

The Krumlov Trinity Transcriptomics Experience



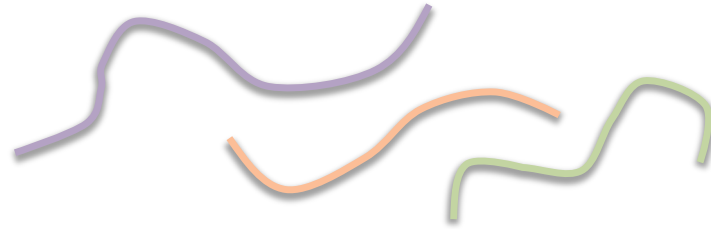
Brian Haas
Broad Institute

Workshop on Genomics, Cesky Krumlov, Jan 2019

Biological Investigations Empowered by Transcriptomics

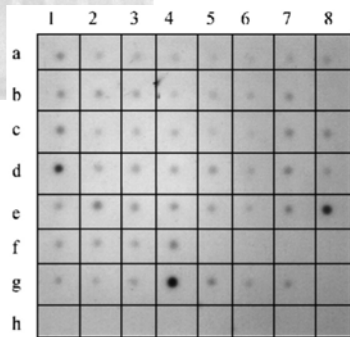
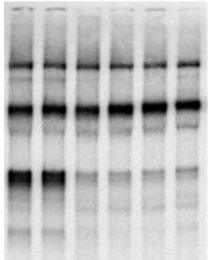


Extract RNA,
... some protocol for processing, ...

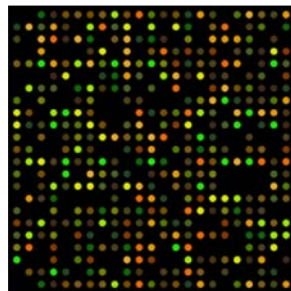


Analysis Method
(pick your favorite)

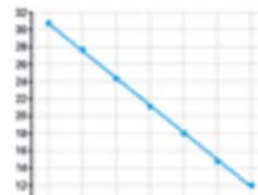
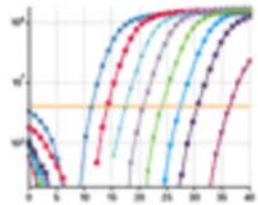
Northern



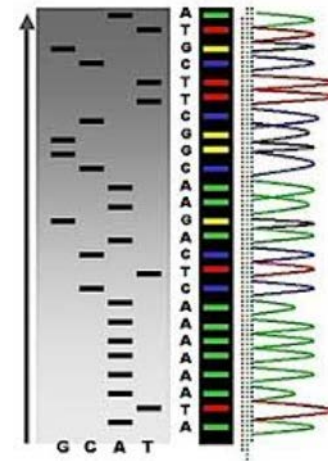
Dot Blot



Microarray



qRT-PCR



Sanger Sequencing



Other...



Minion



MinION Mk1: portable, real time biological analyses

MinION

Historical Timeline of Transcriptomics (from 1970)

Reverse Transcription (1970)

Northern Blot
Sanger Sequencing
(1977)

Expressed Sequence Tags (1992)

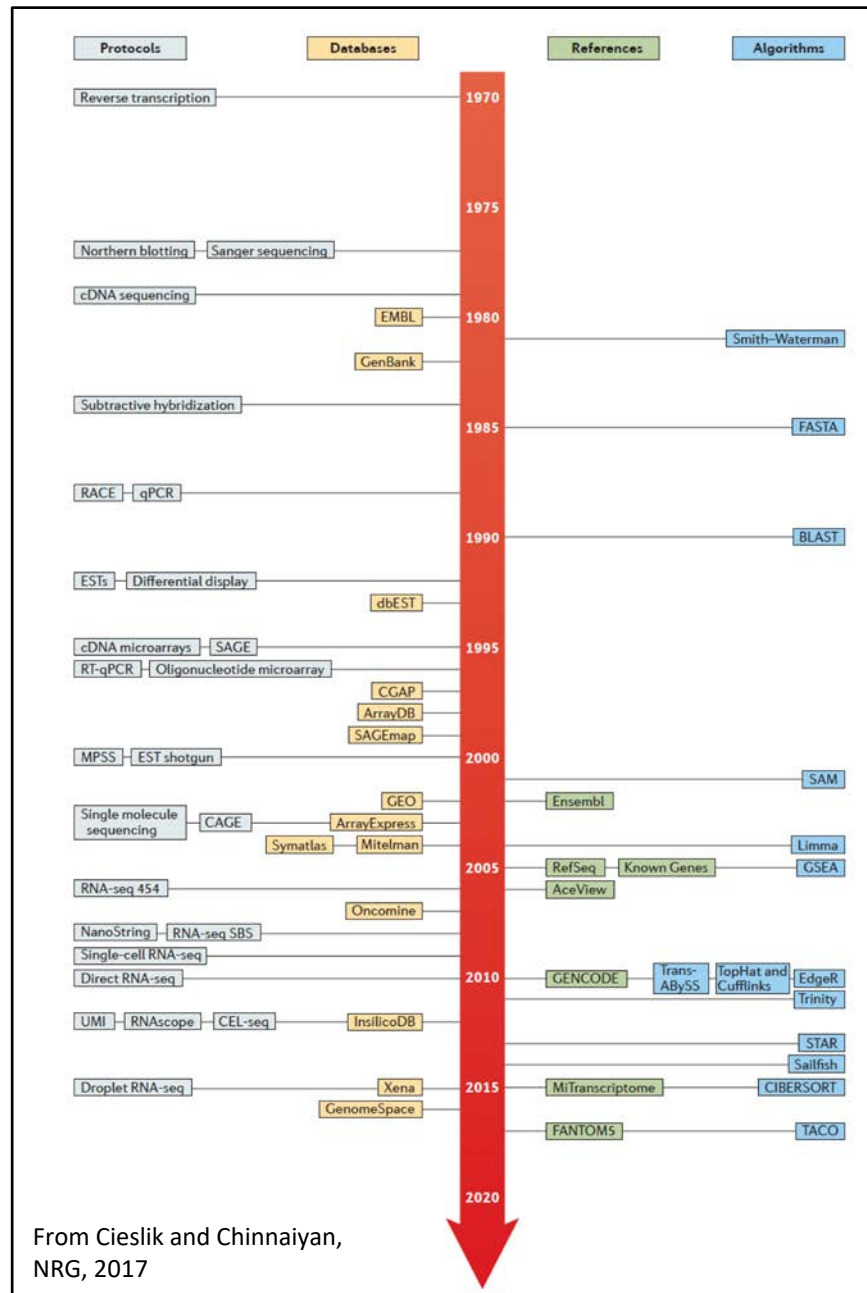
cDNA microarrays (1995)

RNA-Seq (2006-2008)

PacBio IsoSeq (2014)

Droplet single cell RNA-Seq (2015)

Direct RNA Seq Nanopore (2018)



From Cieslik and Chinnaiyan,
NRG, 2017

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

Smith Waterman (1981)

BLAST (1990)

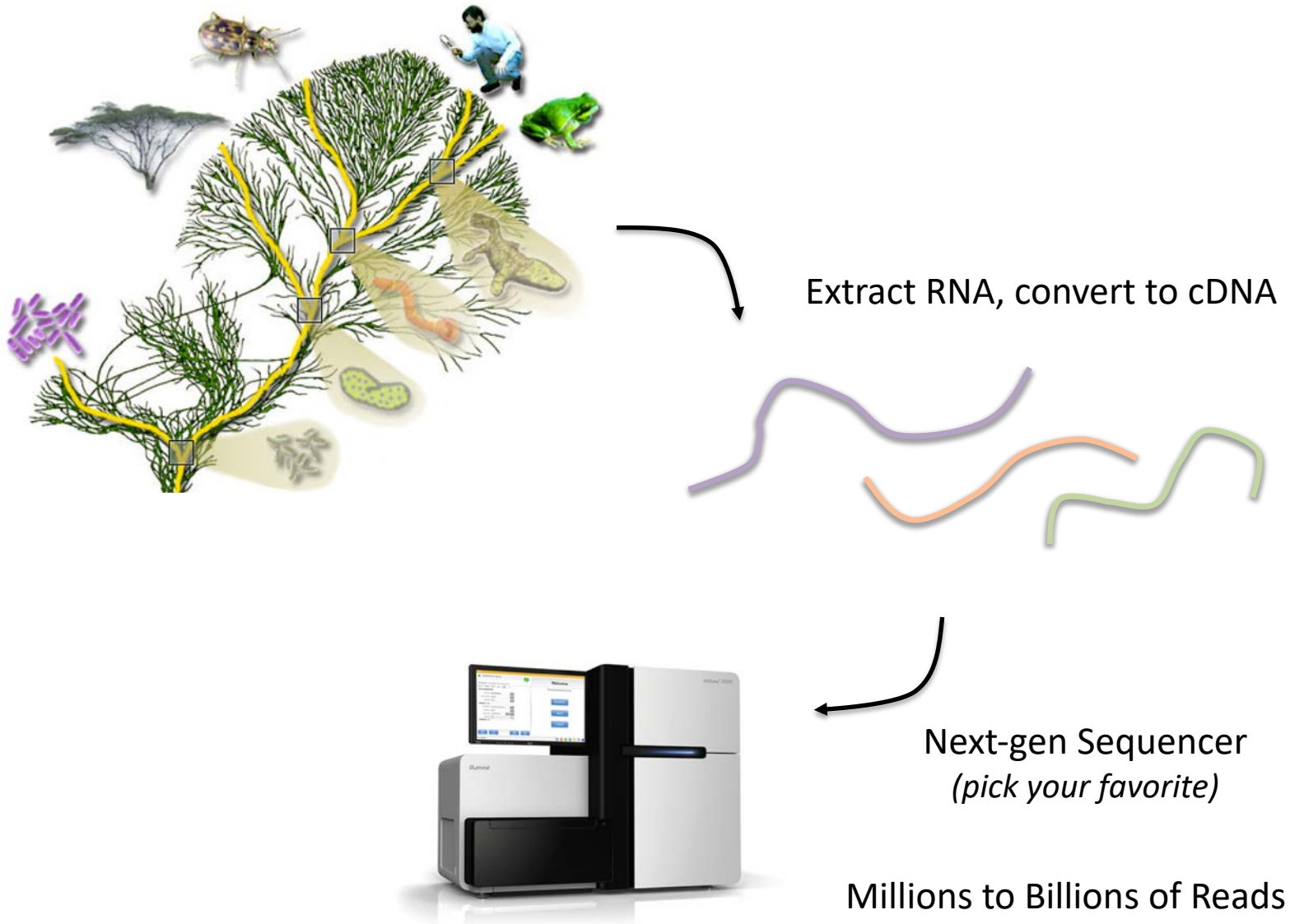
Tophat/Cufflinks (2010)



RSEM
(2011)

Kallisto (2016)
Salmon (2017)

Modern Transcriptome Studies Empowered by RNA-seq



Generating RNA-Seq: *How to Choose?*

Platform	Project Firefly 2018	MiniSeq	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	200	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/ngsspecshared>

#curl "https://docs.google.com/spreadsheets/d/1GMMfhYLK0-q8Xklo3YxlWaZA5vVMuhU1kg41g4xLkXc/export?gid=4&format=csv" | grep -v '^#' | grep -v '^\$' | column -t -s, | less -S

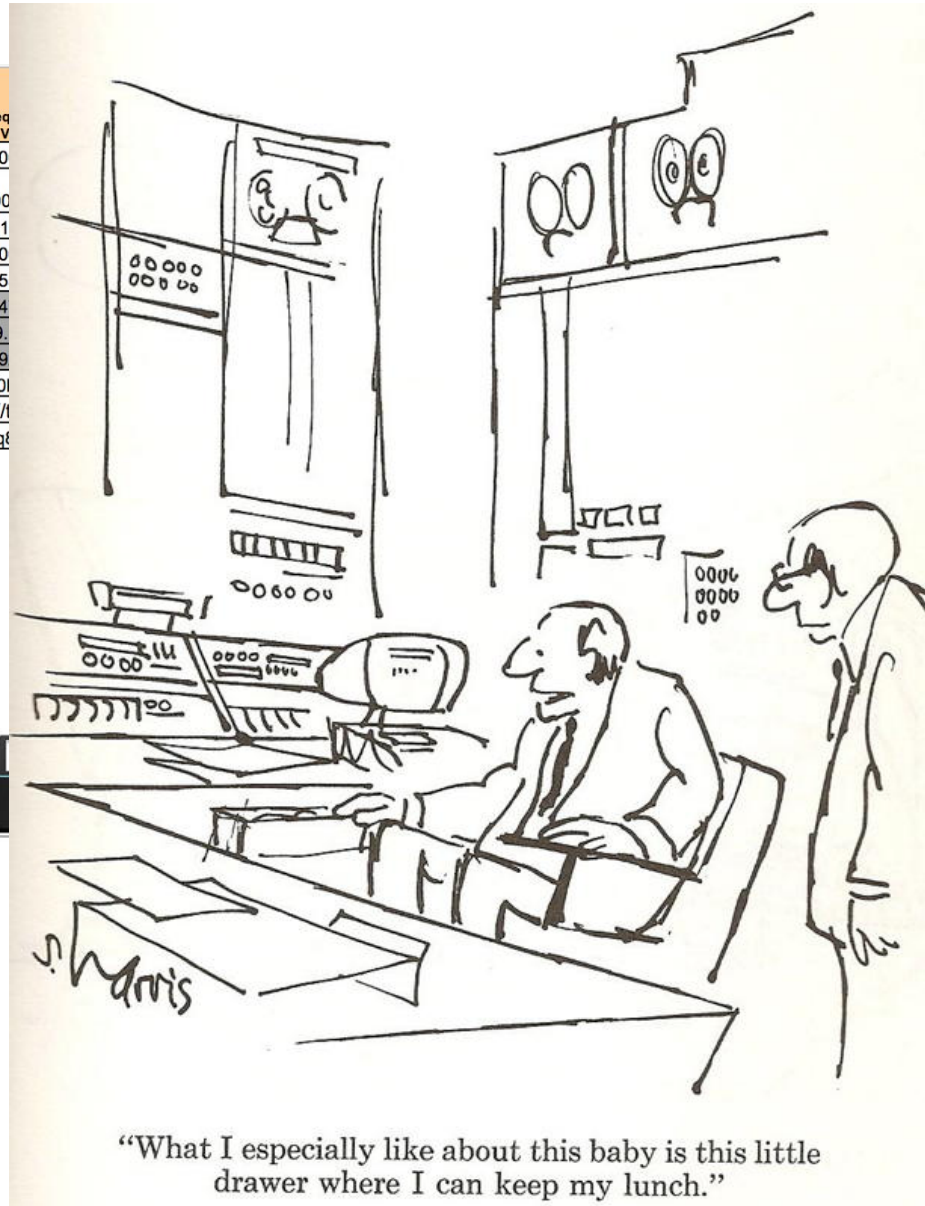


*Not all shown at scale

Generating RNA-Seq: *How to Choose?*

Platform	Project Firefly 2018	MiniSeq	HiSeq	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	690

#Page maintained by <http://twitter.com/albertvilella> <http://i>
[#curl "https://docs.google.com/spreadsheets/d/1GMMfhYLK0-q/](https://docs.google.com/spreadsheets/d/1GMMfhYLK0-q/)



g	Mini ION R9.5	Grid ION X5	Prome thION RnD	Prome thION theor etical	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	--	--



Small to Large

Each has pros/cons



Small to Less Large

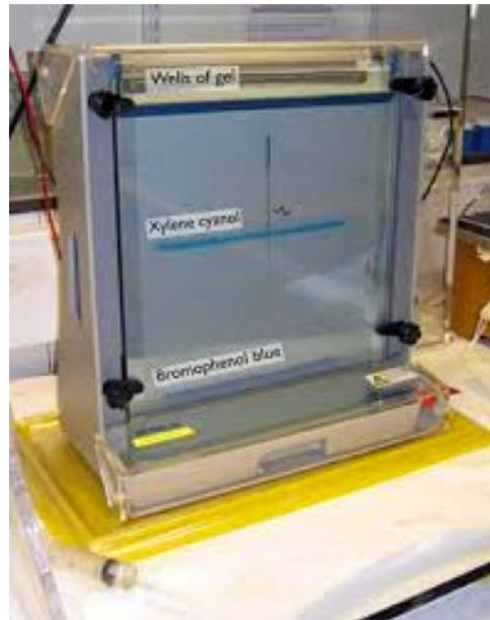
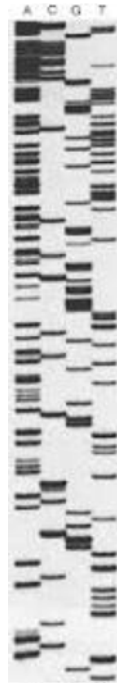


Each technology continues to advance

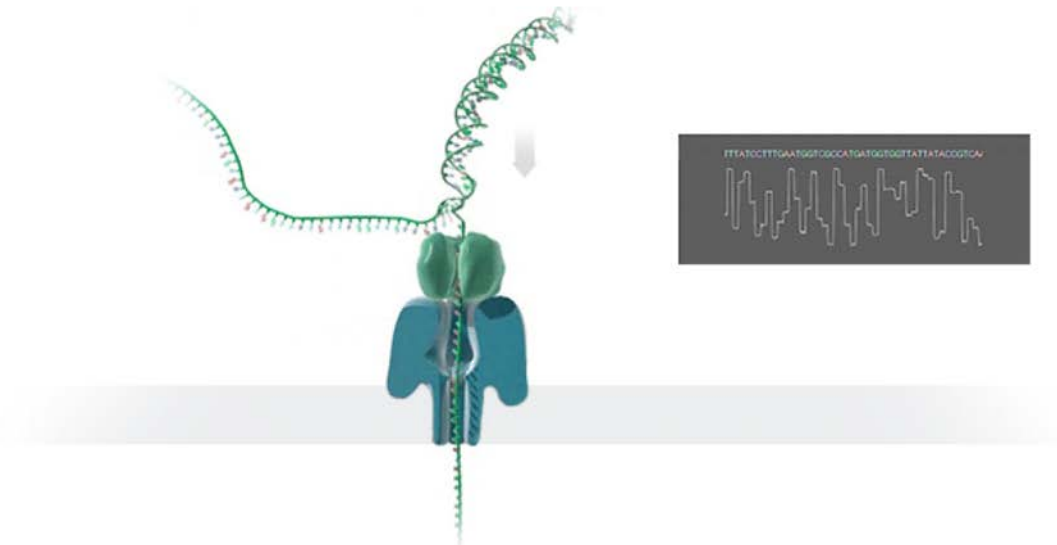


Personal Reflections... (and haunting memories)

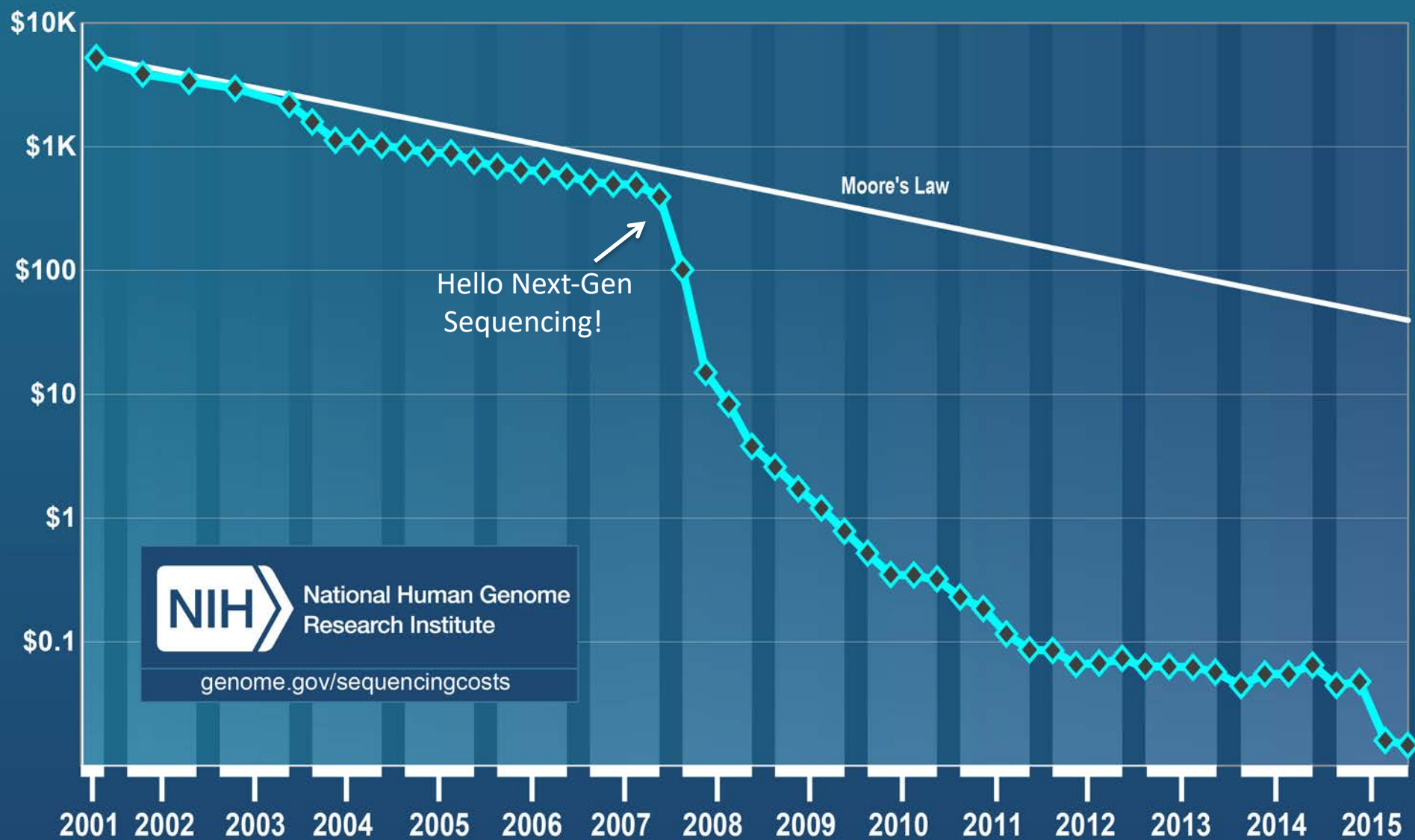
Circa 1995



Now



Cost per Raw Megabase of DNA Sequence



National Human Genome
Research Institute

genome.gov/sequencingcosts

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

A Plethora of Biological Sequence Analyses Enabled by RNA-Seq

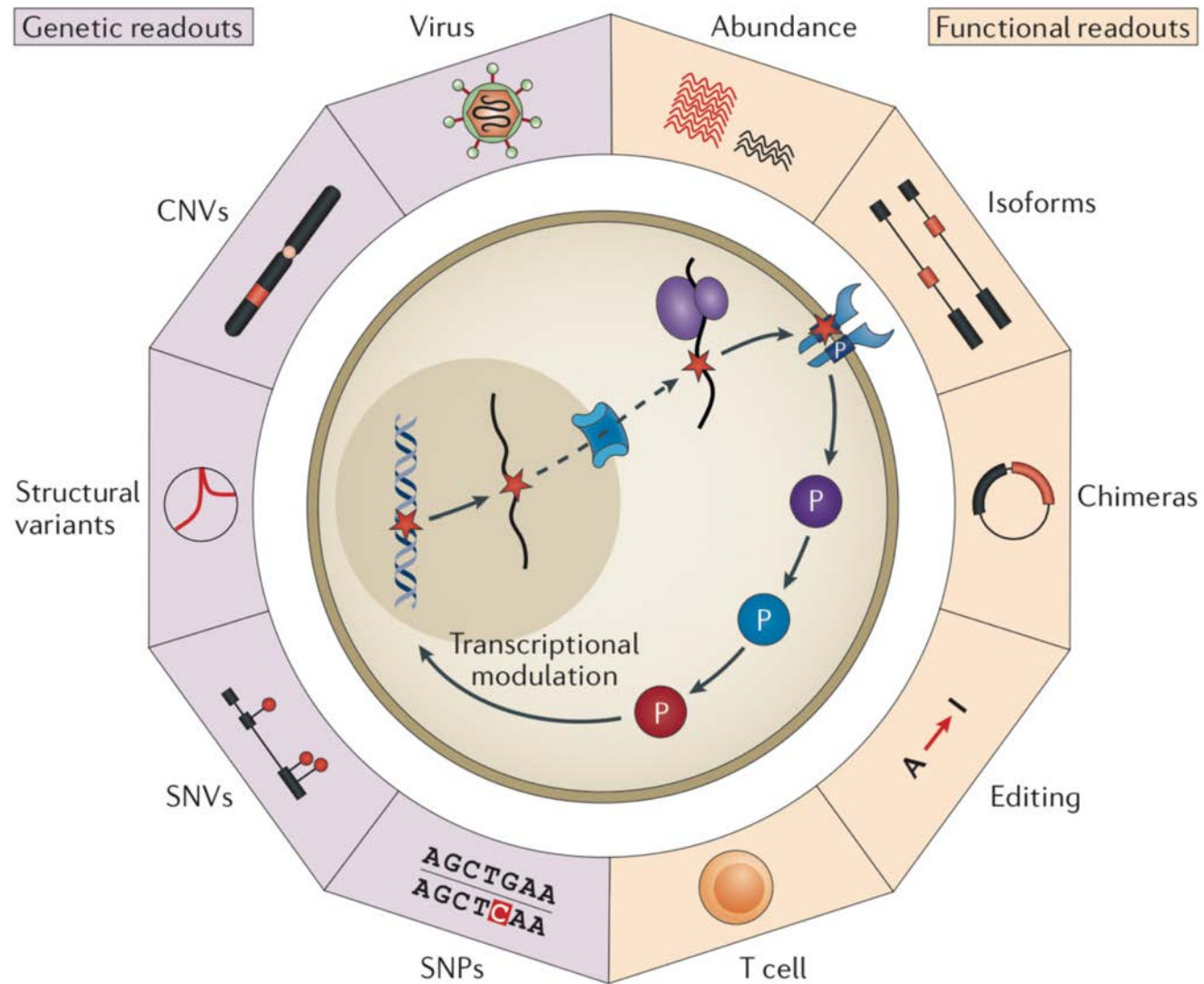
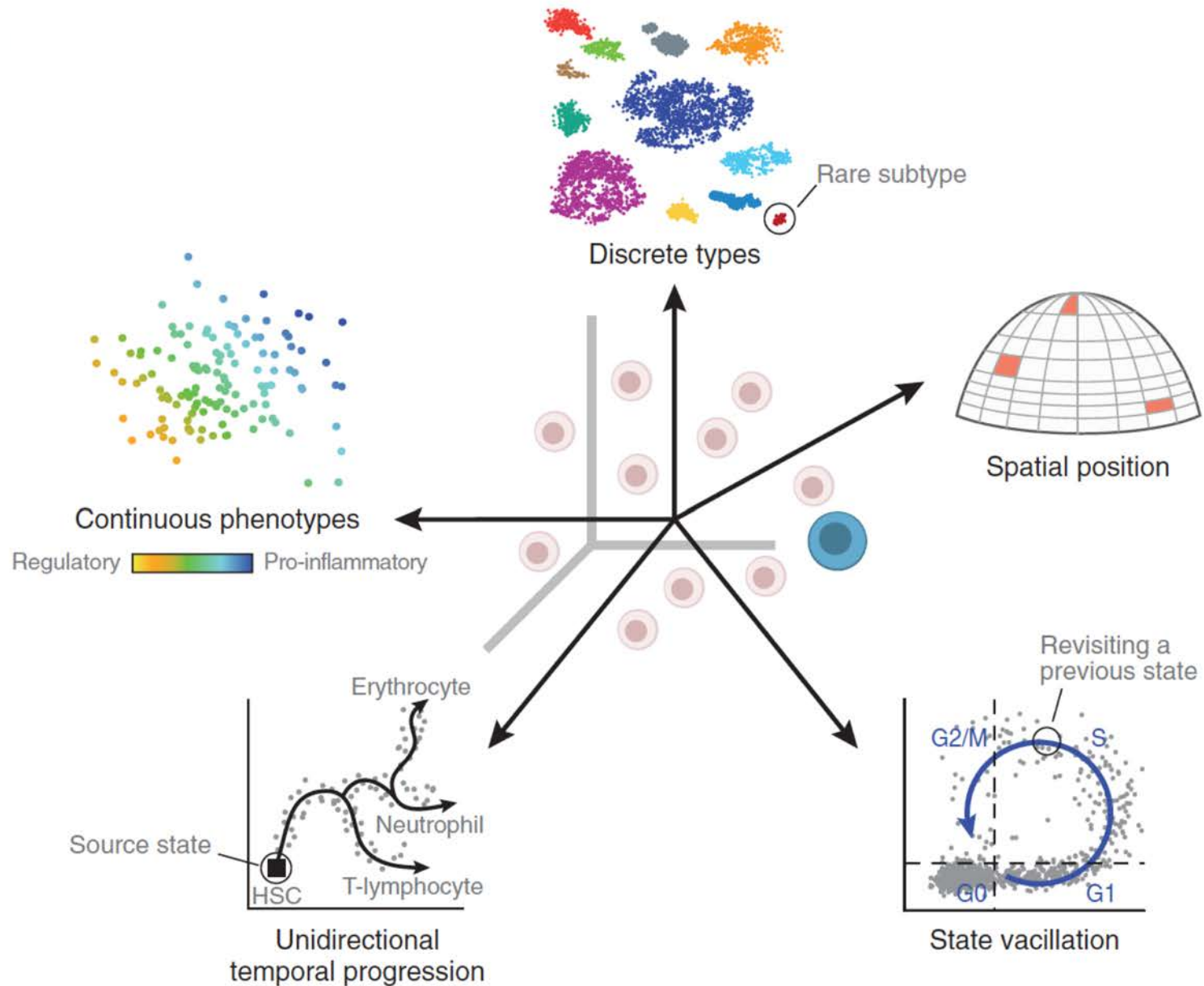
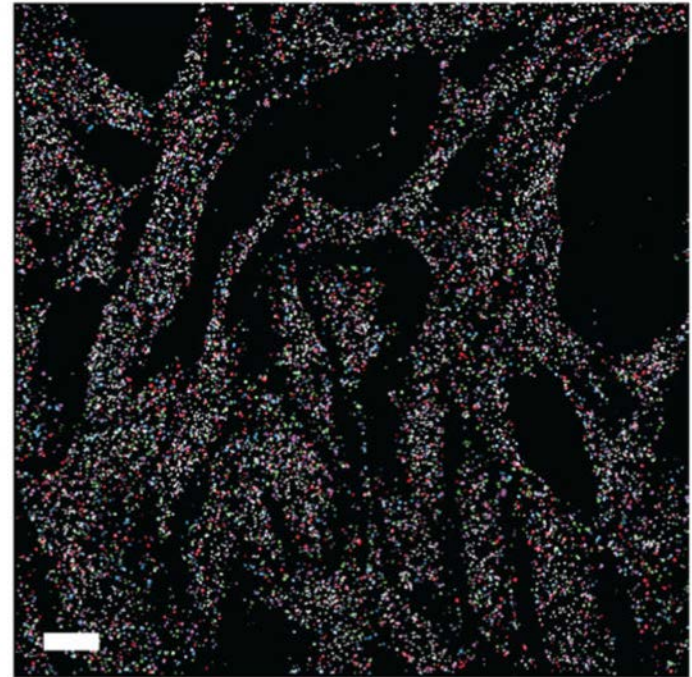
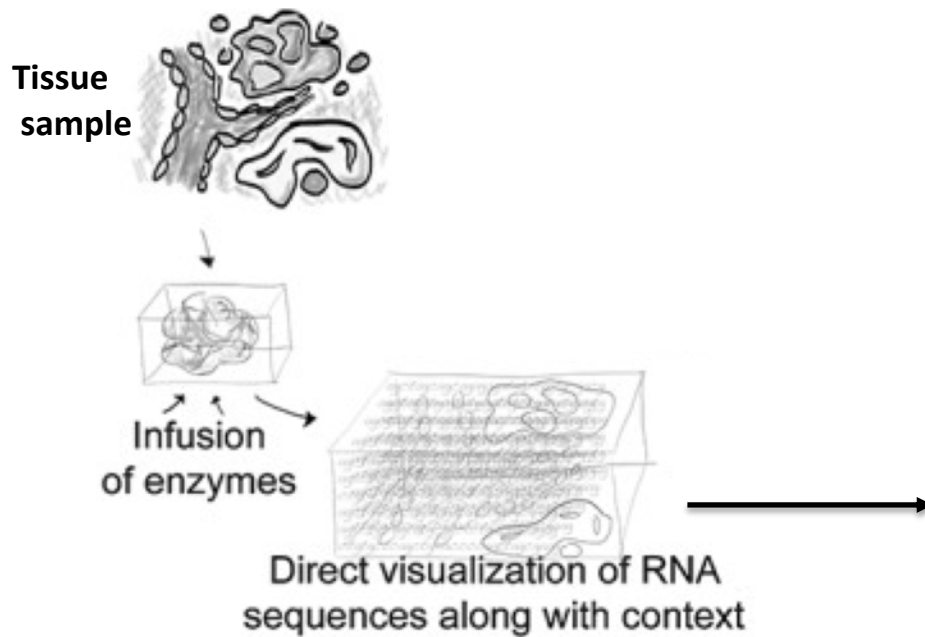


Figure 2 | **Transcriptome profiling for genetic causes and functional phenotypic readouts.**

RNA-Seq is Empowering Discovery at Single Cell Resolution



Fluorescent in situ RNA sequencing (FISSEQ)



Fibroblasts, FISSEQ gene pixels

Adapted from:

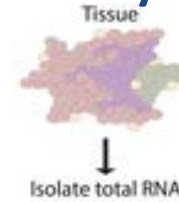
JH Lee, 2017, PMC5315614

JH Lee, 2014, PMC4140943

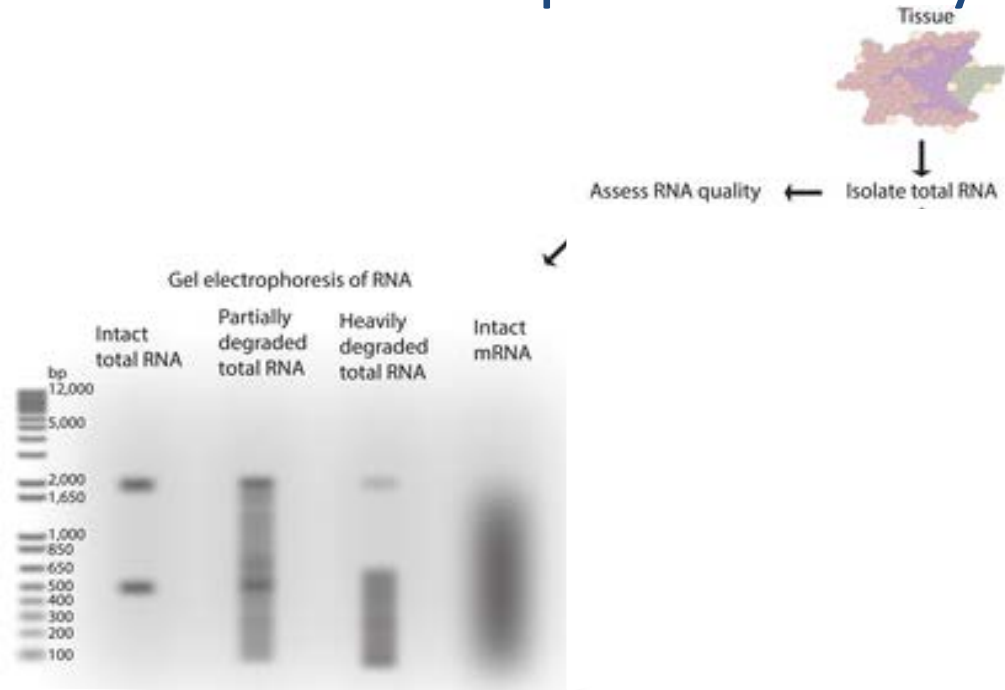
Transcriptomics Lecture Overview

- Overview of RNA-Seq
- Transcript reconstruction methods
- Trinity de novo assembly
- Transcriptome quality assessment
(coffee break)
- Expression quantitation
- Differential expression analysis
- Functional annotation
- Case study: salamander transcriptome

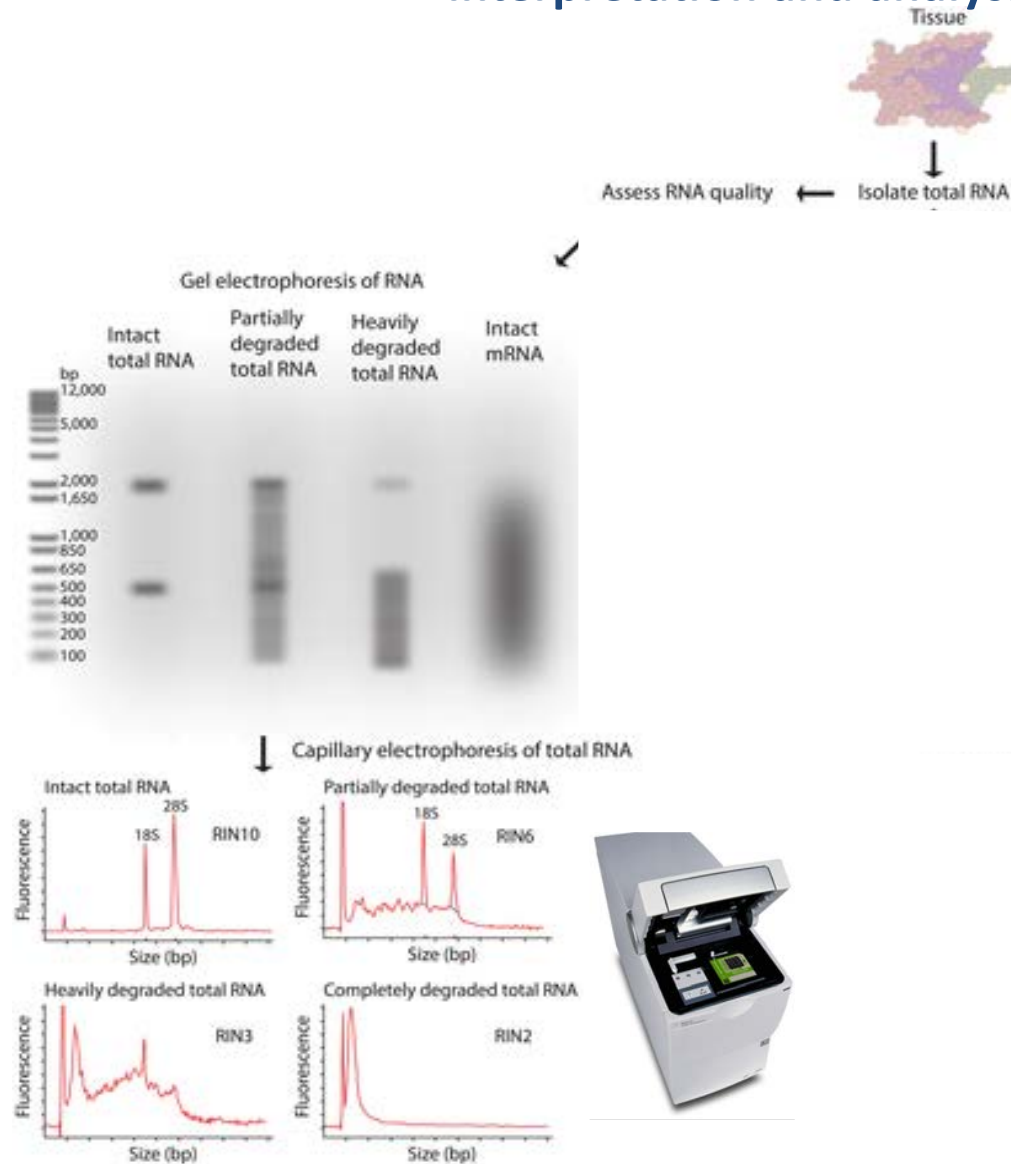
RNA-seq library fragmentation and size selection strategies that influence interpretation and analysis.



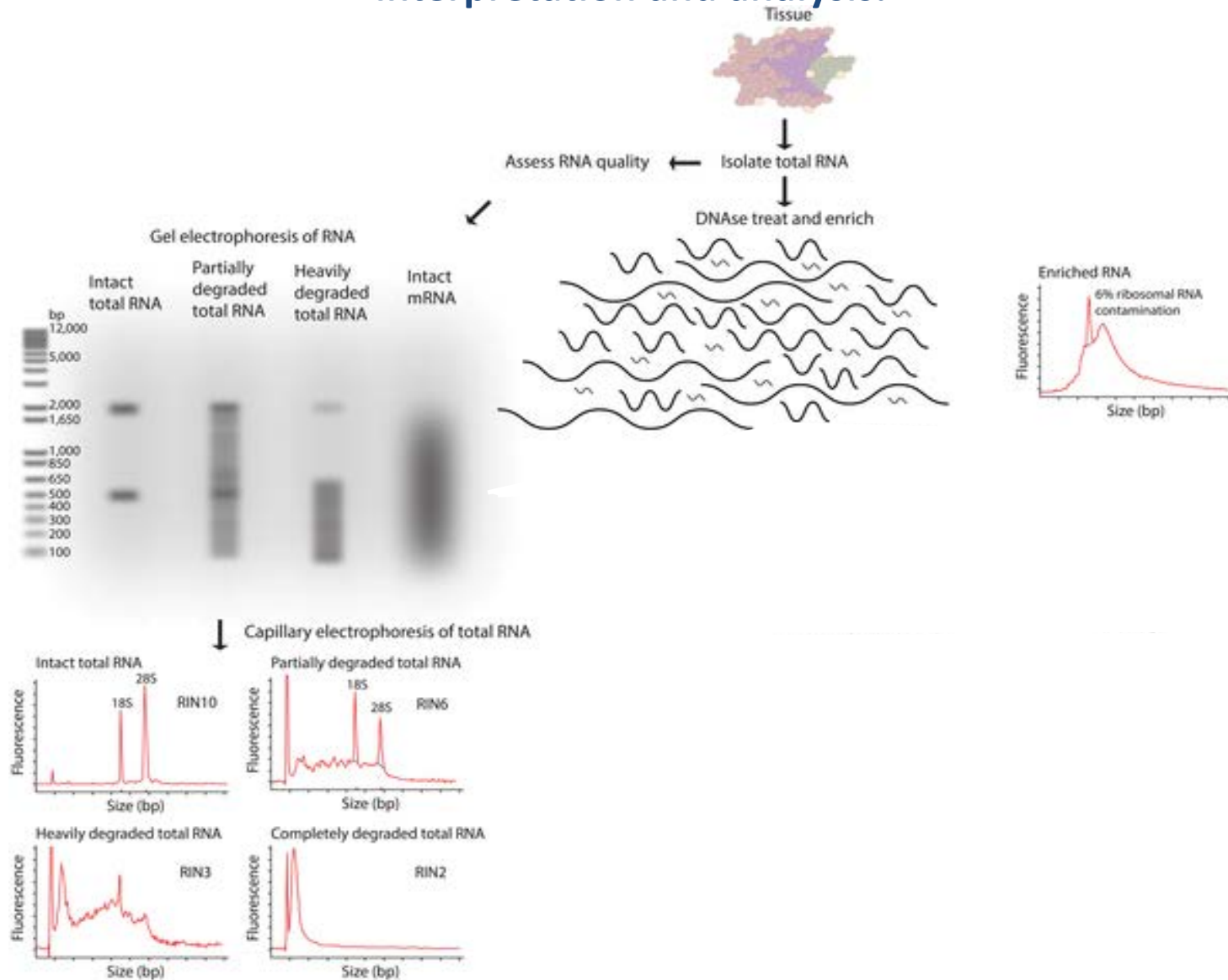
RNA-seq library fragmentation and size selection strategies that influence interpretation and analysis.



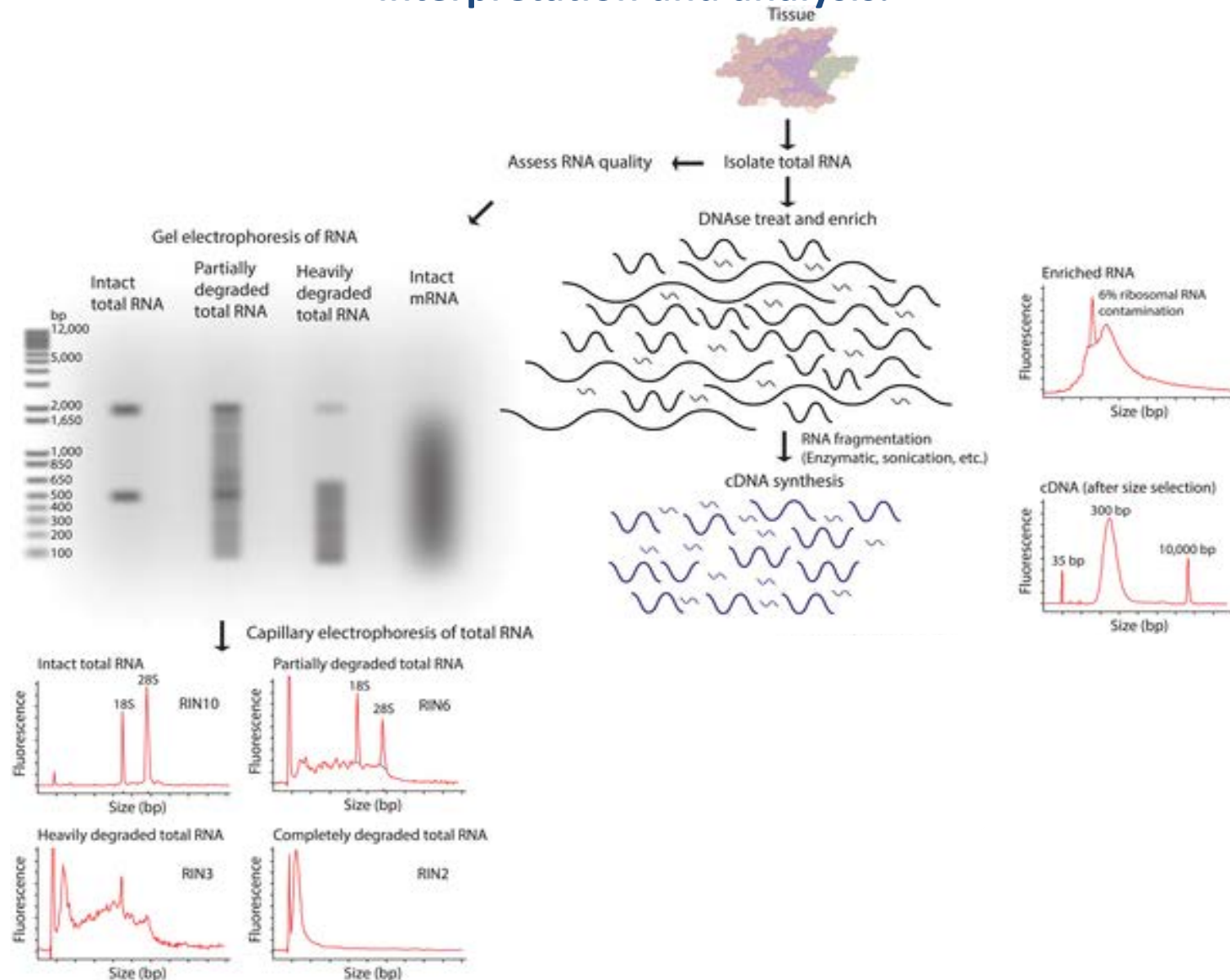
RNA-seq library fragmentation and size selection strategies that influence interpretation and analysis.



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

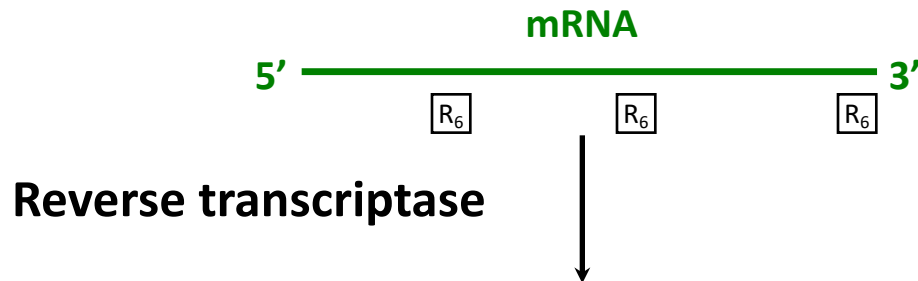


RNA-seq library fragmentation and size selection strategies that influence interpretation and analysis.



RNA-Seq: How do we make cDNA?

Prime with Random Hexamers (R6)



Reverse transcriptase

cDNA First strand synthesis



RNase H

DNA polymerase

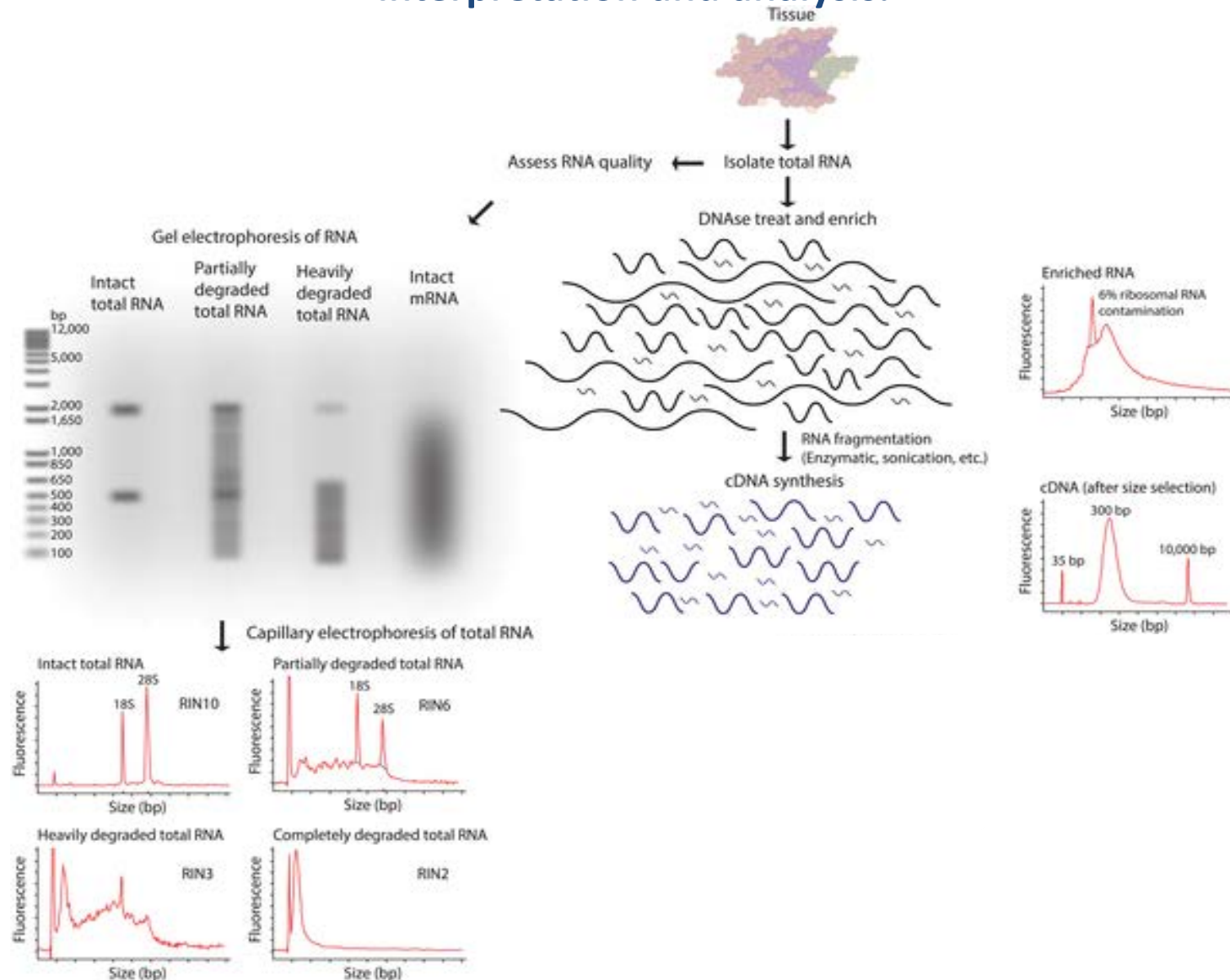
DNA Ligase

cDNA Second strand synthesis

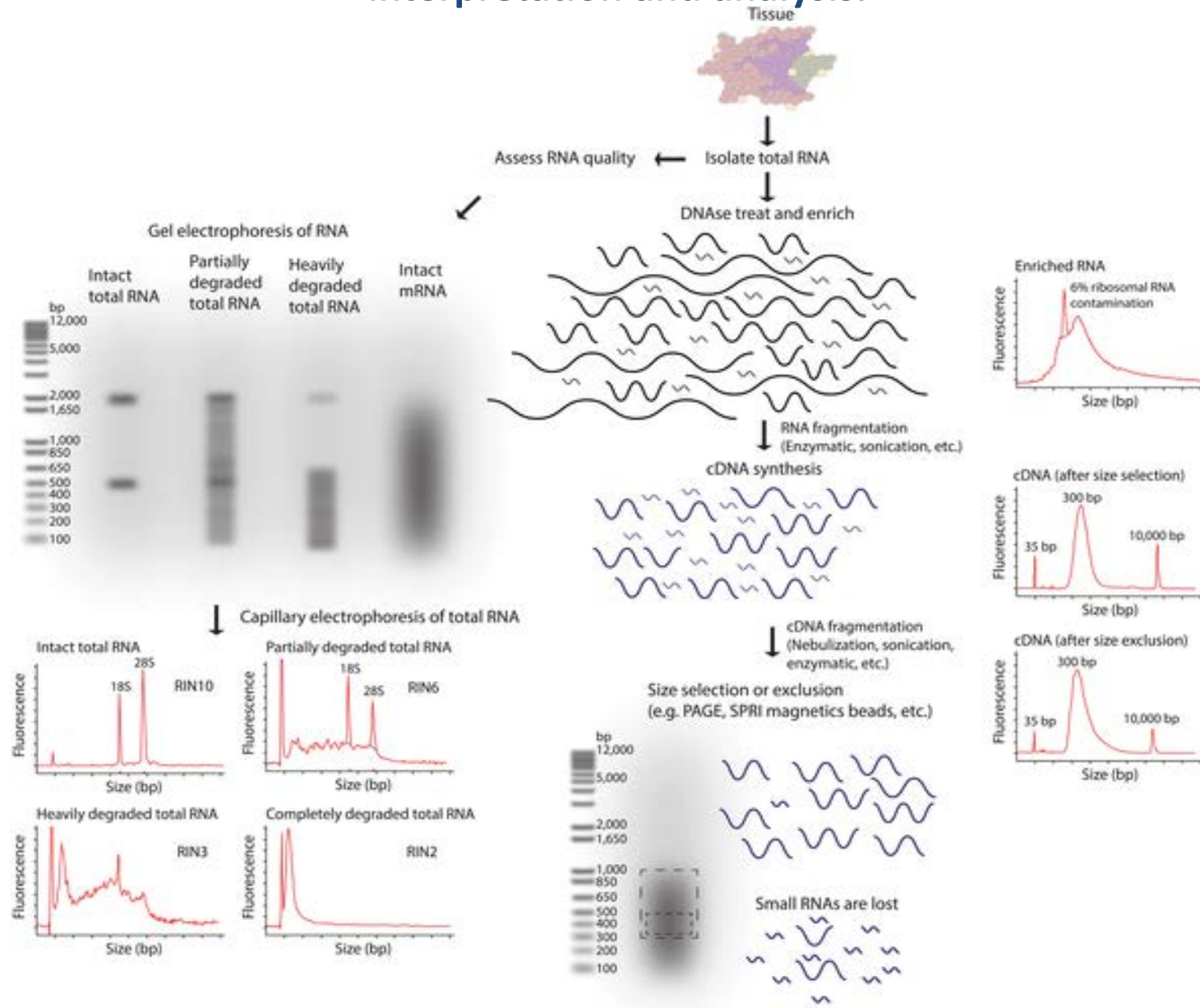


Illumina cDNA Library

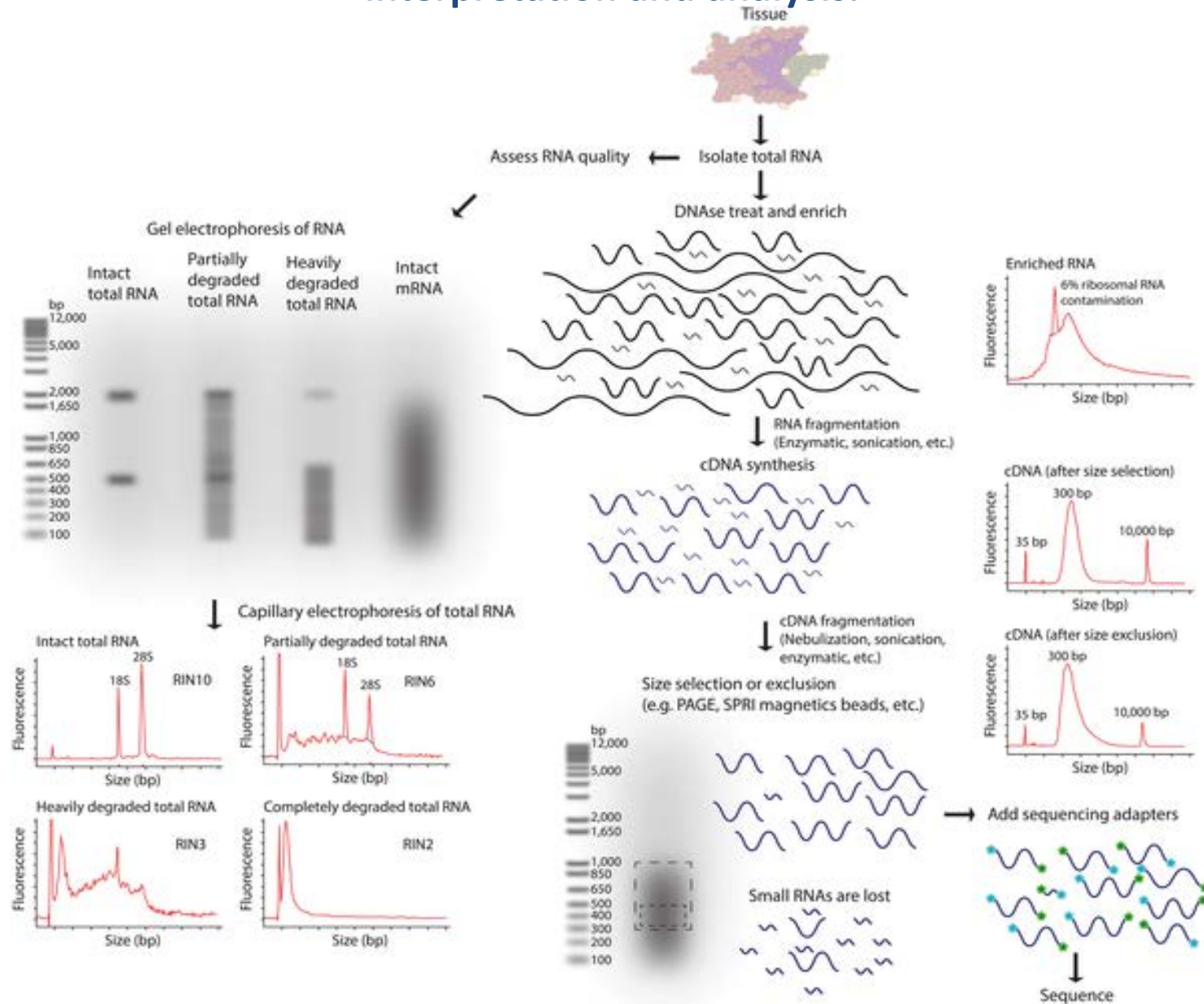
RNA-seq library fragmentation and size selection strategies that influence interpretation and analysis.



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

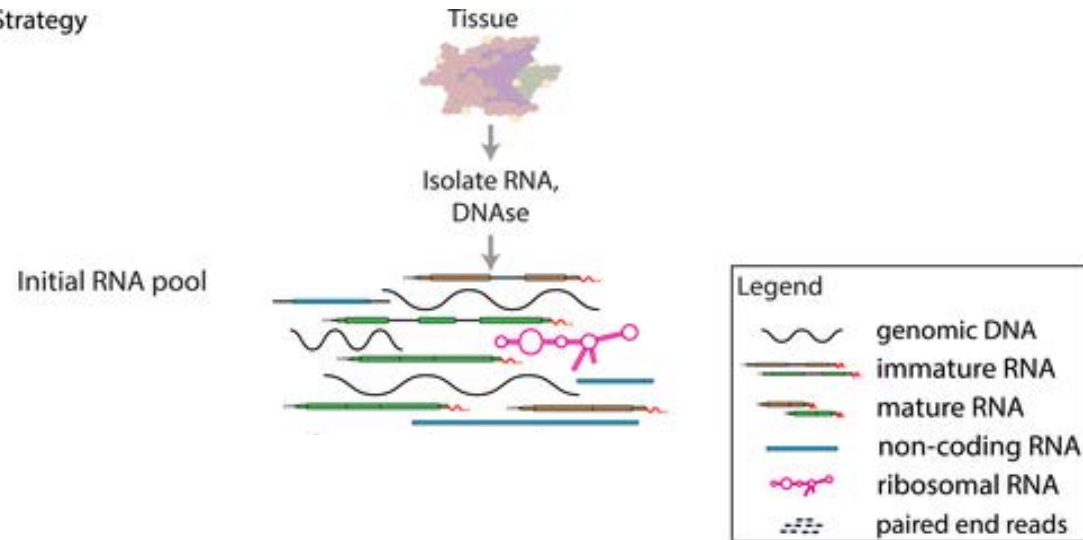


RNA-seq library fragmentation and size selection strategies that influence interpretation and analysis.



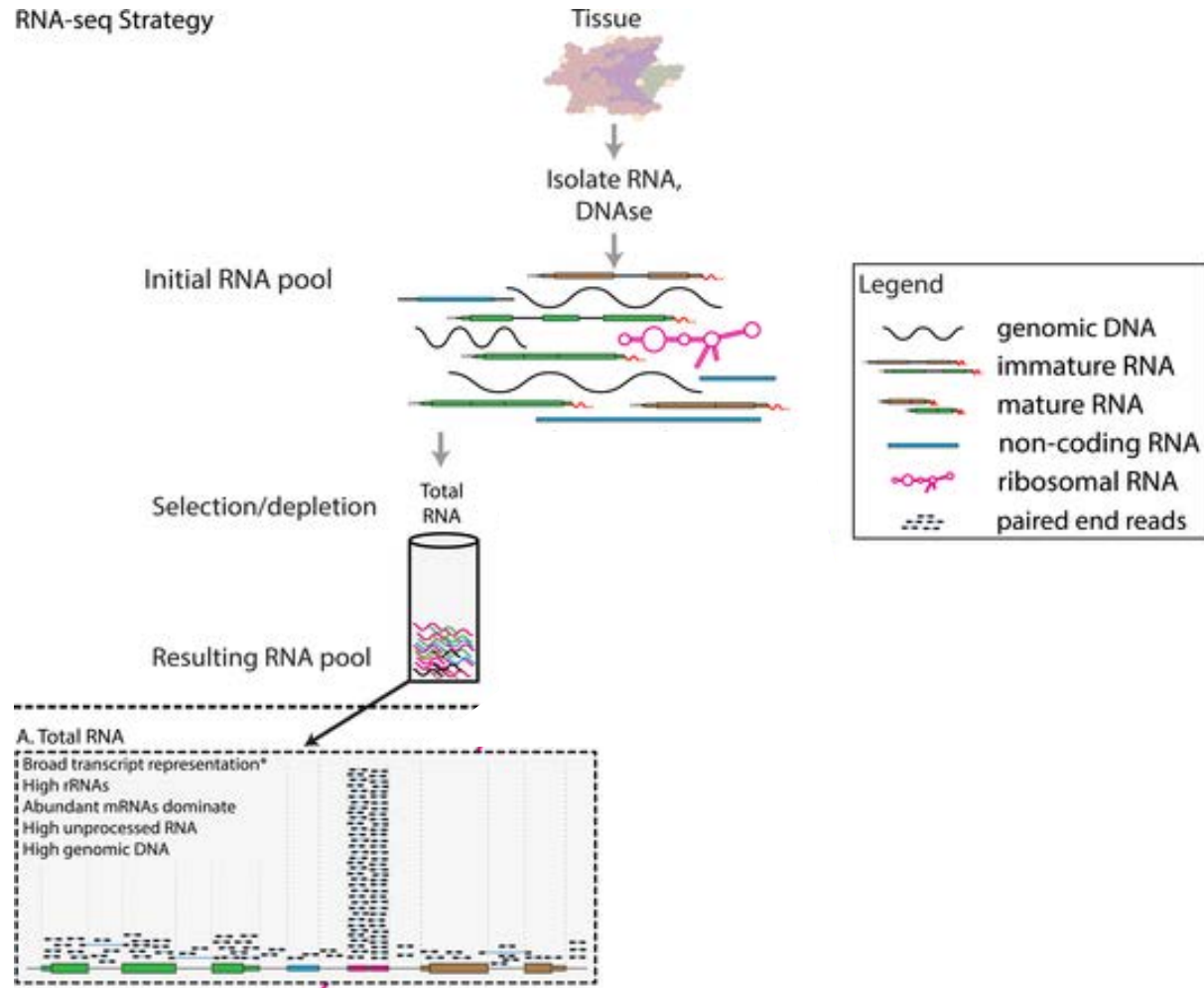
RNA-seq library enrichment strategies that influence interpretation and analysis.

RNA-seq Strategy



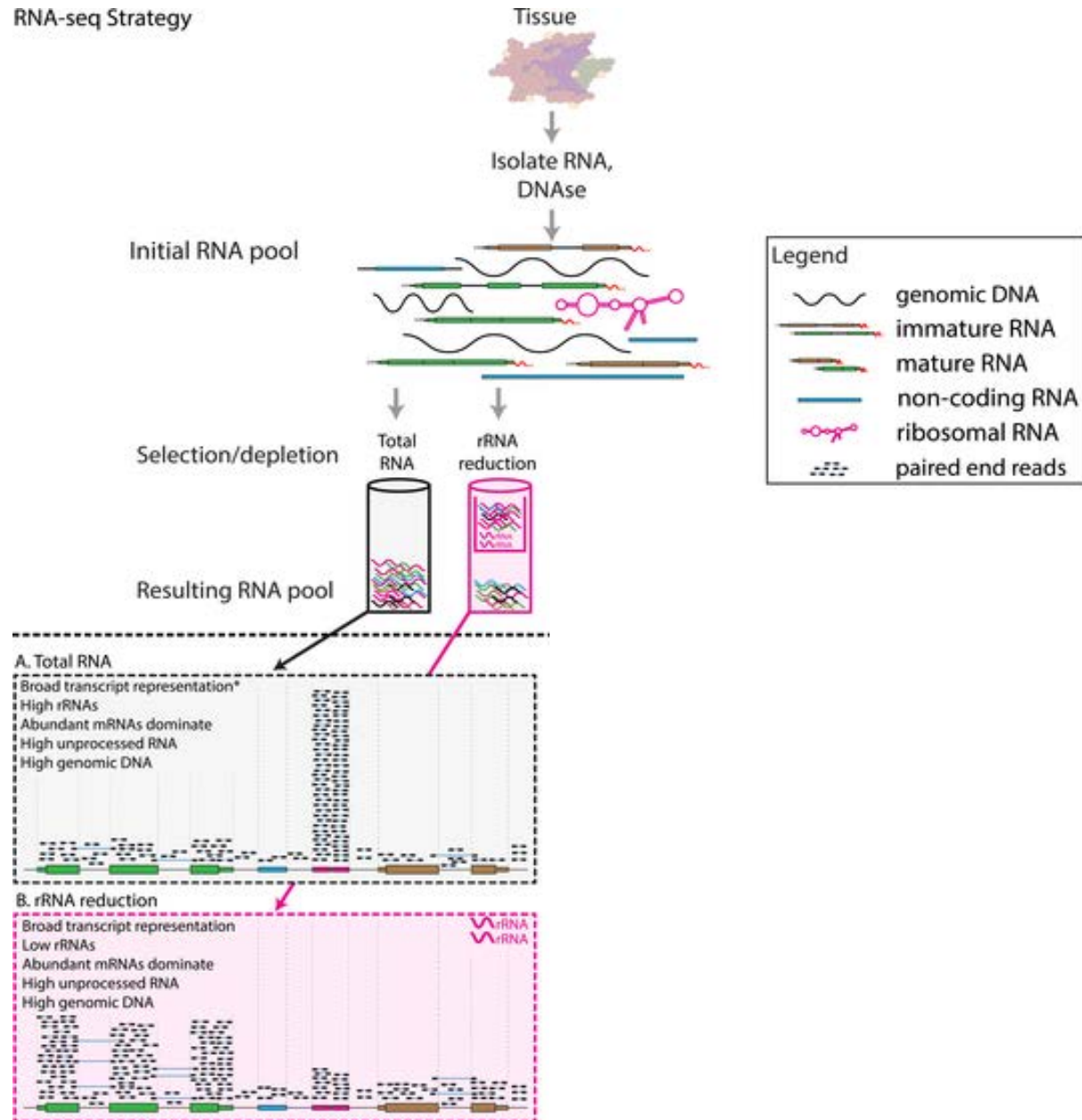
RNA-seq library enrichment strategies that influence interpretation and analysis.

RNA-seq Strategy



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

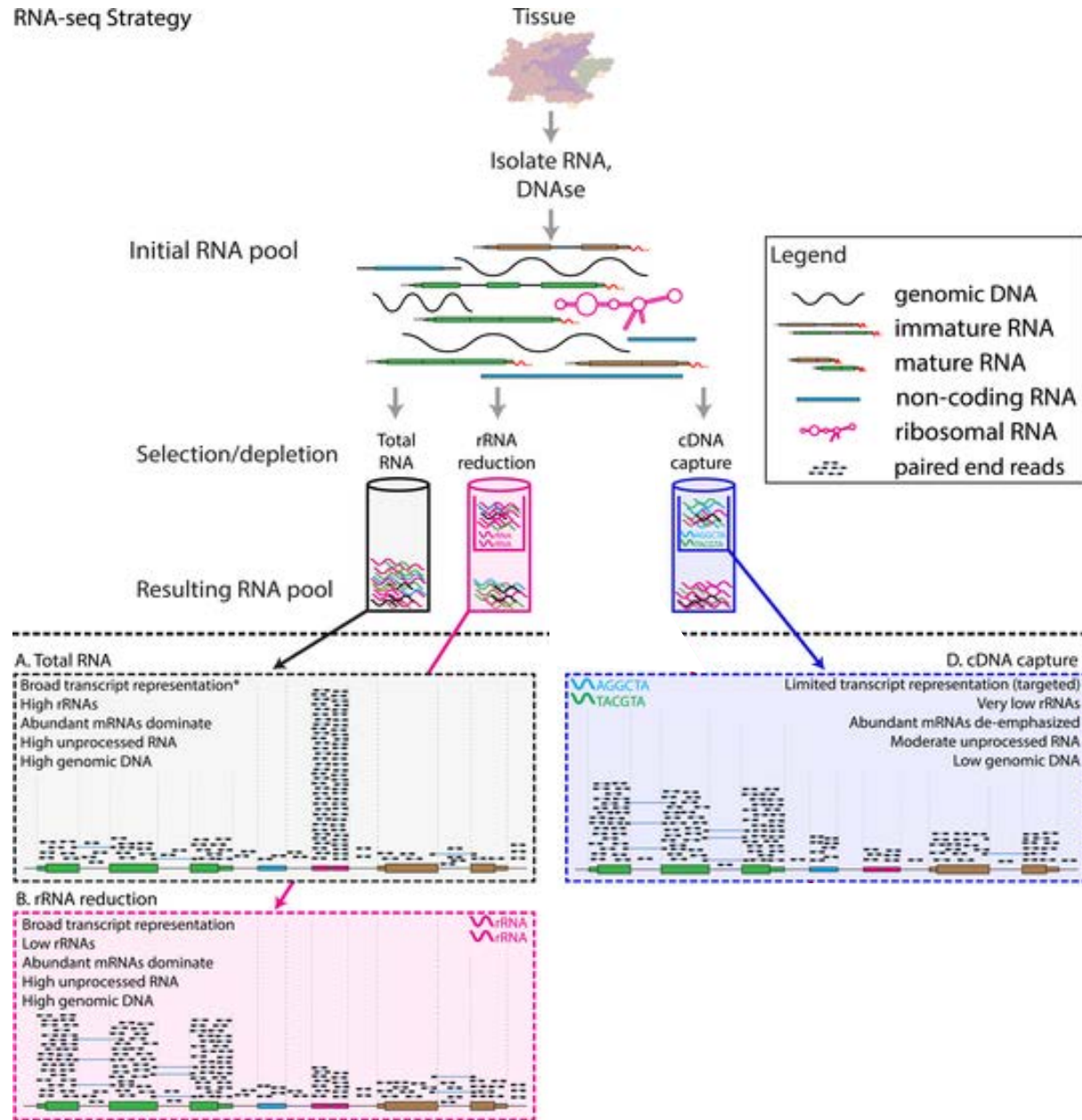
RNA-seq Strategy



Expected Alignments

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

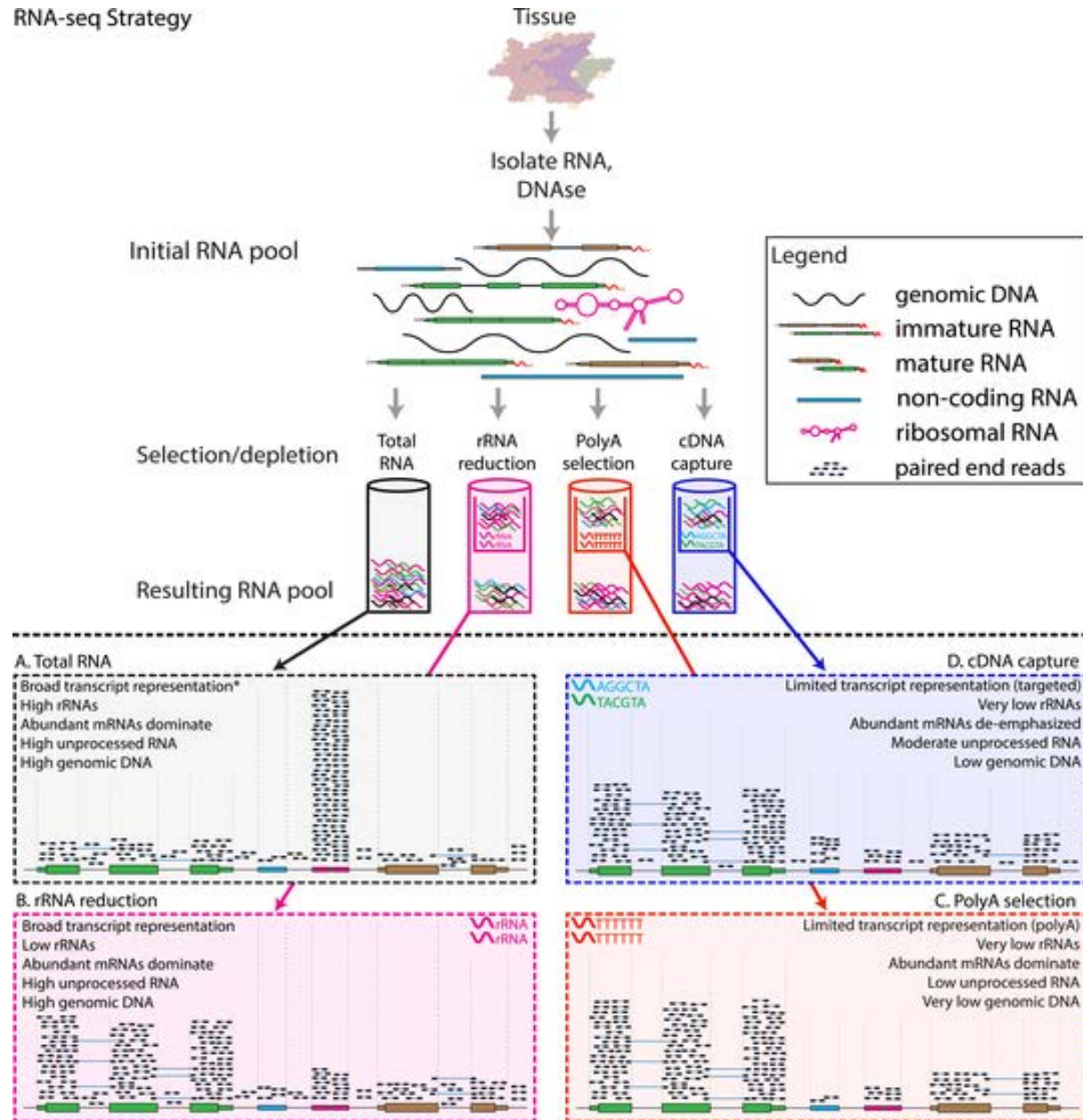
RNA-seq Strategy



Expected Alignments

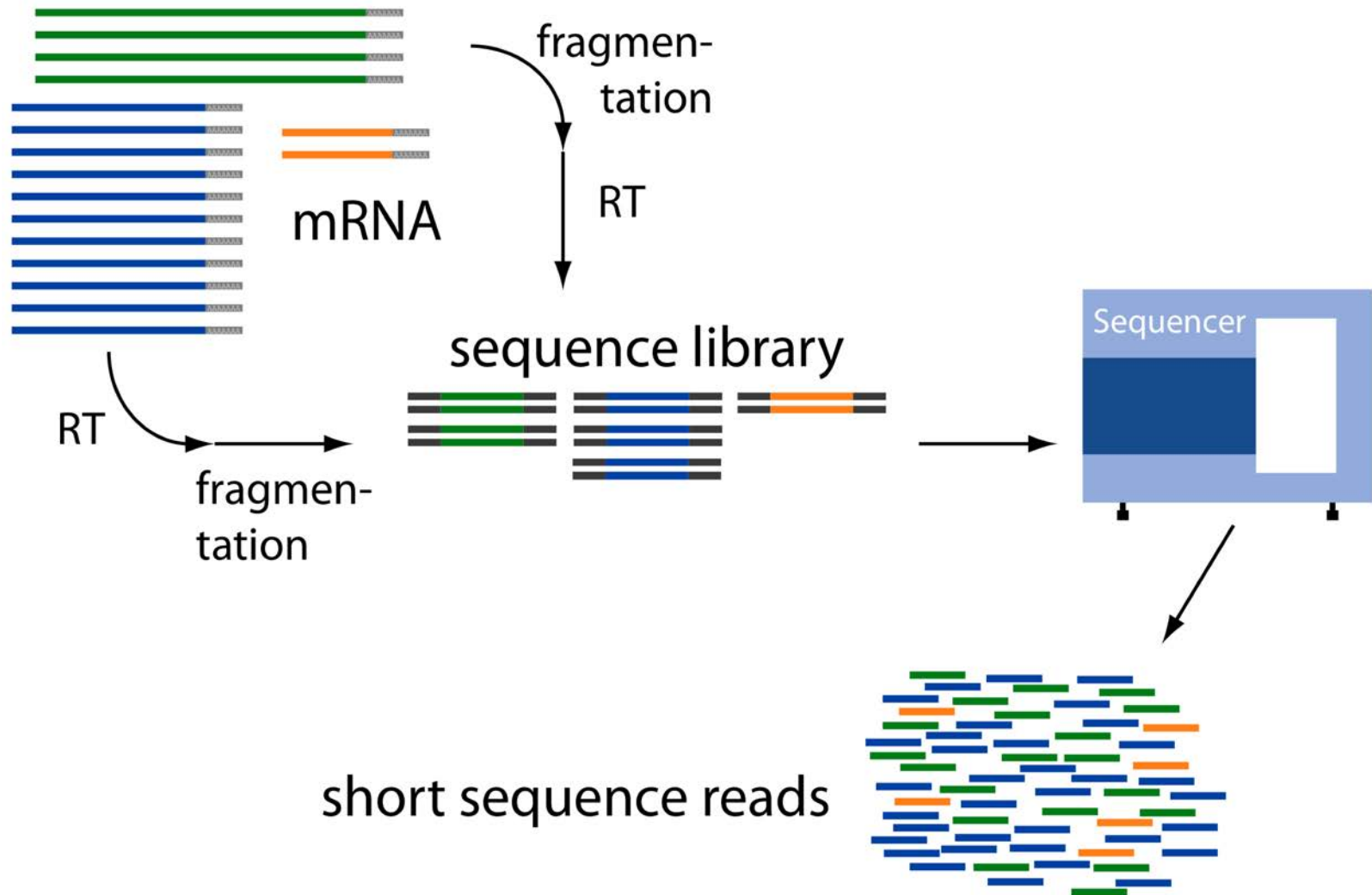
RNA-seq library enrichment strategies that influence interpretation and analysis.

RNA-seq Strategy



Expected Alignments

Overview of RNA-Seq



Common Data Formats for RNA-Seq

FASTA format:

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

FASTQ format:

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

Read

Quality values

$\text{AsciiEncodedQual}(x) = -10 * \log_{10}(\text{Pwrong}(x)) + 33$



$\text{AsciiEncodedQual}('C') = 64$

So, $\text{Pwrong}('C') = 10^{(64-33/(-10))} = 10^{-3.4} = 0.0004$

Paired-end Sequences

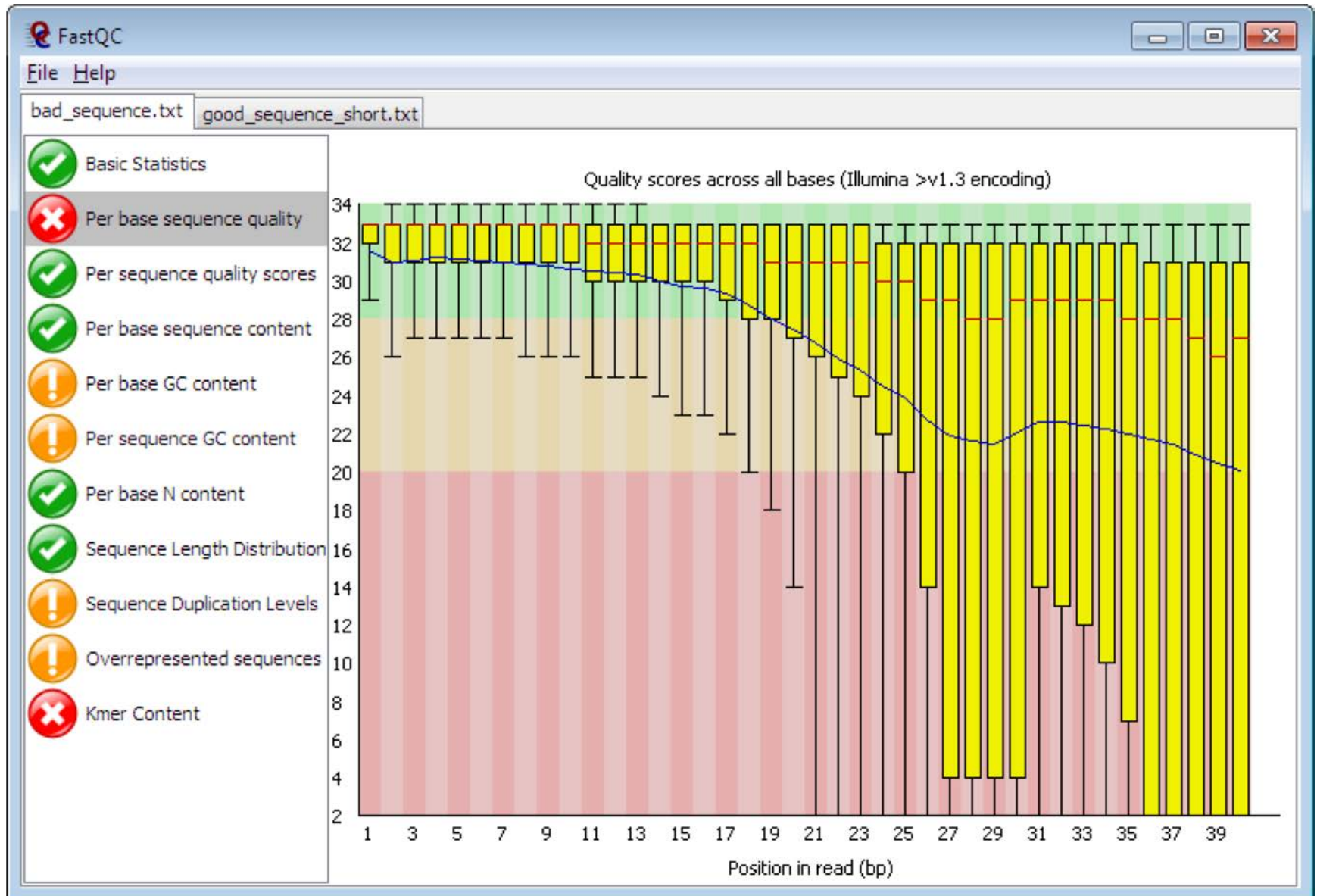


Two FastQ files, read name indicates
left (/1) or right (/2) read of paired-end

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

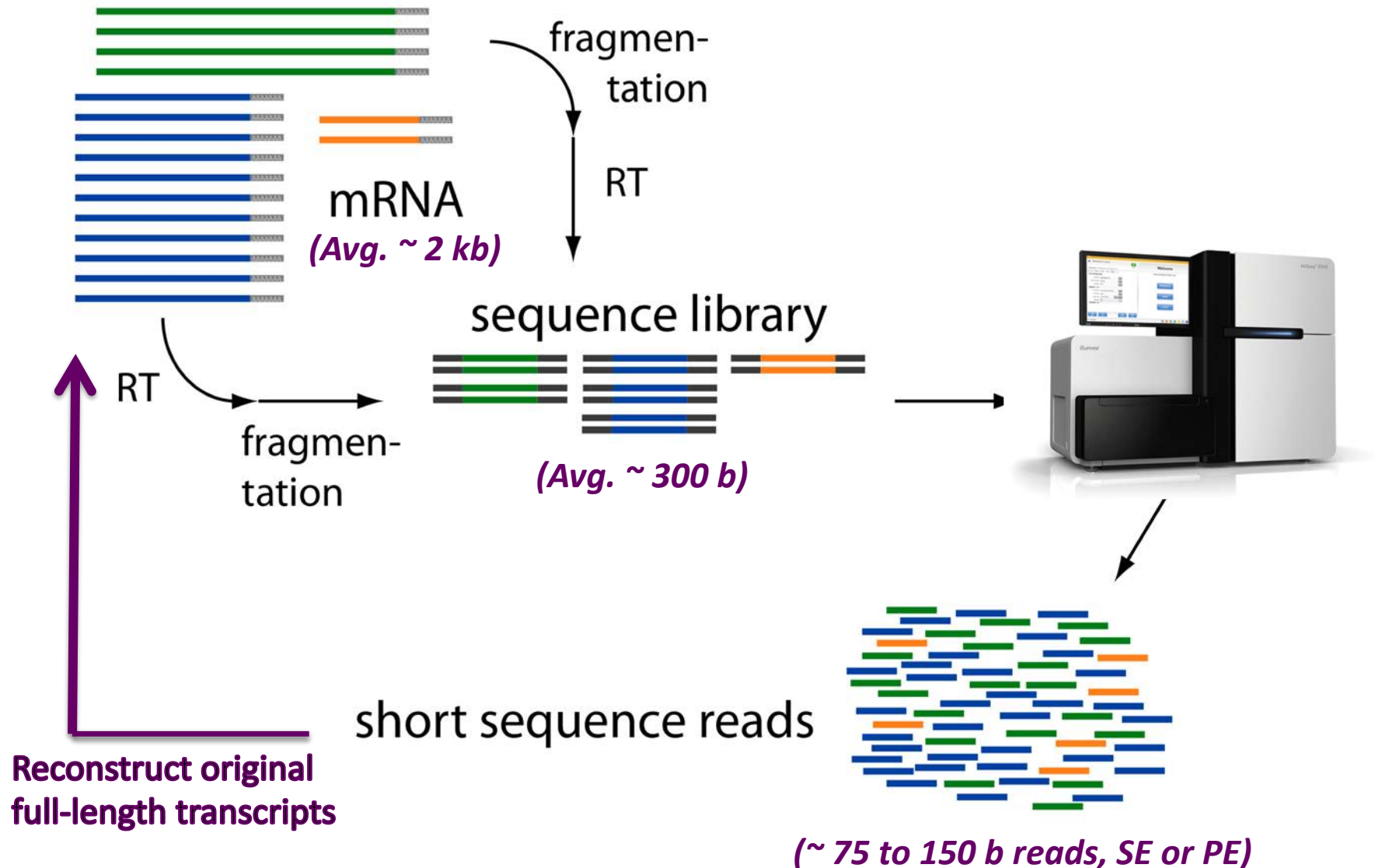
```
@61DFRAAXX100204:1:100:10494:3070/2
CTCAAATGGTTAATTCTCAGGCTGCAAATATTCGTTTCAGGATGGAAGAACA
+
C<CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCBCCCC
```

Read Quality Assessment



From: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

RNA-Seq Challenge: Transcript Reconstruction



Transcript Reconstruction from RNA-Seq Reads



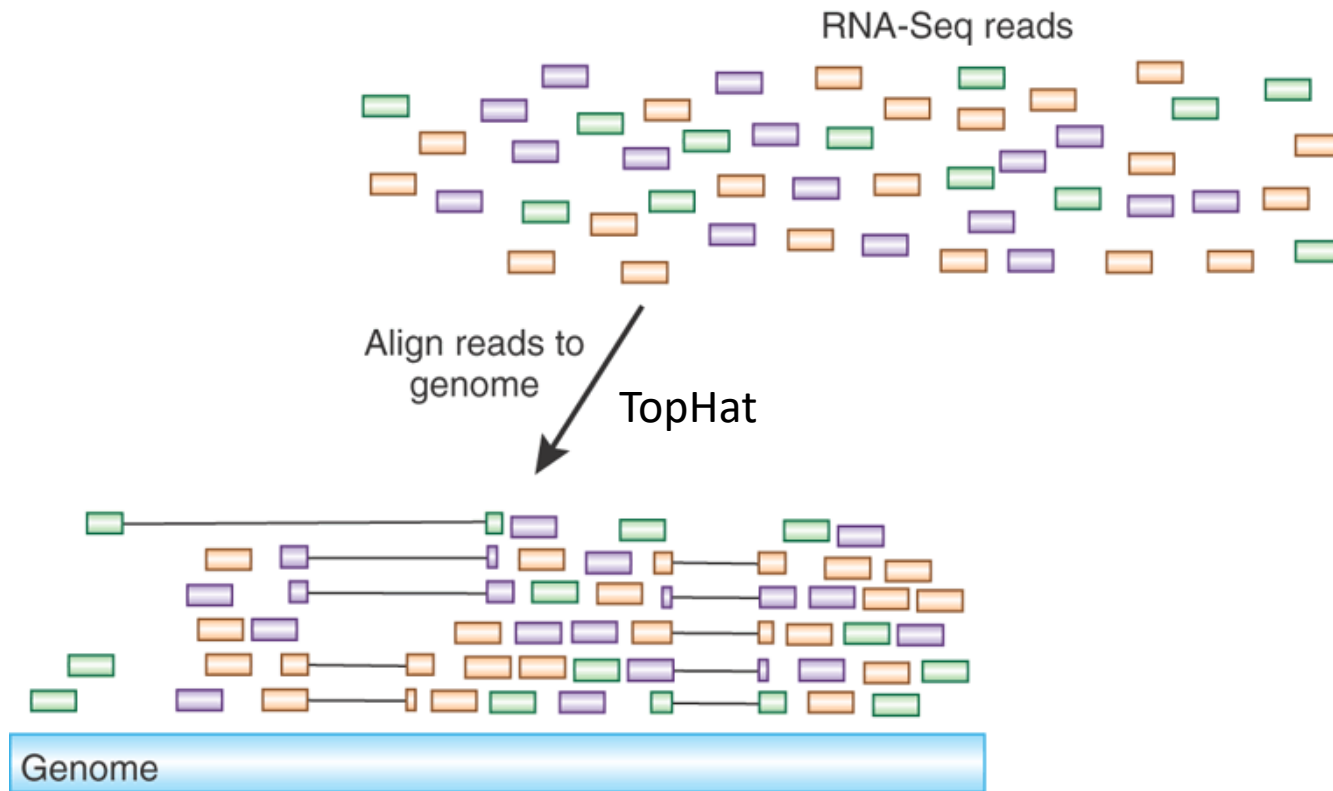
Advancing RNA-Seq analysis

Brian J Haas & Michael C Zody

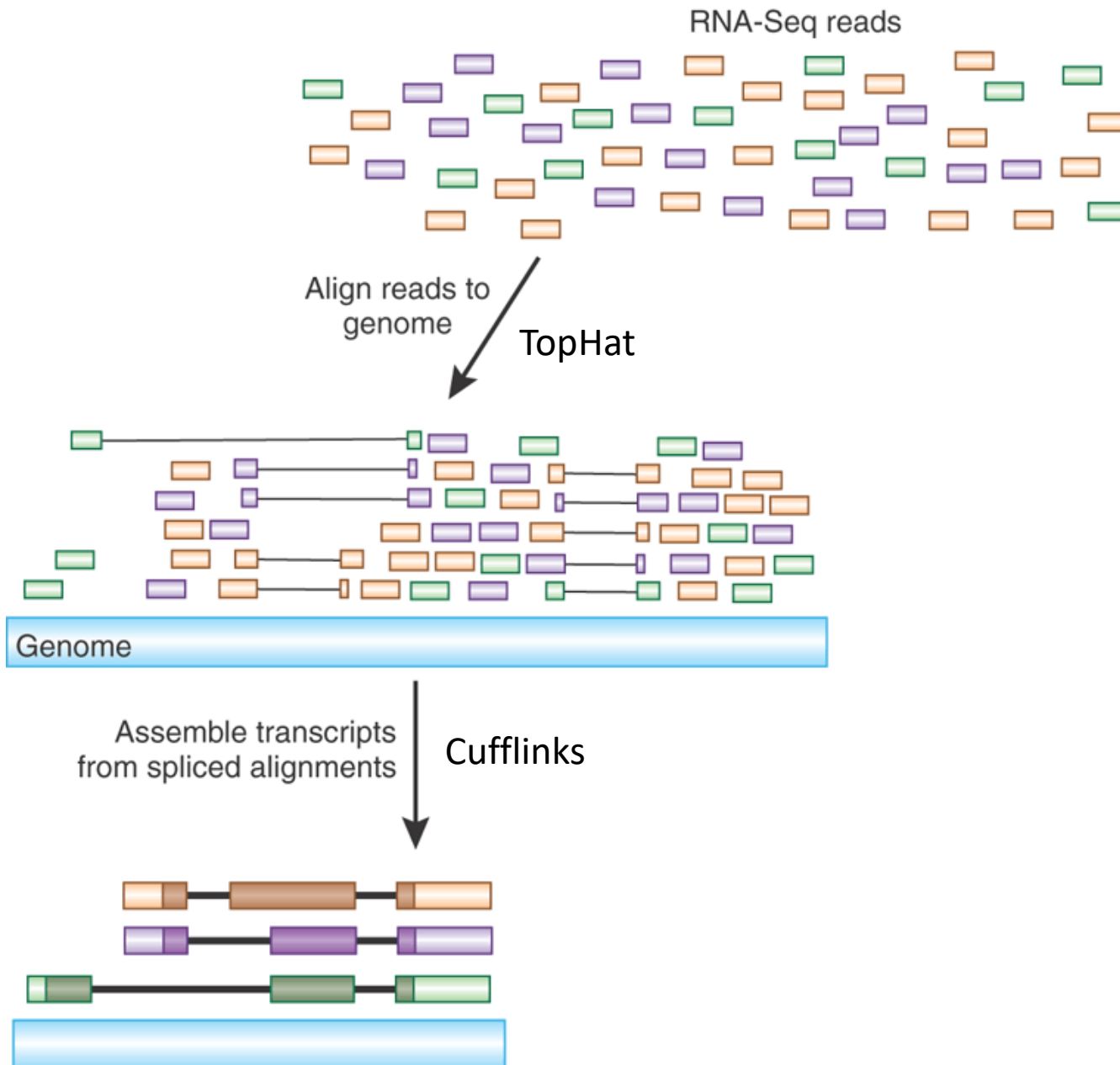
Nature Biotech, 2010

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

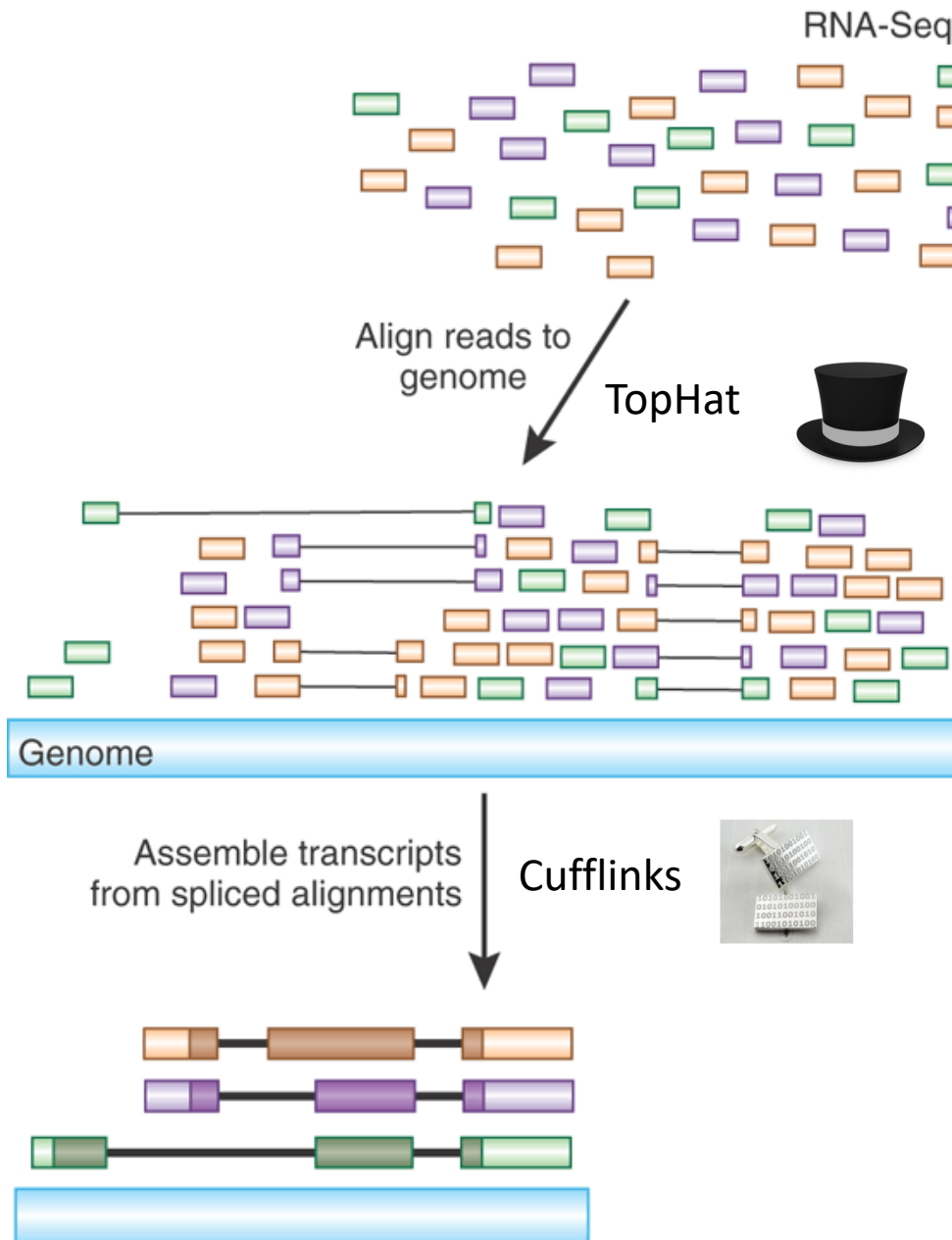
Transcript Reconstruction from RNA-Seq Reads



Transcript Reconstruction from RNA-Seq Reads



Transcript Reconstruction from RNA-Seq Reads



The Tuxedo Suite:

End-to-end **Genome**-based
RNA-Seq Analysis
Software Package

NATURE PROTOCOLS | **PROTOCOL**

Differential gene and transcript
expression analysis of RNA-seq
experiments with TopHat and
Cufflinks

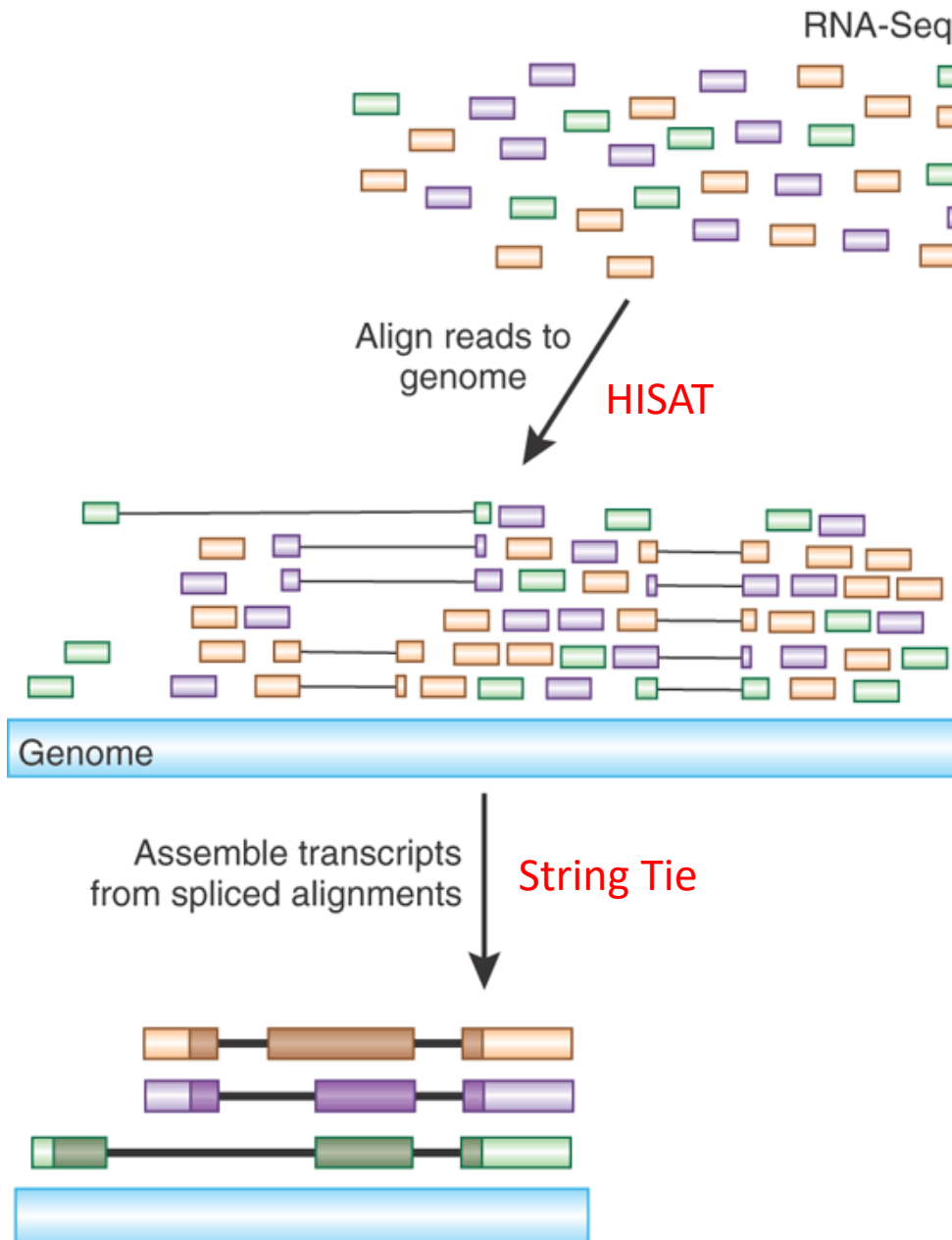
Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn & Lior Pachter

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

Nature Protocols **7**, 562–578 (2012) | doi:10.1038/nprot.2012.016

Published online 01 March 2012

Transcript Reconstruction from RNA-Seq Reads



The “New Tuxedo” Suite:

End-to-end **Genome**-based
RNA-Seq Analysis
Software Package

NATURE PROTOCOLS | PROTOCOL



Transcript-level expression analysis of
RNA-seq experiments with HISAT,
StringTie and Ballgown

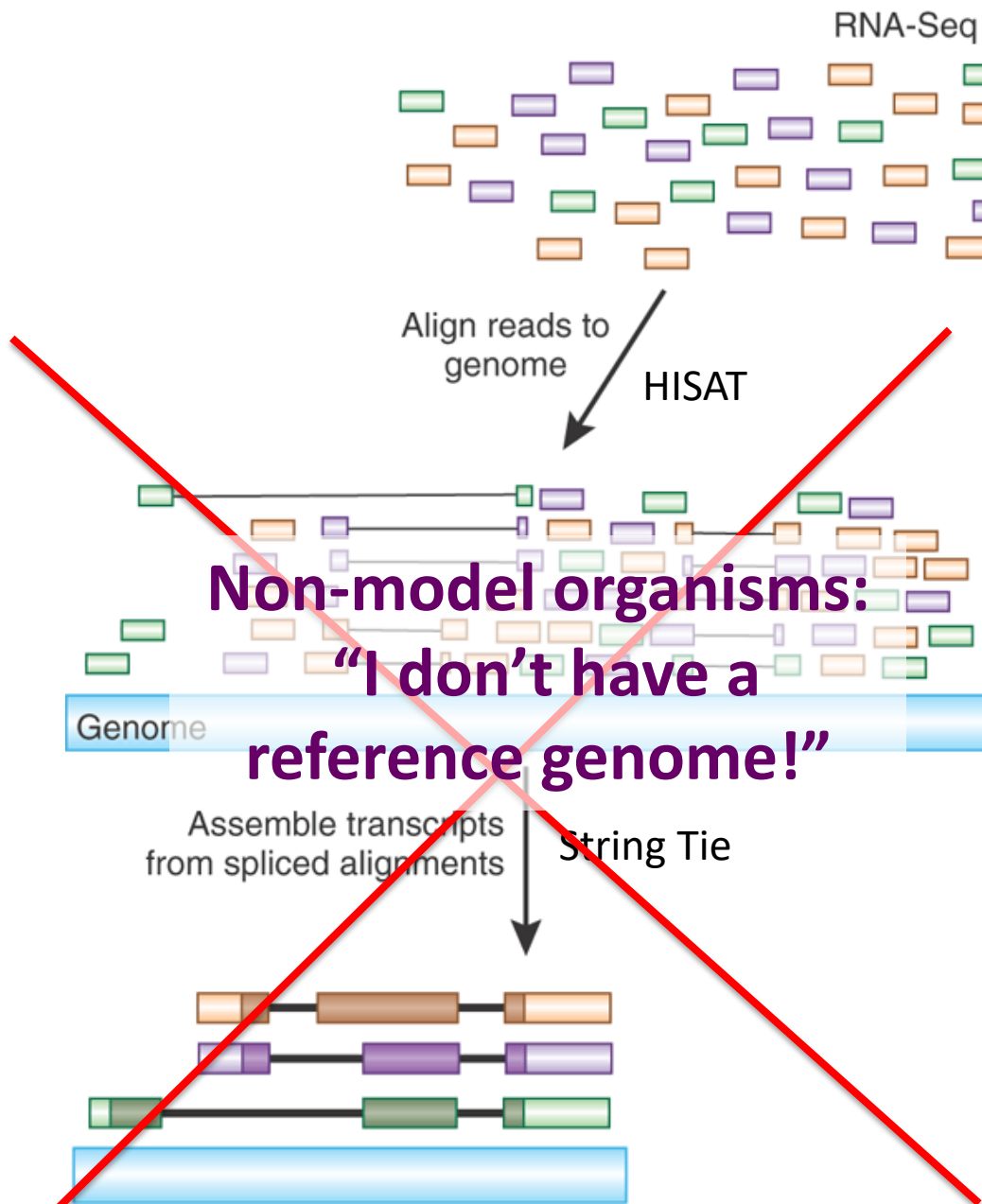
Mihaela Pertea, Daehwan Kim, Geo M Pertea, Jeffrey T Leek & Steven
L Salzberg

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

Nature Protocols **11**, 1650–1667 (2016) | doi:10.1038/nprot.2016.095

Published online 11 August 2016

Transcript Reconstruction from RNA-Seq Reads



The “New Tuxedo” Suite:

End-to-end **Genome**-based
RNA-Seq Analysis
Software Package

NATURE PROTOCOLS | PROTOCOL



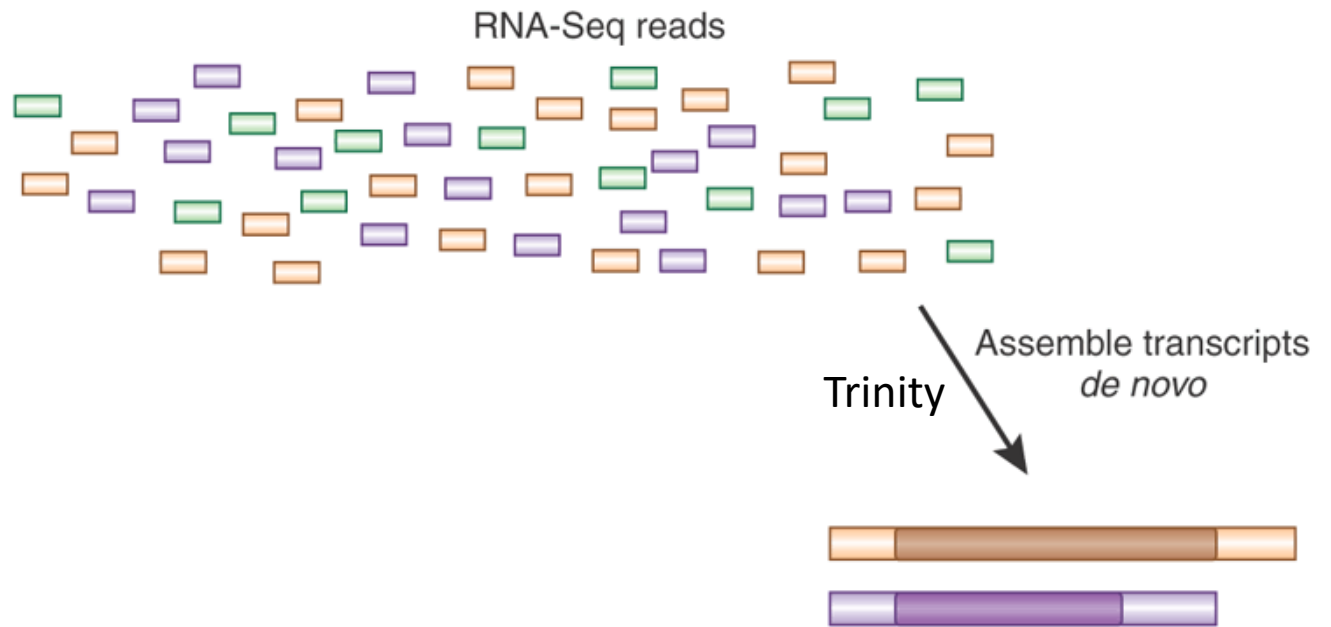
Transcript-level expression analysis of
RNA-seq experiments with HISAT,
StringTie and Ballgown

Mihaela Pertea, Daehwan Kim, Geo M Pertea, Jeffrey T Leek & Steven L Salzberg

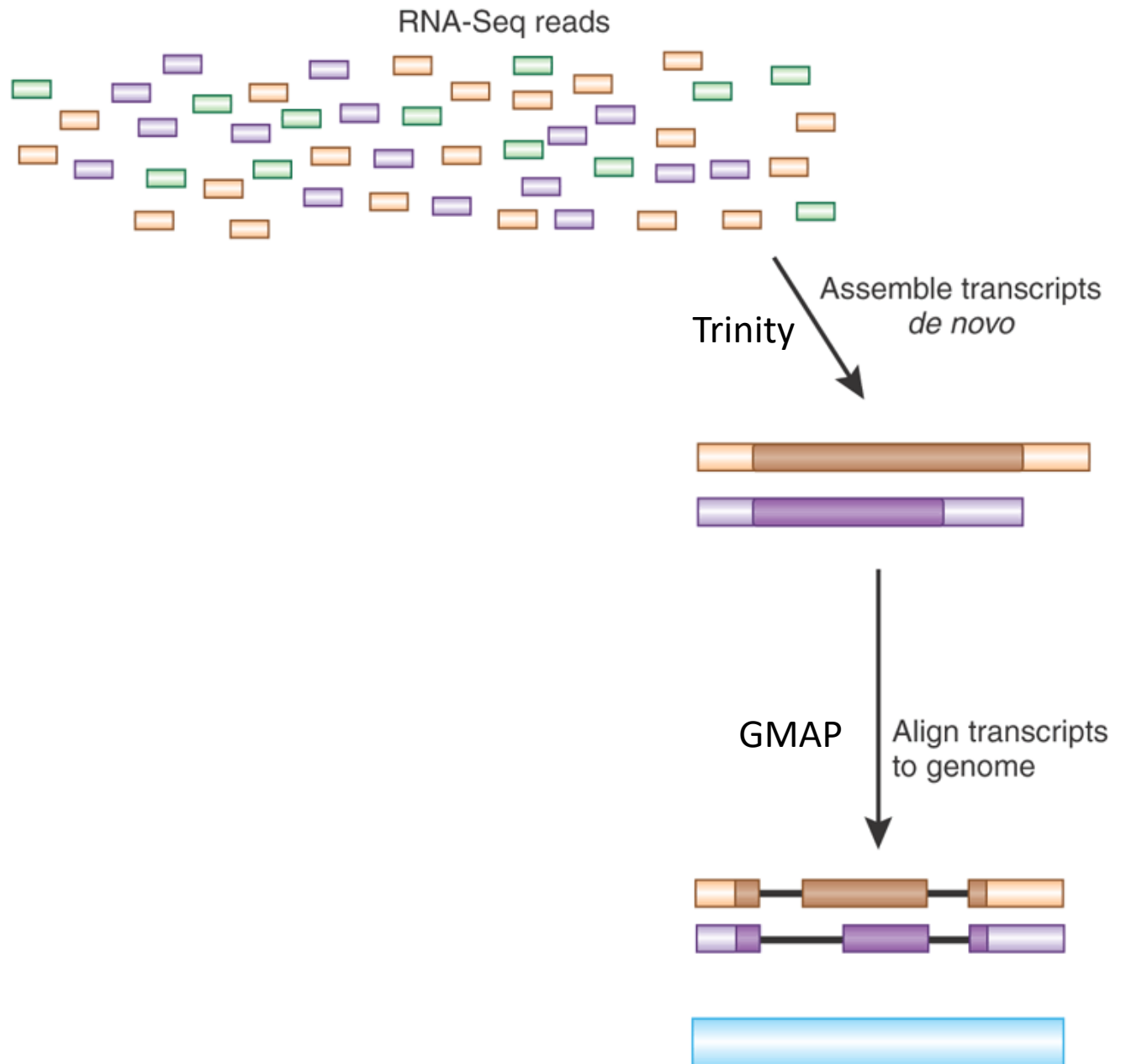
[Affiliations](#) | [Contributions](#) | [Corresponding author](#)

Nature Protocols **11**, 1650–1667 (2016) | doi:10.1038/nprot.2016.095
Published online 11 August 2016

Transcript Reconstruction from RNA-Seq Reads



Transcript Reconstruction from RNA-Seq Reads



Transcript Reconstruction from RNA-Seq Reads



Assemble transcripts
de novo



BROAD
INSTITUTE

Trinity

Align transcripts
to genome



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

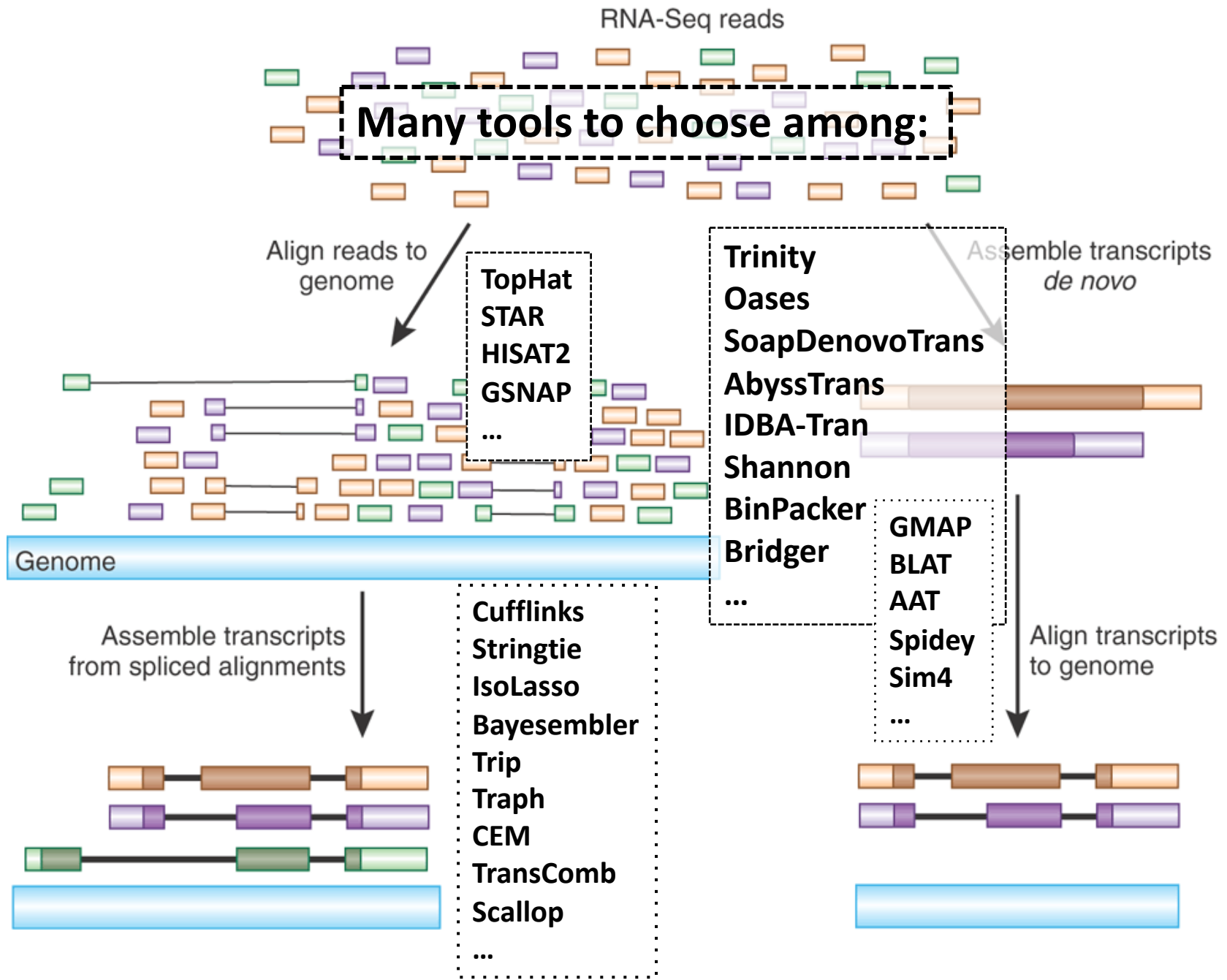
Brian J Haas, Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, Philip D Blood,
Joshua Bowden, Matthew Brian Couger, David Eccles, Bo Li, Matthias Lieber, Matthew D
MacManes, Michael Ott, Joshua Orvis, Nathalie Pochet, Francesco Strozzi, Nathan Weeks,
Rick Westerman, Thomas William, Colin N Dewey,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

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

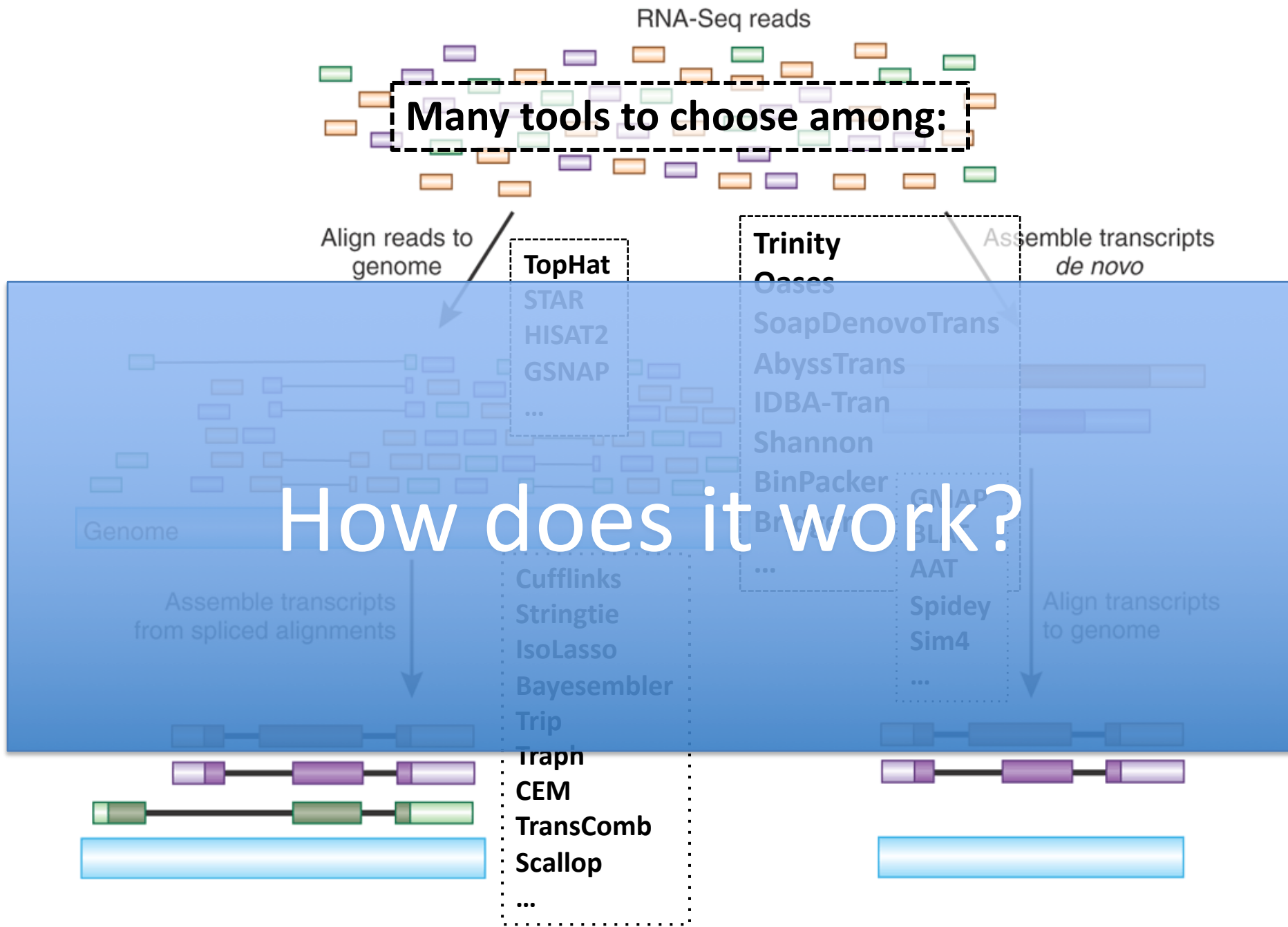
Nature Protocols **8**, 1494–1512 (2013) | doi:10.1038/nprot.2013.084

Published online 11 July 2013

Transcript Reconstruction from RNA-Seq Reads



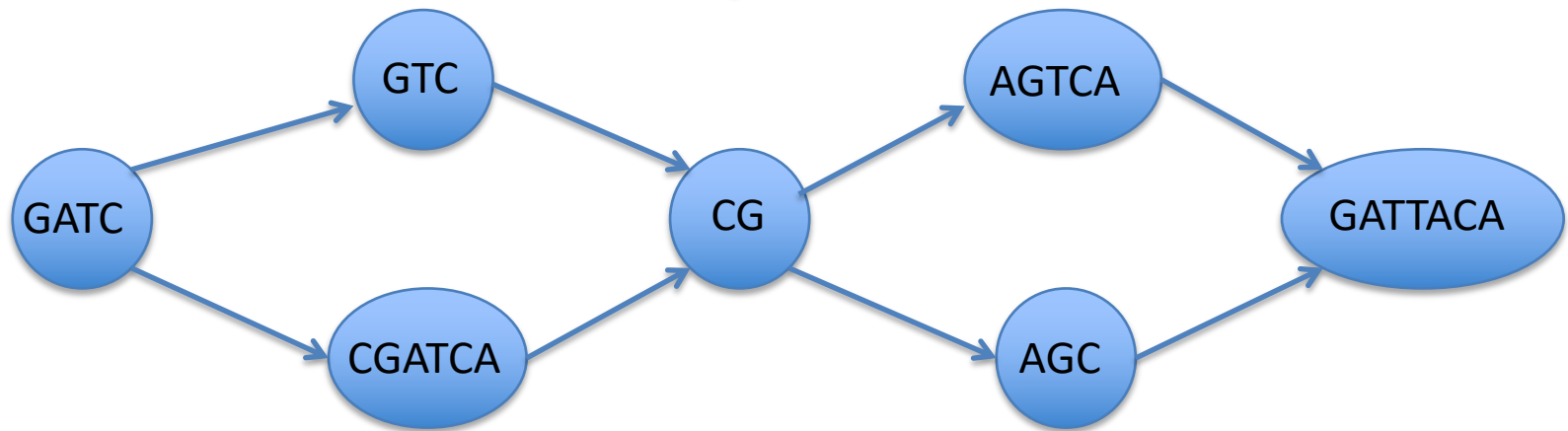
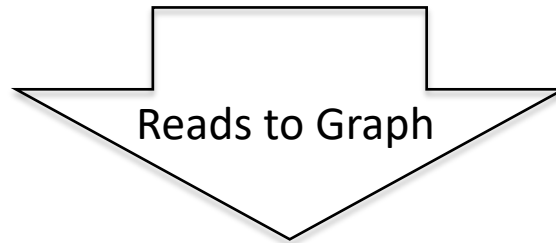
Transcript Reconstruction from RNA-Seq Reads



Graph Data Structures Commonly Used For Assembly



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

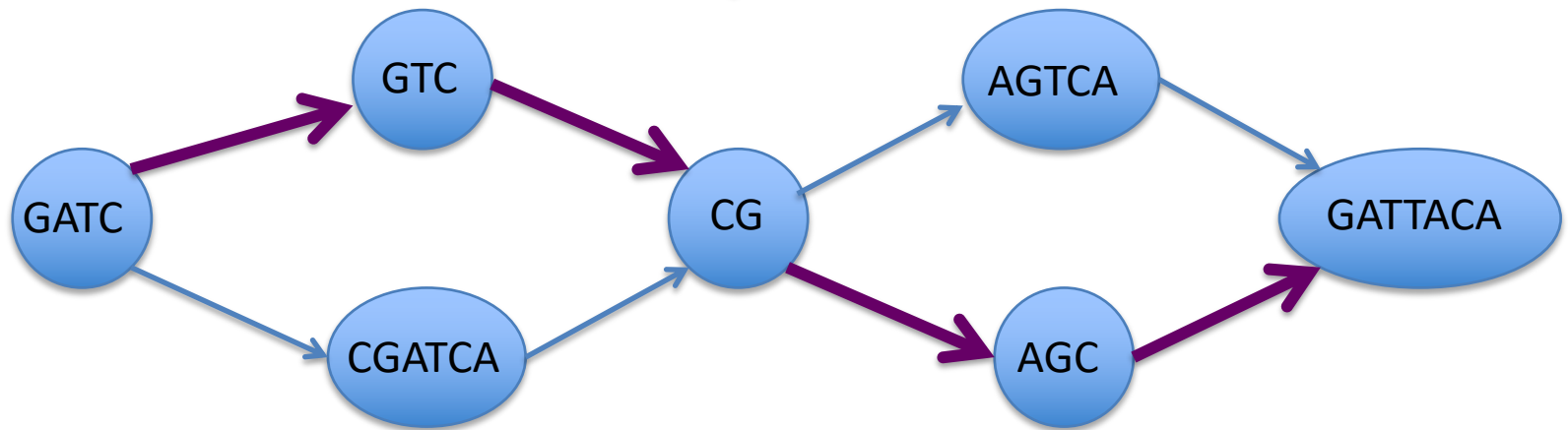
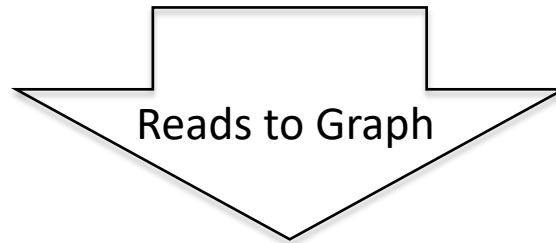


Nodes = sequence (+/-)
Edges = order, overlap

Graph Data Structures Commonly Used For Assembly



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

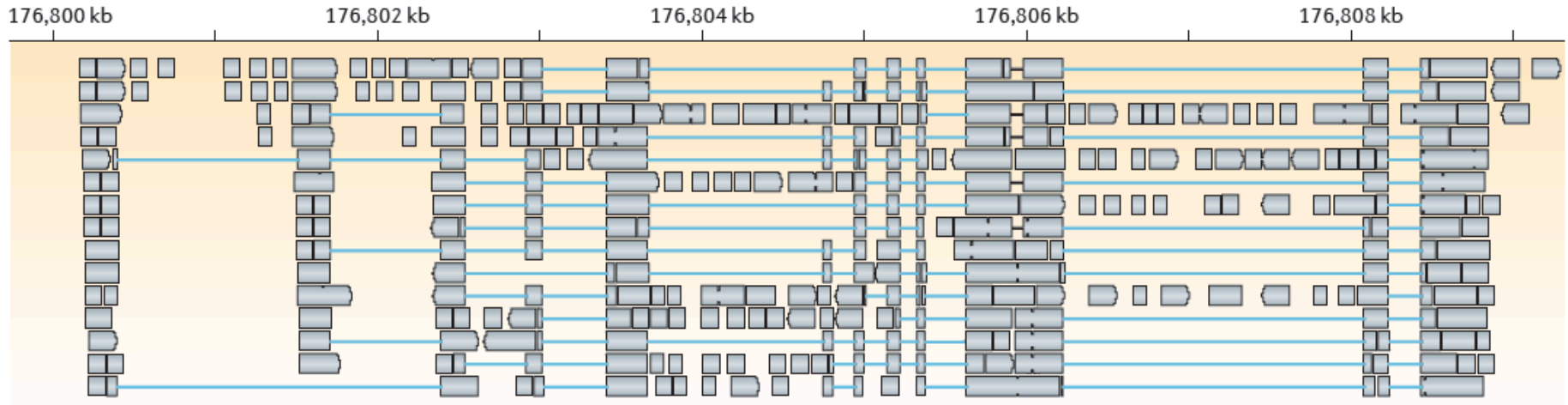


GATCGTCCGAGCGATTACA

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

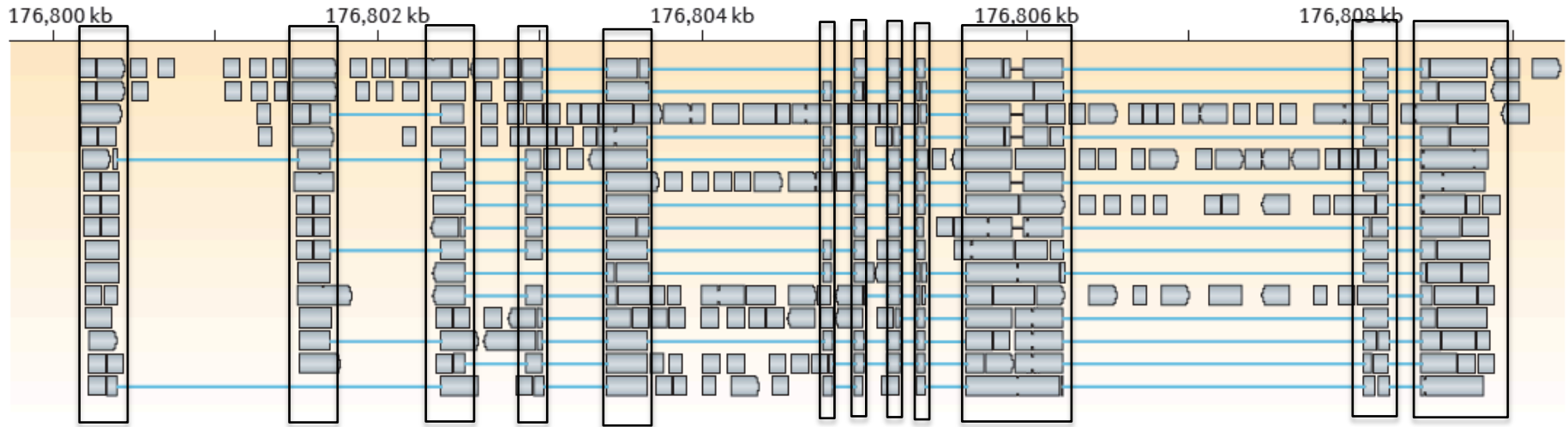
Genome-Guided Transcript Reconstruction

Splice-align reads to the genome



Genome-Guided Transcript Reconstruction

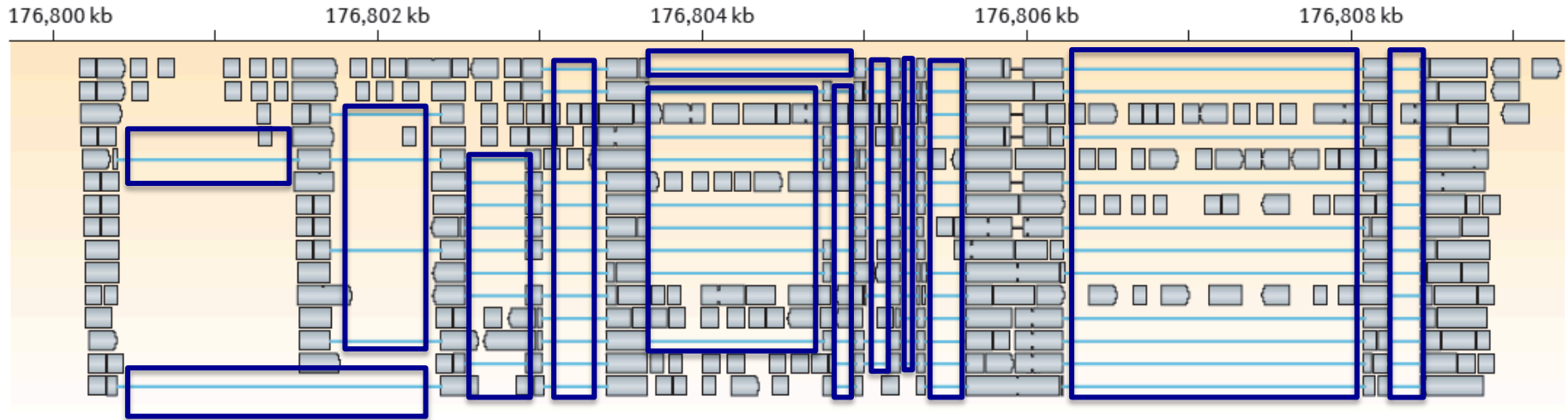
Splice-align reads to the genome



Alignment segment piles => exon regions

Genome-Guided Transcript Reconstruction

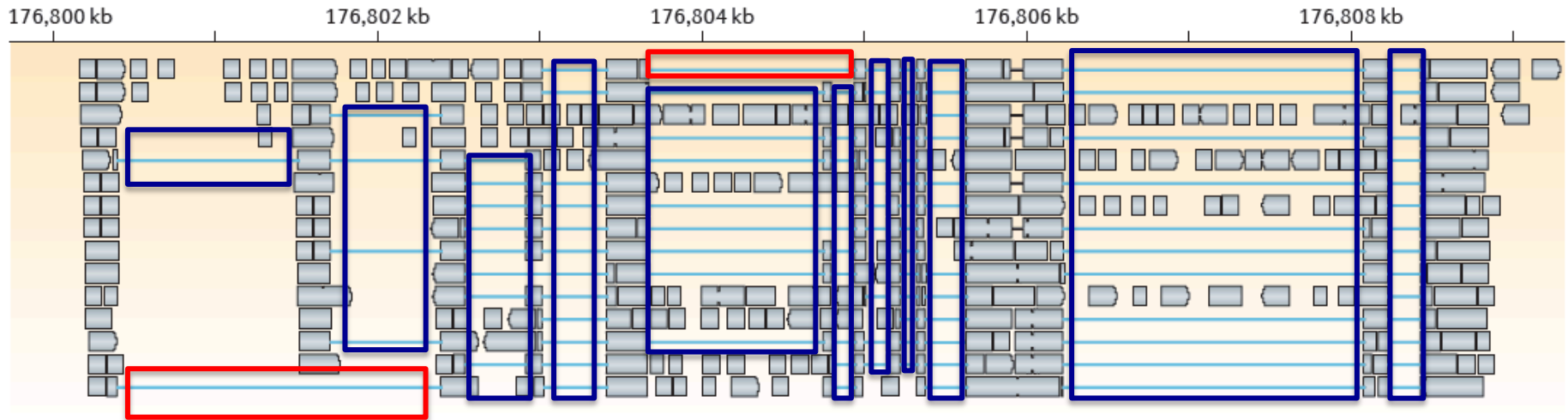
Splice-align reads to the genome



Large alignment gaps => introns

Genome-Guided Transcript Reconstruction

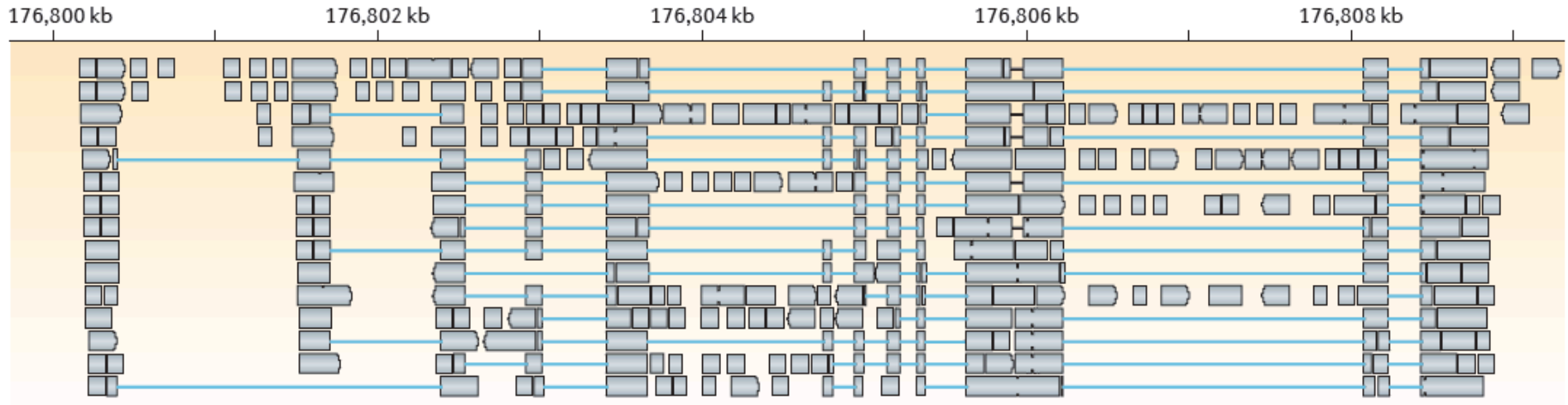
Splice-align reads to the genome



Overlapping but different introns = evidence of alternative splicing

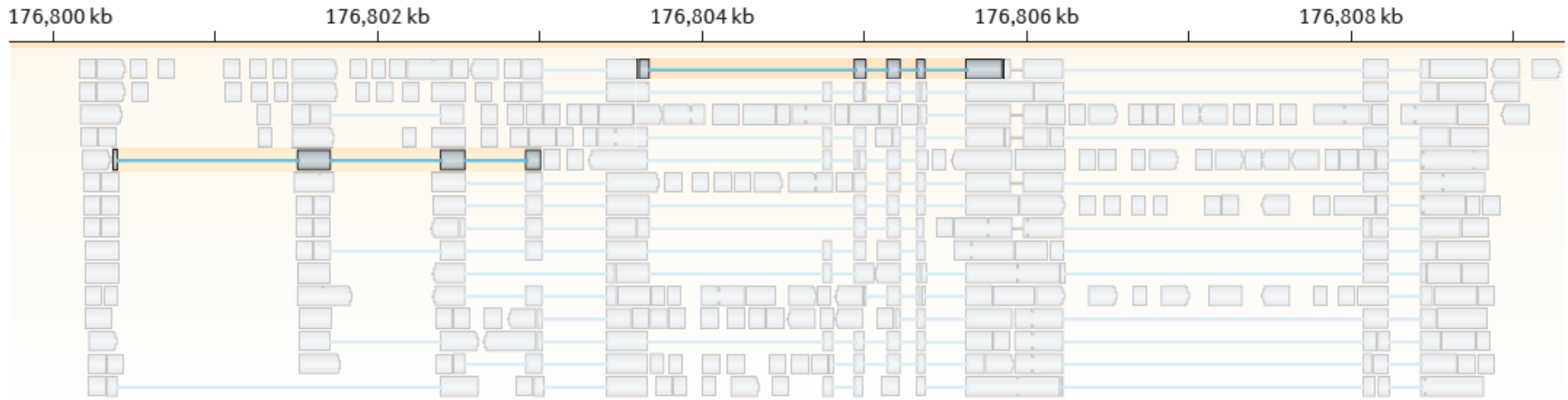
Genome-Guided Transcript Reconstruction

Splice-align reads to the genome



Genome-Guided Transcript Reconstruction

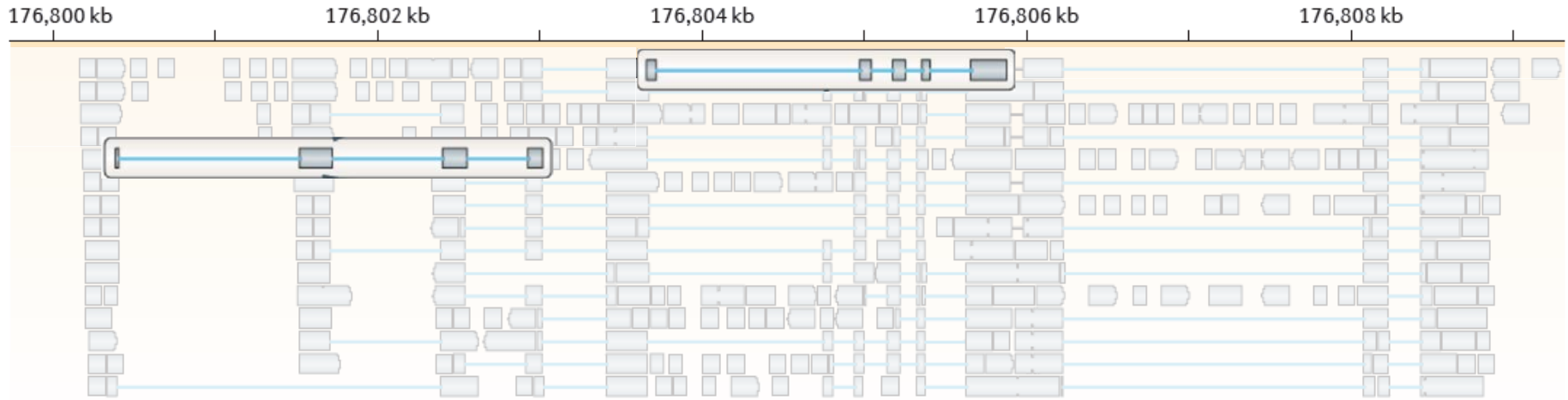
Splice-align reads to the genome



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

Genome-Guided Transcript Reconstruction

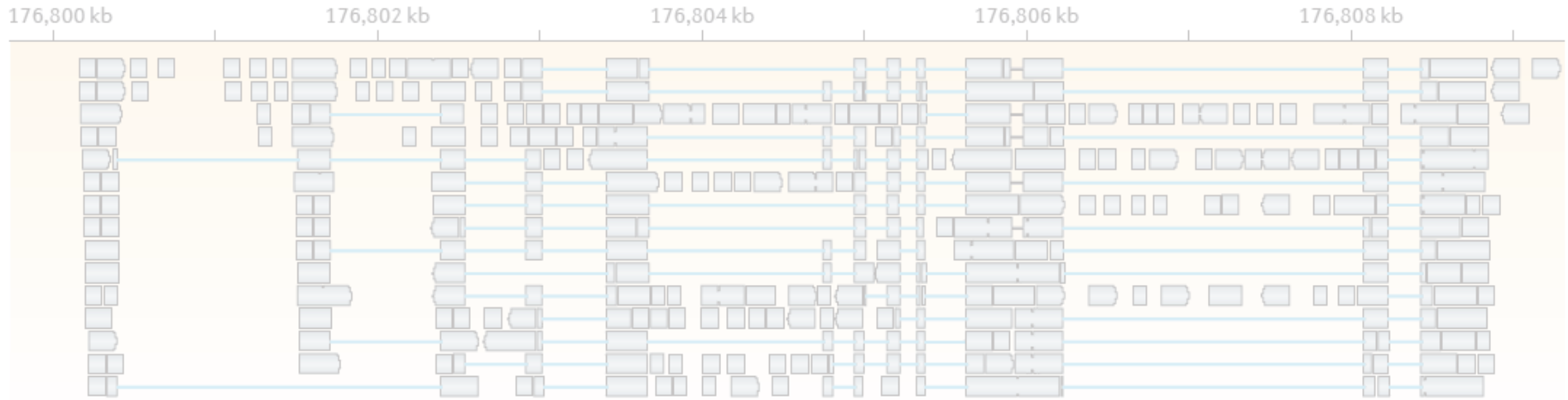
Splice-align reads to the genome



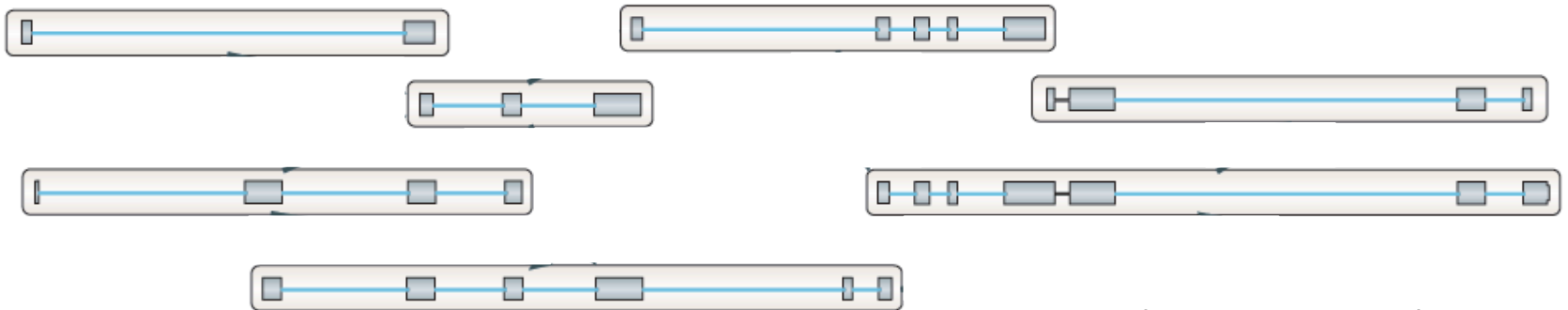
Nodes = unique splice patterns

Genome-Guided Transcript Reconstruction

Splice-align reads to the genome



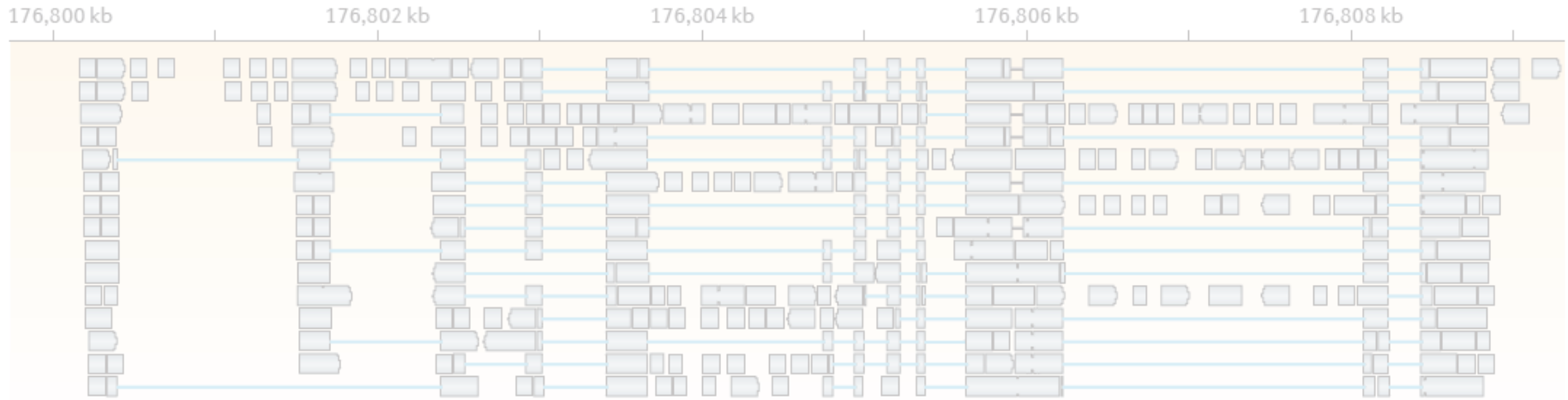
Construct graph from unique splice patterns of aligned reads.



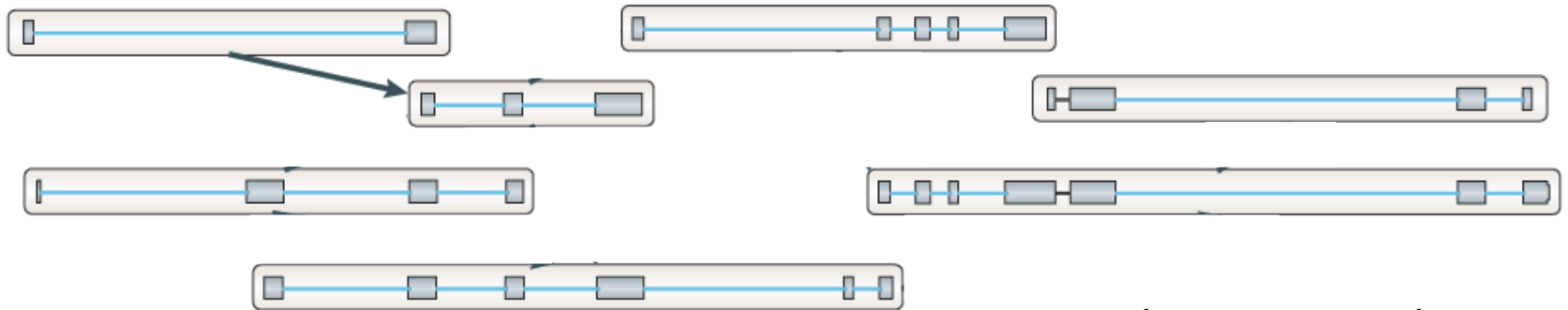
Nodes = unique splice patterns

Genome-Guided Transcript Reconstruction

Splice-align reads to the genome



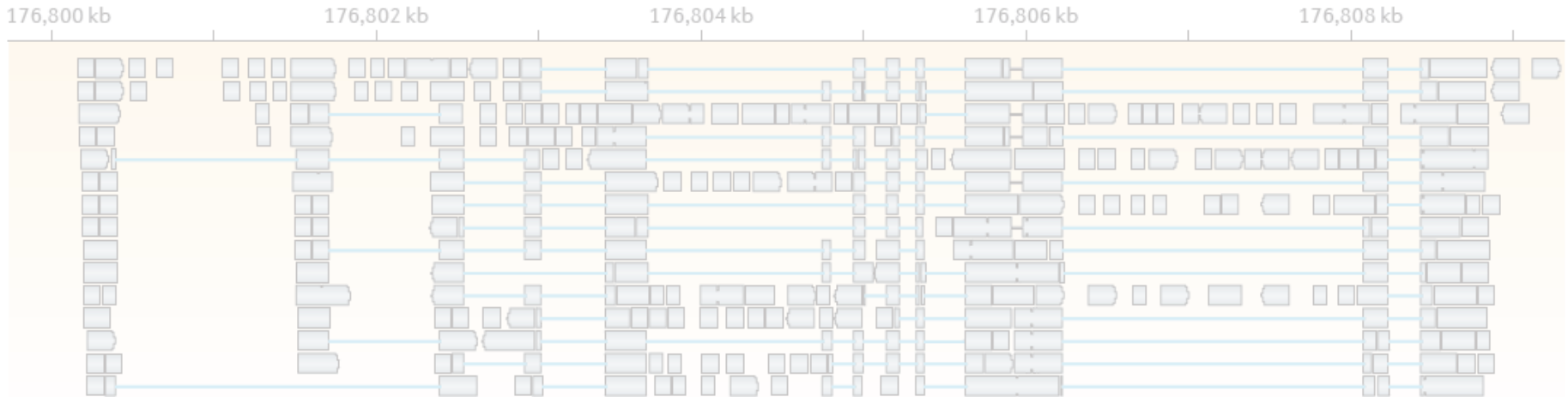
Construct graph from unique splice patterns of aligned reads.



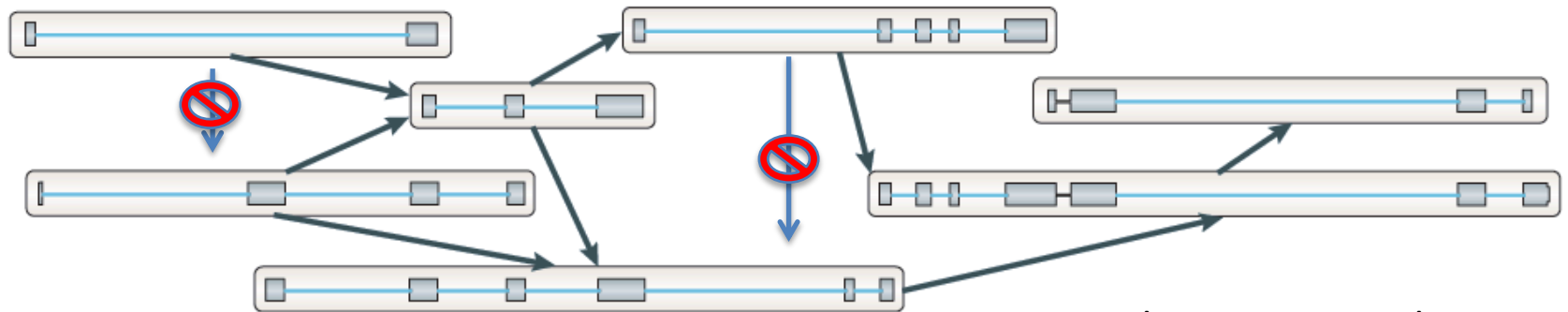
Nodes = unique splice patterns
Edges = compatible patterns

Genome-Guided Transcript Reconstruction

Splice-align reads to the genome



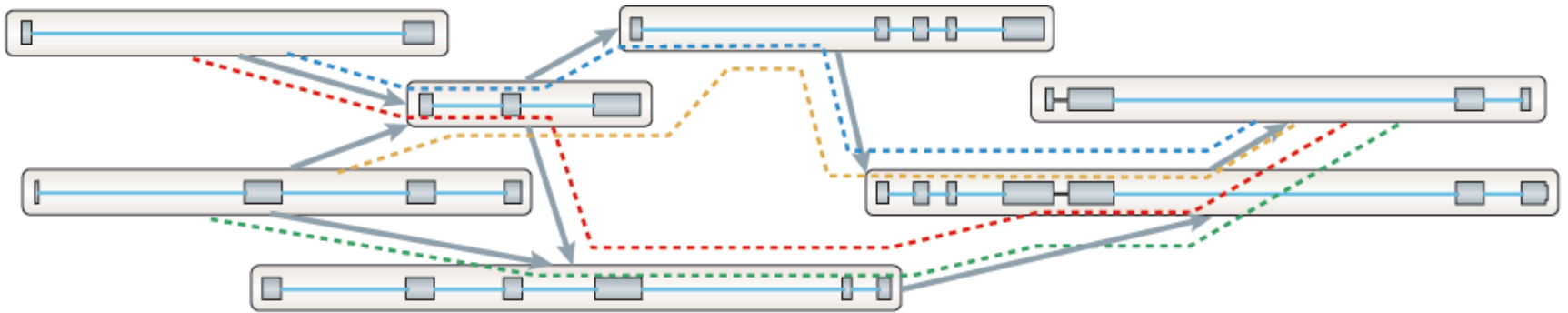
Construct graph from unique splice patterns of aligned reads.



Nodes = unique splice patterns
Edges = compatible patterns

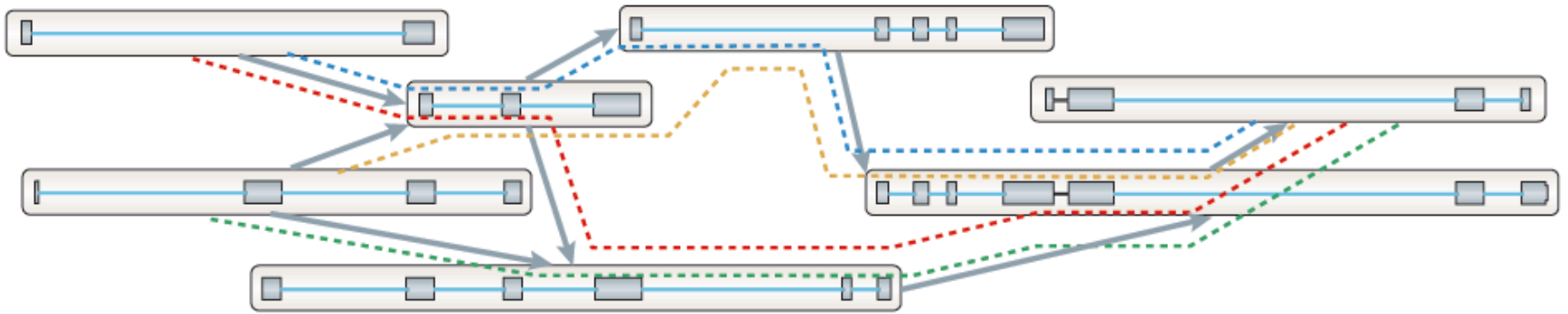
Genome-Guided Transcript Reconstruction

Traverse paths through the graph to assemble transcript isoforms

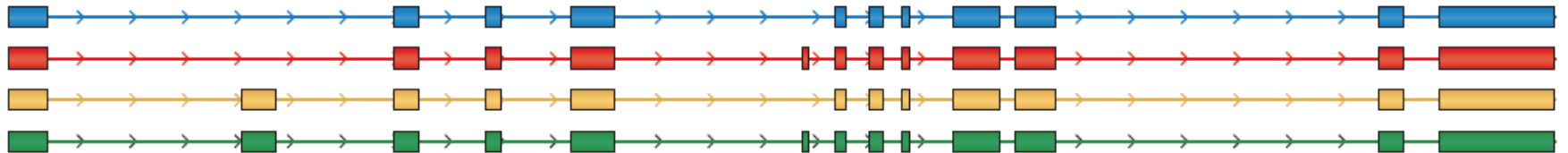


Genome-Guided Transcript Reconstruction

Traverse paths through the graph to assemble transcript isoforms



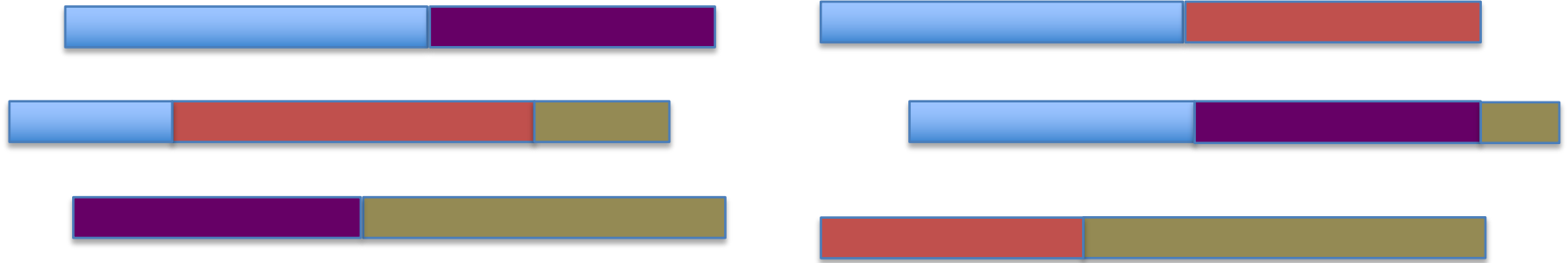
Reconstructed isoforms



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

Genome-free de novo transcript reconstruction to the rescue.

Read Overlap Graph: Reads as nodes, overlaps as edges



Read Overlap Graph: Reads as nodes, overlaps as edges



Node = read
Edge = overlap

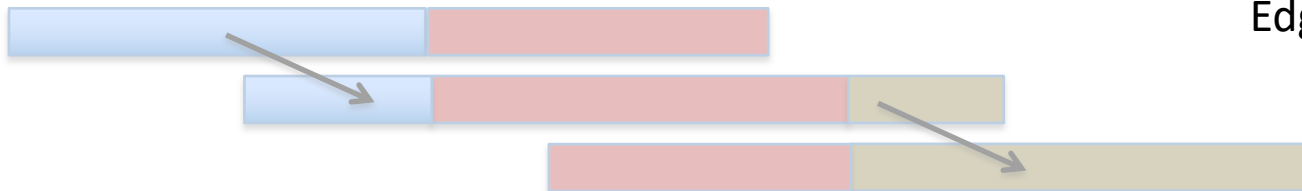
Read Overlap Graph: Reads as nodes, overlaps as edges



Transcript A



Generate consensus sequence where reads overlap

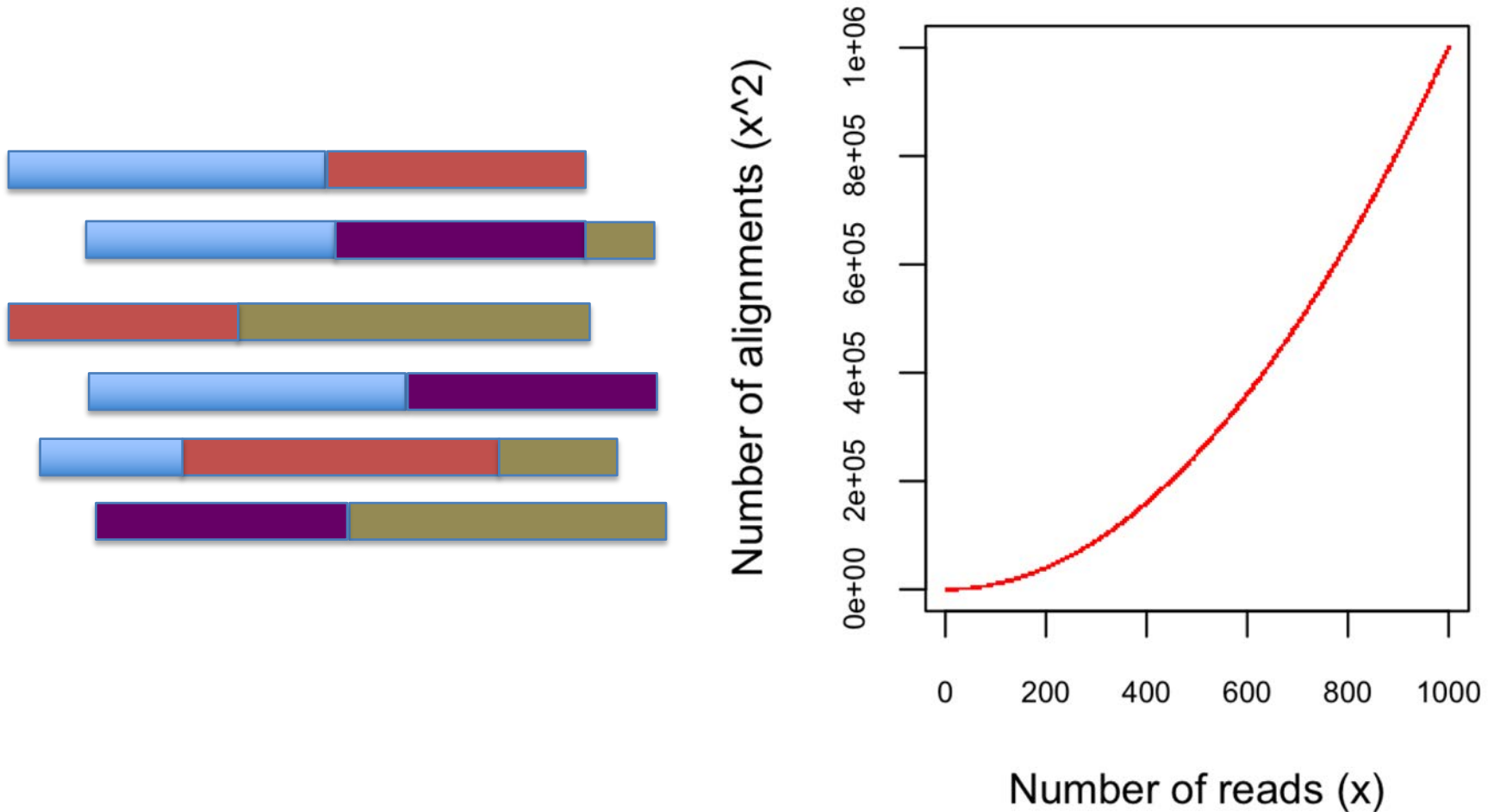


Node = read
Edge = overlap

Transcript B

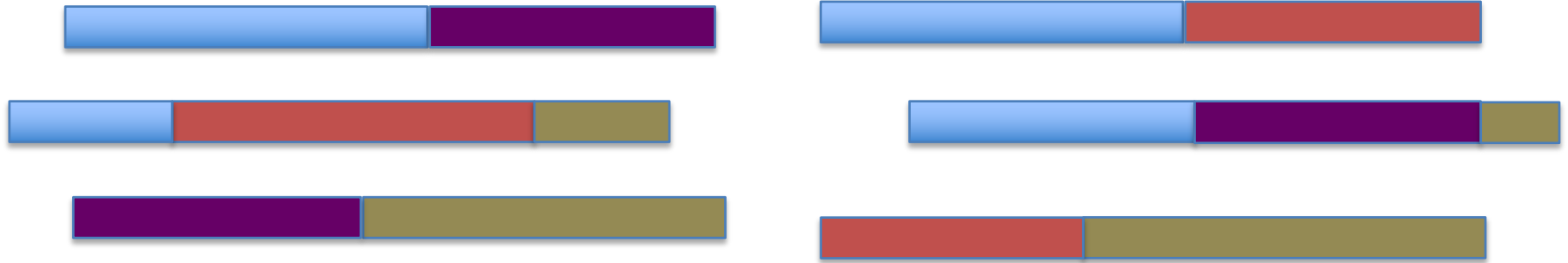


Finding pairwise overlaps between n reads involves $\sim n^2$ comparisons.



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

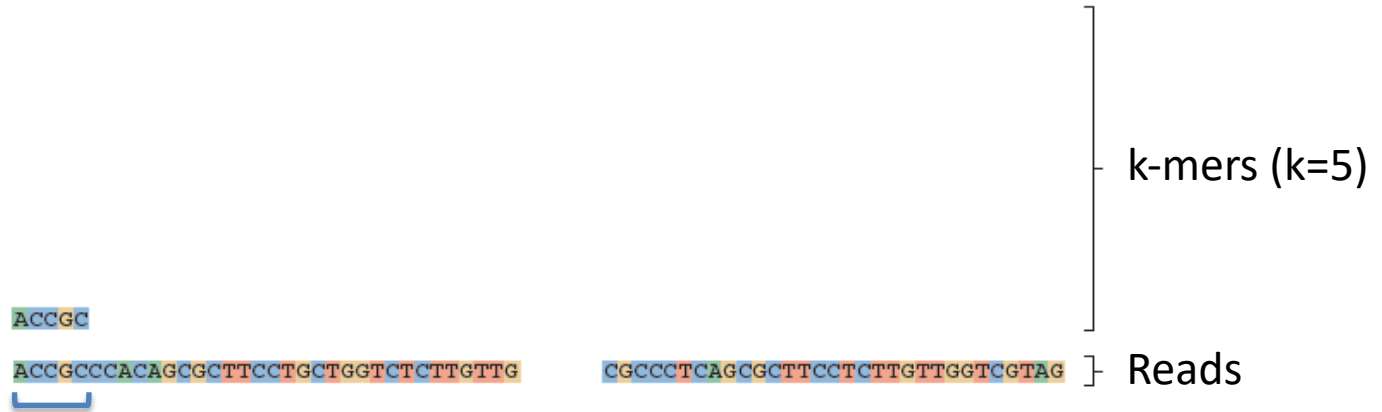
Sequence Assembly via de Bruijn Graphs

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

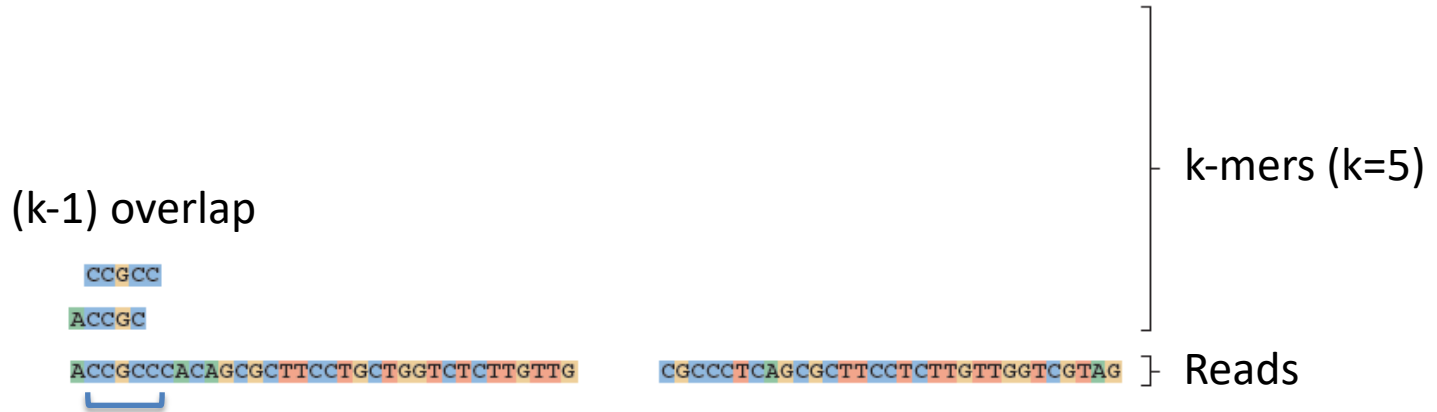


Construct the de Bruijn graph



Sequence Assembly via De Bruijn Graphs

Generate all substrings of length k from the reads

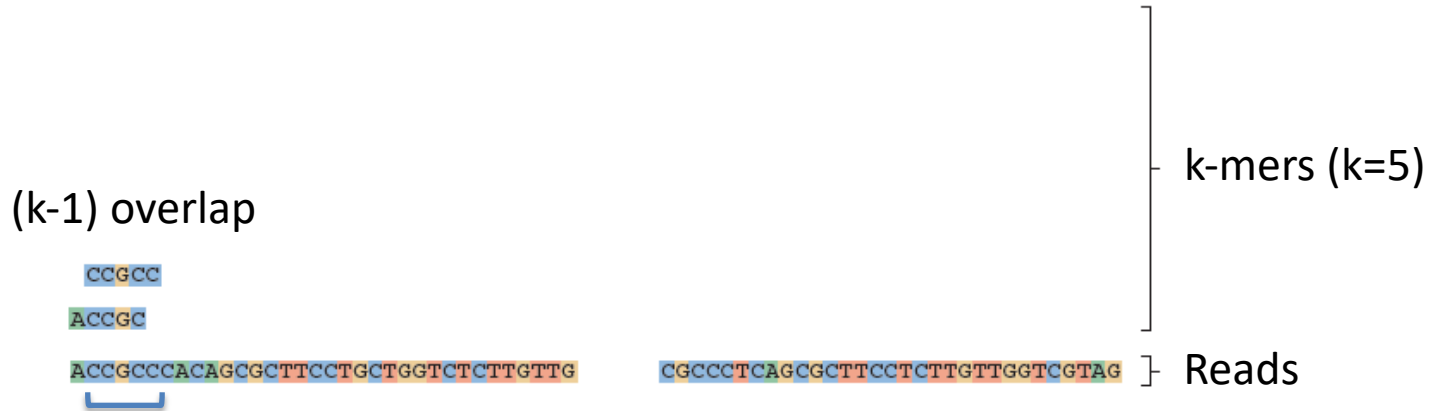


Construct the de Bruijn graph

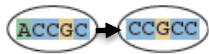


Sequence Assembly via De Bruijn Graphs

Generate all substrings of length k from the reads



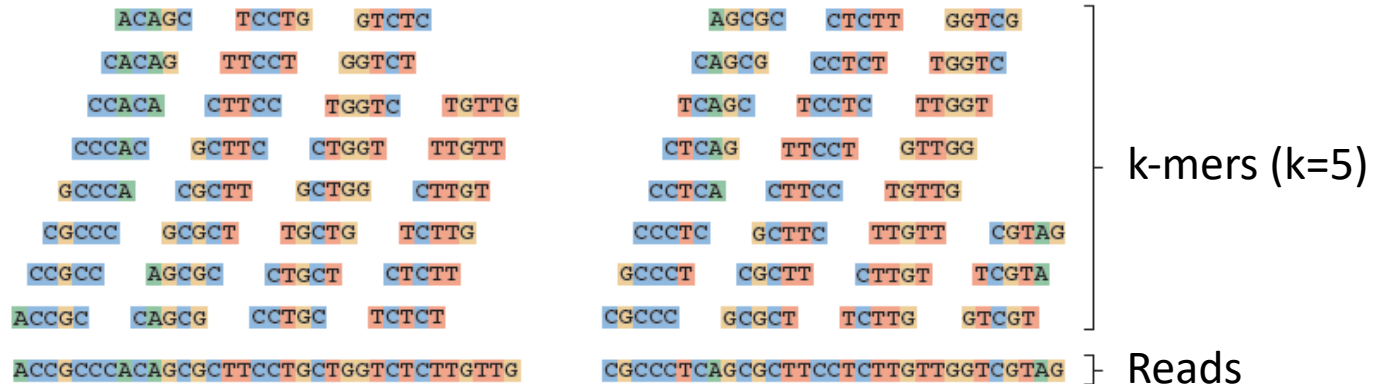
Construct the de Bruijn graph



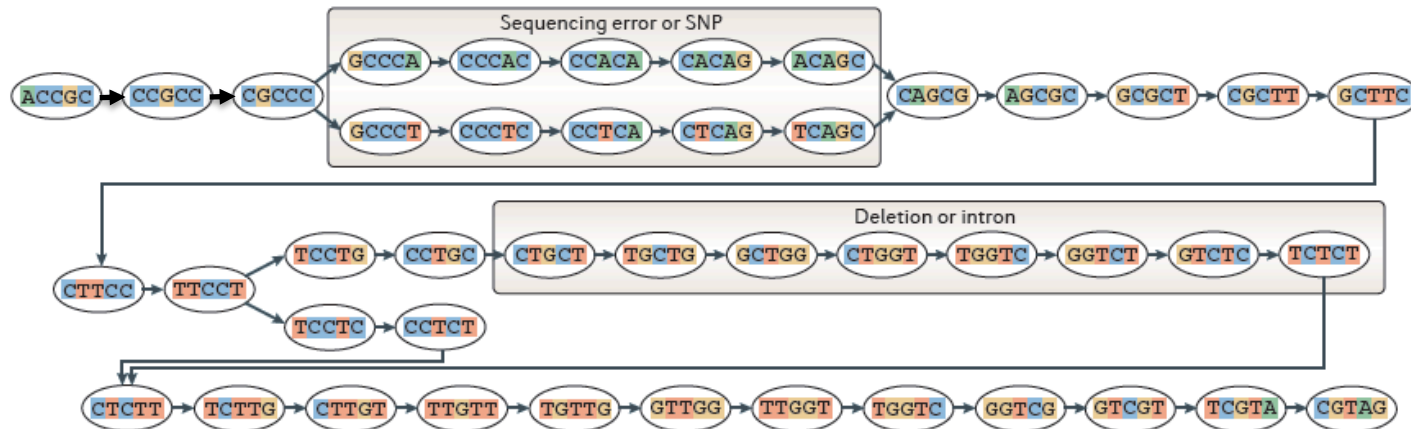
Nodes = unique k -mers
Edges = overlap by ($k-1$)

Sequence Assembly via De Bruijn Graphs

Generate all substrings of length k from the reads

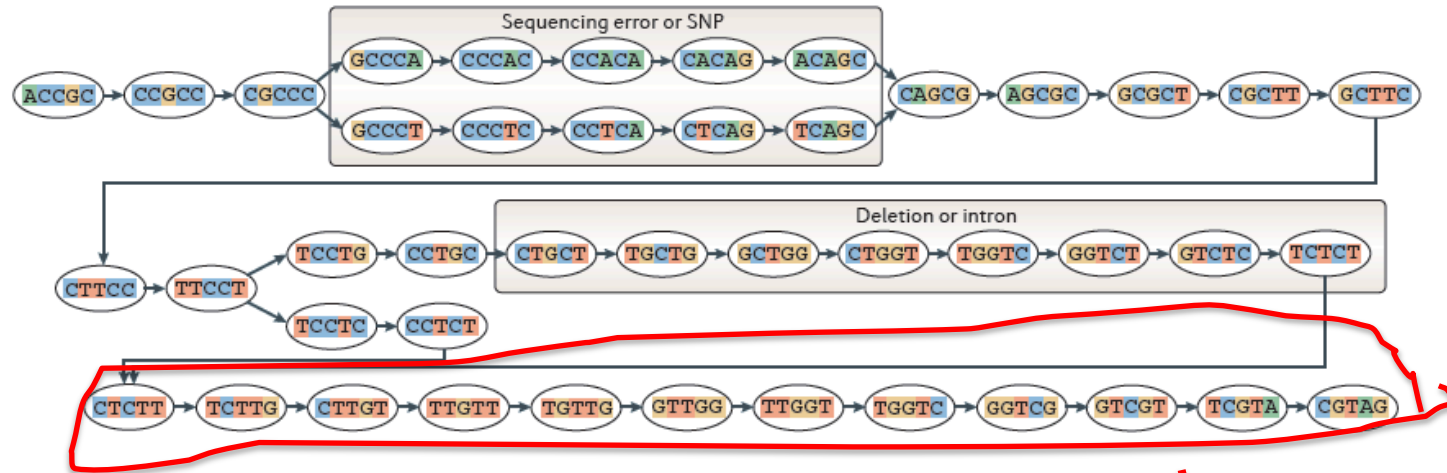


Construct the de Bruijn graph

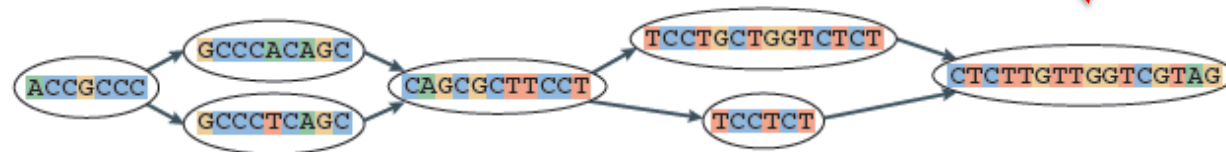


Nodes = unique k-mers
Edges = overlap by $(k-1)$

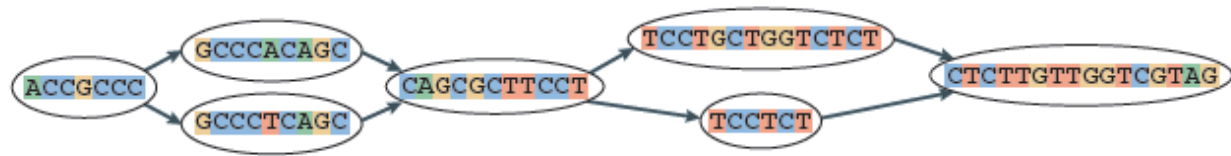
Construct the de Bruijn graph



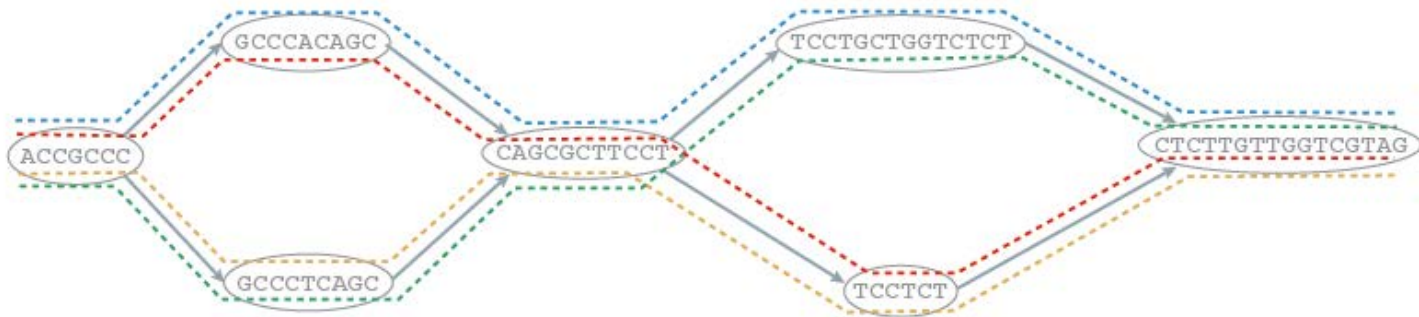
Collapse the de Bruijn graph



Collapse the de Bruijn graph



Traverse the graph



Assemble Transcript Isoforms

```

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

Contrasting Genome and Transcriptome Assembly

Genome Assembly

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

Transcriptome Assembly

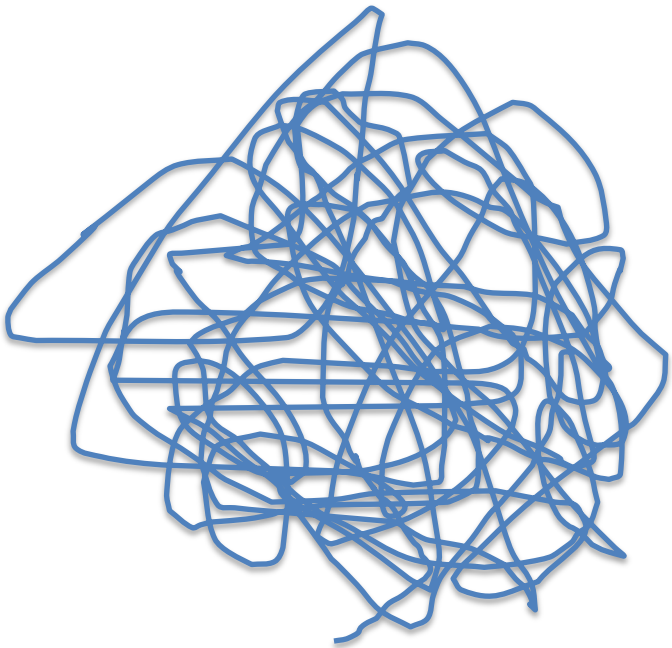
- Exponentially distributed coverage levels
- Multiple contigs per locus (alt splicing)
- Strand-specific



Trinity Aggregates Isolated Transcript Graphs

Genome Assembly

Single Massive Graph



Entire chromosomes represented.

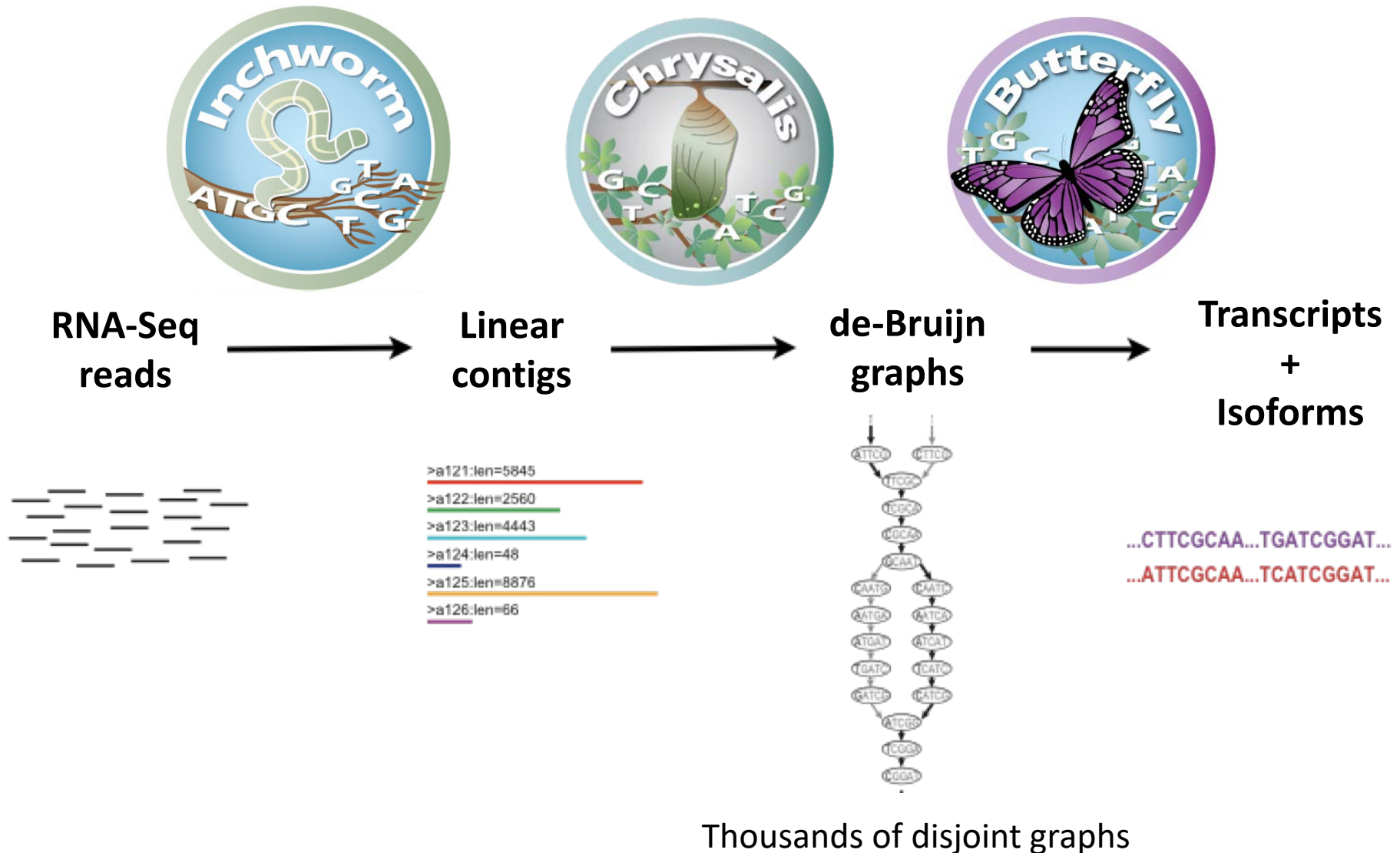
Trinity Transcriptome Assembly

Many Thousands of Small Graphs



Ideally, one graph per expressed gene.

Trinity – How it works:





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

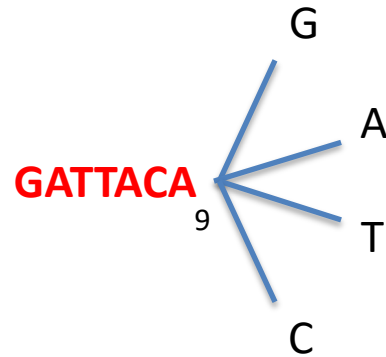
Kmer Catalog (hashtable)

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



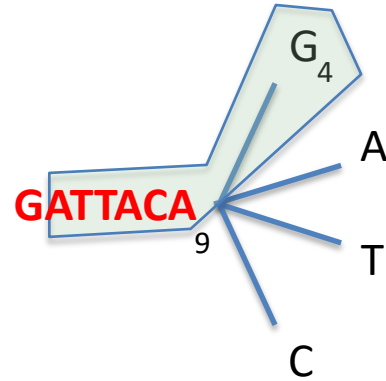
Inchworm Algorithm

- 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.



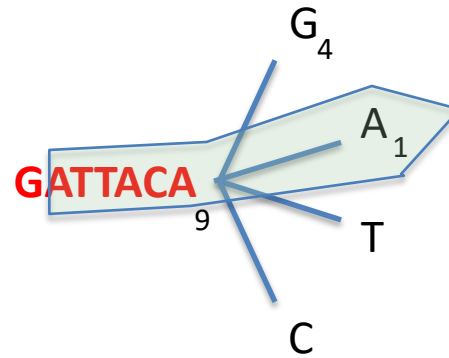


Inchworm Algorithm



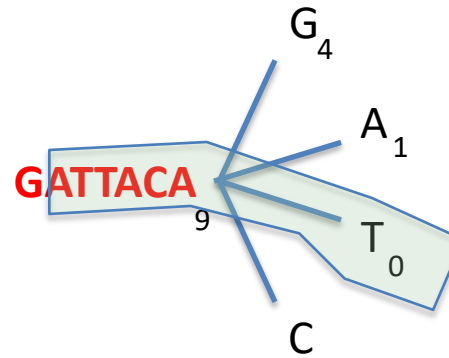


Inchworm Algorithm



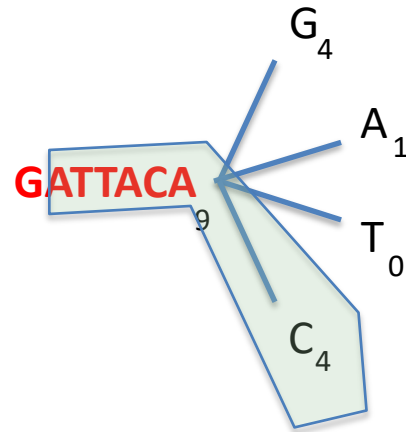


Inchworm Algorithm



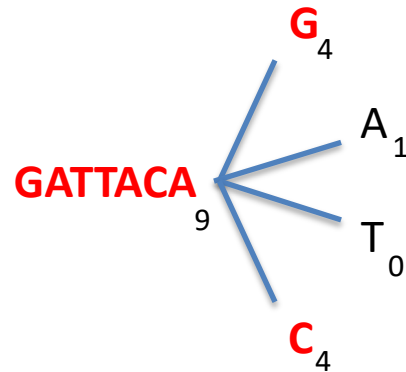


Inchworm Algorithm



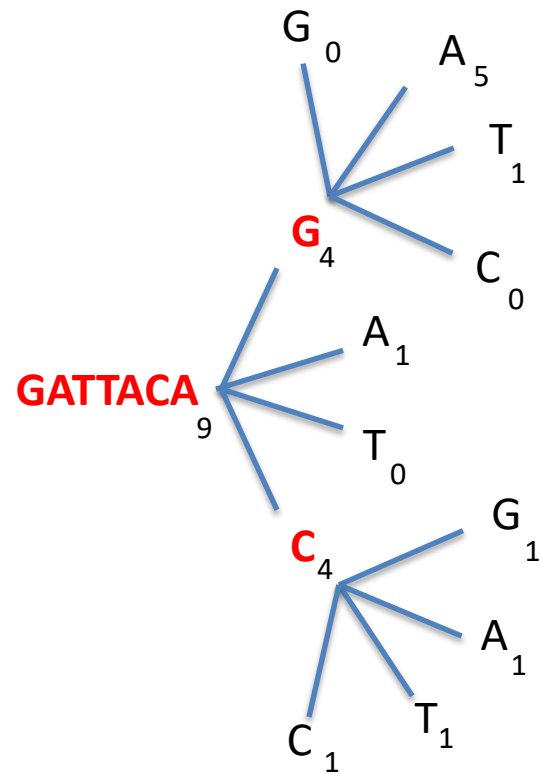


Inchworm Algorithm



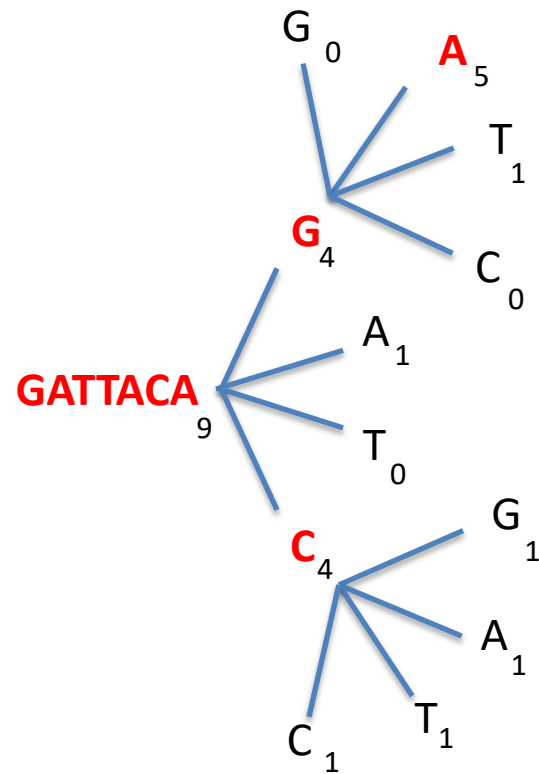


Inchworm Algorithm





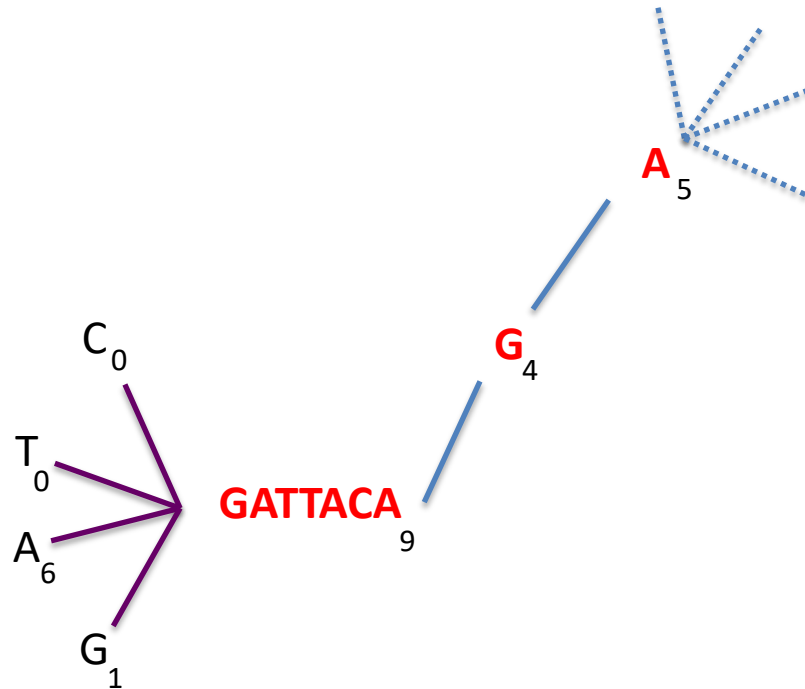
Inchworm Algorithm





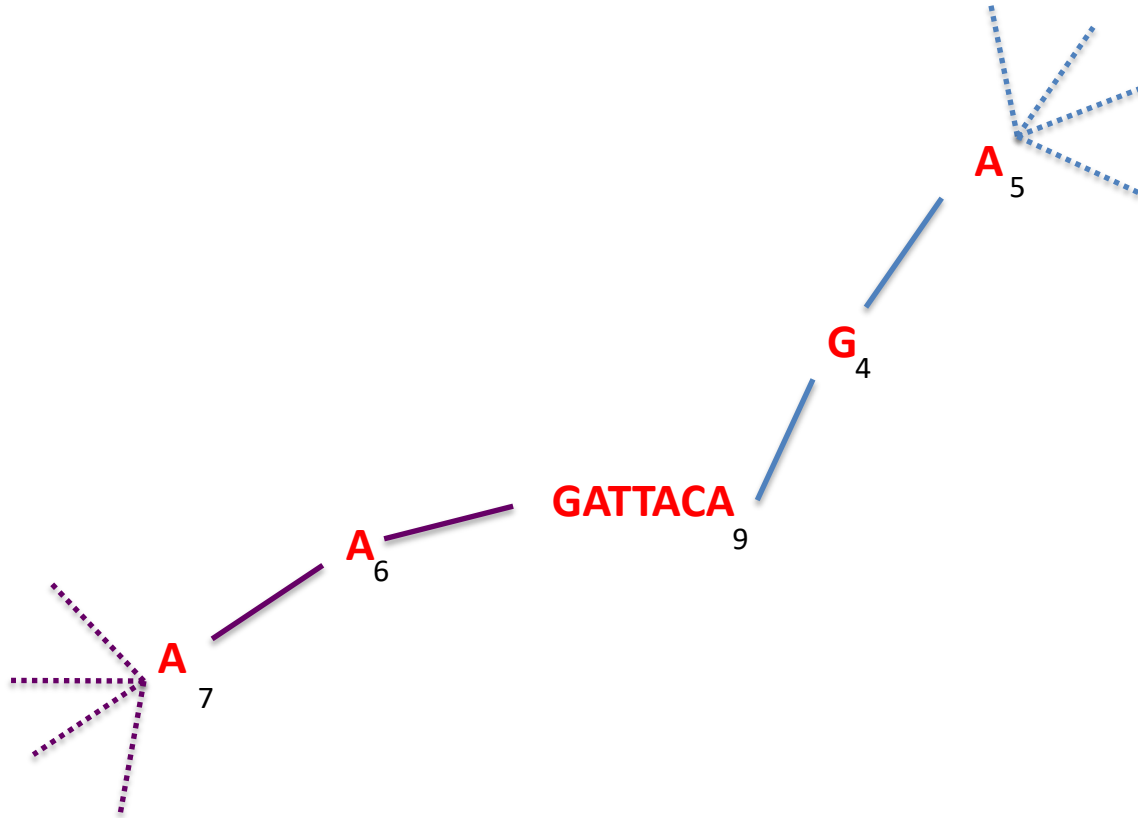


Inchworm Algorithm





Inchworm Algorithm



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

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



Inchworm Contigs from Alt-Spliced Transcripts

Expressed isoforms





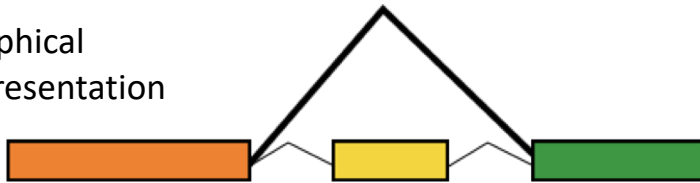
Inchworm Contigs from Alt-Spliced Transcripts

Expressed isoforms

Expression



Graphical
representation



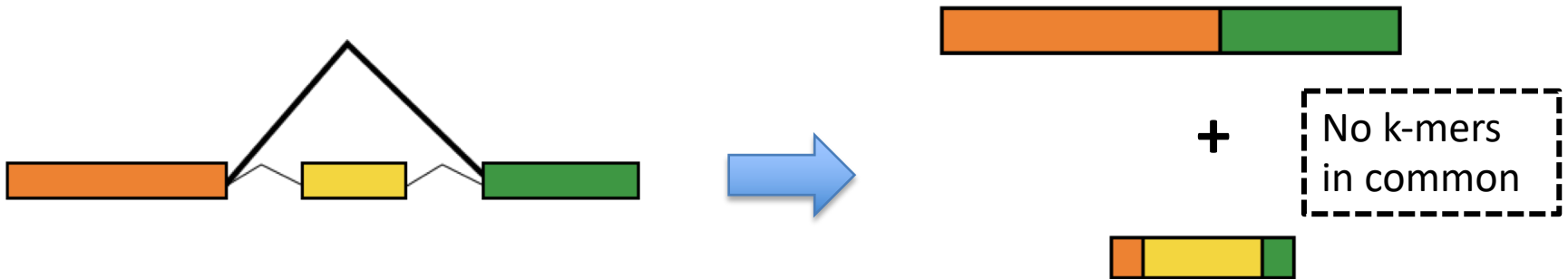


Inchworm Contigs from Alt-Spliced Transcripts



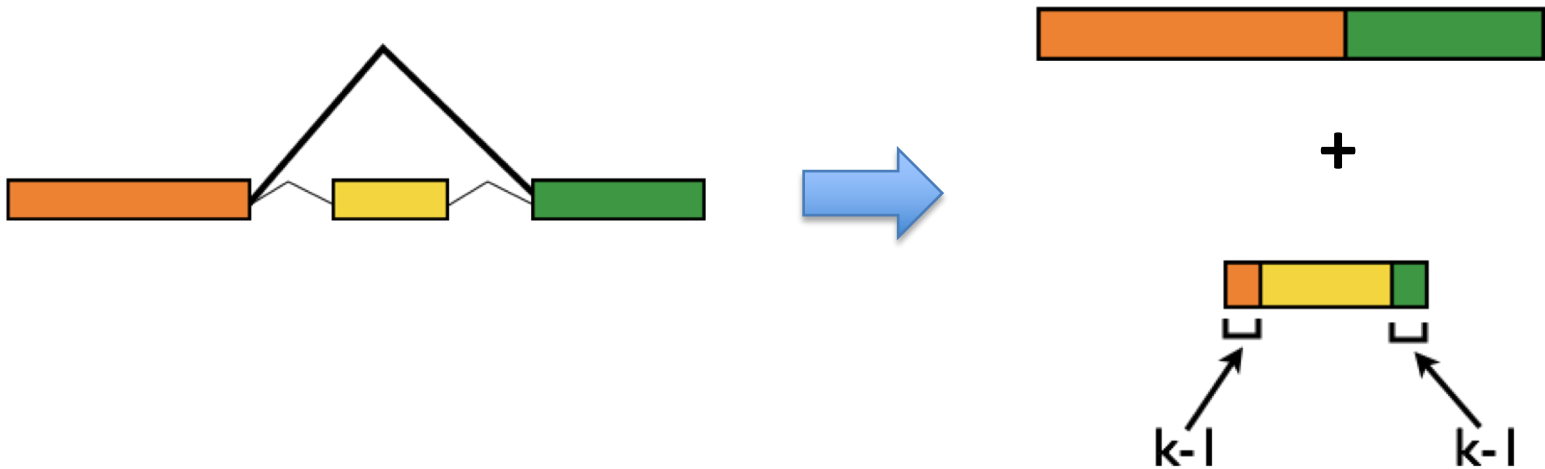


Inchworm Contigs from Alt-Spliced Transcripts

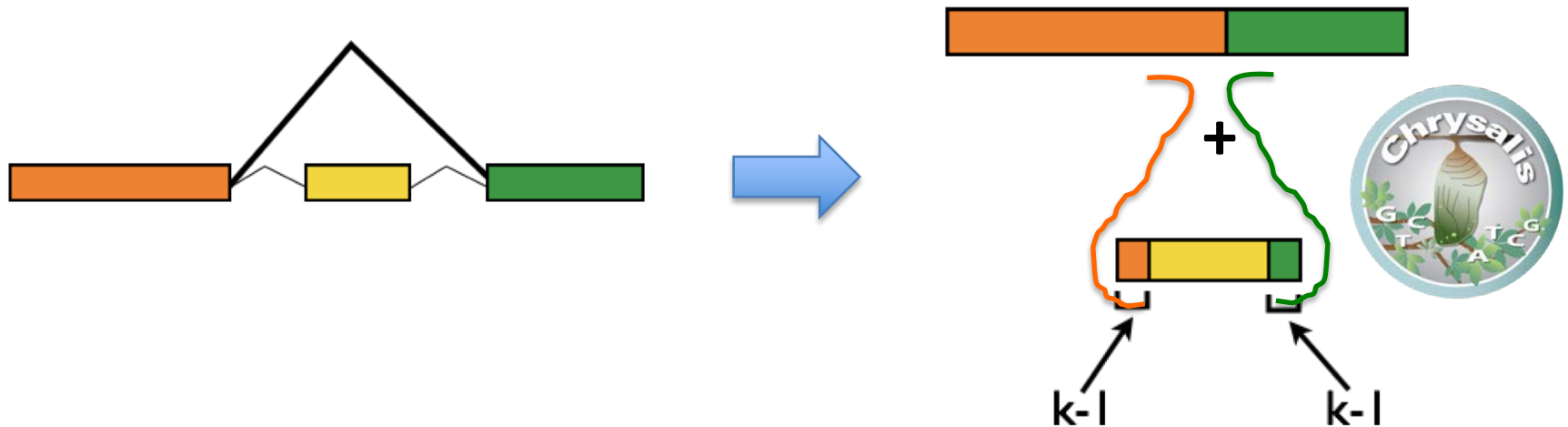




Inchworm Contigs from Alt-Spliced Transcripts



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

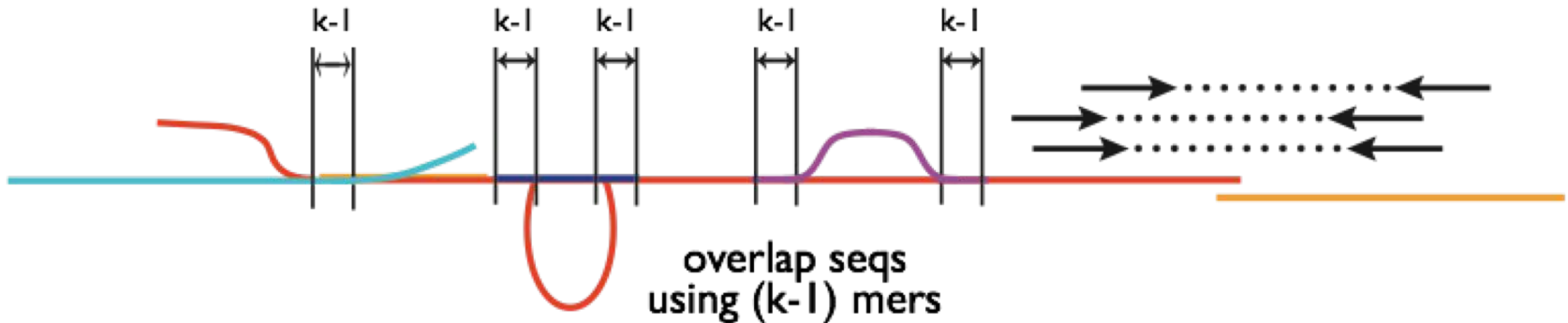
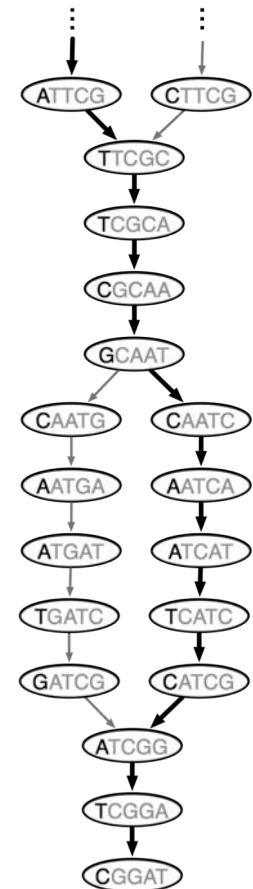
>a125:len=8876

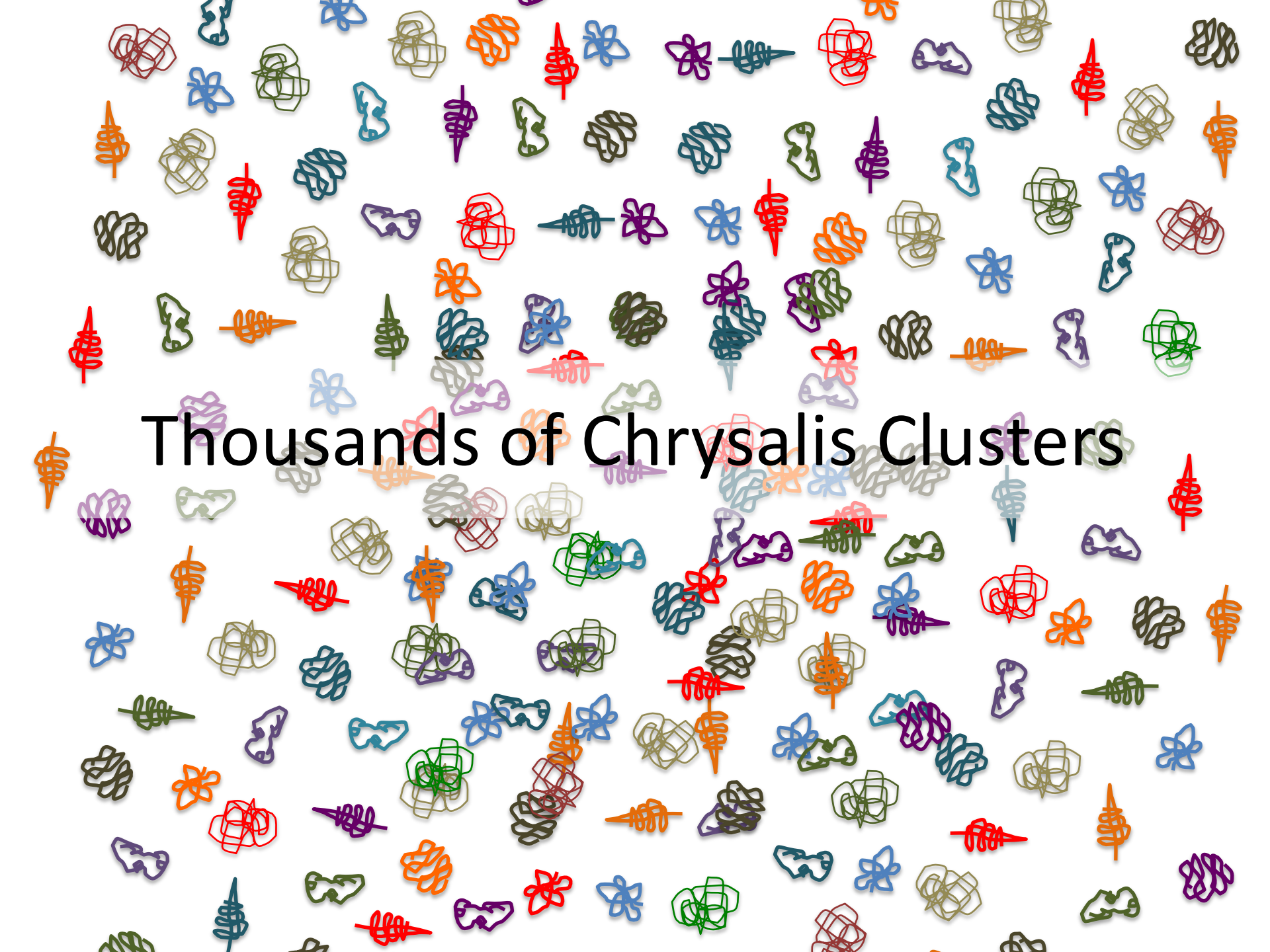
>a126:len=68



Integrate isoforms
via $k-1$ overlaps

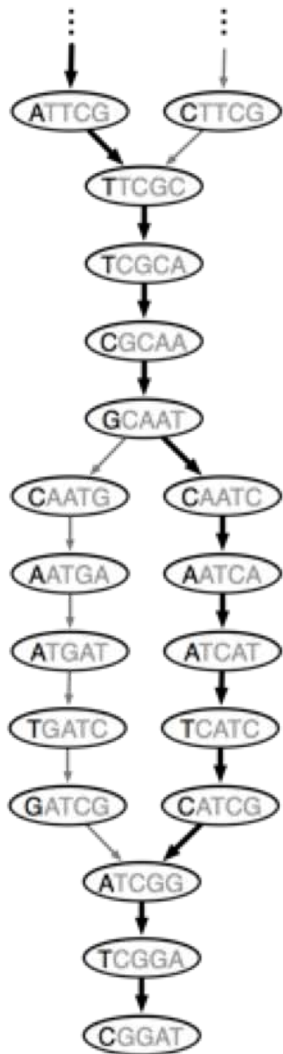
Build de Bruijn Graphs
(ideally, one per gene)



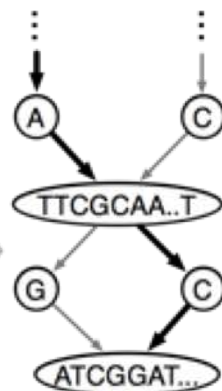


Thousands of Chrysalis Clusters

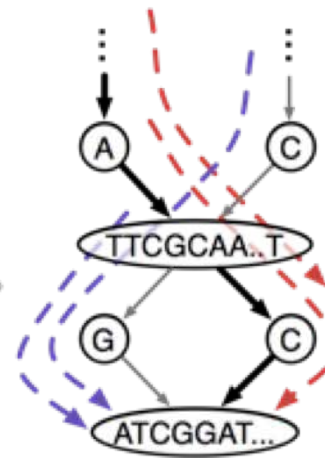
Butterfly



de Bruijn
graph



compact
graph



compact
graph with
reads

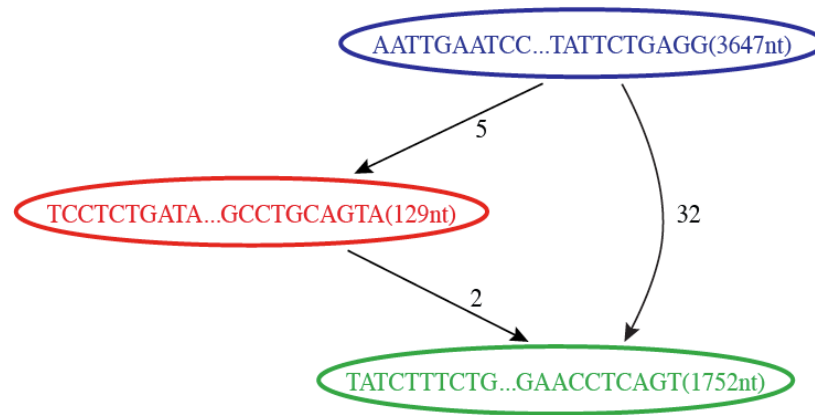


..CTTCGCAA..TGATCGGAT..
..ATTGCAA..TCATCGGAT..

sequences
(isoforms and paralogs)

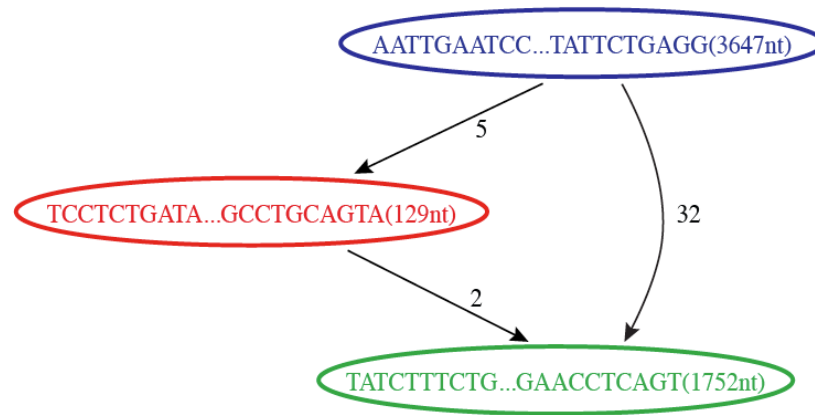
Butterfly Example 1: Reconstruction of Alternatively Spliced Transcripts

Butterfly's Compacted
Sequence Graph



Reconstruction of Alternatively Spliced Transcripts

Butterfly's Compacted
Sequence Graph

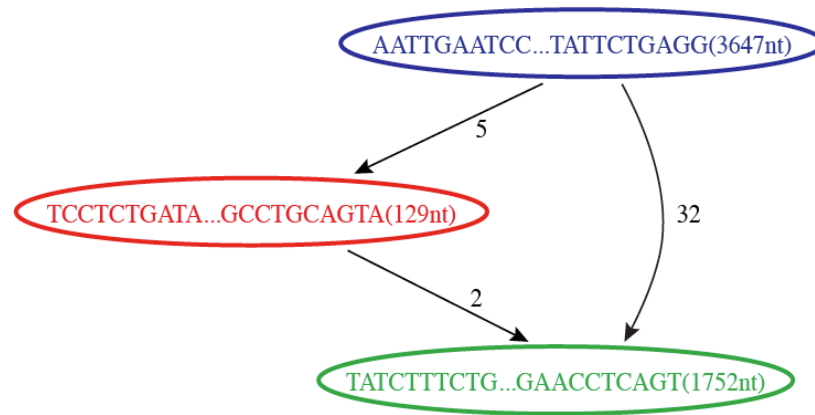


Reconstructed Transcripts



Reconstruction of Alternatively Spliced Transcripts

Butterfly's Compacted
Sequence Graph

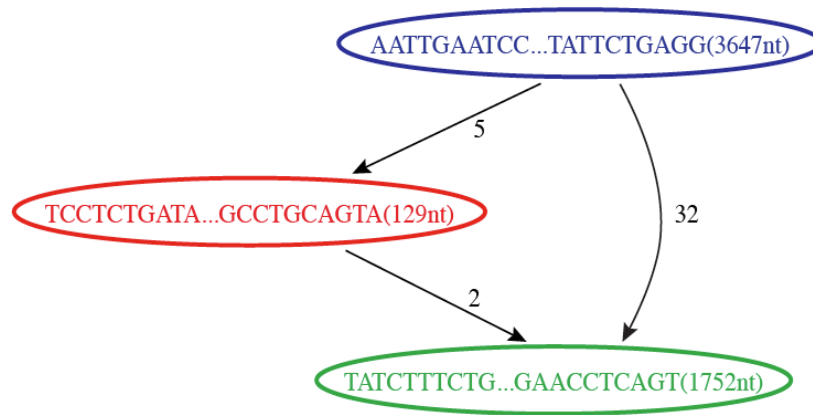


Reconstructed Transcripts



Reconstruction of Alternatively Spliced Transcripts

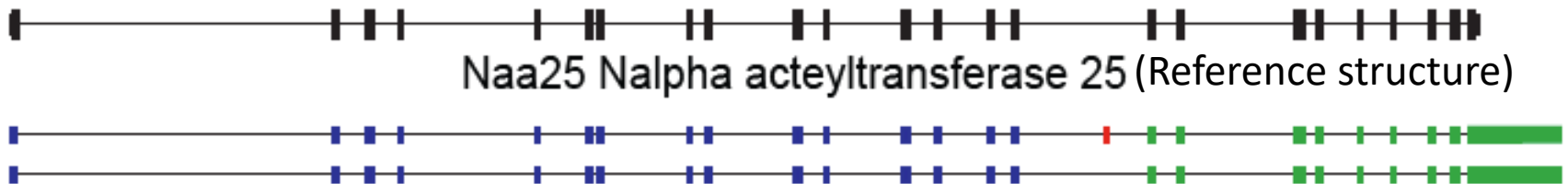
Butterfly's Compacted
Sequence Graph



Reconstructed Transcripts



Aligned to Mouse Genome



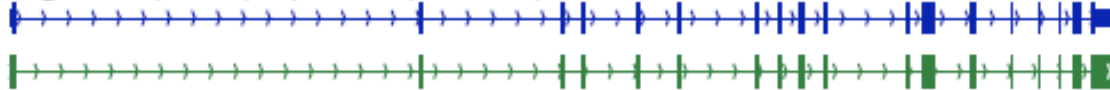
Butterfly Example 2: Teasing Apart Transcripts of Paralogous Genes



Teasing Apart Transcripts of Paralogous Genes

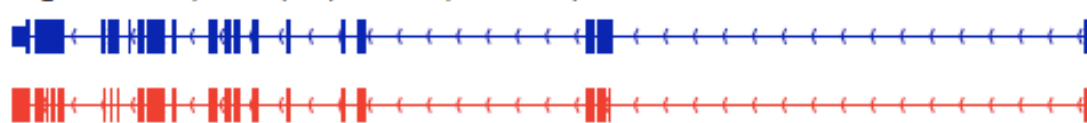
chr7:148,744,197-148,821,437

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



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

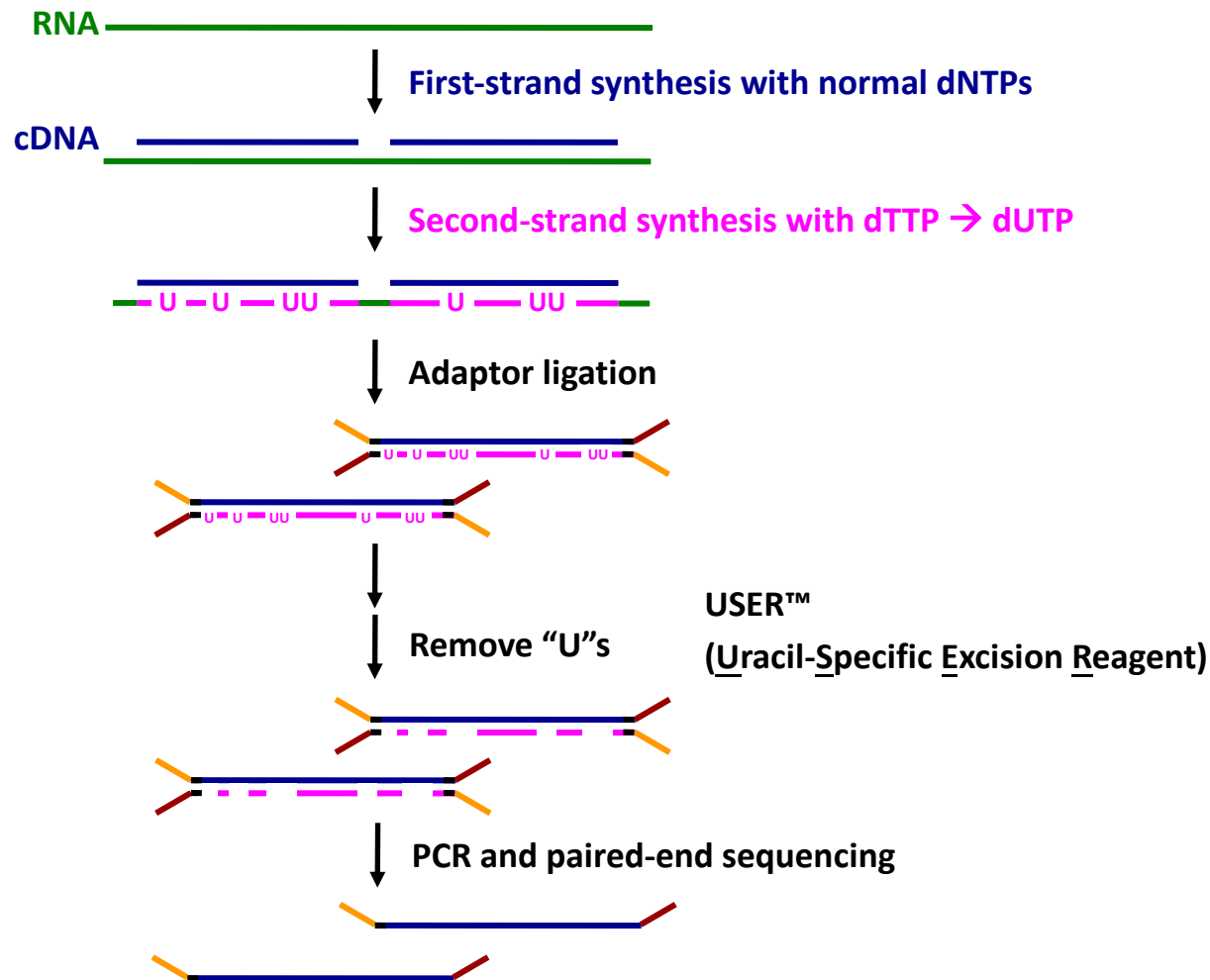
'dUTP second strand marking' identified as the leading protocol

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

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

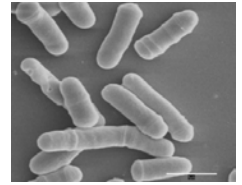
transcribed strand or other noncoding events; demarcate the exact boundaries of adjacent genes transcribed on opposite strands and resolve the correct expression levels of coding or noncoding overlapping transcripts. These tasks are particularly challenging in small microbial genomes, prokaryotic and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

dUTP 2nd Strand Method: Our Favorite

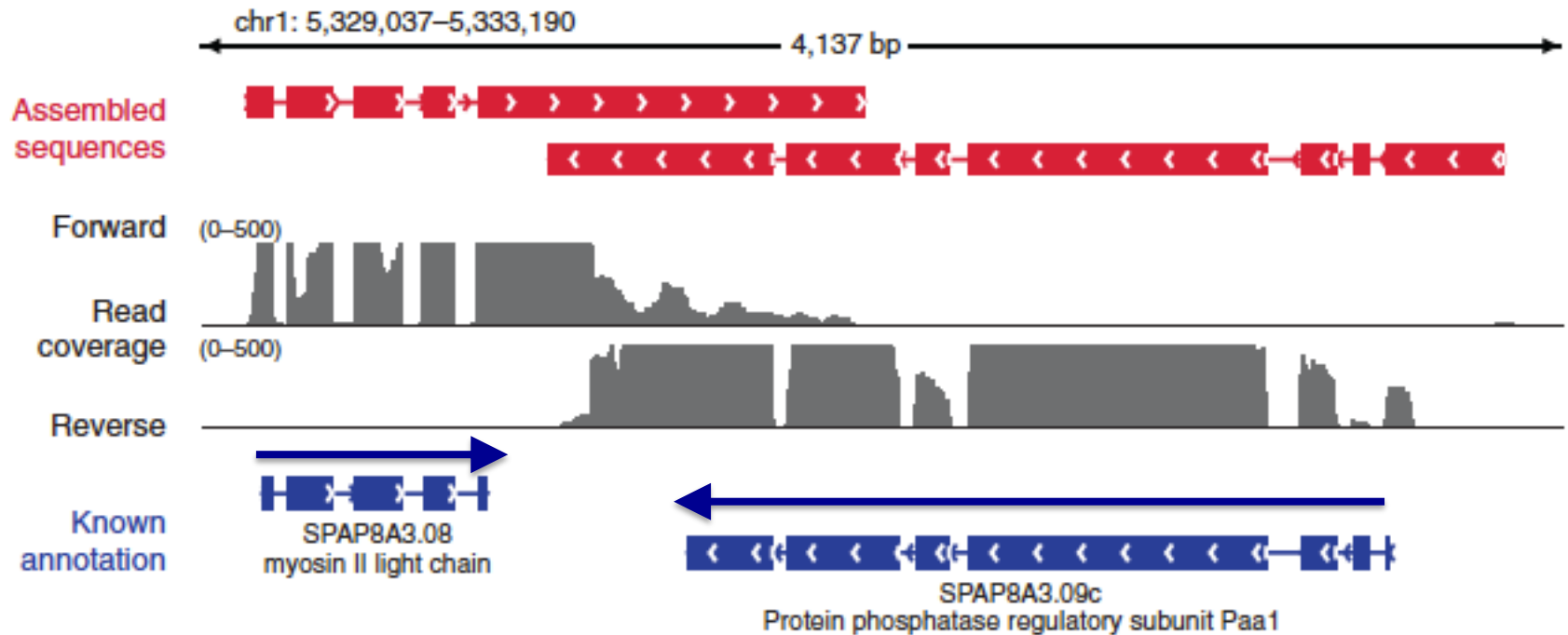


Modified from Parkhomchuk *et al.* (2009) *Nucleic Acids Res.* 37:e123

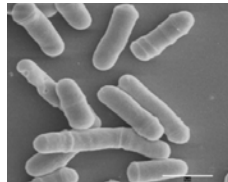
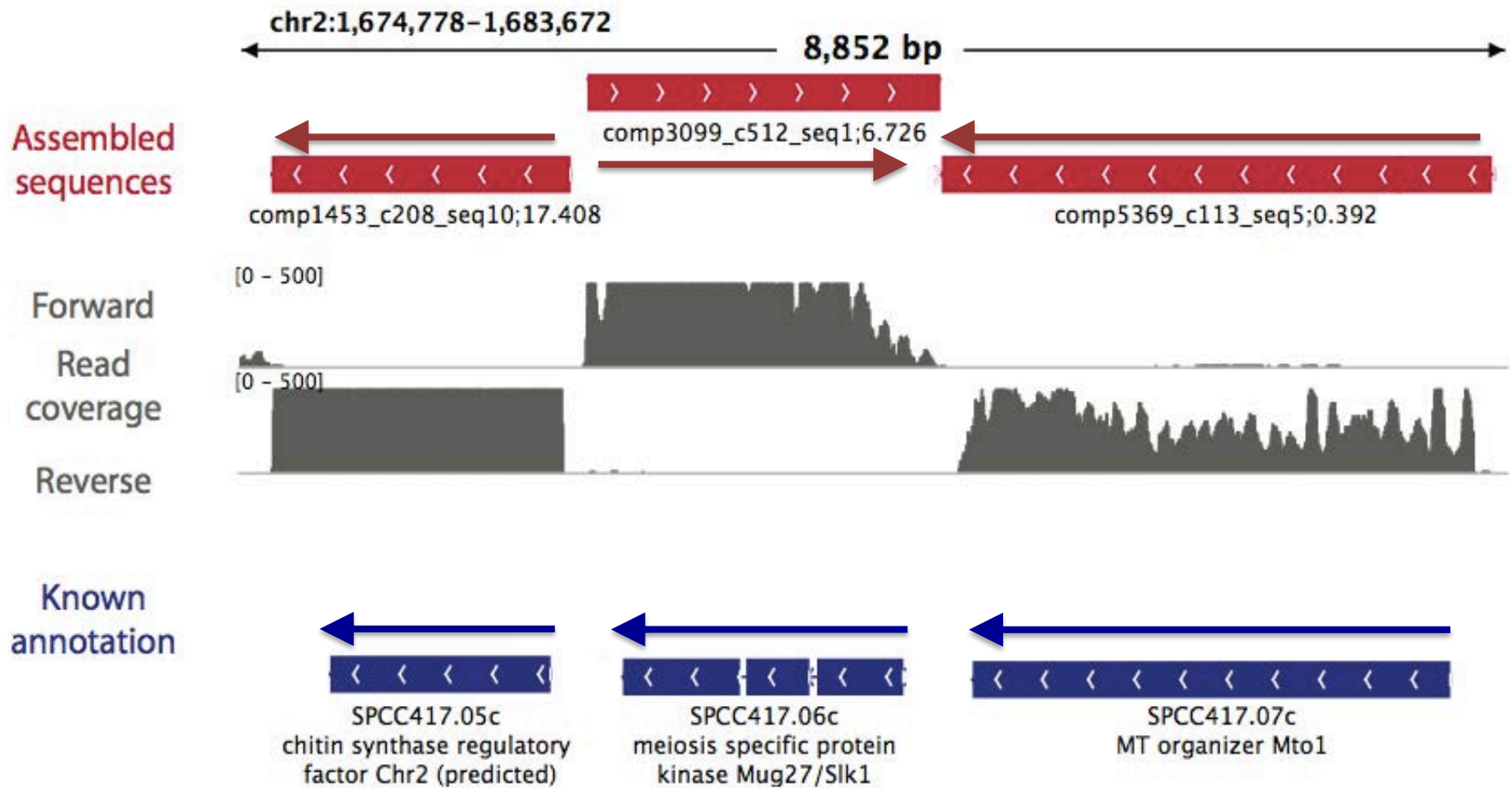
Overlapping UTRs from Opposite Strands



Schizosacharomyces pombe
(fission yeast)



Antisense-dominated Transcription



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

NATURE PROTOCOLS | PROTOCOL

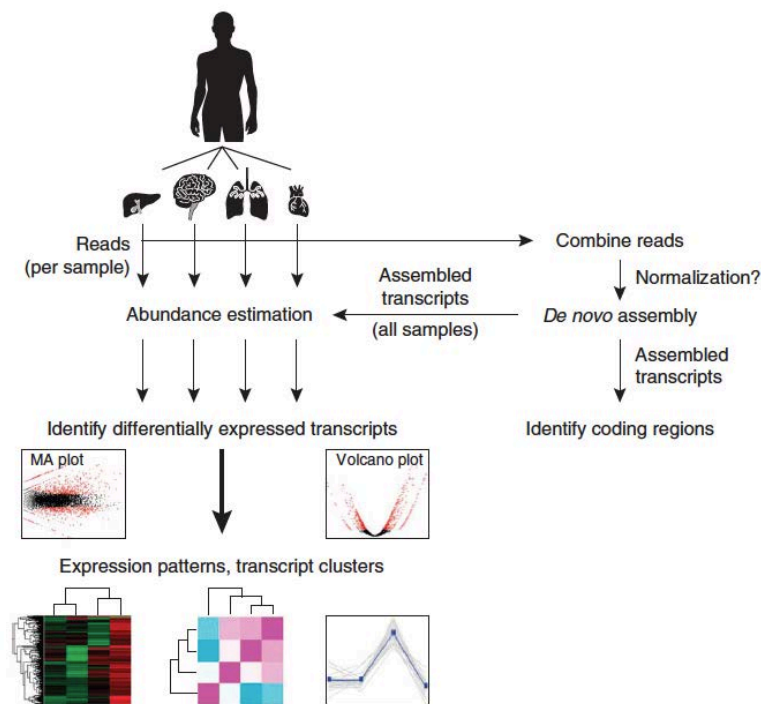
De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis

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](#)

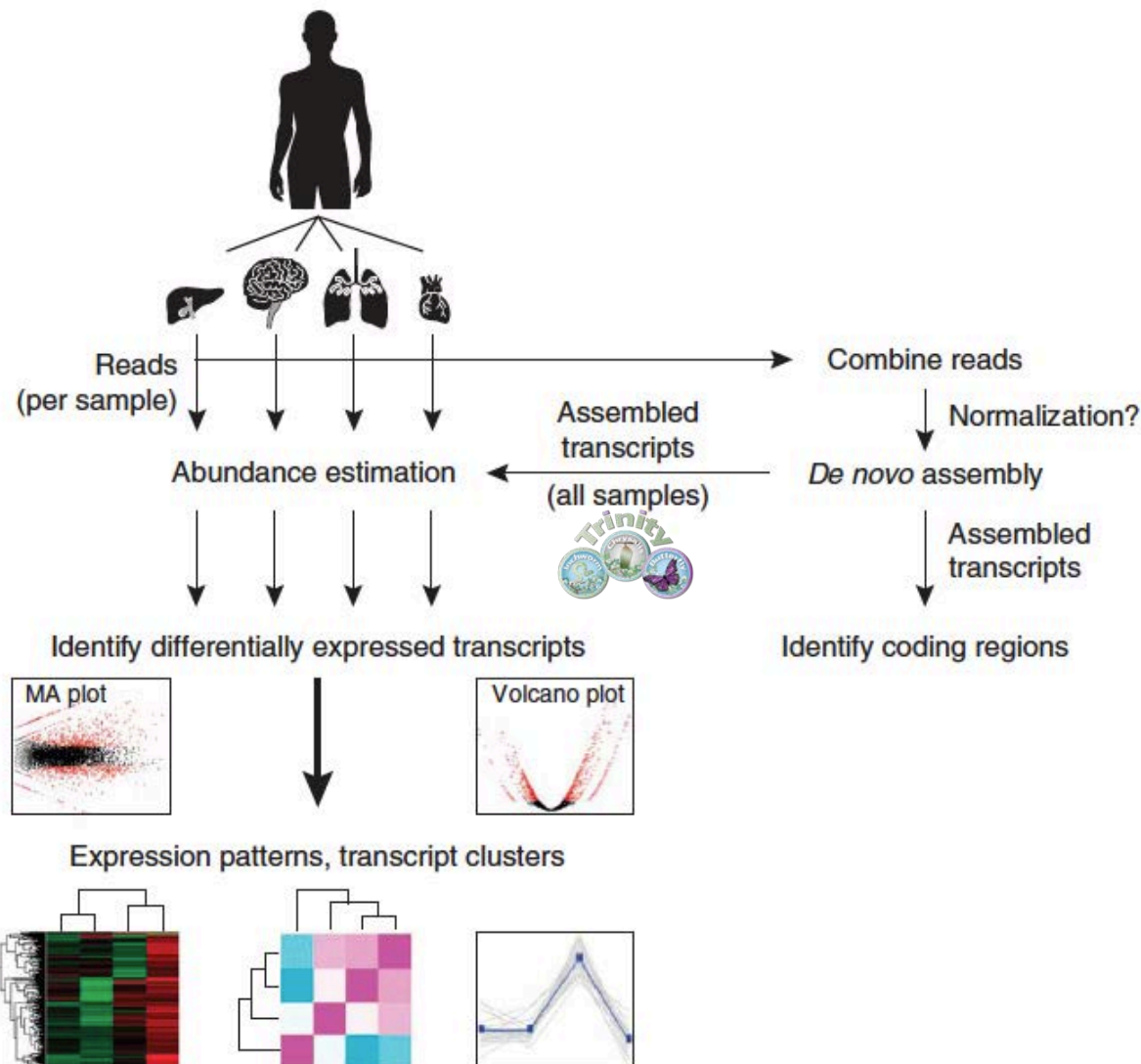
Nature Protocols **8**, 1494–1512 (2013) | doi:10.1038/nprot.2013.084

Published online 11 July 2013



Trinity Framework for De novo Transcriptome Assembly and Analysis

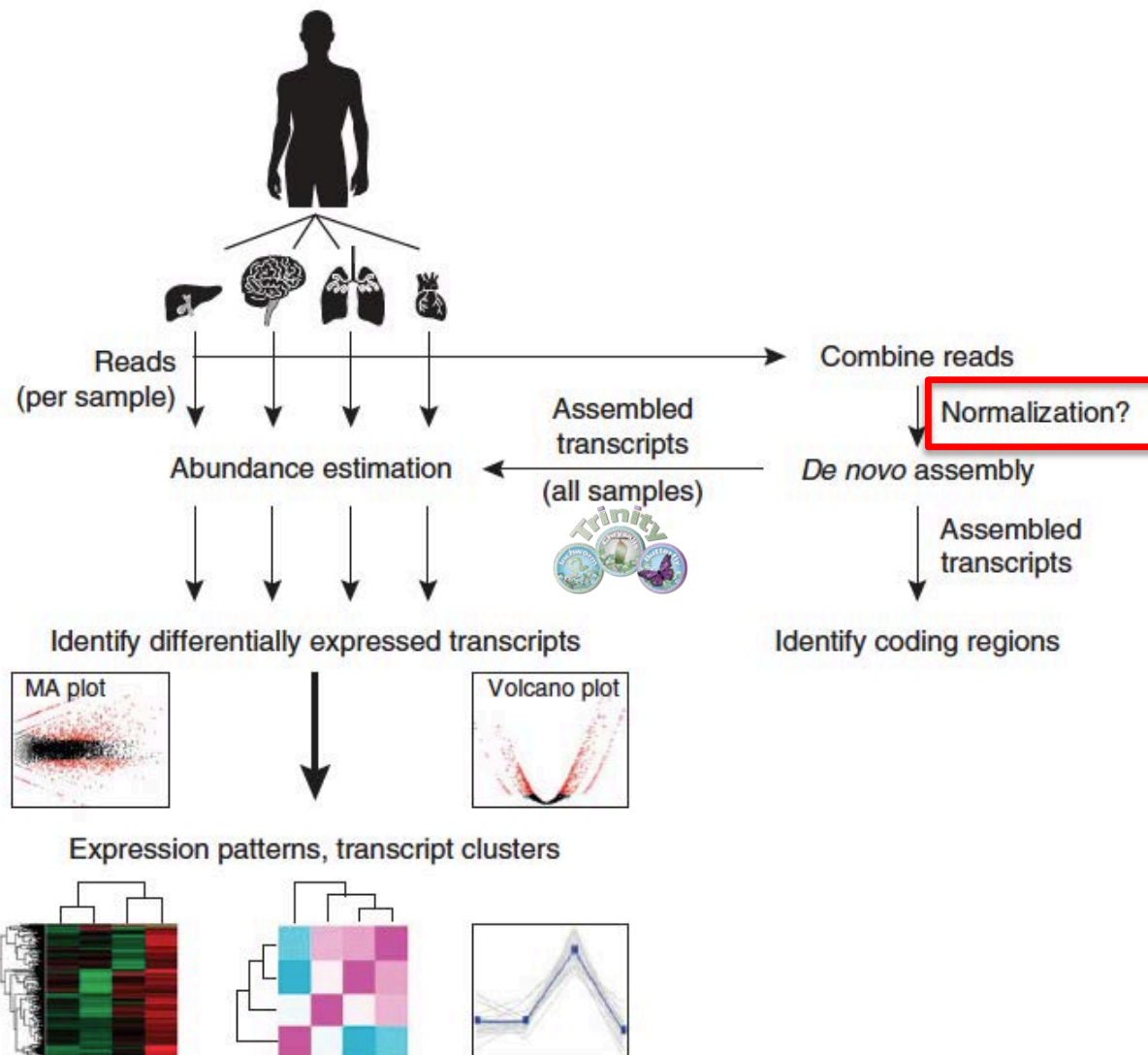
(focus of the transcriptomics lab)



Bioconductor,
& Trinity

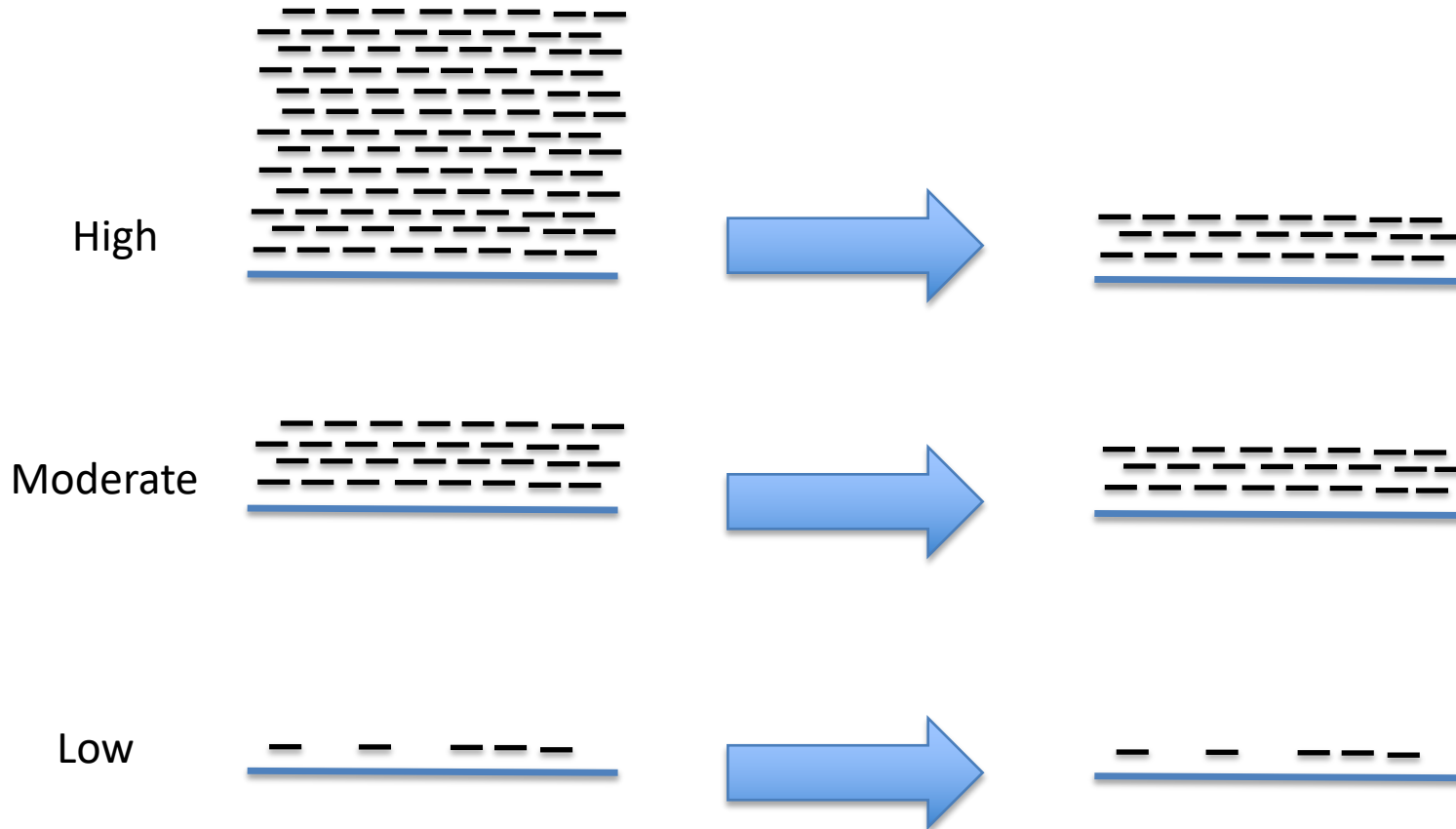
Trinity Framework for De novo Transcriptome Assembly and Analysis

(focus of the transcriptomics lab)

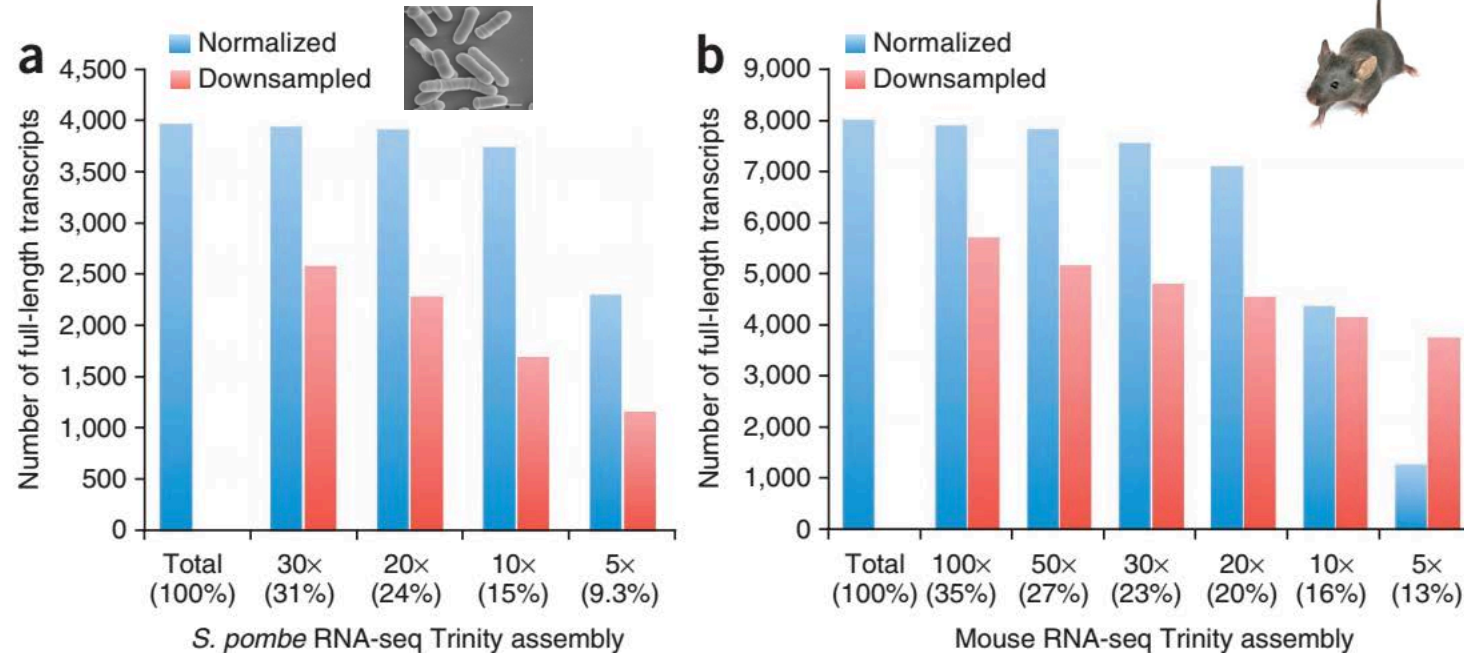


Bioconductor,
& Trinity

In silico normalization of reads

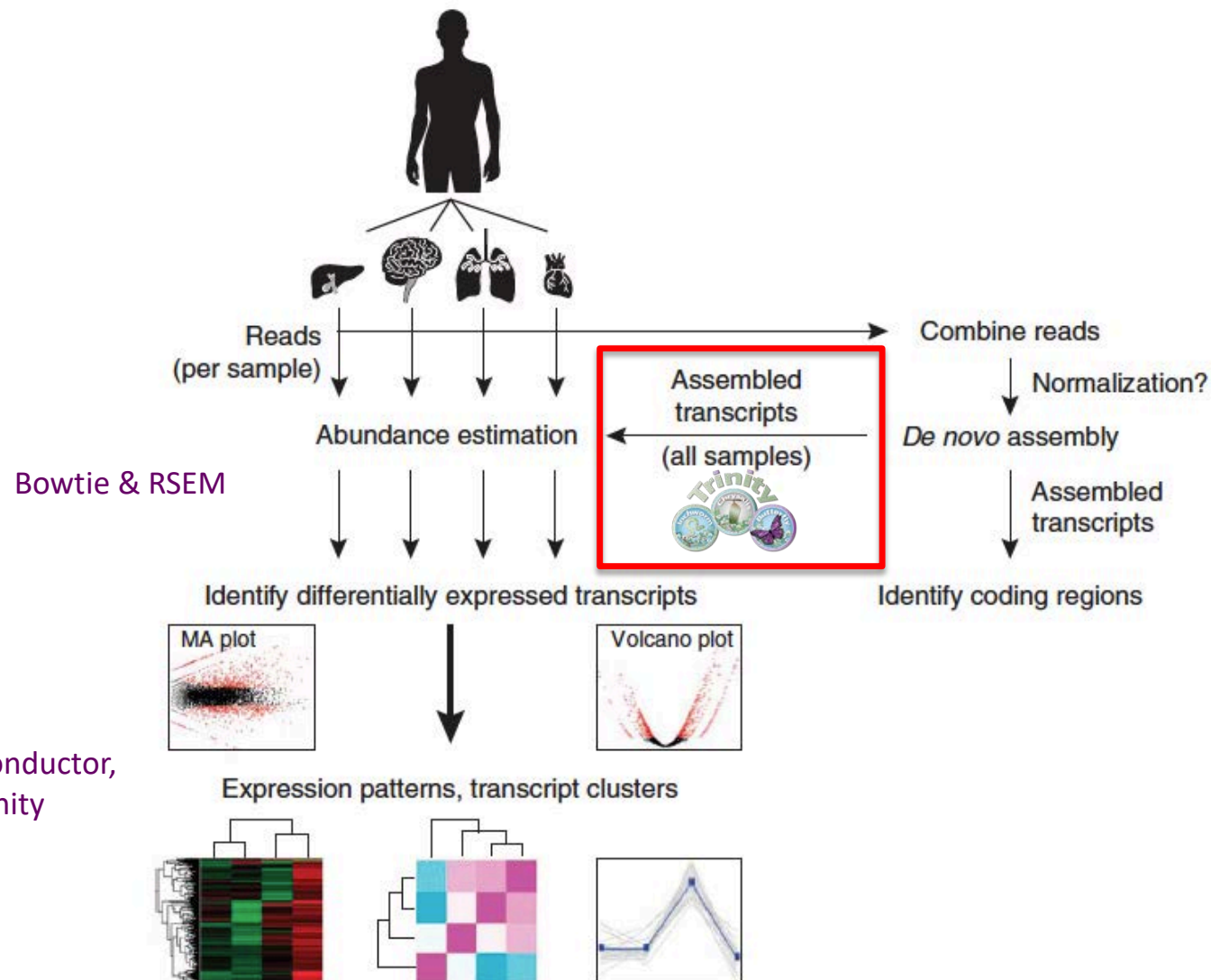


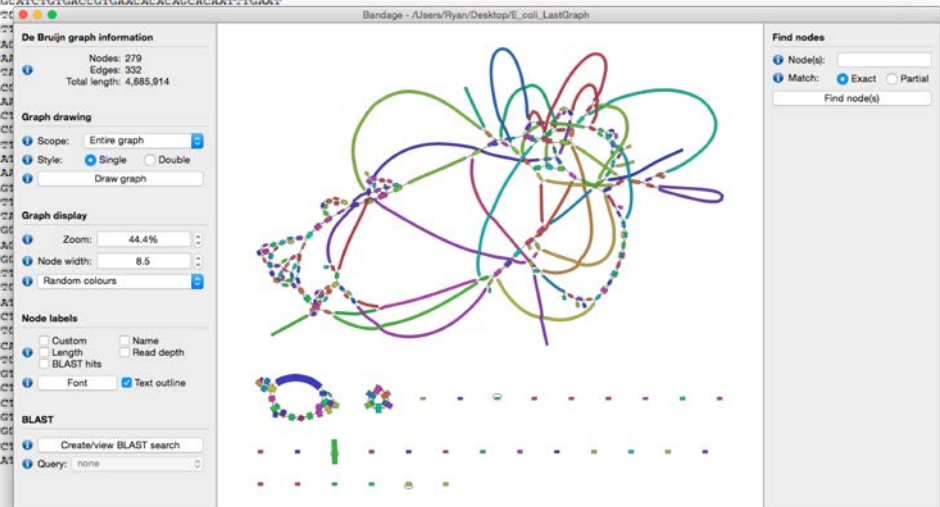
Impact of Normalization on *De novo* Full-length Transcript Reconstruction



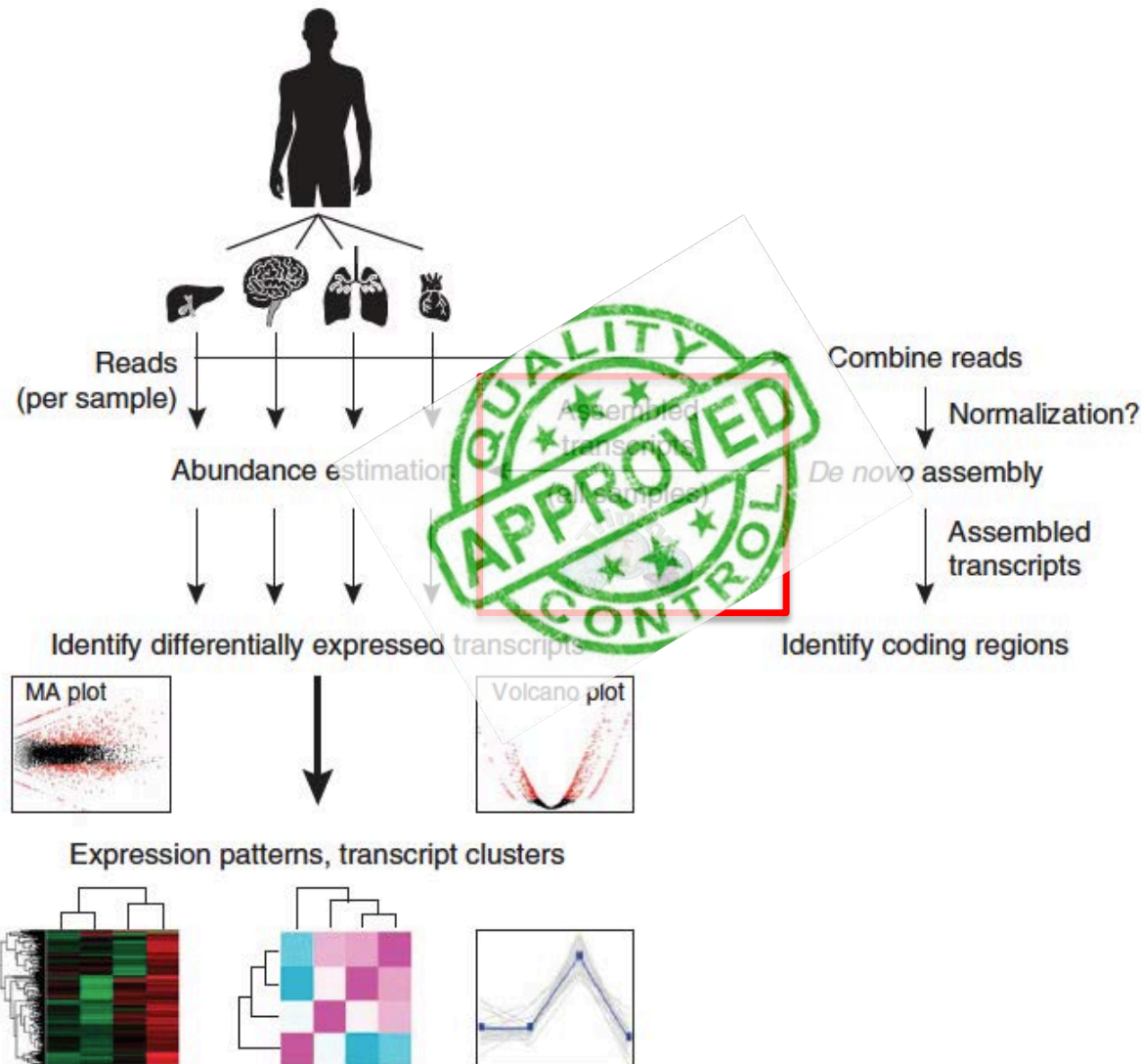
Largely retain full-length reconstruction, but use less RAM and assemble much faster.

Trinity Framework for De novo Transcriptome Assembly and Analysis



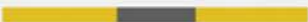


Evaluating the quality of your transcriptome assembly



Bioconductor,
& Trinity

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 



TransRate

1 input data

assembled contigs paired-end reads



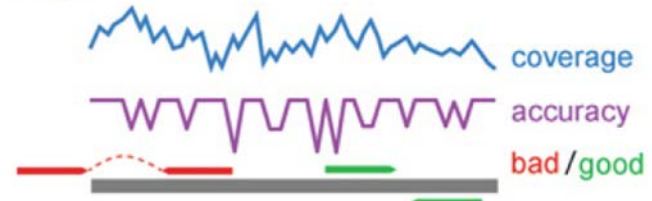
2 align reads to contigs



3 assign multimapping reads



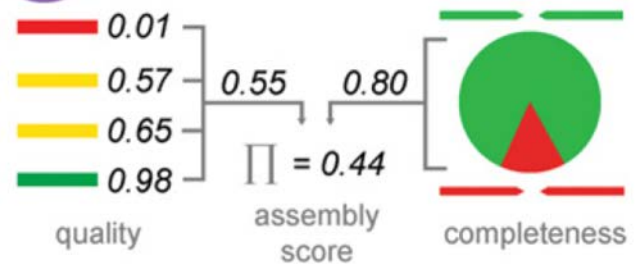
4 collect contig score components



5 calculate contig scores



6 calculate assembly score



Simple Quantitative and Qualitative Assembly Metrics

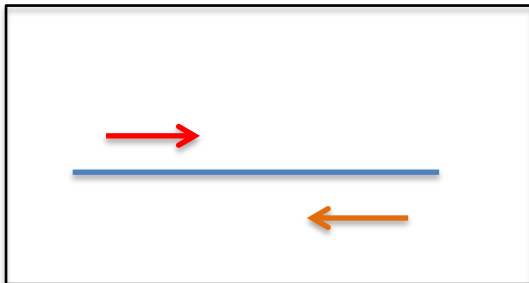
Read representation by assembly

Align reads to the assembled transcripts using Bowtie.

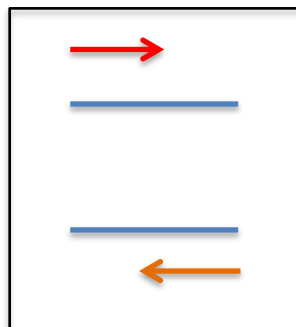
A typical 'good' assembly has ~80 % reads mapping to the assembly and ~80% are properly paired.

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

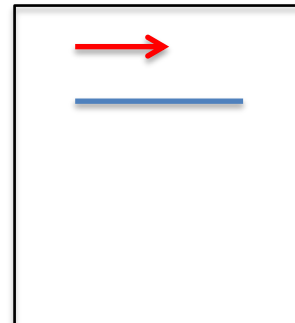
Proper pairs



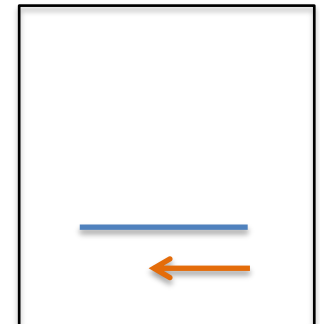
Improper pairs



Left only



Right only



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

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

[illegible]

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

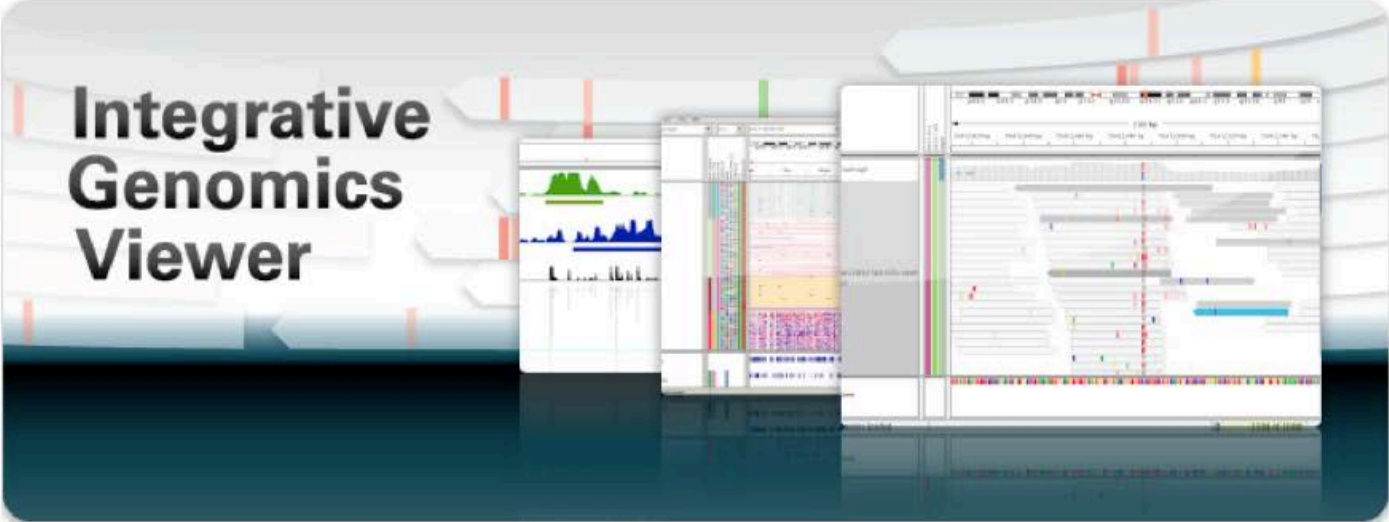
search

[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](#).

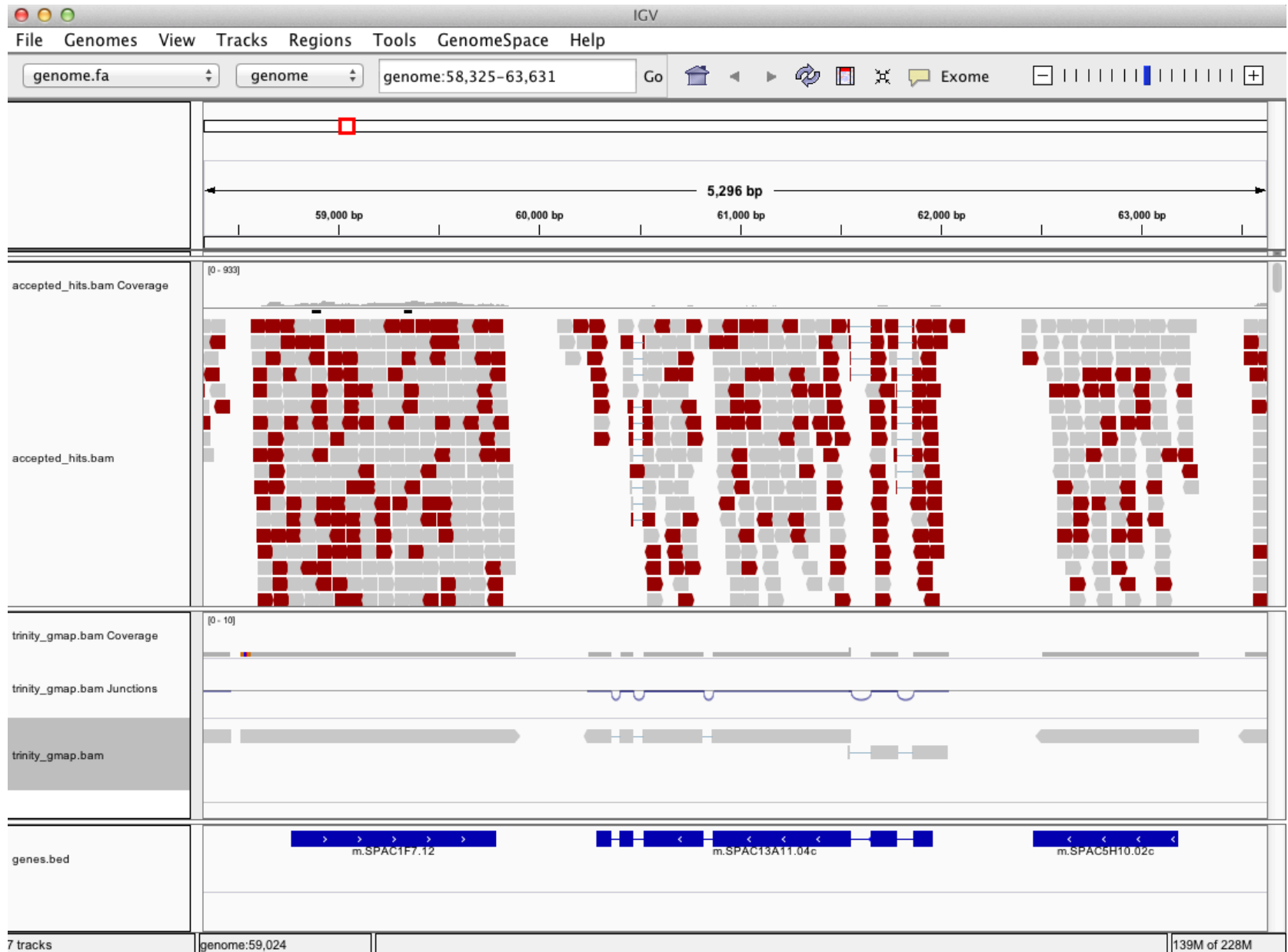
Overview

Can Examine Transcript Read Support Using IGV



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

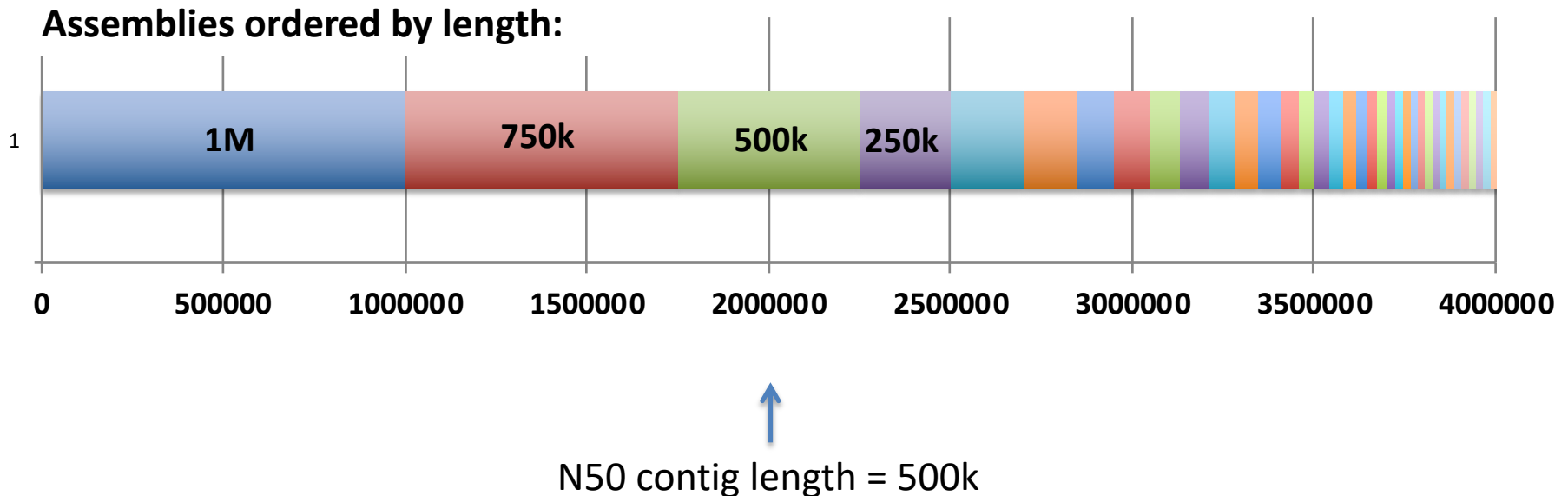
(Trinity transcripts aligned to the genome using GMAP)



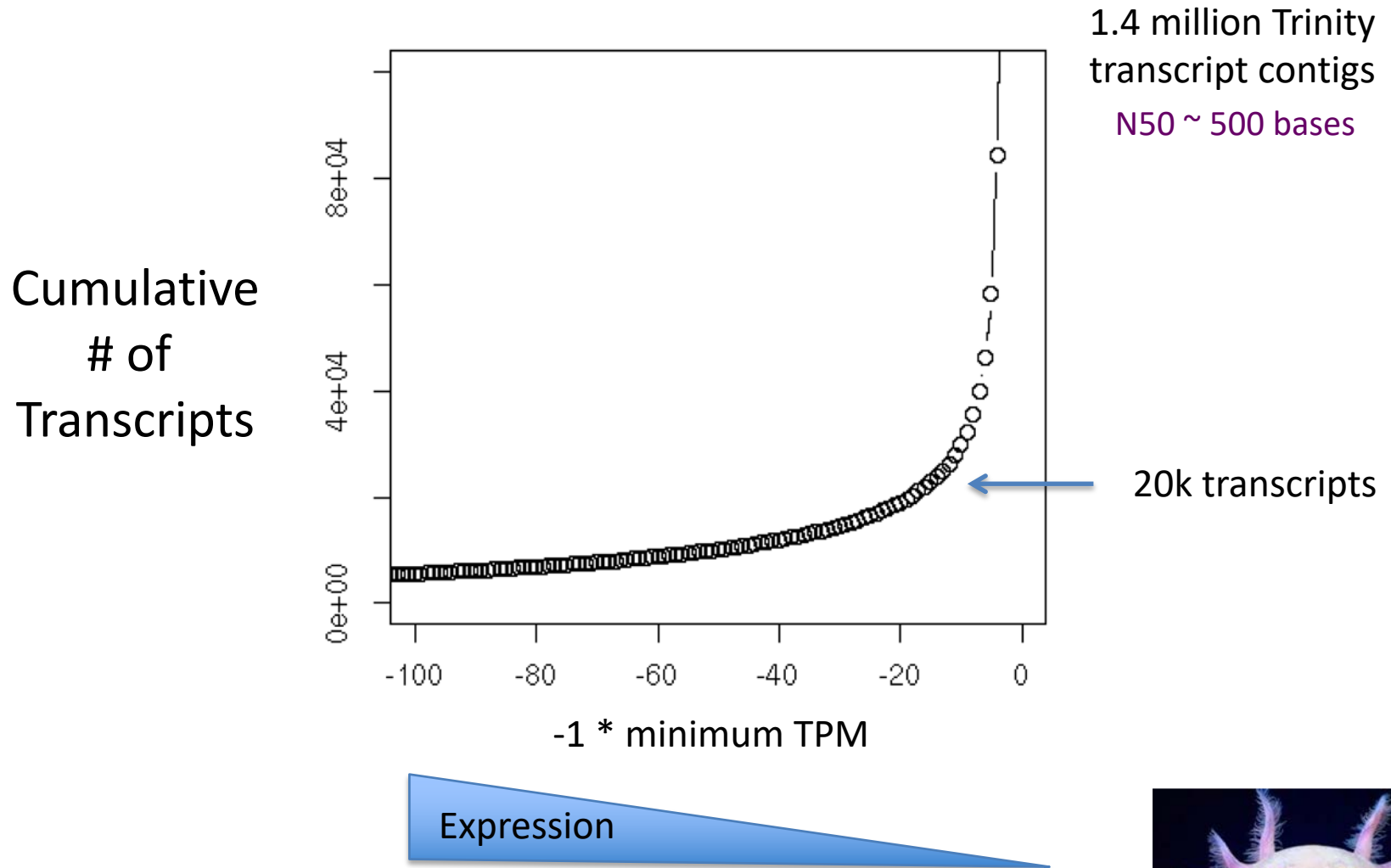
The Contig N50 statistic

“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’



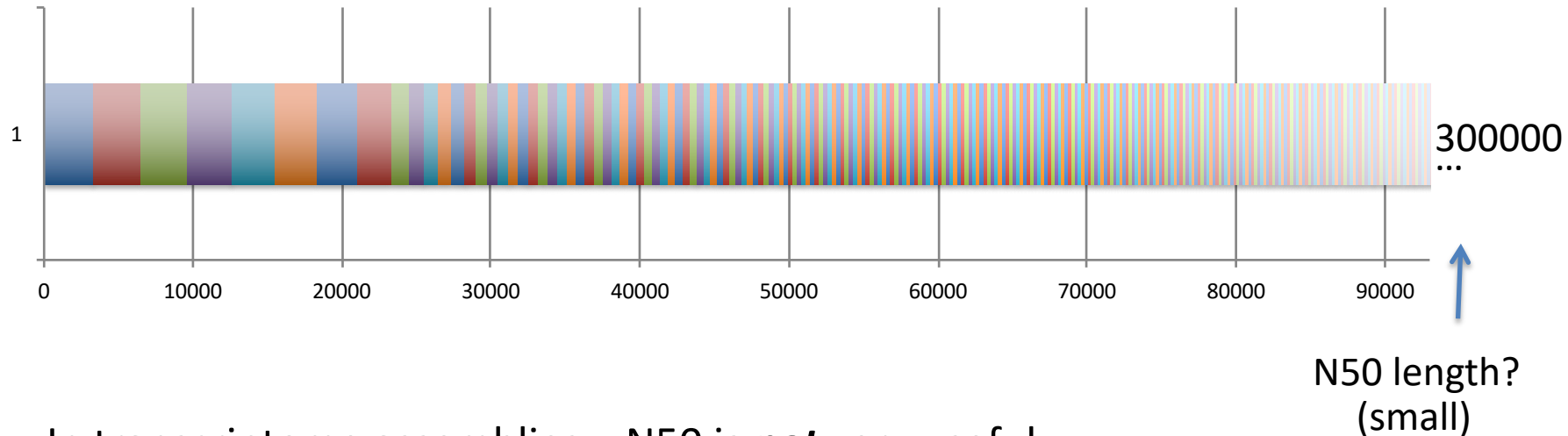
Often, most assembled transcripts are **very lowly expressed**
(How many 'transcripts & genes' are there really?)



* Salamander transcriptome



N50 Calculation for *Transcriptome* Assemblies??



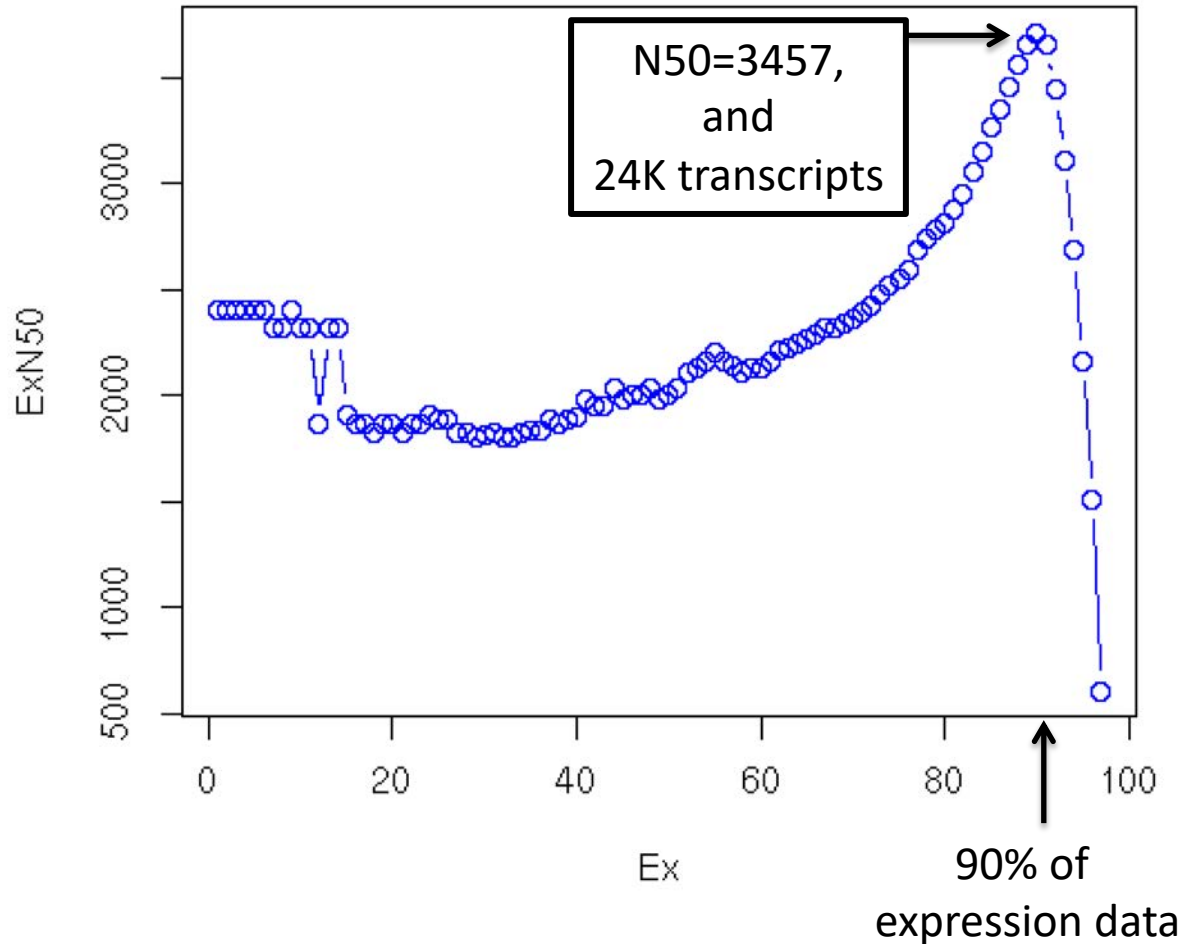
In transcriptome assemblies – N50 is ***not*** very useful.

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

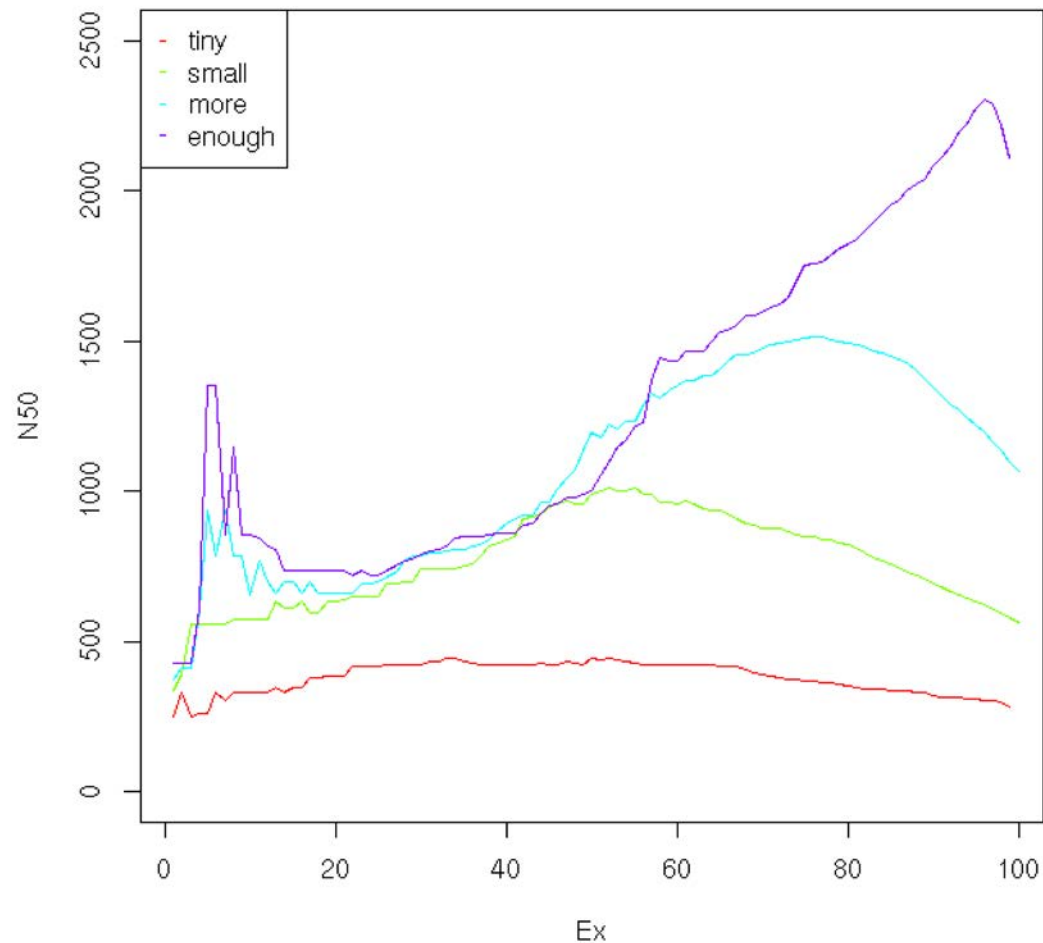
Expression-informed N50 Calculation for Transcriptome Assemblies (ExN50)

Compute N50 Based on the Top-most Highly Expressed Transcripts

- 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

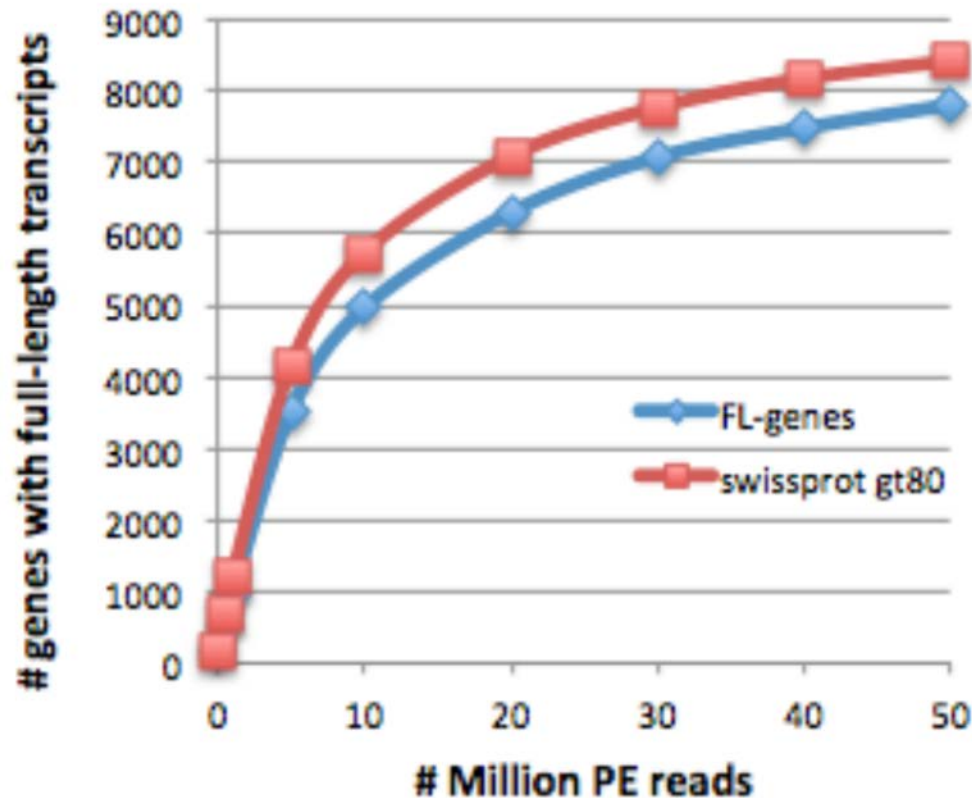
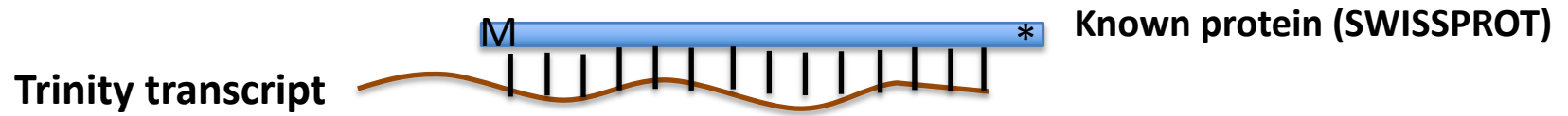


Note shift in ExN50 profiles as you assemble more and more reads.

* 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”

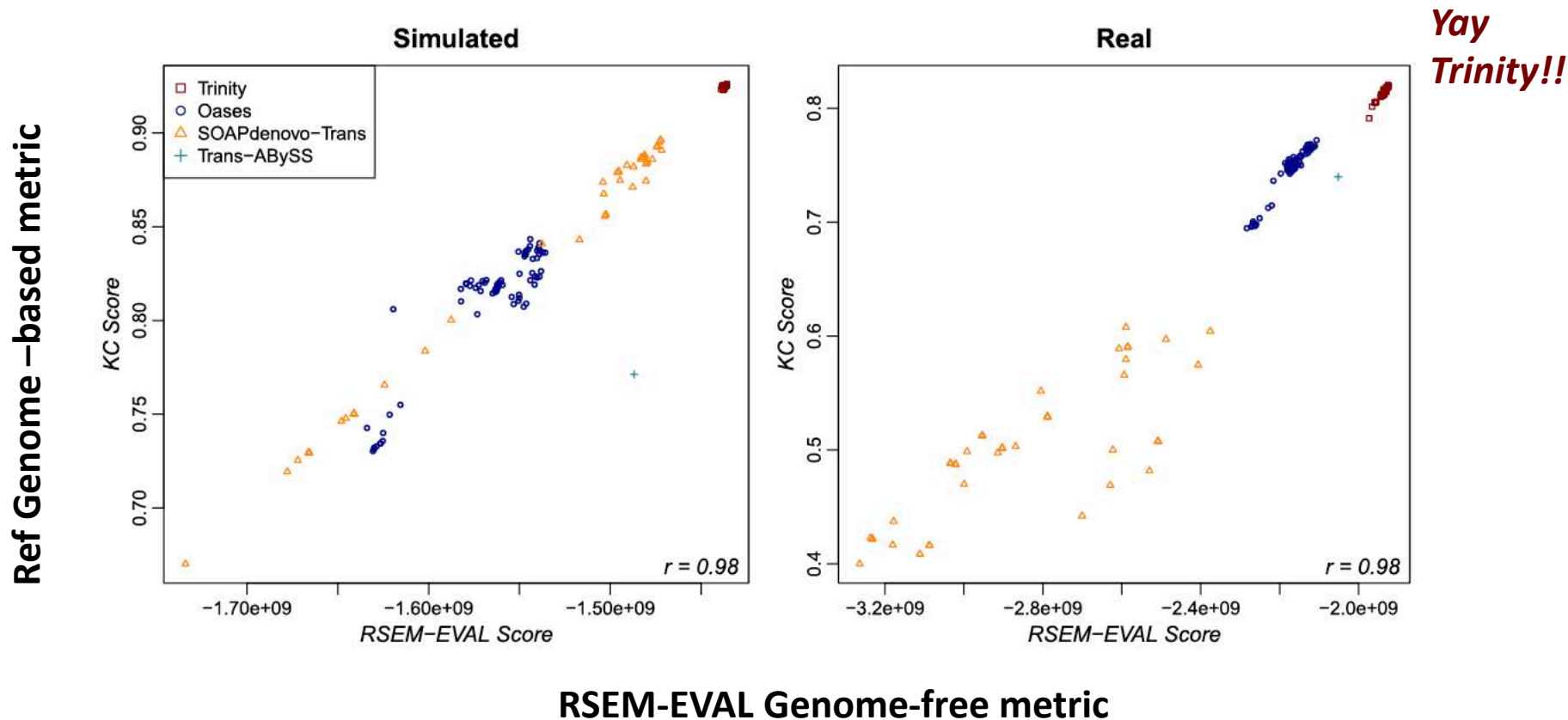
$$\begin{aligned} \log P(A, D) &= \log \int_{\Lambda} P(D|A, \Lambda) P(A|\Lambda) P(\Lambda) d\Lambda \\ &\approx \underbrace{\log P(D|A, \Lambda_{\text{MLE}})}_{\text{likelihood}} + \underbrace{\log P(A|\Lambda_{\text{MLE}})}_{\text{prior}} \end{aligned}$$

Bigger Score = Better Assembly

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

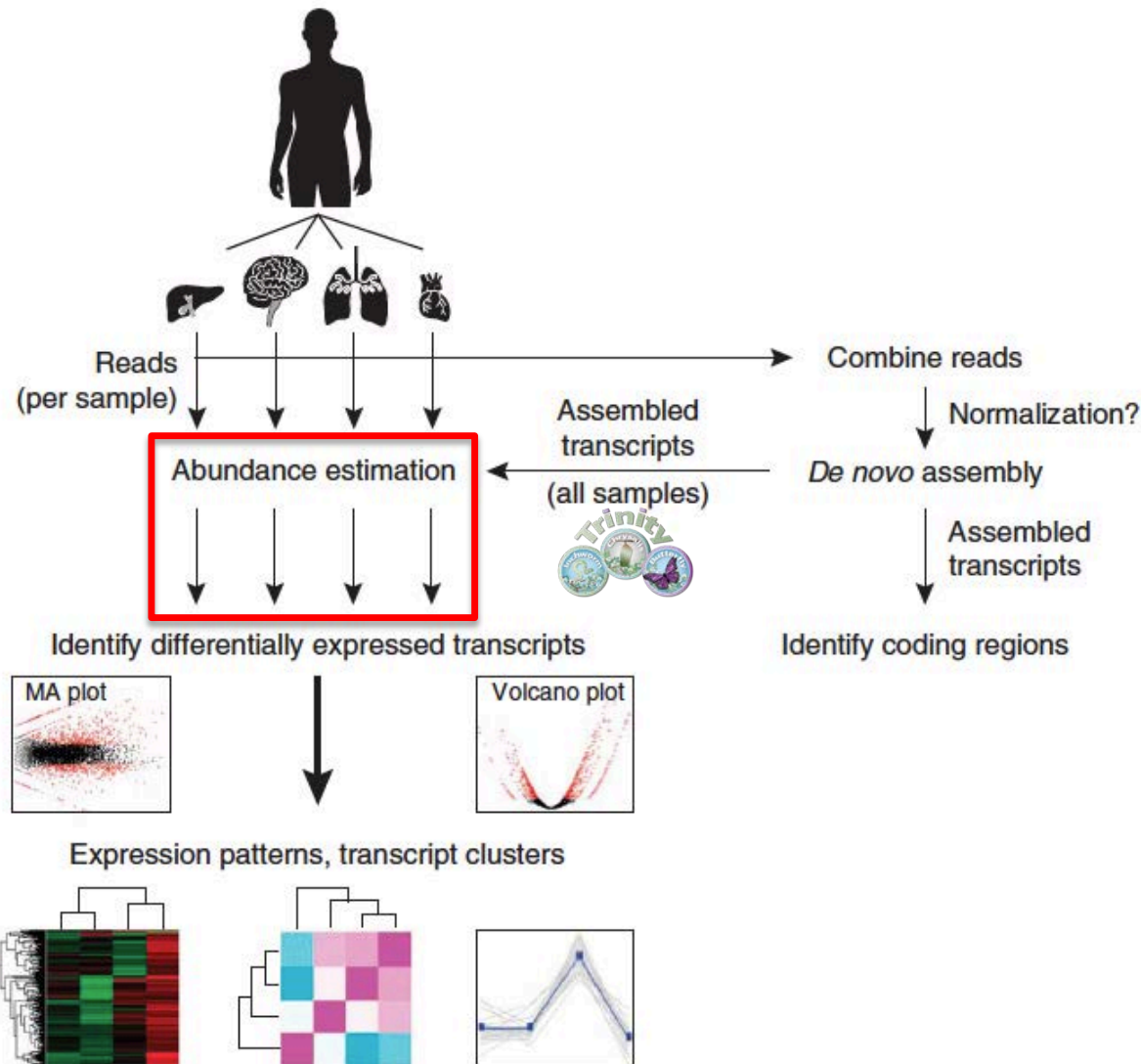
Detonate: Which assembly is better?

“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.”



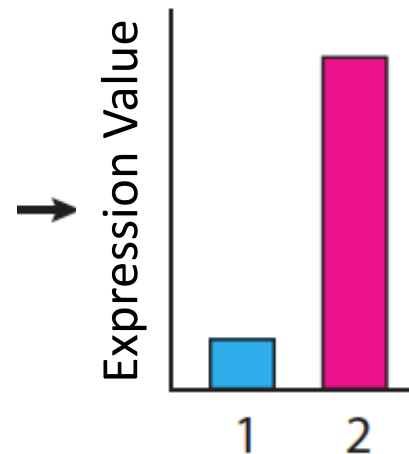
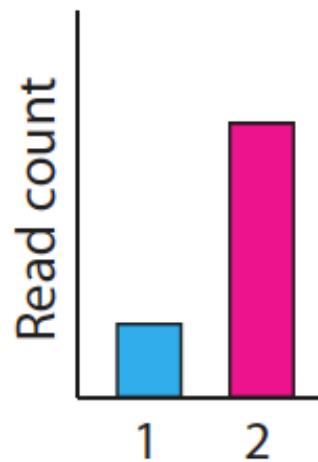
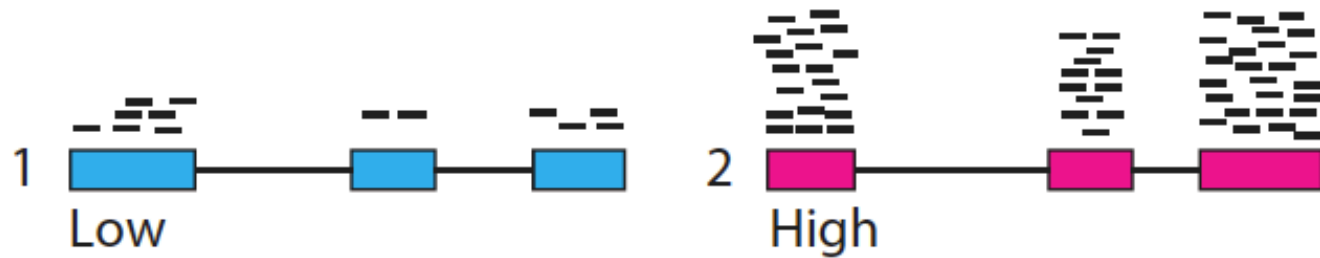
Abundance Estimation

(Aka. Computing Expression Values)

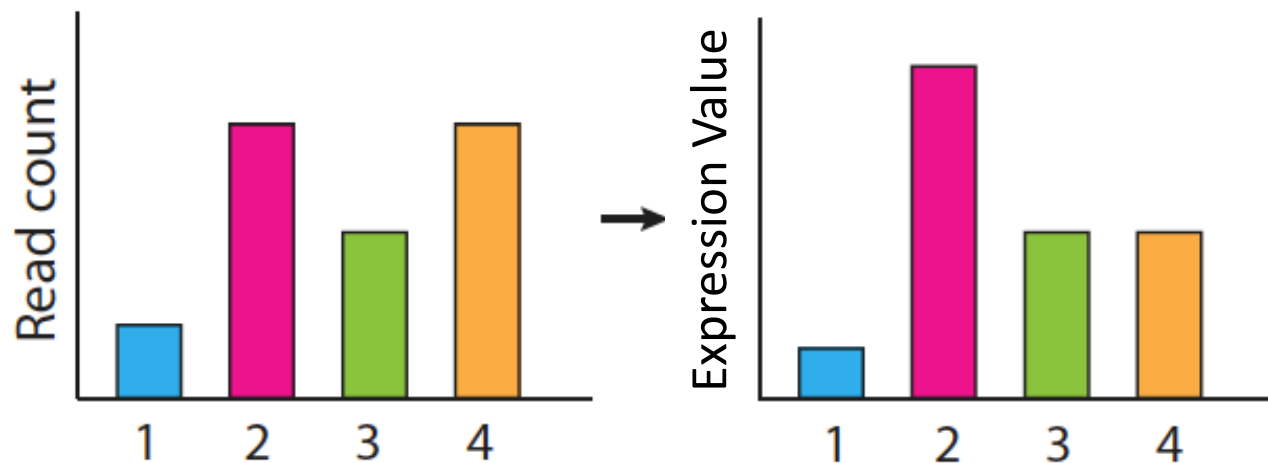
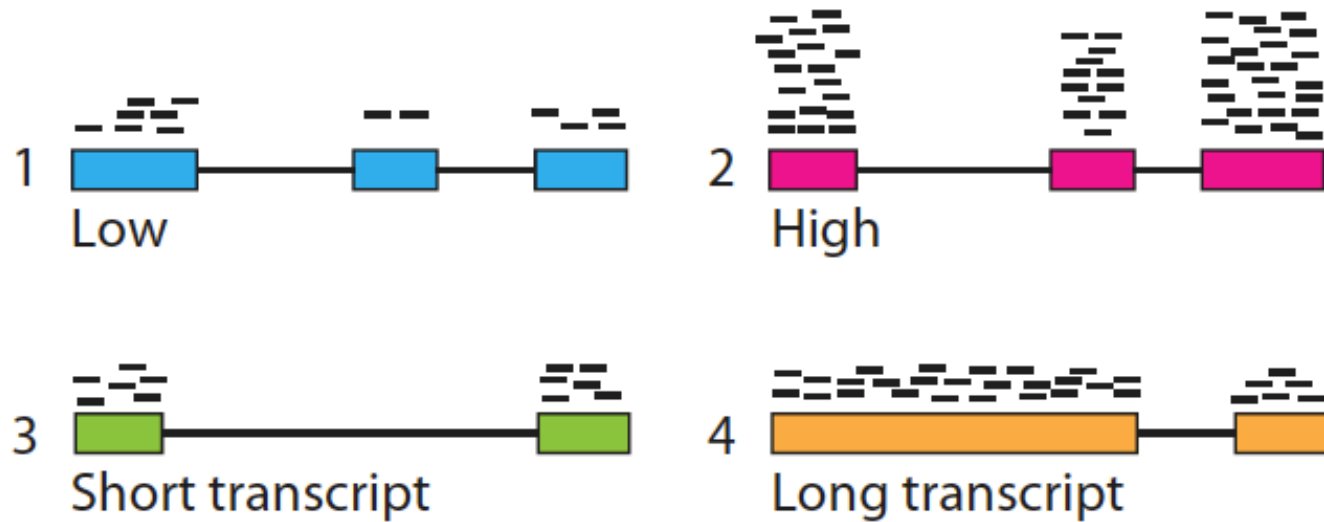


Bioconductor,
& Trinity

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)

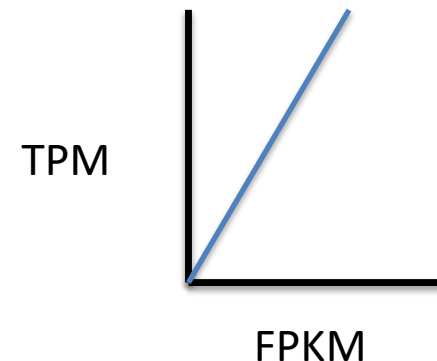
$$TPM_i = \frac{FPKM_i}{\sum_j FPKM_j} * 1e6$$

Preferred metric for measuring expression

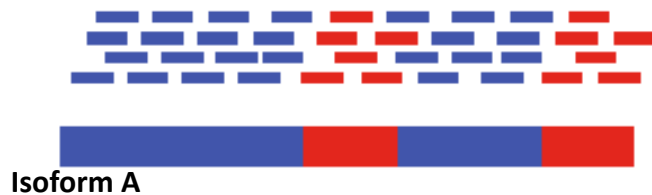
- Better reflects transcript concentration in the sample.
- Nicely sums to 1 million

Linear relationship between TPM and FPKM values.

Both are valid metrics, but best to be consistent.



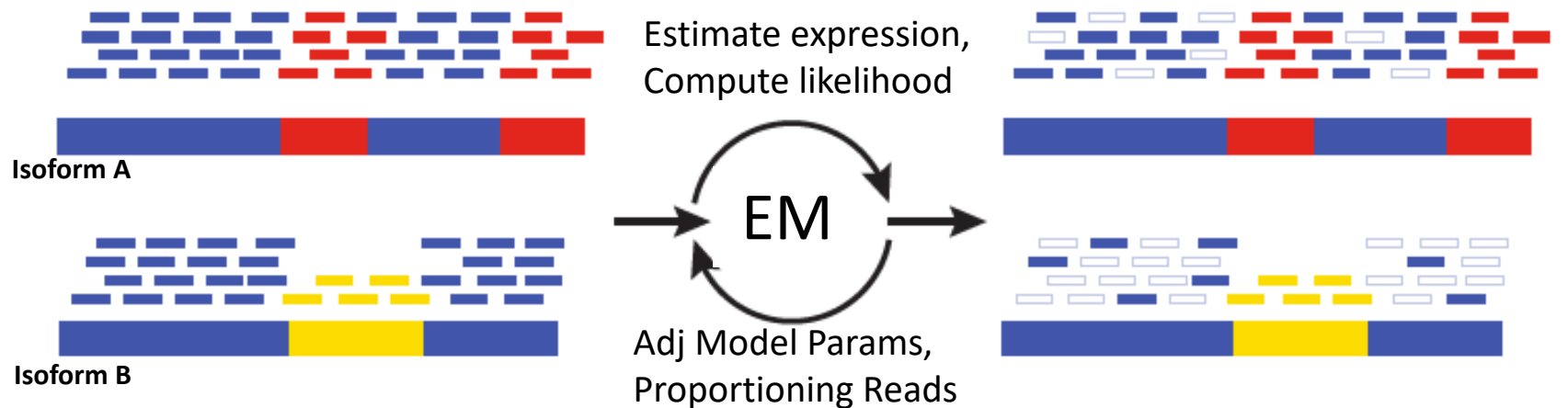
Multiply-mapped Reads Confound Abundance Estimation



Blue = multiply-mapped reads

Red, Yellow = uniquely-mapped reads

Multiply-mapped Reads Confound Abundance Estimation



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

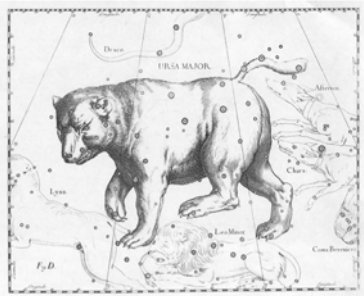
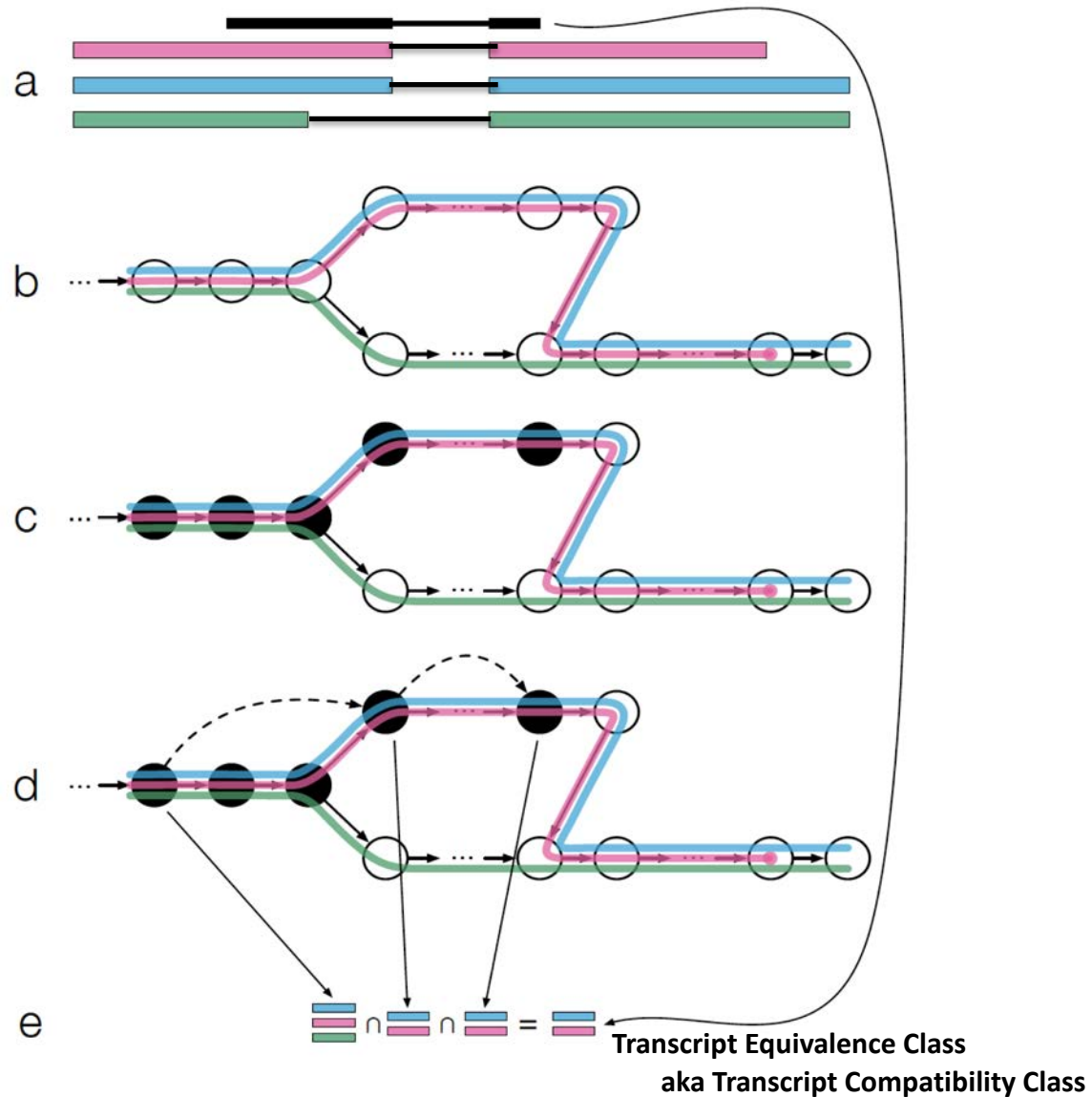
Use Expectation Maximization (EM) to find the most likely assignment of reads to transcripts.

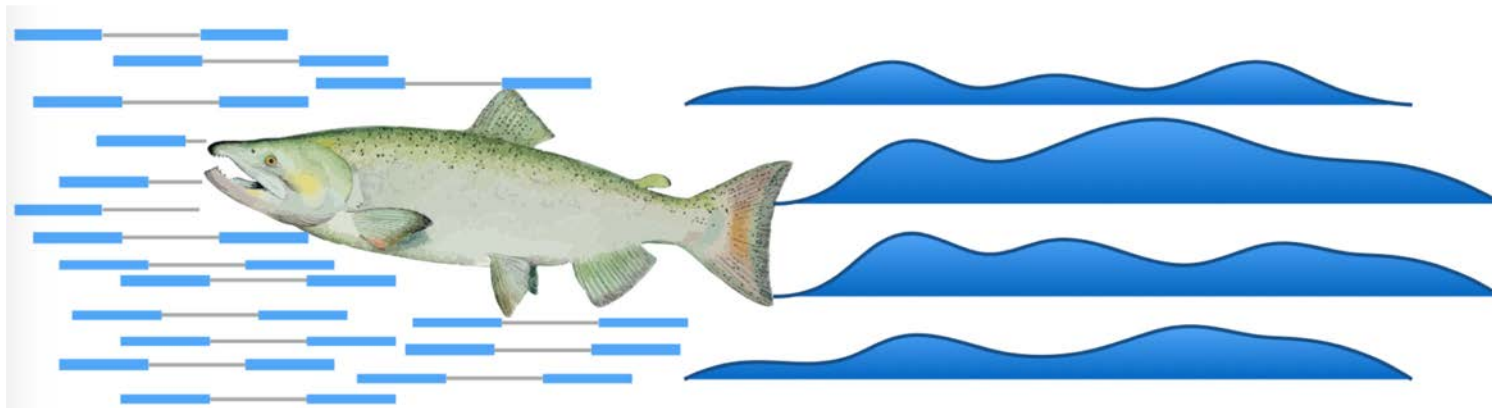
Performed by:

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

Fast Abundance Estimation Using Pseudo-alignments and Equivalence Classes

(Kallisto software, Bray et al., NBT 2016)





Salmon —Don't count . . . quantify!

Uses a suffix array
instead of the
de Bruijn graph

 nature|methods

 Altmetric: 210 Citations: 42 [More detail >>](#)

Brief Communication

Salmon provides fast and bias-aware quantification of transcript expression

Rob Patro , Geet Duggal, Michael I Love, Rafael A Irizarry & Carl Kingsford 

Nature Methods **14**, 417–419 (2017)
doi:10.1038/nmeth.4197
[Download Citation](#)

Received: 29 August 2016
Accepted: 22 January 2017
Published online: 06 March 2017

<https://combine-lab.github.io/salmon/>

Comparing RNA-Seq Samples

Some Cross-sample Normalization May Be Required

eg.

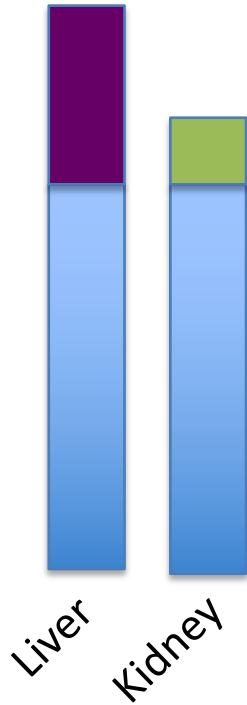


Vs.



Why cross-sample normalization is important

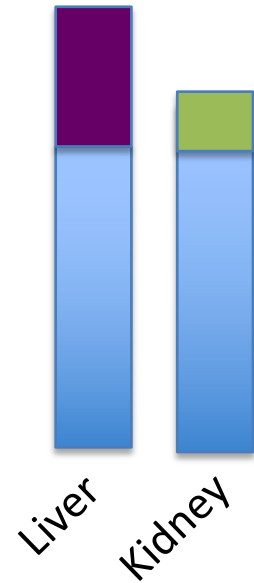
Absolute RNA quantities per cell



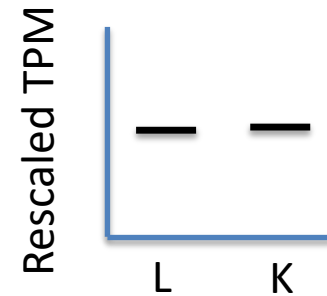
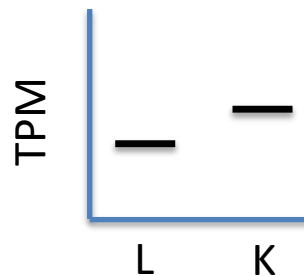
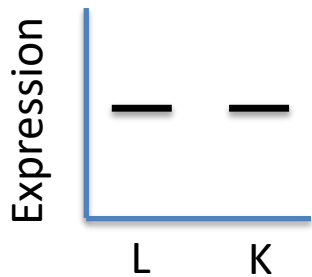
Measured relative abundance via RNA-Seq



Cross-sample normalized (rescaled) relative abundance



eg. Some housekeeping gene's expression level:



Cross-sample Normalization Required

Otherwise, housekeeping genes look diff expressed due to sample composition differences

Subset of genes highly expressed in liver

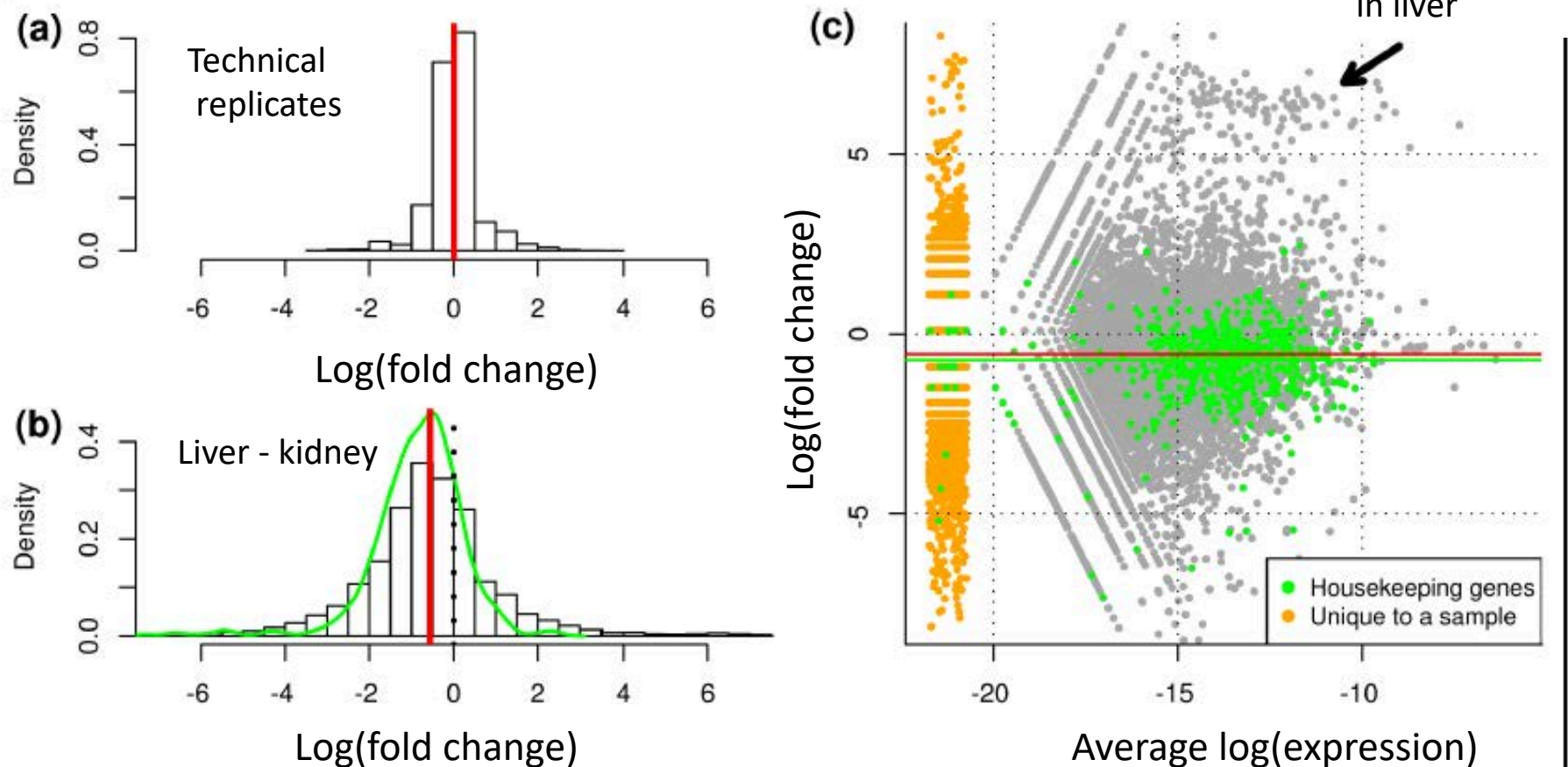
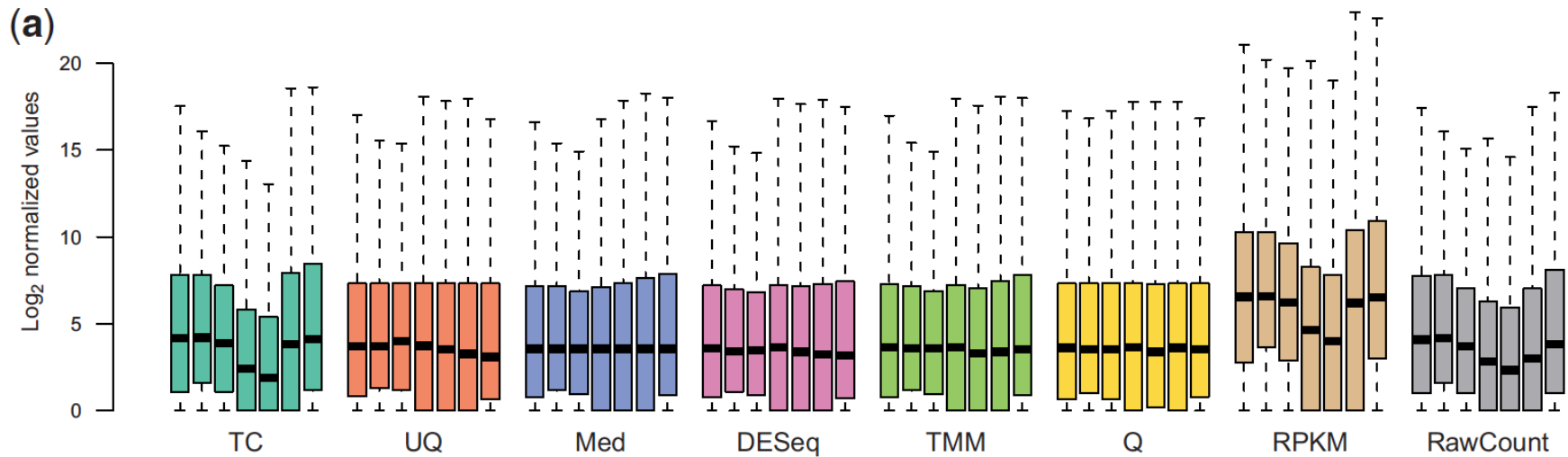


Figure 1 Normalization is required for RNA-seq data. Data from [6] comparing log ratios of (a) technical replicates and (b) liver versus kidney expression levels, after adjusting for the total number of reads in each sample. The green line shows the smoothed distribution of log-fold-changes of the housekeeping genes. (c) An M versus A plot comparing liver and kidney shows a clear offset from zero. Green points indicate 545 housekeeping genes, while the green line signifies the median log-ratio of the housekeeping genes. The red line shows the estimated TMM normalization factor. The smear of orange points highlights the genes that were observed in only one of the liver or kidney the overall bias in log-fold-changes.

Normalization methods for Illumina high-throughput RNA sequencing data analysis.



From “A comprehensive evaluation of normalization methods for Illumina high throughput RNA sequencing data analysis” Brief Bioinform. 2013 Nov;14(6):671-83

<http://www.ncbi.nlm.nih.gov/pubmed/22988256>

Differential Expression Analysis



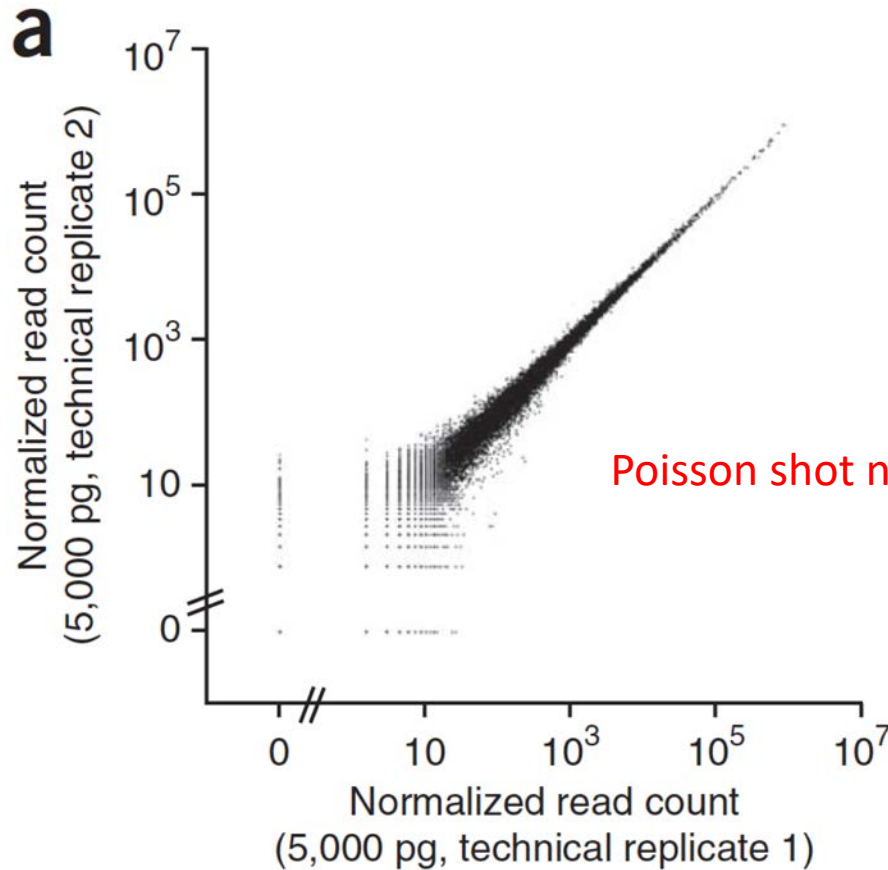
Differential Expression Analysis Involves

- Counting reads mapped to features
- Statistical significance testing

Beware of small counts leading to notable fold changes

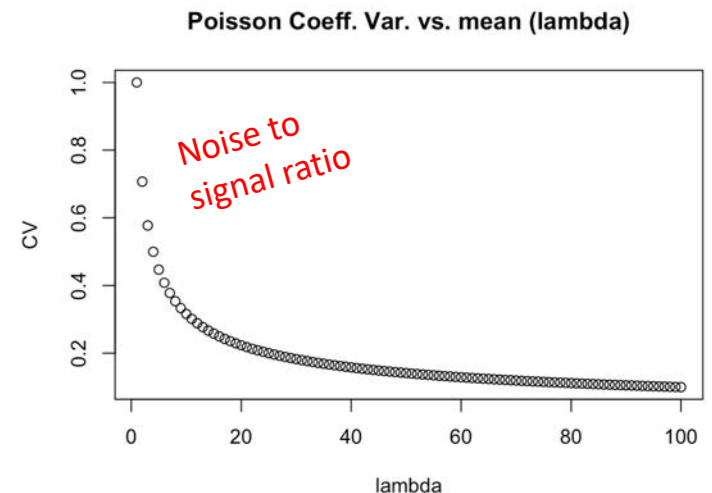
	Sample_A	Sample_B	Fold_Change	Significant?
Gene A	1	2	2-fold	No
Gene B	100	200	2-fold	Yes

Variation Observed Between Technical Replicates



Poisson shot noise is high for small counts.

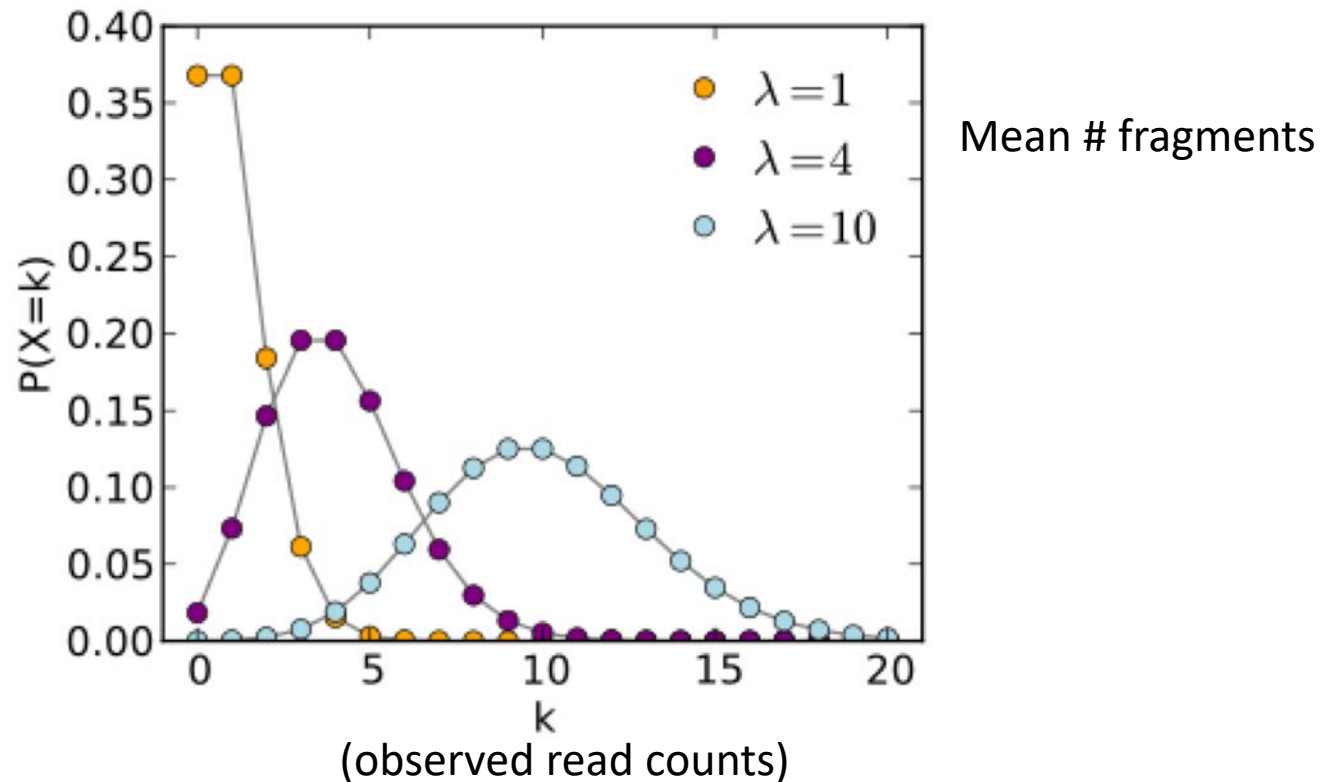
Variation observed is well described by models of random sampling (Poisson Distribution)



* plot from Brennecke, et al. Nature Methods, 2013

Observed RNA-Seq Counts Result from Random Sampling of the Population of Reads

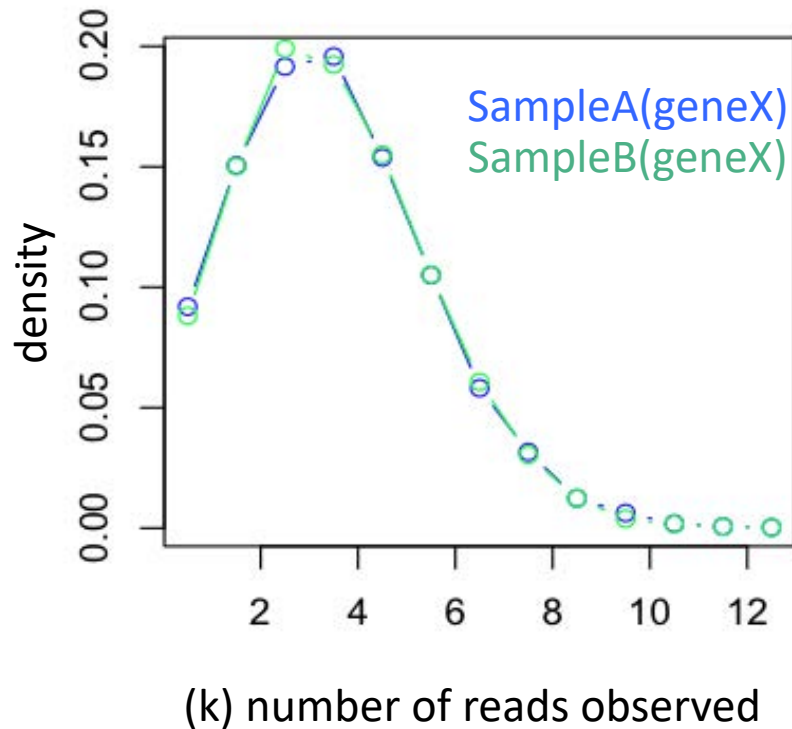
Technical variation in RNA-Seq counts per feature is well modeled by the Poisson distribution



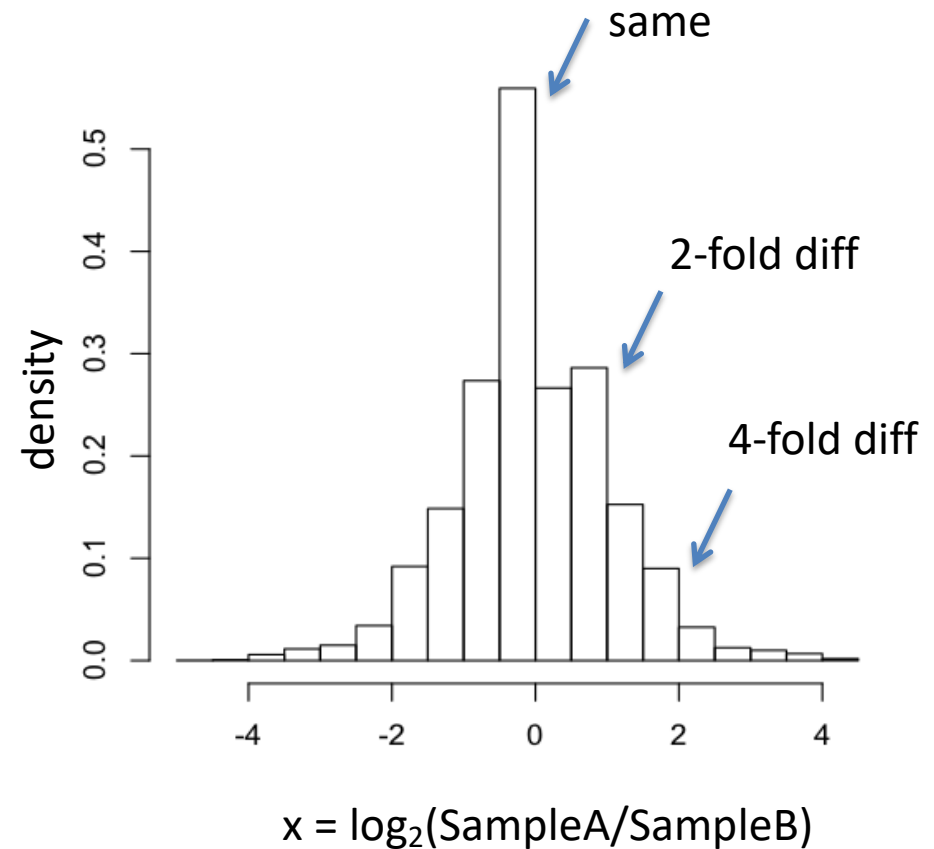
Example: One gene*not* differentially expressed

Example: $\text{SampleA}(\text{gene}) = \text{SampleB}(\text{gene}) = 4$ reads

Distribution of observed counts for single gene
(under Poisson model)

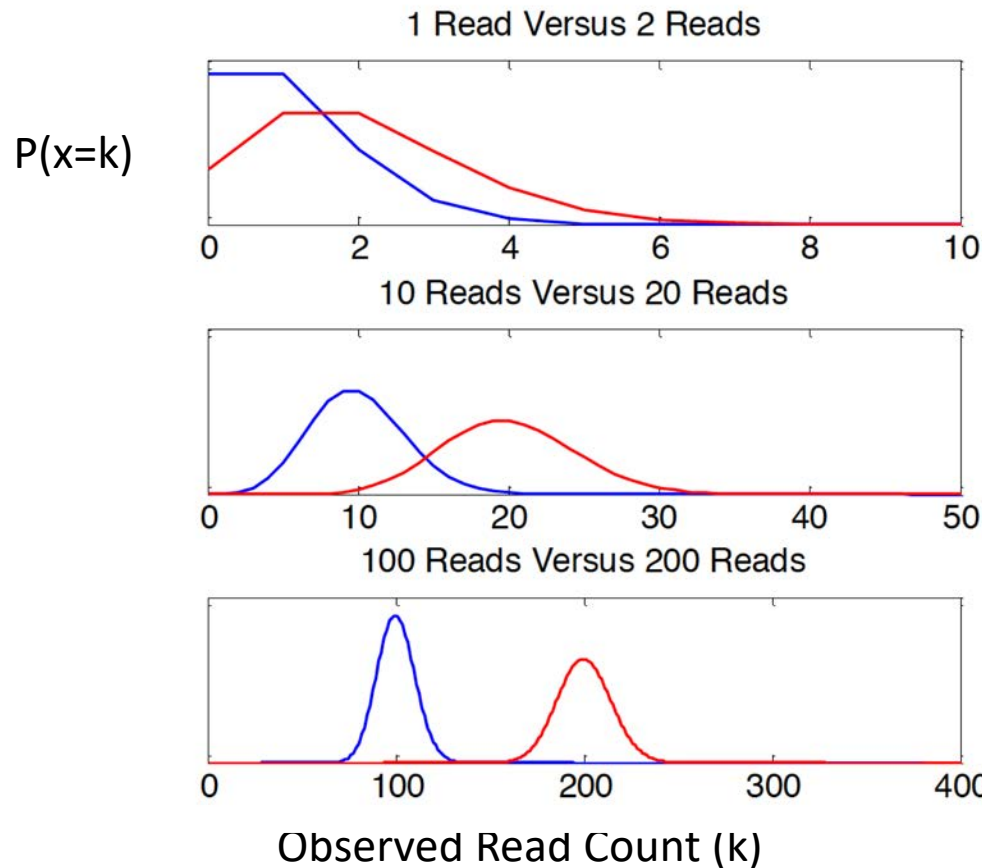


Dist. of $\log_2(\text{fold change})$ values



Sequencing Depth Matters

Poisson distributions for counts based on **2-fold** expression differences



No confidence in 2-fold difference. Likely observed by chance.

High confidence in 2-fold difference. Unlikely observed by chance.

Greater Depth = More Statistical Power

Example: Single gene, reads sampled at different sequencing depths

Reads per sample	Sample A Number of reads	Sample B Number of reads	P-value (Fishers Exact Test)
100,000	1	2	1
1,000,000	10	20	0.099
10,000,000	100	200	8.0e-09

Technical vs. Biological Replicates

RNA-Seq Technical replicates aren't essential

(Technical variation is well-modeled by the Poisson distribution)

“We find that the Illumina sequencing data are highly replicable, with relatively little technical variation, and thus, for many purposes, it may suffice **to sequence each mRNA sample only once**” *Marioni et al., Genome Research, 2008*

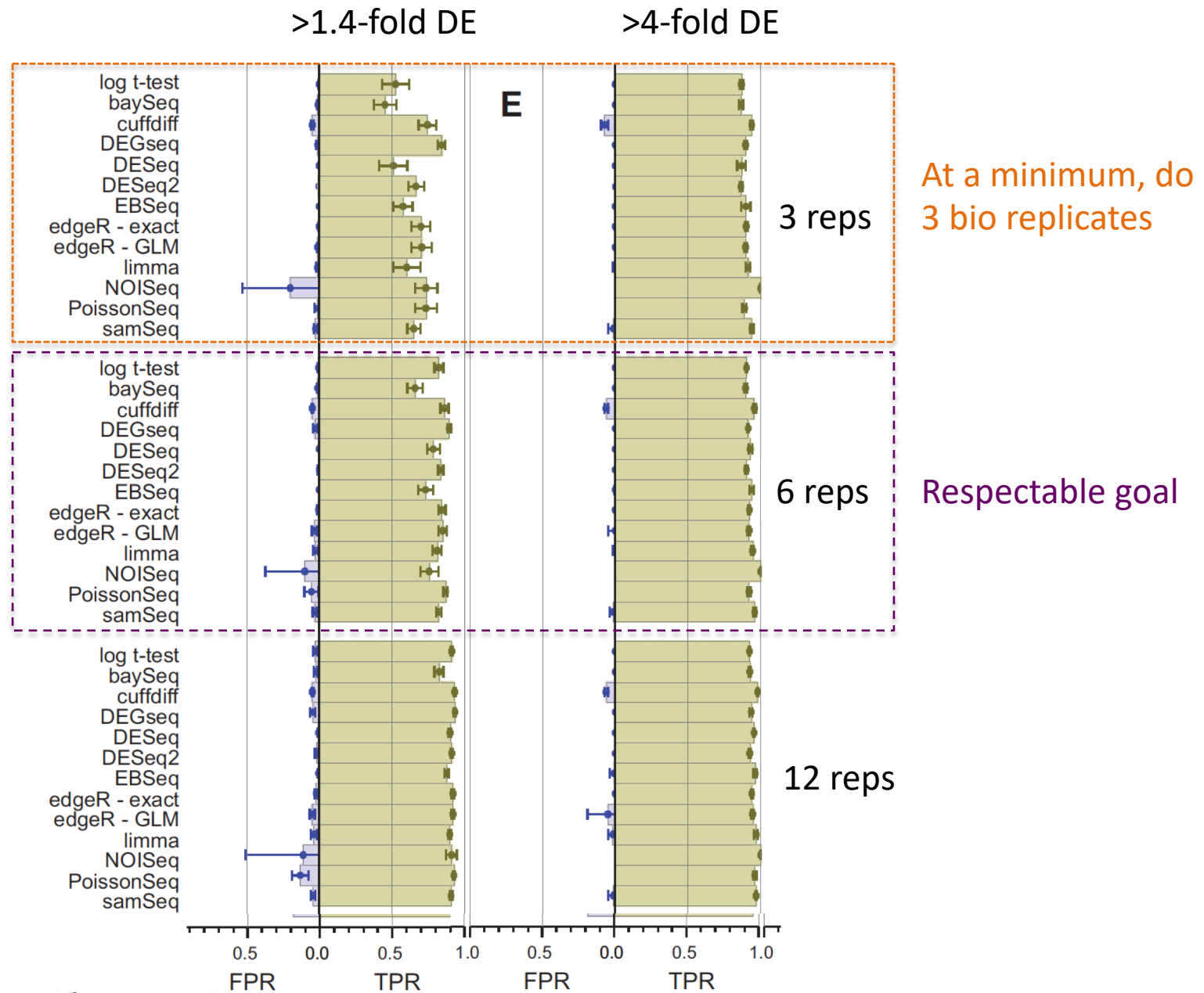
However, biological replicates ***ARE*** essential

$\text{total_variance} = \text{technical_variance} + \text{biological_variance}$

(Total variance well-modeled by negative binomial distribution)

“... **at least six biological replicates should be used**, rising to at least 12 when it is important to identify SDE genes for all fold changes.” *Schurch et al., RNA, 2016*

DE Accuracy Improves with Higher Biological Replication



*Figure taken and adapted from Schurch et al., RNA, 2016

Tools for DE analysis with RNA-Seq



edgeR

ShrinkSeq

DESeq

baySeq

Vsf

Limma/Voom

mmdiff

cuffdiff

ROTS

TSPM

DESeq2

EBSeq

NBPSeq

SAMseq

NoiSeq

Sleuth

*(italicized not in R/Bioconductor
but stand-alone)*

See: <http://www.biomedcentral.com/1471-2105/14/91>

A comparison of methods for differential expression analysis of RNA-seq data

Soneson & Delorenzi, 2013

Typical output from DE analysis

	logFC	logCPM	PValue	FDR
TRINITY_DN876_c0_g1_i1	-7.15049572793027	10.6197708379285	0	0
TRINITY_DN6470_c0_g1_i1	-7.26777912190146	7.03987604865422	1.687485656951e-287	6.46813252309319e-284
TRINITY_DN5186_c0_g1_i1	-7.85623682454322	9.18570464327063	1.17049180235068e-278	2.99099671894011e-275
TRINITY_DN768_c0_g1_i1	7.72884741150304	9.7514619195169	4.32504881419265e-272	8.28895605240022e-269
TRINITY_DN70_c0_g1_i1	-12.7646078189688	7.86482982471445	3.92853491279431e-253	6.02322972829624e-250
TRINITY_DN1587_c0_g1_i1	-5.89392061881667	9.07366563894607	6.32919557933429e-243	8.08660221852944e-240
TRINITY_DN3236_c0_g1_i1	-7.27029815068473	8.02209568234202	3.64955175271959e-235	3.99678053376405e-232
TRINITY_DN4631_c0_g1_i1	-7.45310693639574	6.91664918183241	4.30540921272851e-229	4.1256583780971e-226
TRINITY_DN5082_c0_g5_i1	-5.33154406167545	10.6977538760467	2.74243356676259e-225	2.33594396920022e-222
TRINITY_DN1789_c0_g3_i1	10.2032564835076	7.32607652700285	1.44273728647186e-213	1.10600240380933e-210
TRINITY_DN4204_c0_g1_i1	4.81030233739325	9.88844409410644	9.27180216086162e-205	6.46160321501501e-202
TRINITY_DN799_c0_g1_i1	-4.22044475626154	6.9937398638711	1.24746518421083e-197	7.96922341846683e-195
TRINITY_DN196_c0_g2_i1	4.60597918494257	9.86878463857276	1.9819997623131e-192	1.16877001368402e-189
TRINITY_DN5041_c0_g1_i1	-4.27126549355785	9.70894399883	1.8930437900069e-185	1.03657669244235e-182
TRINITY_DN1619_c0_g1_i1	-4.47156415953777	9.22535948721718	1.76766063029526e-181	9.03392426122899e-179
TRINITY_DN899_c0_g1_i1	-4.90914328409143	7.93768691394594	1.11054513767547e-180	5.32089939088761e-178
TRINITY_DN324_c0_g2_i1	4.87160837667488	6.84850312231775	2.20092562166991e-179	9.92487989160089e-177
TRINITY_DN3241_c0_g1_i1	-4.77760618069256	7.94111259715689	1.60585457735621e-173	6.83915621667372e-171
TRINITY_DN4379_c0_g1_i1	3.85133572453294	7.23712813663389	3.48140532848425e-164	1.4046554341137e-161
TRINITY_DN1919_c0_g1_i1	4.05998814332136	6.95937301668582	1.8588621194715e-161	7.12501850393425e-159
TRINITY_DN2504_c0_g1_i1	-6.92417817059644	6.20370039359785	2.42022459856956e-160	8.83497227268296e-158

...



Up vs. Down regulated

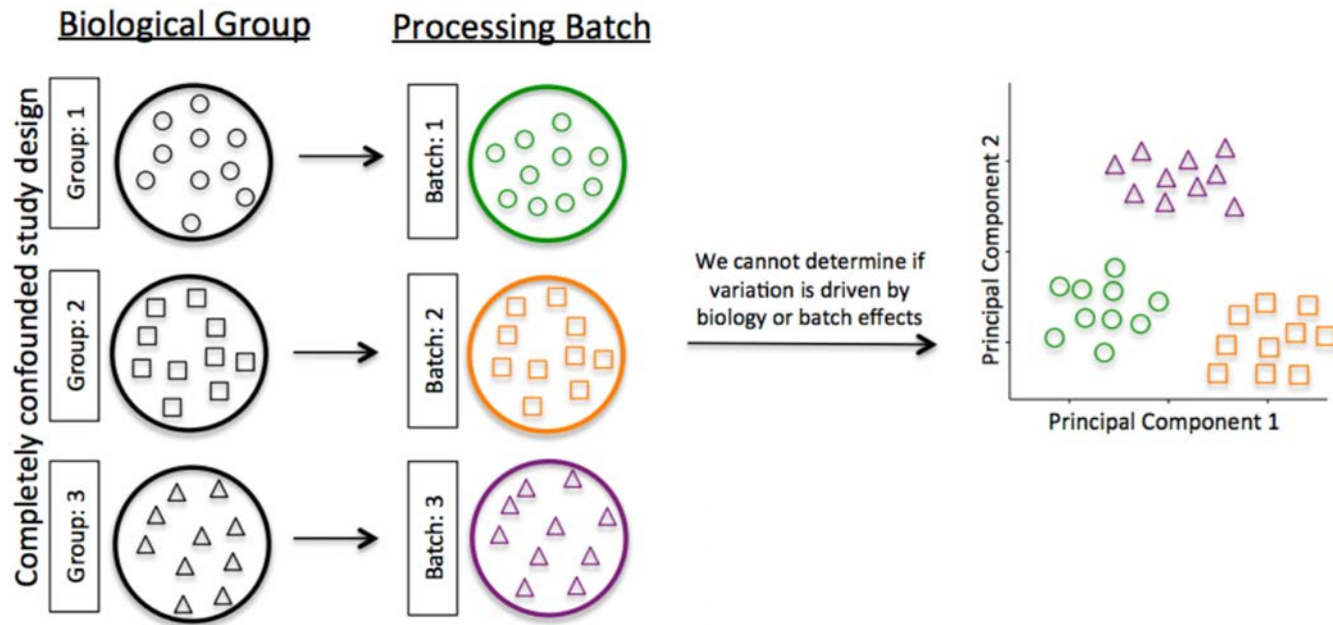


Avg. expression level



Significance

Avoid Batch Effects



Batch variable types:

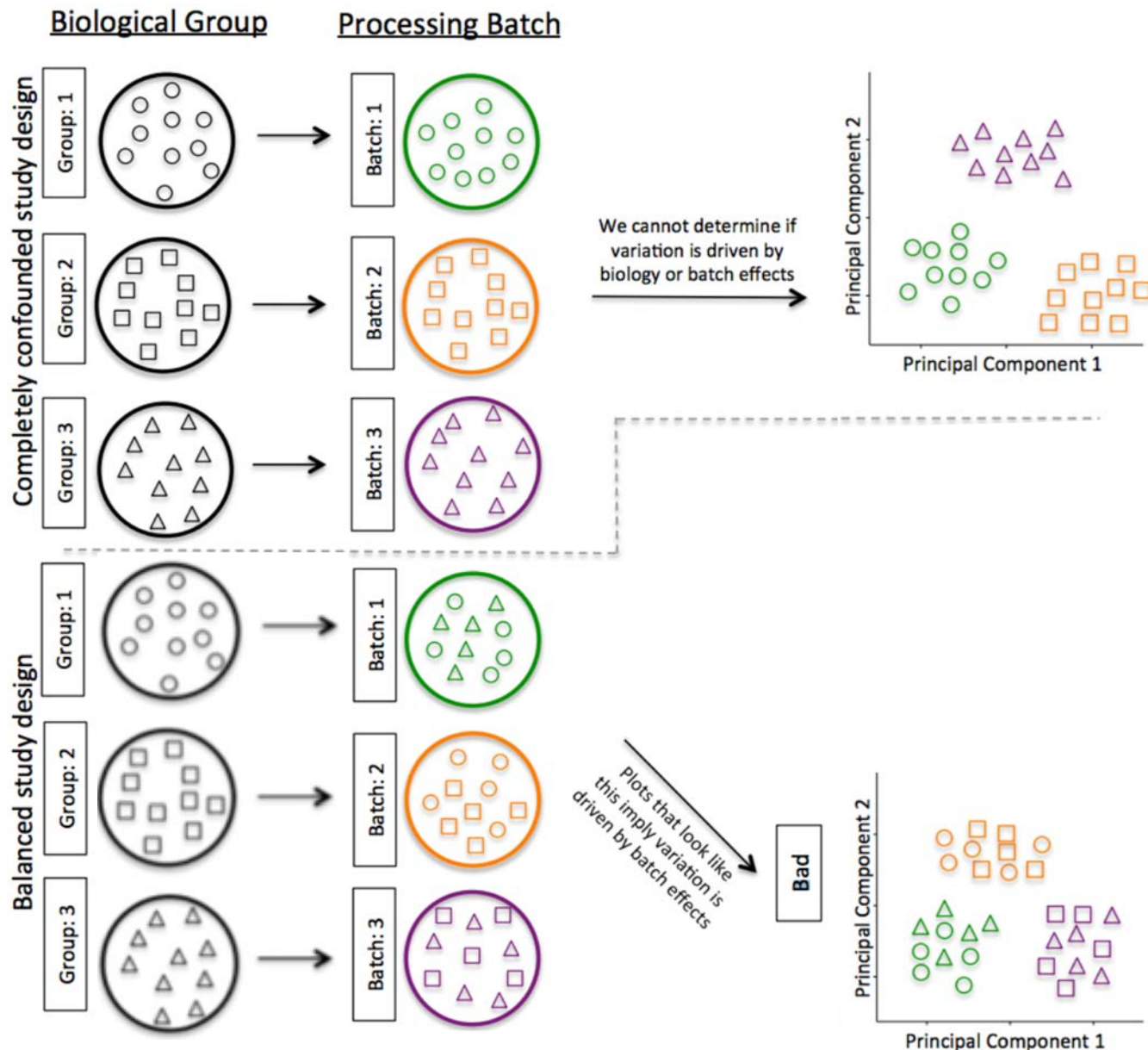
- Times and dates
- Technician processing the samples
- Sequencing machine, or flow cell lane (Illumina)

Adapted from: Stephanie C. Hicks, Mingxiang Teng, Rafael A. Irizarry.

<https://www.biorxiv.org/content/early/2015/09/04/025528>

On the widespread and critical impact of systematic bias and batch effects in single-cell RNA-Seq data.

Avoid Batch Effects



Grouping by
Study or
Batch?



Grouping
by Batch



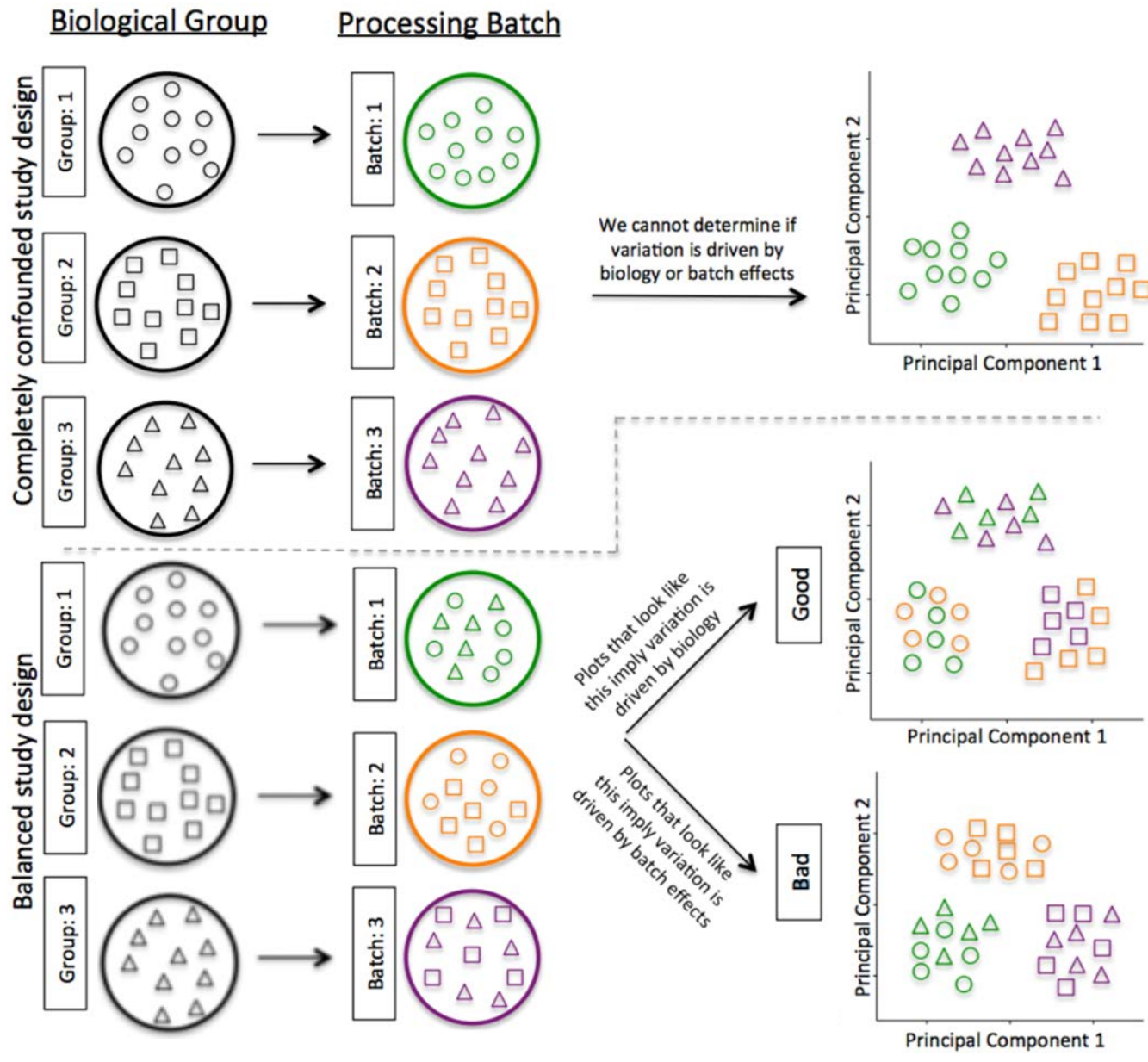
Adapted from: Stephanie C. Hicks, Mingxiang Teng, Rafael A. Irizarry.

<https://www.biorxiv.org/content/early/2015/09/04/025528>

On the widespread and critical impact of systematic bias and batch effects in single-cell RNA-Seq data.

(Explore Batch Removal Techniques)

Avoid Batch Effects



Grouping by Study or Batch?



Grouping by Study



Grouping by Batch



Adapted from: Stephanie C. Hicks, Mingxiang Teng, Rafael A. Irizarry.

<https://www.biorxiv.org/content/early/2015/09/04/025528>

On the widespread and critical impact of systematic bias and batch effects in single-cell RNA-Seq data.

(Explore Batch Removal Techniques)

Mouse and human tissue expression more similar within than between species. ?!?!?



[Proc Natl Acad Sci U S A](#). 2014 Dec 2; 111(48): 17224–17229.

PMCID: PMC4260565

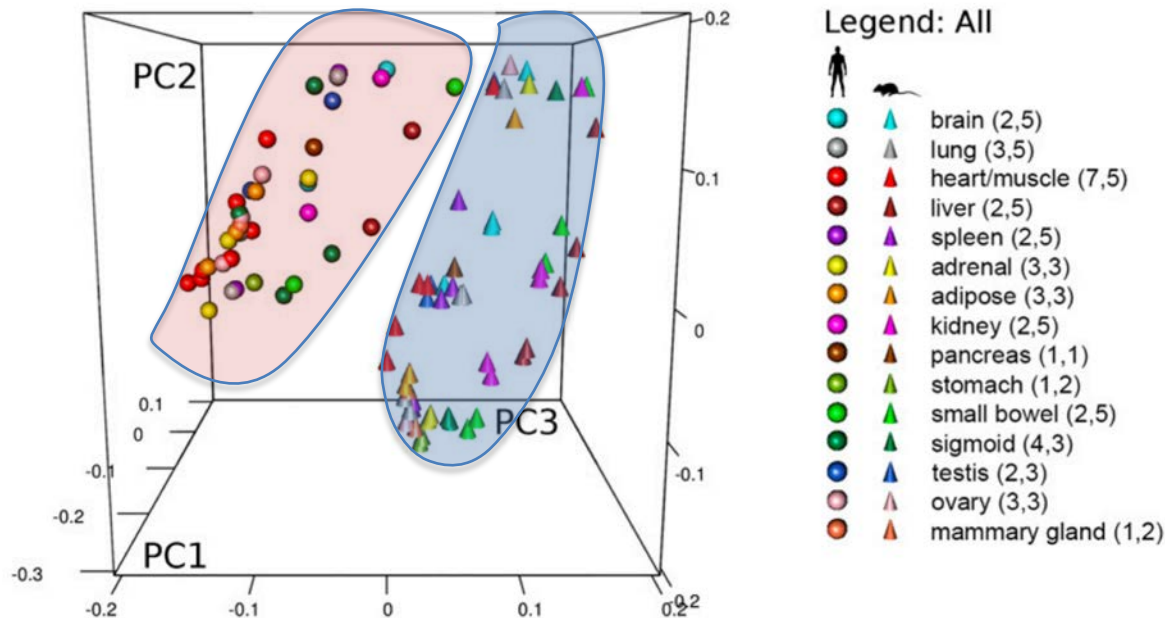
Published online 2014 Nov 20. doi: [10.1073/pnas.1413624111](https://doi.org/10.1073/pnas.1413624111)

PMID: [25413365](https://pubmed.ncbi.nlm.nih.gov/25413365/)

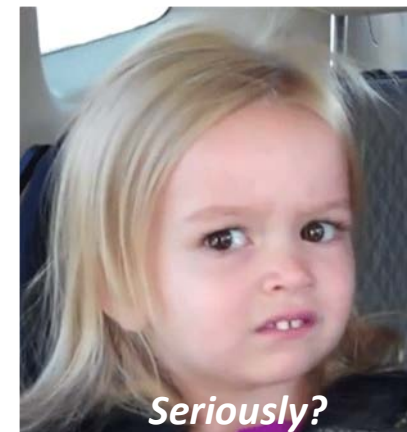
Genetics

Comparison of the transcriptional landscapes between human and mouse tissues

[Shin Lin](#),^{a,b,1} [Yiing Lin](#),^{c,1} [Joseph R. Nery](#),^d [Mark A. Urich](#),^d [Alessandra Breschi](#),^{e,f} [Carrie A. Davis](#),^g
[Alexander Dobin](#),^g [Christopher Zaleski](#),^g [Michael A. Beer](#),^h [William C. Chapman](#),^c
[Thomas R. Gingeras](#),^{g,i} [Joseph R. Ecker](#),^{d,j,2} and [Michael P. Snyder](#)^{a,2}



“... our results indicate that for the human–mouse comparison, tissues appear more similar to one another within the same species than to the comparable organs of other species ...”



~6 months later

RESEARCH ARTICLE

A reanalysis of mouse ENCODE comparative gene expression data [version 1; referees: 3 approved, 1 approved with reservations]

Yoav Gilad, Orna Mizrahi-Man

Department of Human Genetics, University of Chicago, Chicago, IL, 60637, USA

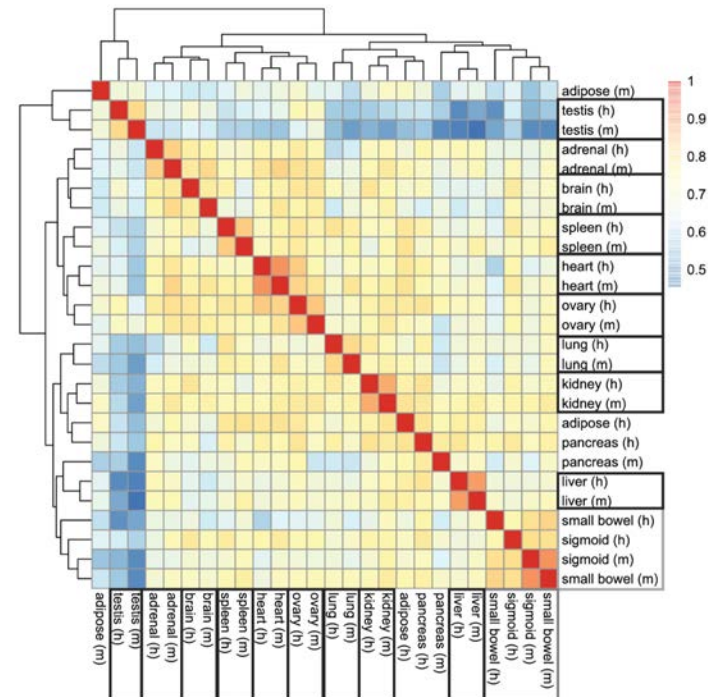
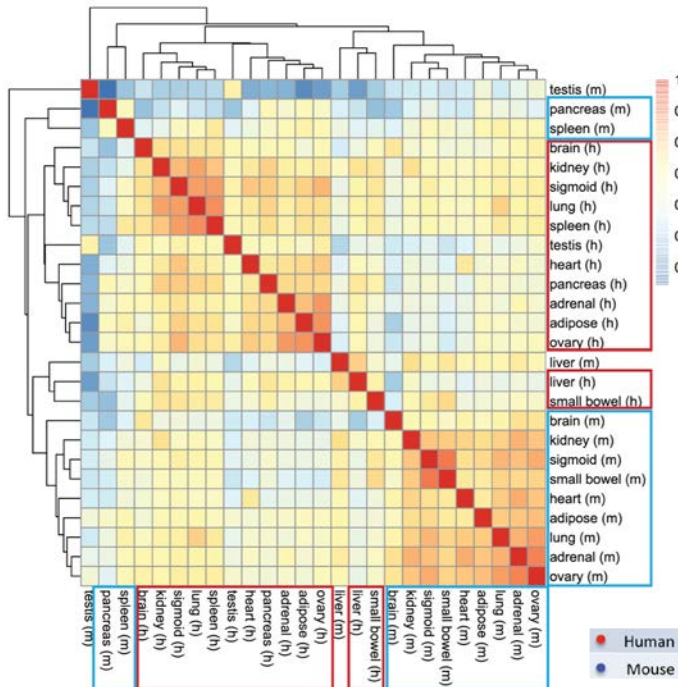
Yes, tissue expression patterns within species more similar than between species, but doesn't make sense and maybe due to a batch effect?

Grouping of samples by Sequencing Batch

D87PMJN1 (run 253, flow cell D2GUAACXX, lane 7)	D87PMJN1 (run 253, flow cell D2GUAACXX , lane 8)	D4LHBFN1 (run 276, flow cell C2HKJACXX , lane 4)	MONK (run 312, flow cell C2GR3ACXX , lane 6)	HWI-ST373 (run 375, flow cell C3172ACXX , lane 7)
heart	adipose	adipose	heart	brain
kidney	adrenal	adrenal	kidney	pancreas
liver	sigmoid colon	sigmoid colon	liver	brain
small bowel	lung	lung	small bowel	spleen
spleen	ovary	ovary	testis	Human
testis		pancreas		Mouse

Post Batch Correction:

Tissue patterns more similar than by species



Flavors of Differential Expression Analyses

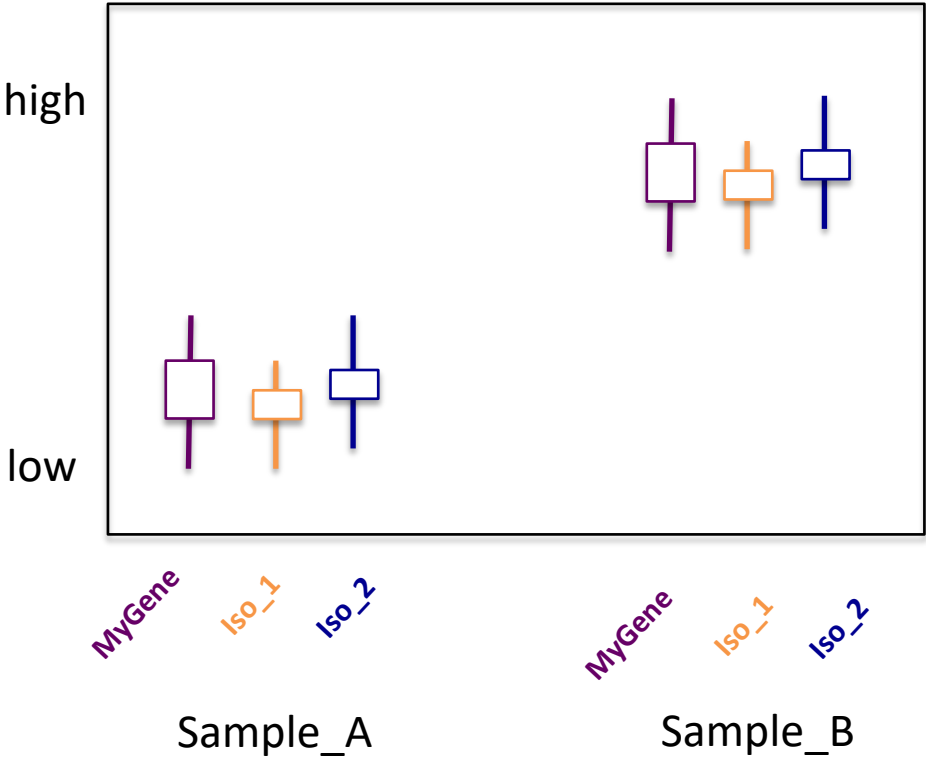
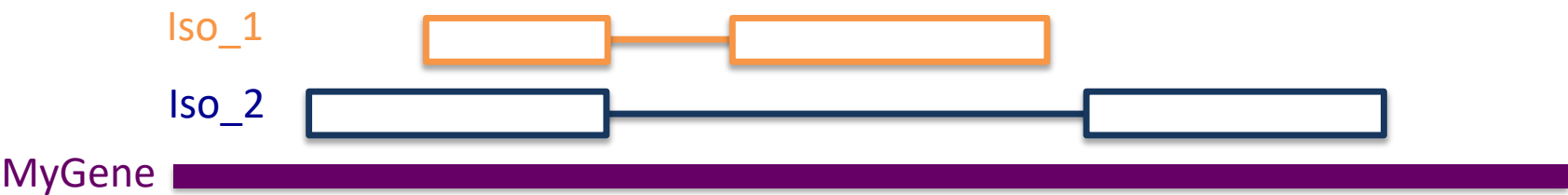
- Transcripts:
 - Differential Transcript Expression (DTE)
 - Differential Transcript Usage (DTU)
 - Differential Exon Usage (DEU)
- Gene:
 - Differential Gene Expression (DGE)
 - Gene Differential Expression (GDE)



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

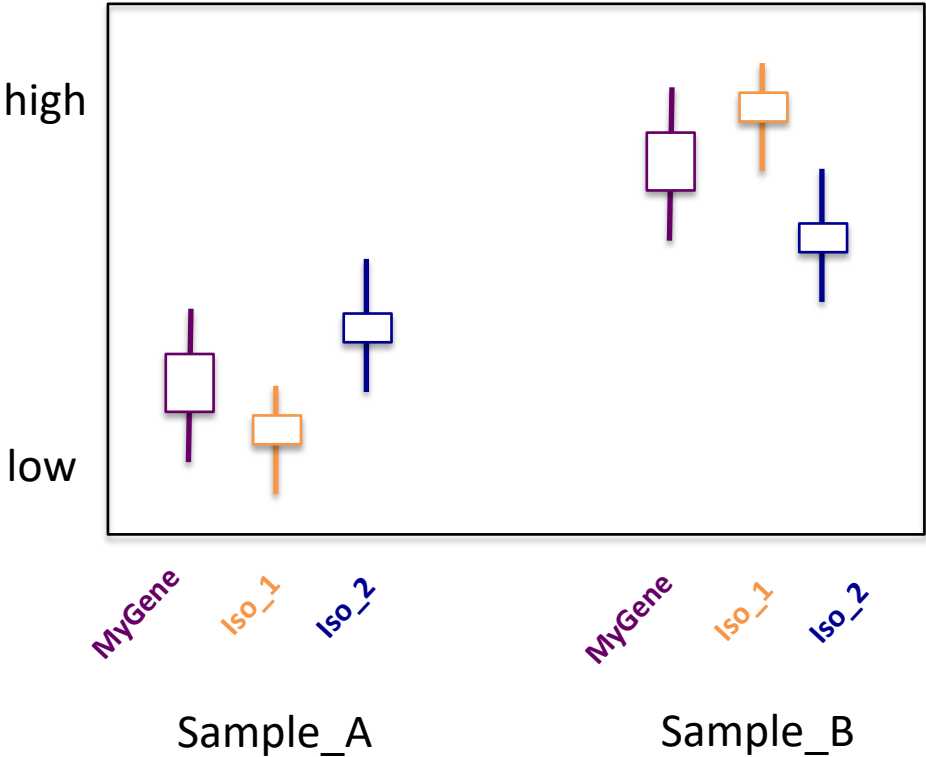
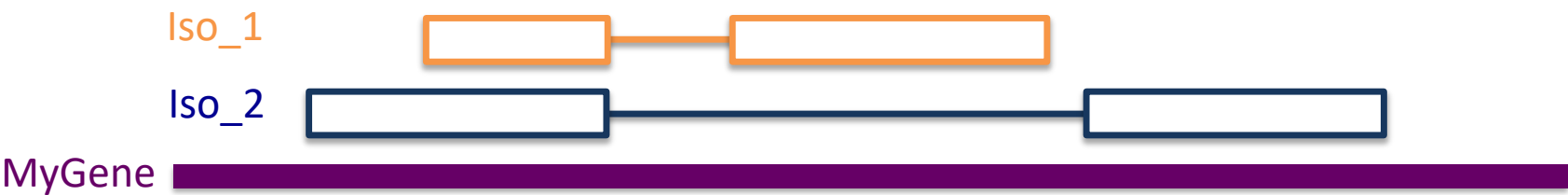


Differential Gene Expression (DGE) and Differential Transcript Expression (DTE) (Example 1)



Feature	Diff Expressed?
MyGene	Yes
Iso_1	Yes
Iso_2	Yes
Diff. Transcript Usage ? (eg. Isoform switching)	No

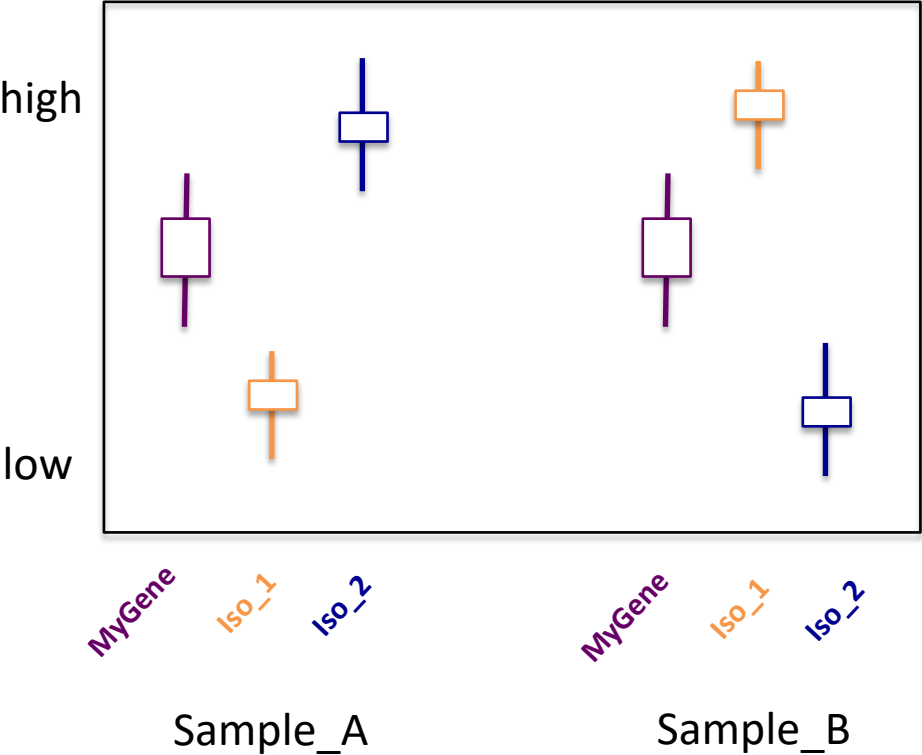
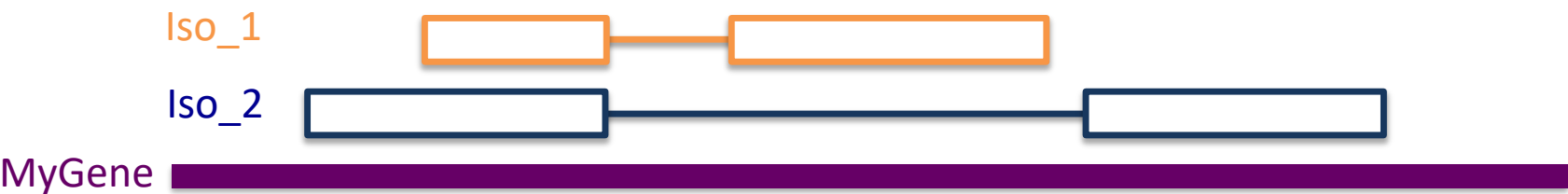
Differential Gene Expression (DGE) and Differential Transcript Expression (DTE) (Example 2)



Feature	Diff Expressed?
MyGene	Yes
Iso_1	Yes
Iso_2	Yes

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

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



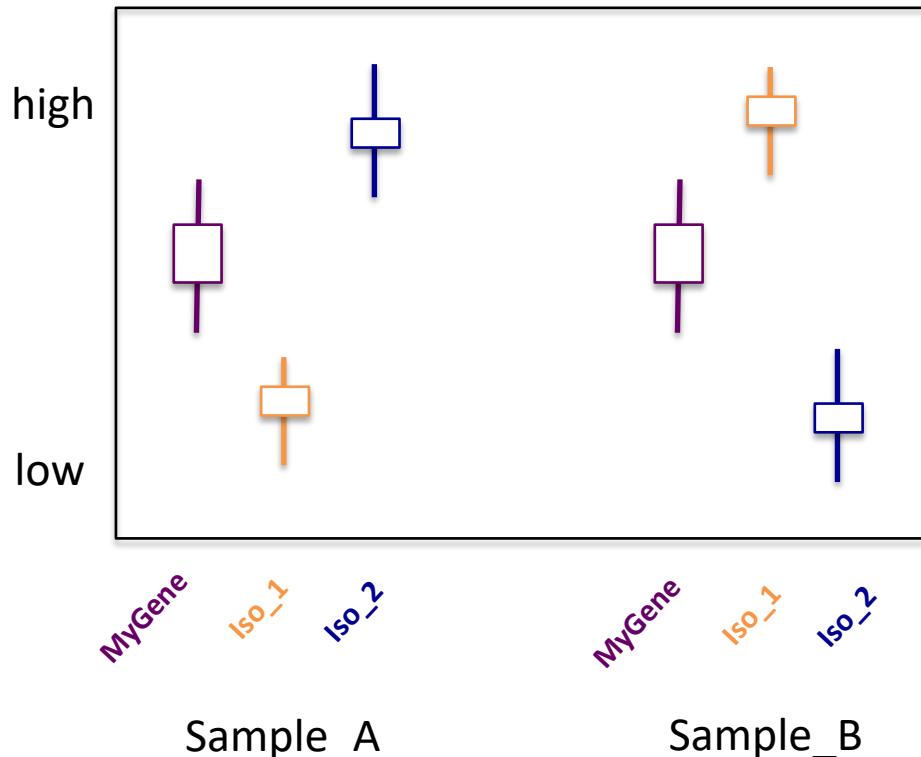
Feature	Diff Expressed?
MyGene	No
Iso_1	Yes
Iso_2	Yes

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

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

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

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



Feature Diff Expressed?

MyGene

No?

Iso_1

Yes

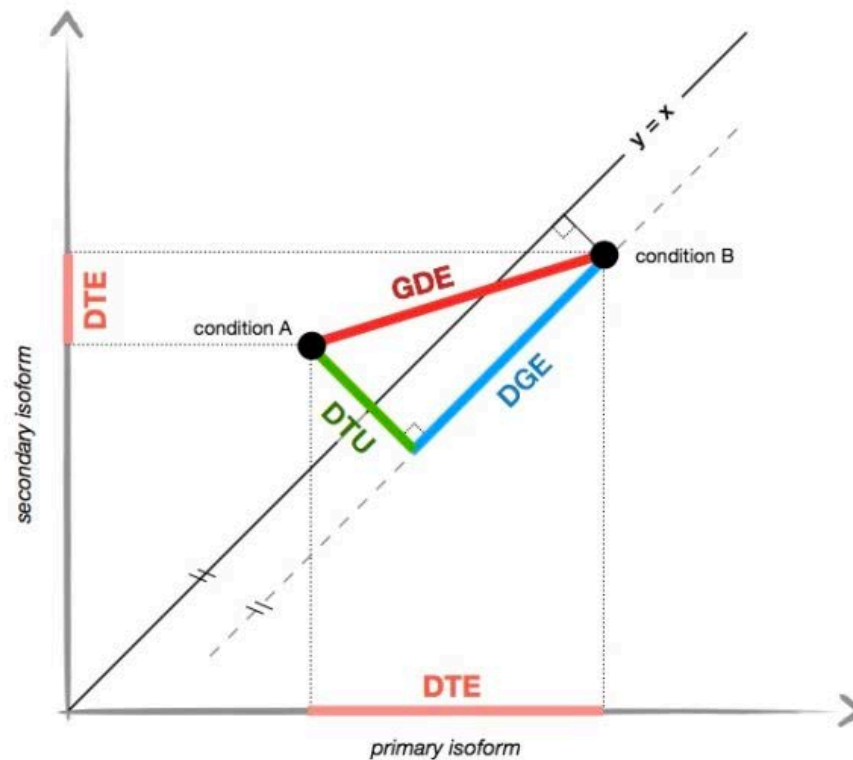
Iso_2

Yes

Diff. Transcript Usage ?
(eg. Isoform switching)

Yes

Clarifying view: (DTE or DTU or DGE) as special cases of Gene Differential Expression (DGE)

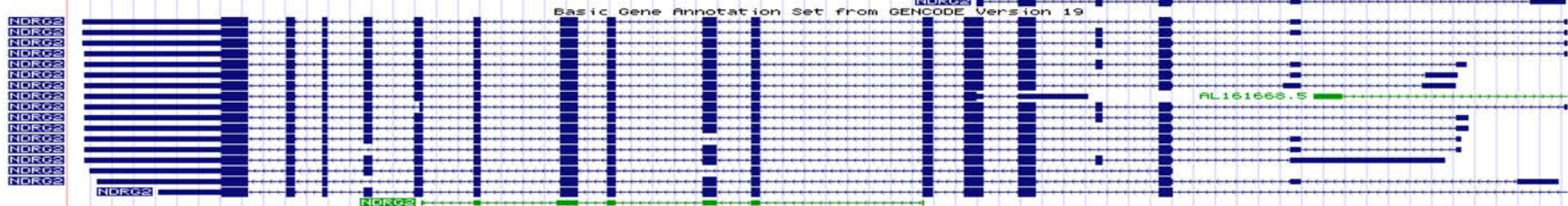


DTE: differential transcript expression
DTU: differential transcript usage
DGE: differential gene expression (gene-level analysis)
GDE: gene differential expression (transcript-level analysis)

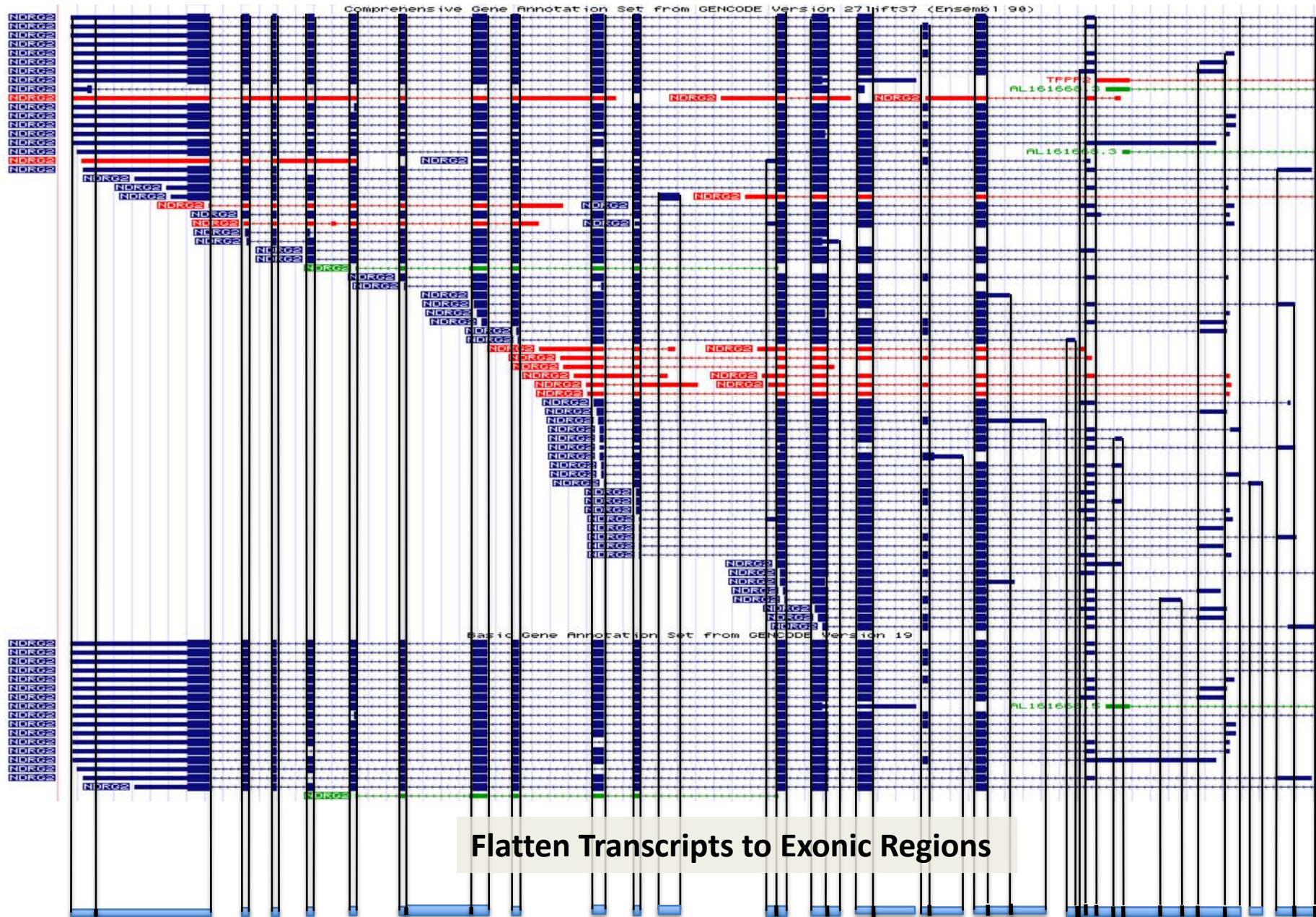
Ntranos, Yi, et al., 2018 – see supp.

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

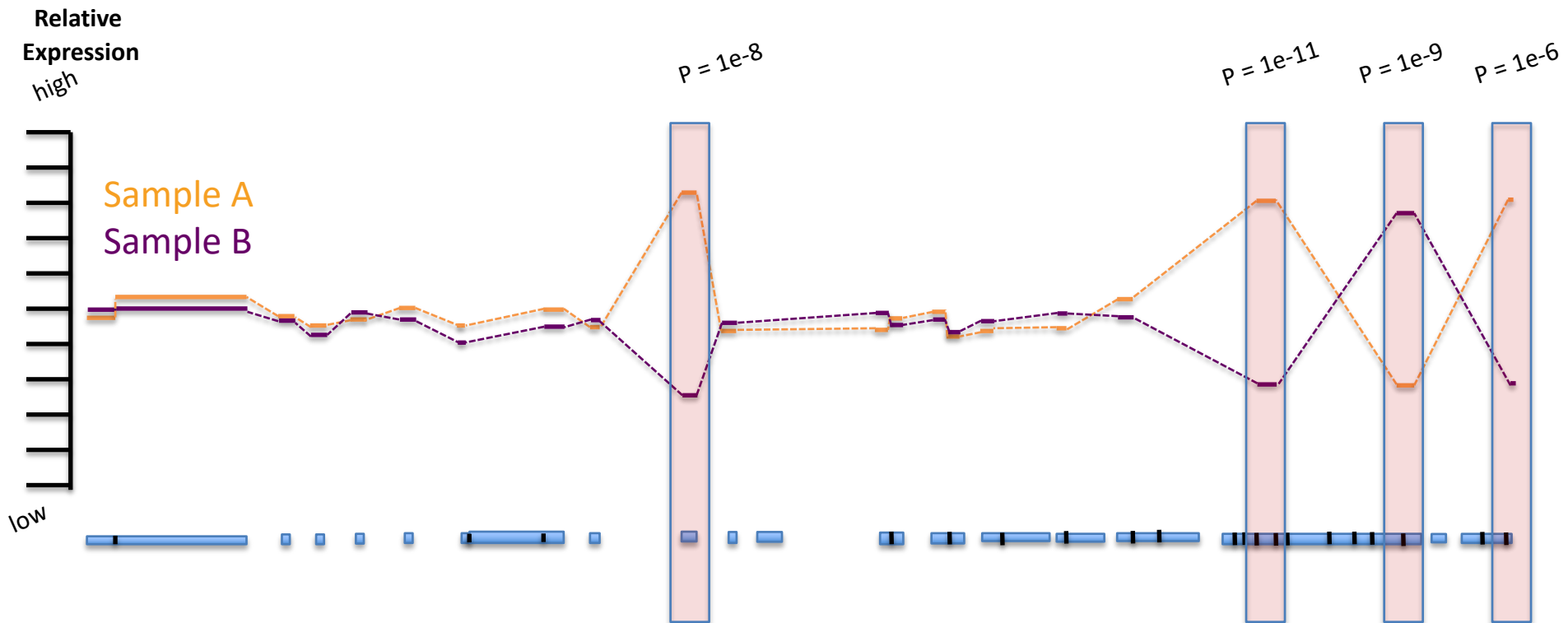
Which isoforms are expressed?
 there evidence of differential transcript usage?



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



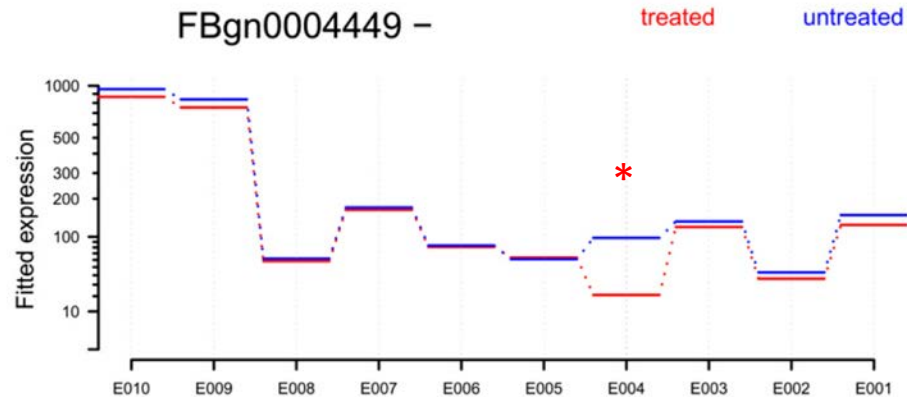
Measure Differential Transcript Usage (DTU) via Differential Exon Usage (DEU)



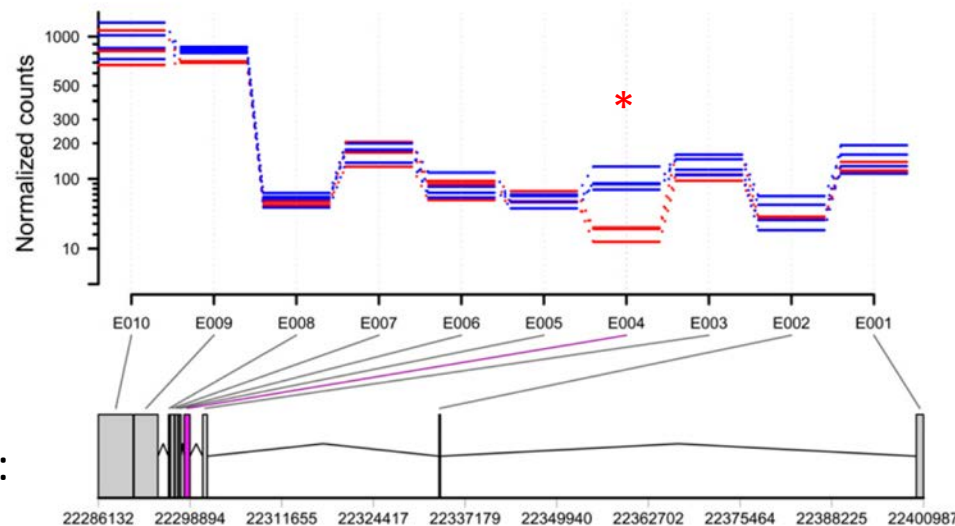
Detecting differential usage of exons from RNA-seq data

Simon Anders,^{1,2} Alejandro Reyes,¹ and Wolfgang Huber

Averaged Replicates



Each Replicate



Flattened gene structure:

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

Enabling Differential Transcript Usage Analysis for De novo Transcriptome Assemblies

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

Genome Biology

METHOD

Open Access

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



Nadia M. Davidson^{1,2*}, Anthony D. K. Hawkins¹ and Alicia Oshlack^{1,2*} 

Enabling Differential Transcript Usage Analysis for De novo Transcriptome Assemblies

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

Genome Biology

METHOD

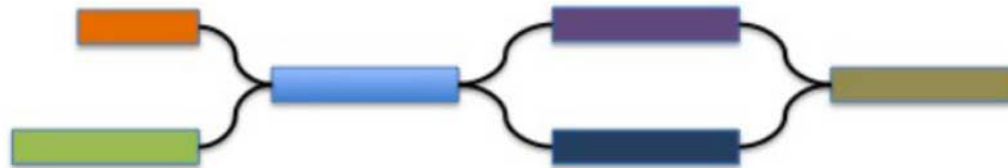
Open Access

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



Nadia M. Davidson^{1,2*}, Anthony D. K. Hawkins¹ and Alicia Oshlack^{1,2*} 

Transcript splice graph:



Similar method and protocols now integrated into Trinity:

<https://github.com/trinityrnaseq/trinityrnaseq/wiki/SuperTranscripts>

Enabling Differential Transcript Usage Analysis for De novo Transcriptome Assemblies

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

Genome Biology

METHOD

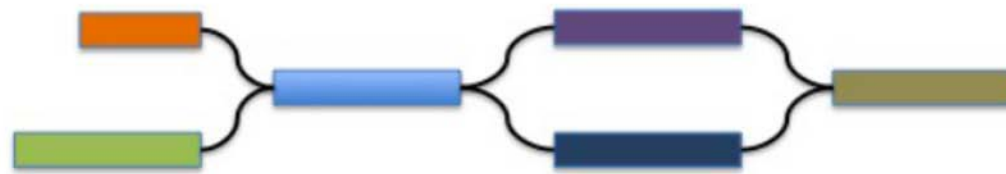
Open Access

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



Nadia M. Davidson^{1,2*}, Anthony D. K. Hawkins¹ and Alicia Oshlack^{1,2*} 

Transcript splice graph:



Linearize graph via topological sorting
or graph multiple alignment

SuperTranscript:



DEXseq for DTU,
GATK for Variant Detection

Similar method and protocols now integrated into Trinity:

<https://github.com/trinityrnaseq/trinityrnaseq/wiki/SuperTranscripts>

Further Pushing the Envelope with RNA-Seq Analysis

Audoux et al. *Genome Biology* (2017) 18:243
DOI 10.1186/s13059-017-1372-2

Genome Biology

METHOD

Open Access

DE-kupl: exhaustive capture of biological variation in RNA-seq data through *k*-mer decomposition

Jérôme Audoux¹, Nicolas Philippe^{2,3}, Rayan Chikhi⁴, Mikaël Salson⁴, Mélina Gallopin⁵, Marc Gabriel^{5,6},
Jérémy Le Coz⁵, Emilie Drouineau⁵, Thérèse Combes^{1,2} and Daniel Gautheret^{5,6*} 

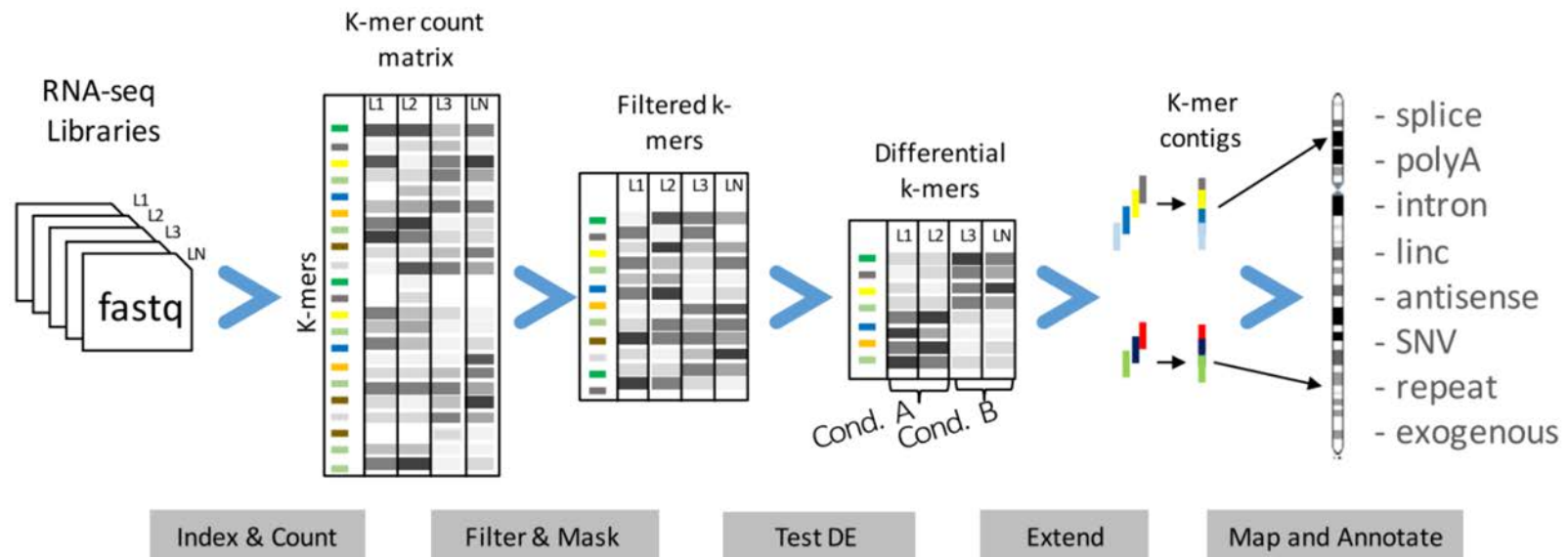


Fig. 4 The DE-kupl pipeline for the discovery and analysis of differentially expressed *k*-mers. First, Jellyfish is applied to count *k*-mers in all libraries. *k*-mers counts are then joined into a count matrix and filtered for low recurrence and matching to the reference transcriptome. Normalization factors are computed from raw *k*-mer counts and the differential expression procedure is applied. Finally, overlapping differentially expressed *k*-mers are extended into contigs and annotated based on their alignment to the reference and overlap with annotated genes

Instead of Gene, Transcript, or Exon Counts (or expression levels... Can we use the newly defined Transcript Equivalence Classes for DE?

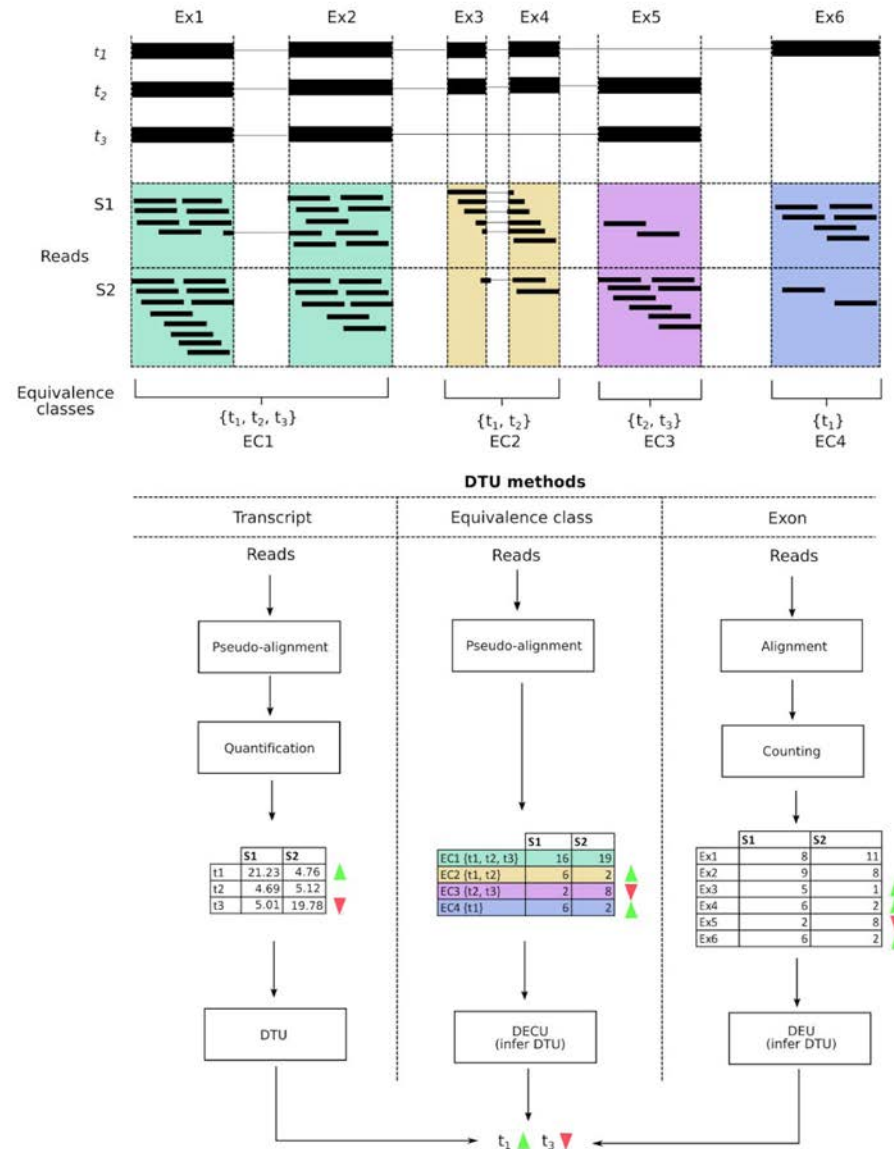


Fig. from Marek Cmero, Nadia M Davidson, Alicia Oshlack
bioRxiv, Dec. 2018

Again, see Lior Pachter's blog post: <https://liorpachter.wordpress.com/2019/01/07/fast-and-accurate-gene-differential-expression-by-testing-transcript-compatibility-counts/>

Which feature to choose for DE analysis?

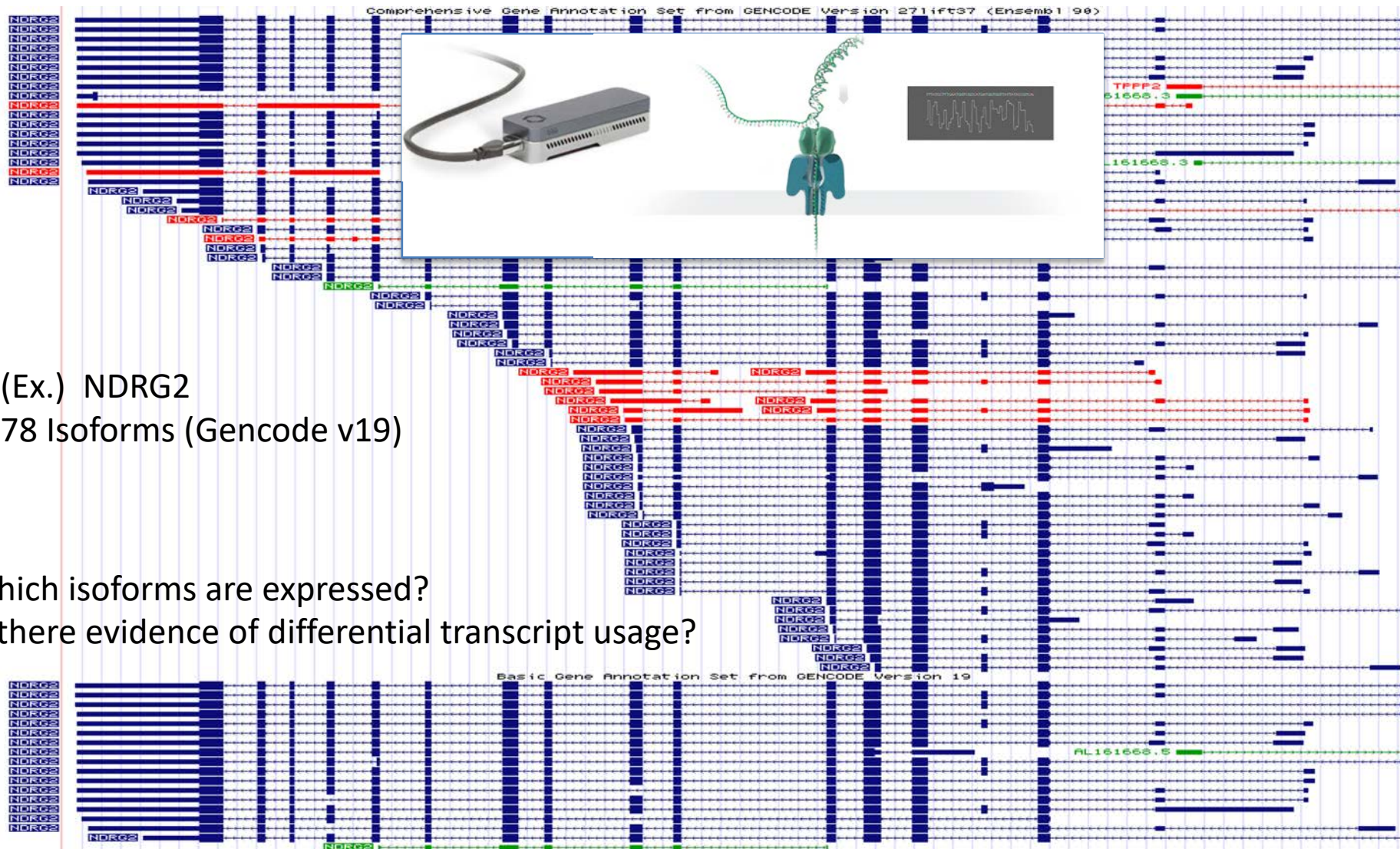
Feature	Pro	Con
Gene	Simplest, easily interpretable	Can hide evidence for DTU (iso switching)
Transcript	Illuminates the biological targets of interest.	Read mapping uncertainty among many isoforms confounding identification.
Exon	Detect evidence of DTU leveraging many isoform structures, and pinpoint exonic regions of interest.	May not easily identify specific isoforms of interest in shared exon situations.
Equivalence Class	Very Fast, not requiring read alignment nor sophisticated abundance estimation.	Biological interpretability can be challenging.

Routine

Semi-specialized

Burgeoning

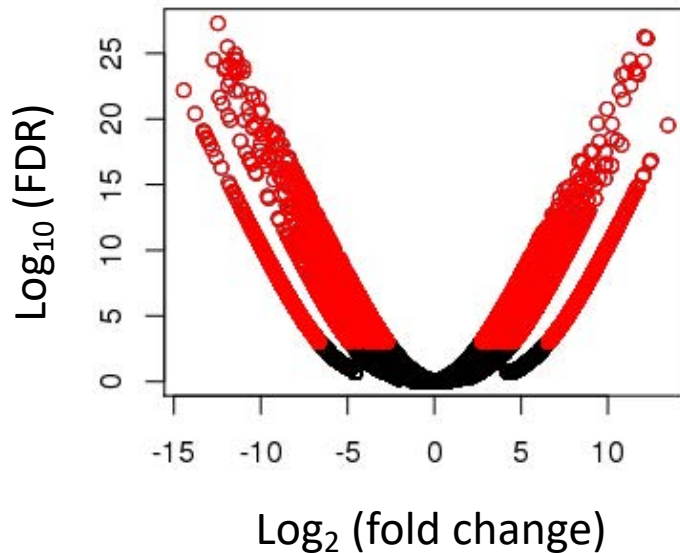
Too complex... don't guess from short reads, use long reads.



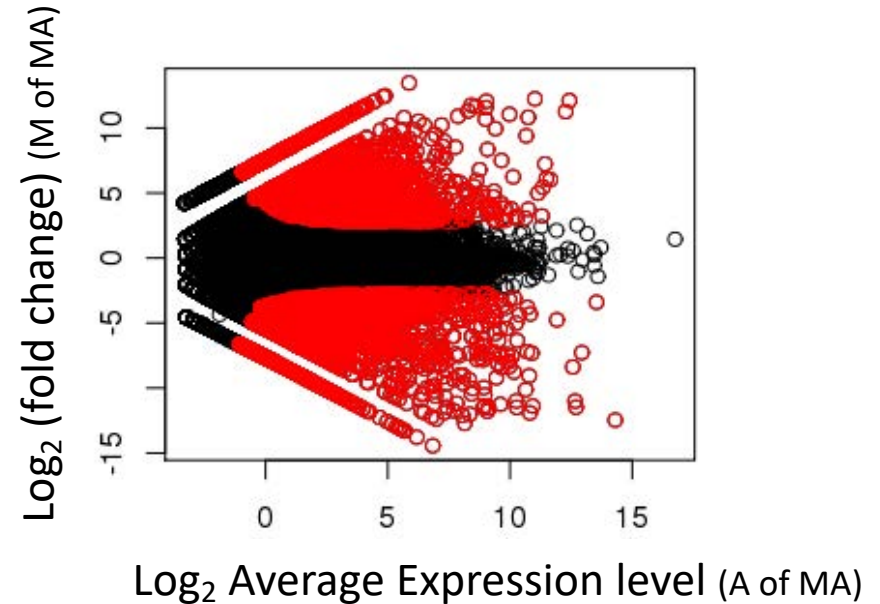
Visualization of DE results and Expression Profiling

Plotting Pairwise Differential Expression Data

Volcano plot
(fold change vs. significance)

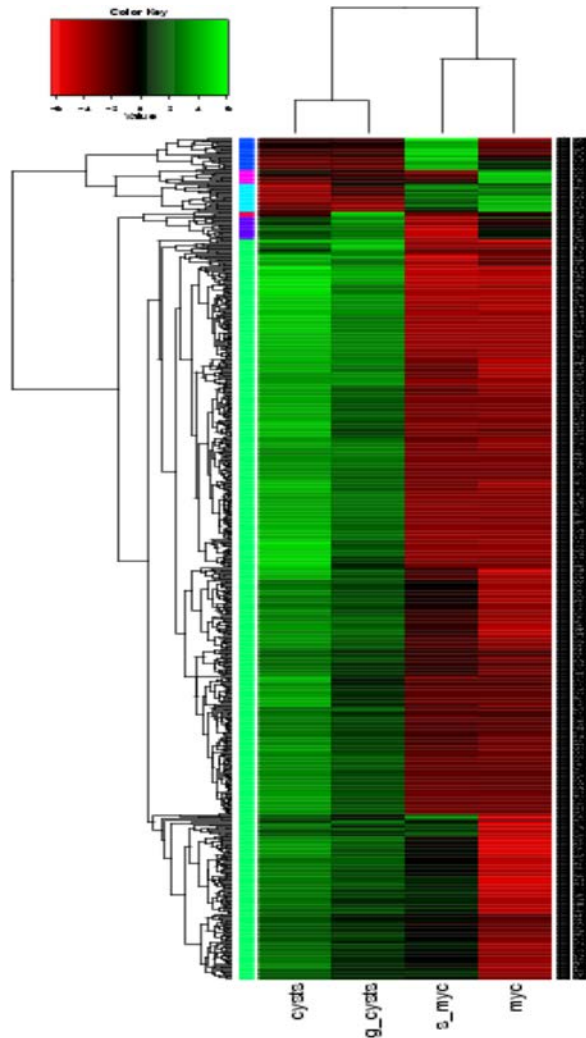


MA plot
(abundance vs. fold change)



Significantly differently expressed transcripts have $\text{FDR} \leq 0.001$
(shown in red)

Comparing Multiple Samples



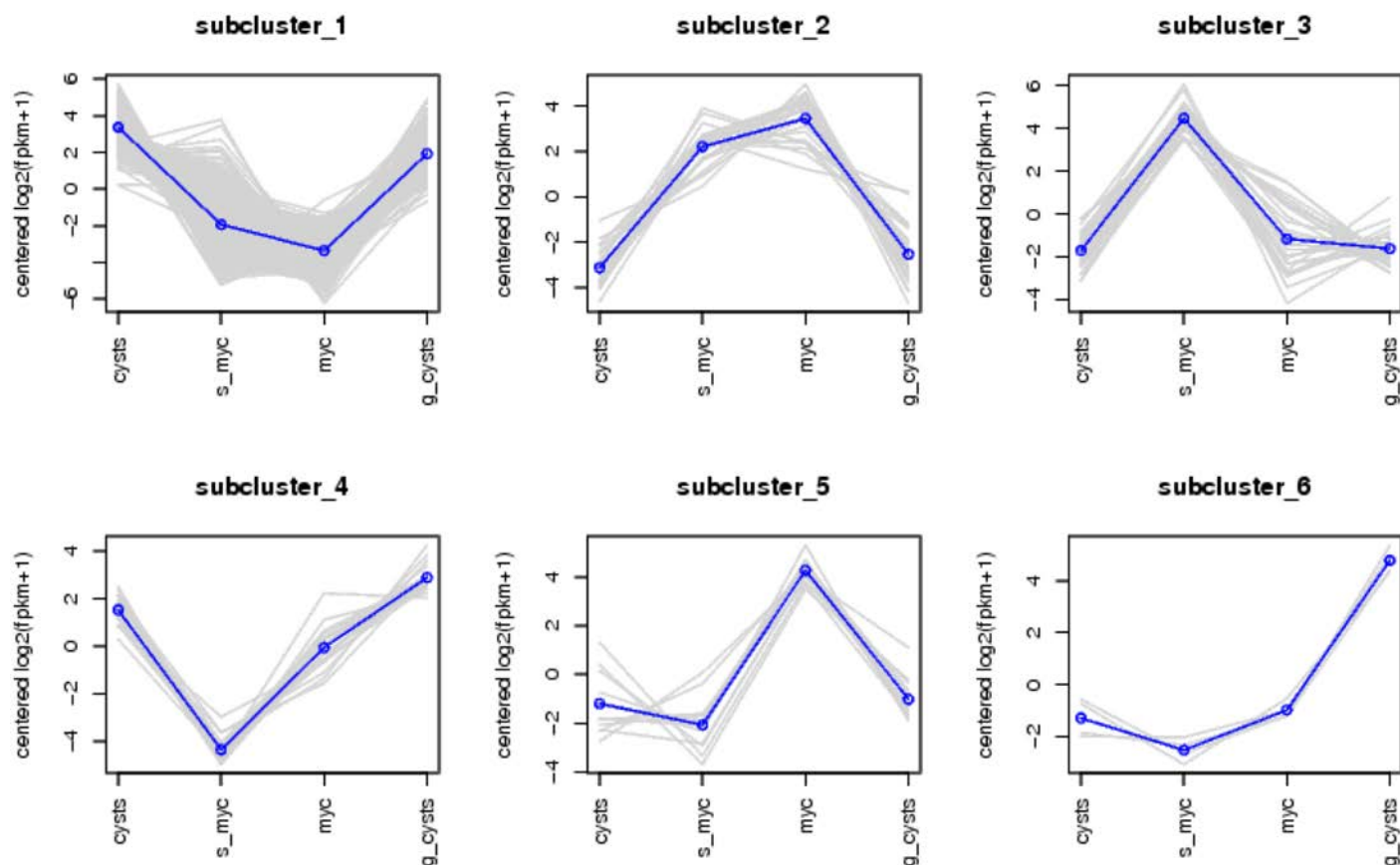
Heatmaps provide an effective tool for navigating differential expression across multiple samples.

Clustering can be performed across both axes:

- cluster transcripts with similar expression patterns.
- cluster samples according to similar expression values among transcripts.

Examining Patterns of Expression Across Samples

Can extract clusters of transcripts and examine them separately.



Transcript Functional Annotation

GGAGCTGGAGGCCCCCAGGCAACTACACCGTCCACGTACCCAGAGGGGGCTGGGCCCCTCCC
ACCAGAGACCACGCCCTGGTGTGCCTTAGGGGGCCCTGGTTTGTTAGTCTCTGAGTGTGCA
GTTGCTGCACATGGGGCCCTGGCGCTTGCTGCACCAACTTCCTGTTGGGCCCCGTGGTCCT
TGGAGGCATGCAGTTCAGCAGACAGTGACTCAGCCATCCACCCAACATGCGGAACGTGTC
TCTTCTGCAGGTCCCGGTCCACAGCAGGATTCCCCCTCTGTGAAAAGGCACGCTGATCTG
TCTGGA TCGAC
TCTCC TCCCA
AAAGAC CCTGG
GGCTTG CCTAA
TGACCT TGCTG
GAAAAG CAGCC
TTGTCA TTCCA
GGAAGCACATAATTGAAGGACTGAAAGCGTCCCTGGAGCGGCTGCAGCTGGAGTACGTGG
ATGTGGTTTTTTGCCAACCGCCCAGACCCCAACACGCCCATGGAAGAGACCGTGCGGGCCA
TGACCCATGTCATCAACCAGGGGATGGCCATGTACTGGGGCACATCACGCTGGAGCTCCA
TGGAGATCATGGAGGCCTACTCGGTGGCTCGGCAGTTCAACCTGATCCCGCCCCATCTGCG
AGCAAGCGGAATATCACATGTTCCAGAGGGAGAAGGTGGAGGTCCAGCTGCCAGAGCTGT
TCCACAAGATAGGAGTAGGTGCCATGACCTGGTCCCCTCTGGCGTGCGGCATCGTCTCAG
GGAAGTATGACAGCGGGATCCCACCCTACTCCAGAGCCTCCCTGAAGGGCTACCAGTGGT
TGAAGGACAAGATCCTGAGTGAGGAGGGTCGCCGCCAGCAGGCCAAGCTGAAGGAACTGC
AGGCCATTGCCGAACGCCTGGGCTGCACCCTACCCCAGCTGGCCATAGCCTGGTGCCTGA
GGAATGAGGGTGTGTCAGCTCCGTGCTTCTGGGTGCTTCCAATGCAGAACAACCTTATGGAGA

Can we gather hints of biological function
from sequence?

Methods used to predict function from sequence

- Sequence homology

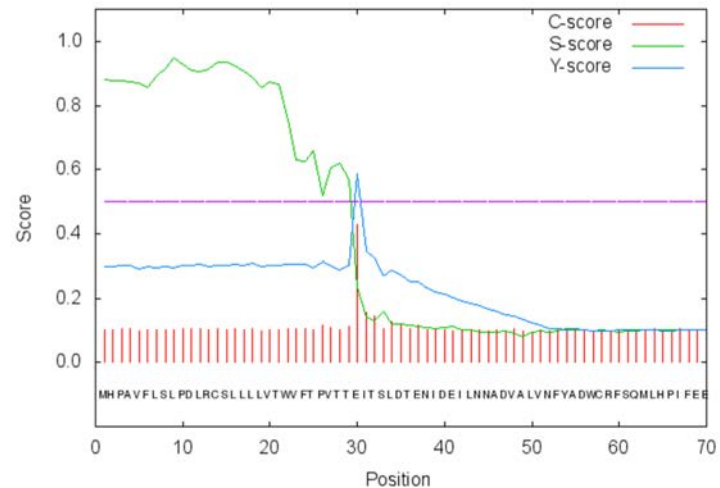
Searching protein database for sequence similarity

```
Query  THVHRPYNEHKSLSGTARYMSINTHLGREQSRDDLESMDGHVFMFLRGSLPW--QGLKA
       T   P + K   GT   Y S + HLG   RR DLE +G   L   LPW   Q L A
Database Match  TGDFKP-DPKKMHNGTIEYTSRDAHLG-VPTRRADLEILGYNLIEWLGAELPWVTQKLLA
```

- Sequence composition

Predict functions of sequence using machine learning methods for pattern recognition.

- Neural Networks
- Hidden Markov Models



Use BLAST to search for sequence similarity to known proteins

← → ↻ [Secure https://blast.ncbi.nlm.nih.gov/Blast.cgi](https://blast.ncbi.nlm.nih.gov/Blast.cgi) ☆             

 U.S. National Library of Medicine  NCBI National Center for Biotechnology Information [Sign in to NCBI](#)

BLAST®

[Home](#)

[Recent Results](#)

[Saved Strategies](#)

[Help](#)

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance.

[Learn more](#)

NEWS

Magic-BLAST 1.2.0 released

A new version of the BLAST RNA-seq mapping tool is now available.

Mon, 27 Feb 2017 14:00:00 EST

 [More BLAST news...](#)

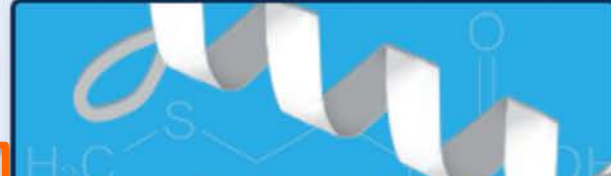
Web BLAST



Nucleotide BLAST
nucleotide ► nucleotide

blastx
translated nucleotide ► protein

tblastn
protein ► translated nucleotide



Protein BLAST
protein ► protein

The Swiss-Prot database is a valuable source of proteins with known functions

← → ↻ <https://www.uniprot.org> ☆ ABP a 7 JB PDF 21m

UniProt Advanced

BLAST Align Retrieve/ID mapping Peptide search Help Contact

The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

UniProtKB
UniProt Knowledgebase

Swiss-Prot
(558,898)
 Manually annotated and reviewed.

TrEMBL
(137,213,158)
 Automatically annotated and not reviewed.

(as of Jan, 2018)

UniRef
Sequence clusters

UniParc
Sequence archive

Proteomes

Supporting data

Literature citations

Taxonomy

Subcellular locations

Cross-ref. databases

Diseases
XXX

Keywords

News

[Forthcoming changes](#)
Planned changes for UniProt

[UniProt release 2018_11](#)
Enhanced enzyme annotation in UniProtKB using Rhea – integrating biology and chemistry

[UniProt release 2018_10](#)
You're not coming in!

[News archive](#)

[Text search](#)
Our basic text search allows you to search all the resources available

[BLAST](#)
Find regions of similarity between your sequences

[Sequence alignments](#)
Align two or more protein sequences using the



UniProt data

[Download latest release](#)
Get the UniProt data

[Statistics](#)
View Swiss-Prot and TrEMBL statistics

[How to cite us](#)
The UniProt Consortium

[Submit your data](#)

Protein spotlight



Ice Whisperer

December 2018

No one likes the cold. Humans wear scarves, fur boots, quilted coats and woollen hats to keep the harshness of winter out while other creatures grow their own fur or line their bodies with a thick layer of blubber. There are those, too, who have a more subtle approach to dealing with extreme temperatures

Example of a Swiss-Prot Record

← → ↻ ⓘ www.uniprot.org/uniprot/Q9H479 ☆

UniProtKB ▾ Advanced ▾ 🔍 Search

BLAST Align Retrieve/ID mapping Peptide search Help Contact

UniProtKB - Q9H479 (FN3K_HUMAN) 🛒 Basket ▾

Display

Entry Publications Feature viewer Feature table

BLAST Align Format Add to basket History Feedback Help video Other tutorials and videos

Protein | **Fructosamine-3-kinase**

Gene | **FN3K**

Organism | *Homo sapiens (Human)*

Status | Reviewed - Annotation score: ●●●●○ - Experimental evidence at protein levelⁱ

None

☒ Function

☒ Names & Taxonomy

☒ Subcell. location

☒ Pathol./Biotech

☒ PTM / Processing

☒ Expression

☒ Interaction

☒ Structure

☒ Family & Domains

☒ Sequence

☒ Cross-references

☒ Entry information

☒ Miscellaneous

Functionⁱ

May initiate a process leading to the deglycation of fructoselysine and of glycosylated proteins. May play a role in the phosphorylation of 1-deoxy-1-morpholino-fructose (DMF), fructoselysine, fructoseglycine, fructose and glycosylated lysozyme.

GO - Molecular functionⁱ

- fructosamine-3-kinase activity Source: UniProtKB ▾
- kinase activity Source: Reactome

Complete GO annotation...

GO - Biological processⁱ

- epithelial cell differentiation Source: UniProtKB ▾
- fructosamine metabolic process Source: GO_Central
- fructoselysine metabolic process Source: UniProtKB ▾
- post-translational protein modification Source: Reactome

Complete GO annotation...

Keywordsⁱ

Molecular Kinase Transferase

Gene Ontology (GO):

Structured vocabulary for defining molecular functions, biological processes, and cellular components.

No significant sequence similarity... What else?

GGAGCTGGAGGCCCCCAGGCAACTACACCGTCCACGTACCCAGAGGGGGCTGGGCCCCTCCC
ACCAGAGACCACGCCCTGGTGTGCCTTAGGGGGCCCTGGTTTGTTAGTCTCTGAGTGTGCA
GTTGCTGCACATGGGGGCCCTGGCGCTTGCTGCACCAACTTCCTGTTGGGCCCCGTGGTCCT
TGGAGGCATGCAGTTCAGCAGACAGTGACTCAGCCATCCACCCAACATGCGGAACGTGTC
TCTTCTGCAGGTCCCGGTCCACAGCAGGATTCCCCCTCTGTGAAAAGGCACGCTGATCTG
TCTGGATAAGTGTGGCCGGCCCCATGTATCCGGAATCAACCACGGGGTCCCCAGCTCGAC
TCTCCCTGCGGCAGACAGGCTCCCCCGGGATGATCTACAGTACTCGTTATGGGAGTCCCA
AAAGACAGCTCCAGTTTTTACAGGAATCTGGGCAAATCTGGCCTTCGGGTCTCCTGCCTGG
GGCTTGGAACATGGGTGACCTTCGGGGGGCCAGATCACGGATGAGATGGCAGAGCACCTAA
TGACCTTGGCCTACGATAATGGCATCAACCTGTTCGATACGGCGGAGGTCTACGCTGCTG
GAAAAGCTGAAGTGGTATTAGGGAACATCATTAAGAAGAAGGGATGGAGACGGTCCAGCC
TTGTCATCACCAACCAAGATCTTCTGGGGTGGAAAAGCGGAGACTGAGAGAGGCCTTTCCA
GGAAGCACATAATTGAAGGACTGAAAGCGTCCCTGGAGCGGCTGCAGCTGGAGTACGTGG
ATGTGGTTTTTTGCCAACCGCCCAGACCCCAACACGCCCATGGAAGAGACCGTGCGGGCCA
TGACCCATGTCATCAACCAGGGGATGGCCATGTACTGGGGCACATCACGCTGGAGCTCCA
TGGAGATCATGGAGGCCTACTCGGTGGCTCGGCAGTTCAACCTGATCCCGCCCATCTGCG
AGCAAGCGGAATATCACATGTTCCAGAGGGAGAAGGTGGAGGTCCAGCTGCCAGAGCTGT
TCCACAAGATAGGAGTAGGTGCCATGACCTGGTCCCCTCTGGCGTGCGGCATCGTCTCAG
GGAAGTATGACAGCGGGATCCCACCCTACTCCAGAGCCTCCCTGAAGGGCTACCAGTGGT
TGAAGGACAAGATCCTGAGTGAGGAGGGTCGCCGCCAGCAGGCCAAGCTGAAGGAACTGC
AGGCCATTGCCGAACGCCTGGGCTGCACCCTACCCCAGCTGGCCATAGCCTGGTGCCTGA
GGAATGAGGGTGTGTCAGCTCCGTGCTTCTGGGTGCTTCCAATGCAGAACAACCTTATGGAGA

Is there an ORF for a potential Coding Region?

GGAGCTGGAGGGCCCCCAGGCAACTACACCGTCCACGTACCCAGAGGGGGCTGGGGCCCTCCC
ACCAGAGACCACGCCCTGGTGTGCCTTAGGGGGCCCTGGTTTGTTAGTCTCTGAGTGTGCA
GTTGCTGCACATGGGGGCCCTGGCGCTTGCTGCACCAACTTCCTGTTGGGGCCCGTGGTCCT
TGGAGGCATGCAGTTCAGCAGACAGTGACTCAGCCATCCACCCAACATGCGGAACGTGTC
TCTTCTGCAGGTCCCGGTCCACAGCAGGATTCCCCCTCTGTGAAAAGGCACGCTGATCTG
TCTGGATAAGTGTGGCCGGCCCCCATGTATCCGGAATCAACCACGGGGTCCCCAGCTCGAC
TCTCCCTGCGGCAGACAGGCTCCCCCGGGATGATCTACAGTACTCGTTATGGGAGTCCCA
AAAGACAGCTCCAGTTTTTACAGGAATCTGGGCAAATCTGGCCTTCGGGTCTCCTGCCTGG
GGCTTGGAACATGGGTGACCTTCGGGGGGCCAGATCACGGATGAGATGGCAGAGCACCTAA
TGACCTTGGCCTACGATAATGGCATCAACCTGTTCGATACGGCGGAGGTCTACGCTGCTG
GAAAAGCTGAAGTGGTATTAGGGAACATCATTAAGAAGAAGGGATGGAGACGGTCCAGCC
TTGTCATCACCAACCAAGATCTTCTGGGGTGGAAAAGCGGAGACTGAGAGAGGCCTTTCCA
GGAAGCACATAATTGAAGGACTGAAAGCGTCCCTGGAGCGGCTGCAGCTGGAGTACGTGG
ATGTGGTTTTTTGCCAACCGCCCAGACCCCAACACGCCCATGGAAGAGACCGTGCGGGCCA
TGACCCATGTCATCAACCAGGGGATGGCCATGTACTGGGGCACATCACGCTGGAGCTCCA
TGGAGATCATGGAGGCCTACTCGGTGGCTCGGCAGTTCAACCTGATCCCGCCCATCTGCG
AGCAAGCGGAATATCACATGTTCCAGAGGGAGAAGGTGGAGGTCCAGCTGCCAGAGCTGT
TCCACAAGATAGGAGTAGGTGCCATGACCTGGTCCCCTCTGGCGTGCGGCATCGTCTCAG
GGAAGTATGACAGCGGGATCCCACCCTACTCCAGAGCCTCCCTGAAGGGCTACCAGTGGT
TGAAGGACAAGATCCTGAGTGAGGAGGGTCGCCGCCAGCAGGCCAAGCTGAAGGAACTGC
AGGCCATTGCCGAACGCCTGGGCTGCACCCTACCCCAGCTGGCCATAGCCTGGTGCCTGA
GGAATGAGGGTGTGTCAGCTCCGTGCTTCTGGGTGCTTCCAATGCAGAACAACCTTATGGAGA

Is there an ORF for a potential Coding Region?

GGAGCTGGAGGGCCCCCAGGCAACTACACCGTCCACGTACCCAGAGGGGGCTGGGGCCCTCCC
ACCAGAGACCACGCCCTGGTGTGCCTTAGGGGGCCCTGGTTTGTTAGTCTCTGAGTGTGCA
GTTGCTGCAC**ATGGGGCCCTGGCGCTTGCTGCACCAACTTCCTGTTGGGGCCCGTGGTCCT**
TGGAGGCATGCAGTTCAGCAGACAGTGACTCAGCCATCCACCCCAACATGCGGAACGTGTC
TCTTCTGCAGGTCCCGGTCCACAGCAGGATTCCCCCTCTGTGAAAAGGCACGCTGATCTG
TCTGGATAAGTGTGGCCGGCCCCCATGTATCCGGAATCAACCACGGGGTCCCCAGCTCGAC
TCTCCCTGCGGCAGACAGGCTCCCCCGGGATGATCTACAGTACTCGTTATGGGAGTCCCA
AAAGACAGCTCCAGTTTTTACAGGAATCTGGGCAAATCTGGCCTTCGGGTCTCCTGCCTGG
GGCTTGGAACATGGGTGACCTTCGGGGGGCCAGATCACGGATGAGATGGCAGAGCACCTAA
TGACCTTGGCCTACGATAATGGCATCAACCTGTTCGATACGGCGGAGGTCTACGCTGCTG
GAAAAGCTGAAGTGGTATTAGGGAACATCATTAAGAAGAAGGGATGGAGACGGTCCAGCC
TTGTCATCACCAACCAAGATCTTCTGGGGTGGAAAAGCGGAGACTGAGAGAGGCCTTTCCA
GGAAGCACATAATTGAAGGACTGAAAGCGTCCCTGGAGCGGCTGCAGCTGGAGTACGTGG
ATGTGGTTTTTTGCCAACCGCCCAGACCCCAACACGCCCATGGAAGAGACCGTGCGGGCCA
TGACCCATGTCATCAACCAGGGGATGGCCATGTACTGGGGCACATCACGCTGGAGCTCCA
TGGAGATCATGGAGGCCTACTCGGTGGCTCGGCAGTTCAACCTGATCCCGCCCATCTGCG
AGCAAGCGGAATATCACATGTTCCAGAGGGAGAAGGTGGAGGTCCAGCTGCCAGAGCTGT
TCCACAAGATAGGAGTAGGTGCCATGACCTGGTCCCCTCTGGCGTGCGGCATCGTCTCAG
GGAAGTATGACAGCGGGATCCCACCCTACTCCAGAGCCTCCCTGAAGGGCTACCAGTGGT
TGAAGGACAAGATCCTGAGTGAGGAGGGTCGCCGCCAGCAGGCCAAGCTGAAGGAACTGC
AGGCCATTGCCGAACGCCTGGGCTGCACCCTACCCCAGCTGGCCATAGCCTGGTGCCTGA
GGAATGAGGGTGTGTCAGCTCCGTGCTTCTGGGTGCTTCCAATGCAGAACAACCTTATGGAGA

ORF finder searches for open reading frames (ORFs) in the DNA sequence you enter. The program returns the range of each ORF, along with its protein translation. Use ORF finder to search newly sequenced DNA for potential protein encoding segments, verify predicted protein using newly developed SMART BLAST or regular BLASTP.

Examples (click to set values, then click Submit button) :

- 
- ORFinder

Enter accession number, gi, or nucleotide sequence in FASTA format:

From: **To:**

ORFfinder finds all open reading frames and provides translations

Secure <https://www.ncbi.nlm.nih.gov/orffinder/>

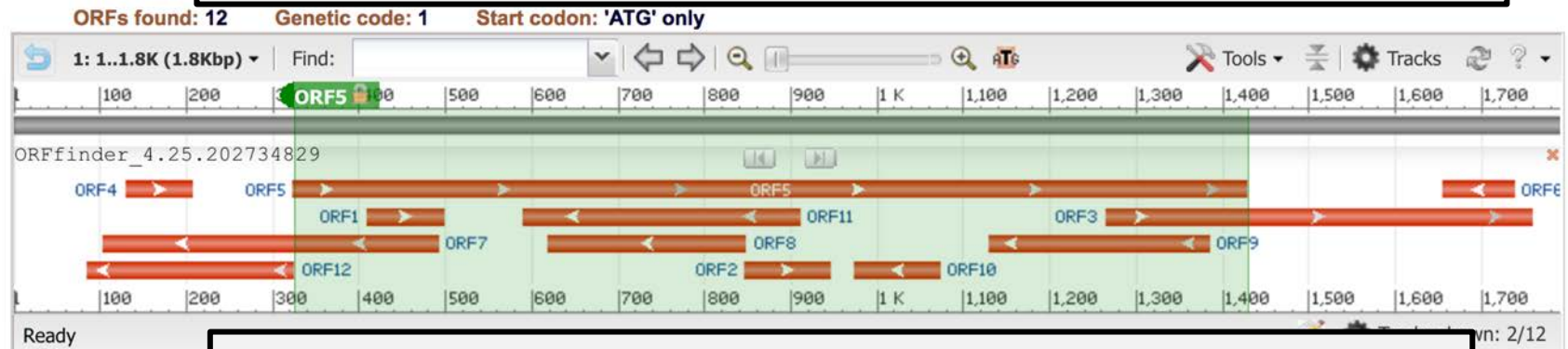
NCBI Resources How To Sign in to NCBI

ORFfinder PubMed Search

Open Reading Frame Viewer

Sequence

ORFs can appear in random sequence – so further analysis is required



Predict coding vs. non-coding ORFs: <http://TransDecoder.github.io>

ORF5 (367 aa)

Display ORF as...

Mark

```
>lcl|ORF5
MYPESTTGSPARLSLRQTGSPGMIYSTRYGSPKRQLQFYR
NLGKSGLRVSCGLGTWVTFGGQITDEMAEHLMTLAYDNG
INLFDTAEVYAAGKAEEVVLGNIIKKKGWRRSSLVITTKIF
WGGKAETERGLSRKHIEGLKASLERLQLEYVDVVFANRP
DPNTPMEETVRAMTHVINQGMAMYWGTSRWSSMEIMEAYS
VARQFNLIPIICEQAETHMFQREKVEVQLPELFHKIGVGA
MTWSPLACGIVSGKYDSGIPPYSRASLKGQWLKDKILSE
EGRRQQAQKLKELQAIARLGCTLPQLAIWCLRNQGVSSV
LLGASNAEQLMENIGAIQVLPKLSSSIVHEIDSILGNKPY
SKKDYRS
```

Add six-frame translation track

Mark subset...

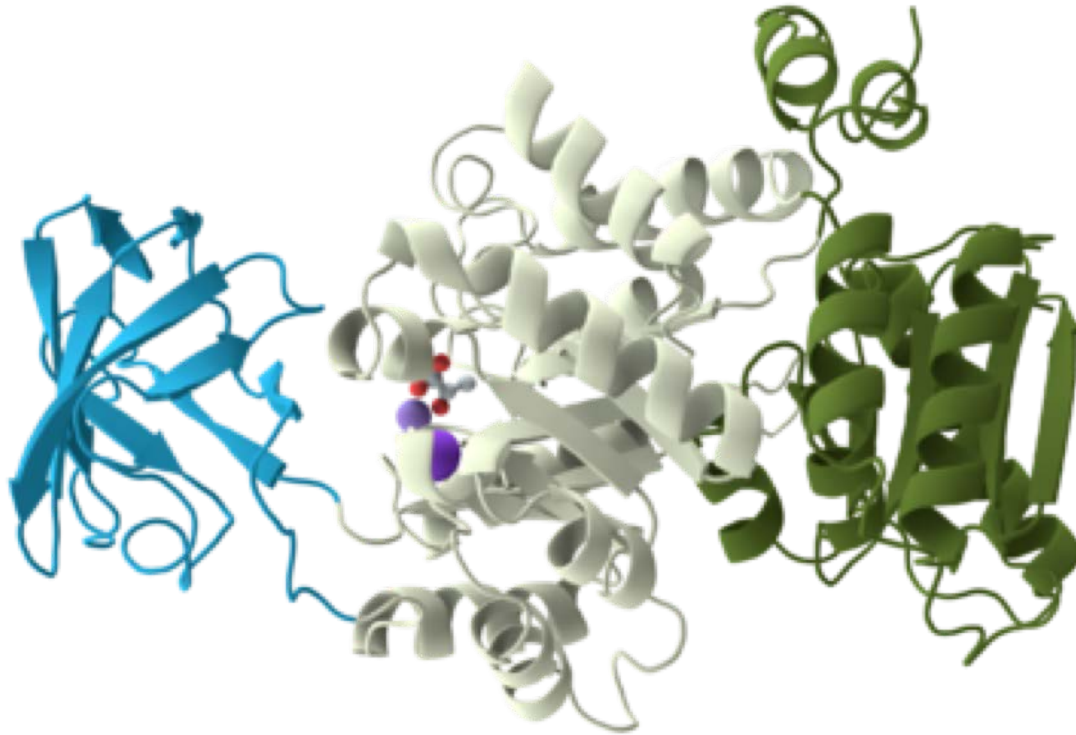
Marked: 0

Download marked set

as Protein FA

Label	Strand	Frame	Start	Stop	Length (nt)
ORF5	+	3	324	1427	1104 36
ORF3	+	1	1264	1758	495 16
ORF7	-	1	492	103	390 12
ORF11	-	3	910	590	321 10
ORF9	-	3	1384	1130	255 8
ORF12	-	3	325	86	240 7
ORF8	-	2	848	618	231 7

Can we recognize functional domains in putative coding regions?



Hints at substrate binding or catalytic activity

DNA, RNA, calcium,
phosphate, etc.

Glycosylase, methylase, kinase, nuclease,
lipase, protease, etc.

Search the Pfam library of HMMs to identify potential functional domains

← → ↺ pfam.xfam.org

EMBL-EBI



HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Pfam

keyword search

Pfam 31.0 (March 2017, 16712 entries)

The Pfam database is a large collection of protein families, each represented by **multiple sequence alignments** and **hidden Markov models (HMMs)**. [More...](#)

QUICK LINKS

SEQUENCE SEARCH

VIEW A PFAM ENTRY

VIEW A CLAN

VIEW A SEQUENCE

VIEW A STRUCTURE

KEYWORD SEARCH

JUMP TO

ANALYZE YOUR PROTEIN SEQUENCE FOR PFAM MATCHES

Paste your protein sequence here to find matching Pfam entries.

Go

Example

METGGRARTGTPQPAAPGVWRARPAGGGGGGASSWLLDGNWLLCYGFLY
LALYAQVVSQSKPCERTGSCFSGRCVNSTCLCDPGWVGDCQHCQGRFKLT
EPSGYLTDGPINYYKYTKCTWLIEGYPNAVLRLRFNFHATECSWDHMYVY
DGD SIYAPLIAVL SGLVPEIRGNETVPEVVTTSGYALLHFFSDAAYNLT
GFNIFYSINSCPNNGSGHKGCTTSVSVSPQVYCECDKYWKGEACDIPYCK
ANCGSPDHGYCDLTGEKLCVCNDSWQGPDCSLNVPSTESYWLPNVKPFS
PSVGRASHKAVLHGKFMWVIGGYTFNYSSFQMVNLNYLESSIWNVGTPSR
GPLQRYGHSLALYQENIFMYGGRIETNDGNVTDELWVFNIHSQSWSTKTP
TVLGHGQQYAVEGHS AHIMELDSRDVVMIIIFGYSAIYGYTSSIQEYHIS
SNTWLVPETKGAIVQGGYGHSTSVYDEITKSIYVHGGYKALPGNKYGLVDD
LYKYEVTNTKTWILKESGFARYLHSAVLINGAMLIFGGNTHNDTSLNGA
KCFSAFLAYDIACDEWKILPKPNLHRDVNRFHGS AVVINGSMYIFGGFS
SVLLNDILVYKPPNCKAFRDEELCKNAGPGIKCVWNKNHCESWESGNTN
ILRAKCPPKTAASDDRCYRYADCASCTANTGCQWCDKCKCISANSNCM
SVKNYTKCHVRNEQICNKLTSCKSCSLNLCQWDQRQCEQALPAHLGCE
GWSHIGDACLRVNSSRENYDNAKLYCYNLSGNLASLTTSKEVEFLDEIQ
KYTQQKVPSPWGLRKINISYWGWEDEMSPTNTTLQWLPGEPNDSGFCAYL
ERAAVAGLKANPCTSMANGLVCEKPVVSPNQNARPCPKPCSLRTSCSNCT
SNGMECMWCSSTKRCVDSNAYIISFPYGCLEWQTATCSPQNCSSGLRTCG
QCLEQPGCGWCNDPSNTGRGHCIEGSSRGPMKLIGHMHSEMVLDTNLCPK
EKNYEWSFIQCPCACQCNHSTCINNVCQCKNLTGKQCQDCMPGYGDD
PTNGGQCTACTCSGHANICHLHTGKCFCTTKGIKGDQCQLCDSENRYVGN
PLRGTCYYSLLIDYQFTFSLLQEDDRHHTAINFIANPEQSNKNDISINA
SNNFNLTITWSVGSTAGTISGEETSIVSKNNIKEYRDSFSYEKFNRSP
NITFYVYVSNSFWPIKIQIAFSQHNTIMDLVQFFVTFSCFLSLLVAAV
VWKIKQTCWASRRREQLLRERQQMASRPFASVDVALEVGAEQTEFLRGPL
EGAPKPIAIEPCAGNRAAVLTVFLCLPRGSSGAPPPGQSGLAIASALIDI
SQQKASDSKDKTSGVNRNRKHLSTRQGTCV

This search will use an E-value of 1.0. You can set your own search parameters and perform a range of other searches [here](#).

Example Pfam report illustrating modular domain architecture

← → ↺ pfam.xfam.org/search/sequence ☆ a ABP JB PDF PDF PDF PDF PDF PDF PDF PDF PDF PDF PDF PDF PDF PDF PDF PDF

EMBL-EBI  HOME | SEARCH | BROWSE | FTP | HELP | ABOUT  keyword search Go

Sequence search results

[Show](#) the detailed description of this results page.

We found **9** Pfam-A matches to your search sequence (**all** significant)



[Show](#) the search options and sequence that you submitted.

[Return](#) to the search form to look for Pfam domains on a new sequence.

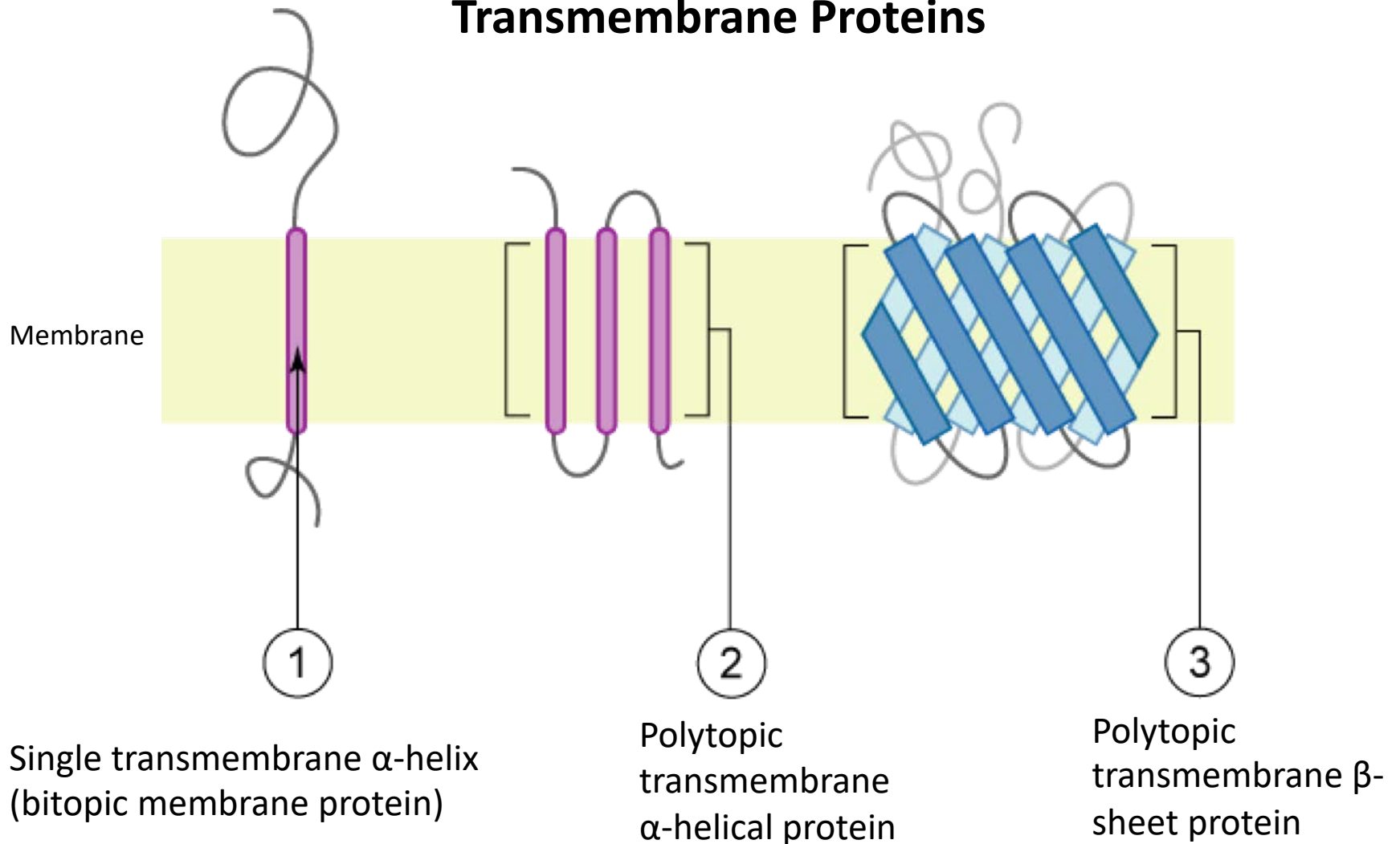
Significant Pfam-A Matches

[Show](#) or [hide](#) all alignments.

Family	Description	Entry type	Clan	Envelope		Alignment		HMM		HMM length	Bit score	E-value	Predicted active sites	Show/hide alignment
				Start	End	Start	End	From	To					
CUB	CUB domain	Domain	CL0164	93	206	93	206	1	110	110	42.2	7.7e-11	n/a	Show
EGF_2	EGF-like domain	Domain	CL0001	249	280	249	280	1	32	32	22.5	0.0001	n/a	Show
Kelch_5	Kelch motif	Repeat	CL0186	351	393	352	392	2	41	42	33.7	2.2e-08	n/a	Show
Kelch_4	Galactose oxidase, central domain	Repeat	CL0186	466	518	468	514	3	44	49	20.6	0.0003	n/a	Show
Kelch_1	Kelch motif	Repeat	CL0186	520	574	520	573	1	45	46	20.0	0.00033	n/a	Show
Kelch_5	Kelch motif	Repeat	CL0186	579	614	581	613	5	40	42	25.3	9.7e-06	n/a	Show
Lectin_C	Lectin C-type domain	Domain	CL0056	765	874	766	874	2	108	108	70.2	2e-19	n/a	Show
PSI	Plexin repeat	Family	CL0630	889	939	890	938	2	50	51	27.8	2.5e-06	n/a	Show
PSI	Plexin repeat	Family	CL0630	942	1012	942	1012	1	51	51	50.0	2.9e-13	n/a	Show

Comments or questions on the site? Send a mail to pfam-help@ebi.ac.uk.
European Molecular Biology Laboratory

Transmembrane Proteins



Using TMHMM to identify putative transmembrane proteins

[←](#) [→](#) [↻](#) [www.cbs.dtu.dk/services/TMHMM/](#) [☆](#) [a](#) [ABP](#) [JB](#) [📄](#) [🔍](#) [📺](#) [🔧](#) [🔑](#) [🔗](#) [🔗](#) [🔗](#) [🔗](#)

CENTER FOR BIOLOGICAL SEQUENCE ANALYSIS CBS

EVENTS

NEWS

RESEARCH GROUPS

CBS PREDICTION SERVERS

CBS DATA SETS

PUBLICATIONS

EDUCATION

STAFF

CONTACT

ABOUT CBS

INTERNAL

CBS BIOINFORMATICS TOOLS

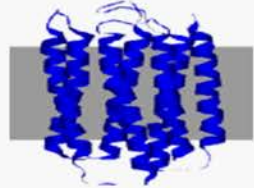
CBS COURSES

OTHER BIOINFORMATICS LINKS

[CBS](#) >> [CBS Prediction Servers](#) >> [TMHMM](#)

TMHMM Server v. 2.0

Prediction of transmembrane helices in proteins



Instructions

SUBMISSION

Submission of a local file in **FASTA** format (HTML 3.0 or higher)

Choose File No file chosen

OR by pasting sequence(s) in **FASTA** format:

MEILCEDNTSLSSIPNSLMQVDGDSGLYRNDNFNSRDANSSDASNWTIDGENRTNLSFEG
YLPPTCLSLHLQEKNSWALLTAVVIITAGNIVIMAVSLEKKLQATNYFLMSLAIDMLL
GFLVMPVSMLTILYGYRWPLPSKLCVWYLDVLFSTASIMHLCAISLDRYVAIQNPIHHSR
FNSRTKAFLKIIAVWTISVGVSMPIPVFGLQDDSKVFKQGSCLLADDNFVLIGSFVAFFIPLTI
MVITYFLTIKSLQKEATLCVSDLSTRAKLASFSL

Output format:

☒ Extensive, with graphics

☐ Extensive, no graphics

☐ One line per protein

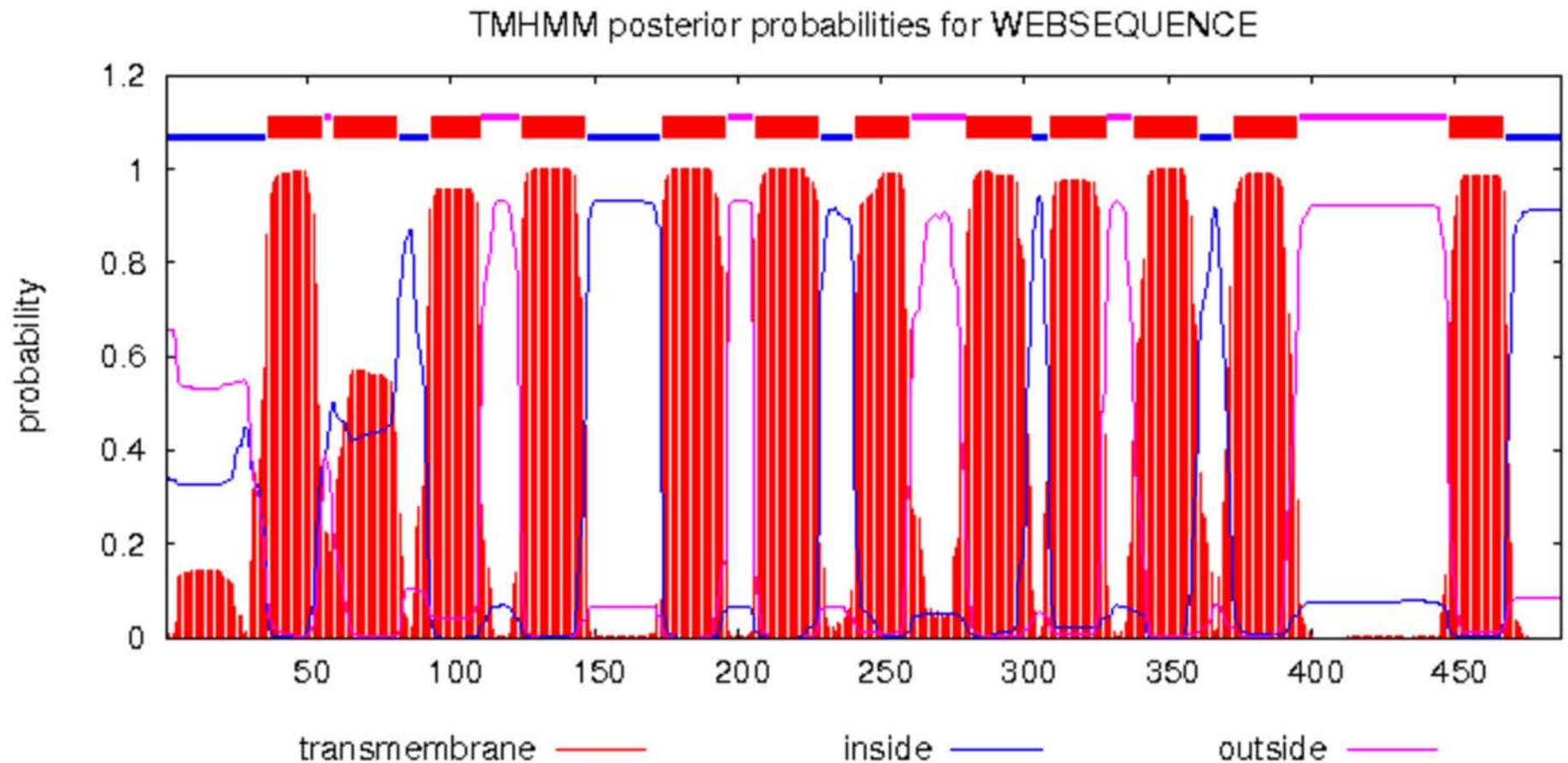
Other options:

☐ Use old model (version 1)

Submit Clear

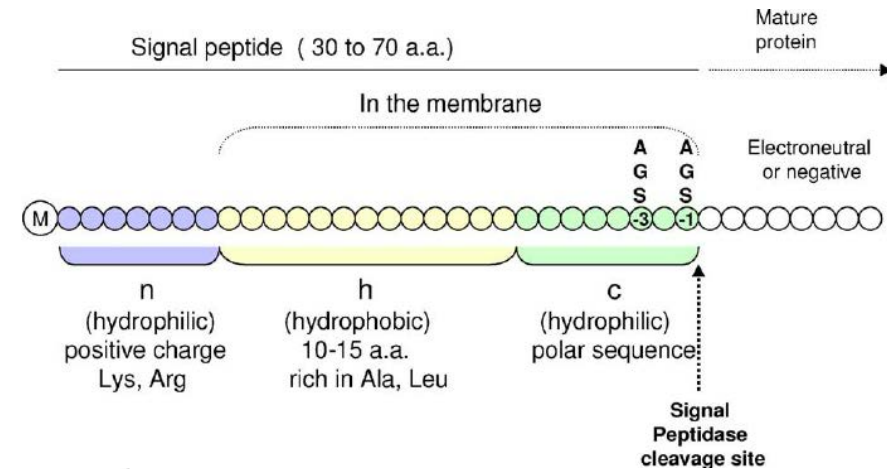
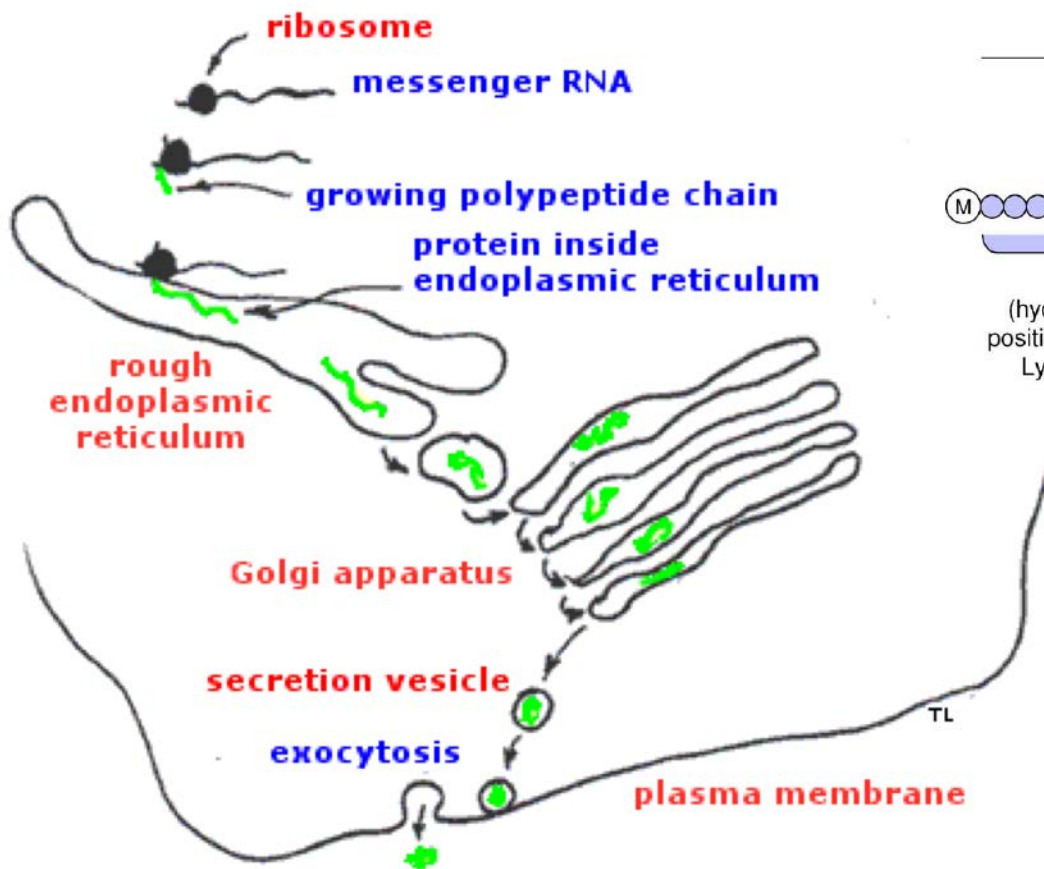
CENTER FOR BIOLOGICAL SEQUENCE ANALYSIS || TECHNICAL UNIVERSITY OF DENMARK DTU

Trans-membrane Domains via TmHMM



Topology=i36-55o59-81i93-110o125-147i174-196o206-228i241-260o280-302i309-328o338-360i373-395o448-467i

Predicting Secreted Proteins



(from: Vaccine 23(15):1770-8)

(from: <https://courses.washington.edu/coni/cell/secretion.htm>)

SignalP: Prediction of N-terminal signal peptides (predict secreted proteins)

← → ↻ ⓘ www.cbs.dtu.dk/services/SignalP/

CENTER FOR BIOLOGICAL SEQUENCE ANALYSIS CBS

EVENTS

STAFF

NEWS

CONTACT

RESEARCH GROUPS

ABOUT CBS

CBS PREDICTION SERVERS

INTERNAL

CBS DATA SETS

BIOINFORMATICS TOOLS

PUBLICATIONS

CBS COURSES

EDUCATION

OTHER BIOINFORMATICS LINKS

CBS >> [CBS Prediction Servers](#) >> SignalP

SignalP 4.1 Server

SignalP 4.1 server predicts the presence and location of signal peptide cleavage sites in amino acid sequences from different organisms: Gram-positive prokaryotes, Gram-negative prokaryotes, and eukaryotes. The method incorporates a prediction of cleavage sites and a signal peptide/non-signal peptide prediction based on a combination of several artificial neural networks.

View the [version history](#) of this server. All the previous versions are available on line, for comparison and reference.

NEW: The portable version of SignalP 4.1, previously only available for Mac (Darwin), Linux, and IRIX, is now also available for Windows systems. Academic users: select the "CYGWIN" option at the [download page](#). [Cygwin](#) or [MobaXterm](#) is required to install SignalP under Windows. For details, read the [installation instructions](#).

FAQ

Article abstracts

Instructions

Output format

Performance

Data

SUBMISSION

Paste a single amino acid sequence or several sequences in **FASTA** format into the field below:

MHPAVFLSLPDLRCSLLLVTVWVFTPTVTEITSLDTENIDEILNNADVALVNFYADWCRFSQMLHPHFEEASDVKEEFPNENQVVFARVDCDQHSIDIAQRYRISKYPTLKLFRNGMMM
KREYRGQRSVKALADYIRQKQSDPIQEIRDLAEITTLDRSKRNIIGYFEQKQSDNYRVFERVANILHDDCAFLSAFGDVSKPERYSGDNIYKPPGHSAPDMVYLGAMTNFDVTVNWIQ
DKCVPLVREITFENGEELEGLPFLILFHMKEDTESLEIFQNEVARQLISEKGTINFLHADCDKFRHPLLHIQKTPADCPVIAIDSFHMYVFGDFKDVLPGLKQFVFDLHSGKLHREF
HHGPDPTDTAPGEQAQDVASSPPESFQKLAPSEYRYTLLRDRDEL

Submit a file in **FASTA** format directly from your local disk:

Choose File No file chosen

Organism group [\(explain\)](#)

☒ Eukaryotes
☐ Gram-negative bacteria
☐ Gram-positive bacteria

D-cutoff values [\(explain\)](#)

☒ Default (optimized for correlation)
☐ Sensitive (reproduce SignalP 3.0's sensitivity)
☐ User defined:

D-cutoff for SignalP-noTM networks
 D-cutoff for SignalP-TM networks

Graphics output [\(explain\)](#)

☐ No graphics
☒ PNG (inline)
☐ PNG (inline) and EPS (as links)

Output format [\(explain\)](#)

☒ Standard
☐ Short (no graphics)
☐ Long
☐ All - SignalP-noTM and SignalP-TM output (no graphics)

Method [\(explain\)](#)

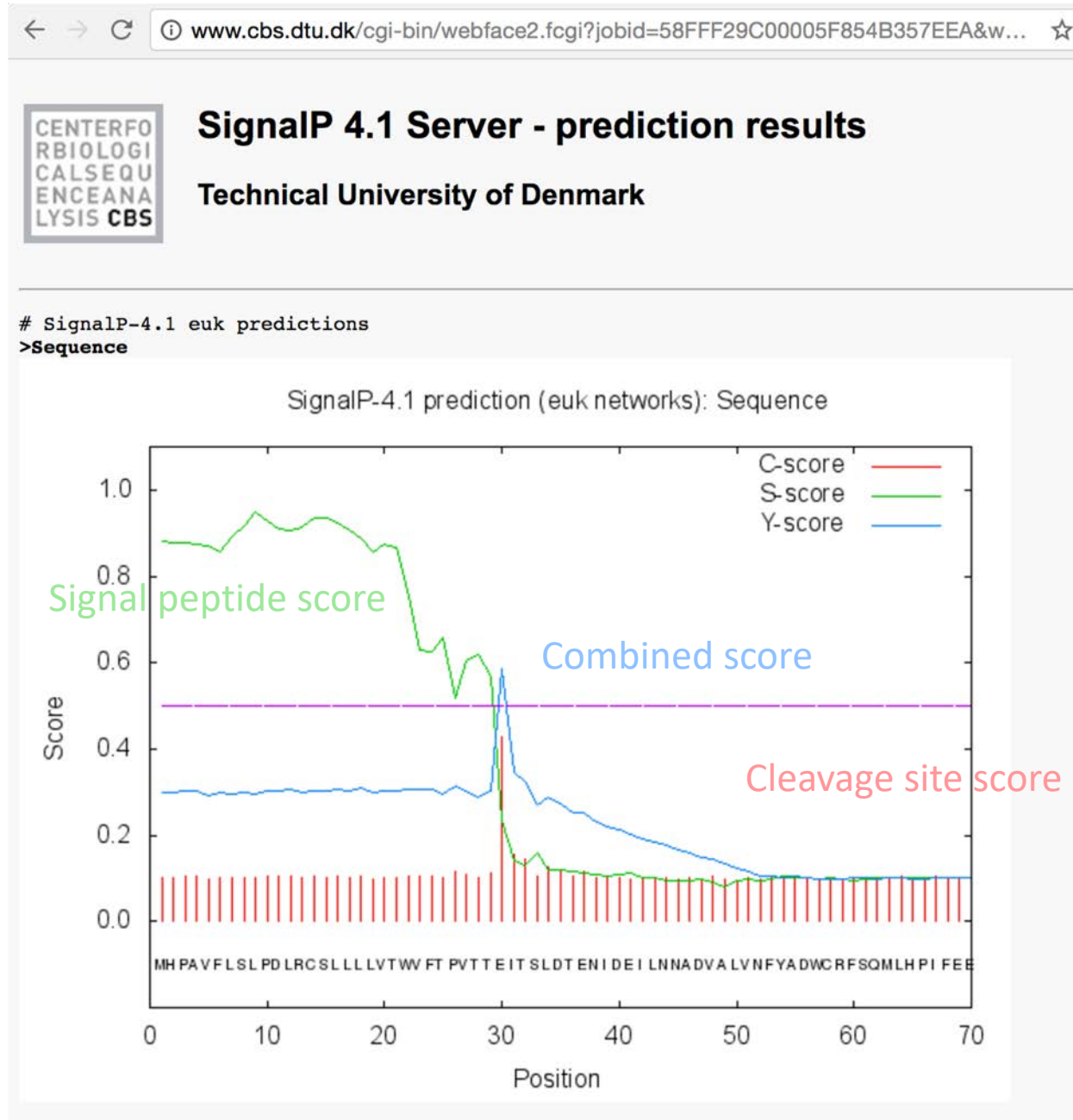
☒ Input sequences may include TM regions
☐ Input sequences do not include TM regions

Positional limits [\(explain\)](#)

☐ Minimal predicted signal peptide length. *Default: 10*
N-terminal truncation of input sequence (0 means no truncation).
Default: Truncate sequence to a length of 70 aa
☐

TECHNICAL UNIVERSITY OF DENMARK DTU

Example SignalP predicted signal peptide



Transcriptome-scale functional annotation using Trinotate



Trinotate: Transcriptome Functional Annotation and Analysis

Trinotate

TransDecoder



TMHMM

SignalP

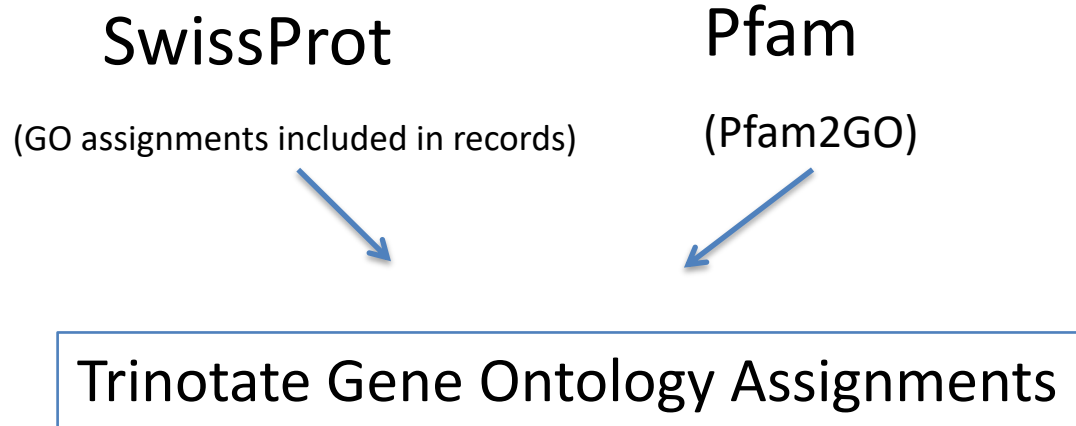


eggNOG
version 3.0



RNA-Seq ➡ Trinity ➡ Transcripts/Proteins ➡ Functional Data ➡ Discovery

GoSeq for Functional Enrichment Testing



METHOD | OPEN ACCESS

Gene ontology analysis for RNA-seq: accounting for selection bias

Matthew D Young, Matthew J Wakefield, Gordon K Smyth and Alicia Oshlack 

Genome Biology 2010 11:R14 | DOI: 10.1186/gb-2010-11-2-r14 | © Young et al.; licensee BioMed Central Ltd. 2010

Gene ontology functional enrichment

	(+) Differentially Expressed	(-) Not Differentially Expressed	Totals
+ Gene Ontology	50	200	250
- Gene Ontology	1950	17800	19750
Totals	2000	18000	20000

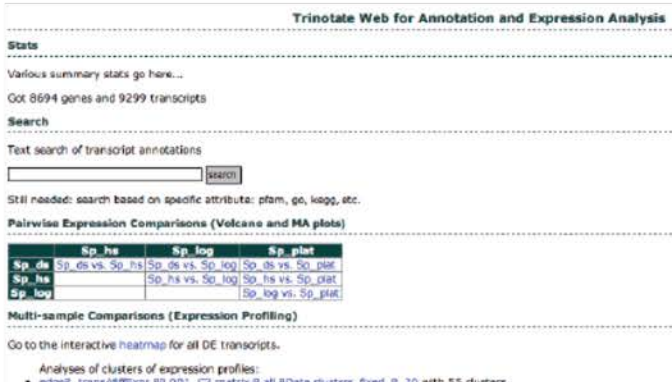
	drawn	not drawn	total
green marbles	k	$K - k$	K
red marbles	$n - k$	$N + k - n - K$	$N - K$
total	n	$N - n$	N

The probability of drawing exactly k green marbles can be calculated by the formula

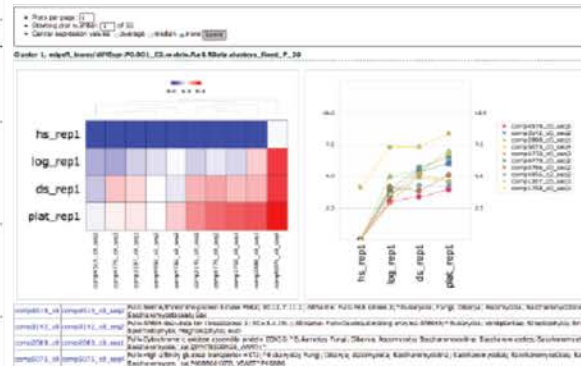
$$P(X = k) = f(k; N, K, n) = \frac{\binom{K}{k} \binom{N-K}{n-k}}{\binom{N}{n}}.$$

Trinotate Web for Interactive Analysis

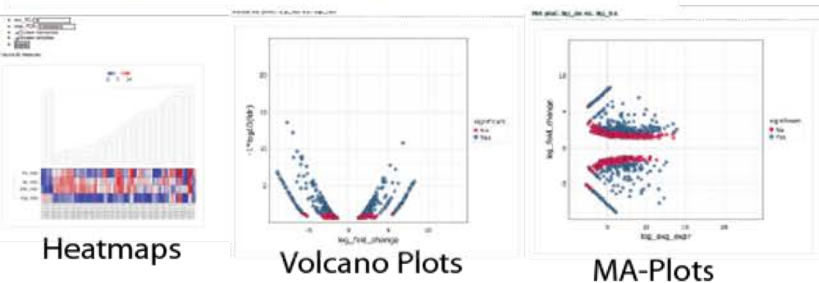
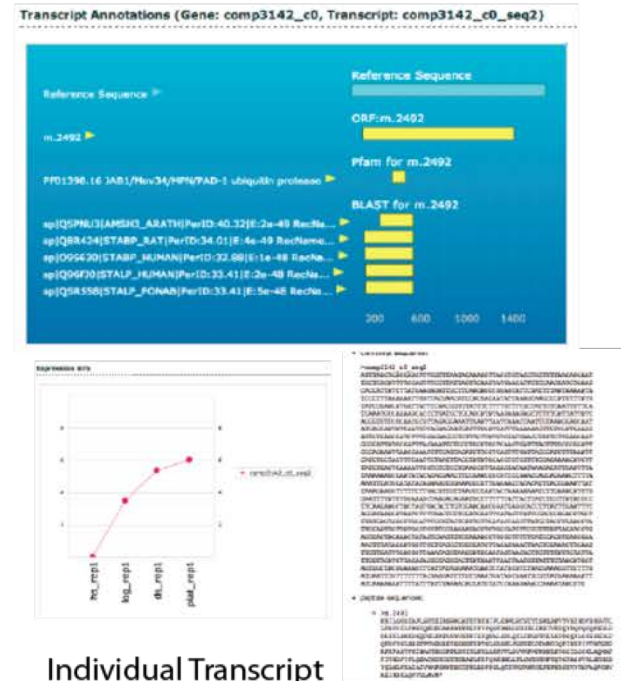
TrinotateWeb Entry Point



Clustered Expression Profiles



Transcript/Protein Annotation Report Blast Hits, Pfam Domains, etc.



*Very Early Release and
Just Scratching the Surface*

Individual Transcript Expression Profiles

Transcript and Protein Sequences

Deciphering the Cell Circuitry of Limb Regeneration Via Single Cell Transcriptome Studies



Work done in collaboration with
Jessica Whited's lab



Axolotl (*Ambystoma mexicanum*) Transcriptomics

Axolotl "water monster", aka Mexican salamander or Mexican walking fish.

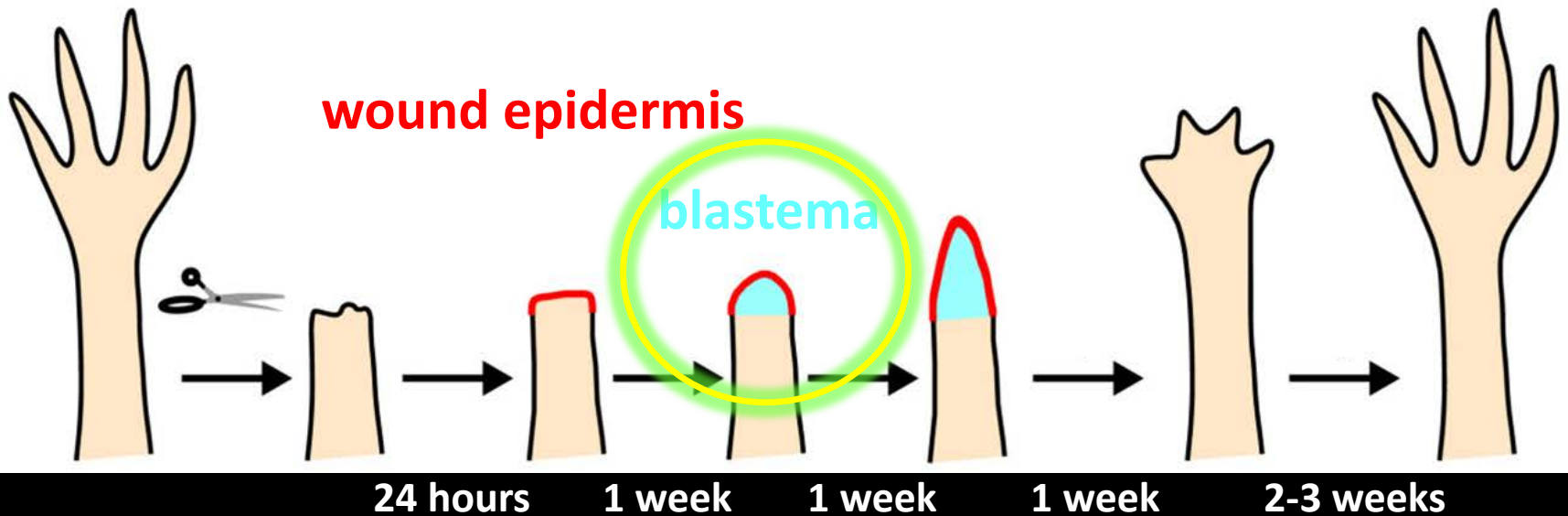
- Model for vertebrate studies of tissue regeneration
- Short generation time
- Can fully regenerate a severed limb in just weeks.
- Genome estimated at ~30 Gb (not yet sequenced)



Google Anonymous Axolotl Icon



Key morphological steps during limb regeneration





1. Building a reference Axolotl transcriptome

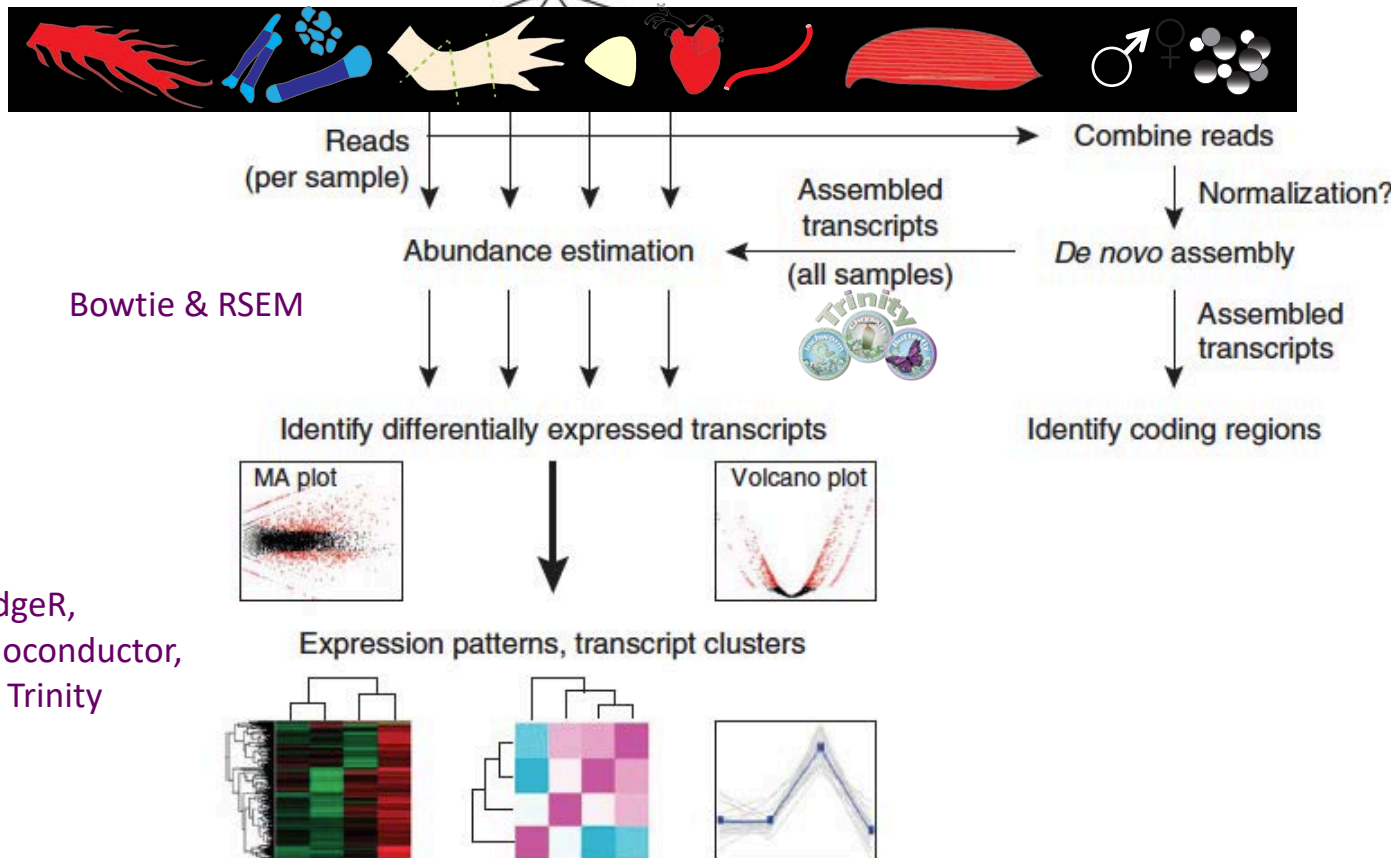


limb tissues and select
other tissues with
biological replicates

1.3 billion of
100 bp paired-end
Illumina reads



Framework for De novo Transcriptome Assembly and Analysis



1.3 Billion
Total Reads

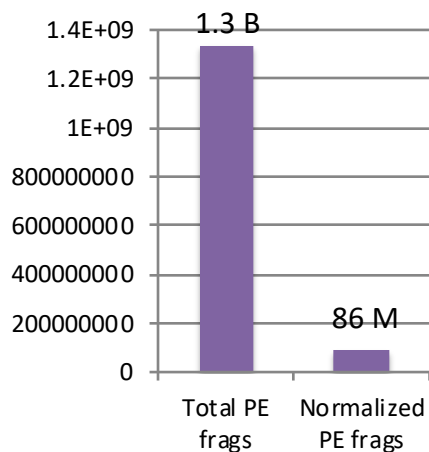
86 Million
Normalized Reads

EdgeR,
Bioconductor,
& Trinity



Axolotl Transcriptome De novo Assembly Statistics And Quality Assessment

In silico Normalization

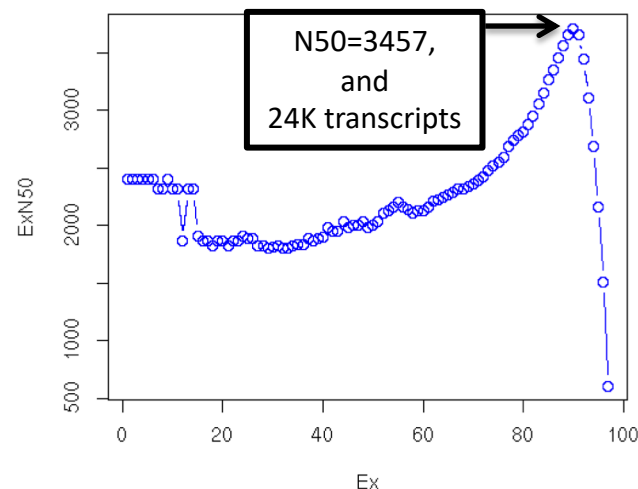


Counts of Transcripts

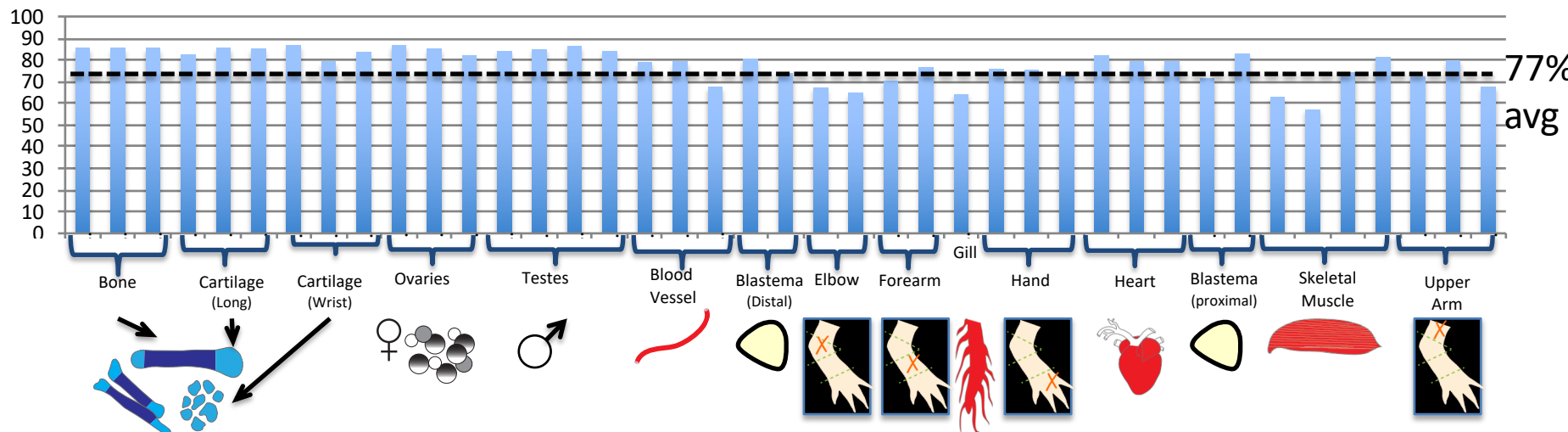
Trinity contigs (transcripts)	1,554,055
Trinity components (genes)	1,388,798

Min. length 200 bases

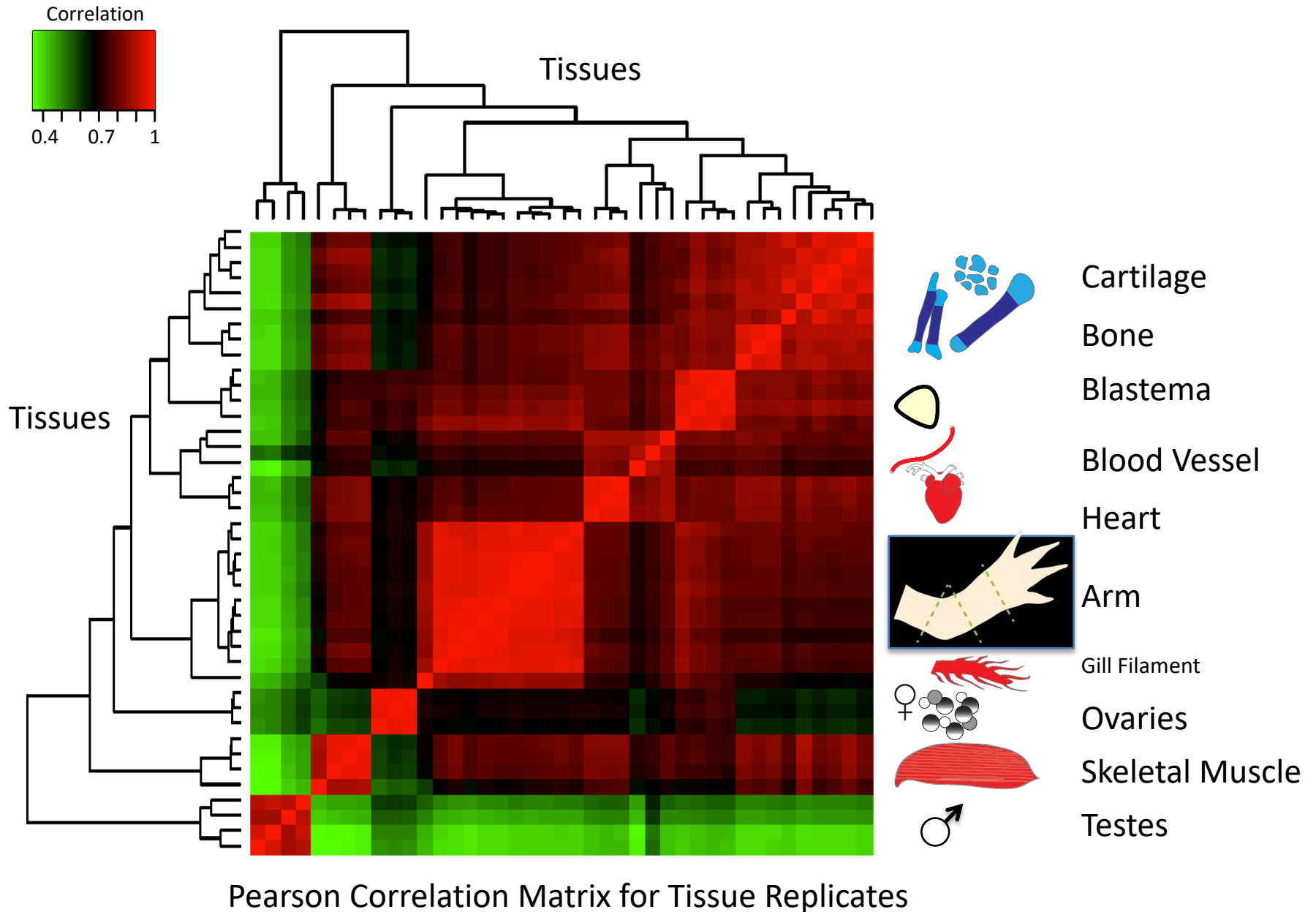
ExN50 looks good!



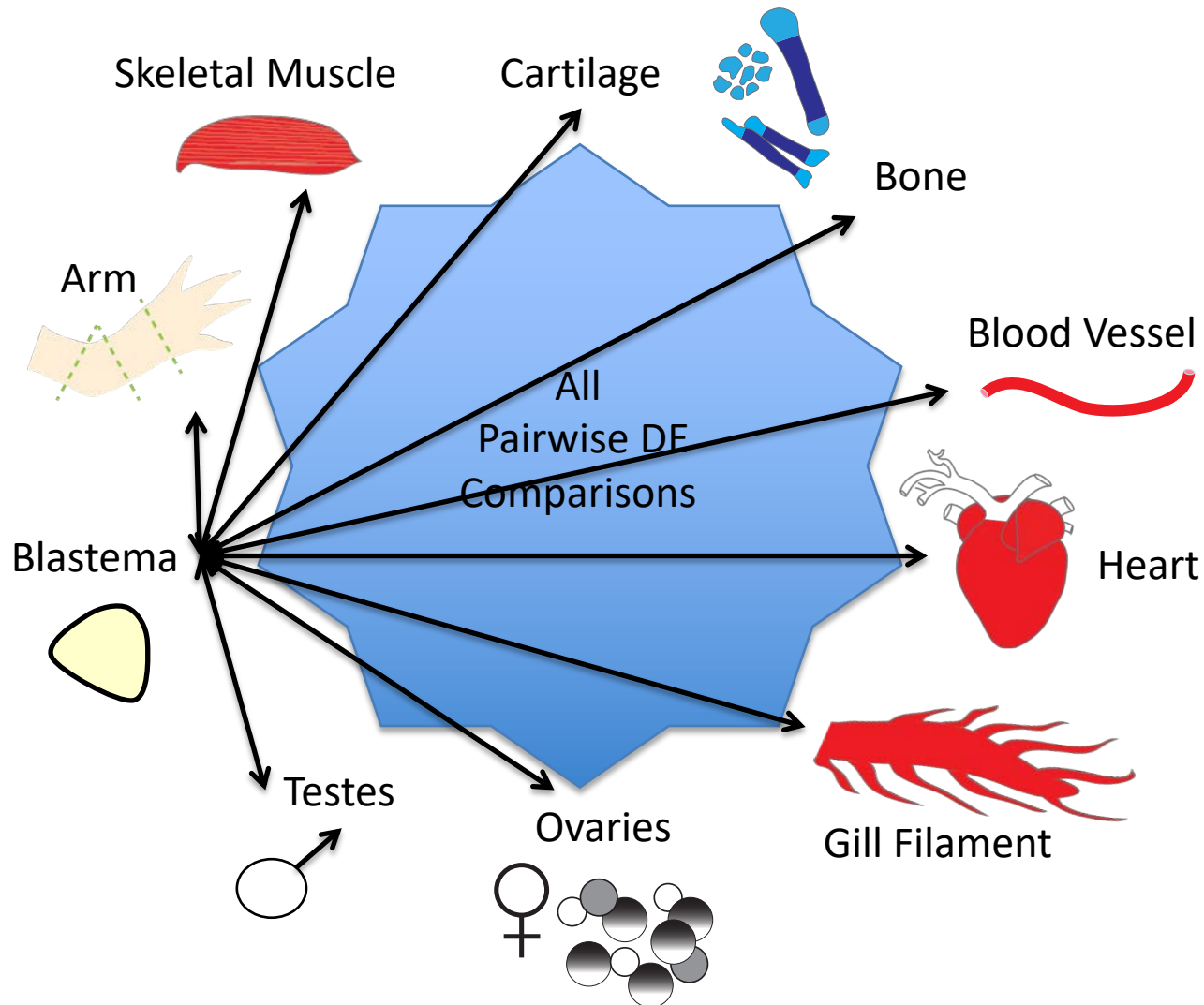
Percent of Non-normalized Fragments Mapping as Properly Paired to Transcriptome



Biological Replicates Cluster According to Sample

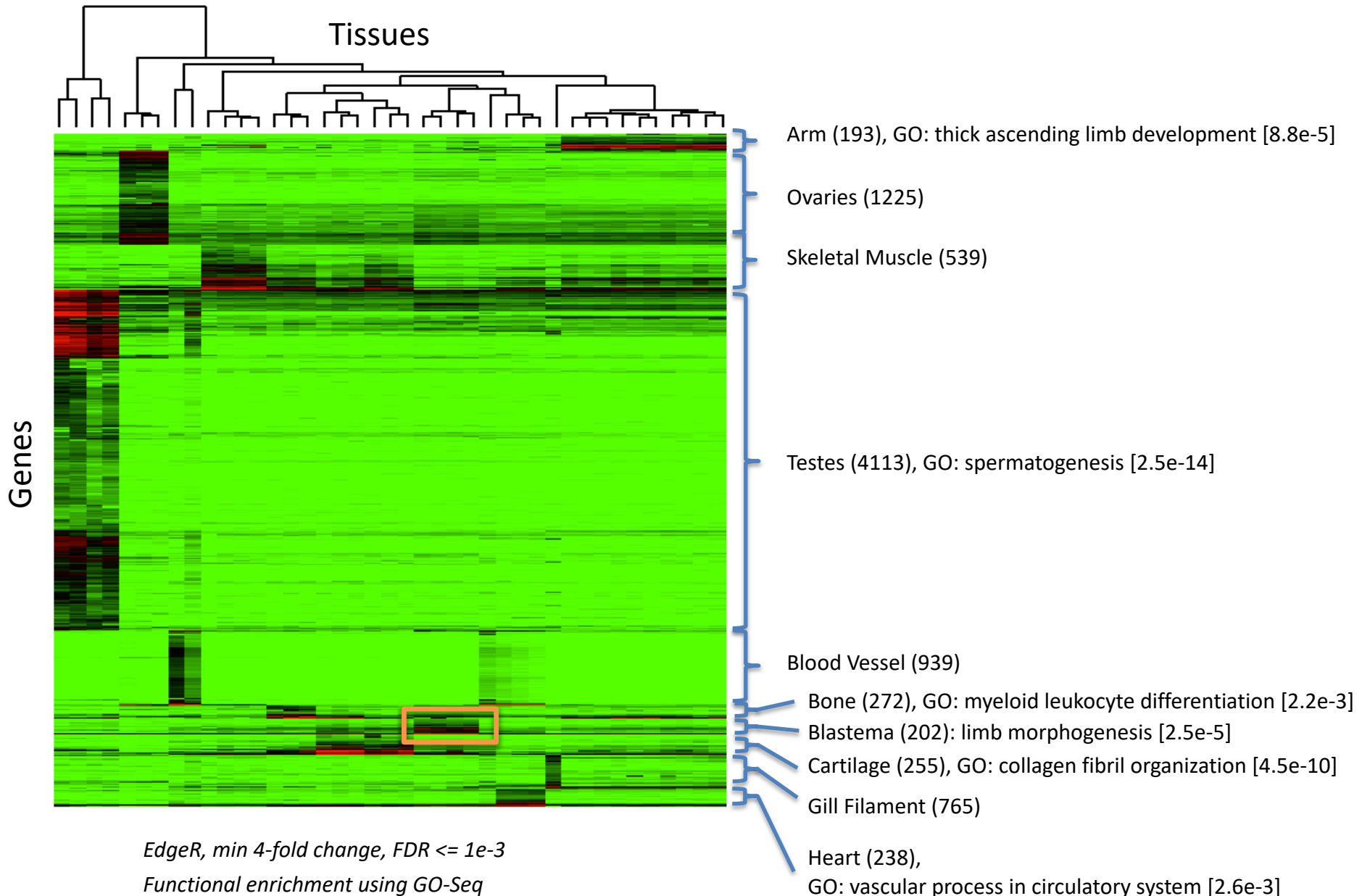


2. Identification of Tissue-enriched Expression

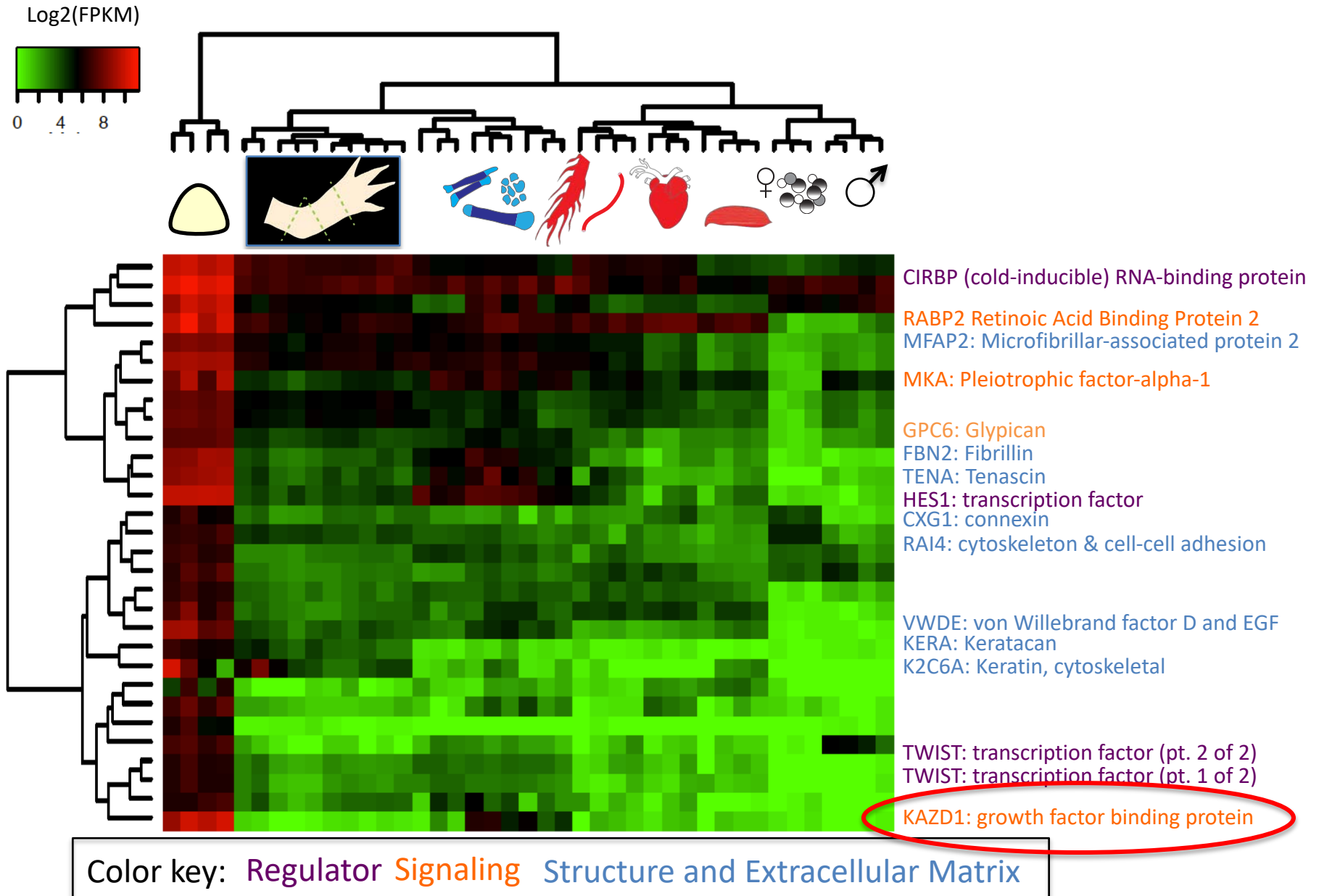


EdgeR, min 4-fold change, FDR $\leq 1e-3$

Identification of Tissue-enriched Gene Expression

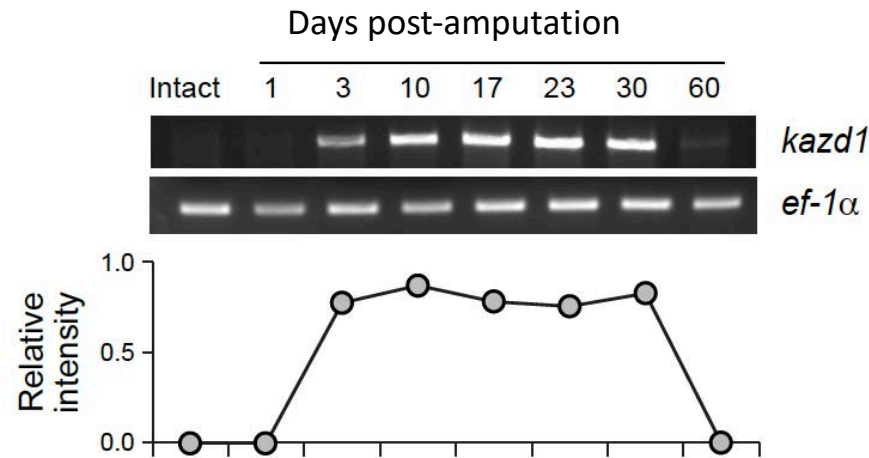


Most Highly Expressed Blastema-enriched Genes

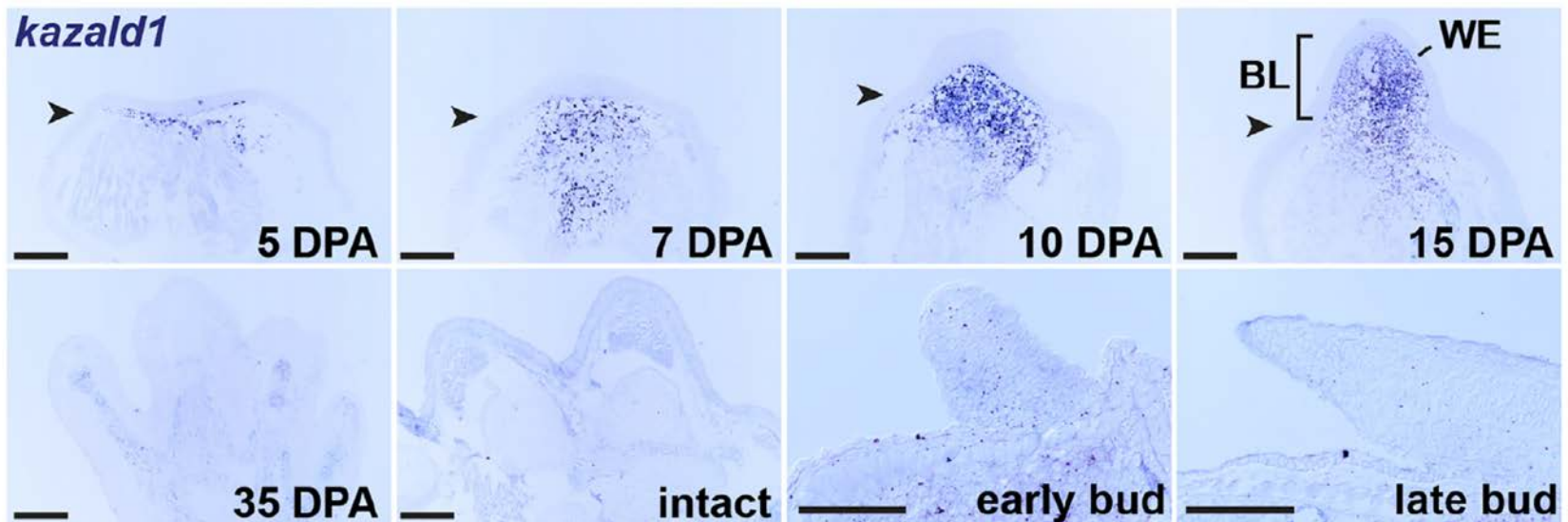


Functional Characterization of Blastema-enriched KAZD1

RT-PCR Timecourse of Kazald1 Expression



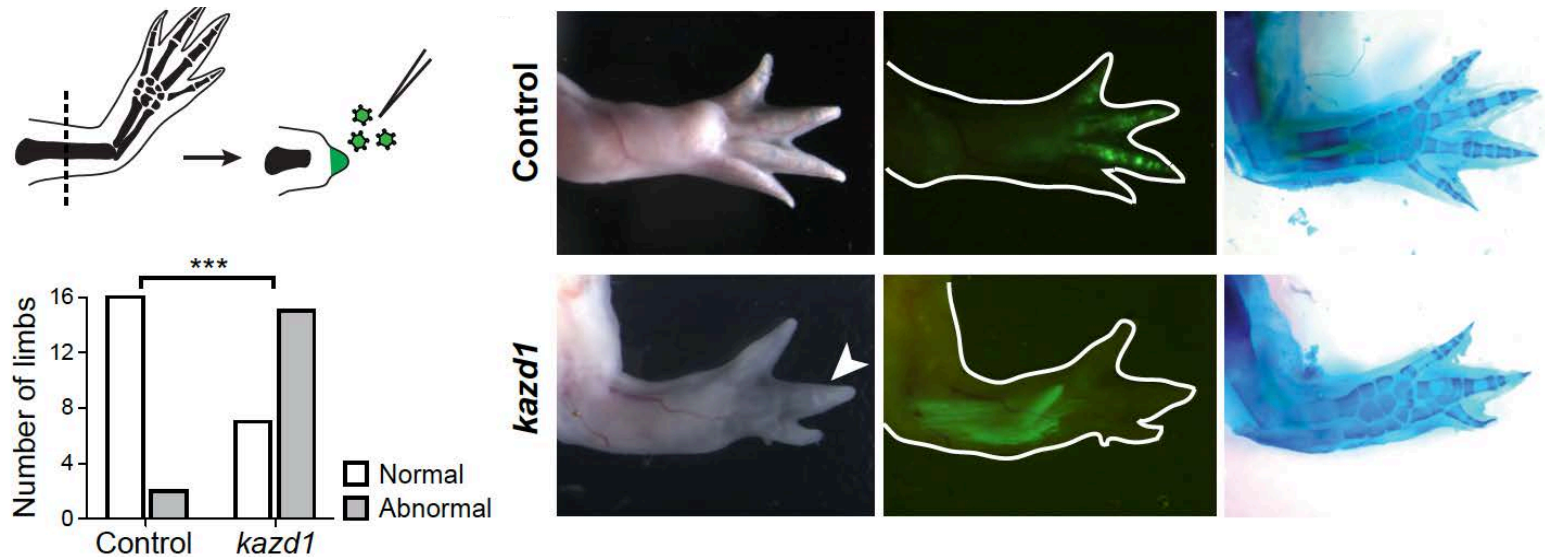
In situ hybridization of kazald1 over course of regeneration



Morpholino Knockdown of Kazald1 Expression



Viral-based Delivered Over-expression of KAZD1 Leads to Regeneration Defects



Cell Reports

Volume 18
Number 3

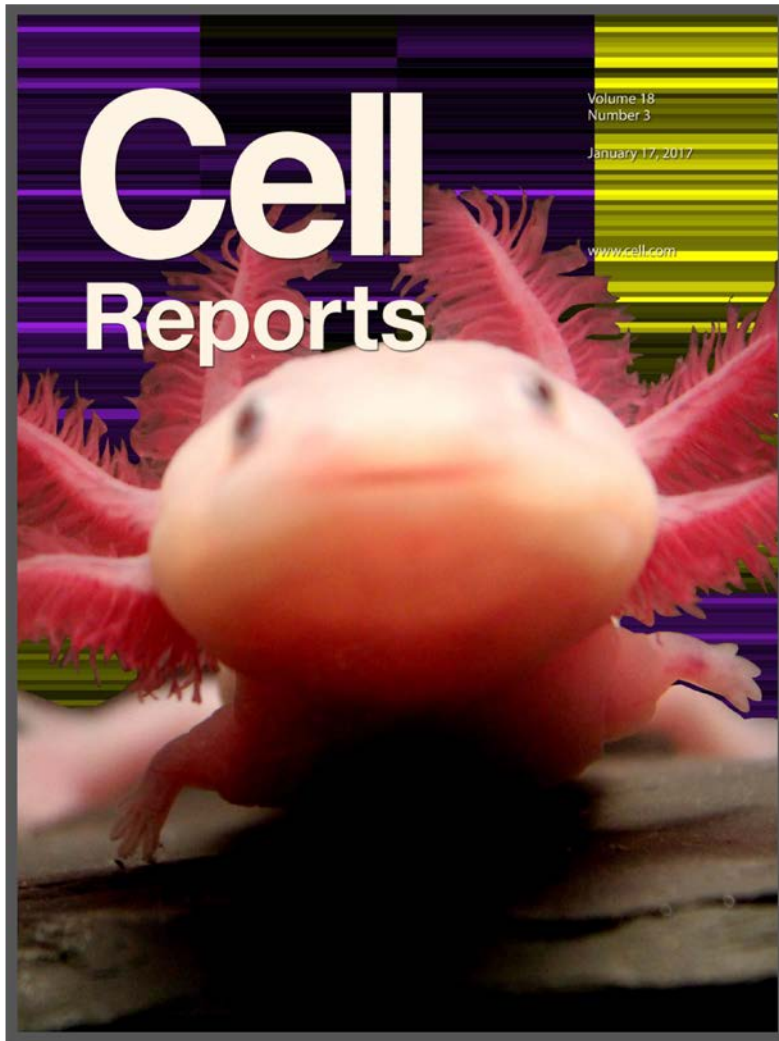
January 17, 2017

www.cell.com

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

Jan 17, 2017

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

Summary of Key Points

- RNA-Seq is a versatile method for transcriptome analysis enabling quantification and novel transcript discovery.
- Expression quantification is based on sampling and counting reads derived from transcripts
- Fold changes based on few read counts lack statistical significance.
- Trinity assembly and supported downstream computational analysis tools facilitate transcriptome studies.
- The Trinity framework can empower transcriptome studies for organisms lacking reference genome sequences (ex. Axolotl) or suboptimal references (ex. cancer).

Summary of Current Trends

- Quantification without read alignment (pseudalignment – kallisto, salmon).
- Differential expression w/o expression estimation (transcript equivalence classes)
- Leverage longer reads (no assembly required?) (pacbio, nanopore)

Acknowledgements



Current and Former Trinity Contributors

Aviv Regev

* Brian Haas

Moran Yassour

Manfred Grabherr

Asma Bankapur

Tim Tickle

Christophe Georgescu

Vrushali Fangal

Maxwell Brown



Jill Mesirov

James Robinson

Trinotate & TrinoateWeb

Brian Couger

Leonardo Gonzalez



Salamander limb regeneration

Jessica Whited

Nick Leigh

Kim Johnson

Donald Bryant

Tia DiTommaso

Tae Lee

Anna Guzikowski

Trinity is funded by:



Informatics Technology
for Cancer Research



Transcriptomics Lab

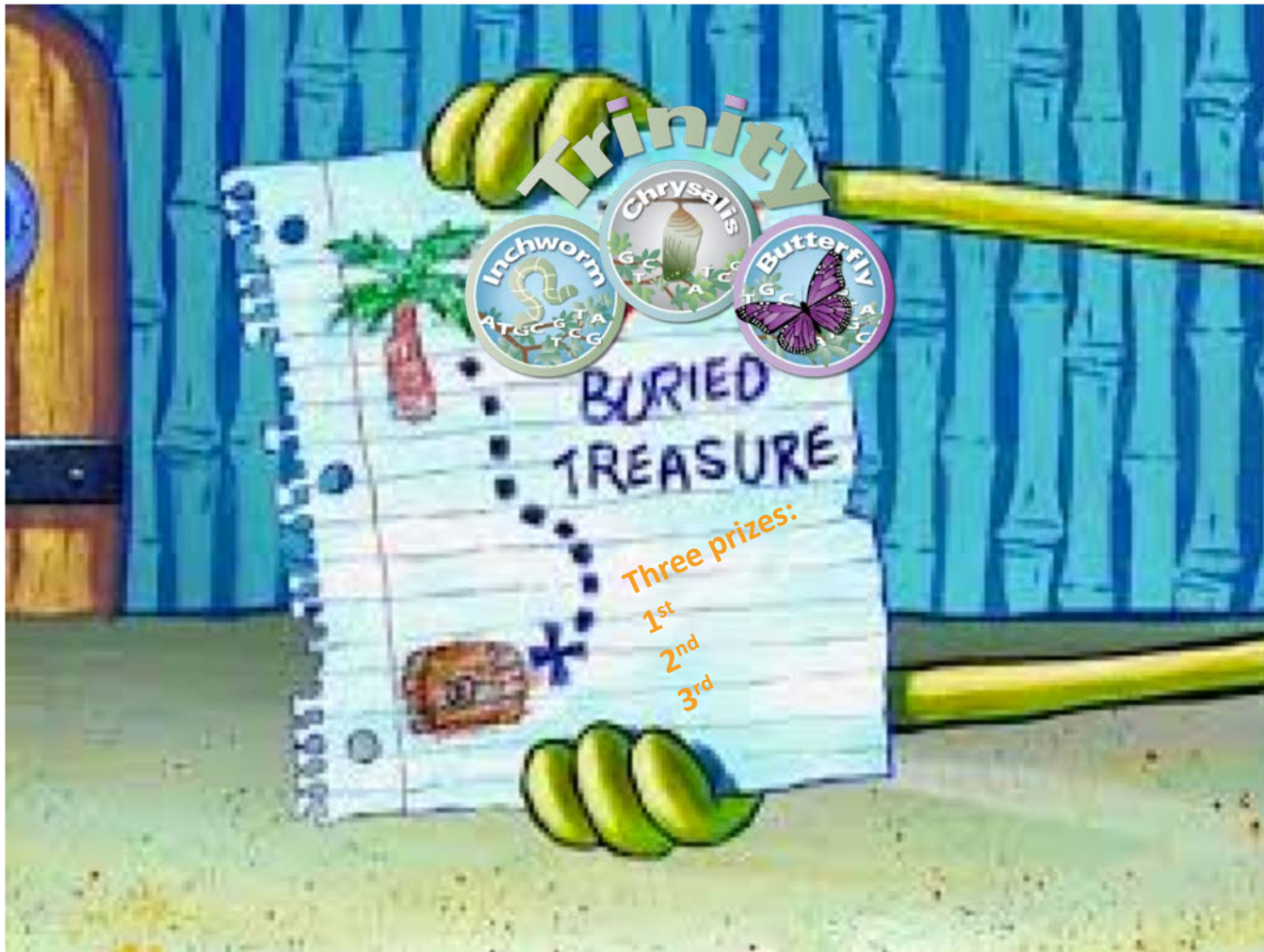
(Krumlov Prelate, 2-5pm)

De novo RNA-Seq Assembly, Annotation, and Analysis Using Trinity and Trinotate

The following details the steps involved in:

- Generating a Trinity de novo RNA-Seq assembly
- Evaluating the quality of the assembly
- Quantifying transcript expression levels
- Identifying differentially expressed (DE) transcripts
- Functionally annotating transcripts using Trinotate and predicting coding regions using TransDecoder
- Examining functional enrichments for DE transcripts using GOseq
- Interactively Exploring annotations and expression data via TrinotateWeb

Trinity Treasure Hunt!!! 😊



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

Slack channel: [#transcriptomicslab](#)