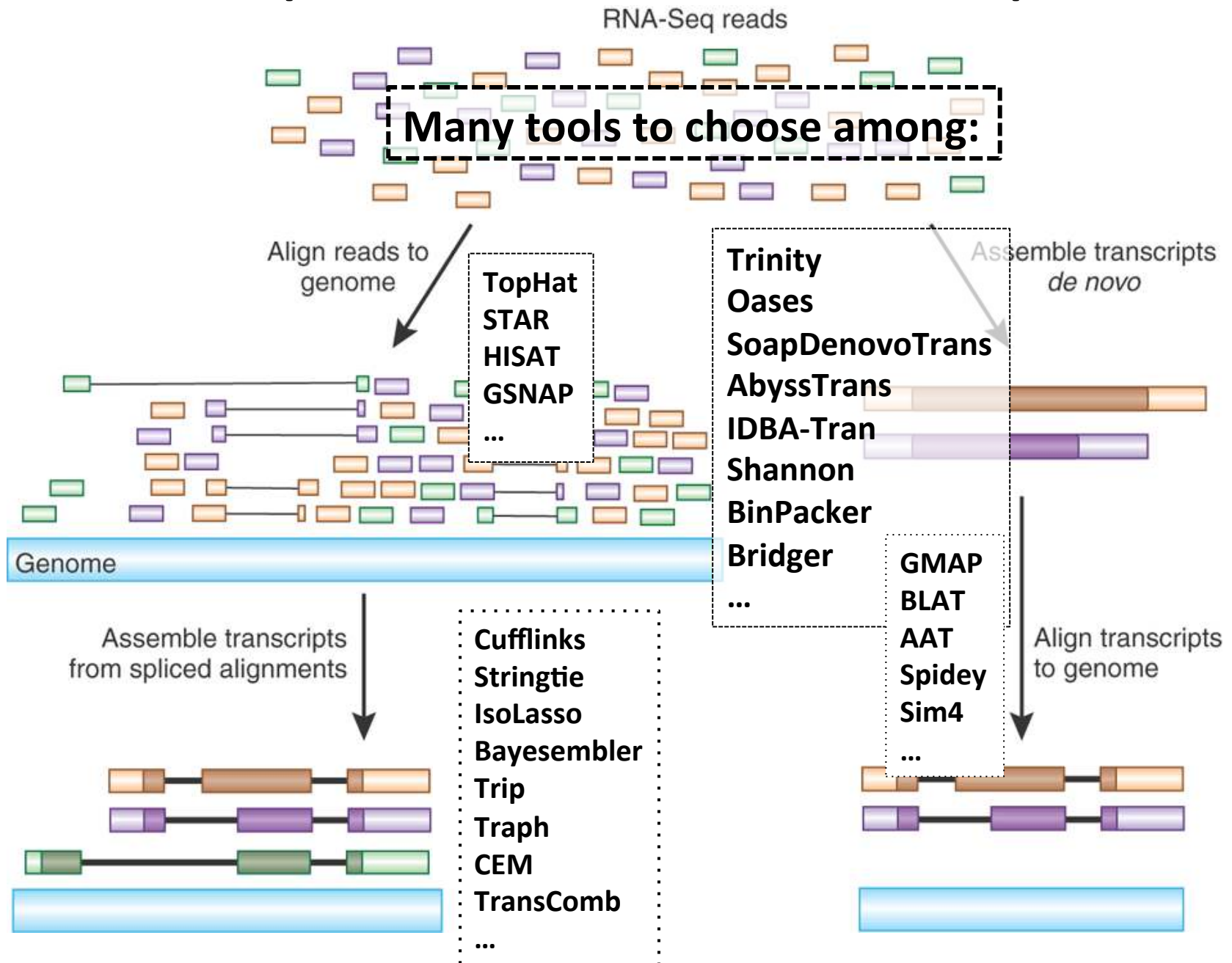
Transcript Reconstruction from RNA-Seq Reads



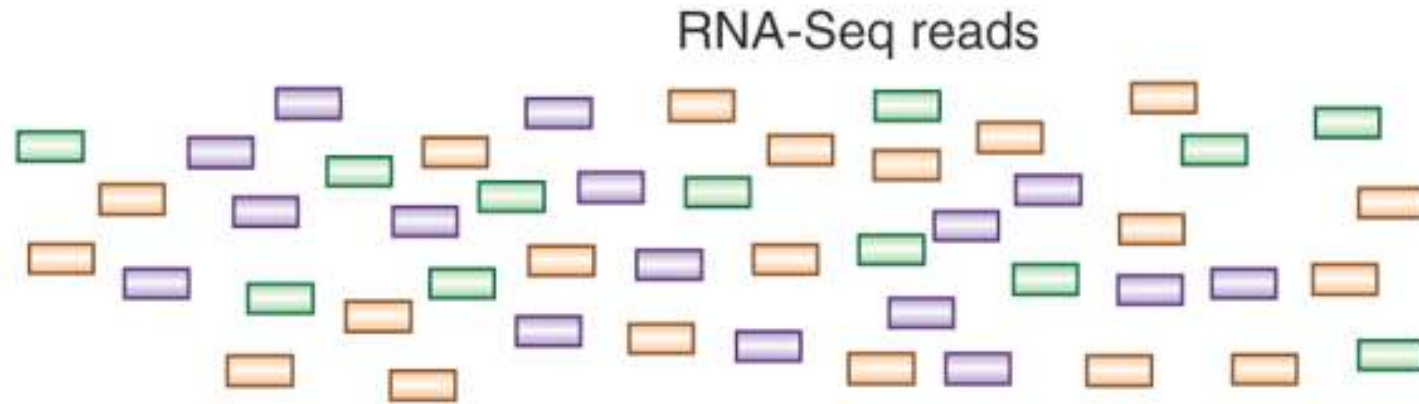
Transcript Reconstruction from RNA-Seq Reads



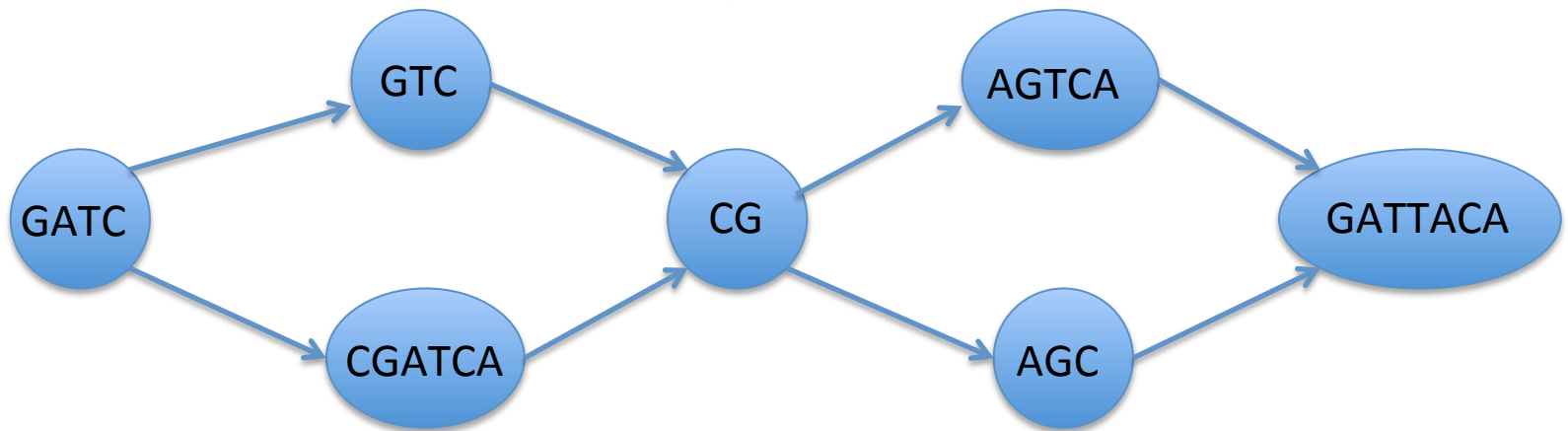
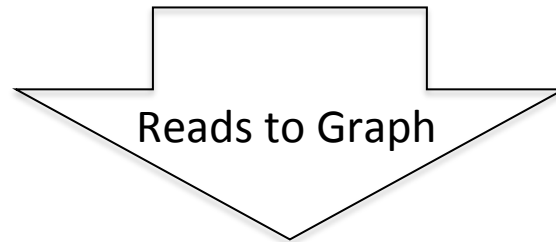
Transcript Reconstruction from RNA-Seq Reads



Graph Data Structures Commonly Used For Assembly

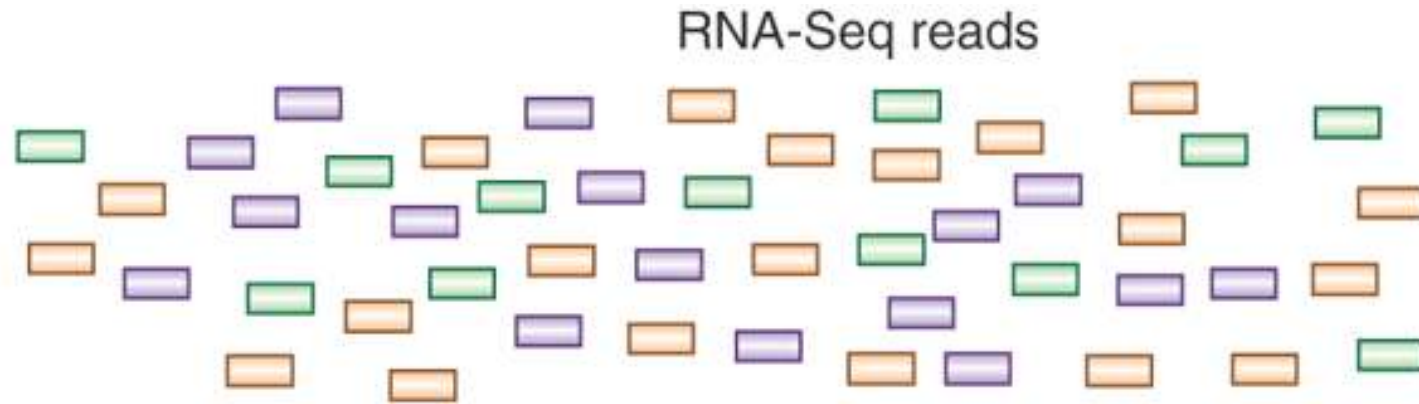


- Sequence
- Order
- Orientation (+, -)
- Overlap

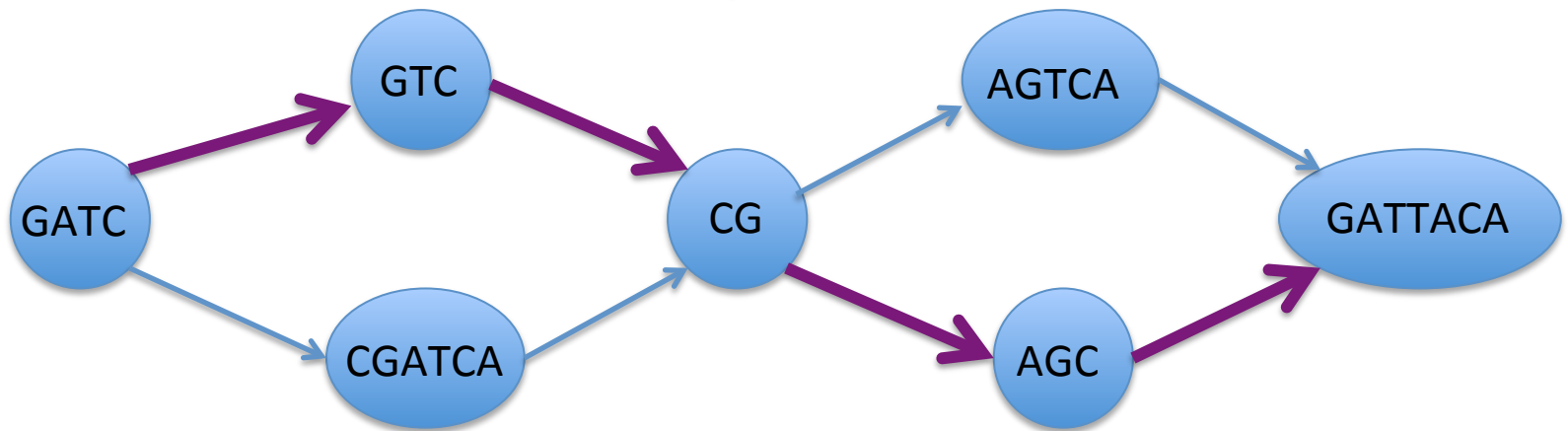
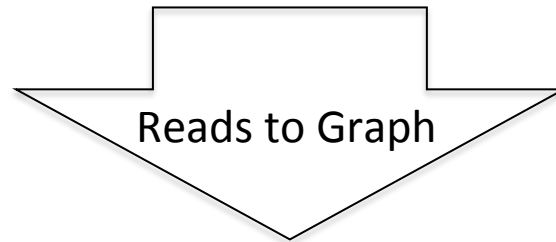


Nodes = sequence (+/-)
Edges = order, overlap

Graph Data Structures Commonly Used For Assembly



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

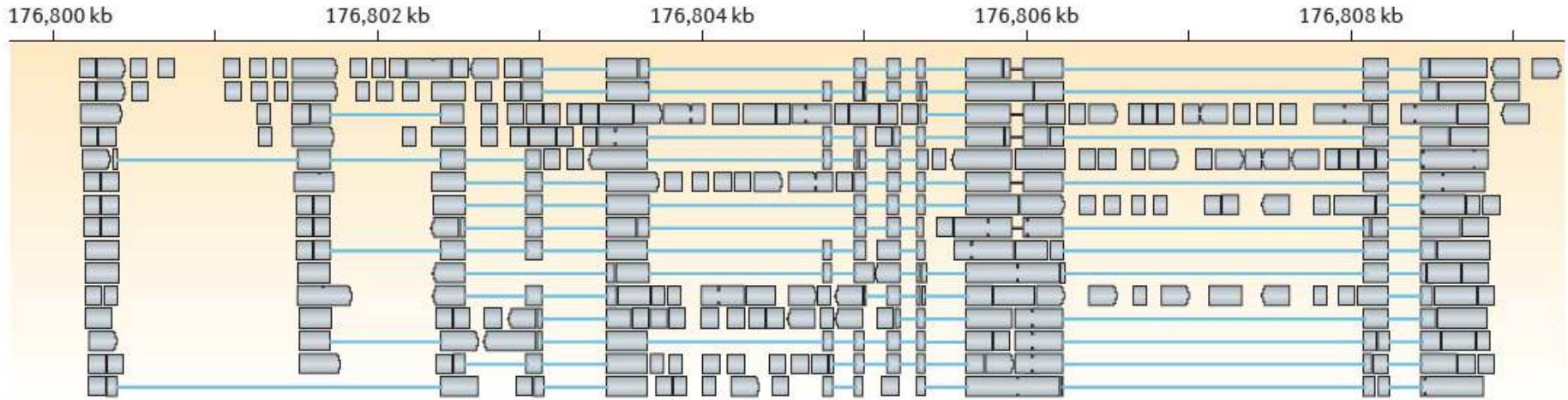


GATCGTCCGAGCGATTACA

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

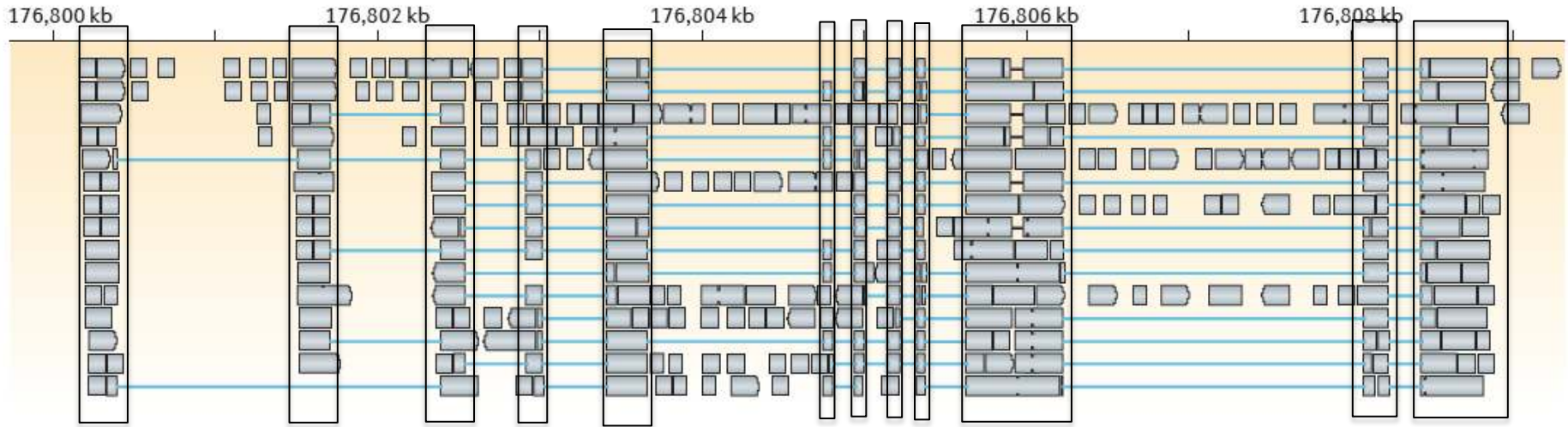
Genome-Guided Transcript Reconstruction

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

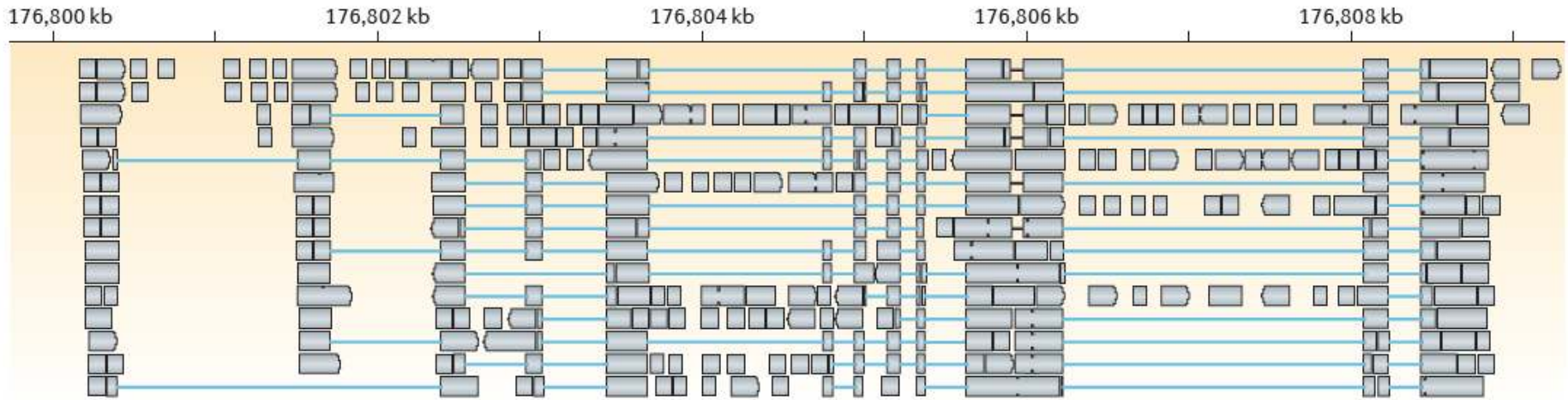
Splice-align reads to the genome



Overlapping but different introns = evidence of alternative splicing

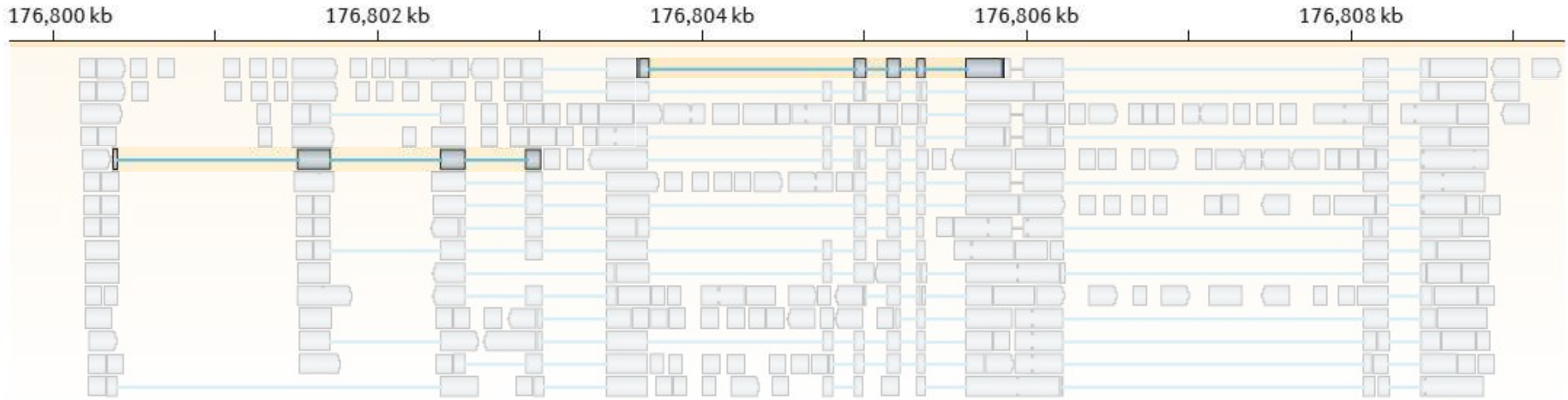
Genome-Guided Transcript Reconstruction

Splice-align reads to the genome



Genome-Guided Transcript Reconstruction

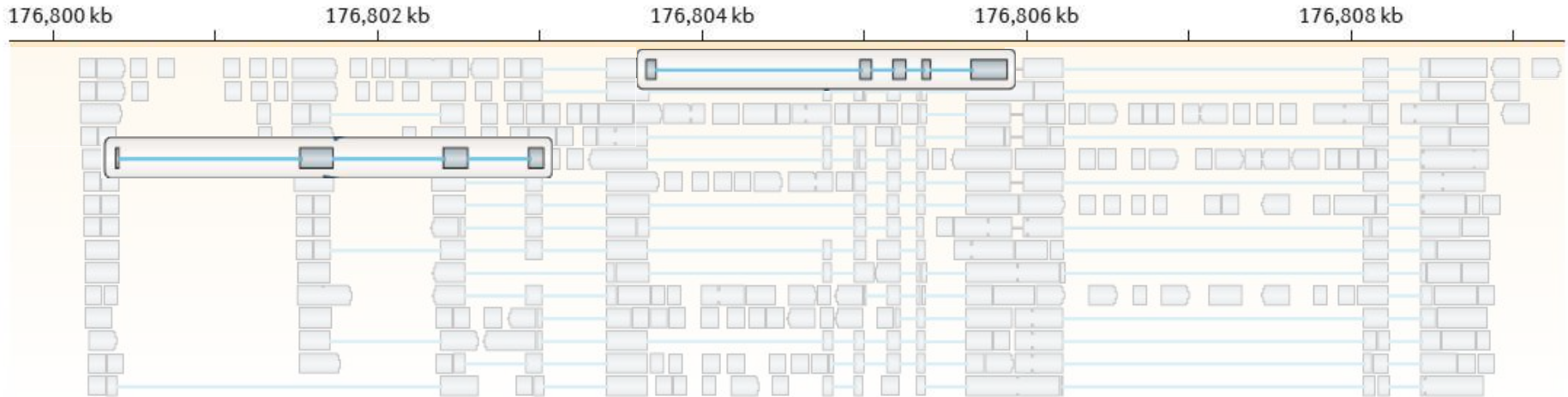
Splice-align reads to the genome



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

Genome-Guided Transcript Reconstruction

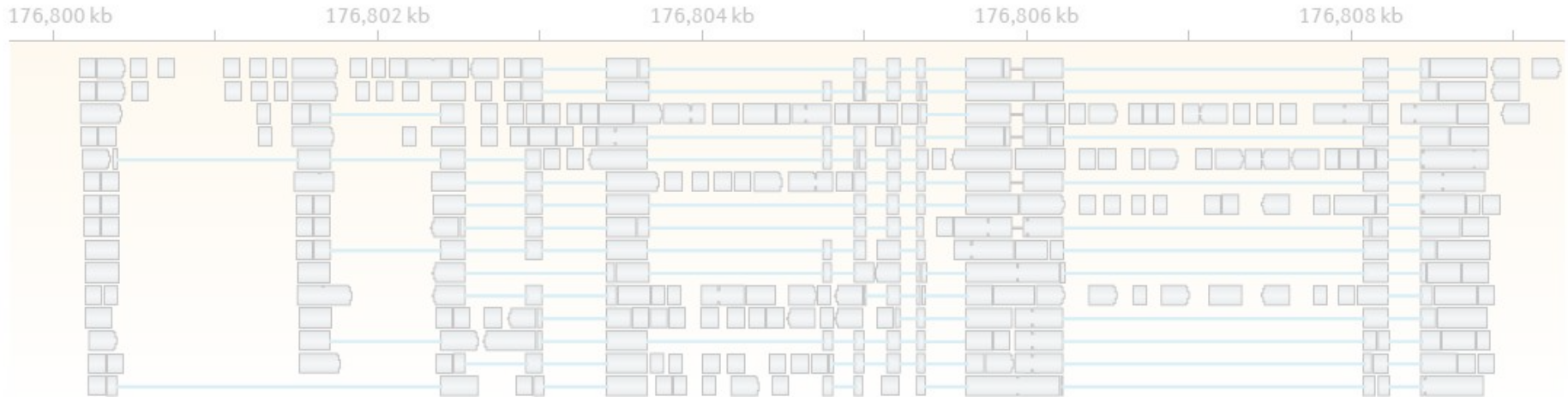
Splice-align reads to the genome



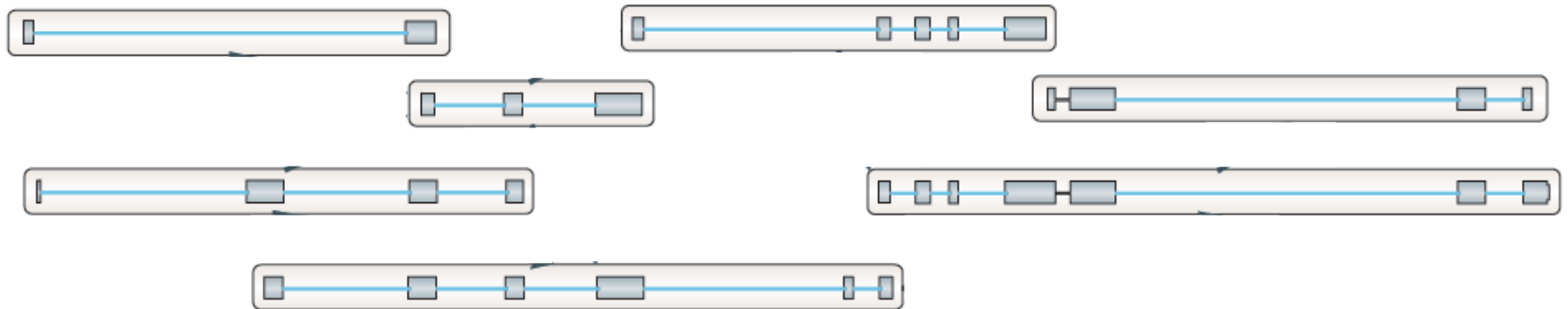
Nodes = unique splice patterns

Genome-Guided Transcript Reconstruction

Splice-align reads to the genome



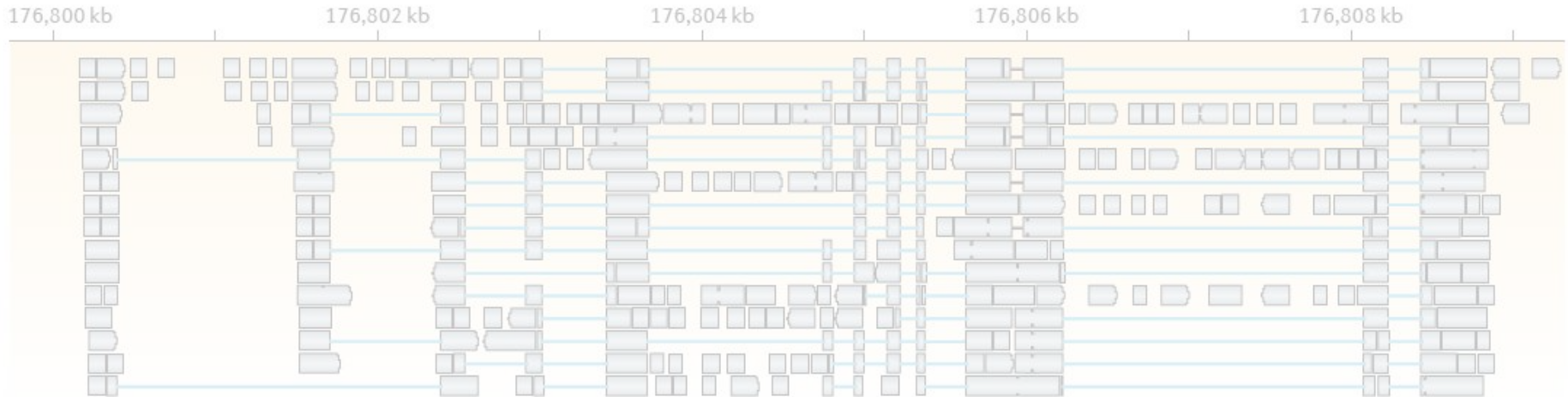
Construct graph from unique splice patterns of aligned reads.



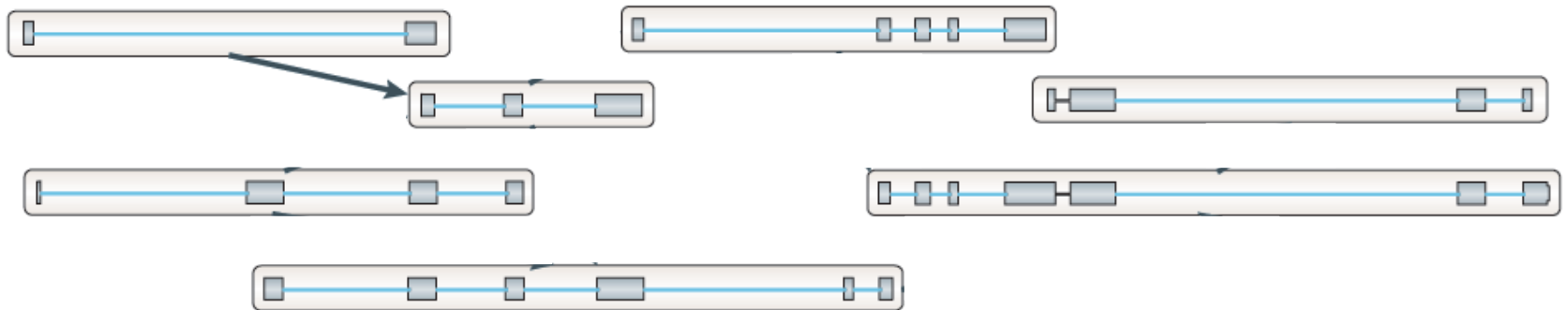
Nodes = unique splice patterns

Genome-Guided Transcript Reconstruction

Splice-align reads to the genome



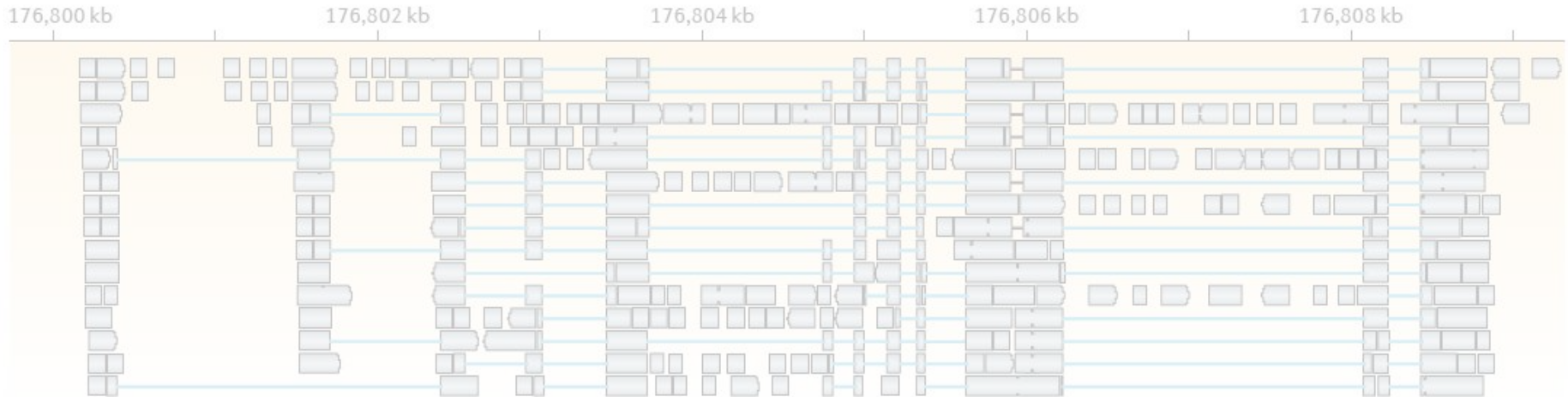
Construct graph from unique splice patterns of aligned reads.



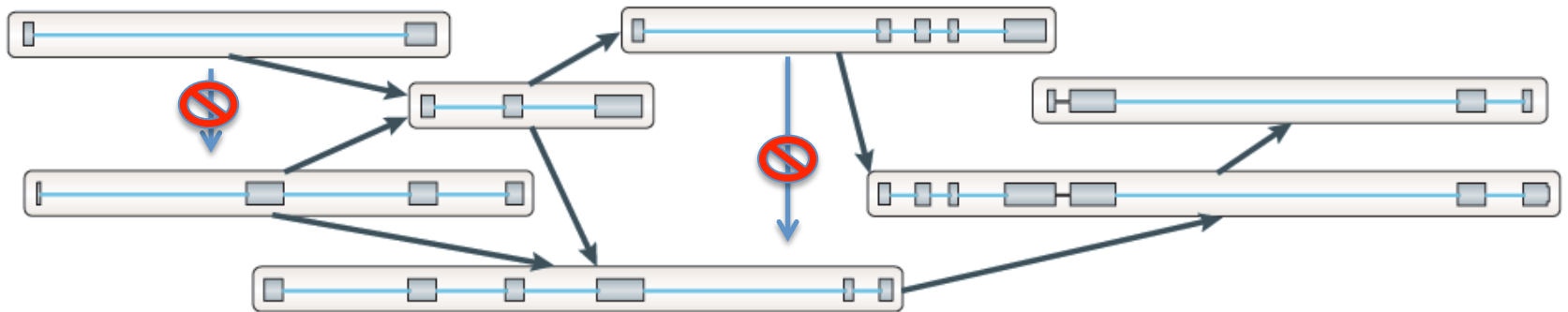
Nodes = unique splice patterns
Edges = compatible patterns

Genome-Guided Transcript Reconstruction

Splice-align reads to the genome



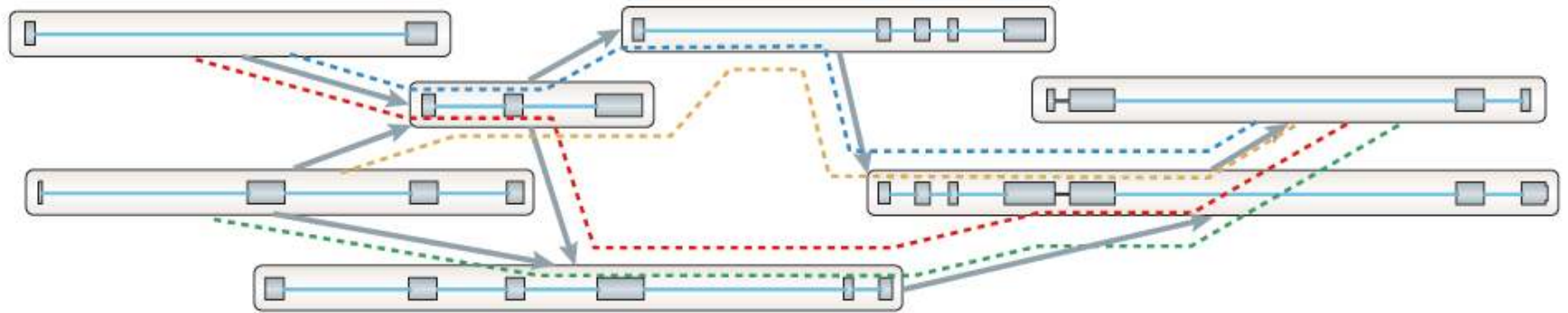
Construct graph from unique splice patterns of aligned reads.



Nodes = unique splice patterns
Edges = compatible patterns

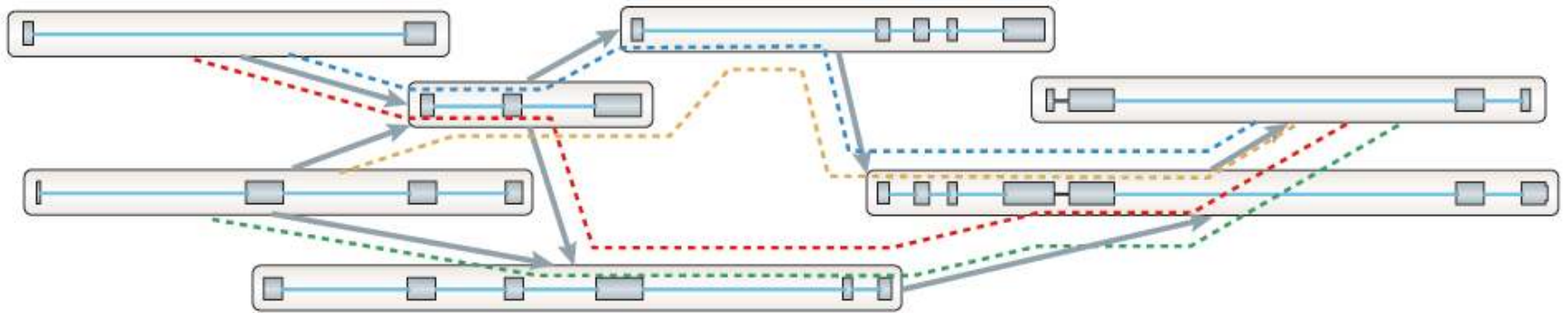
Genome-Guided Transcript Reconstruction

Traverse paths through the graph to assemble transcript isoforms

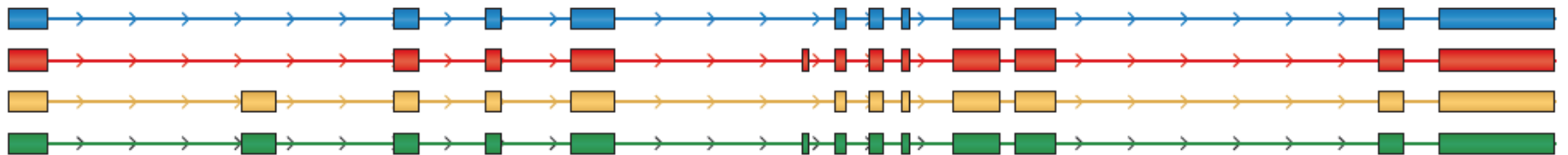


Genome-Guided Transcript Reconstruction

Traverse paths through the graph to assemble transcript isoforms

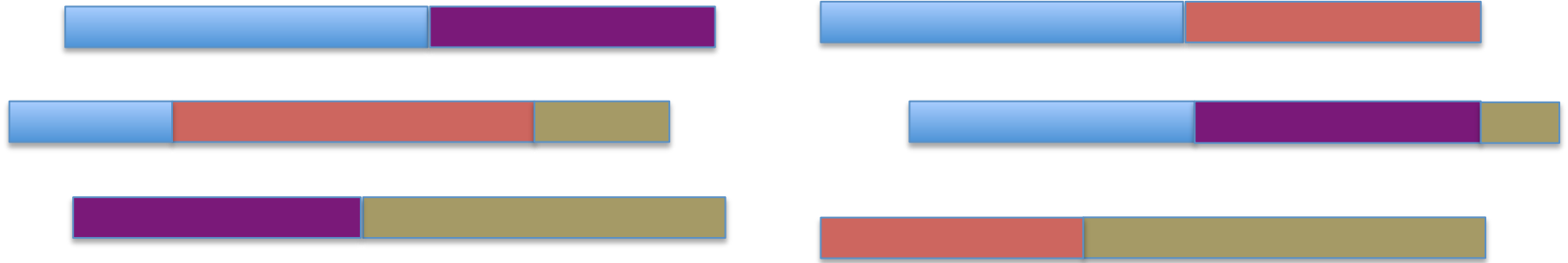


Reconstructed isoforms



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

Read Overlap Graph: Reads as nodes, overlaps as edges

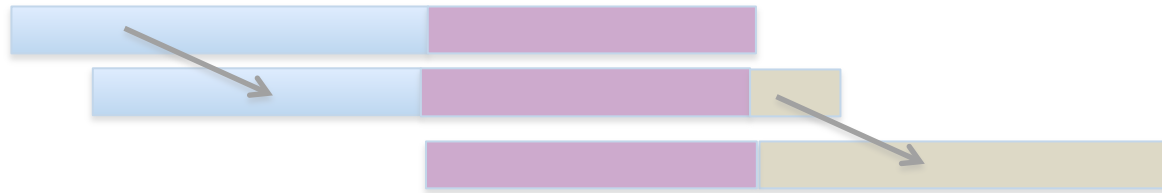


Read Overlap Graph: Reads as nodes, overlaps as edges



Node = read
Edge = overlap

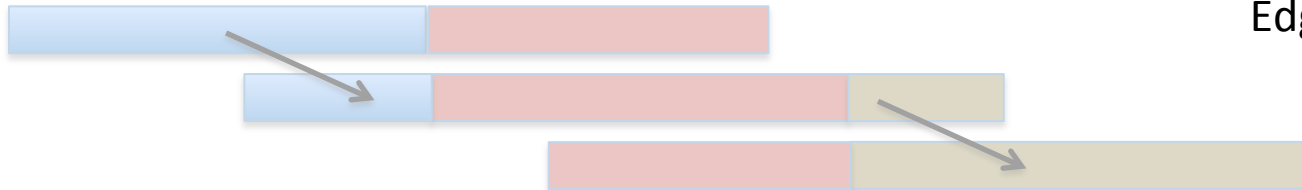
Read Overlap Graph: Reads as nodes, overlaps as edges



Transcript A



Generate consensus sequence where reads overlap

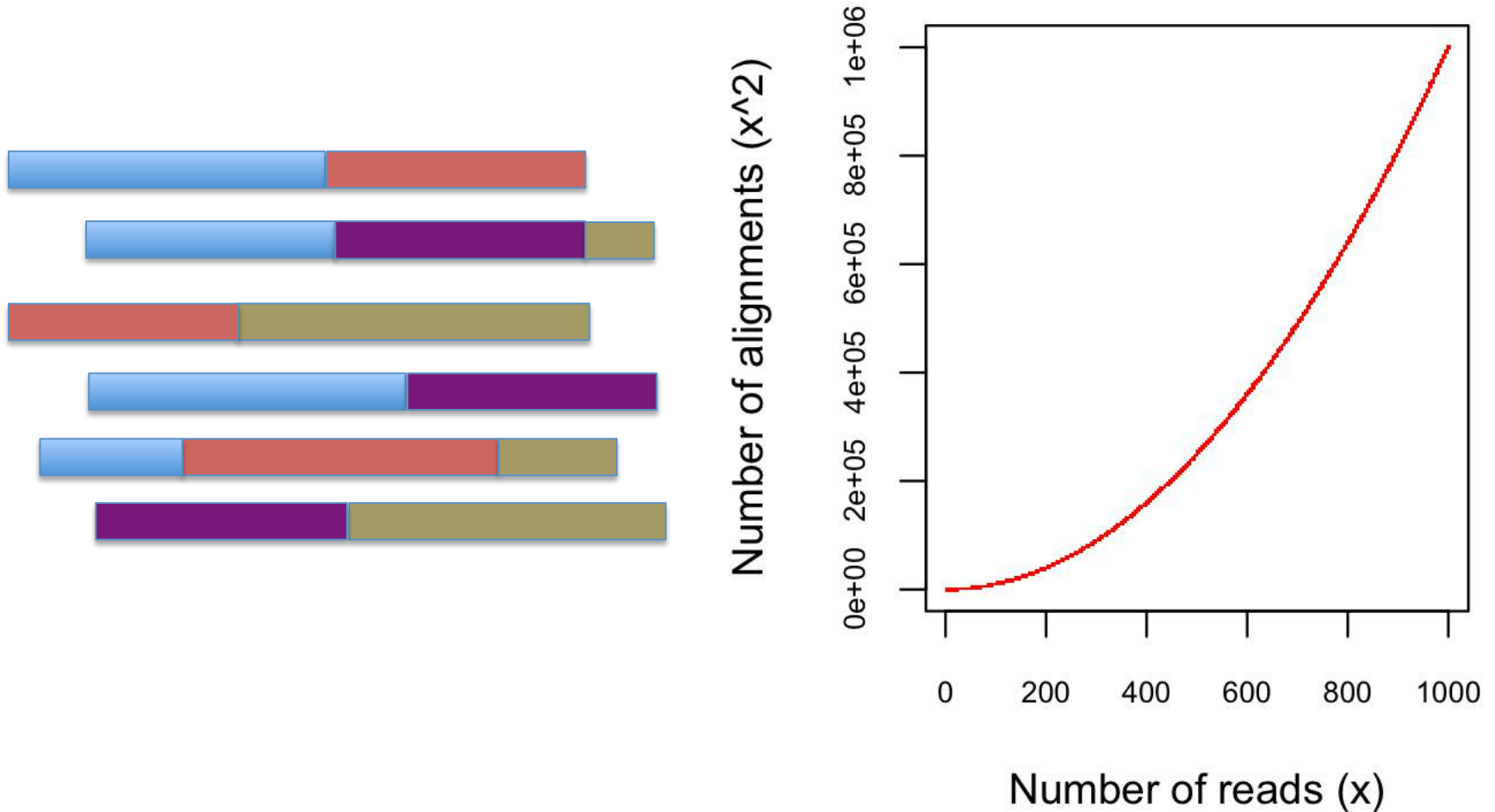


Node = read
Edge = overlap

Transcript B

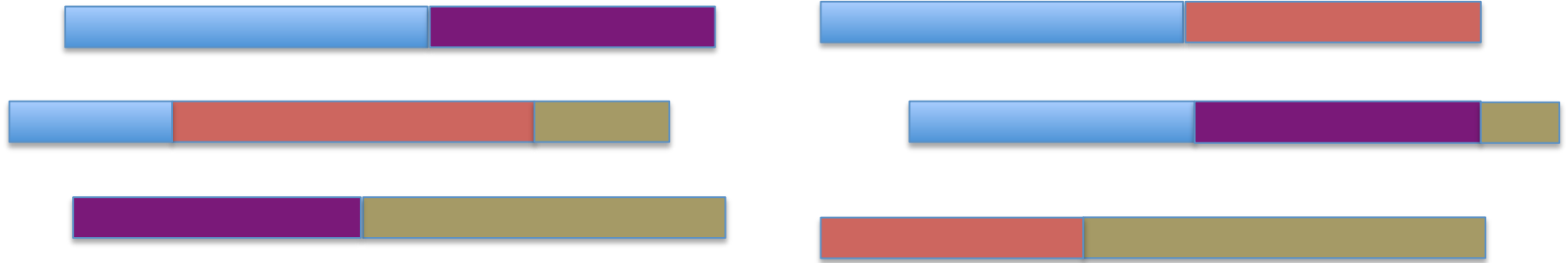


Finding pairwise overlaps between n reads involves $\sim n^2$ comparisons.



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

No genome to align to... De novo assembly required

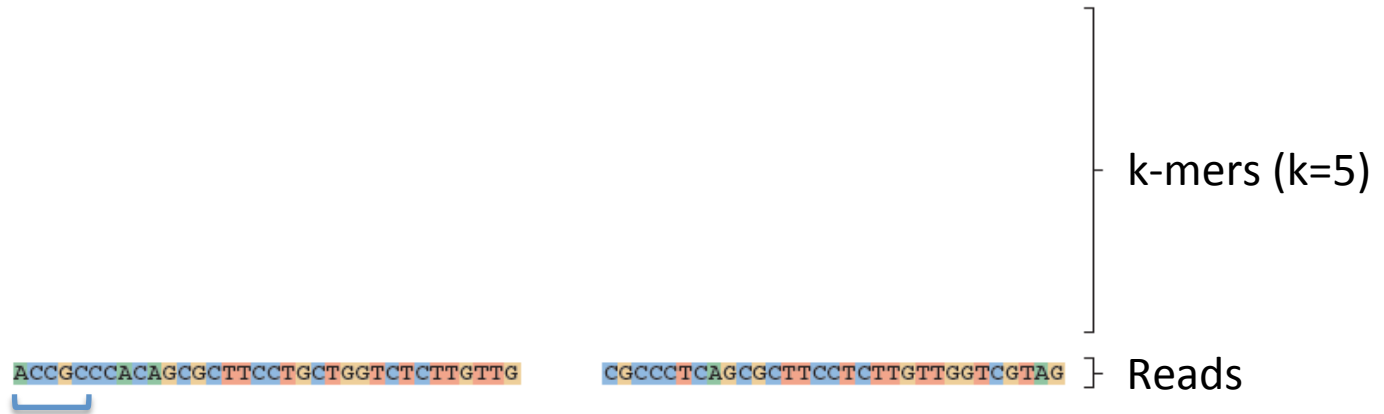


Want to avoid n^2 read alignments to define overlaps

Use a de Bruijn graph

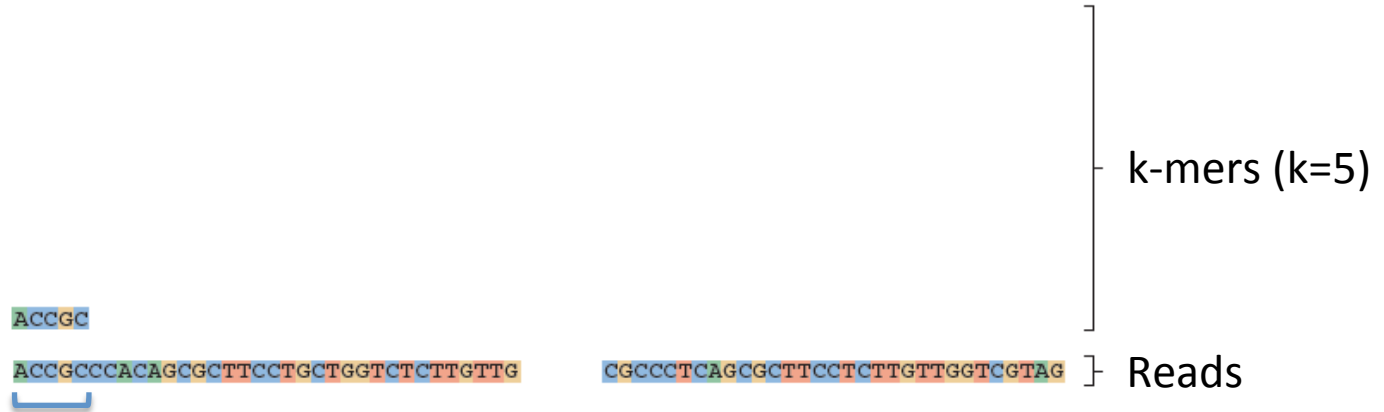
Sequence Assembly via de Bruijn Graphs

Generate all substrings of length k from the reads



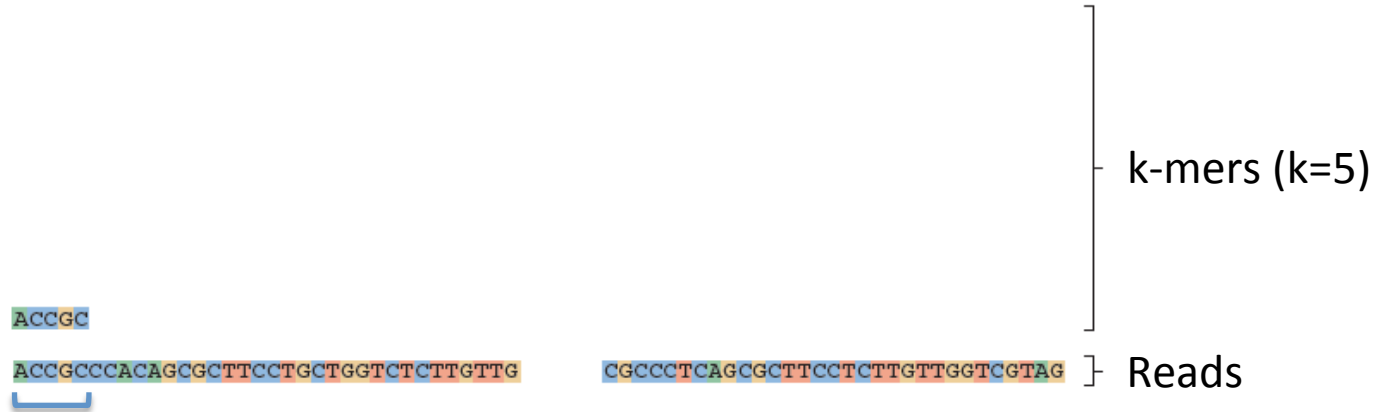
Sequence Assembly via De Bruijn Graphs

Generate all substrings of length k from the reads



Sequence Assembly via De Bruijn Graphs

Generate all substrings of length k from the reads



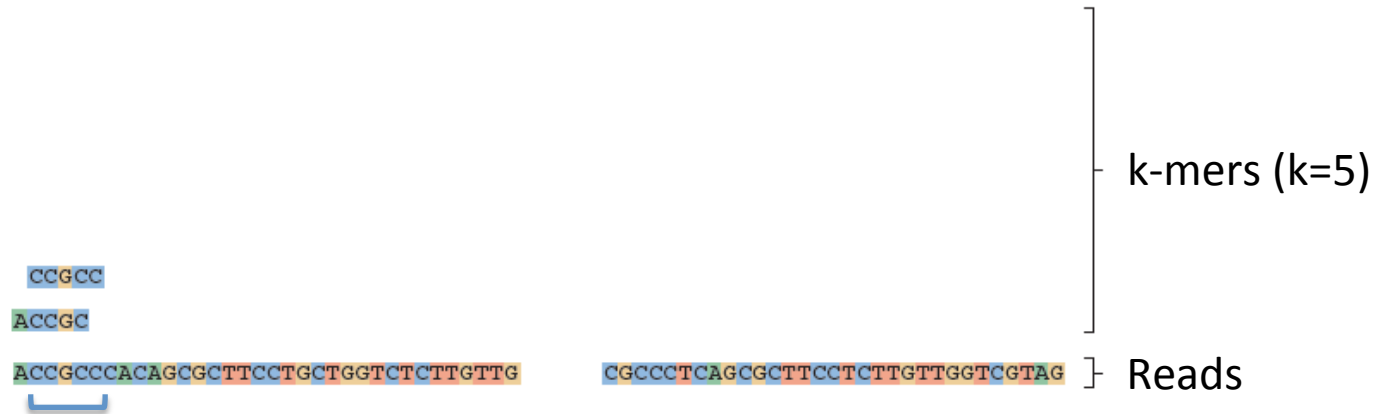
Construct the de Bruijn graph



Nodes = unique k-mers

Sequence Assembly via De Bruijn Graphs

Generate all substrings of length k from the reads

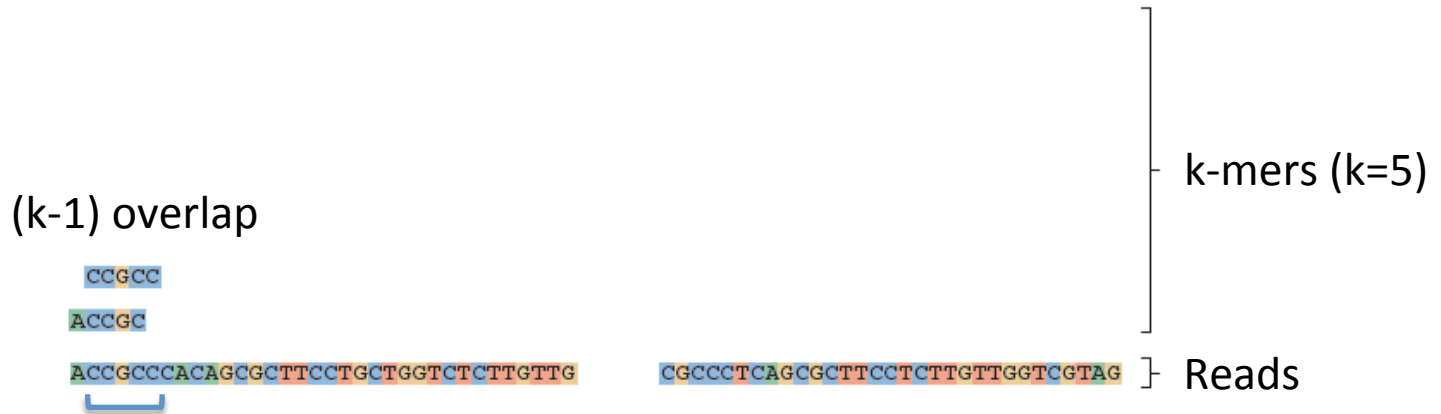


Construct the de Bruijn graph



Sequence Assembly via De Bruijn Graphs

Generate all substrings of length k from the reads

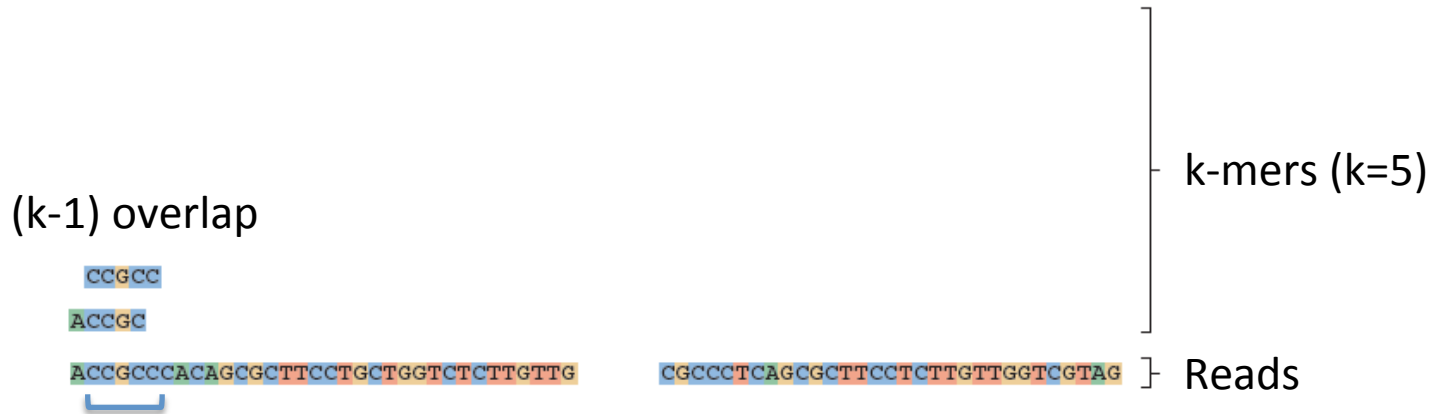


Construct the de Bruijn graph

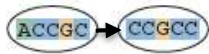


Sequence Assembly via De Bruijn Graphs

Generate all substrings of length k from the reads

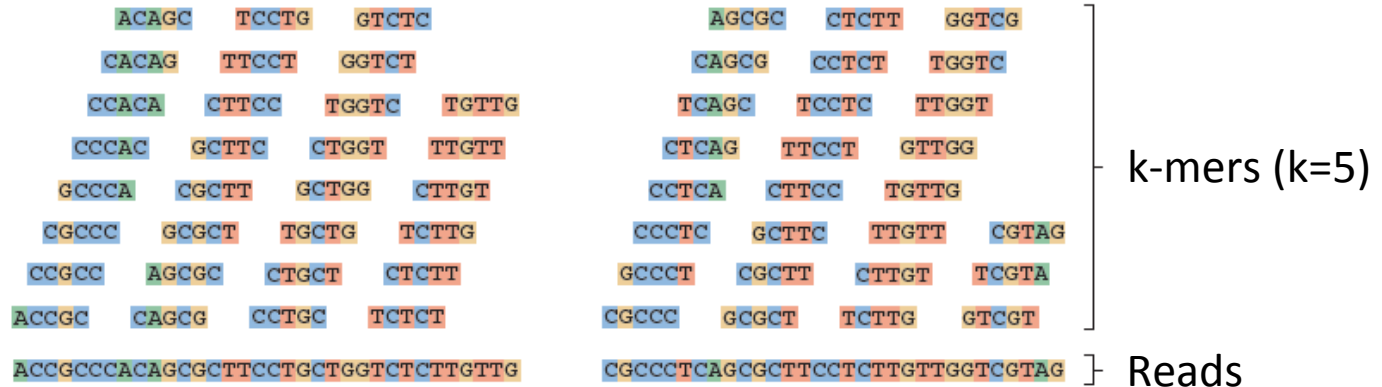


Construct the de Bruijn graph

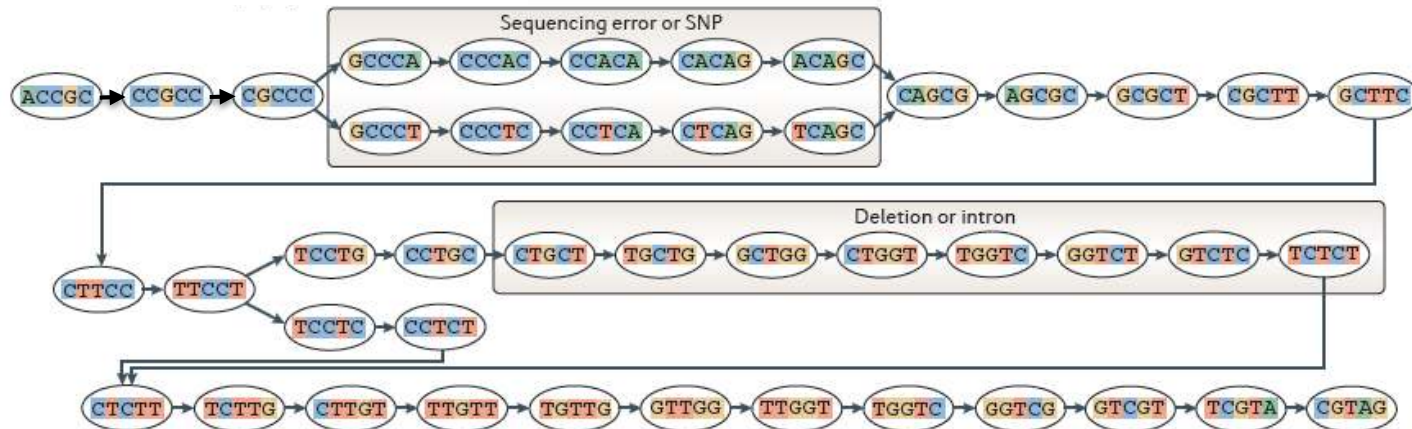


Sequence Assembly via De Bruijn Graphs

Generate all substrings of length k from the reads

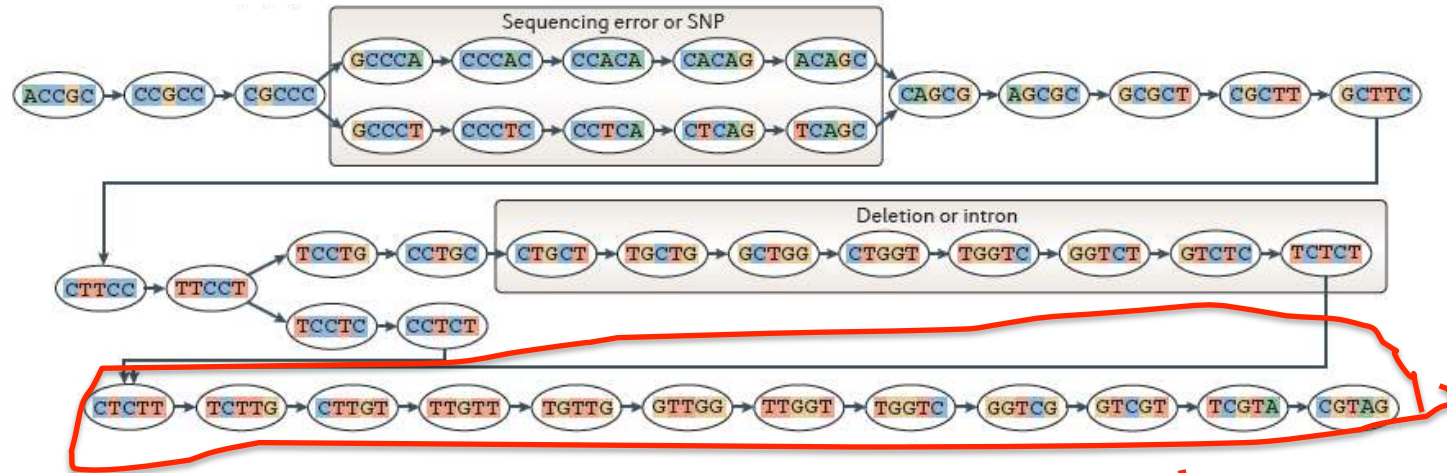


Construct the de Bruijn graph

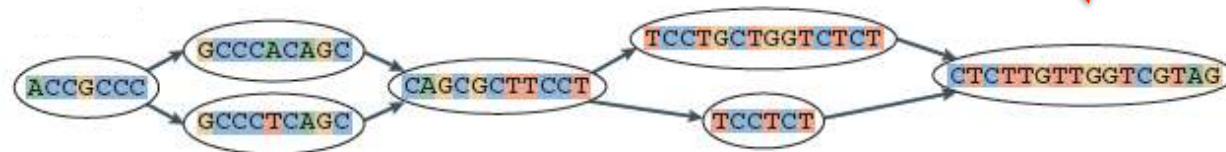


Nodes = unique k-mers
Edges = overlap by $(k-1)$

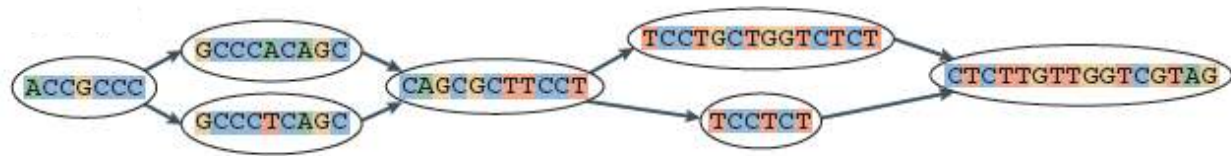
Construct the de Bruijn graph



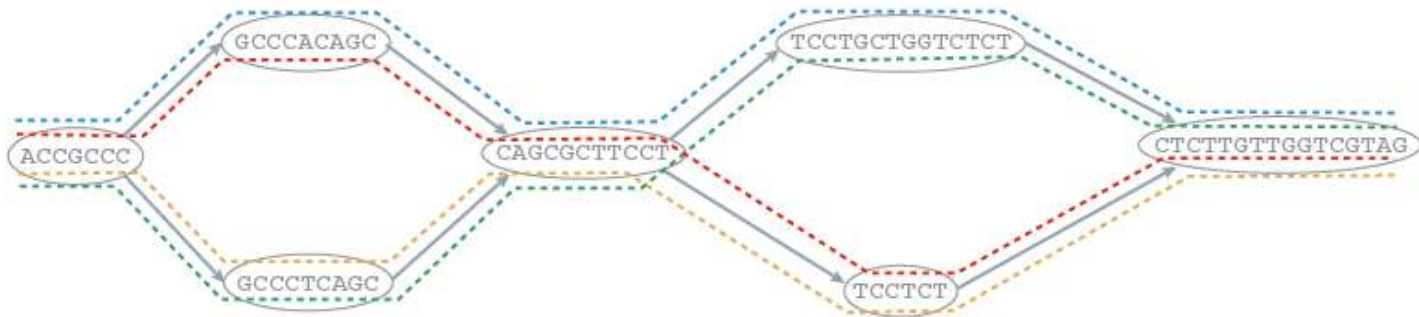
Collapse the de Bruijn graph



Collapse the de Bruijn graph



Traverse the graph



Assemble Transcript Isoforms

----- ACCGCCACAGCGCTTCCTGCTGGTCTCTTGTTGGTCGTAG
 ----- ACCGCCACAGCGCTTCCT ----- CTTGTTGGTCGTAG
 ----- ACCGCCCTCAGCGCTTCCT ----- CTTGTTGGTCGTAG
 ----- ACCGCCCTCAGCGCTTCCTGCTGGTCTCTTGTTGGTCGTAG

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

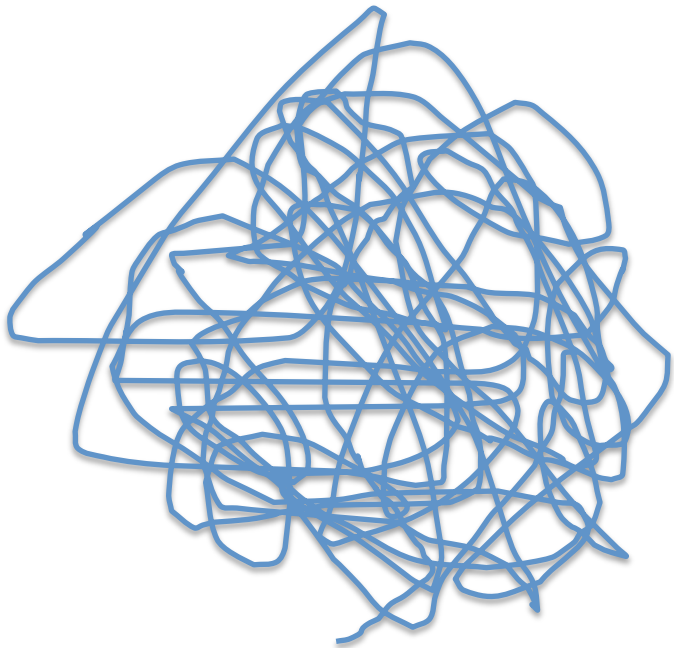
- Exponentially distributed coverage levels
- Multiple contigs per locus (alt splicing)
- Assemble many thousands of Kb-length transcripts
- Strand-specific data available



Trinity Aggregates Isolated Transcript Graphs

Genome Assembly

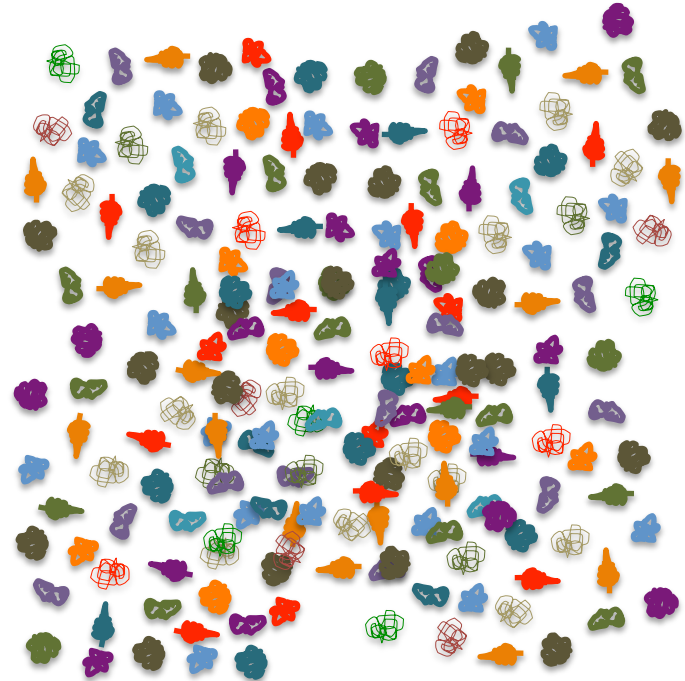
Single Massive Graph



Entire chromosomes represented.

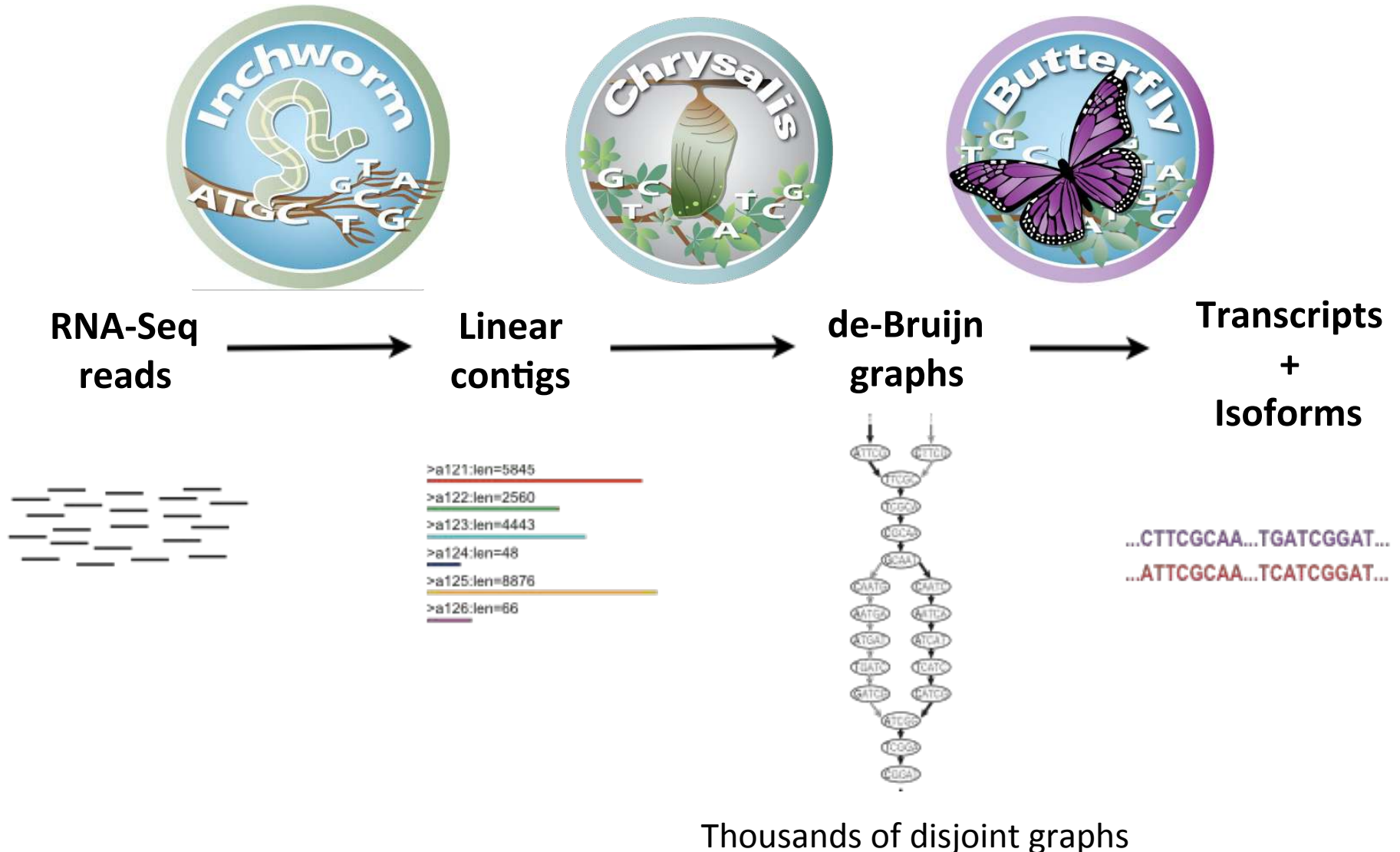
Trinity Transcriptome Assembly

Many Thousands of Small Graphs



Ideally, one graph per expressed gene.

Trinity – How it works:





Inchworm Algorithm

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

Read: **AATGTGAAA**ACTGGATTACATGCTGGTATGTC...

AATGTGA

ATGTGAA

TGTGAAA

...

Overlapping kmers of length (k)

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.

GATTACA
9

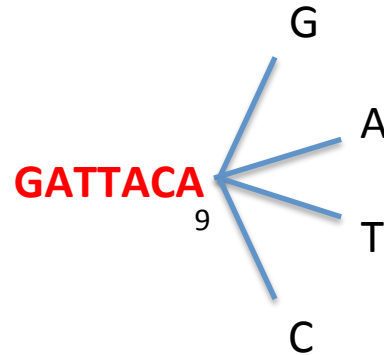
Kmer Catalog (hashtable)

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



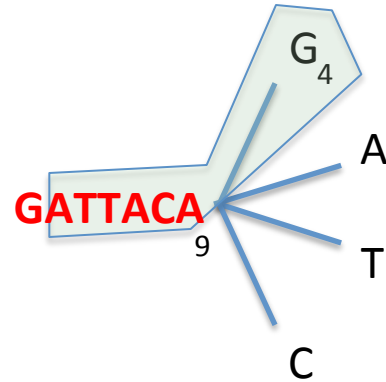
Inchworm Algorithm

- 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.



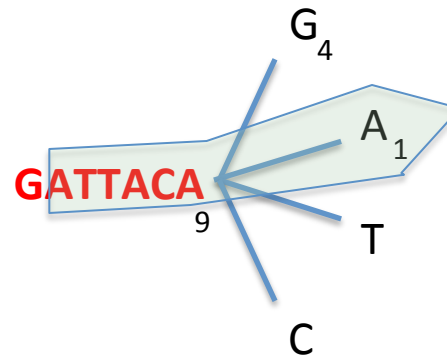


Inchworm Algorithm





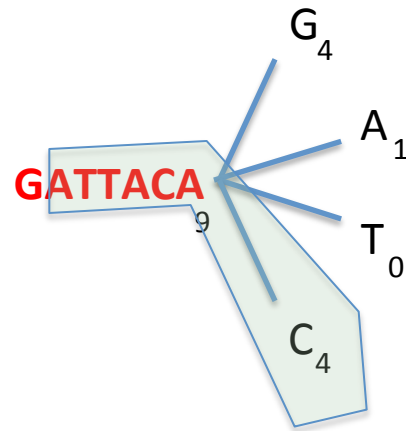
Inchworm Algorithm





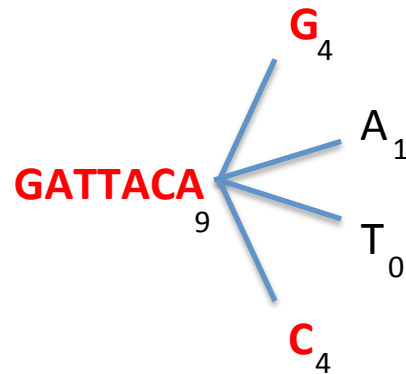


Inchworm Algorithm



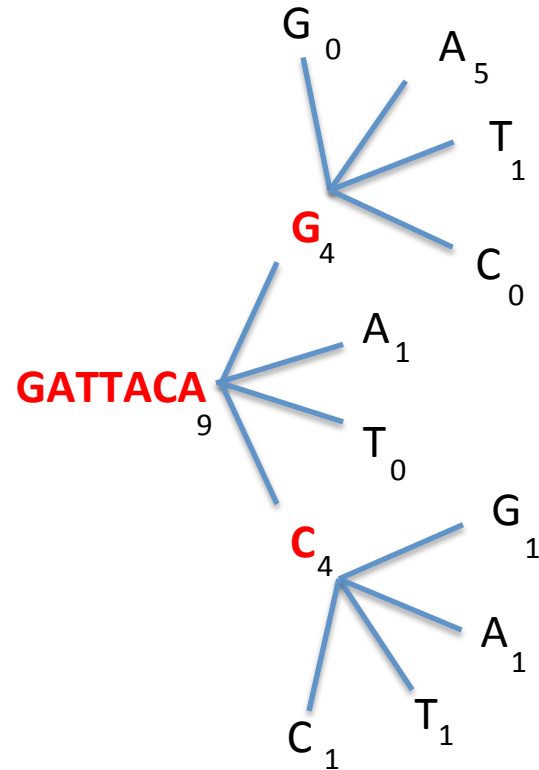


Inchworm Algorithm



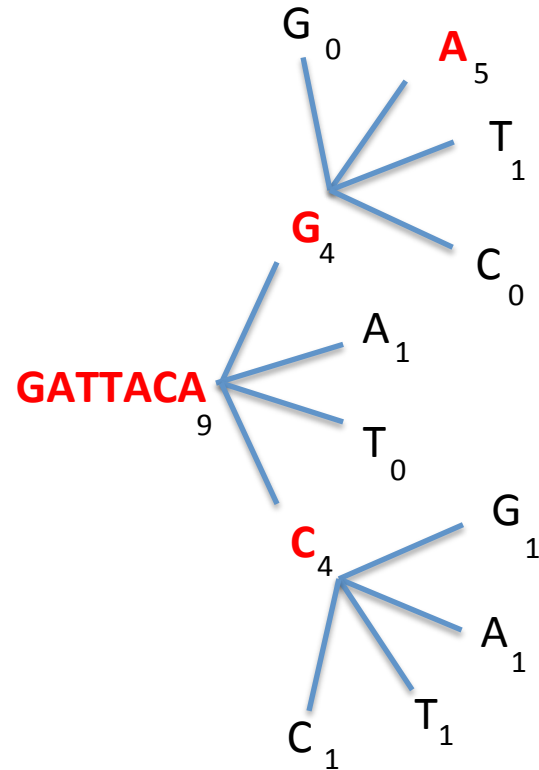


Inchworm Algorithm



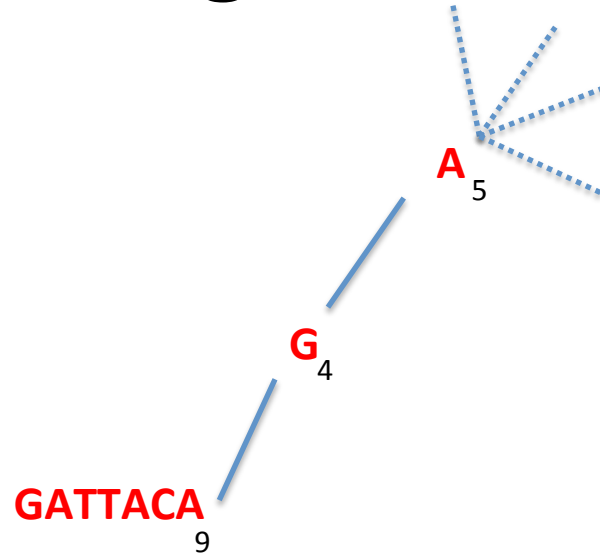


Inchworm Algorithm



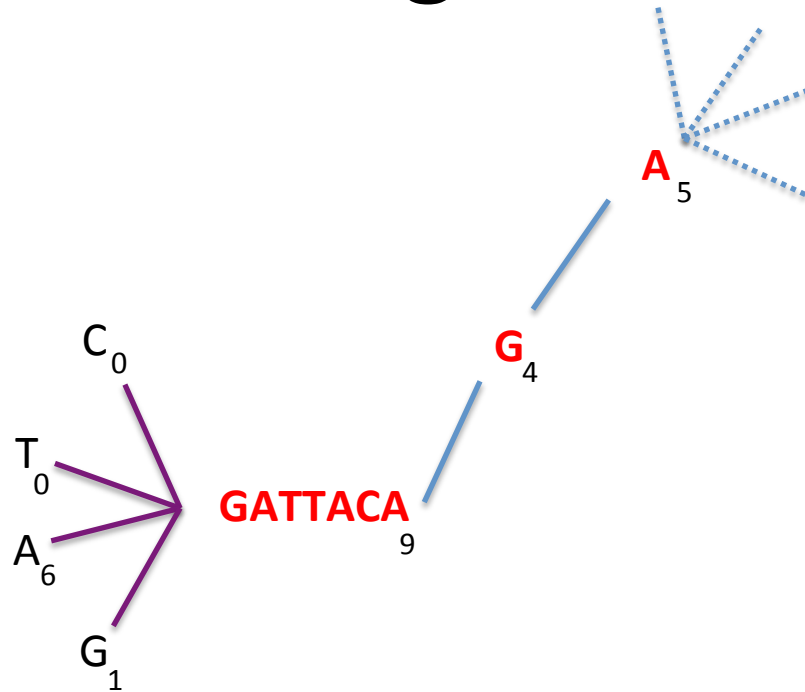


Inchworm Algorithm



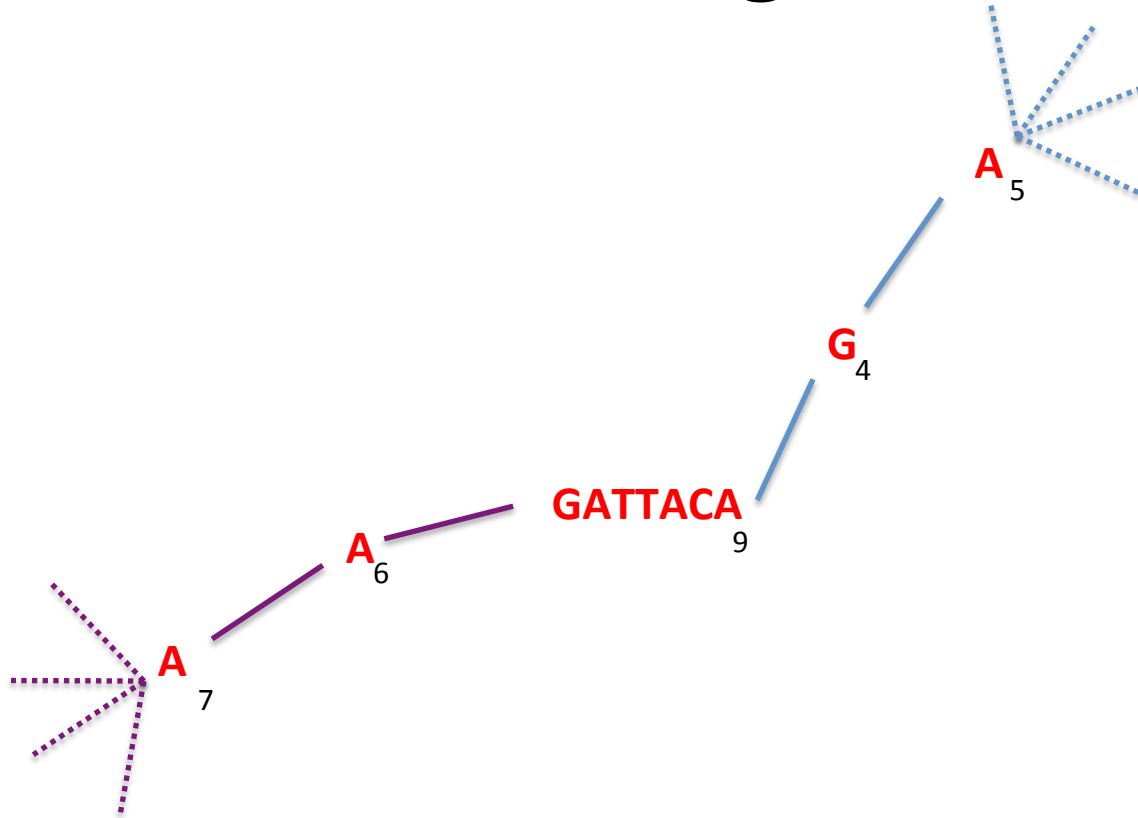


Inchworm Algorithm





Inchworm Algorithm



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

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



Inchworm Contigs from Alt-Spliced Transcripts

Expressed isoforms





Inchworm Contigs from Alt-Spliced Transcripts

Expressed isoforms

Expression



Graphical
representation



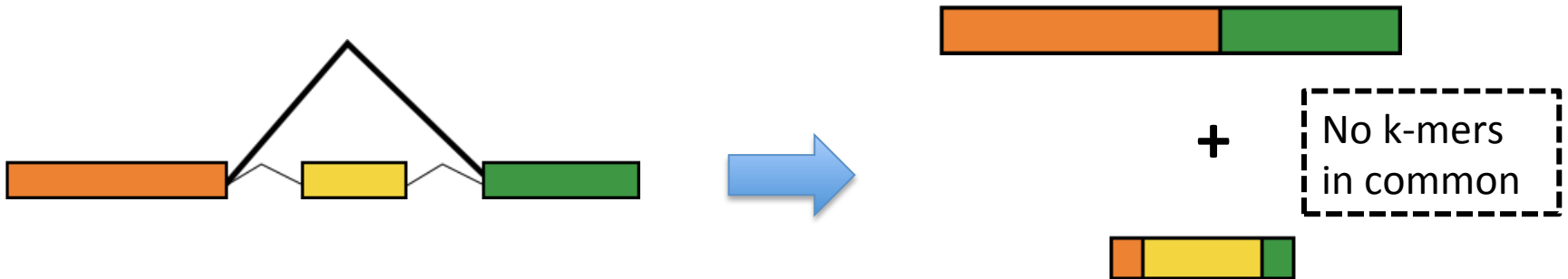


Inchworm Contigs from Alt-Spliced Transcripts



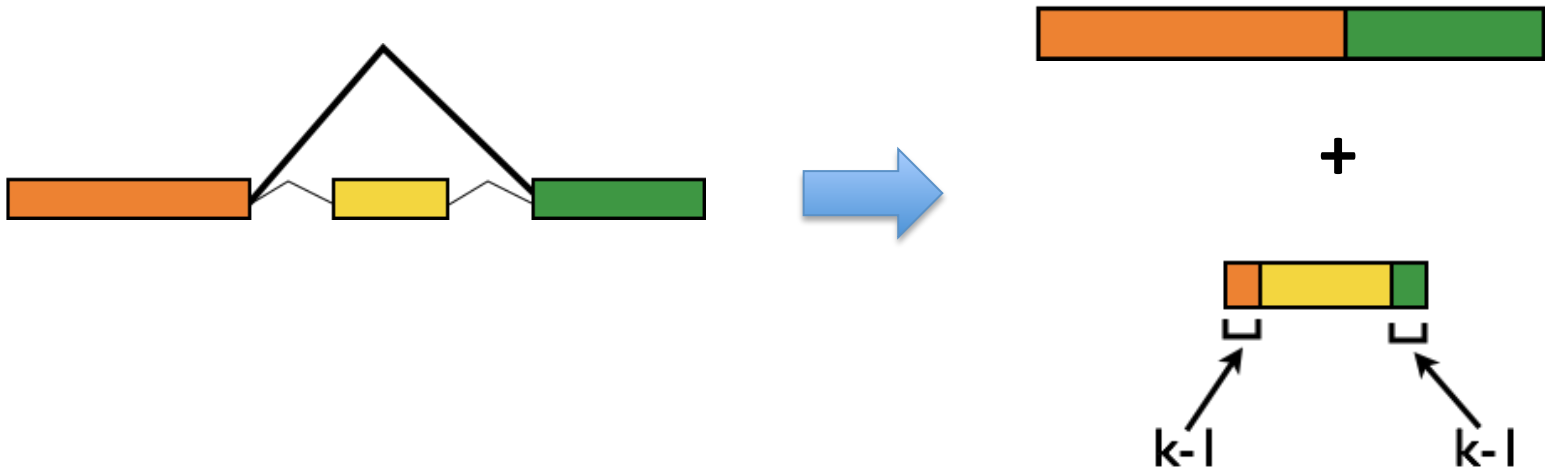


Inchworm Contigs from Alt-Spliced Transcripts

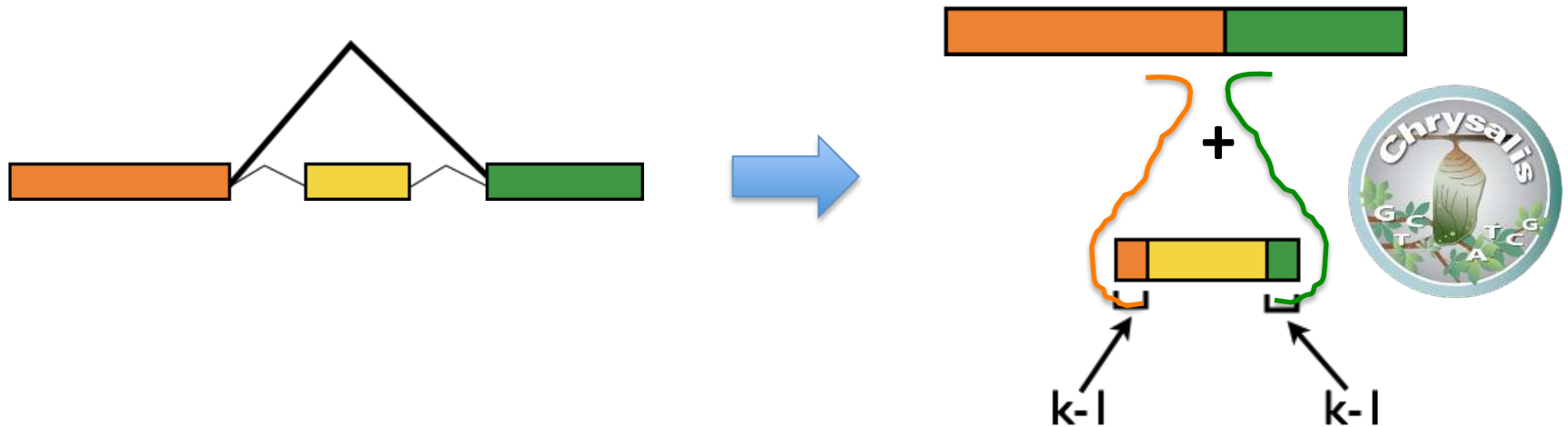




Inchworm Contigs from Alt-Spliced Transcripts



Chrysalis Re-groups Related Inchworm Contigs



Chrysalis uses $(k-1)$ overlaps and read support to link related Inchworm contigs

Chrysalis

>a121:len=5845

>a122:len=2560

>a123:len=4443

>a124:len=48

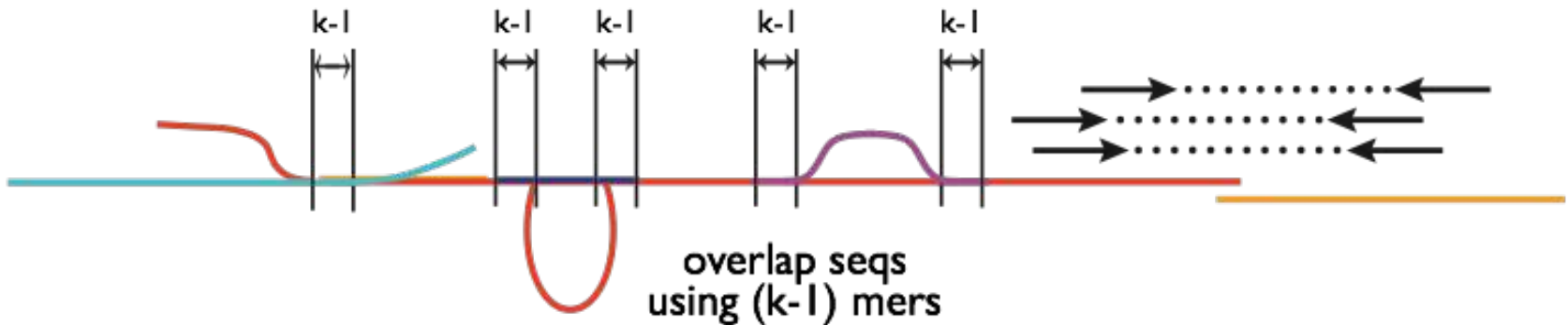
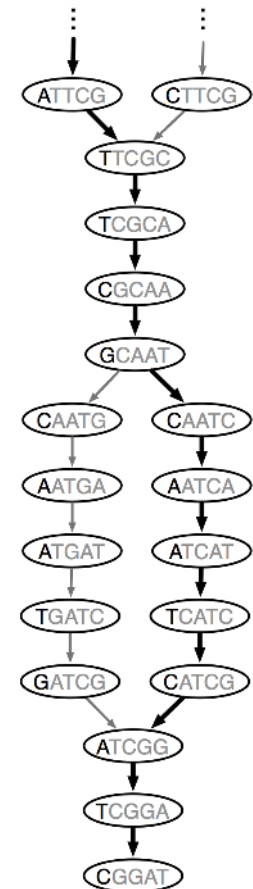
>a125:len=8876

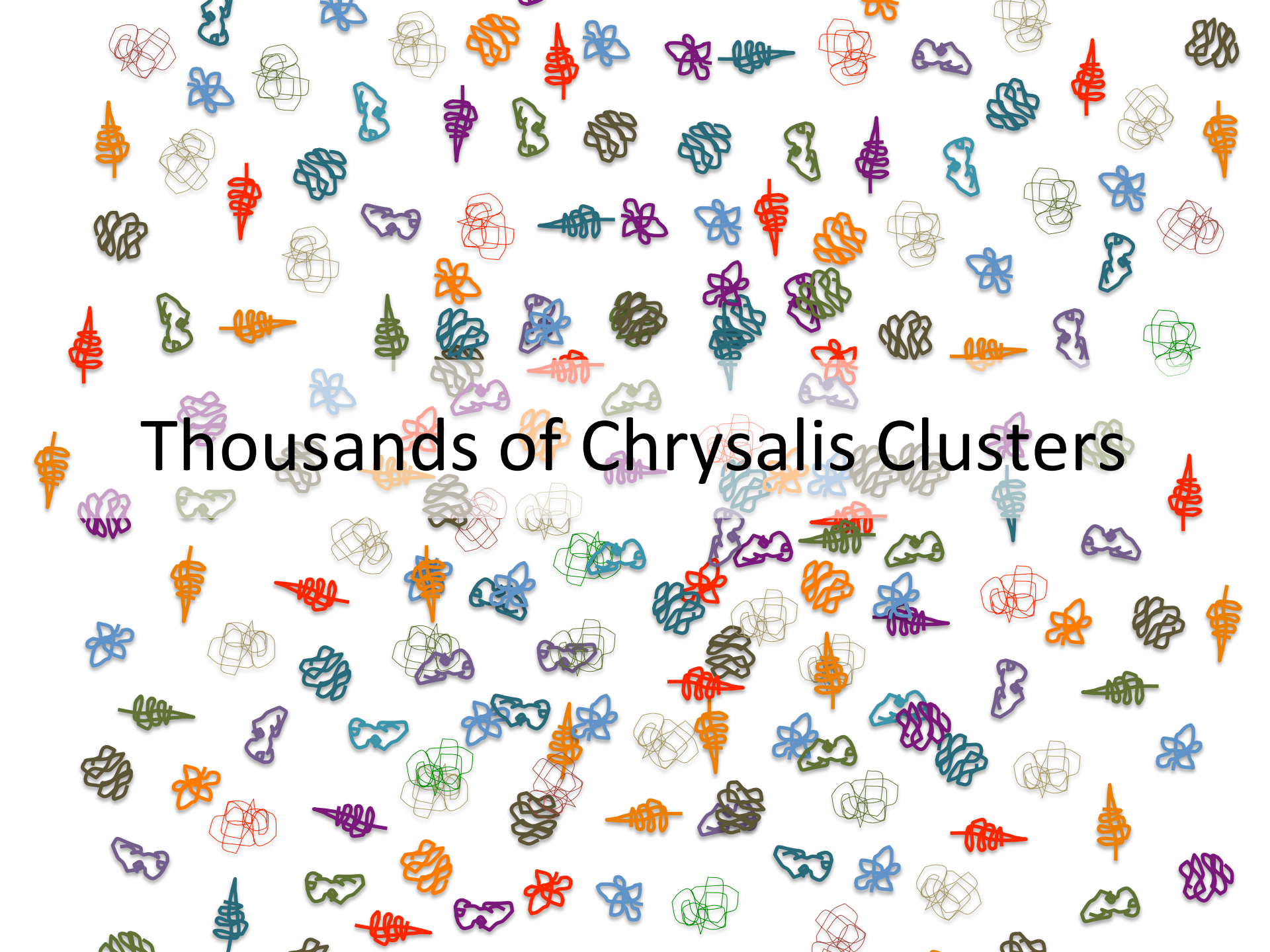
>a126:len=68



Integrate isoforms
via $k-1$ overlaps

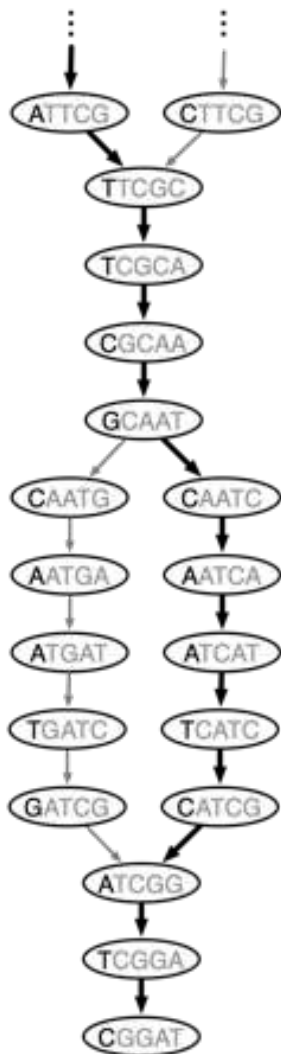
Build de Bruijn Graphs
(ideally, one per gene)



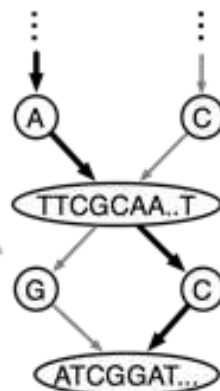


Thousands of Chrysalis Clusters

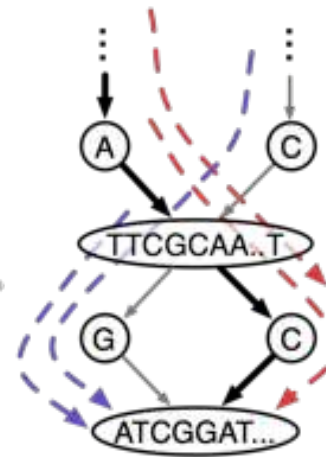
Butterfly



de Bruijn
graph



compact
graph



compact
graph with
reads

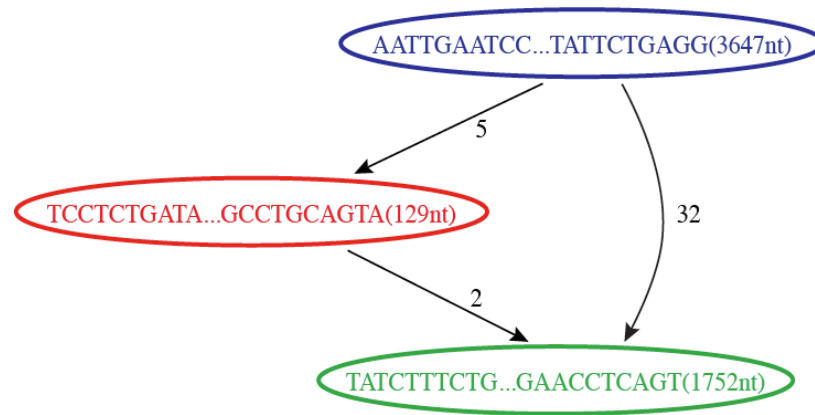


..**CTTCGCAA..TGATCGGAT...**
..**ATTCGCAA..TCATCGGAT...**

sequences
(isoforms and paralogs)

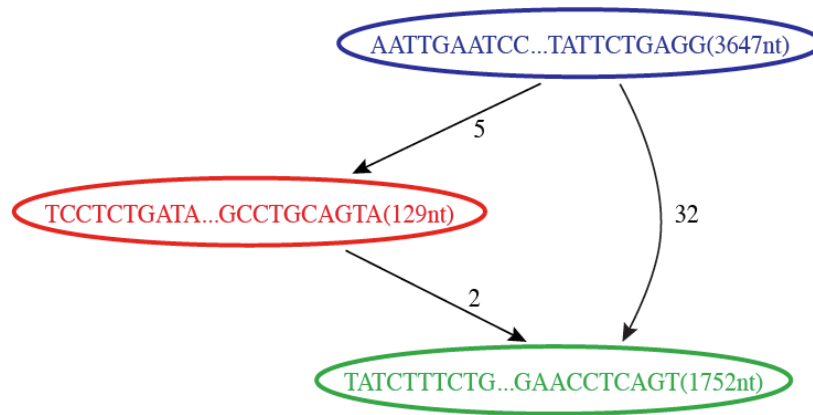
Butterfly Example 1: Reconstruction of Alternatively Spliced Transcripts

Butterfly's Compacted
Sequence Graph



Reconstruction of Alternatively Spliced Transcripts

Butterfly's Compacted
Sequence Graph

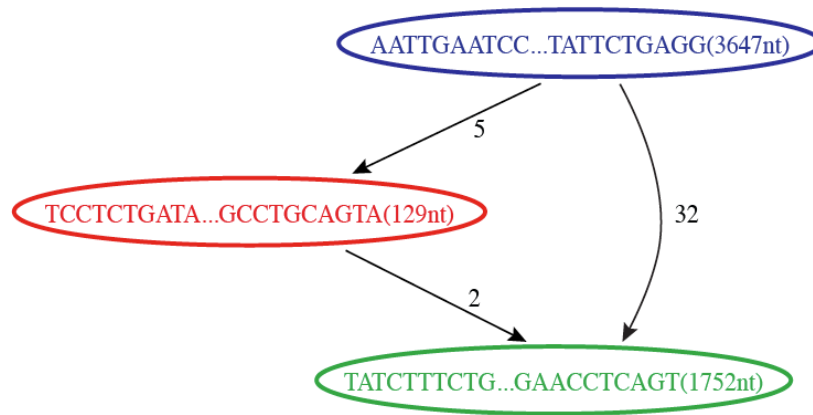


Reconstructed Transcripts



Reconstruction of Alternatively Spliced Transcripts

Butterfly's Compacted
Sequence Graph

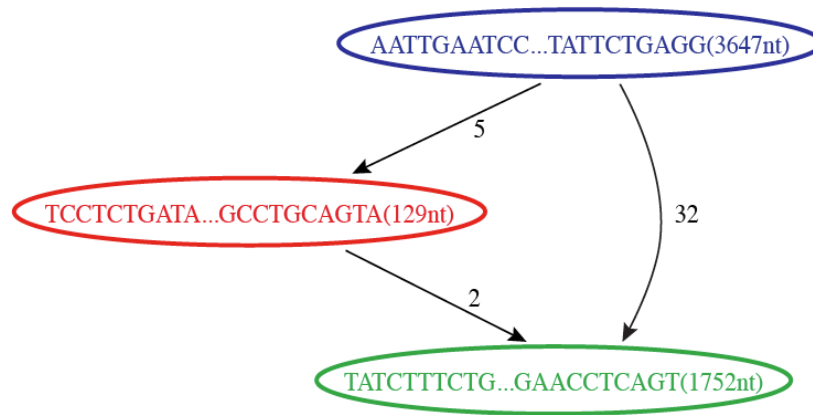


Reconstructed Transcripts



Reconstruction of Alternatively Spliced Transcripts

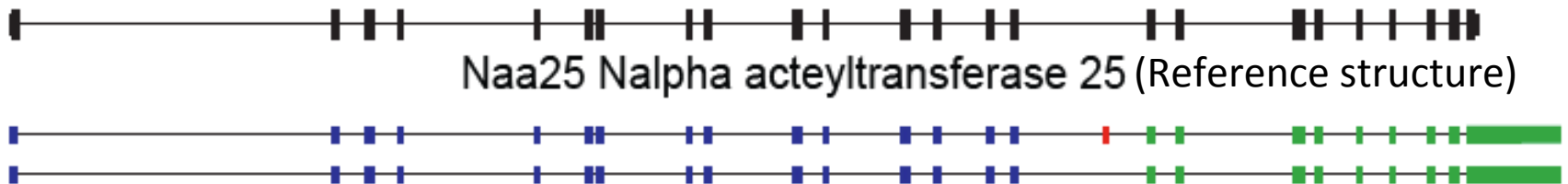
Butterfly's Compacted
Sequence Graph



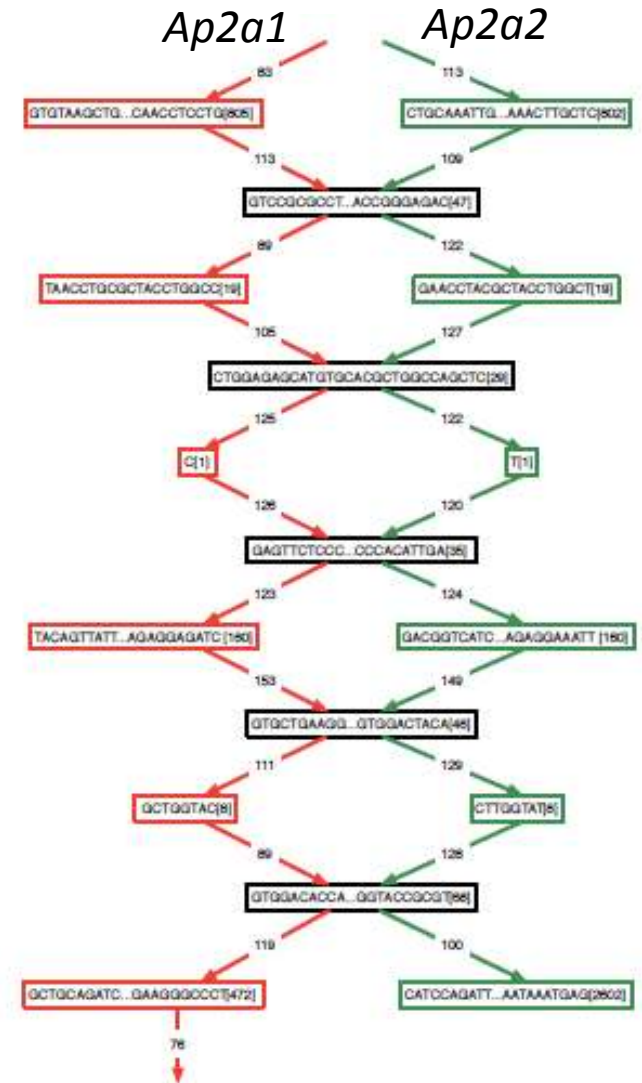
Reconstructed Transcripts



Aligned to Mouse Genome



Butterfly Example 2: Teasing Apart Transcripts of Paralogous Genes



Teasing Apart Transcripts of Paralogous Genes

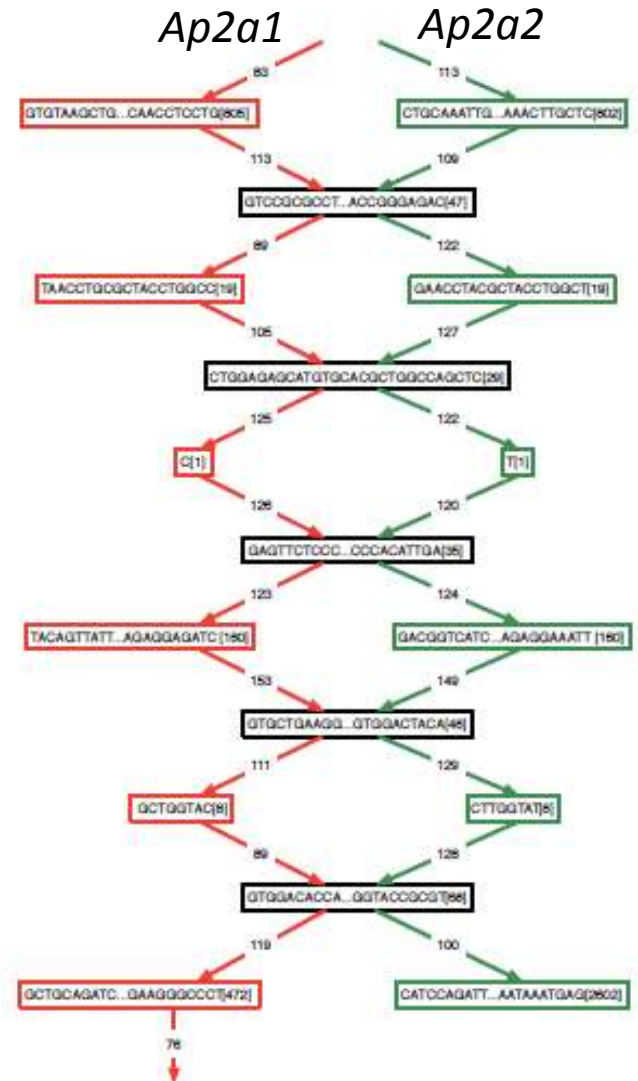
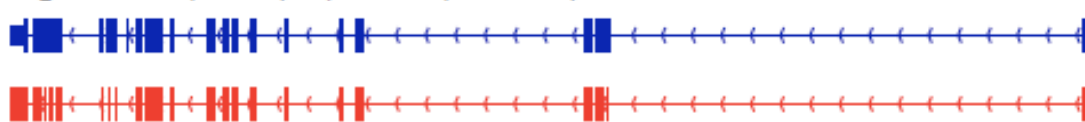
chr7:148,744,197-148,821,437

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



```
>comp0 c0 seq1 len=5528 path=[1:0-3646 10775:3647-3775 3648:3776-5527]
```

[illegible]

```
>comp0 c0 seq2 len=5399 path=[1:0-3646 3648:3647-5398]
```

[illegible]

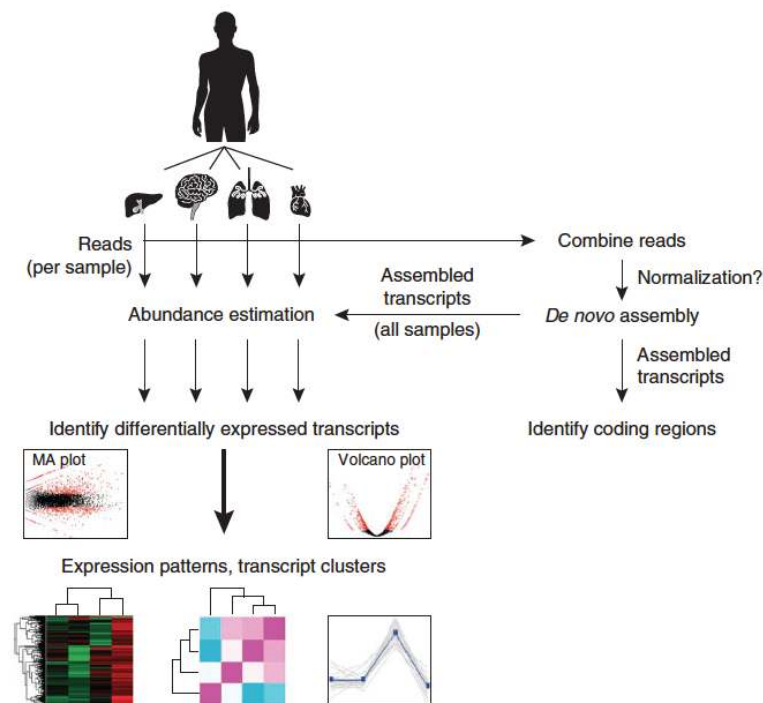
De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis

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



RNA-Seq De novo Assembly Using Trinity

► Pages 27



Quick Guide for the Impatient

Trinity assembles transcript sequences from Illumina RNA-Seq data.

Download Trinity [here](#).

Build Trinity by typing 'make' in the base installation directory.

Assemble RNA-Seq data like so:

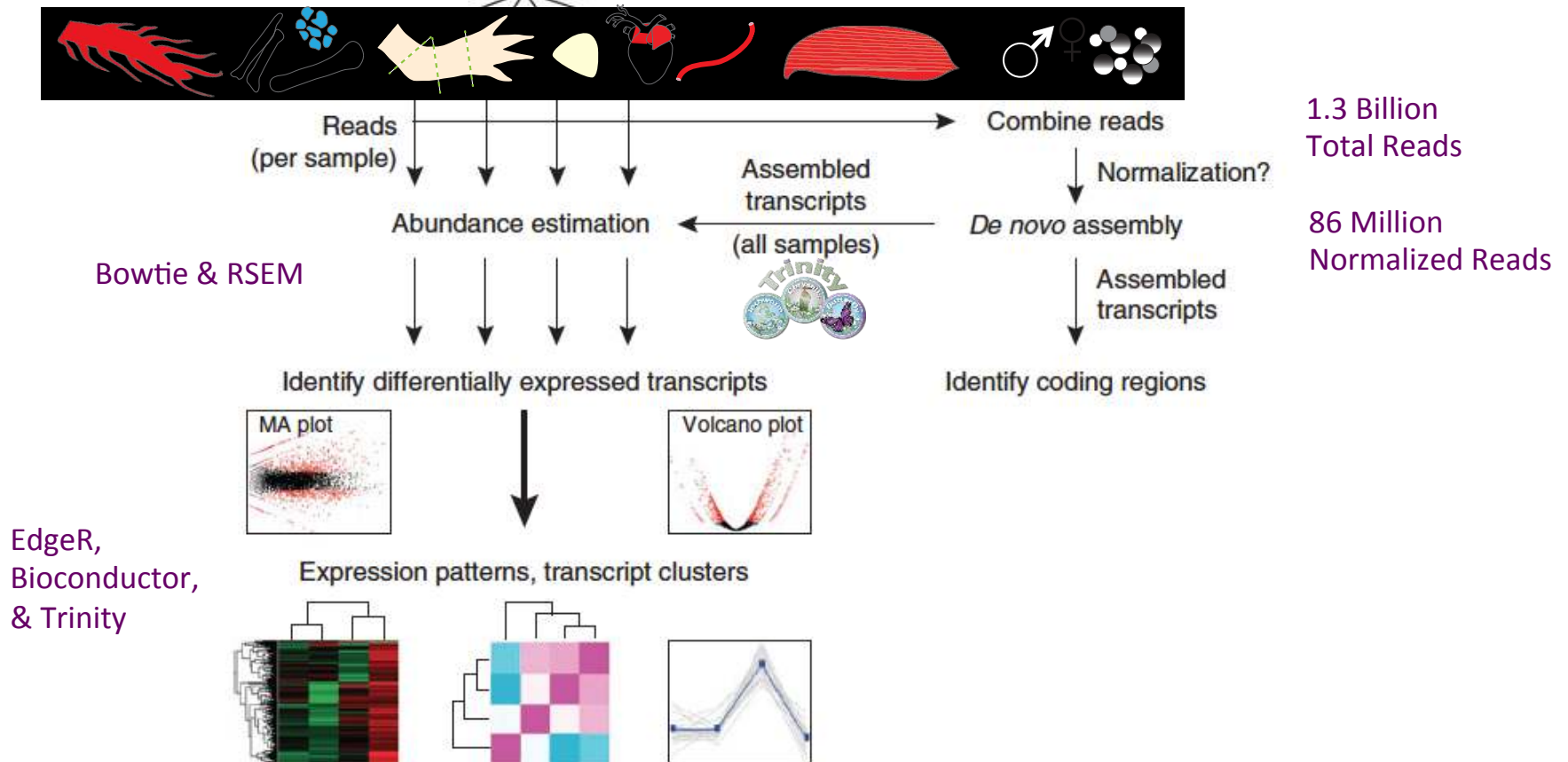
```
Trinity --seqType fq --left reads_1.fq --right reads_2.fq --CPU 6 --max_memory 20G
```

Find assembled transcripts as: 'trinity_out_dir/Trinity.fasta'

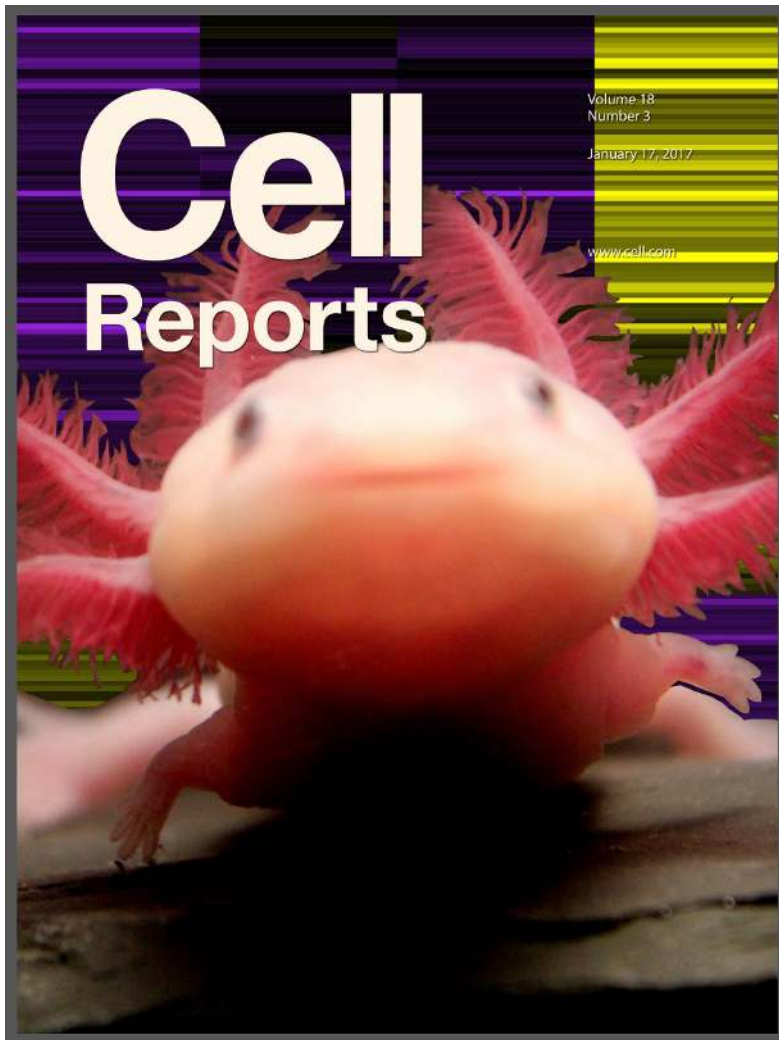
Use the documentation links in the right-sidebar to navigate this documentation, and contact our [Google group for technical support](#).

- [Trinity Wiki Home](#)
- [Installing Trinity](#)
 - [Trinity Computing Requirements](#)
 - [Accessing Trinity on Publicly Available Compute Resources](#)
 - [Run Trinity using Docker](#)
- [Running Trinity](#)
 - [Genome Guided Trinity Transcriptome Assembly](#)
 - [Gene Structure Annotation of Genomes](#)
- [Trinity process and resource monitoring](#)
 - [Monitoring Progress During a Trinity Run](#)
 - [Examining Resource Usage at the End of a Trinity Run](#)
- [Output of Trinity Assembly](#)
- [Assembly Quality Assessment](#)
 - [Counting Full-length Transcripts](#)
 - [RNA-Seq Read Representation](#)
 - [Contig Nx and ExN50 stats](#)
 - [Examine strand-specificity of reads](#)
- [Downstream Analyses](#)

Framework for De novo Transcriptome Assembly and Analysis



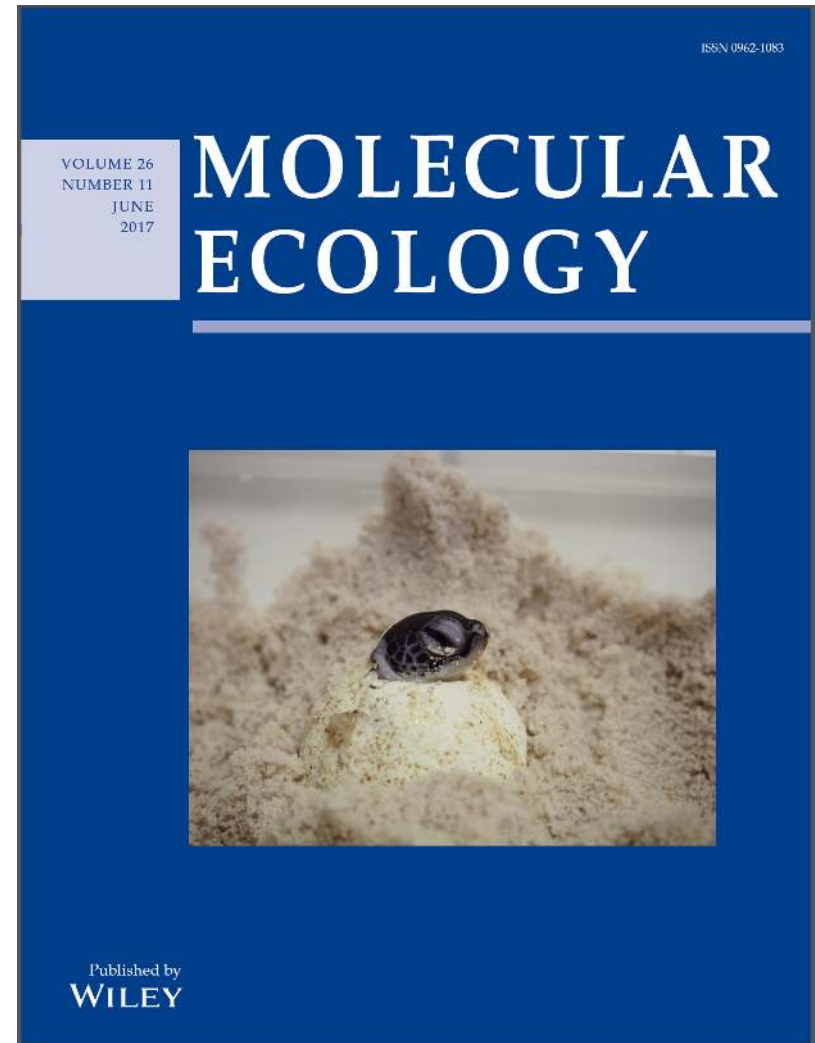
Example Applications of the Trinity RNA-Seq Protocol



Resource

A Tissue-Mapped Axolotl De Novo Transcriptome Enables Identification of Limb Regeneration Factors

Donald M. Bryant^{1,6}, Kimberly Johnson^{1,6}, Tia DiTommaso¹, Timothy Tickle², Matthew Brian Couger³, Duygu Payzin-Dogru¹, Tae J. Lee¹, Nicholas D. Leigh¹, Tzu-Hsing Kuo¹, Francis G. Davis¹, Joel Bateman¹, Sevara Bryant¹, Anna R. Guzikowski¹, Stephanie L. Tsai⁴, Steven Coyne¹, William W. Ye¹, Robert M. Freeman Jr.⁵, Leonid Peshkin⁵, Clifford J. Tabin⁴, Aviv Regev², Brian J. Haas²,  , Jessica L. Whited^{1,7}.



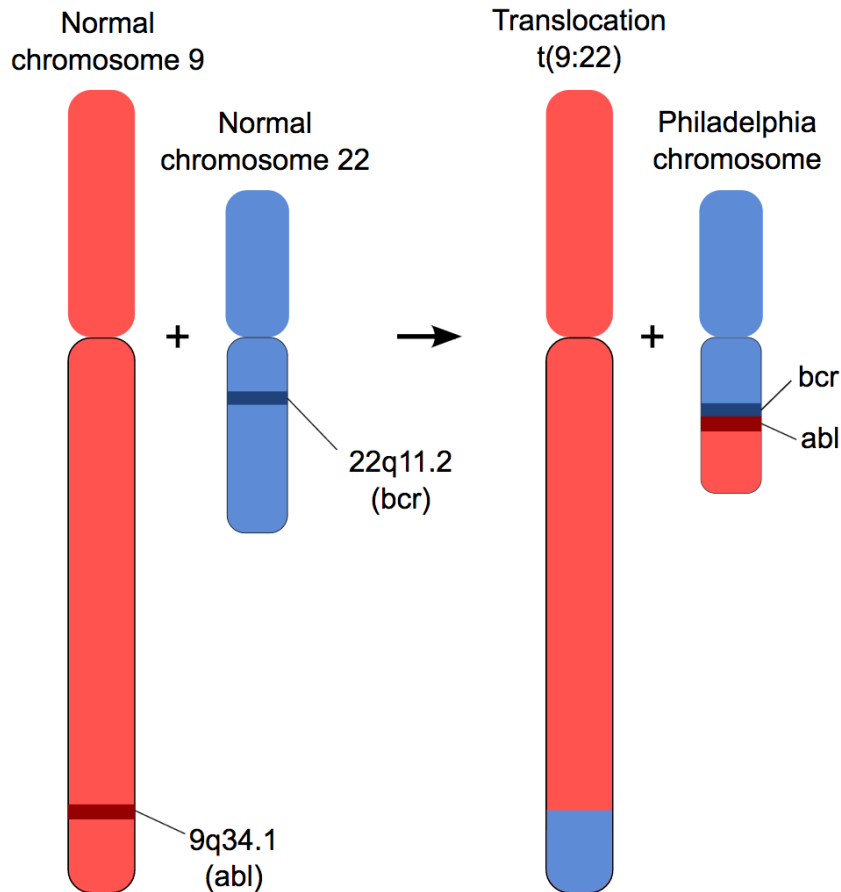
Original Article

Loggerhead sea turtle embryos (*Caretta caretta*) regulate expression of stress response and developmental genes when exposed to a biologically realistic heat stress

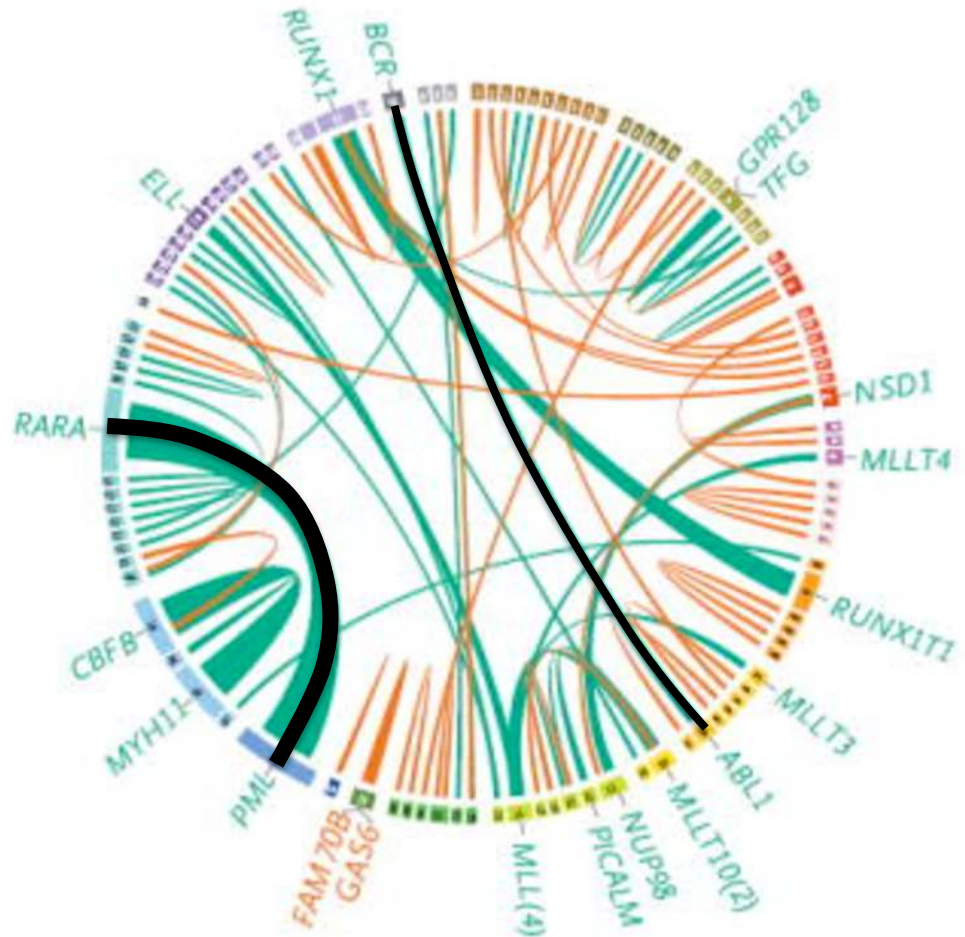
Blair P. Bentley , Brian J. Haas, Jamie N. Tedeschi, Oliver Berry

Biomedical Applications for *de Novo* Transcriptome Assembly

Fusion transcripts in Cancer



BCR--ABL1 fusion in ~95% of chronic myelogenous leukemias (CML)

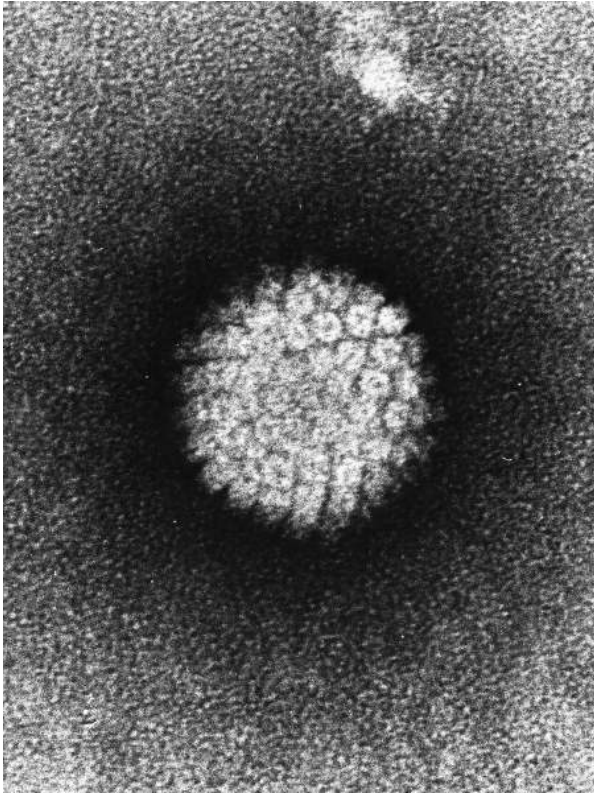


Fusions Identified in a cohort of acute myeloid leukemias (AML) using *de novo* transcriptome assembly.

N Engl J Med. 2013 May 30; 368(22)

Biomedical Applications for *de Novo* Transcriptome Assembly

Detection & Reconstruction of Viral and Microbial Transcripts in Cancer



HPV

Tumor Viruses

- Human papilloma virus (HPV) in cervical cancer
- Hepatitis B & C in liver cancer
- Epstein Barr Virus in lymphomas
- T-lymphotrophic virus in adult T-cell leukemia

Bacterial / Cancer Associations

- Helicobacter pylori / stomach cancer
- Fusobacterium nucleatum / colon cancer

Contrasting Genome-guided and De novo Assembly

Genome-guided reconstruction is
more sensitive than genome-free methods

reads $(k-1)$
Genome-free *de novo* assembly

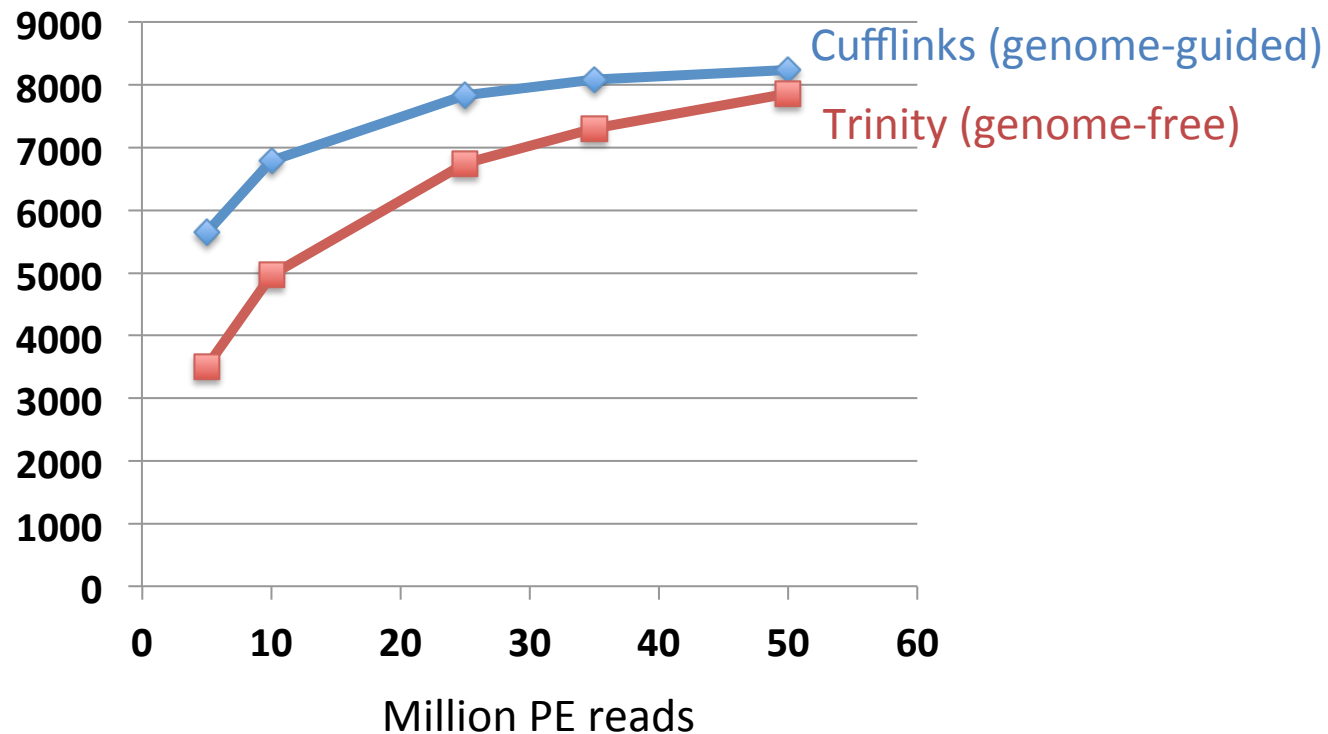
Overlap requirements
>>

reads
Genome-guided assembly

Genes w/ fully
reconstructed
transcripts



Mouse data



Summary

- Transcript reconstruction from RNA-Seq data may leverage genome-guided or de novo assembly
- Transcriptome assembly uses directed graph data structures and path traversal
- Advantages and disadvantages to assembly approaches
 - Genome-guided: well-matched samples and very sensitive
 - *De novo*: almost any sample will do, but requires higher depth of read coverage
- Biomedical applications for *de novo* transcriptome assembly
 - Cancer research: fusion transcripts & pathogen detection

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 |



Comprehensive comparative analysis of strand-specific RNA sequencing methods

Joshua Z Levin^{1,6}, Moran Yassour^{1-3,6}, Xian Adiconis¹, Chad Nusbaum¹, Dawn Anne Thompson¹, Nir Friedman^{3,4}, Andreas Gnirke¹ & Aviv Regev^{1,2,5}

Strand-specific, massively parallel cDNA sequencing (RNA-seq) is a powerful tool for transcript discovery, genome annotation

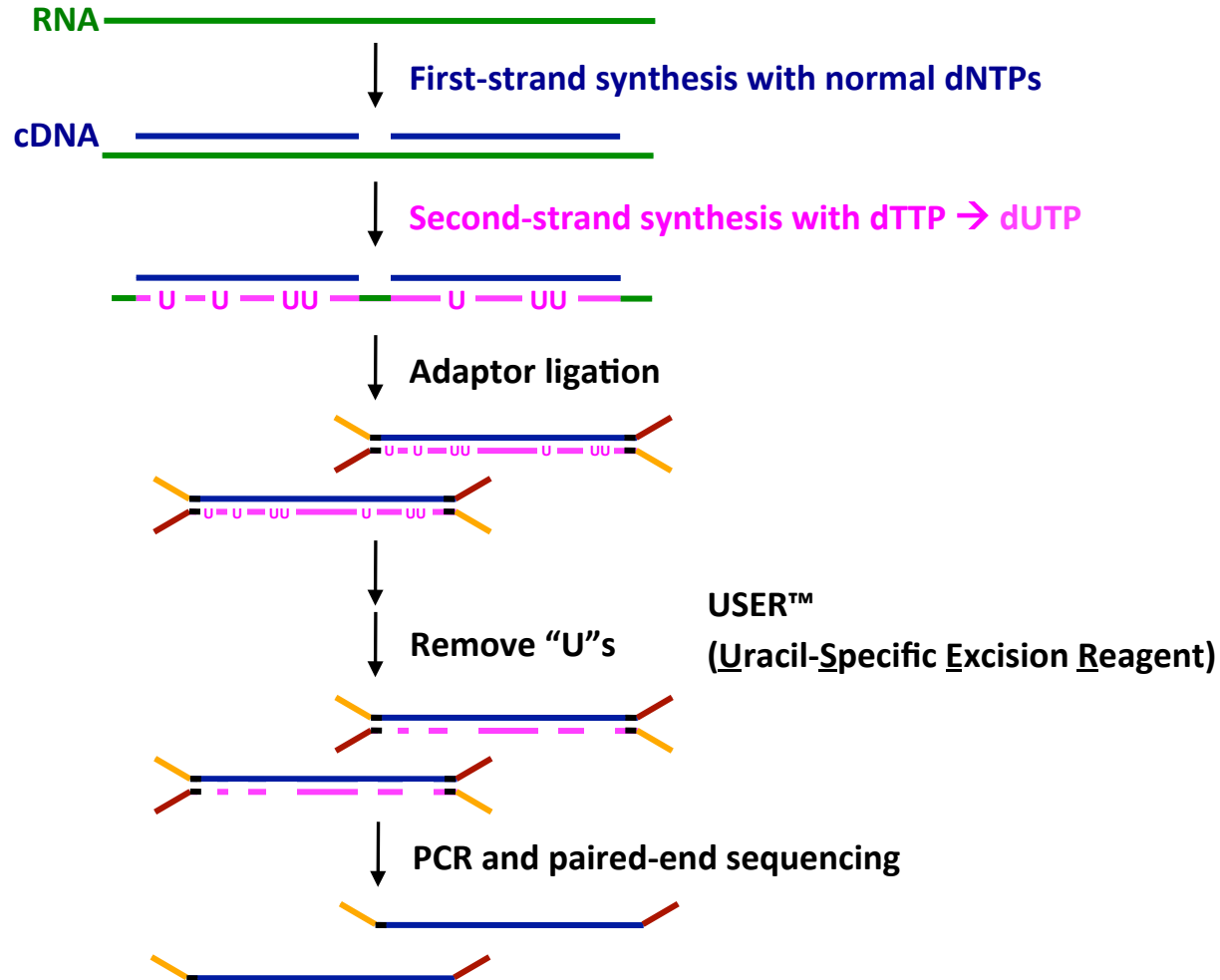
Nevertheless, direct information on the originating strand can substantially enhance the value of an RNA-seq experiment. For

'dUTP second strand marking' identified as the leading protocol

to choose between them, here we developed a comprehensive 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

transcribed strand or other noncoding RNAs; demarcate the exact 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 2nd Strand Method: Our Favorite

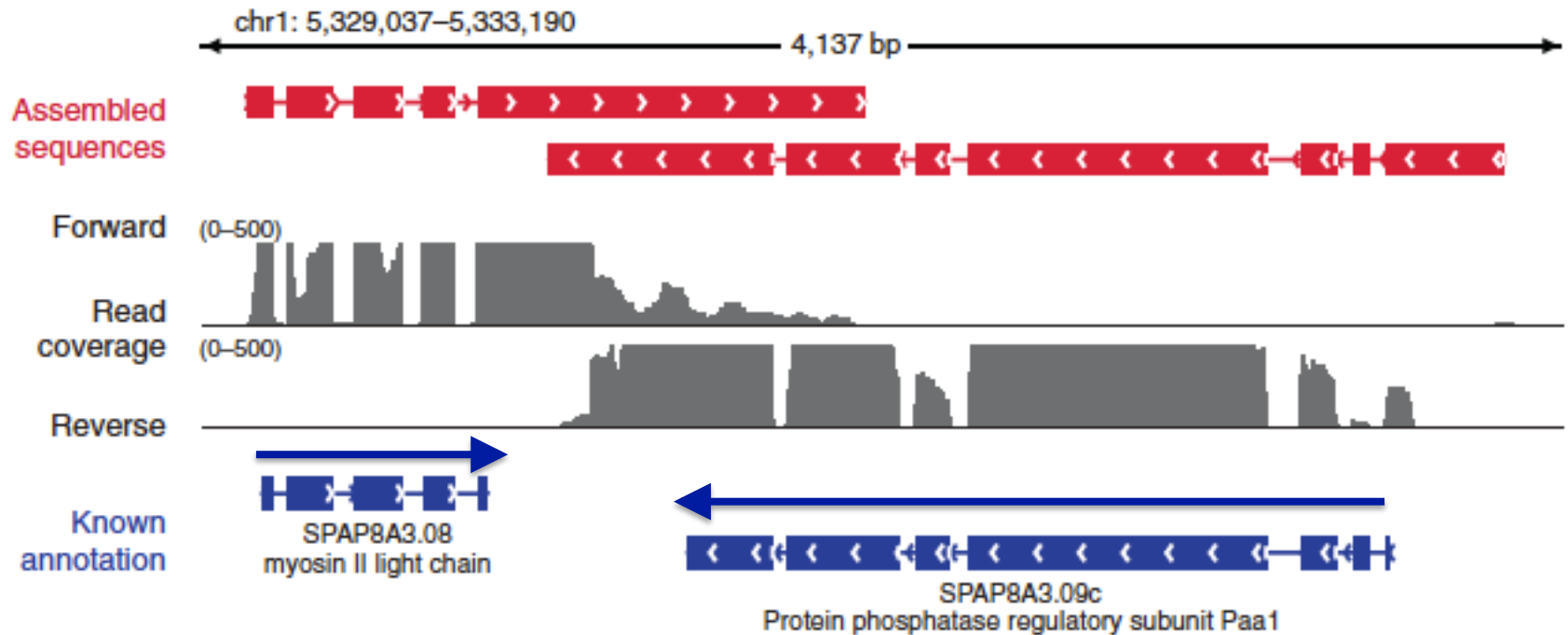


Modified from Parkhomchuk *et al.* (2009) *Nucleic Acids Res.* 37:e123

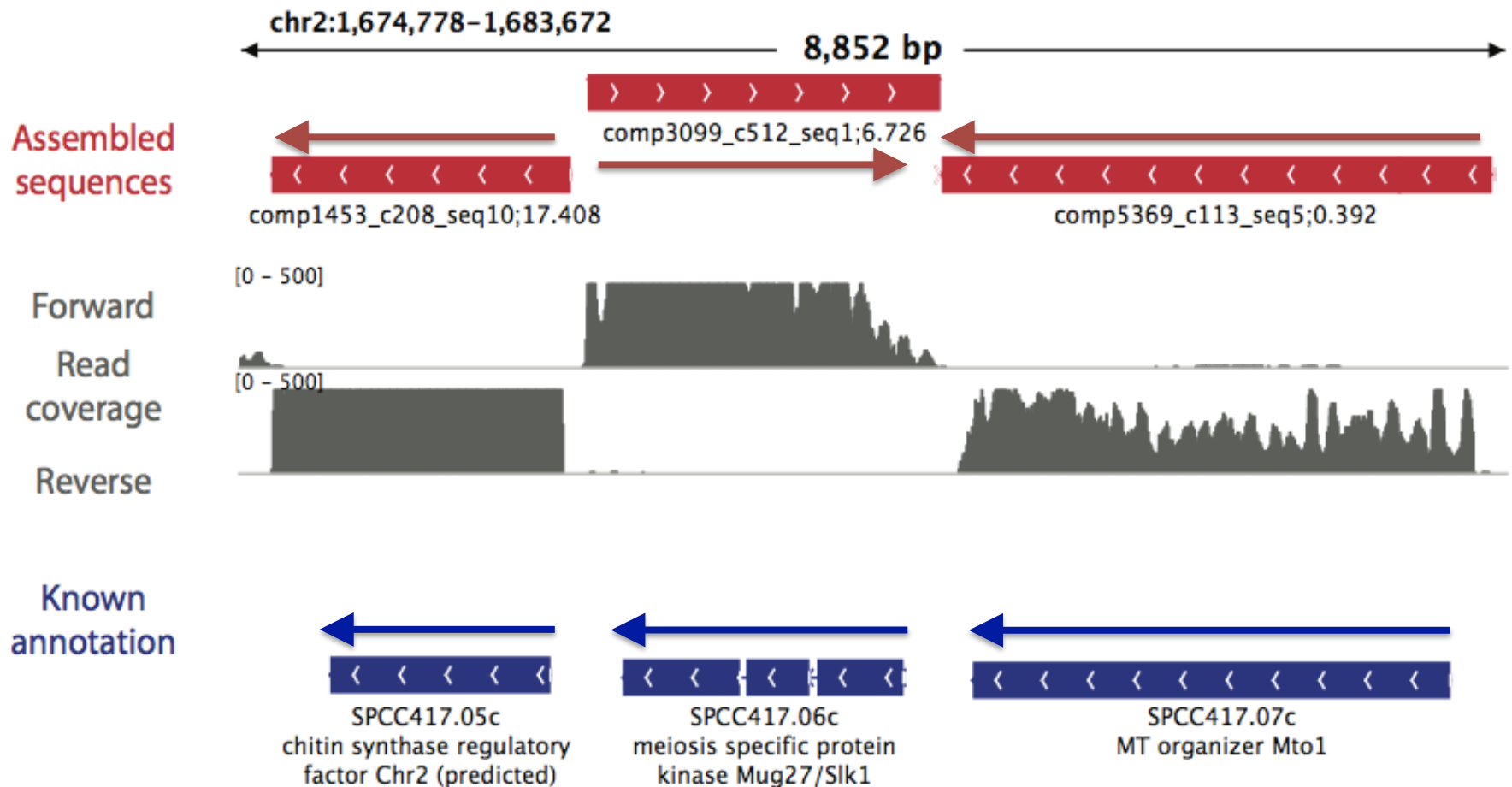
Overlapping UTRs from Opposite Strands



Schizosacharomyces pombe
(fission yeast)



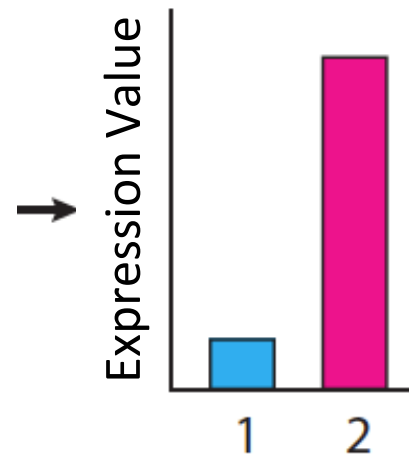
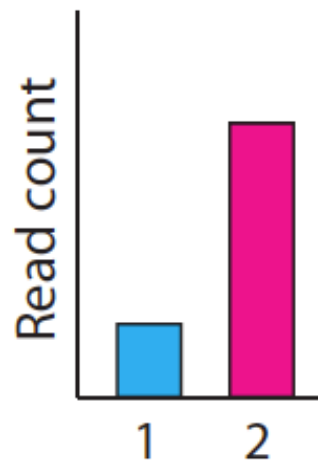
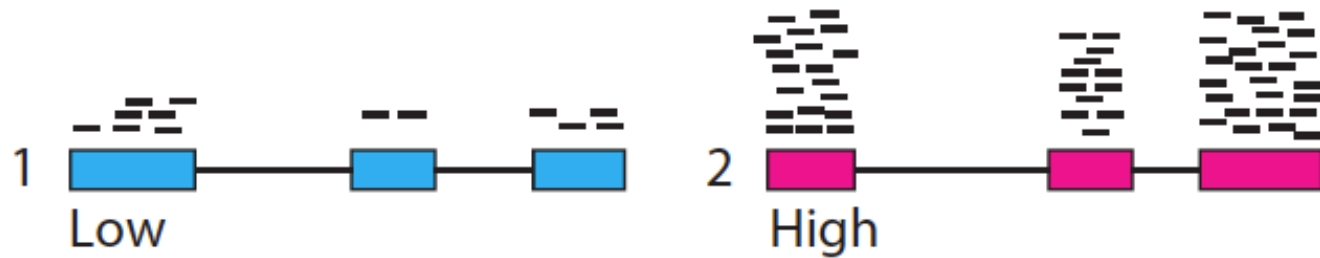
Antisense-dominated Transcription



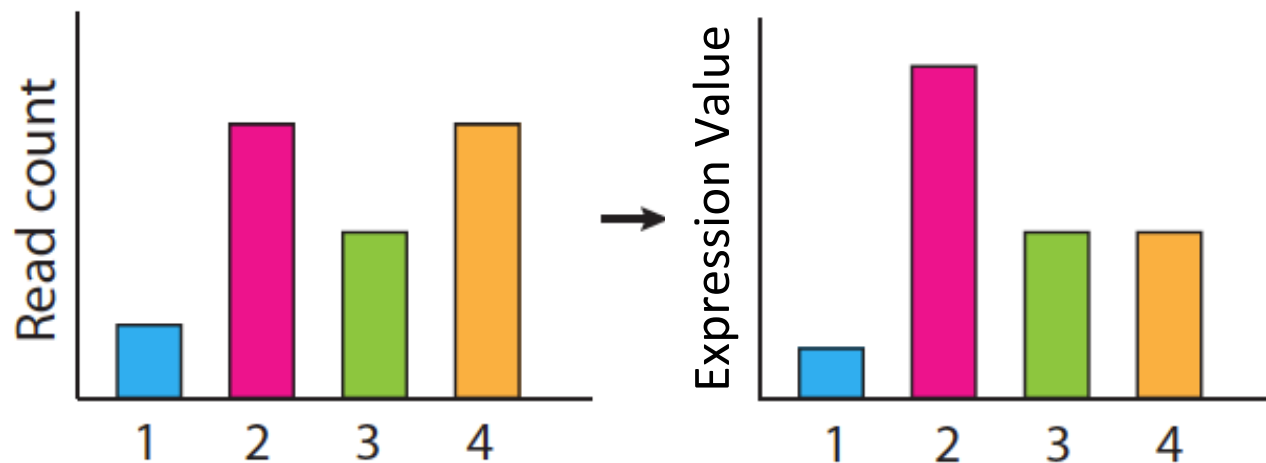
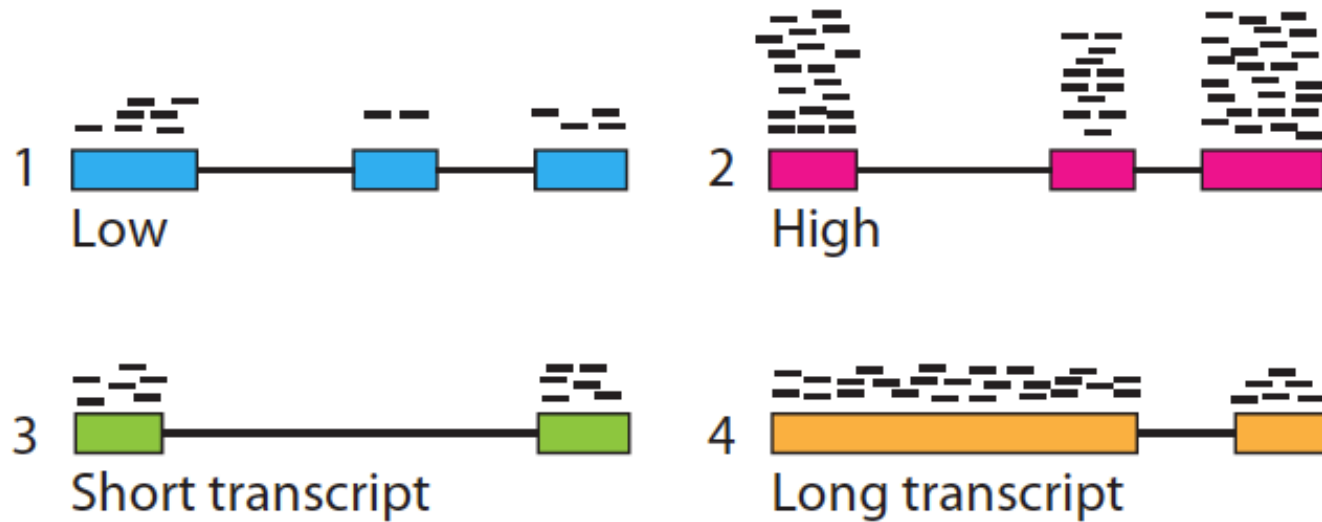
Abundance Estimation

(Aka. Computing Expression Values)

Calculating expression of genes and transcripts



Calculating expression of genes and transcripts



Normalized Expression Values

- Transcript-mapped read counts are normalized for both length of the transcript and total depth of sequencing.
- Reported as: Number of RNA-Seq **F**ragments
Per **K**ilobase of transcript
per total **M**illion fragments mapped
FPKM

RPKM (reads per kb per M) used with Single-end RNA-Seq reads
FPKM used with Paired-end RNA-Seq reads.

Transcripts per Million (TPM)

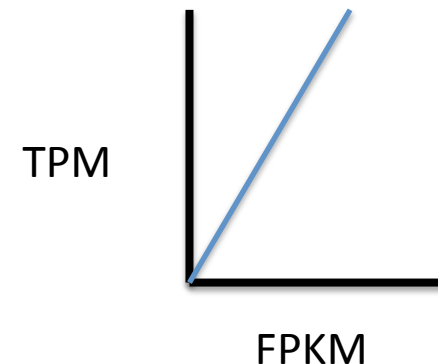
$$TPM_i = \frac{FPKM_i}{\sum_j FPKM} * 1e6$$

Preferred metric for measuring expression

- Better reflects transcript concentration in the sample.
- Nicely sums to 1 million

Linear relationship between TPM and FPKM values.

Both are valid metrics, but best to be consistent.

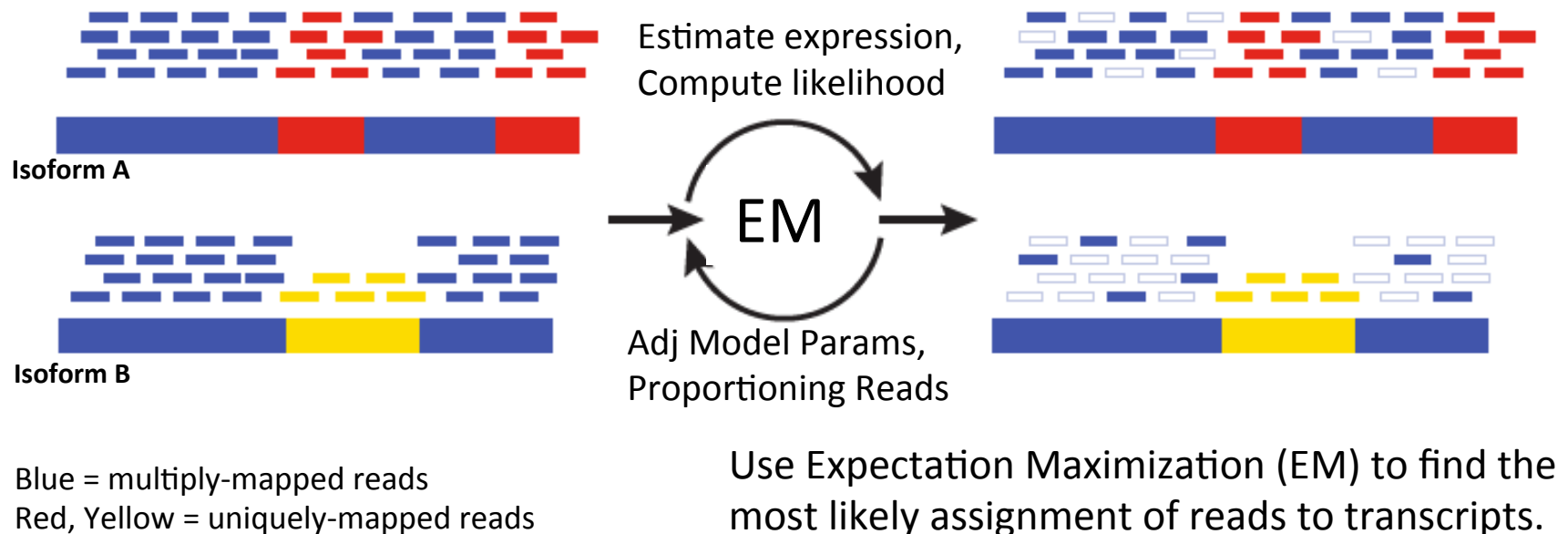


Multiply-mapped Reads Confound Abundance Estimation



Blue = multiply-mapped reads
Red, Yellow = uniquely-mapped reads

Multiply-mapped Reads Confound Abundance Estimation



Performed by:

- Cufflinks, String Tie (Tuxedo)
- RSEM, eXpress (genome-free)
- Kallisto, Salmon (alignment-free)

Expression Quantification Results

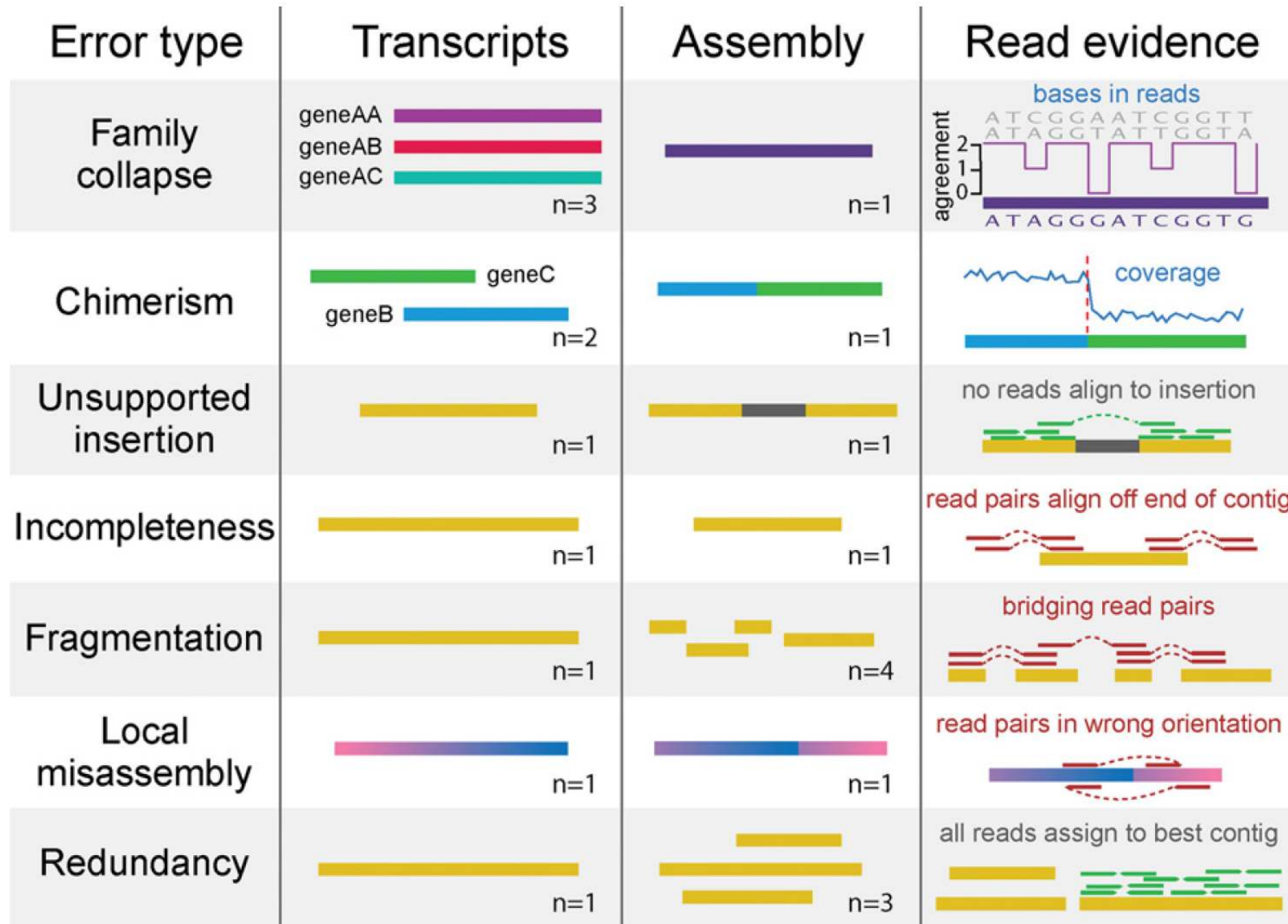
(ex. from Kallisto)

target_id	length	eff_length	est_counts	tpm
TRINITY_DN10_c0_g1_i1	334	100.489	13	4186.62
TRINITY_DN11_c0_g1_i1	319	87.9968	0	0
TRINITY_DN12_c0_g1_i1	244	38.2208	2	1693.43
TRINITY_DN17_c0_g1_i1	229	30.2382	5	5351.21
TRINITY_DN18_c0_g1_i1	633	384.493	19	1599.2
TRINITY_DN18_c1_g1_i1	289	65.795	1	491.864
TRINITY_DN19_c0_g1_i1	283	61.0618	10	5299.91

Evaluating the quality of your transcriptome assembly



De novo Transcriptome Assembly is Prone to Certain Types of Errors





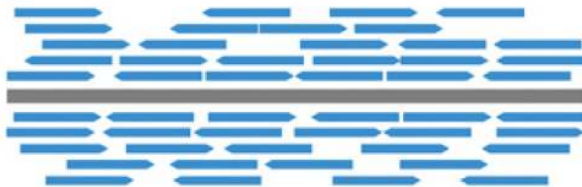
TransRate

1 input data

assembled contigs paired-end reads



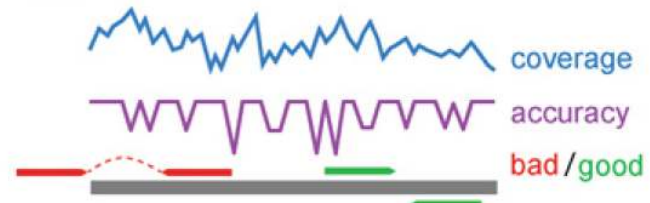
2 align reads to contigs



3 assign multimapping reads



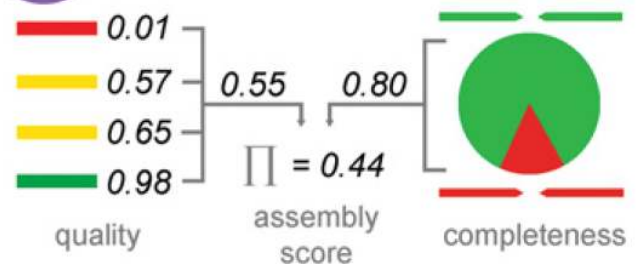
4 collect contig score components



5 calculate contig scores



6 calculate assembly score



Simple Quantitative and Qualitative Assembly Metrics

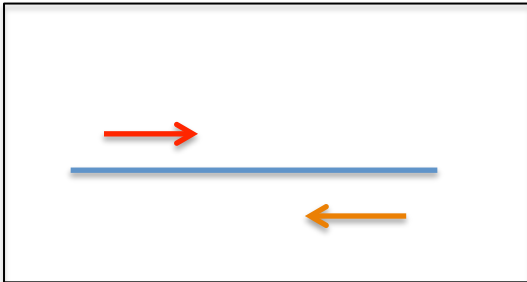
Read representation by assembly

Align reads to the assembled transcripts using Bowtie.

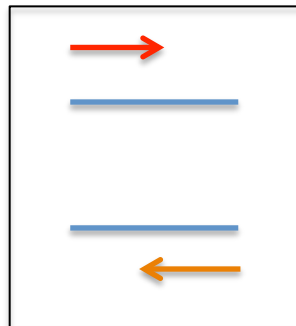
A typical 'good' assembly has ~80 % reads mapping to the assembly and ~80% are properly paired.

Given read pair:   Possible mapping contexts in the Trinity assembly are reported:

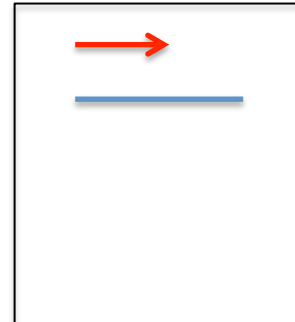
Proper pairs



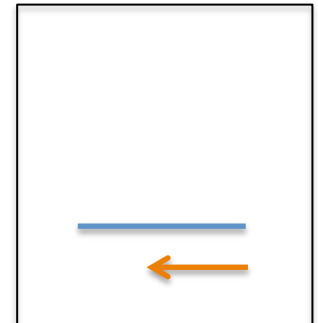
Improper pairs



Left only



Right only



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

% samtools tview alignments.bam target.fasta

	911	921	931	941	951	961	971	981	991	1001	1011	1021	1031	1041	1051	1061	1071
	GTAGGTTTAATTTTCATCTTCTAATTTAGAATCTTGCCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTACATTACTATAAATTGGTGTATCGGGTCTTCCAACCTCTCCATTCAAGACTTAATTGACTCTGT																
GT	GTTTAATTTTCATCTTCTAATTTAGAATCTTGCCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAAC								ctgcttctgagattctaagtagcttagatgccaagtacattactataaattggtgttatcgggtcttcc				ctctccattcaagacttaattgactctgt				
GT	ATTTTCATCTTCTAATTTAGAATCTTGCCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAAC								tgcttctgagattctaagtagcttagatgccaagtacattactataaattggtgttatcgggtcttcca				ctctccattcaagacttaattgactctgt				
GT	atttcatcttctaatttagaattcttgccaatcaagccctctcgaagttggcaatctataactcaac								GCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTACATTACTATAAATTGGTGTATCGGGTCTTCCAA				ctctccattcaagacttaattgactctgt				
GT	atttcatcttctaatttagaattcttgccaatcaagccctctcgaagttggcaatctataactcaac								GCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTACATTACTATAAATTGGTGTATCGGGTCTTCCAA				ctctccattcaagacttaattgactctgt				
GTAGGTTTAAT	aatcttgccaatcaagccctctcgaagttggcaatctataactcaacctctgcttctgagattcta								CTTAGATGCCAAGTACATTACTATAAATTGGTGTATCGGGTCTTCCAACCTCTCCATTCAAGACTTAA				ctgt				
GTAGGTTTAATTT	tcttgccaatcaagccctctcgaagttggcaatctataactcaacctctgcttctgagattctaag								CTTAGATGCCAAGTACATTACTATAAATTGGTGTATCGGGTCTTCCAACCTCTCCATTCAAGACTTAA								
GTAGGTTTAATTTTCATCTT	cttgccaatcaagccctctcgaagttggcaatctataactcaacctctgcttctgagattctaag								TTAGATGCCAAGTACATTACTATAAATTGGTGTATCGGGTCTTCCAACCTCTCCATTCAAGACTTAAT								
GTAGGTTTAATTTTCATCTTC	TGCCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTAC								ATGCCAAGTACATTACTATAAATTGGTGTATCGGGTCTTCCAACCTCTCCATTCAAGACTTAATTGAC								
GTAGGTTTAATTTTCATCTTCTAAT	TGCCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTAC								GCCAAGTACATTACTATAAATTGGTGTATCGGGTCTTCCAACCTCTCCATTCAAGACTTAATTGACTC								
gtaggtttaatttcattcttctaatttag	TGCCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTAC								CATTACTATAAATTGGTGTATCGGGTCTTCCAACCTCTCCATTCAAGACTTAATTGACTCTGT								
GTAGGTTTAATTTTCATCTTCTAATTTAG	GCCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTACC								cattactataaattggtgttatcgggtcttccaactctctccattcaagacttaattgactctgt								
GTAGGTTTAATTTTCATCTTCTAATTTAG	CAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTACC								tgttatcgggtcttccaactctctccattcaagacttaattgactctgt								
GTAGGTTTAATTTTCATCTTCTAATTTAG	CAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTACCCTT																
GTAGGTTTAATTTTCATCTTCTAATTTAG	gacctctcgaagttggcaatctataactcaacctctgcttctgagattctaagtagcttagatgcc																
GTAGGTTTAATTTTCATCTTCTAATTTAGAAT	CCCTCTCGAAGTTGGCAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTACCTTAGATGCCA																
GTAGGTTTAATTTTCATCTTCTAATTTAGAATCT	ctctcgaagttggcaatctataactcaacctctgcttctgagattctaagtagcttagatgccaag																
GTAGGTTTAATTTTCATCTTCTAATTTAGAATCT	CTCGAAGTTGGCAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTAC																
GTAGGTTTAATTTTCATCTTCTAATTTAGAATCT	CGAAGTTGGCAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTAC																
GTAGGTTTAATTTTCATCTTCTAATTTAGAATCT	AAGTTGGCAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTACATT																
gtaggtttaatttcattcttctaatttagaattctg	CAATATCTATAACTCAACCTCTGCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTACATTACTATAA																
GTAGGTTTAATTTTCATCTTCTAATTTAGAATCTTGCCA	CTATAACTCAACCTCTGCTTCTGAGATTCTAAGTACCTTAGATGCCAAGTACATTACTATAAATTGGT																
GTAGGTTTAATTTTCATCTTCTAATTTAGAATCTTGCCA									cttctgagattctaagtagcttagatgccaagtacattactataaattggtgttatcgggtcttccaac				CTCCATTCAAGACTTAATTGACTCTGT				
GTAGGTTTAATTTTCATCTTCTAATTTAGAATCTTGCCAATCAAGCC									cttctgagattctaagtagcttagatgccaagtacattactataaattggtgttatcgggtcttccaac				tccattcaagacttaattgactctgt				
gtaggtttaatttcattcttctaatttagaattcttgccaatcaagccc									tctctgagattctaagtagcttagatgccaagtacattactataaattggtgttatcgggtcttccaac				tccattcaagacttaattgactctgt				
GTAGGTTTAATTTTCATCTTCTAATTTAGAATCTTGCCAATCAAGCC									tgagattctaagtagcttagatgccaagtacattactataaattggtgttatcgggtcttccaactcc				cattcaagacttaattgactctgt				
GTAGGTTTAATTTTCATCTTCTAATTTAGAATCTTGCCAATCAAGCCCTC									tgagattctaagtagcttagatgccaagtacattactataaattggtgttatcgggtcttccaactctc				tcaagacttaattgactctgt				
GTAGGTTTAATTTTCATCTTCTAATTTAGAATCTTGCCAATCAAGCCCTCTCGAAG									gagattctaagtagcttagatgccaagtacattactataaattggtgttatcgggtcttccaactctcc				AAGACTTAATTGACTCTGT				
GTAGGTTTAATTTTCATCTTCTAATTTAGAATCTTGCCAATCAAGCCCTCTCGAAG									agattctaagtagcttagatgccaagtacattactataaattggtgttatcgggtcttccaactctcc				cttaattgactctgt				
ATTTTCATCTTCTAATTTAGAATCTTGCCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAAC									AGATTCTAAGTACCTTAGATGCCAAGTACATTACTATAAATTGGTGTATCGGGTCTTCCAACCTCTCC				attgactctgt				
TTTCATCTTCTAATTTAGAATCTTGCCAATCAAGCCCTCTCGAAGTTGGCAATATCTATAACTCAACCT									gattcctaagtagcttagatgccaagtacattactataaattggtgttatcgggtcttccaactctcca								
									gattcctaattaccttagatgccaagtacattactataaattggtgttatcgggtcttccaactctcca								
									aagtagcttagatgccaagtacattactataaattggtgttatcgggtcttccaactctccattcaag								
													cttccaactctccattcaagacttaattgactctgt				
													TTCCAACCTCTCCATTCAAGACTTAATTGACTCTGT				
													TCCAACCTCTCCATTCAAGACTTAATTGACTCTGT				
													caactctccattcaagacttaattgactctgt				
													caactctccattcaagacttaattgactctgt				
													aactctccattcaagacttaattgactctgt				
													aactctccattcaagacttaattgactctgt				
													tccattcaagacttaattgactctgt				
													ccattcaagacttaattgactctgt				
													ccattcaagacttaattgactctgt				

IGV

← → ↻ www.broadinstitute.org/igv/ ☆ a

**Integrative Genomics Viewer**

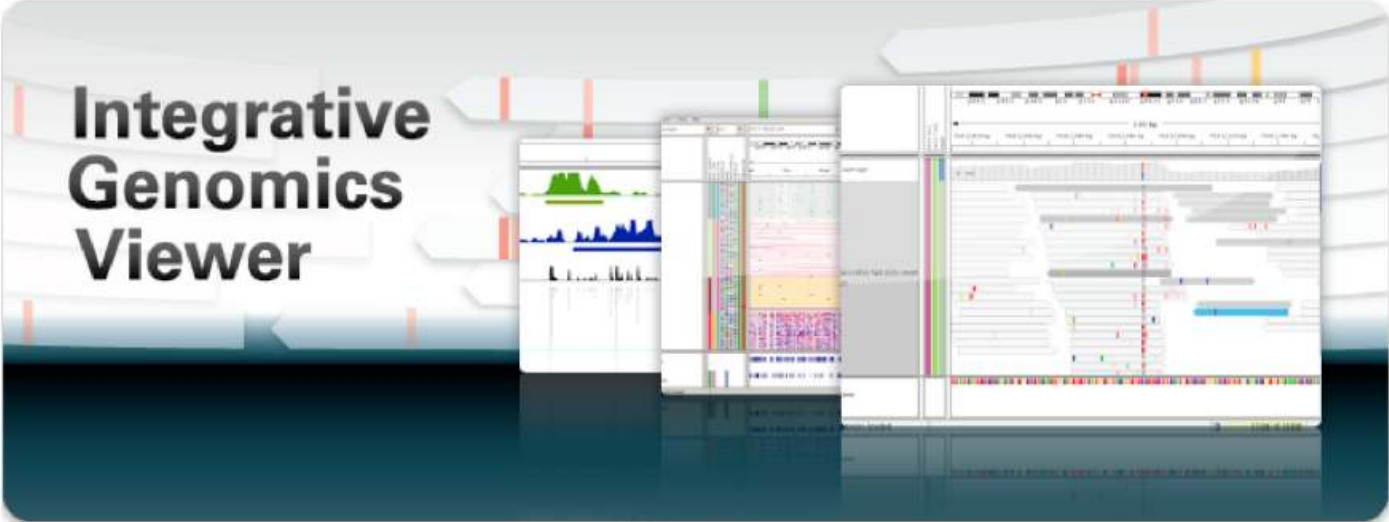
[Home](#)
[Downloads](#)
[Documents](#)
→ [Hosted Genomes](#)
→ [FAQ](#)
+ [IGV User Guide](#)
+ [File Formats](#)
+ [Release Notes](#)
→ [Credits](#)
@ [Contact](#)

Search website

[Broad Home](#)
[Cancer Program](#)

© 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.

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](#).

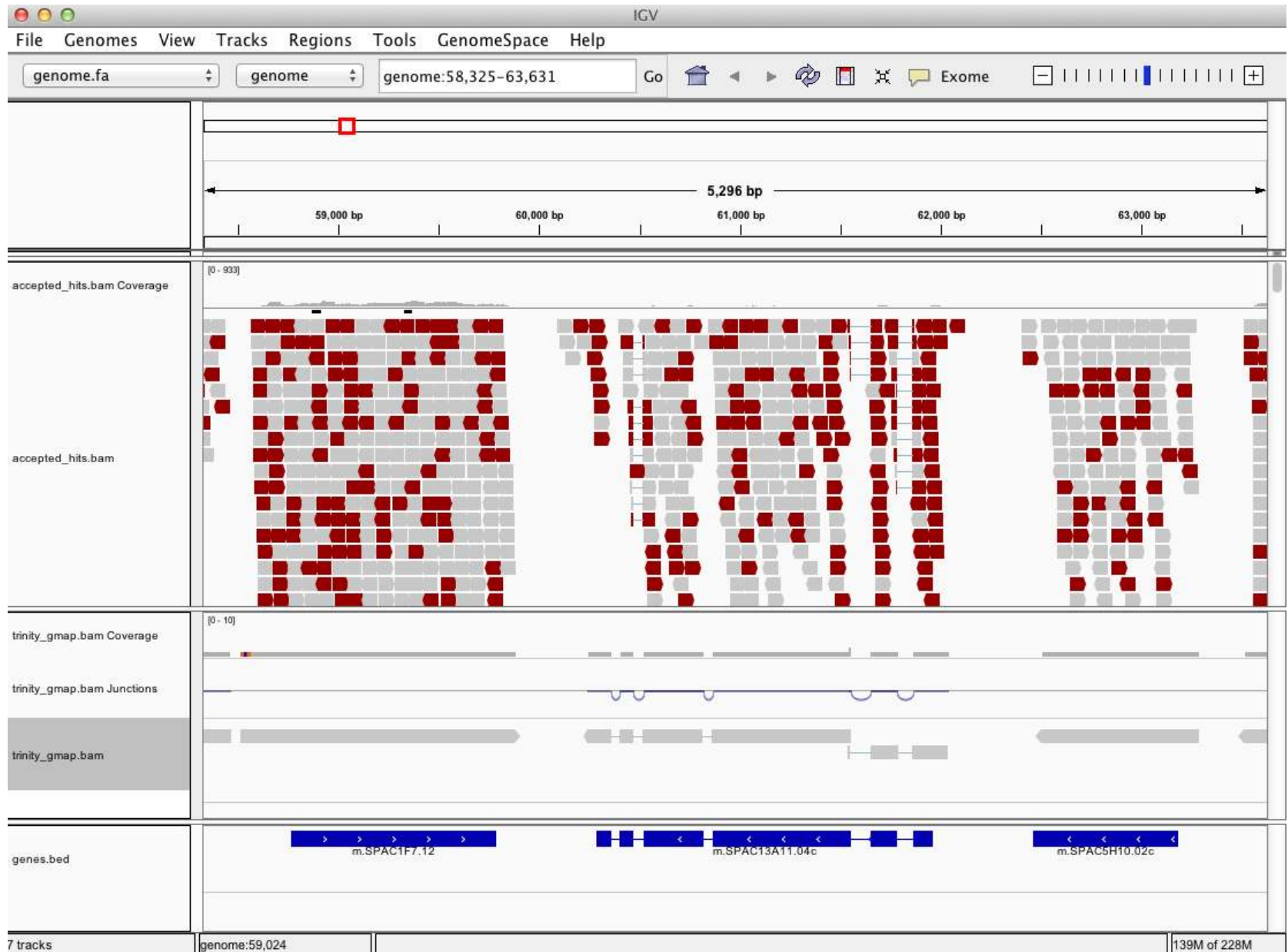
Overview

Can Examine Transcript Read Support Using IGV



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

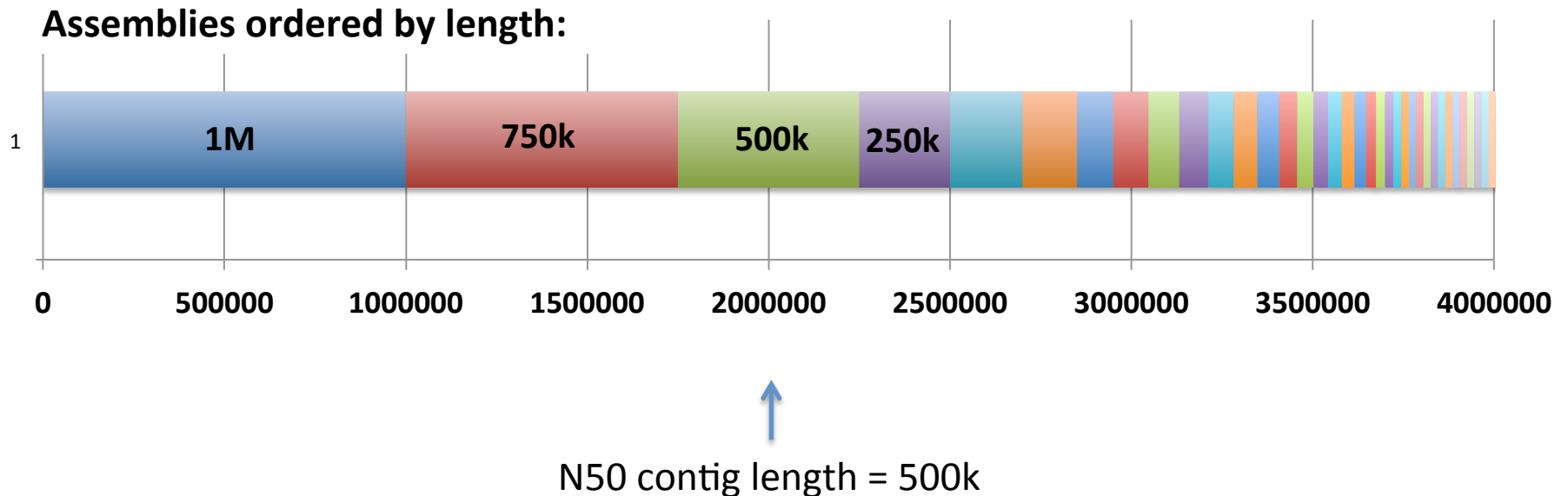
(Trinity transcripts aligned to the genome using GMAP)



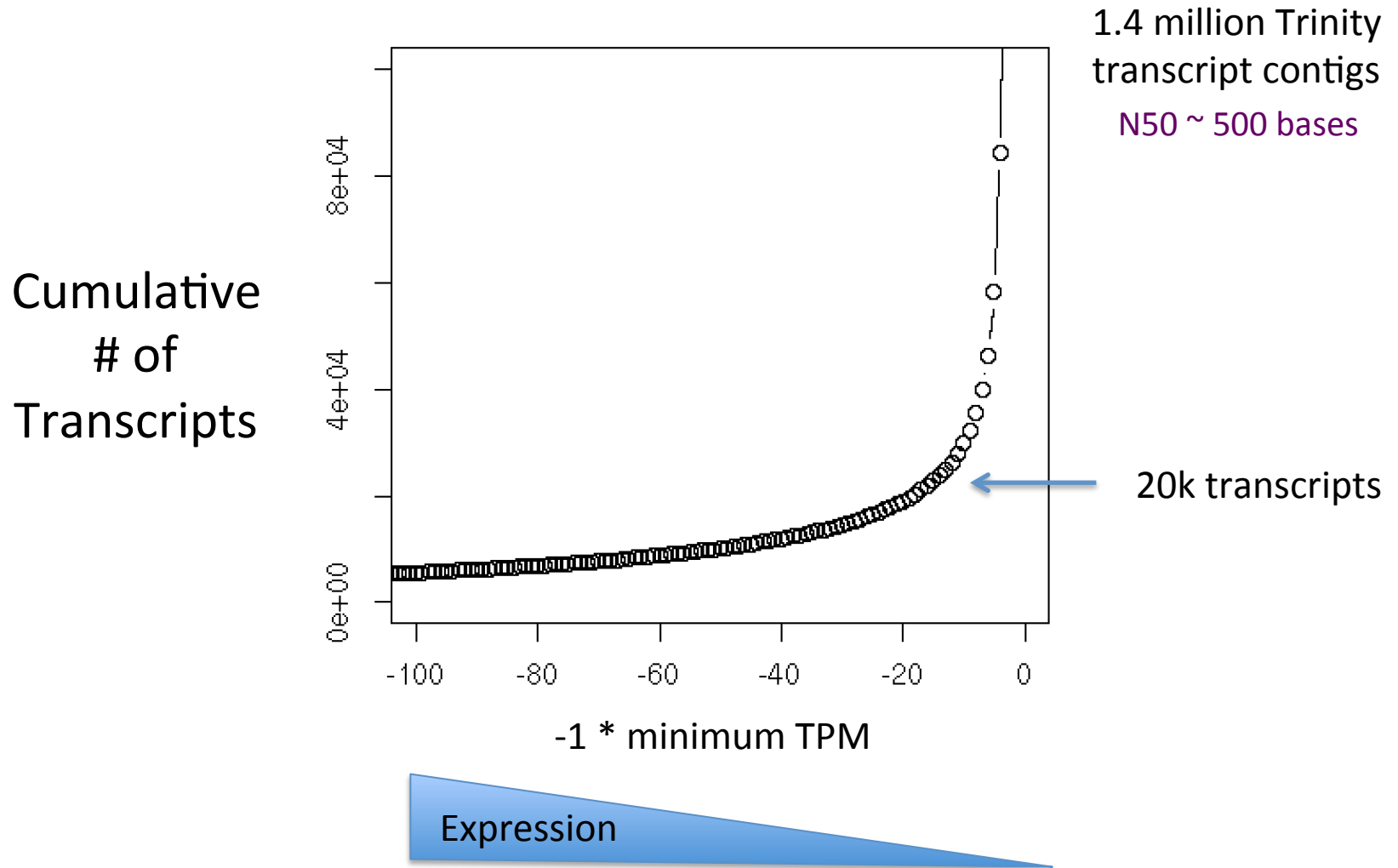
The Contig N50 statistic

“At least half of assembled bases are in contigs that are at least **N50** bases in length”

In genome assemblies – used often to judge ‘which assembly is better’

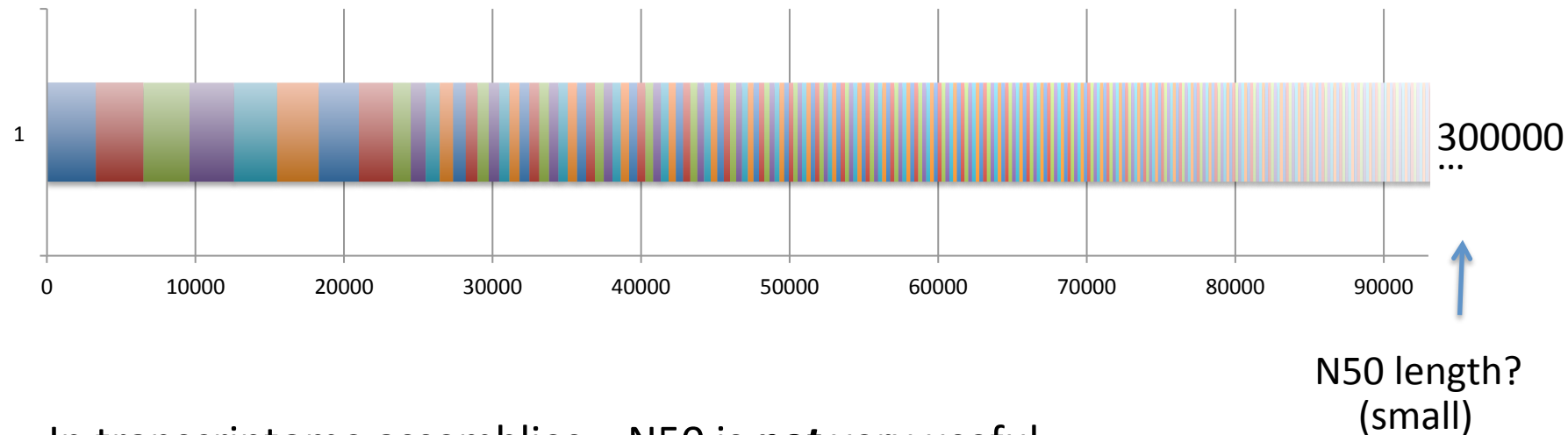


Often, most assembled transcripts are **very lowly expressed**
(How many 'transcripts & genes' are there really?)



* Salamander transcriptome

N50 Calculation for *Transcriptome* Assemblies??

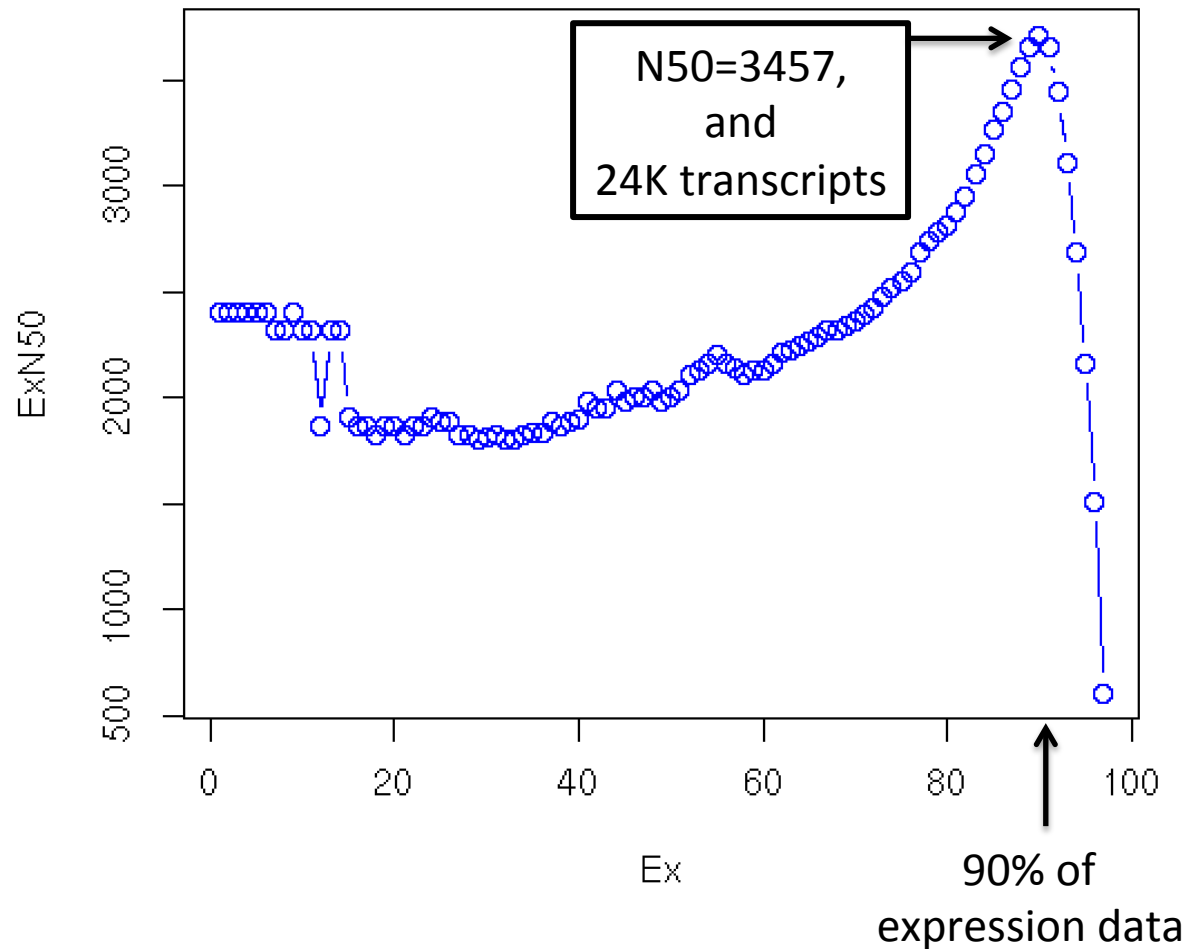


In transcriptome assemblies – N50 is **not** very useful.

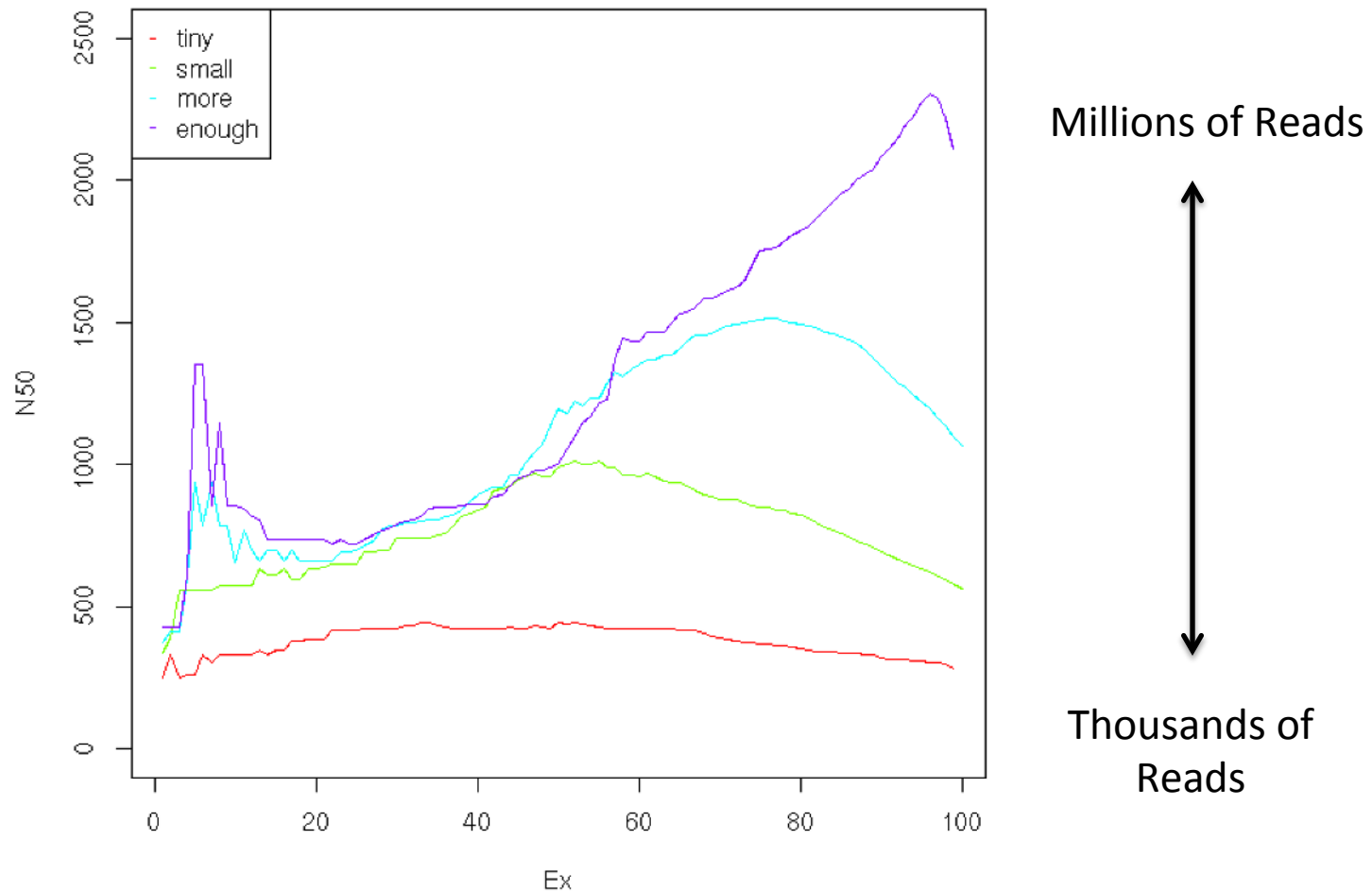
- Overzealous isoform annotation for long transcripts drives higher N50
- Very sensitive reconstruction for short lowly expressed transcripts drives lower N50

Compute N50 Based on the Top-most Highly Expressed Transcripts (ExN50)

- Sort contigs by expression value, descendingly.
- Compute N50 given minimum % total expression data thresholds => ExN50



ExN50 Profiles for Different Trinity Assemblies Using Different Read Depths

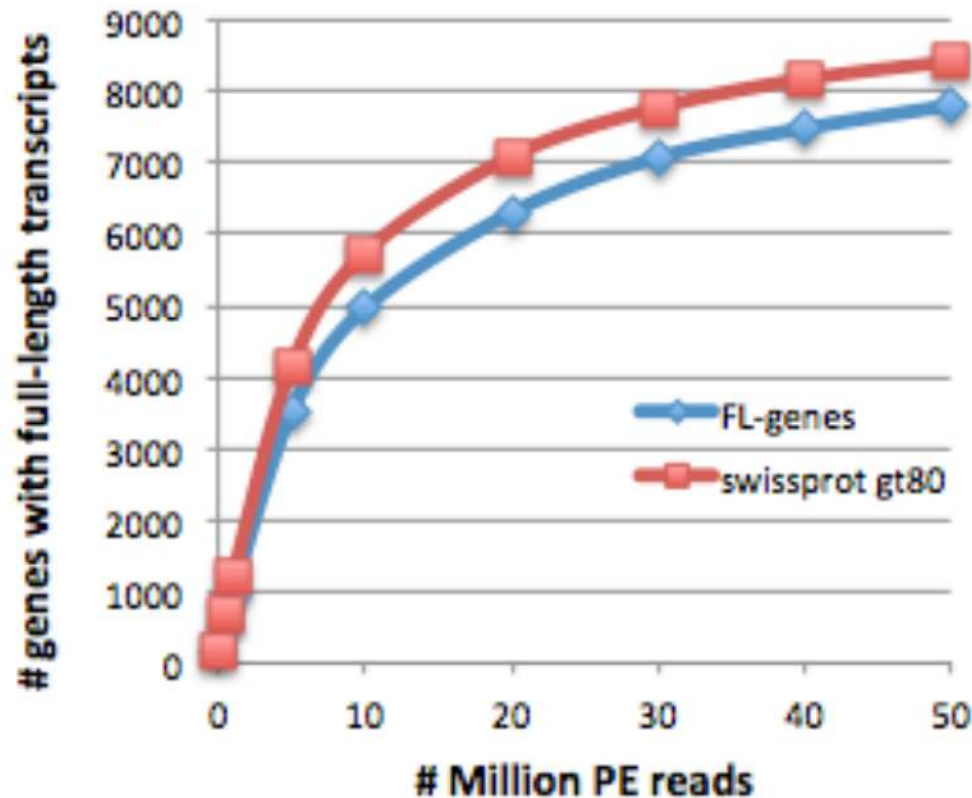


Note shift in ExN50 profiles as you assemble more and more reads.

* Candida transcriptome

Evaluating the quality of your transcriptome assembly

Full-length Transcript Detection via BLASTX



Have you
sequenced
deeply
enough?



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.



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

Detonate: Which assembly is better?

“RSEM-EVAL [sic] uses a novel probabilistic model-based method to compute the joint probability of both an assembly and the RNA-Seq data as an evaluation score.”

$$\text{score}_{\text{RSEM-EVAL}}(A) = \log P(A, D)$$

“the RSEM-EVAL score of an assembly is defined as the log joint probability of the assembly A and the reads D used to construct it”

$$\begin{aligned} \log P(A, D) &= \log \int_{\Lambda} P(D|A, \Lambda) P(A|\Lambda) P(\Lambda) d\Lambda \\ &\approx \underbrace{\log P(D|A, \Lambda_{\text{MLE}})}_{\text{likelihood}} + \underbrace{\log P(A|\Lambda_{\text{MLE}})}_{\text{assembly prior}} \\ &\quad - \underbrace{\frac{1}{2}(M+1) \log N}_{\text{BIC penalty}}, \end{aligned}$$

Detonate: Which assembly is better?

“RSEM-EVAL [sic] uses a novel probabilistic model-based method to compute the joint probability of both an assembly and the RNA-Seq data as an evaluation score.”

$$\text{score}_{\text{RSEM-EVAL}}(A) = \log P(A, D)$$

“the RSEM-EVAL score of an assembly is defined as the log joint probability of the assembly A and the reads D used to construct it”

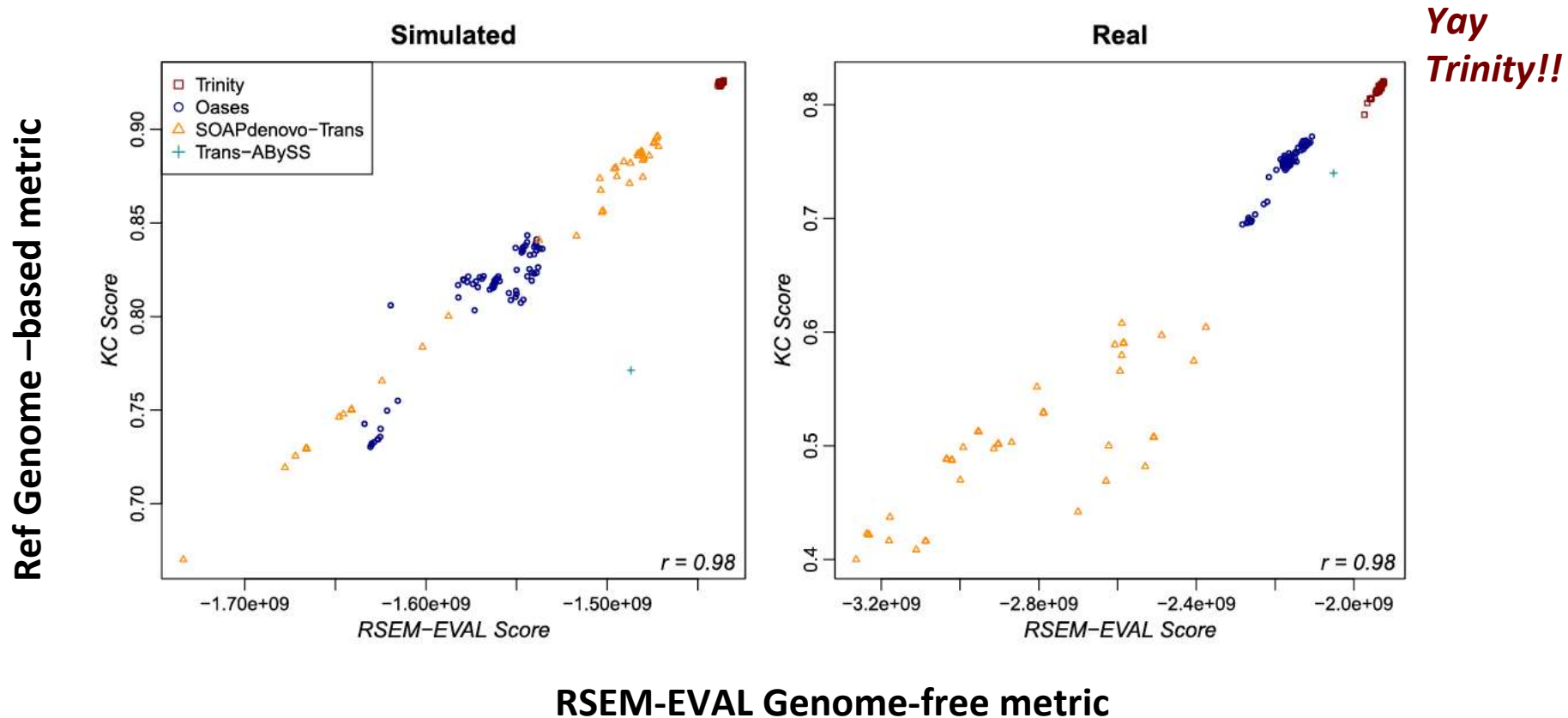
$$\begin{aligned} \log P(A, D) &= \log \int_{\Lambda} P(D|A, \Lambda) P(A|\Lambda) P(\Lambda) d\Lambda \\ &\approx \underbrace{\log P(D|A, \Lambda_{\text{MLE}})} + \underbrace{\log P(A|\Lambda_{\text{MLE}})} \end{aligned}$$

Bigger Score = Better Assembly

$$- \underbrace{\frac{1}{2}(M+1) \log N}_{\text{BIC penalty}}$$

Detonate: Which assembly is better?

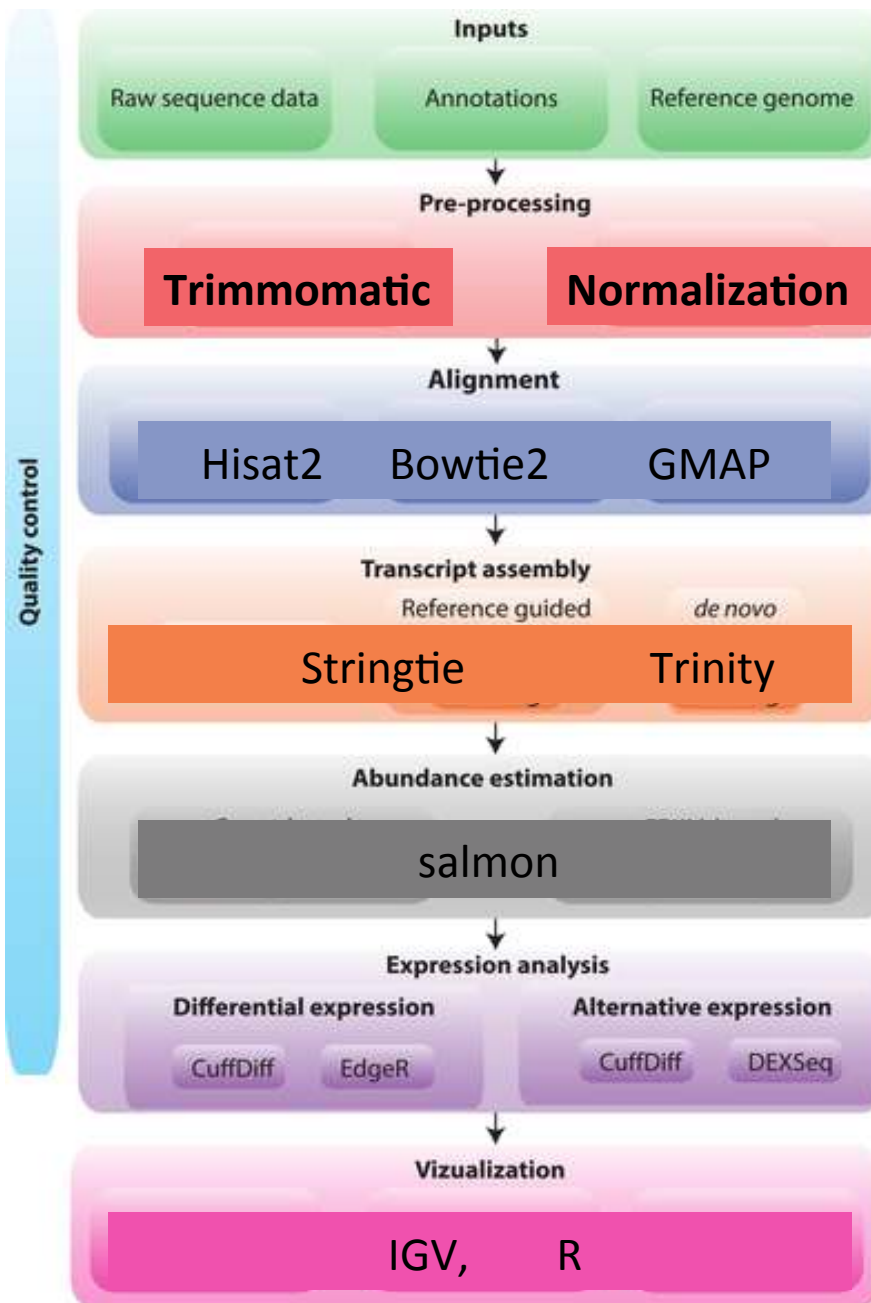
“RSEM-EVAL [sic] uses a novel probabilistic model-based method to compute the joint probability of both an assembly and the RNA-Seq data as an evaluation score.”



Hands-on Workshop Activities

FastQC, MultiQC

Full-length BlastX,
ExN50



(Tomorrow)