The Krumlov Trinity Transcriptomics Experience

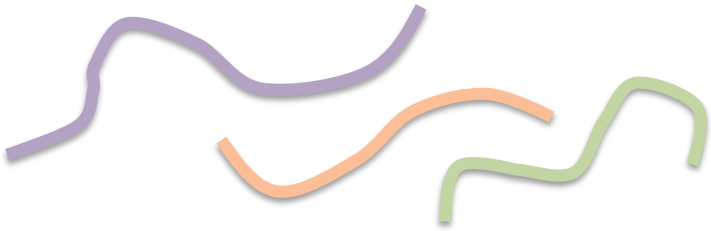
Brian Haas

Broad Institute

Workshop on Genomics, Cesky Krumlov, May 2022

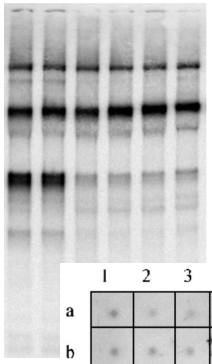
Biological Investigations Empowered by Transcriptomics

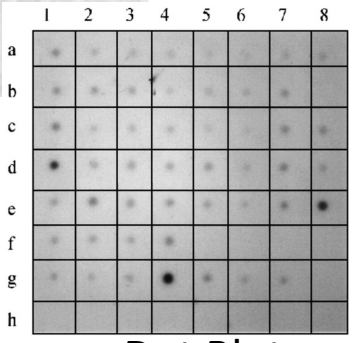
Extract RNA,
… some protocol for processing, ...



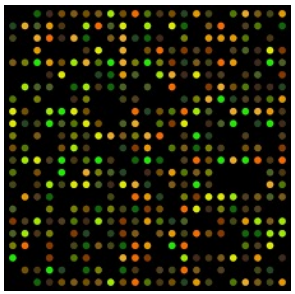
Analysis Method
(pick your favorite)

Northern

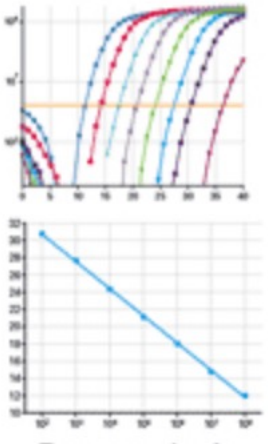




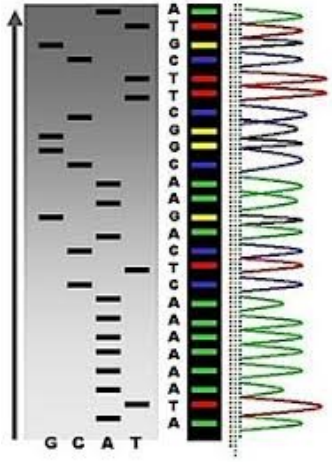
Dot Blot



Microarray



qRT-PCR



Sanger Sequencing

Other…

Minion

MinION

Historical Timeline to Modern Transcriptomics (from 1970)

Reverse Transcription (1970)

Northern Blot
Sanger Sequencing
(1977)

Expressed Sequence Tags (1992)

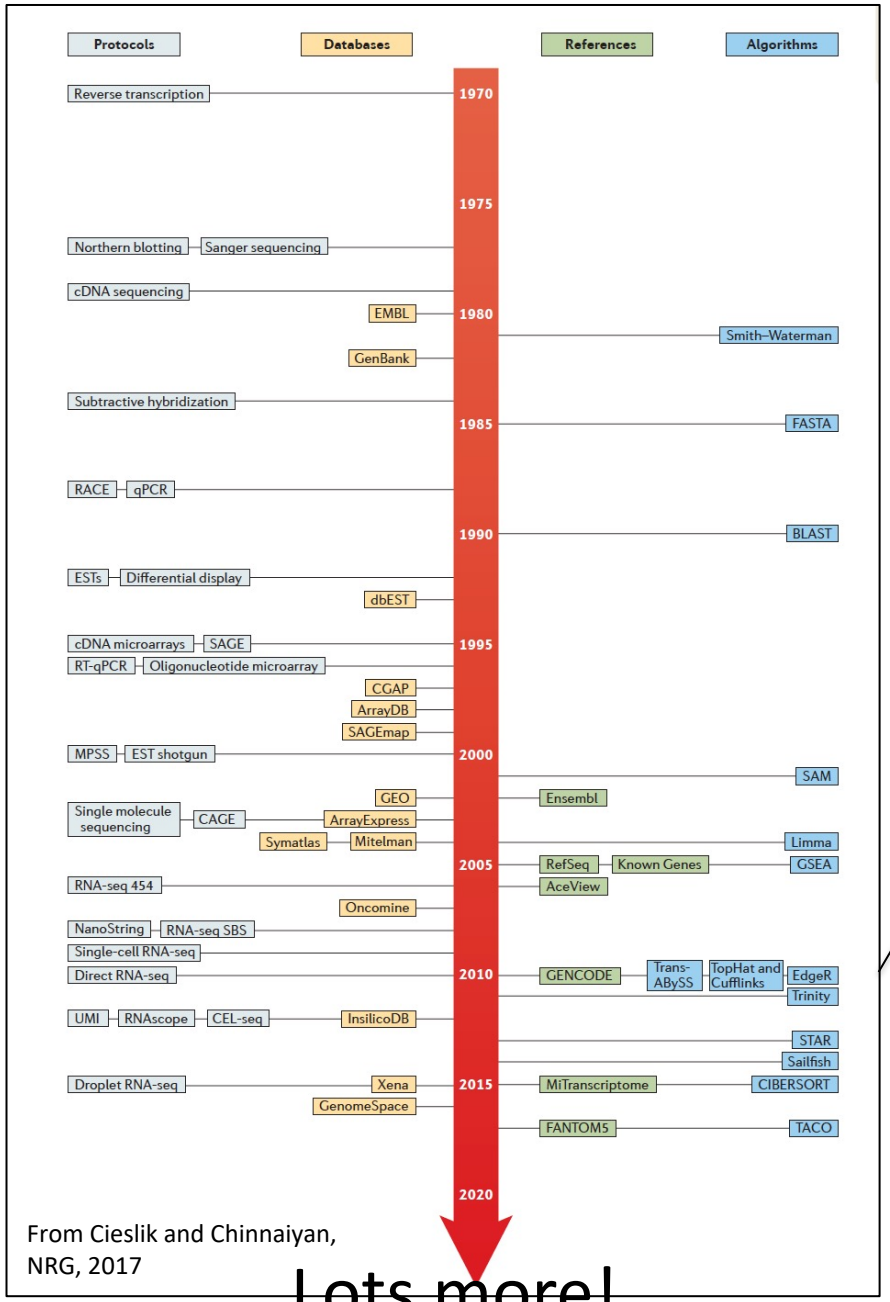
cDNA microarrays (1995)

RNA-Seq (2006-2008)

PacBio IsoSeq (2014)

Droplet single cell RNA-Seq (2015)

Direct RNA Seq Nanopore (2018)



From Cieslik and Chinnaiyan, NRG, 2017

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

Smith Waterman (1981)

BLAST (1990)

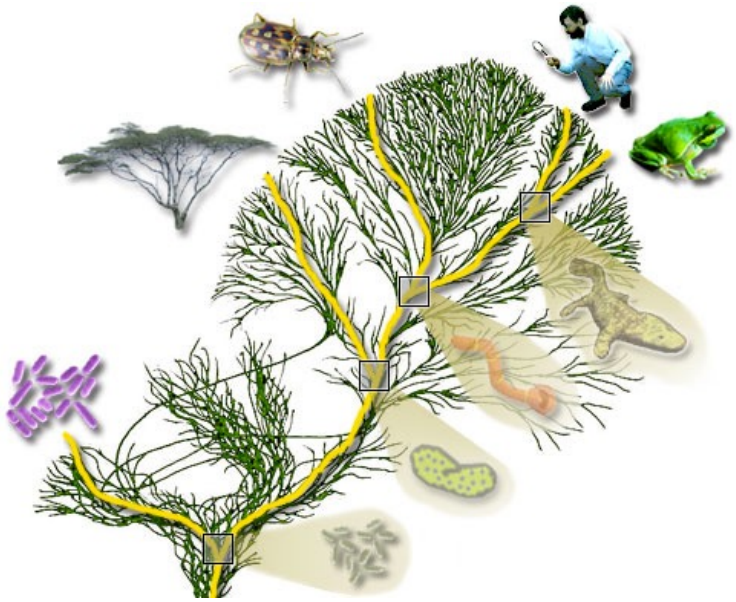
Tophat/Cufflinks (2010)

RSEM (2011)

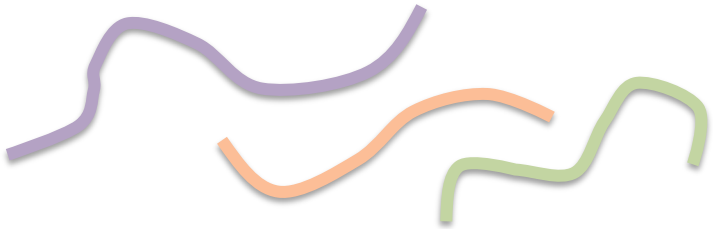
Kallisto (2016)
Salmon (2017)

Lots more!

Modern Transcriptome Studies Empowered by RNA-seq



Extract RNA, convert to cDNA

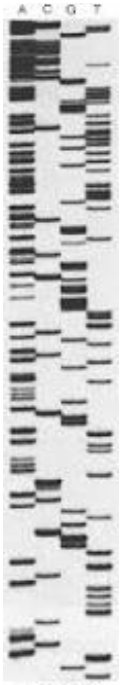


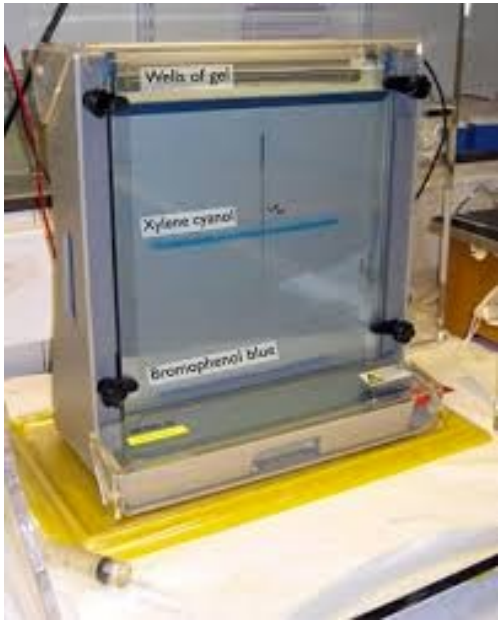
Next-gen Sequencer
(pick your favorite)

Millions to Billions of Reads

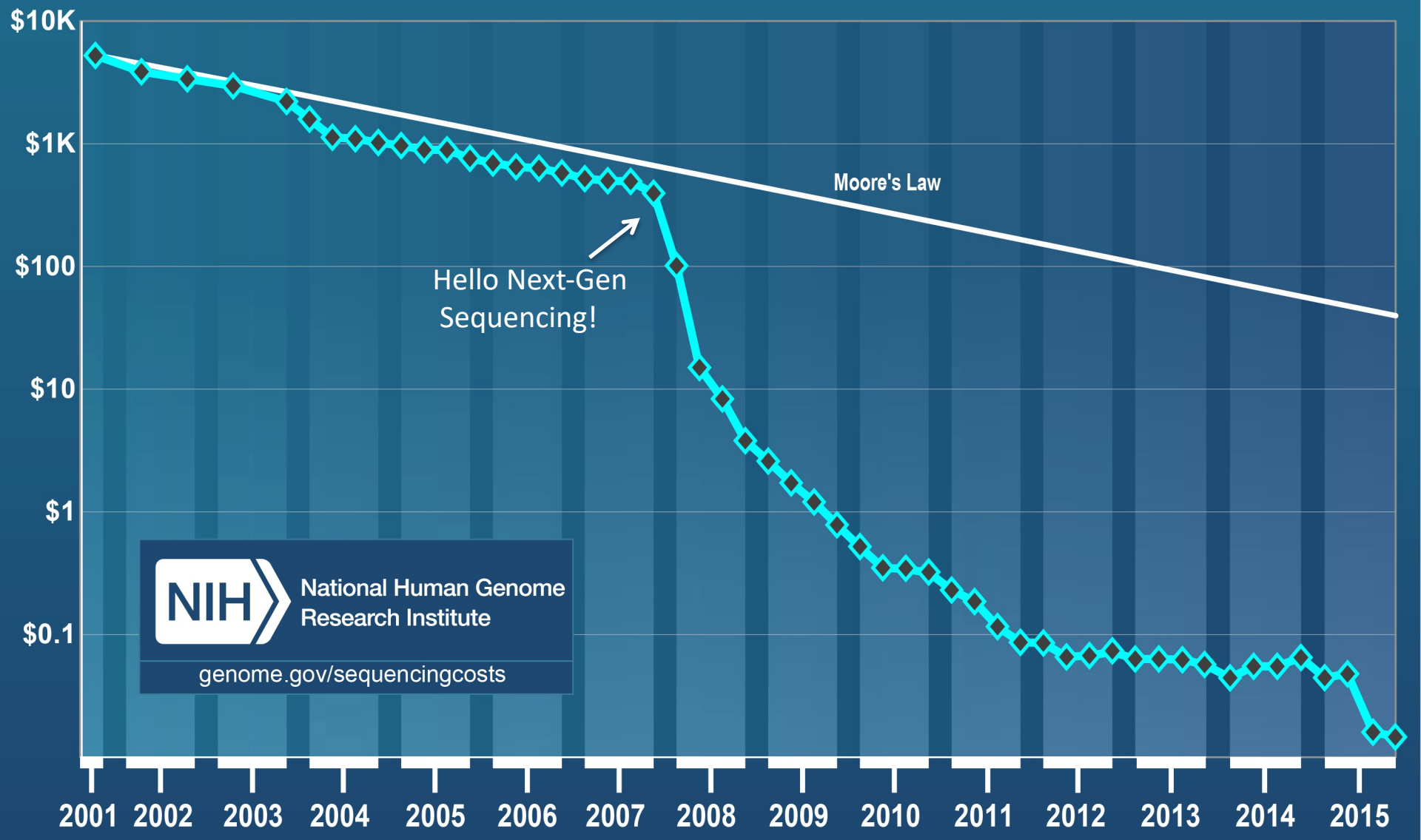
Personal Reflections...

Circa 1995





Cost per Raw Megabase of DNA Sequence



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

Generating RNA-Seq: *How to Choose?*

Platform	iSeq 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	Prome thION RnD	Prome thION theor etical	QiaGen Gene Reader	BGI SEQ 500	BGI SEQ 50	#
Reads: (M)	4	25	25	400	600	3000	4000	5000	6000	3300	6600	20000	1400	3-5	15-20	165	60-80	5.5	38.5	--	--	--	--	--	--	400	1600	1600	--
Read length: (paired-end*)	150*	150*	300*	150*	100*	100*	125*	150*	150*	150*	150*	150*	60	200 400	200 400	200	200	15K	12K	32K	--	--	--	--	--	--	100*	50	--
Run time: (d)	0.54	1	2	1.2	1.125	11	6	3.5	3	1.66	1.66	1.66	7	0.37	0.16	--	0.16	4.3	--	--	--	2	2	2	--	--	1	0.4	--
Yield: (Gb)	1	7.5	15	120	120	600	1000	1500	1800	1000	2000	6000	180	1.5	7	10	12	12	5	150	4	8	40	2400	11000	80	200	8	--
Rate: (Gb/d)	1.85	7.5	7.5	100	106.6	55	166	400	600	600	1200	3600	30	5.5	50	--	93.75	2.8	--	--	--	4	20	1200	5500	--	200	20	--
Reagents: ($K)	0.1	1.75	1	5	6.145	23.47	29.9	--	--	--	--	--	10.5	0.6	--	1	1.2	2.4	--	1	--	0.5	1.5	--	--	0.5	--	--	--
per-Gb: ($)	100	233	66	50	51.2	39.1	31.7	20.5	7.08	18	15	5.8	58.33	--	--	100	--	200	80	6.6	--	62.5	37.5	20	4.3	--	--	--	--
hg-30x: ($)	12000	28000	8000	5000	6144	4692	3804	2460	849.6	1800	1564	700	7000	--	--	12000	--	24000	9600	1000	--	7500	4500	2400	500	--	600	--	
Machine: ($)	30K	49.5K	99K	250K	740K	690K	690K	900K	1M	999K	999K	999K	595K	50K	65K	243K	242K	695K	350K	350K	--	--	125K	75K	75K	--	200K	--	--
#Page maintained by http://twitter.com/albertvilella http://tinyurl.com/ngslytics #Editable version: http://tinyurl.com/ngsspecsshared																													
#curl "https://docs.google.com/spreadsheets/d/1GMMfhyLK0-q8XkIo3YxlWaZA5vVMuhU1kg41g4xLkXc/export?gid=4&format=csv" | grep -v '^#' | grep -v '^"' | column -t -s\, | less -S																													

Stats circa 2018

For current, see: **https://tinyurl.com/wbgcs65**

*Not all shown at scale

Generating RNA-Seq: *How to Choose?*

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

#Page maintained by http://twitter.com/albertvilella http://t

#curl "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	--	--

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

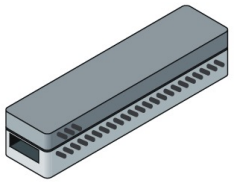
Thx Joshua Levin, for the cartoon. ☺

Each has pros/cons

Today's Most Popular Sequencing Technologies

Illumina

Pacific Biosciences

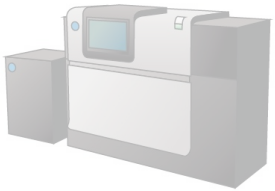


Oxford Nanopore

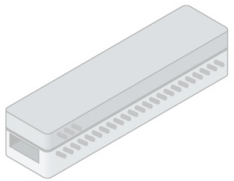
Images from "RNA sequencing: the teenage years"
Rory Stark, Marta Grzelak & James Hadfield
Nature Reviews Genetics volume 20, pages631–656(2019)

Today's Most Popular Sequencing Technologies

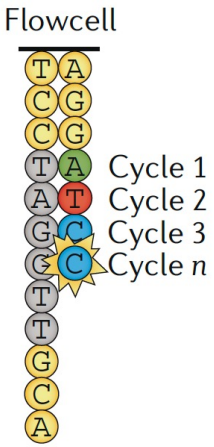
Illumina

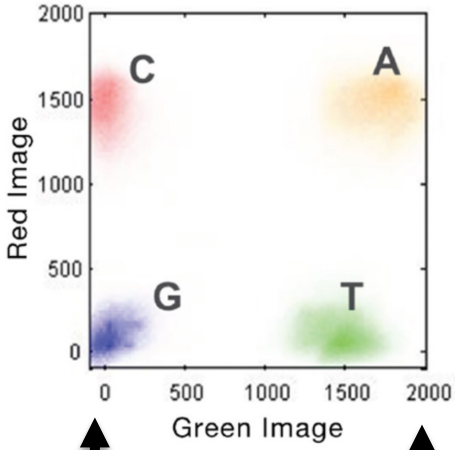


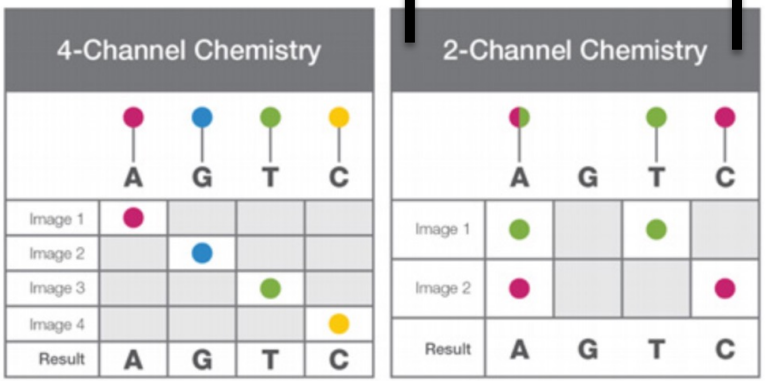
Pacific Biosciences



Oxford Nanopore





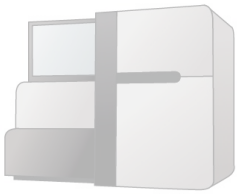


Hundreds of millions to billions of highly accurate but shorter reads. ($)

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

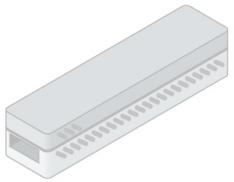
Video at: https://youtu.be/fCd6B5HRaZ8

Today's Most Popular Sequencing Technologies

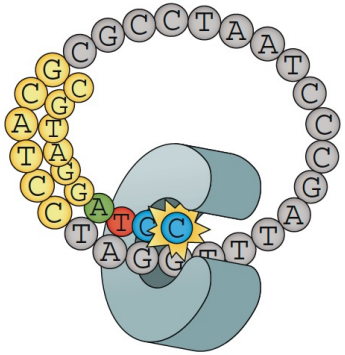


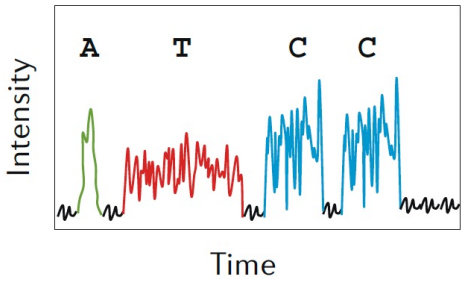
Illumina

Pacific Biosciences



Oxford Nanopore

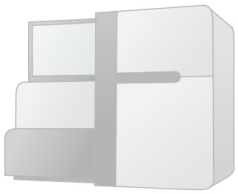




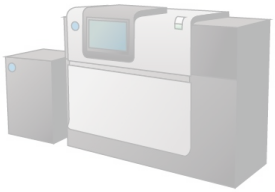
Limited sequencing depth, but highly accurate full-length single molecule reads. ($$$)

Video at: https://www.youtube.com/watch?v=WMZmG00uhwU

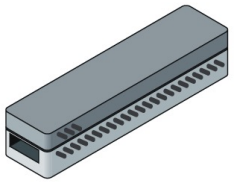
Today's Most Popular Sequencing Technologies



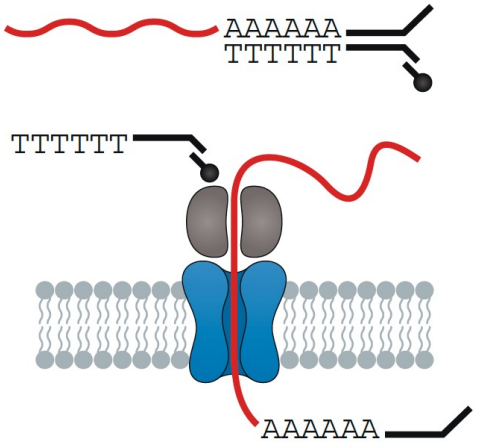
Illumina



Pacific Biosciences



Oxford Nanopore



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

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

Can do direct RNA sequencing! and find evidence for methylation

A Plethora of Biological Sequence Analyses Enabled by RNA-Seq

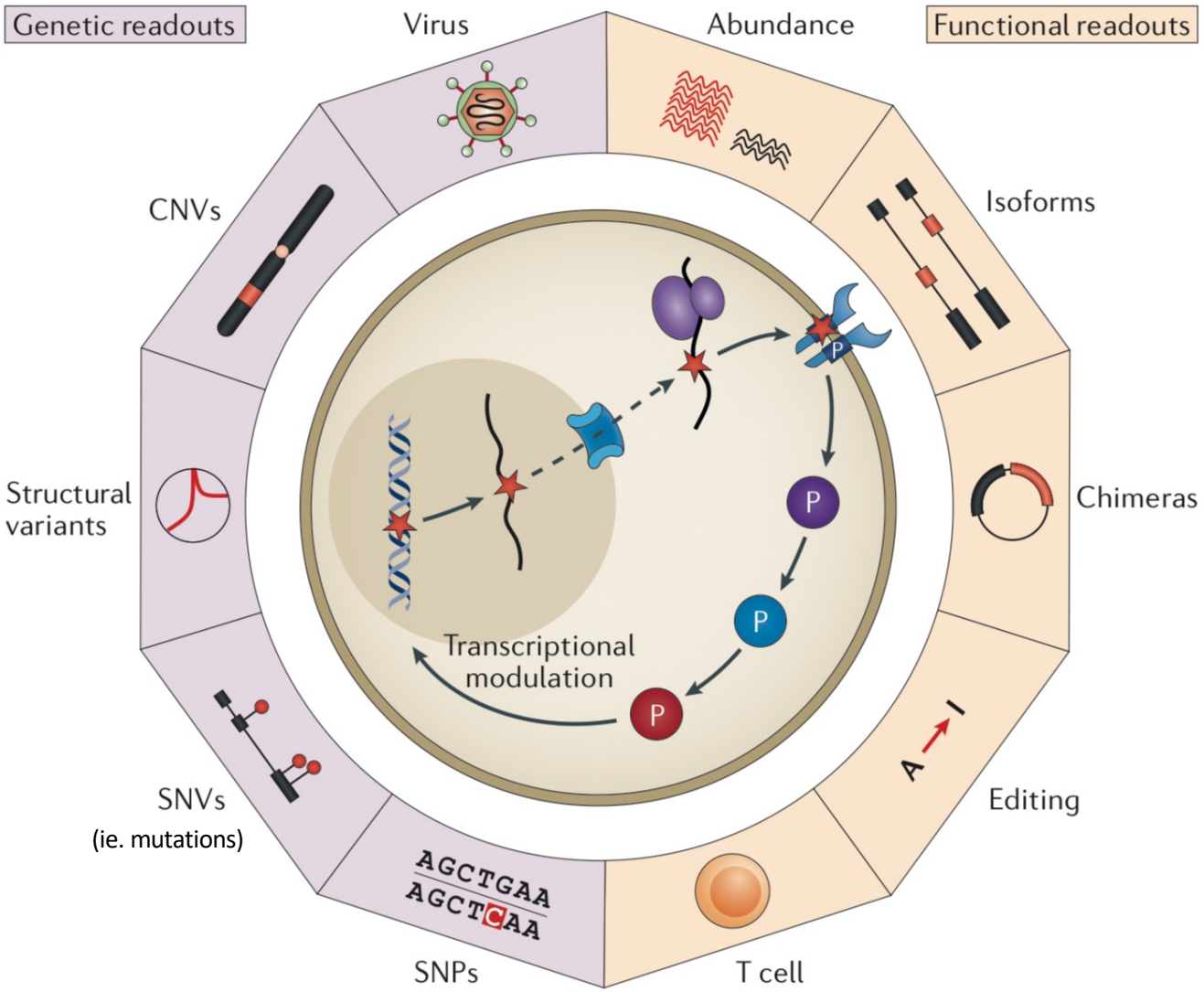
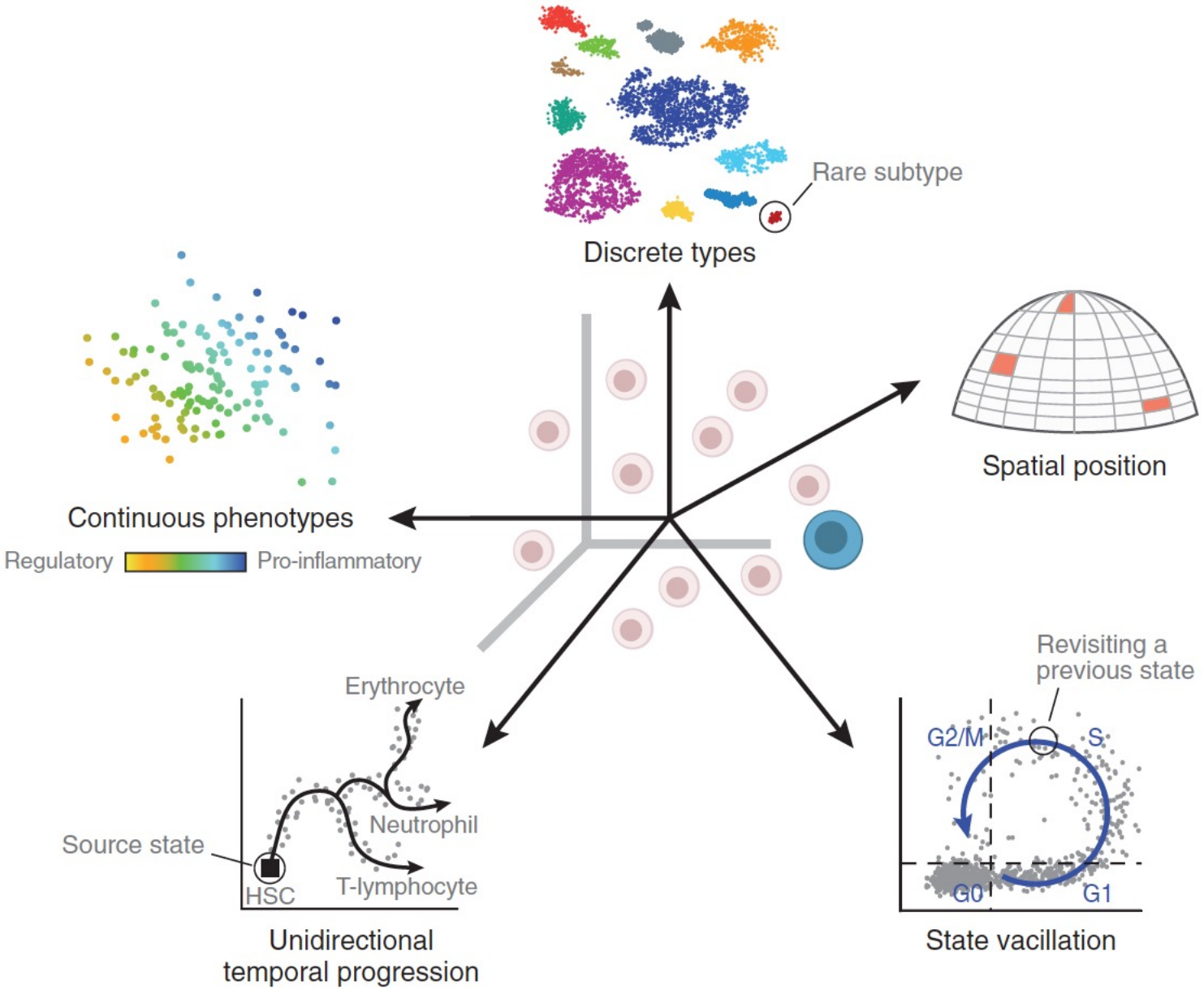


Figure 2 | **Transcriptome profiling for genetic causes and functional phenotypic readouts.**

From Cieslik and Chinnaiyan, NRG, 2017

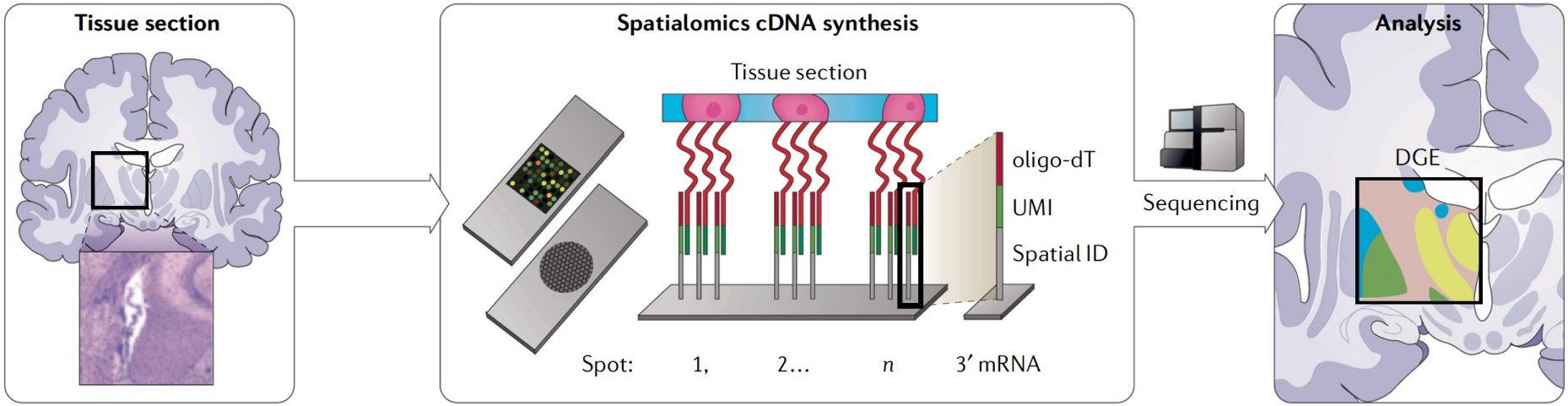
RNA-Seq is Empowering Discovery at Single Cell Resolution



Wagner, Regev, and Yosef. NBT 2016

Spatial Transcriptomics

Spatial Encoding



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

A Myriad of Other Specialized RNA-seq -based Applications

RNA-Sequencing as your lens towards biological discovery

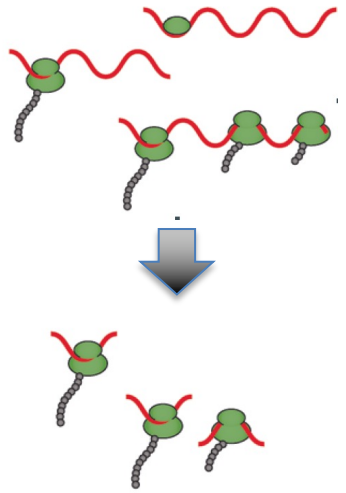
UV crosslink Biotin

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

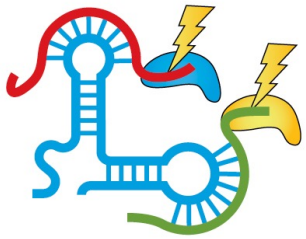
Adapted from "RNA sequencing: the teenage years"
Rory Stark, Marta Grzelak & James Hadfield
Nature Reviews Genetics volume 20, pages631–656(2019)

A Myriad of Other Specialized RNA-seq -based Applications

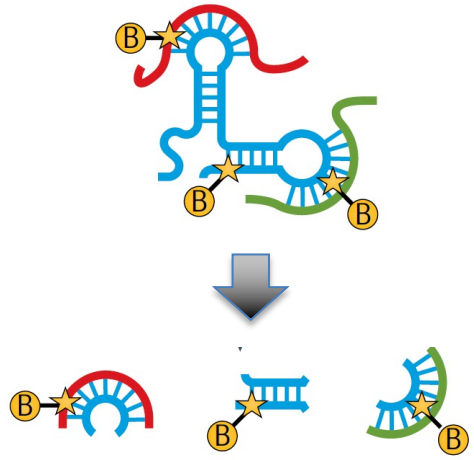
Ribosomal profiling

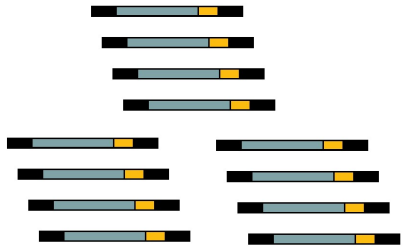


RNA-Protein Interactions

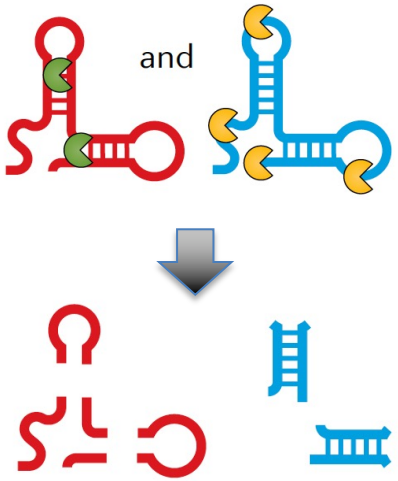


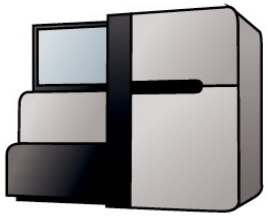
RNA-RNA interactions





RNA Structuromics





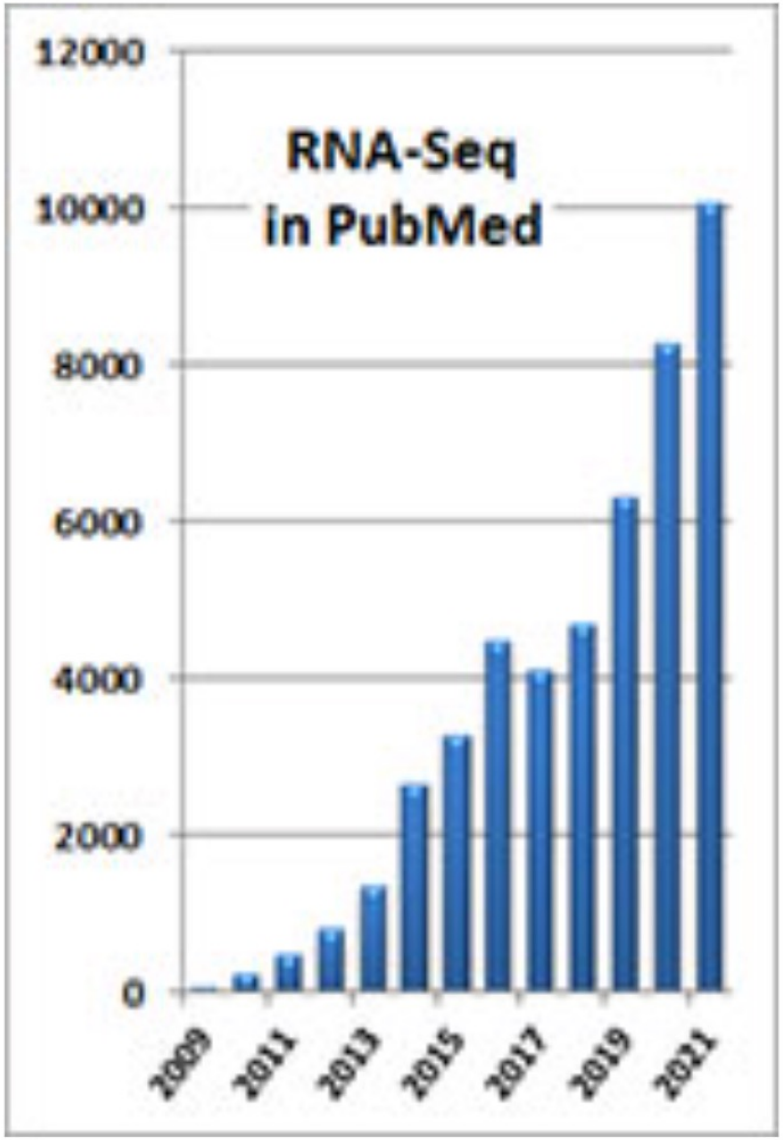
UV crosslink B Biotin

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

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

PUBLICATIONS TREND

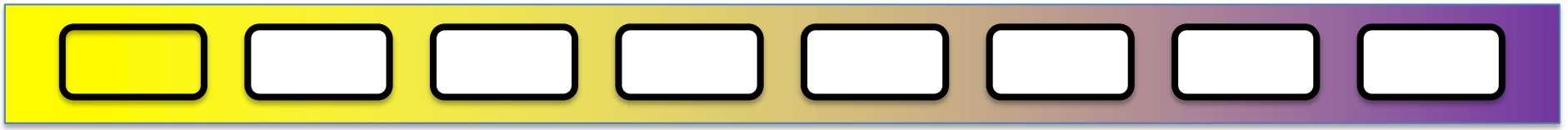
RNA-Seq
Transcriptome Sequencing Research & Industry News



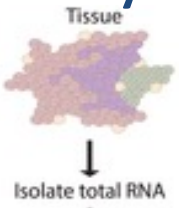
Transcriptomics Lecture Overview

1. Overview of RNA-Seq
2. Transcript reconstruction methods
3. Trinity de novo assembly
4. Transcriptome quality assessment
   (coffee break)
5. Expression quantification
6. Differential expression analysis
7. Functional annotation
8. Case study: salamander transcriptome

Part 1. Overview of RNA-Seq



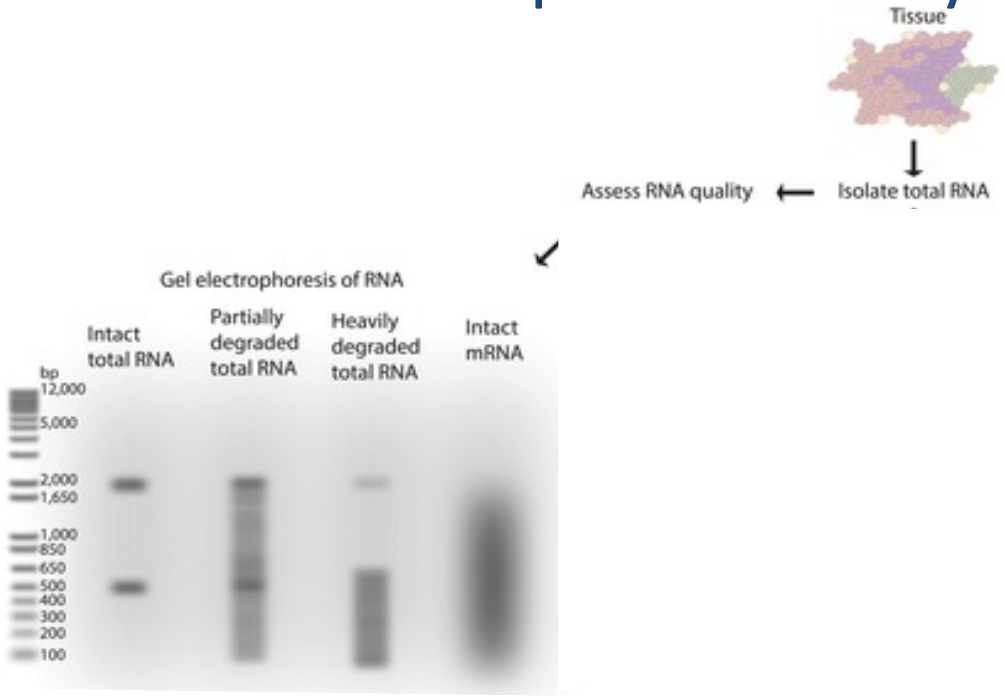
RNA-seq library fragmentation and size selection strategies that influence interpretation and analysis.



http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004393 Griffith et al., 2015

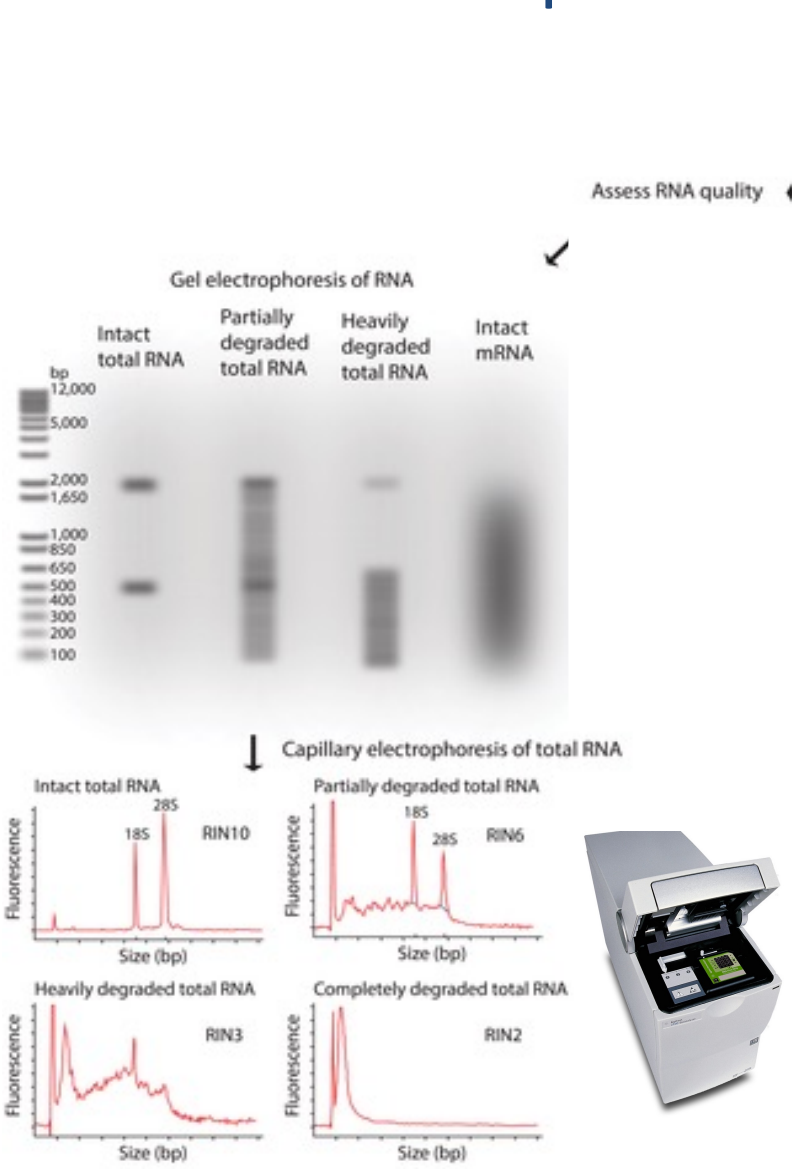
PLOS | COMPUTATIONAL BIOLOGY

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



http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004393 Griffith et al., 2015

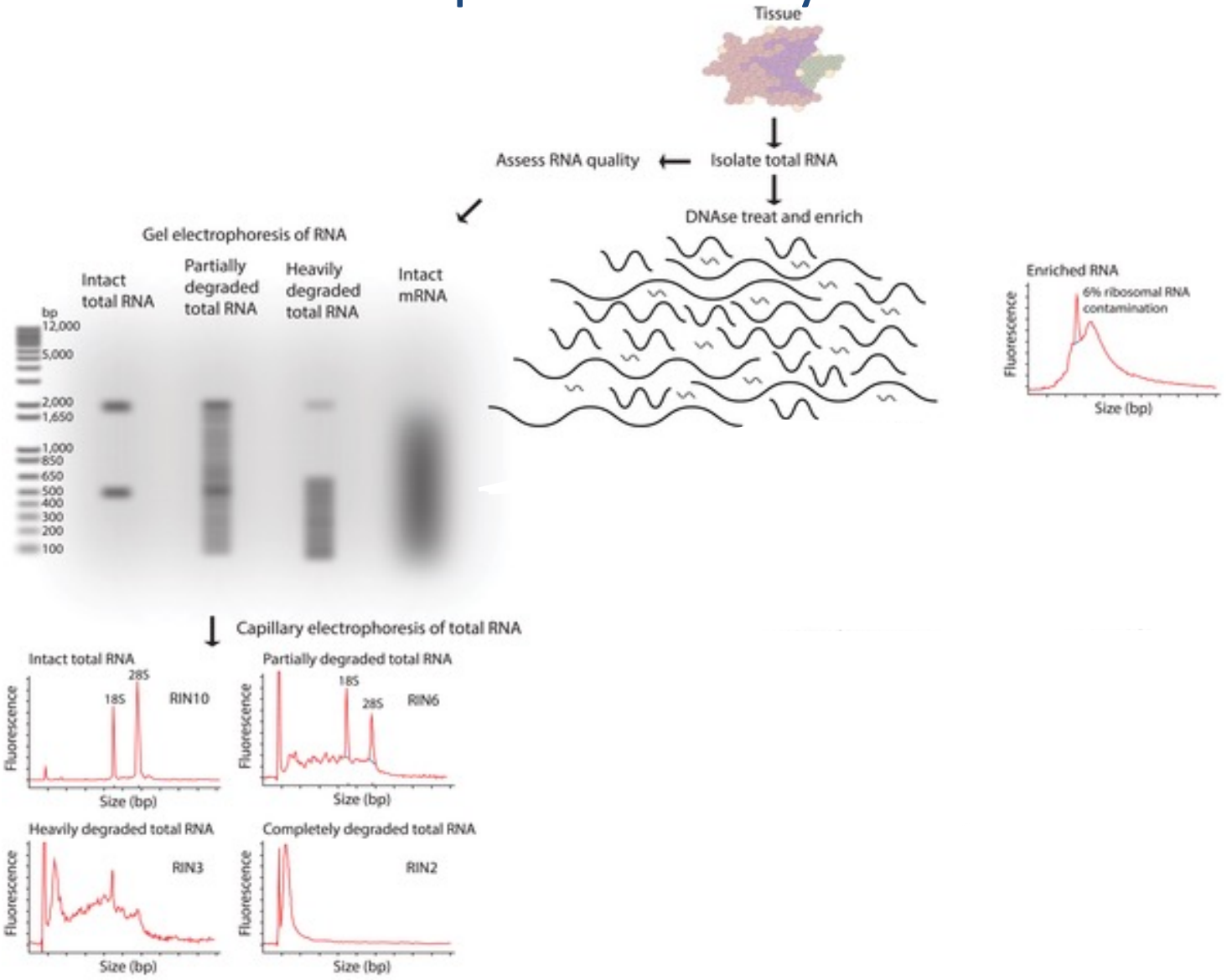
RNA-seq library fragmentation and size selection strategies that influence interpretation and analysis.



http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004393 Griffith et al., 2015

PLOS | COMPUTATIONAL BIOLOGY

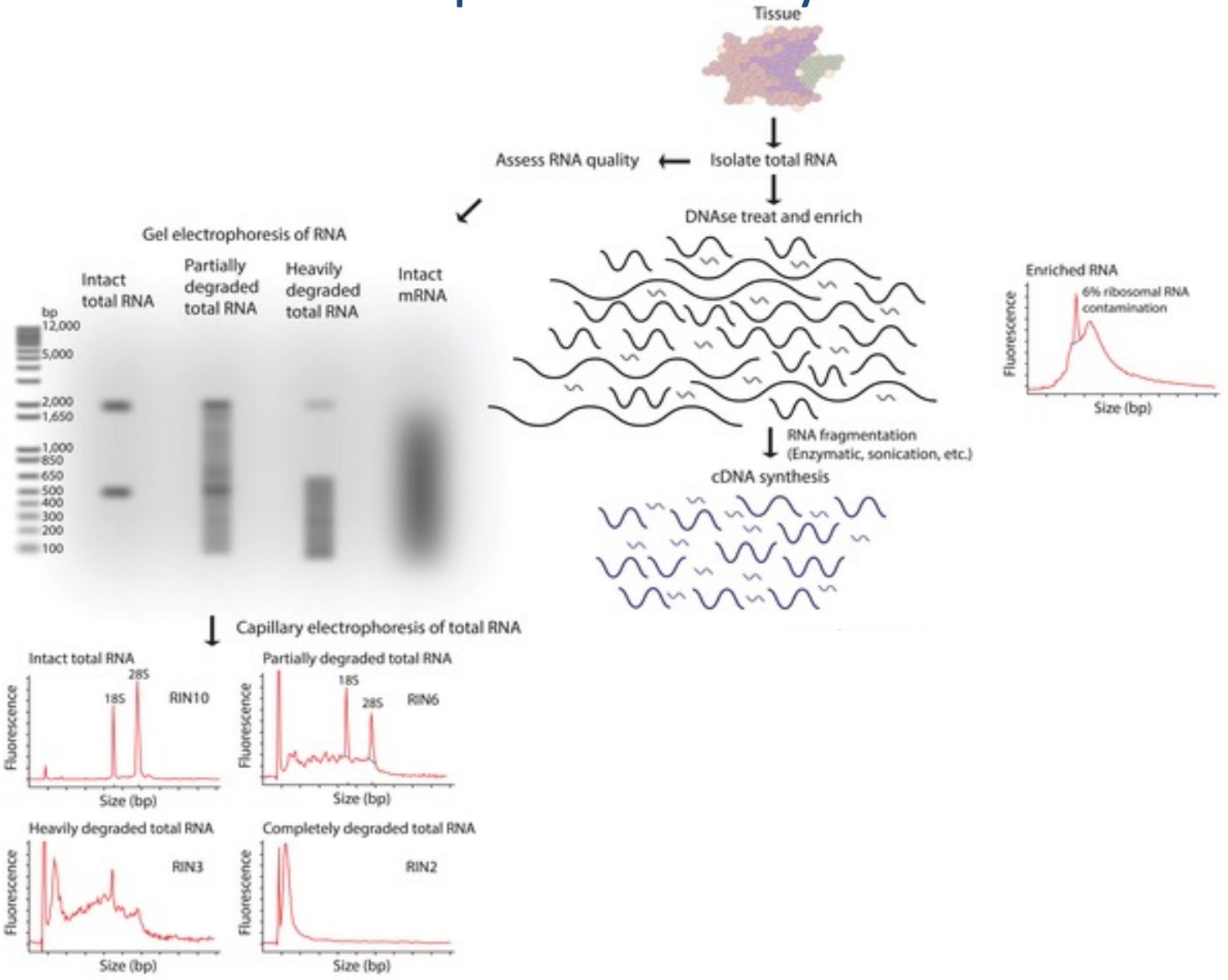
RNA-seq library fragmentation and size selection strategies that influence interpretation and analysis.



http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004393 Griffith et al., 2015

PLOS | COMPUTATIONAL BIOLOGY

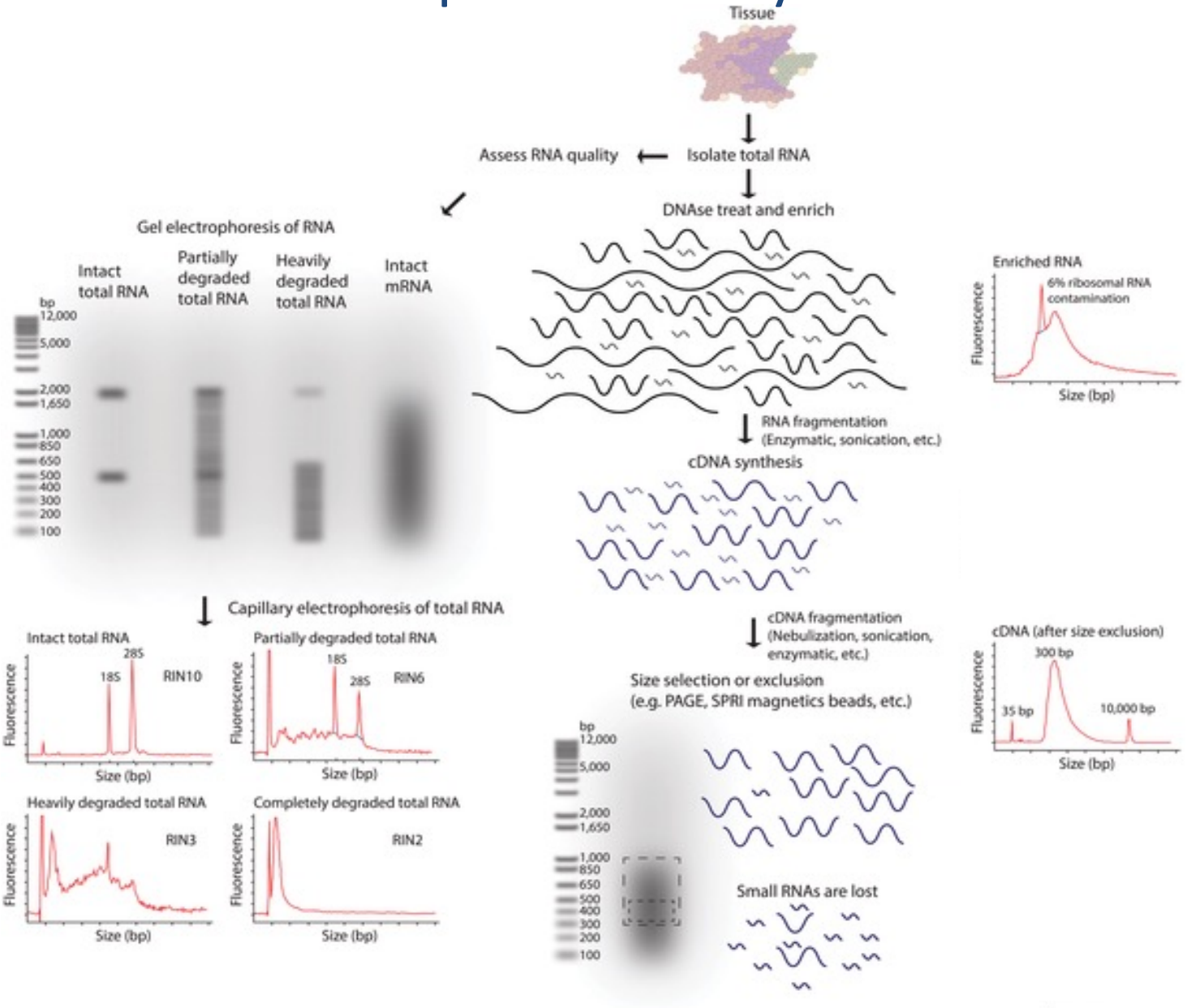
RNA-seq library fragmentation and size selection strategies that influence interpretation and analysis.



http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004393 Griffith et al., 2015

PLOS | COMPUTATIONAL BIOLOGY

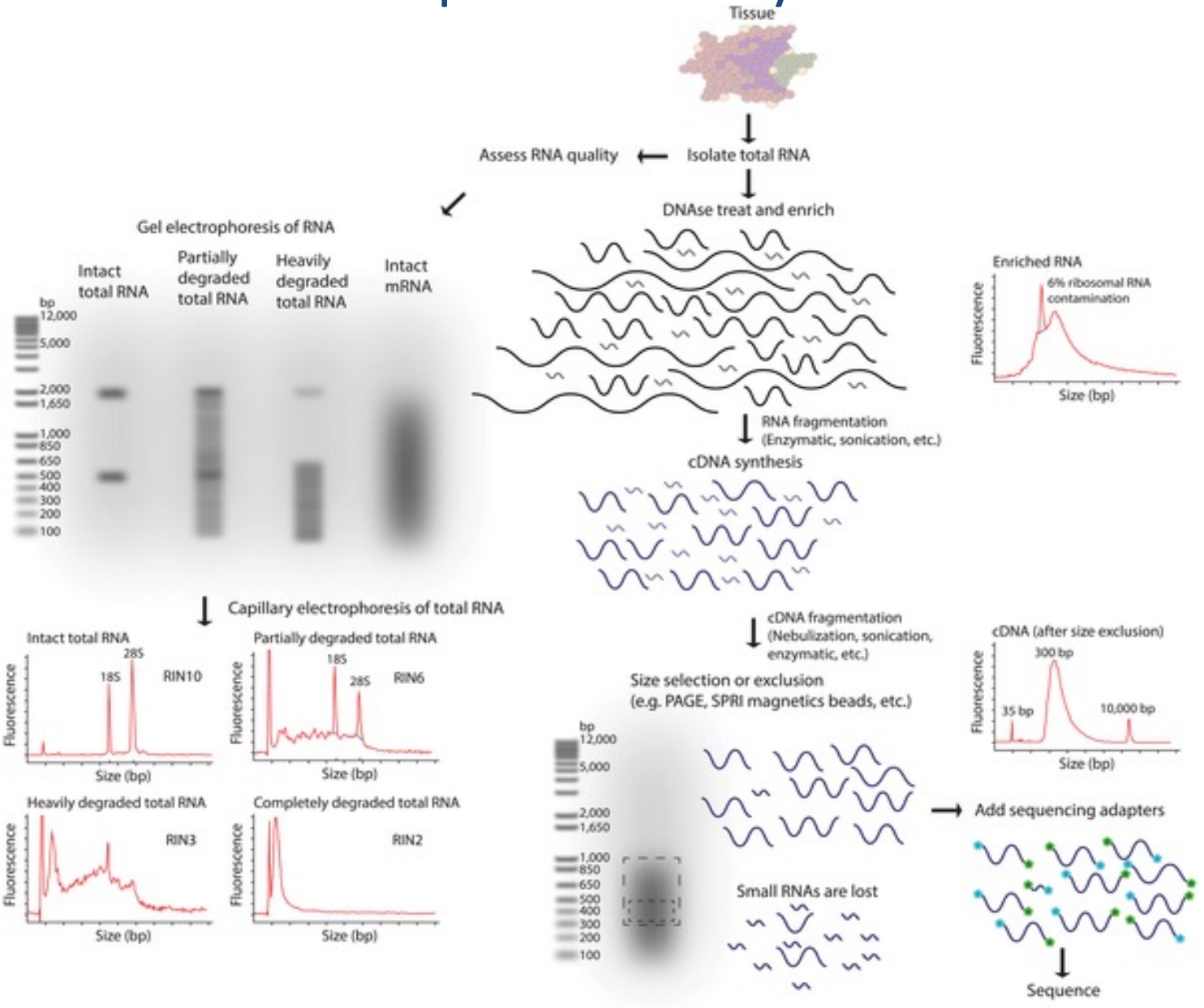
RNA-seq library fragmentation and size selection strategies that influence interpretation and analysis.



http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004393 Griffith et al., 2015

PLOS | COMPUTATIONAL BIOLOGY

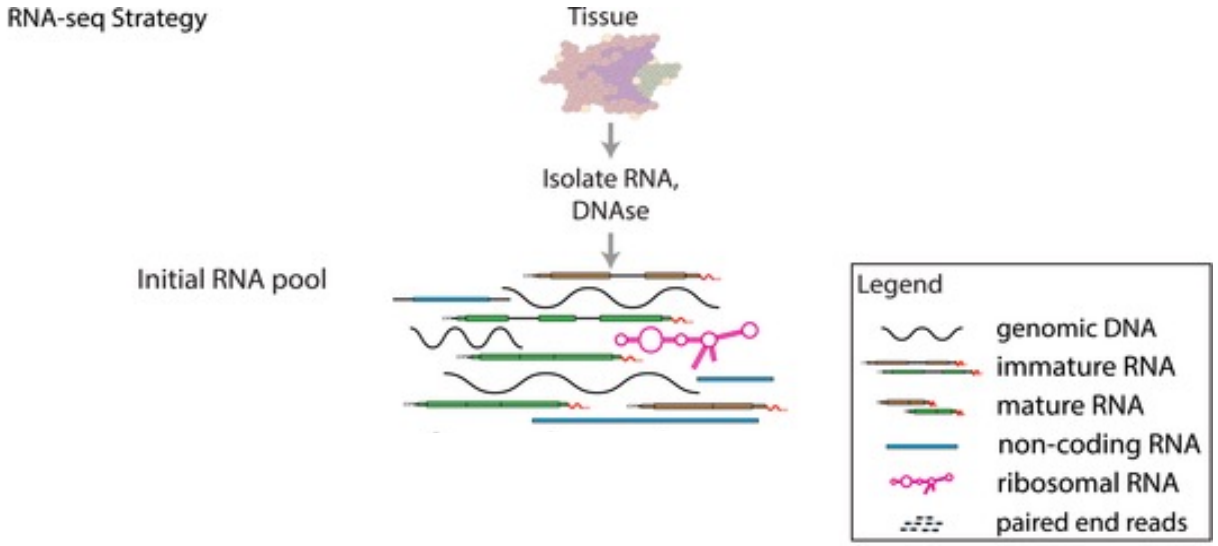
RNA-seq library fragmentation and size selection strategies that influence interpretation and analysis.



http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004393 Griffith et al., 2015

PLOS | COMPUTATIONAL BIOLOGY

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

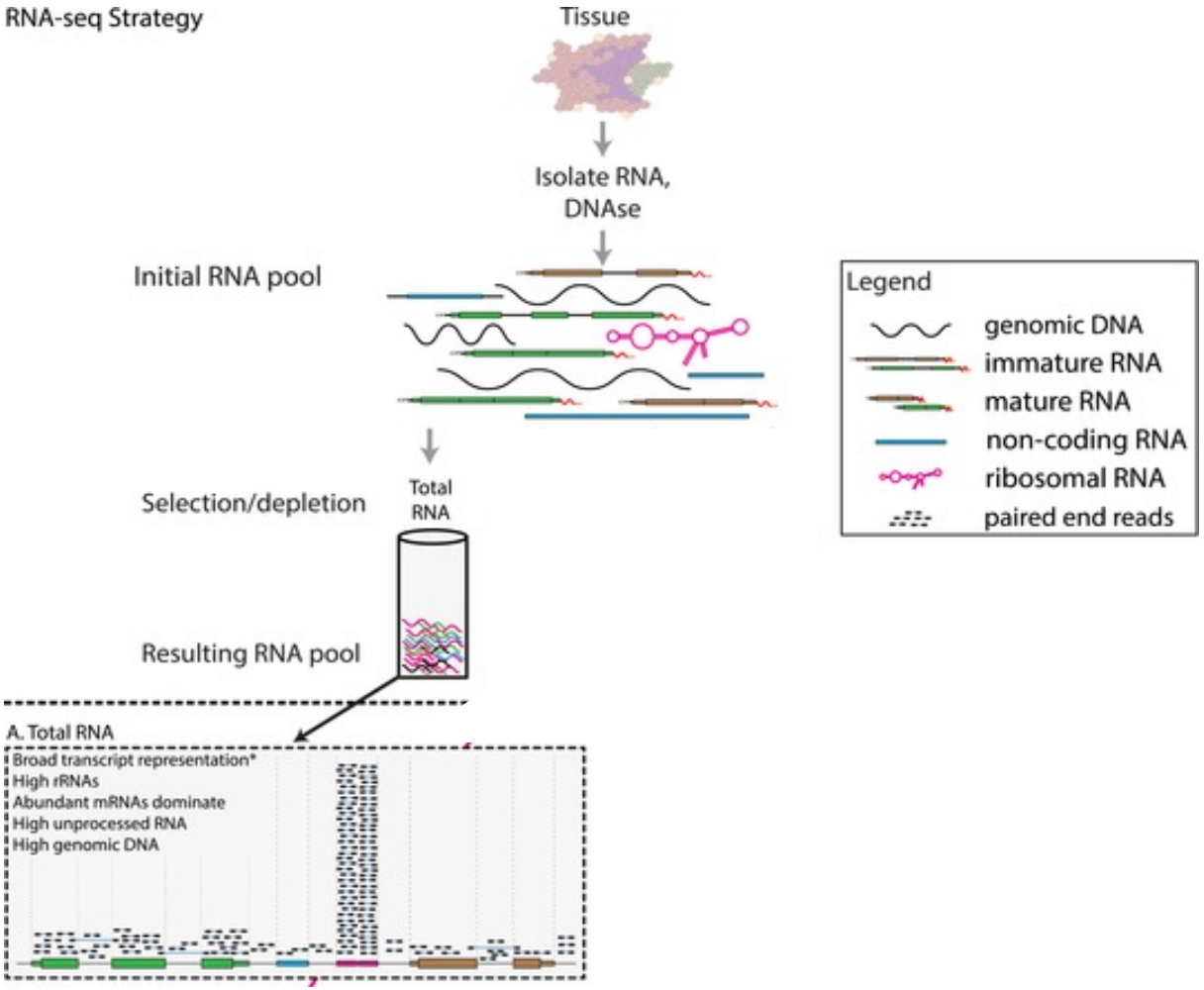


http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004393

Griffith et al., 2015

PLOS | COMPUTATIONAL BIOLOGY

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

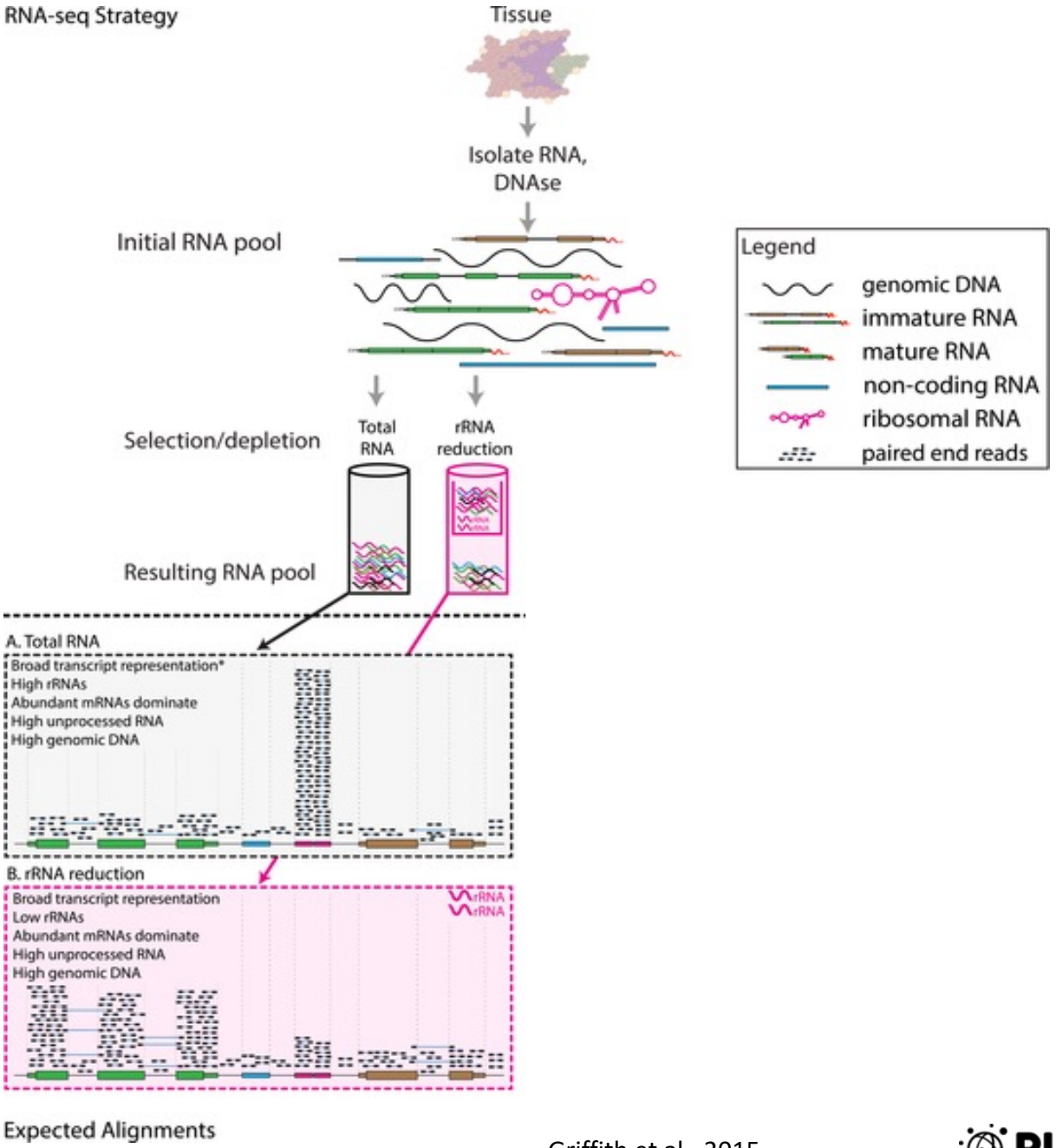


http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004393

Griffith et al., 2015

PLOS | COMPUTATIONAL BIOLOGY

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

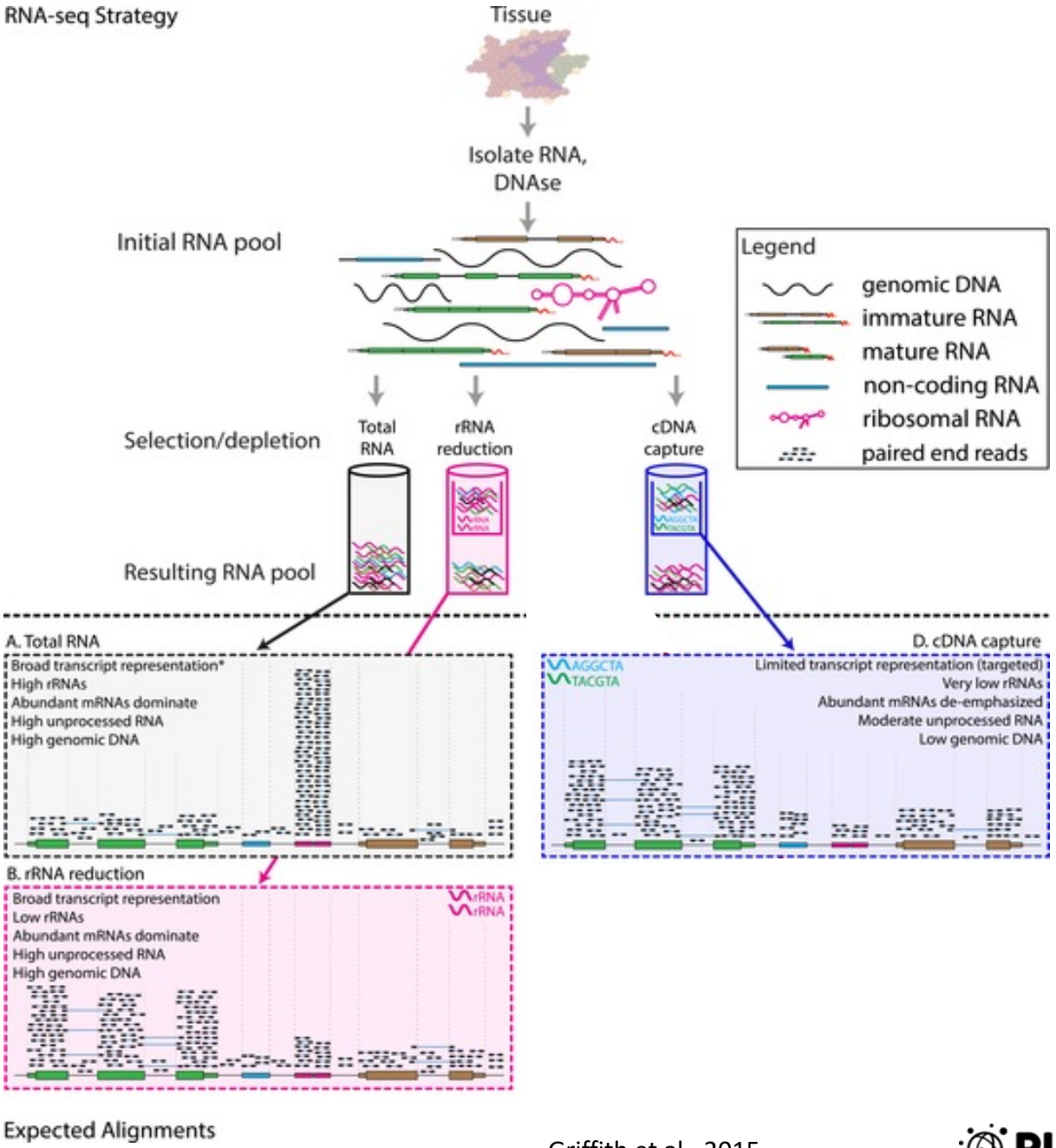


http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004393

Griffith et al., 2015

PLOS | COMPUTATIONAL BIOLOGY

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

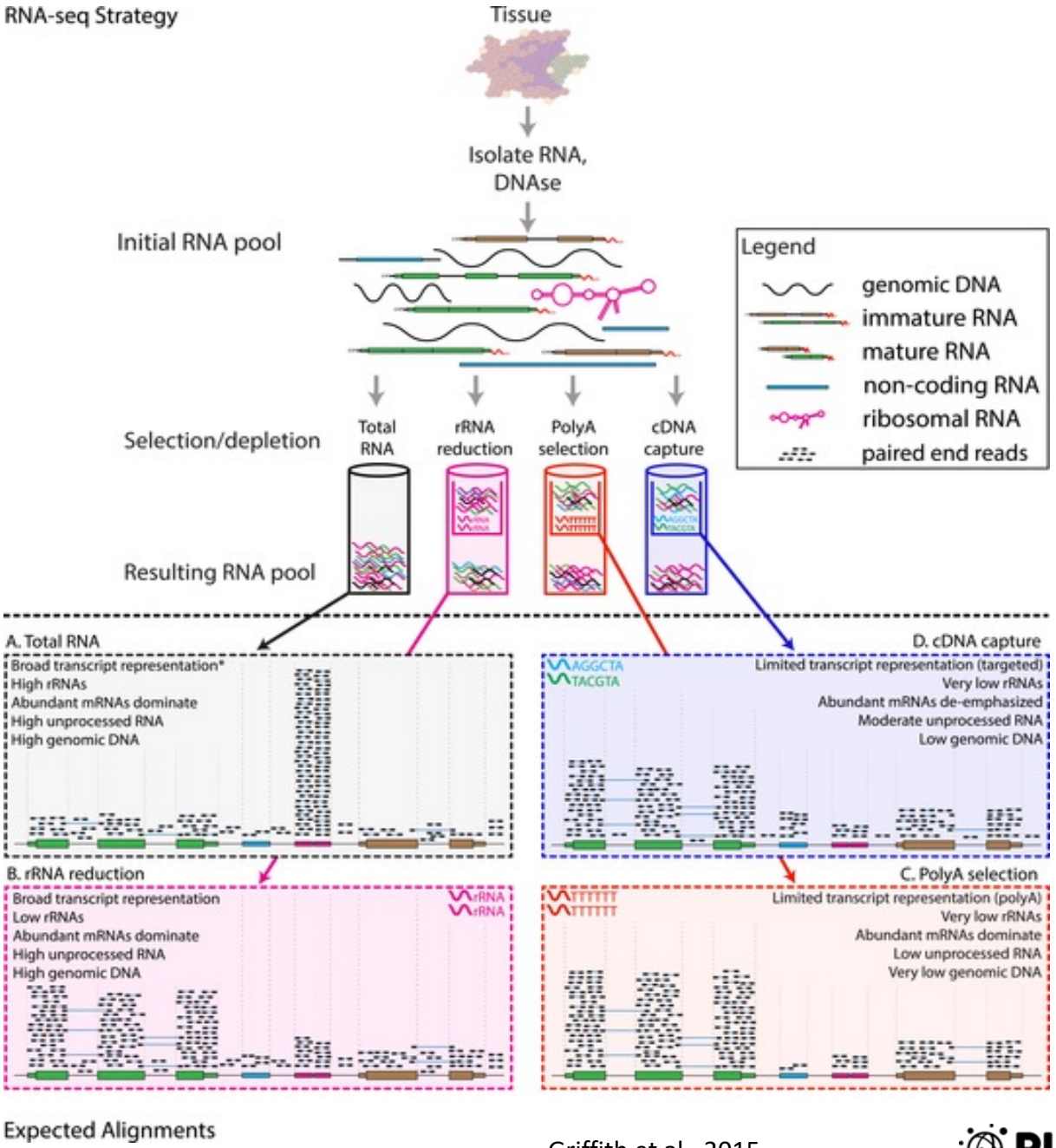


http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004393

Griffith et al., 2015

PLOS | COMPUTATIONAL BIOLOGY

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

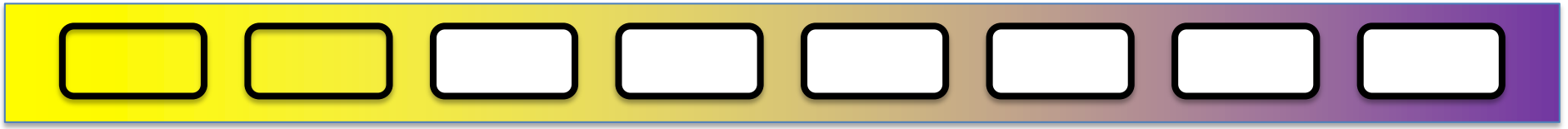


http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004393

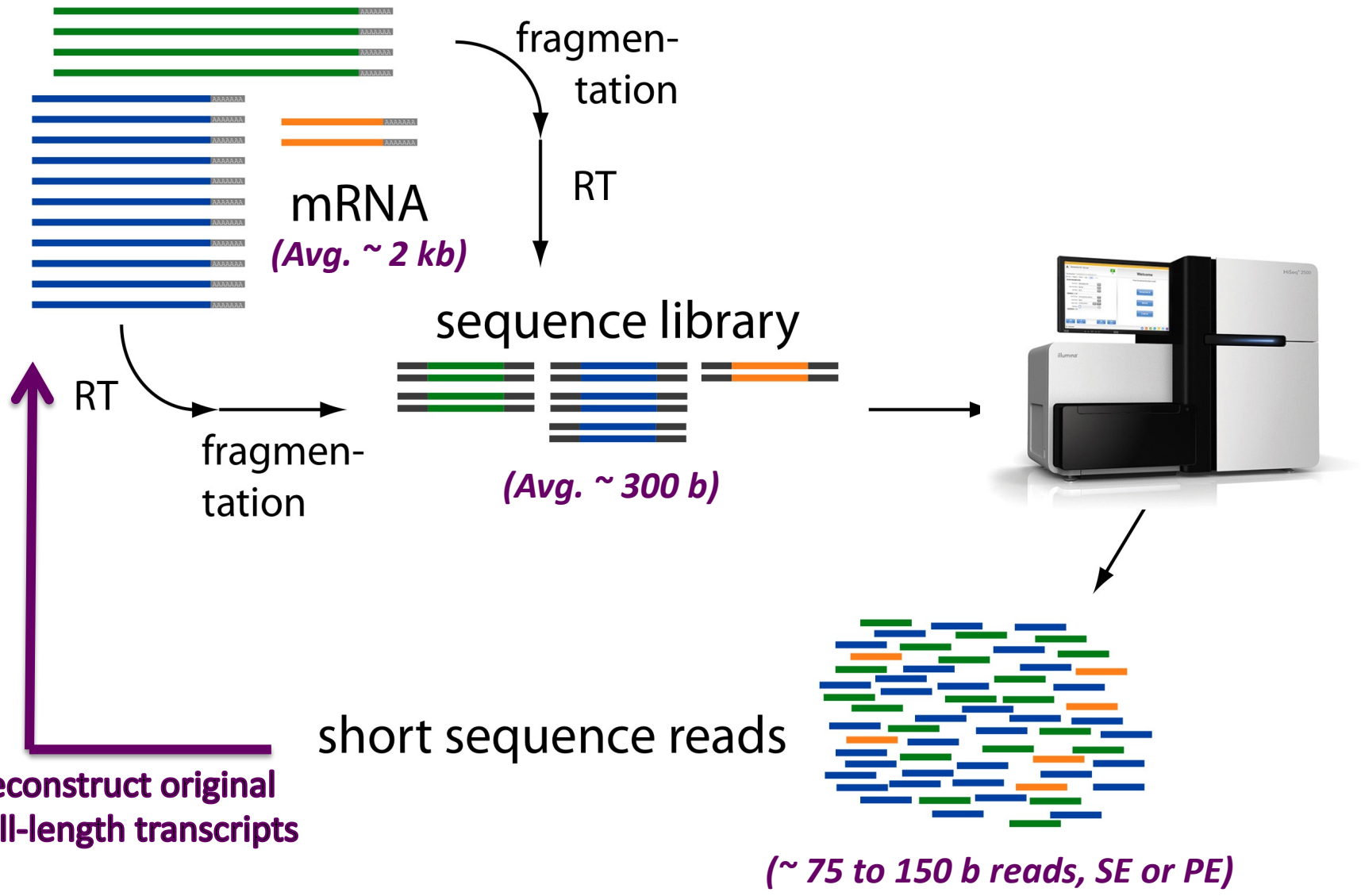
Griffith et al., 2015

PLOS | COMPUTATIONAL BIOLOGY

Part 2. Transcript Reconstruction Methods

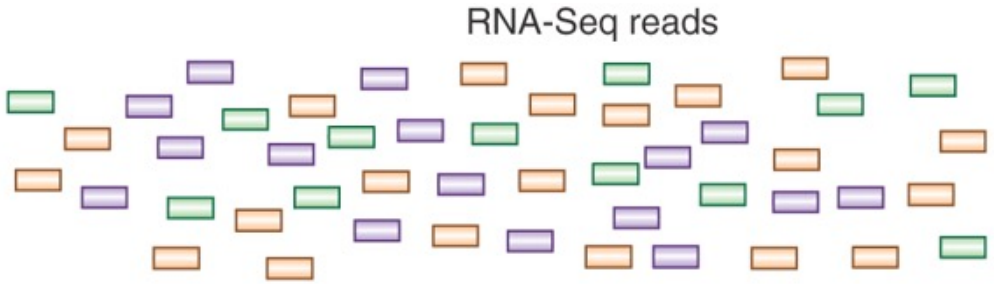


RNA-Seq Challenge: Transcript Reconstruction



Adapted from: http://www2.fml.tuebingen.mpg.de/raetsch/members/research/transcriptomics.html

Transcript Reconstruction from RNA-Seq Reads



Advancing RNA-Seq analysis

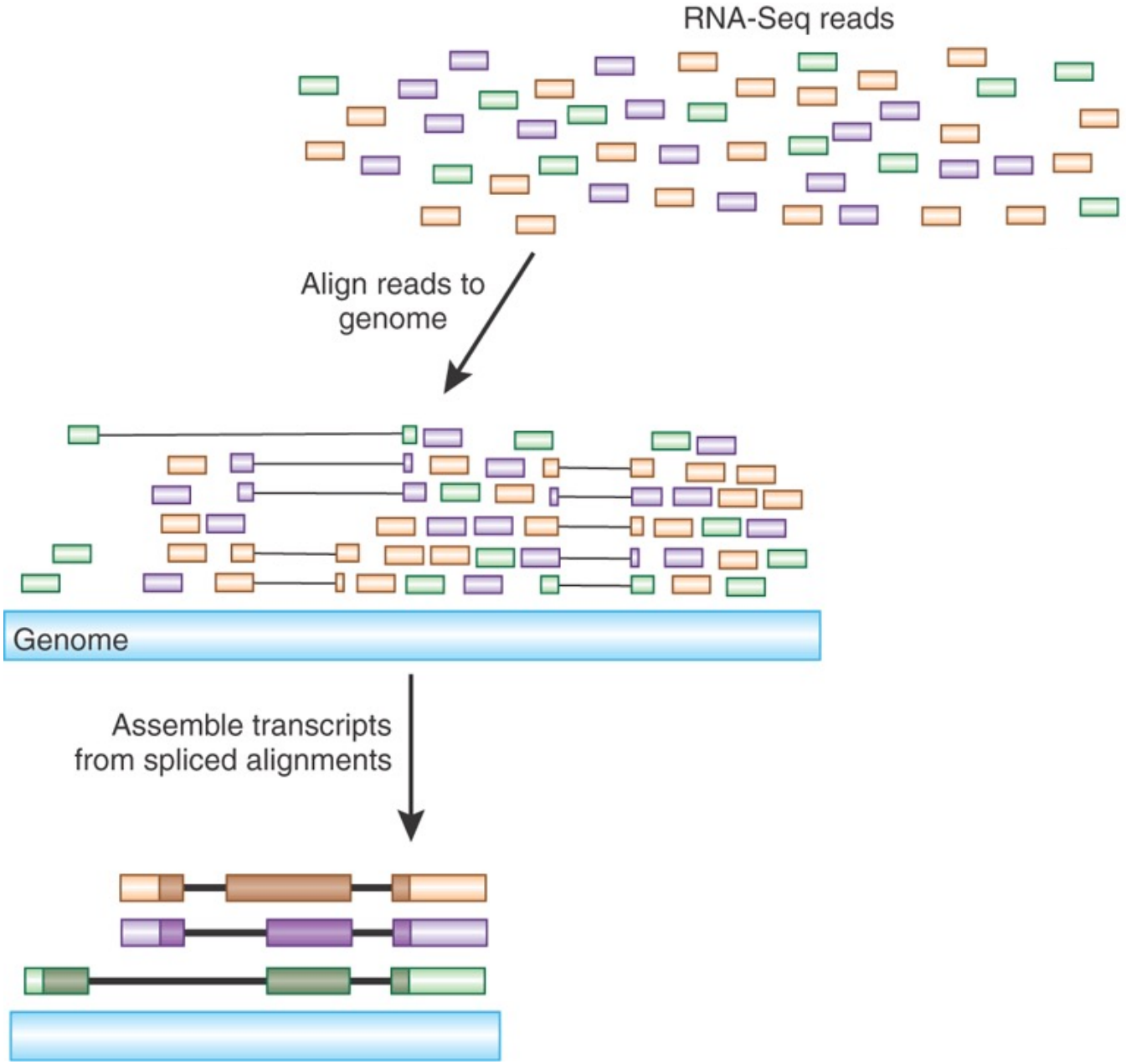
Brian J Haas & Michael C Zody

Nature Biotech, 2010

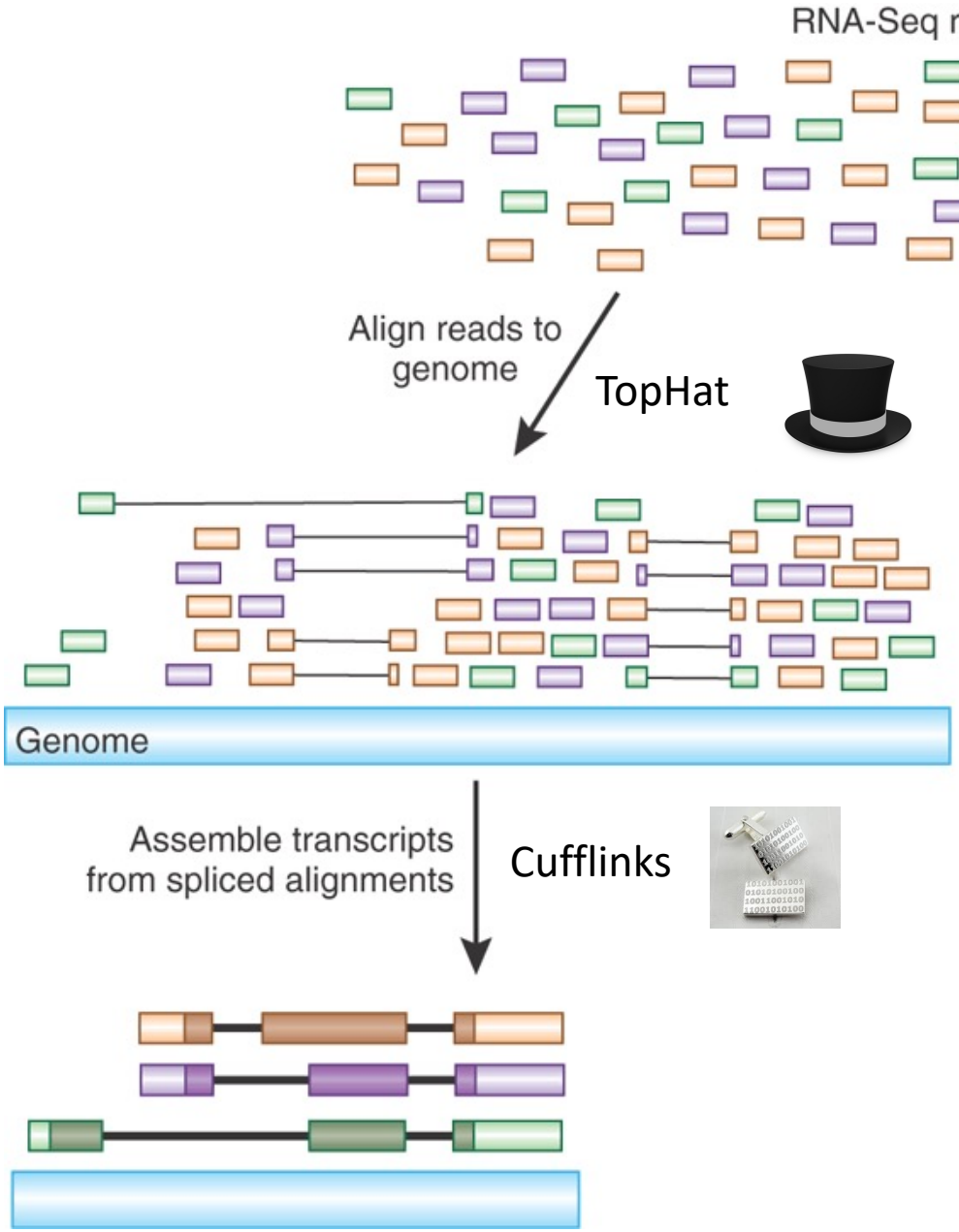
New methods for analyzing RNA-Seq data enable *de novo* reconstruction of the transcriptome.

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

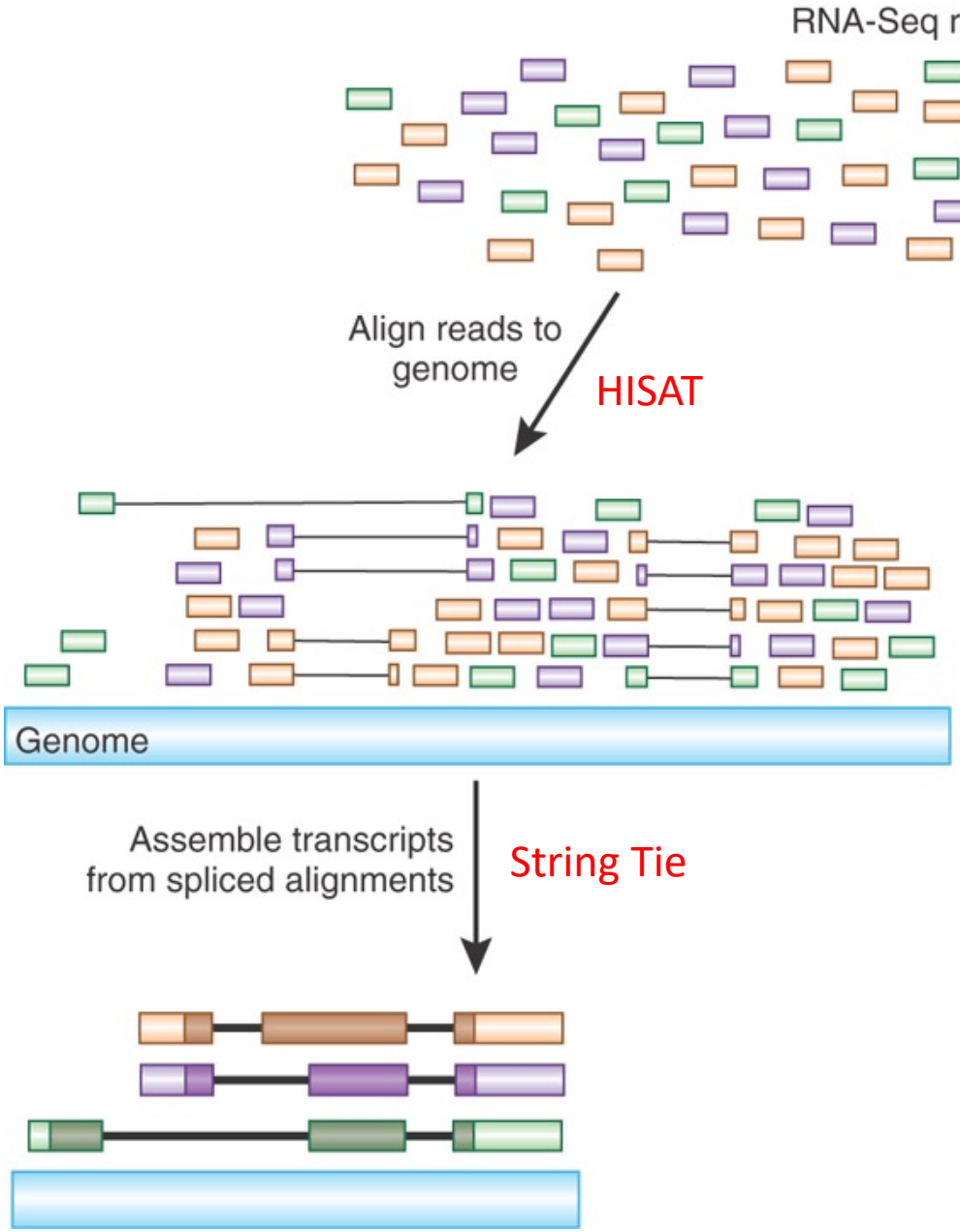
Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks

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

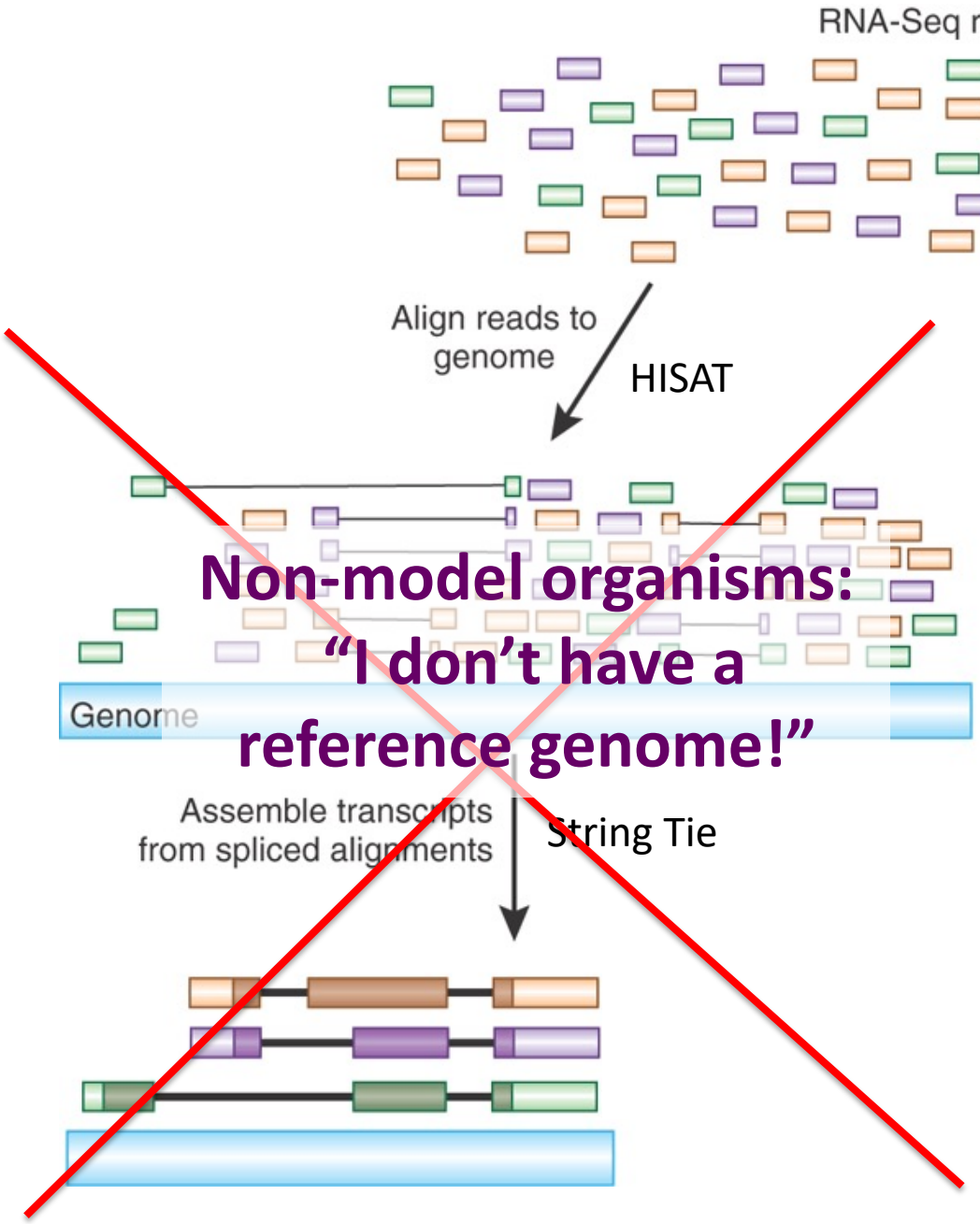
Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown

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



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

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

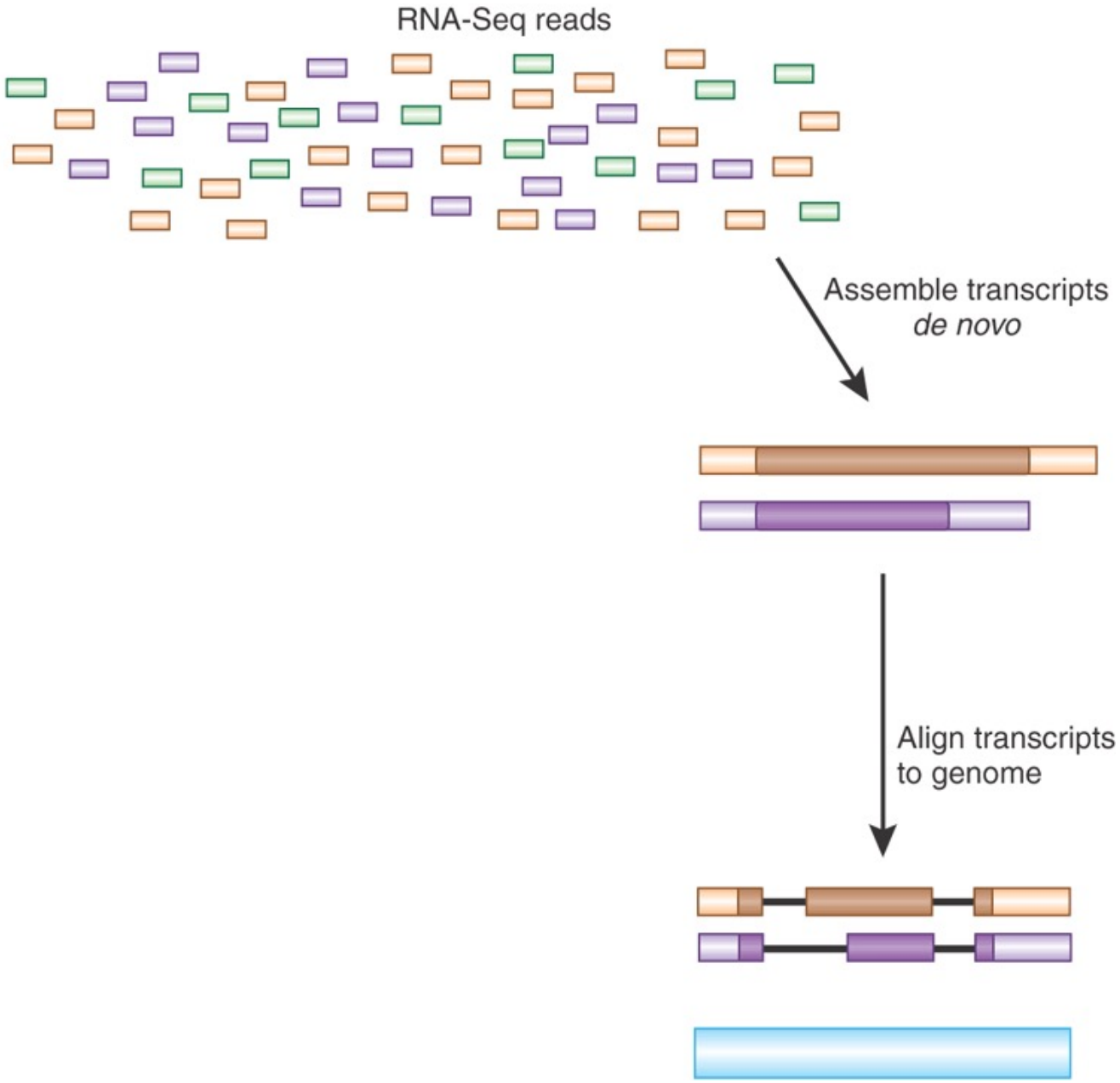
Mihaela Pertea, Daehwan Kim, Geo M Pertea, Jeffrey T Leek & Steven L Salzberg

Affiliations | Contributions | Corresponding author

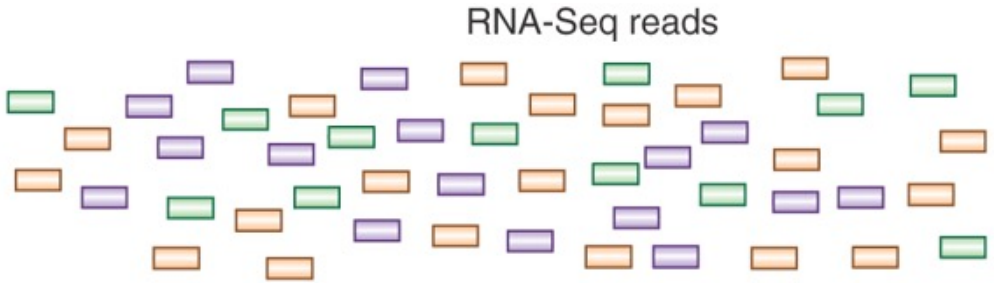
Nature Protocols **11**, 1650–1667 (2016) | doi:10.1038/nprot.2016.095
Published online 11 August 2016

Transcript Reconstruction from RNA-Seq Reads

Transcript Reconstruction from RNA-Seq Reads



Transcript Reconstruction from RNA-Seq Reads



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

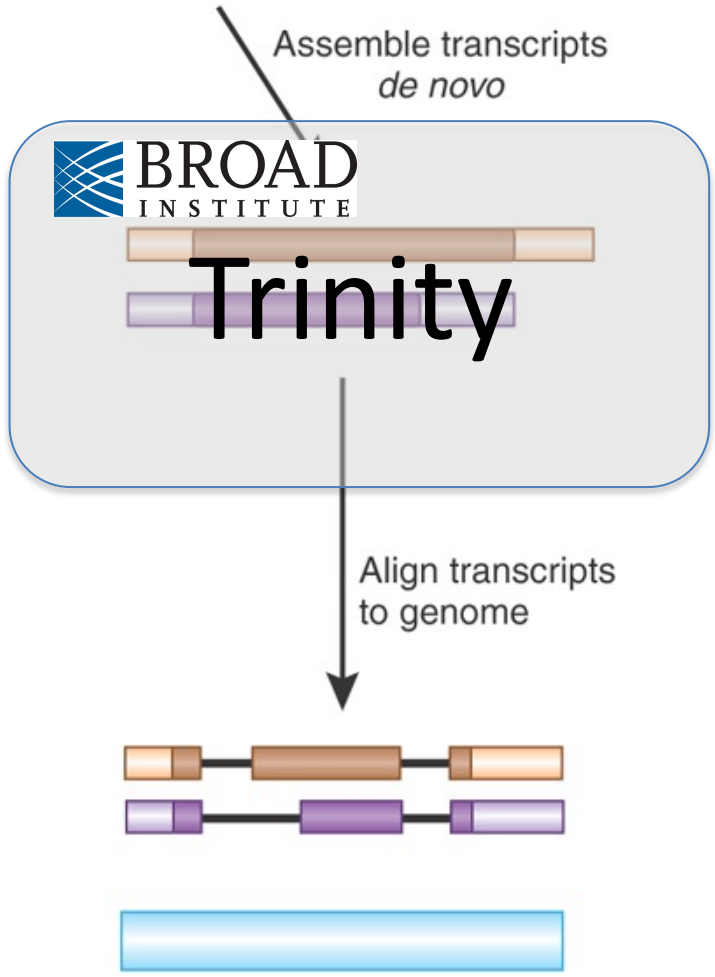
NATURE PROTOCOLS | PROTOCOL

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

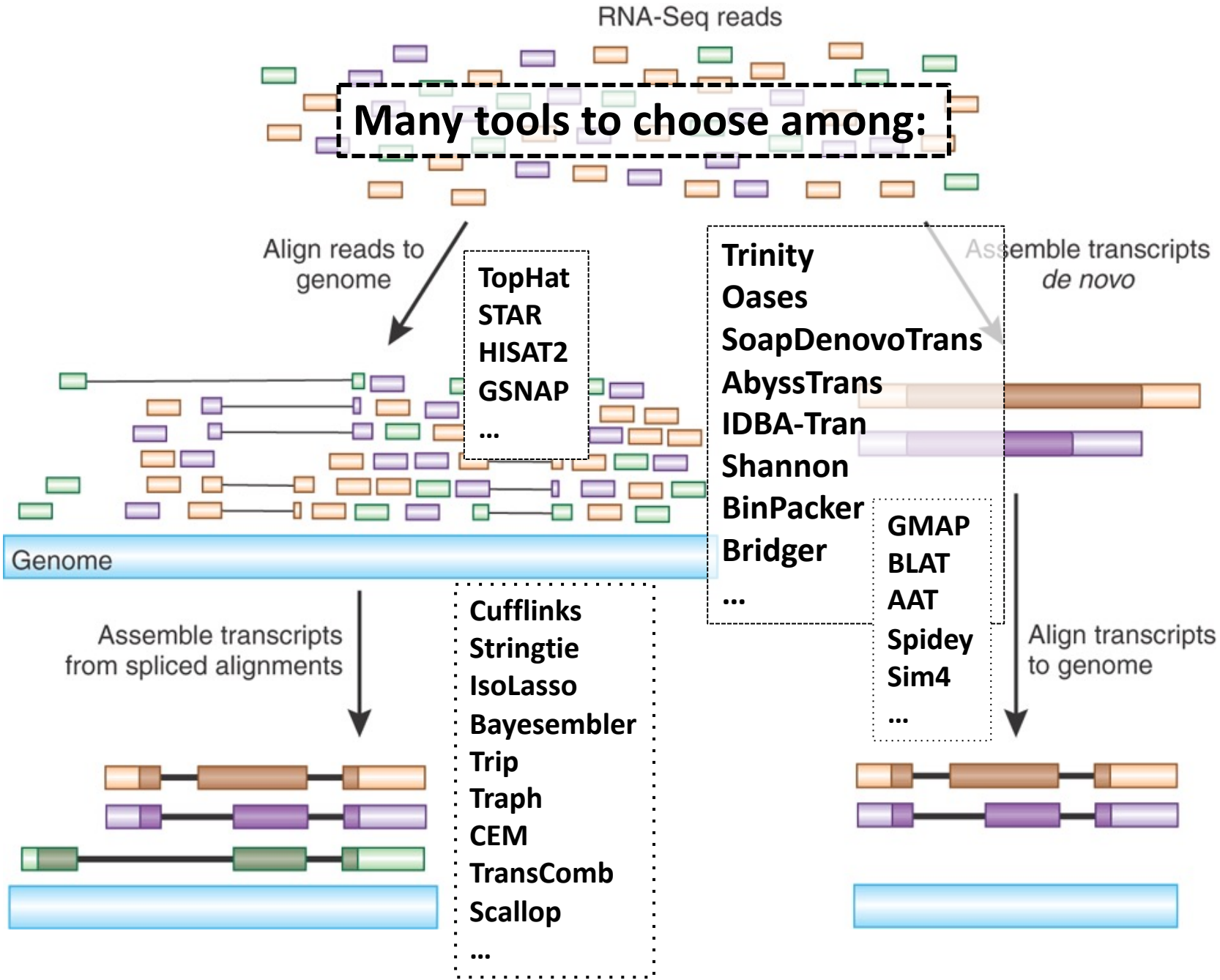
Brian J Haas, Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, Philip D Blood, Joshua Bowden, Matthew Brian Couger, David Eccles, Bo Li, Matthias Lieber, Matthew D MacManes, Michael Ott, Joshua Orvis, Nathalie Pochet, Francesco Strozzi, Nathan Weeks, Rick Westerman, Thomas William, Colin N Dewey,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

Affiliations | Contributions | Corresponding authors

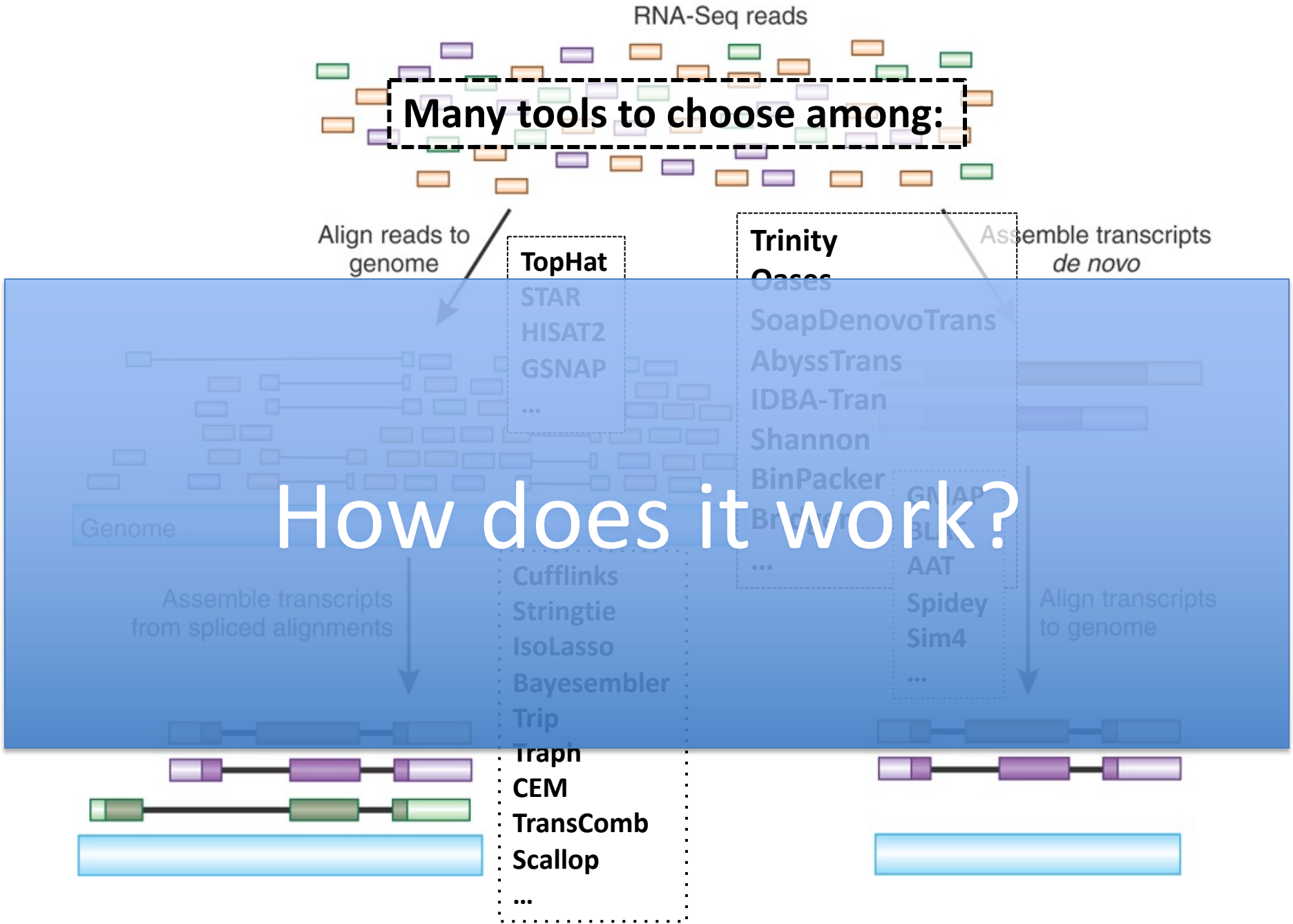
Nature Protocols **8**, 1494–1512 (2013) | doi:10.1038/nprot.2013.084
Published online 11 July 2013



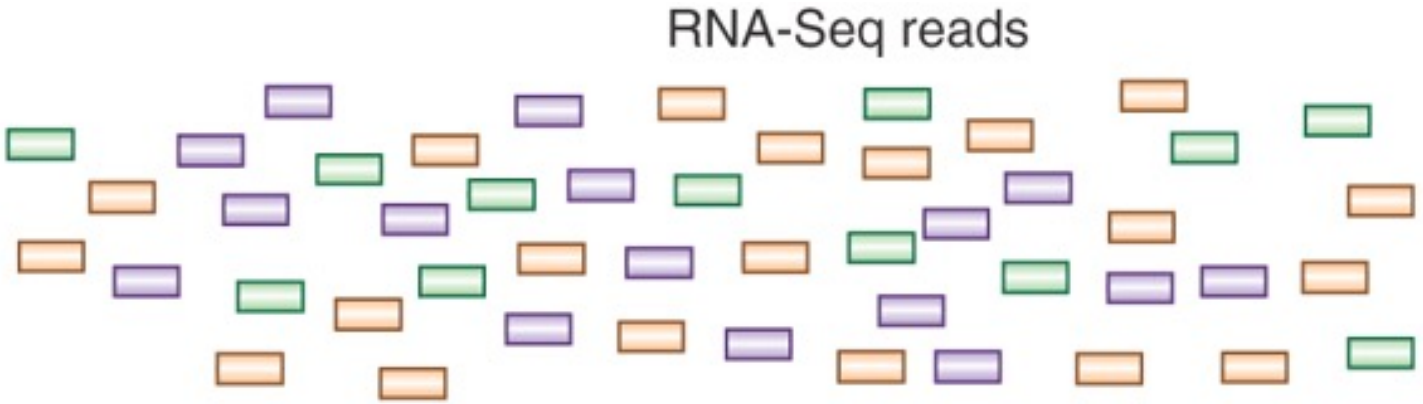
Transcript Reconstruction from RNA-Seq Reads



Transcript Reconstruction from RNA-Seq Reads

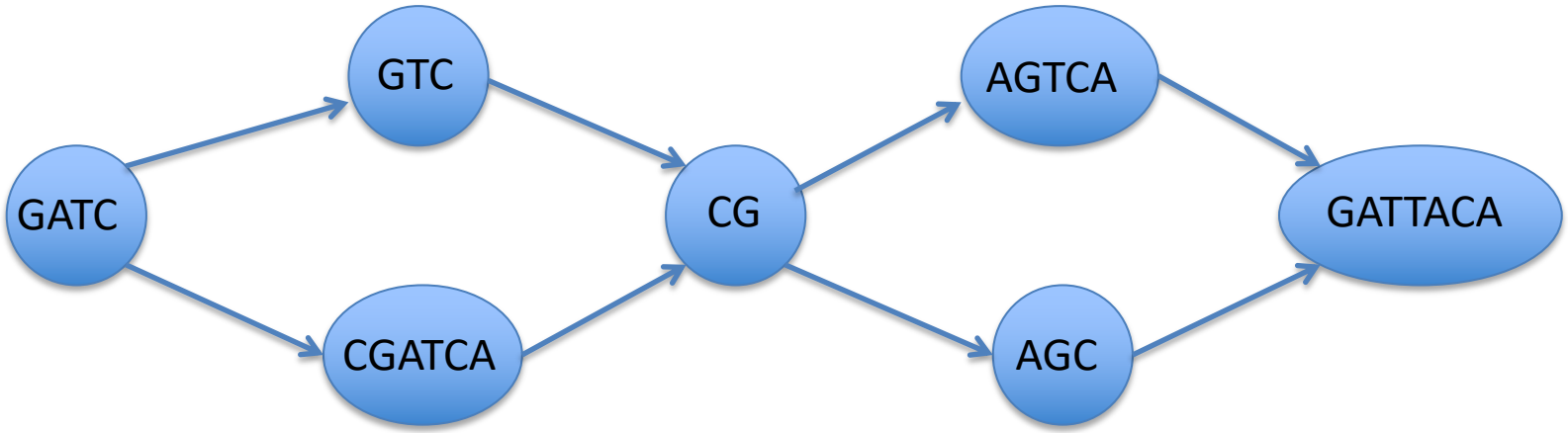


Graph Data Structures Commonly Used For Assembly



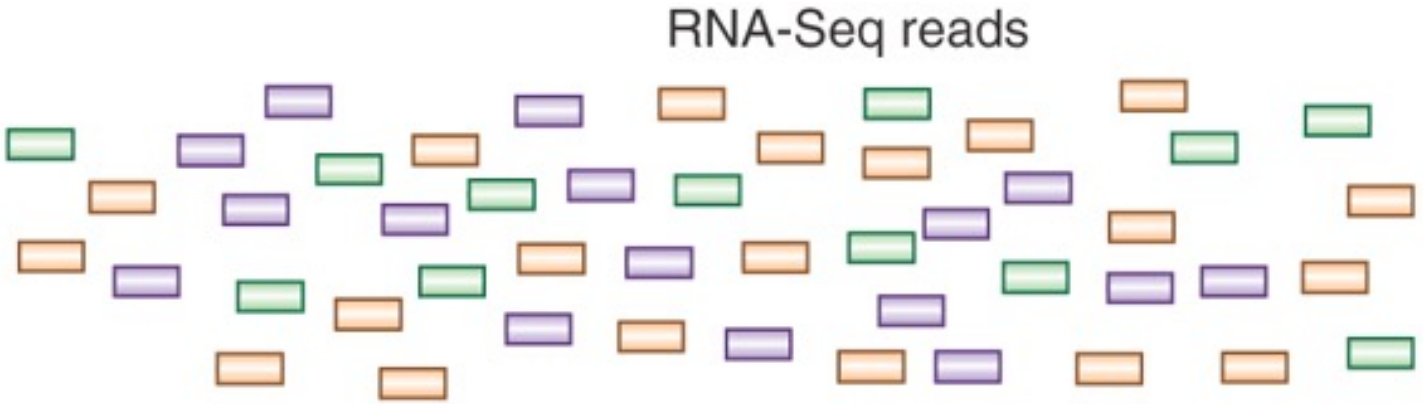
- Sequence
- Order
- Orientation (+, -)
- Overlap

Reads to Graph



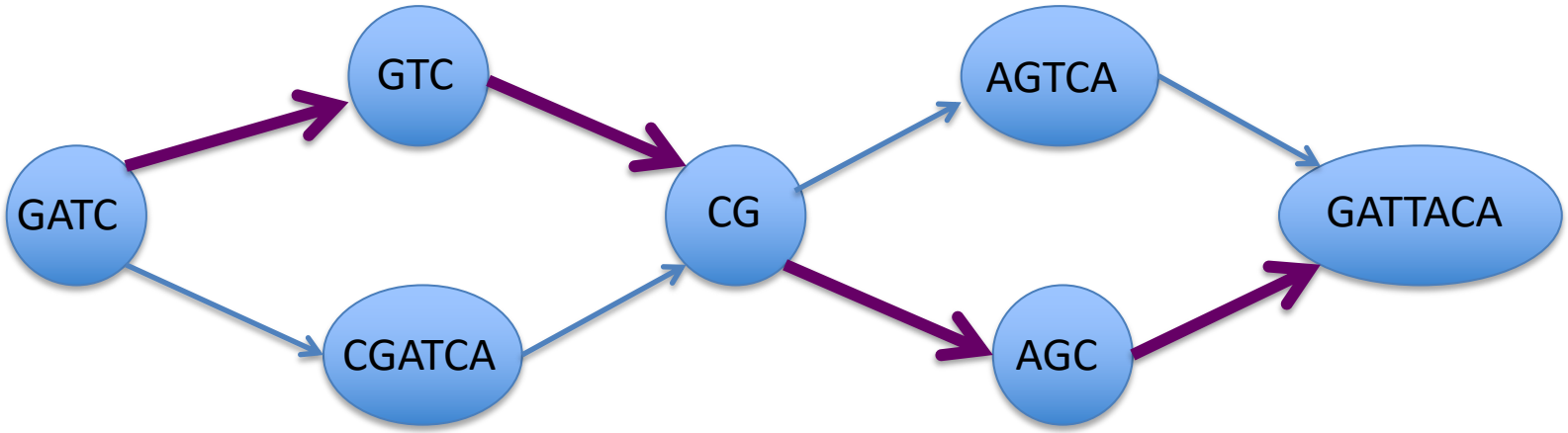
Nodes = sequence (+/-)
Edges = order, overlap

Graph Data Structures Commonly Used For Assembly



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

Reads to Graph

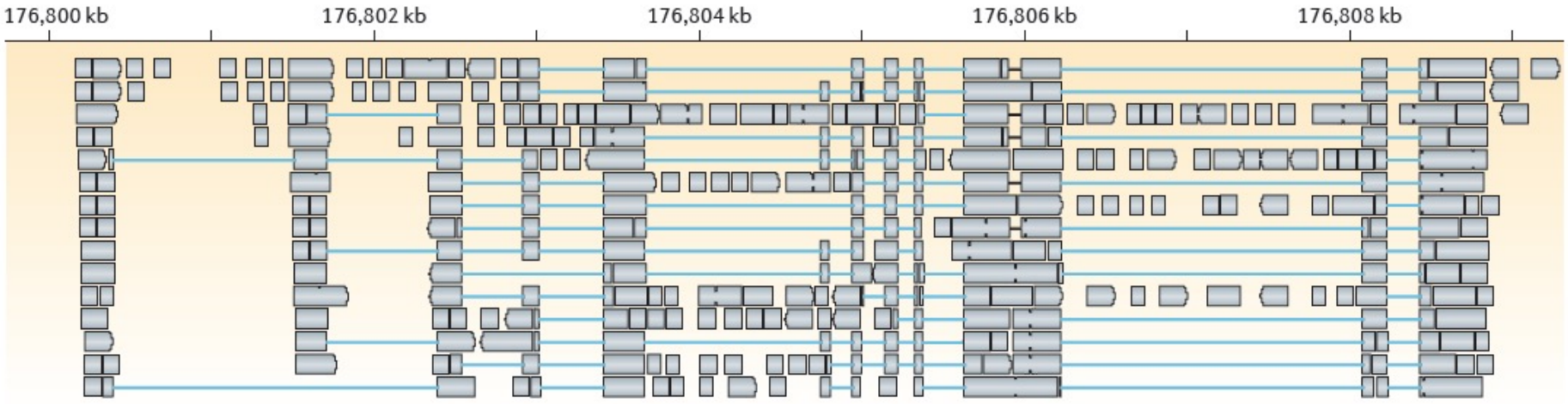


GATCGTCCGAGCGATTACA

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

Genome-Guided Transcript Reconstruction

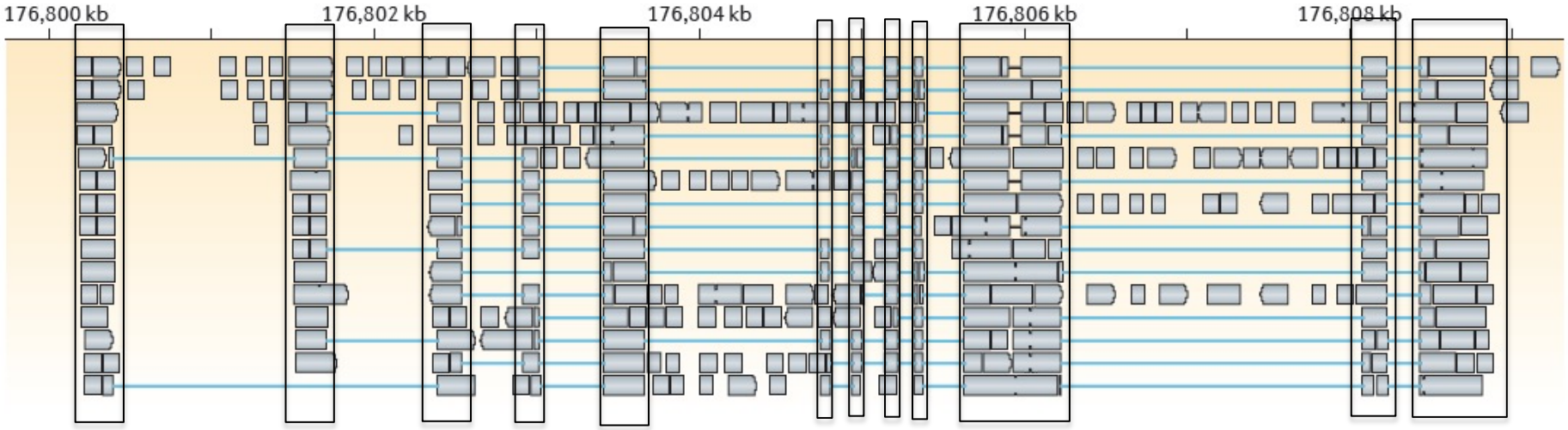
Splice-align reads to the genome



From Martin & Wang. Nature Reviews in Genetics. 2011

Genome-Guided Transcript Reconstruction

Splice-align reads to the genome

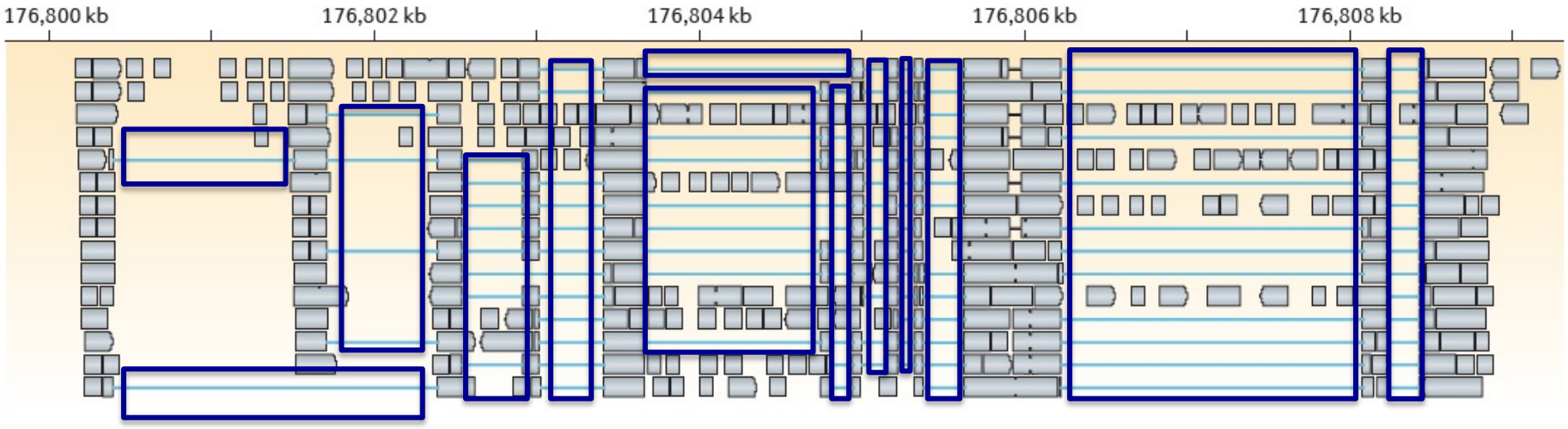


Alignment segment piles => exon regions

From Martin & Wang. Nature Reviews in Genetics. 2011

Genome-Guided Transcript Reconstruction

Splice-align reads to the genome

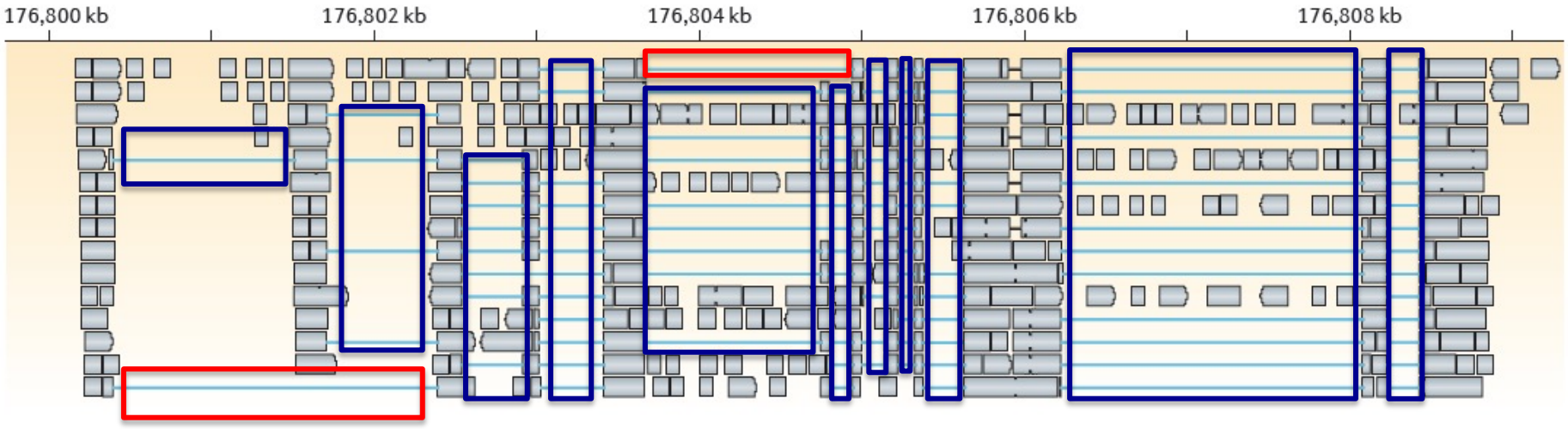


Large alignment gaps => introns

From Martin & Wang. Nature Reviews in Genetics. 2011

Genome-Guided Transcript Reconstruction

Splice-align reads to the genome

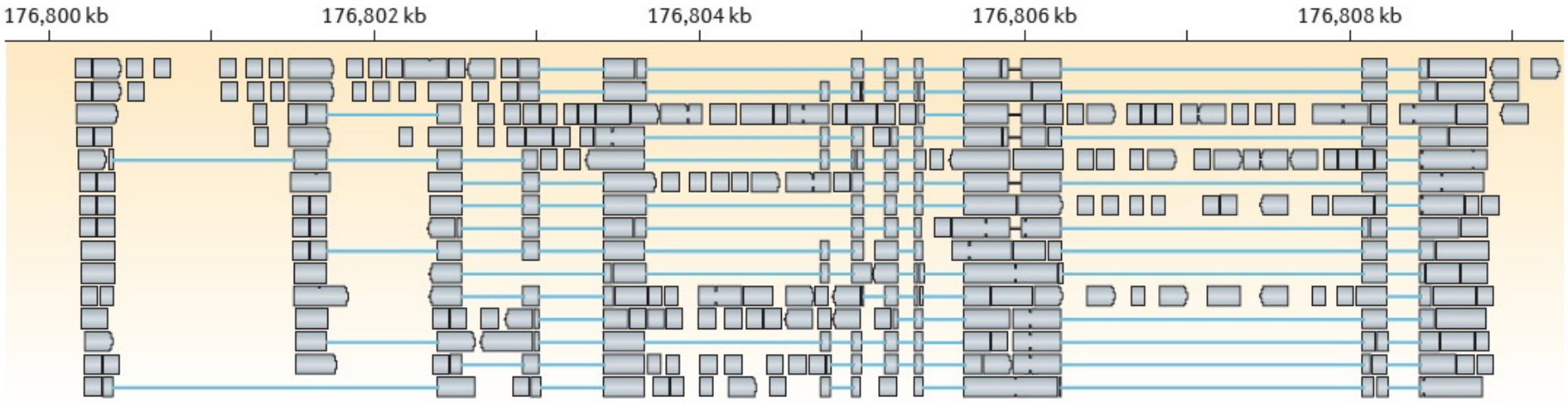


Overlapping but different introns = evidence of alternative splicing

From Martin & Wang. Nature Reviews in Genetics. 2011

Genome-Guided Transcript Reconstruction

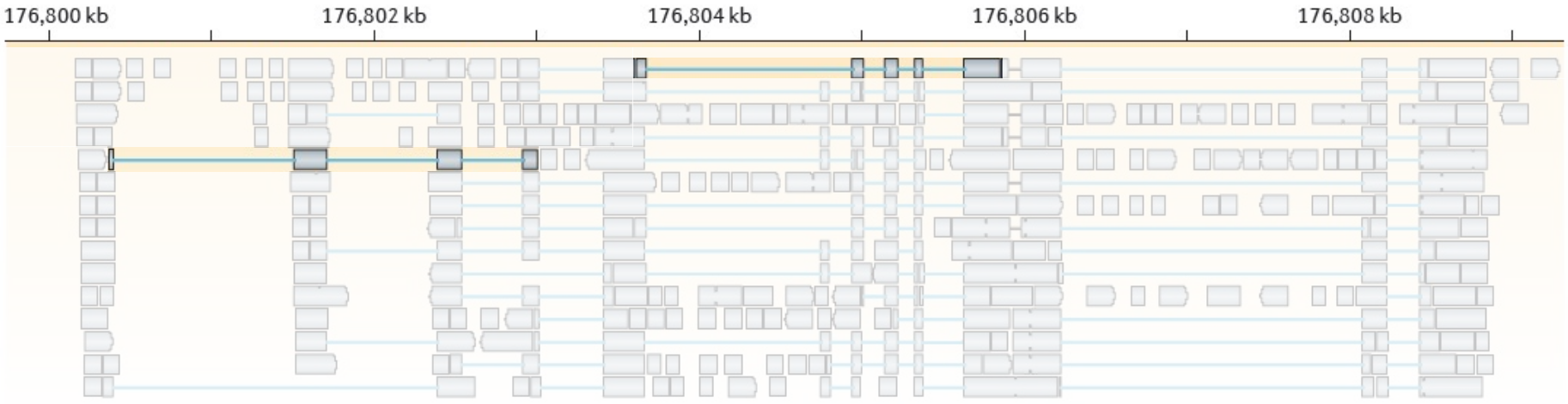
Splice-align reads to the genome



From Martin & Wang. Nature Reviews in Genetics. 2011

Genome-Guided Transcript Reconstruction

Splice-align reads to the genome

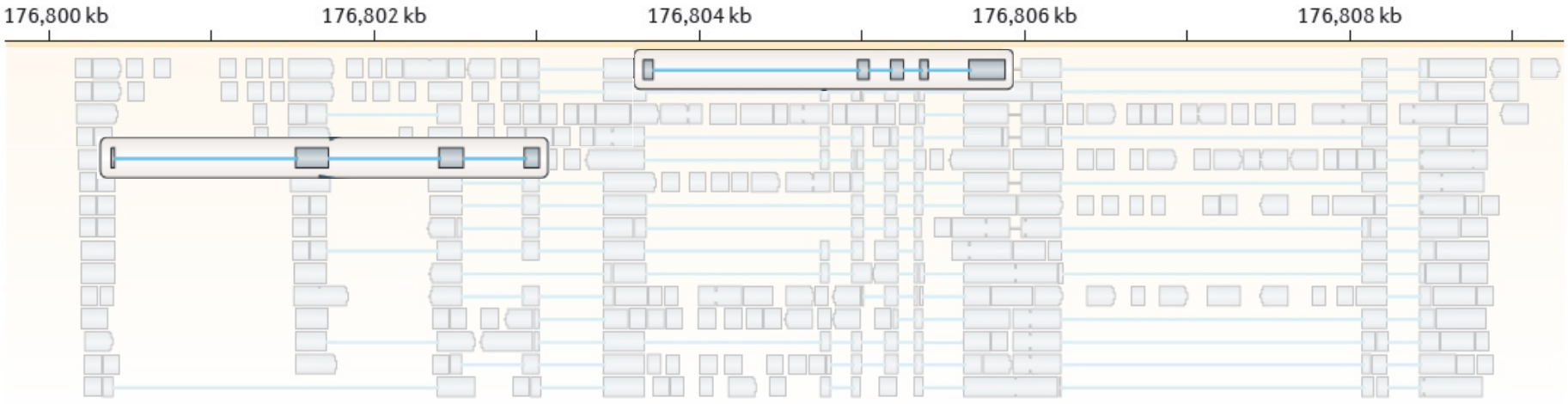


Individual reads can yield multiple exon and intron segments (splice patterns)

From Martin & Wang. Nature Reviews in Genetics. 2011

Genome-Guided Transcript Reconstruction

Splice-align reads to the genome

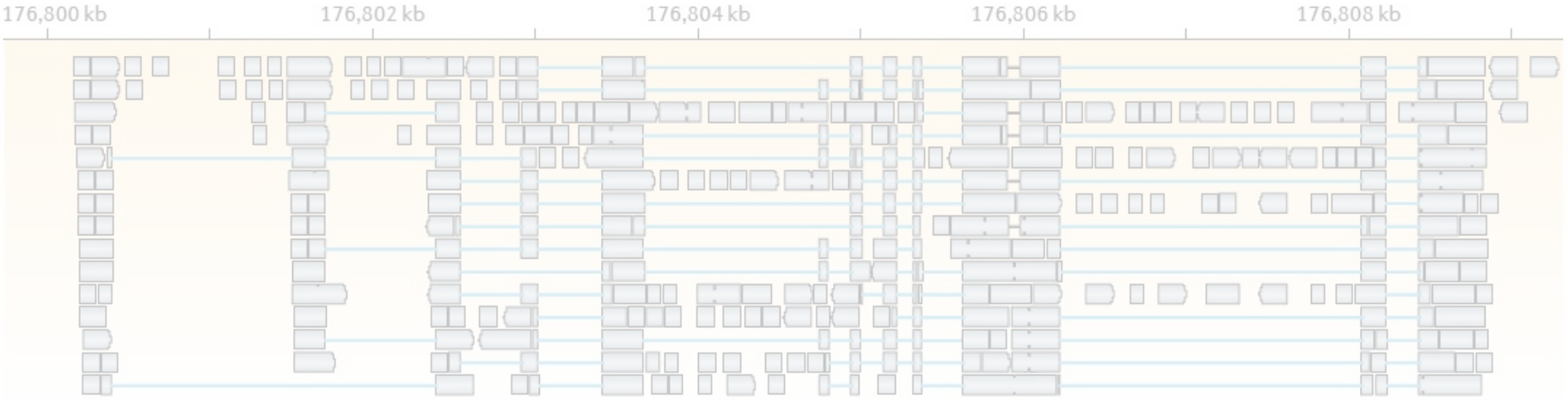


Nodes = unique splice patterns

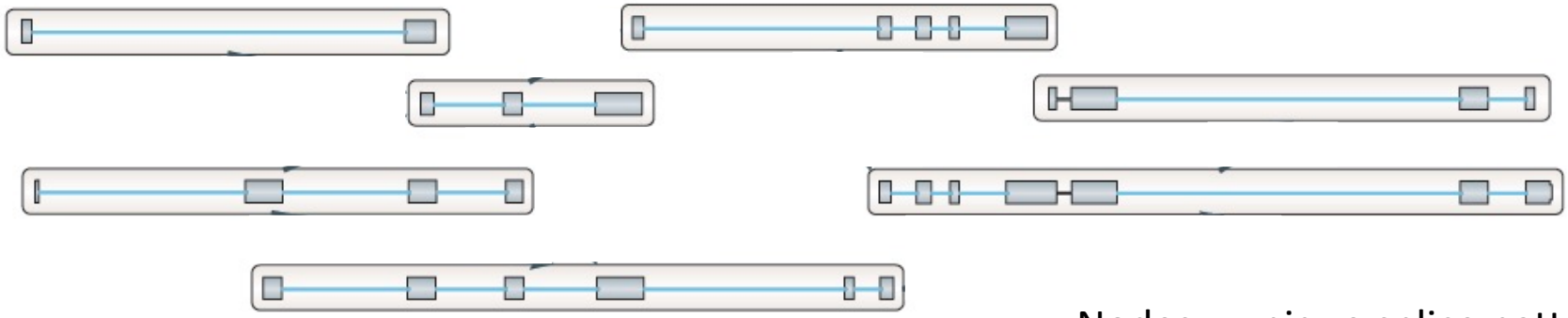
From Martin & Wang. Nature Reviews in Genetics. 2011

Genome-Guided Transcript Reconstruction

Splice-align reads to the genome



Construct graph from unique splice patterns of aligned reads.

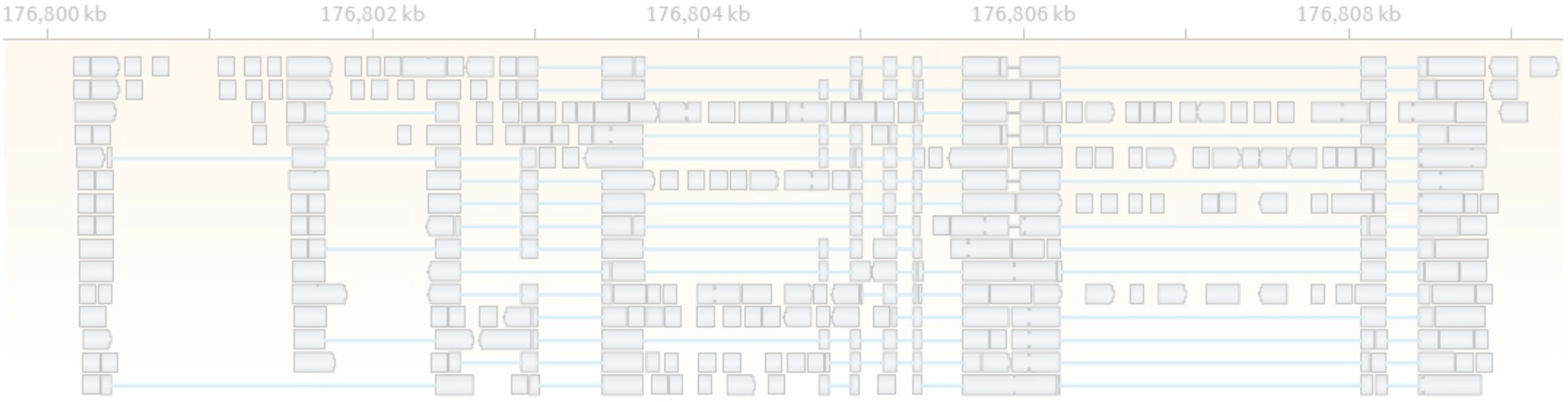


Nodes = unique splice patterns

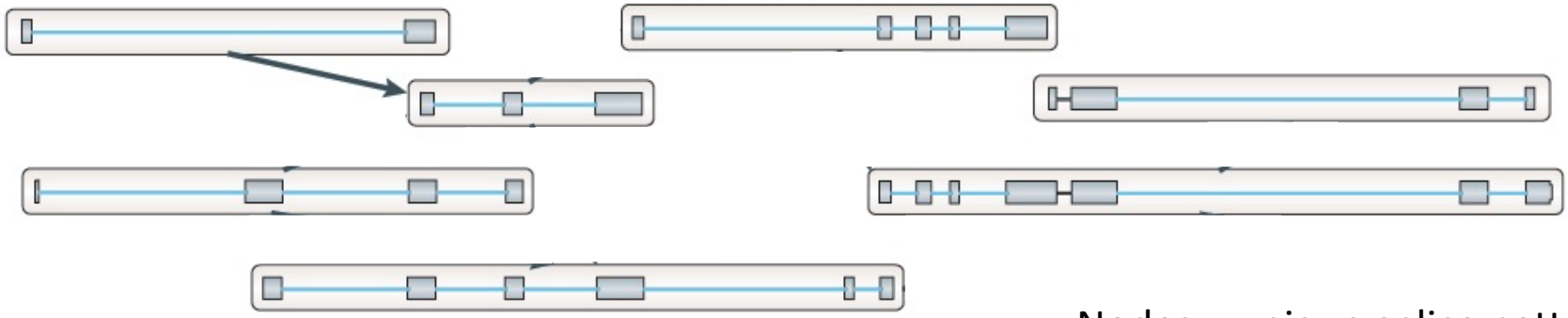
From Martin & Wang. Nature Reviews in Genetics. 2011

Genome-Guided Transcript Reconstruction

Splice-align reads to the genome



Construct graph from unique splice patterns of aligned reads.

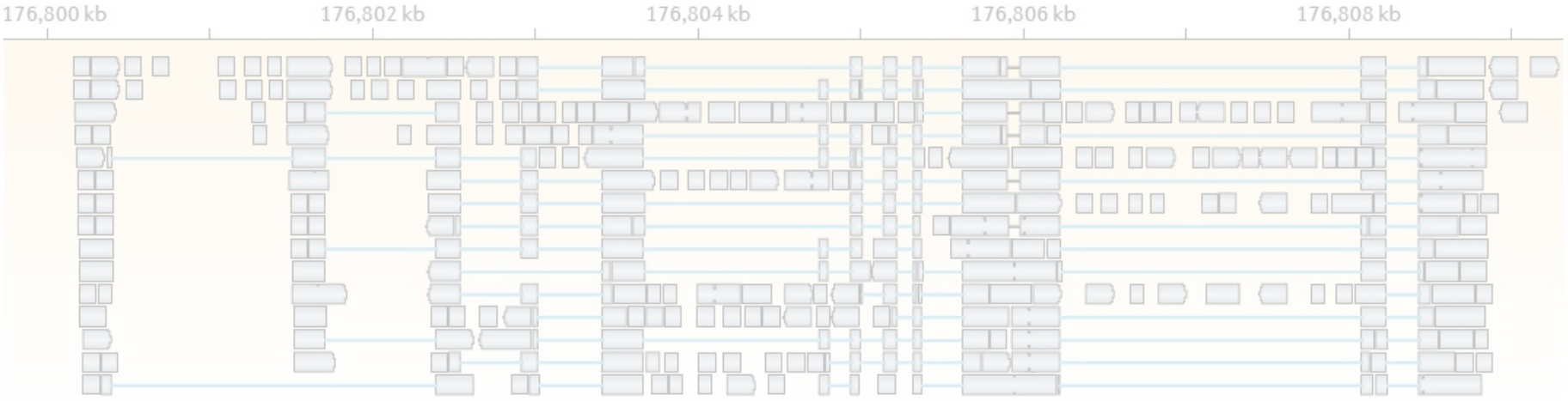


Nodes = unique splice patterns
Edges = compatible patterns

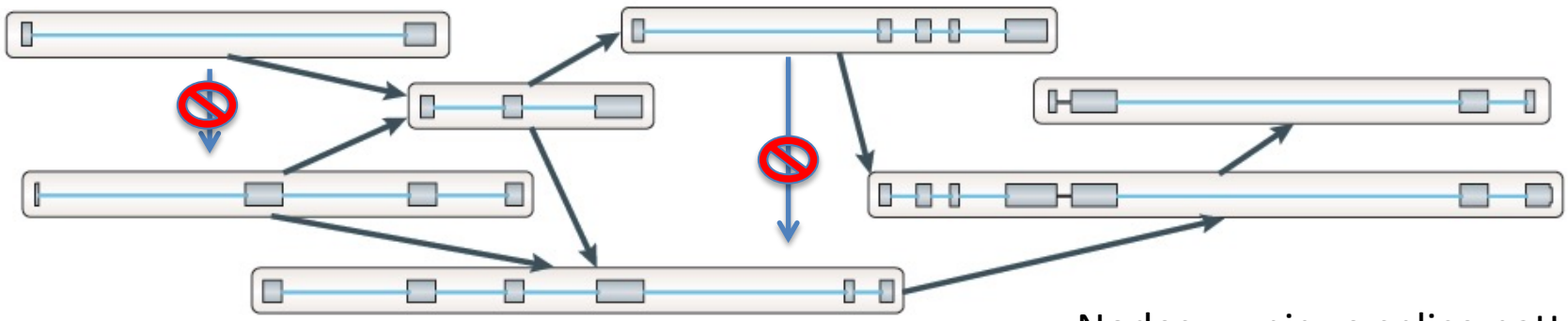
From Martin & Wang. Nature Reviews in Genetics. 2011

Genome-Guided Transcript Reconstruction

Splice-align reads to the genome



Construct graph from unique splice patterns of aligned reads.

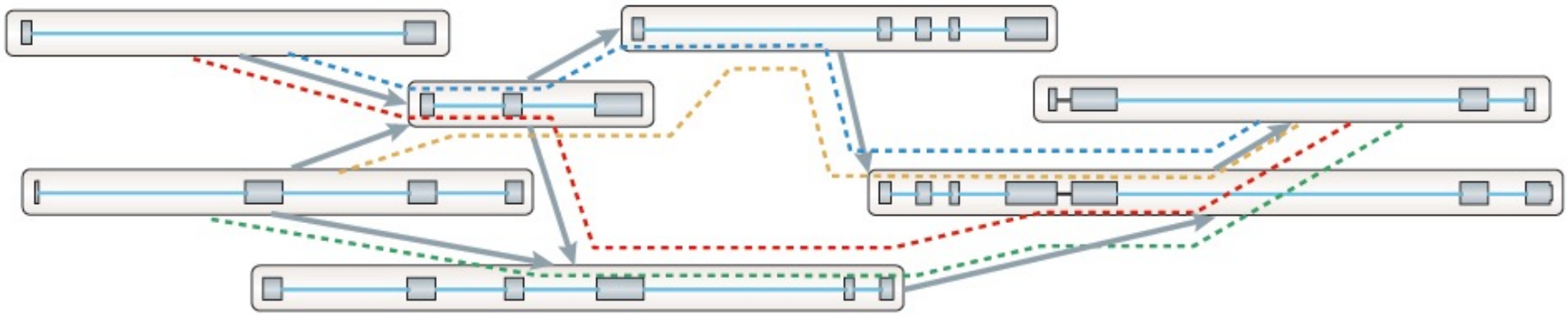


Nodes = unique splice patterns
Edges = compatible patterns

From Martin & Wang. Nature Reviews in Genetics. 2011

Genome-Guided Transcript Reconstruction

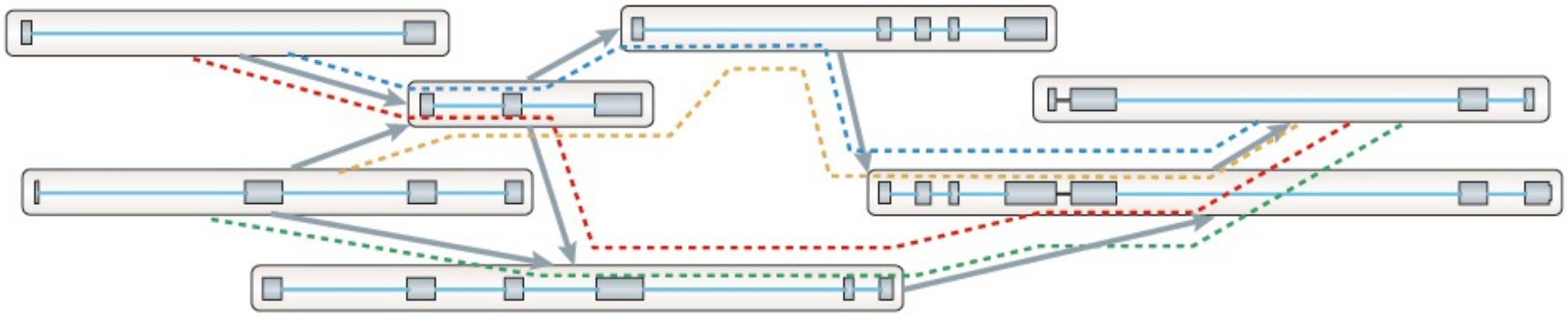
Traverse paths through the graph to assemble transcript isoforms



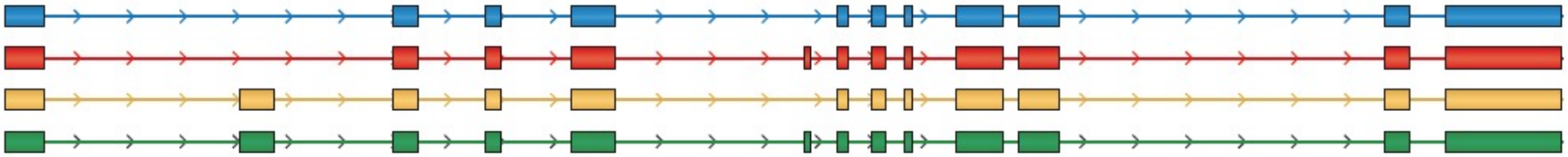
From Martin & Wang. Nature Reviews in Genetics. 2011

Genome-Guided Transcript Reconstruction

Traverse paths through the graph to assemble transcript isoforms



Reconstructed isoforms

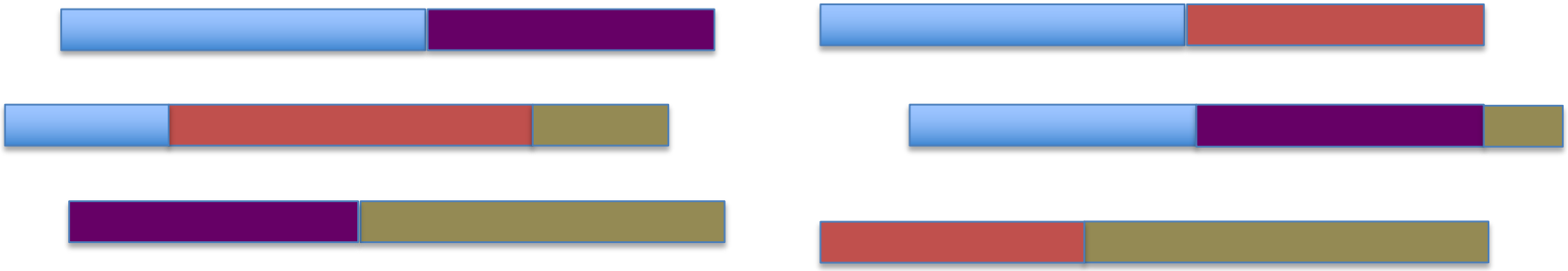


From Martin & Wang. Nature Reviews in Genetics. 2011

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

Genome-free de novo transcript reconstruction to the rescue.

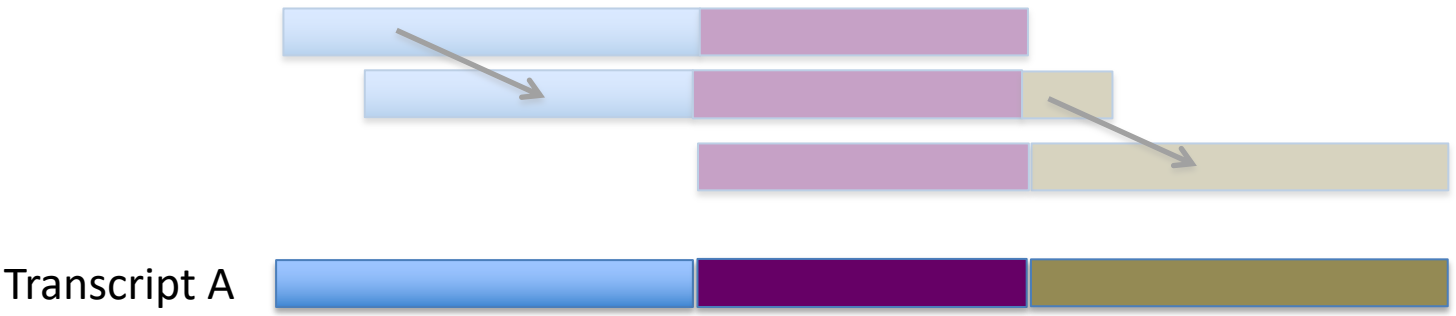
Read Overlap Graph: Reads as nodes, overlaps as edges



Read Overlap Graph: Reads as nodes, overlaps as edges

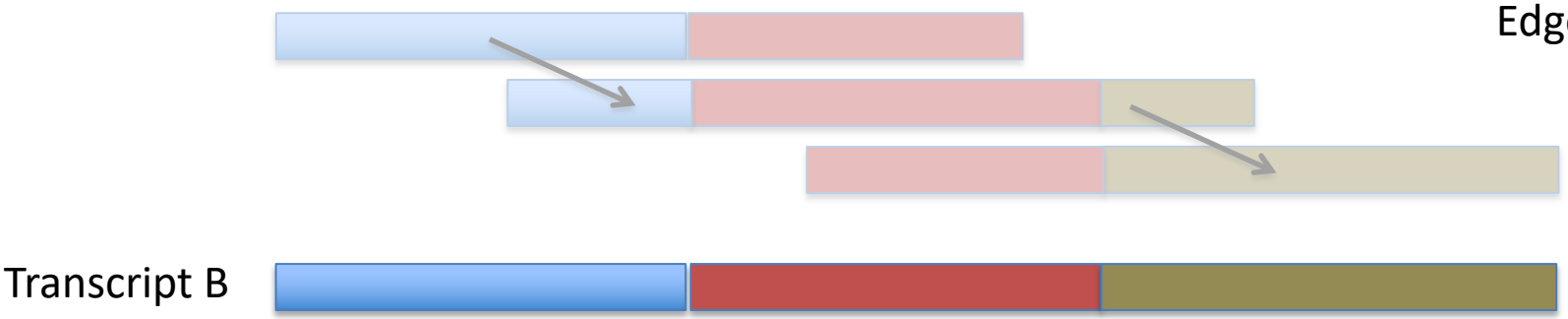
Node = read
Edge = overlap

Read Overlap Graph: Reads as nodes, overlaps as edges

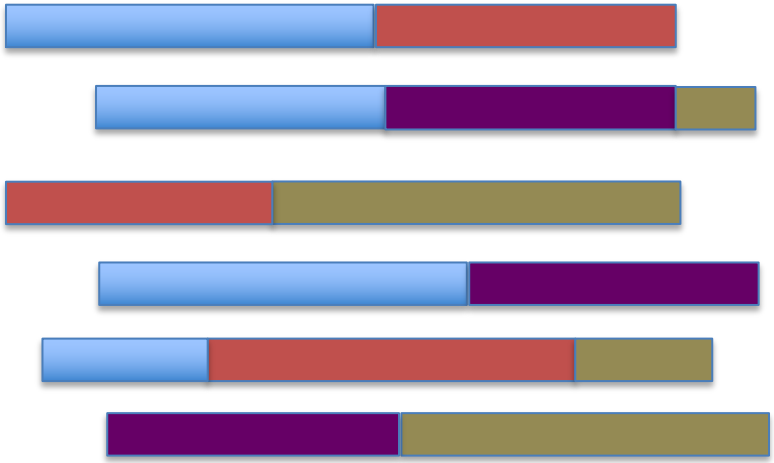


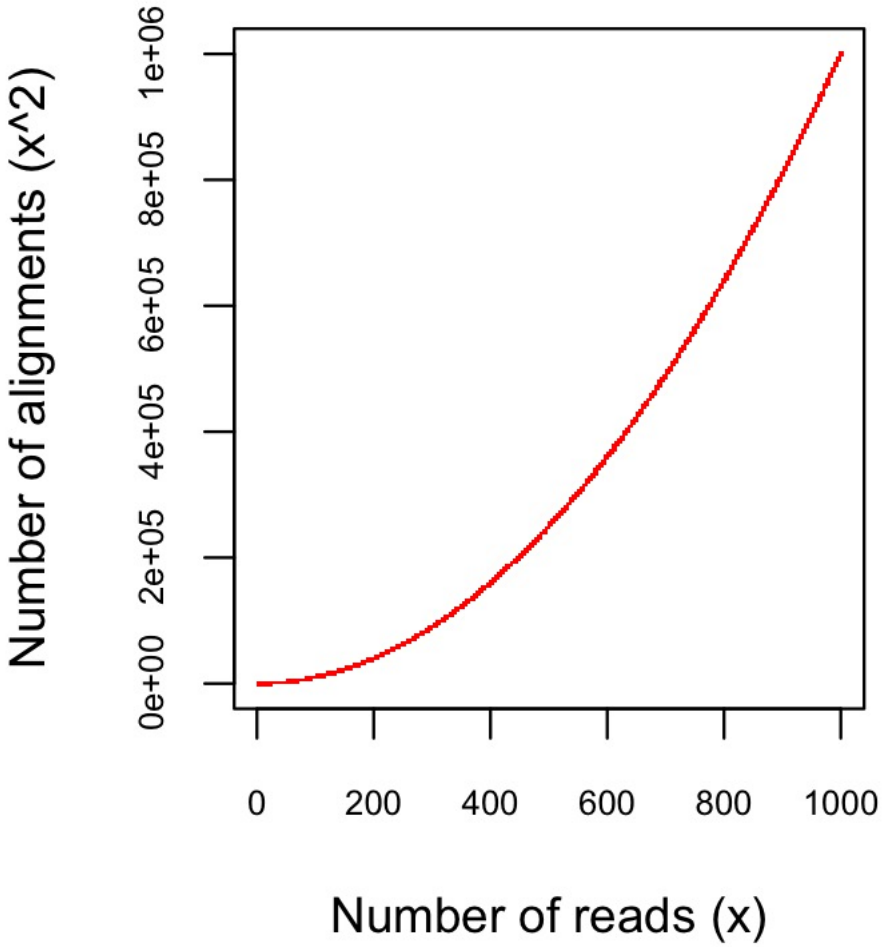
Generate consensus sequence where reads overlap

Node = read
Edge = overlap



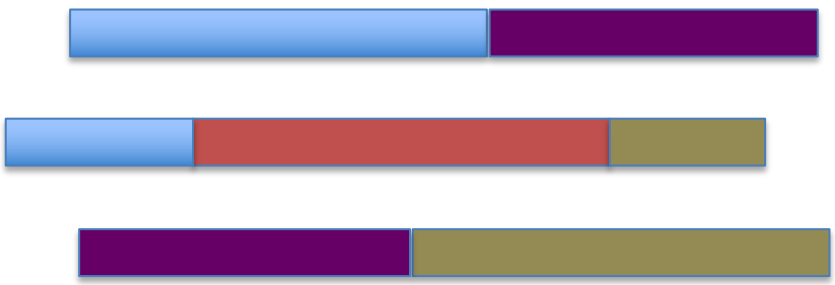
Finding pairwise overlaps between *n* reads involves ~ n^2 comparisons.



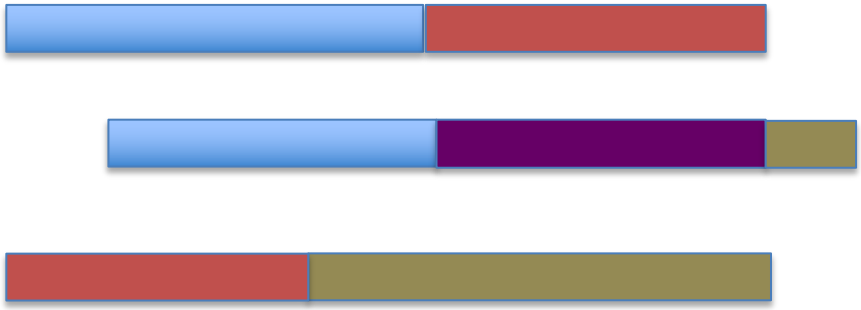


Impractical for typical RNA-Seq data (50M reads)

No genome to align to…



De novo assembly required



Want to avoid n^2 read alignments to define overlaps

Use a de Bruijn graph

Have you learned about the de Bruijn graph already?

Sequence Assembly via de Bruijn Graphs

Generate all substrings of length k from the reads

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

Sequence Assembly via De Bruijn Graphs

Generate all substrings of length k from the reads

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

Sequence Assembly via De Bruijn Graphs

Generate all substrings of length k from the reads

Construct the de Bruijn graph

Nodes = unique k-mers

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

Sequence Assembly via De Bruijn Graphs

Generate all substrings of length k from the reads

Construct the de Bruijn graph

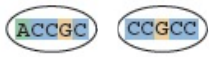
Nodes = unique k-mers
Edges = overlap by (k-1)

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

Sequence Assembly via De Bruijn Graphs

Generate all substrings of length k from the reads

Construct the de Bruijn graph



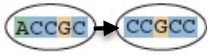
From Martin & Wang, Nat. Rev. Genet. 2011

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

Sequence Assembly via De Bruijn Graphs

Generate all substrings of length k from the reads

Construct the de Bruijn graph

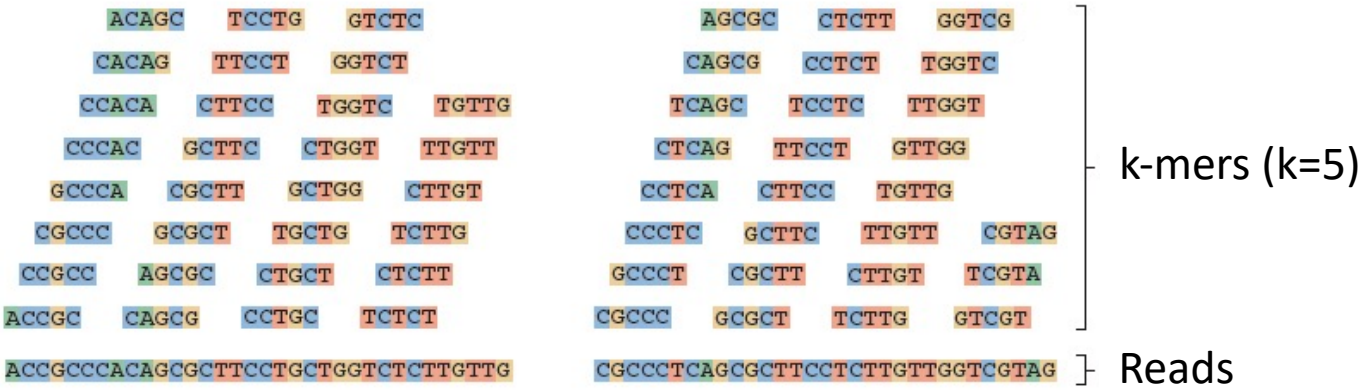


From Martin & Wang, Nat. Rev. Genet. 2011

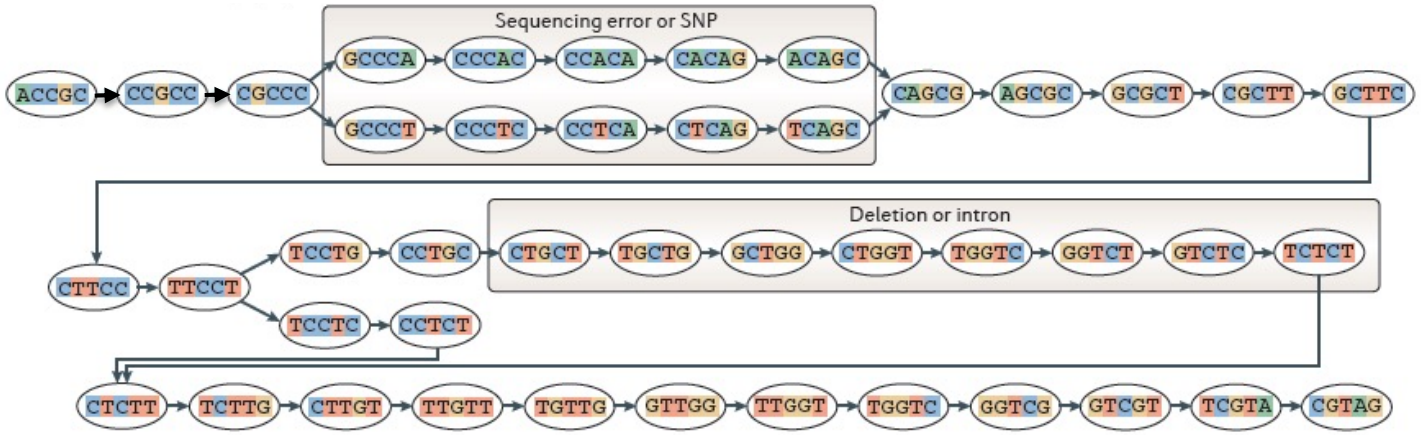
Nodes = unique k-mers
Edges = overlap by (k-1)

Sequence Assembly via De Bruijn Graphs

Generate all substrings of length k from the reads



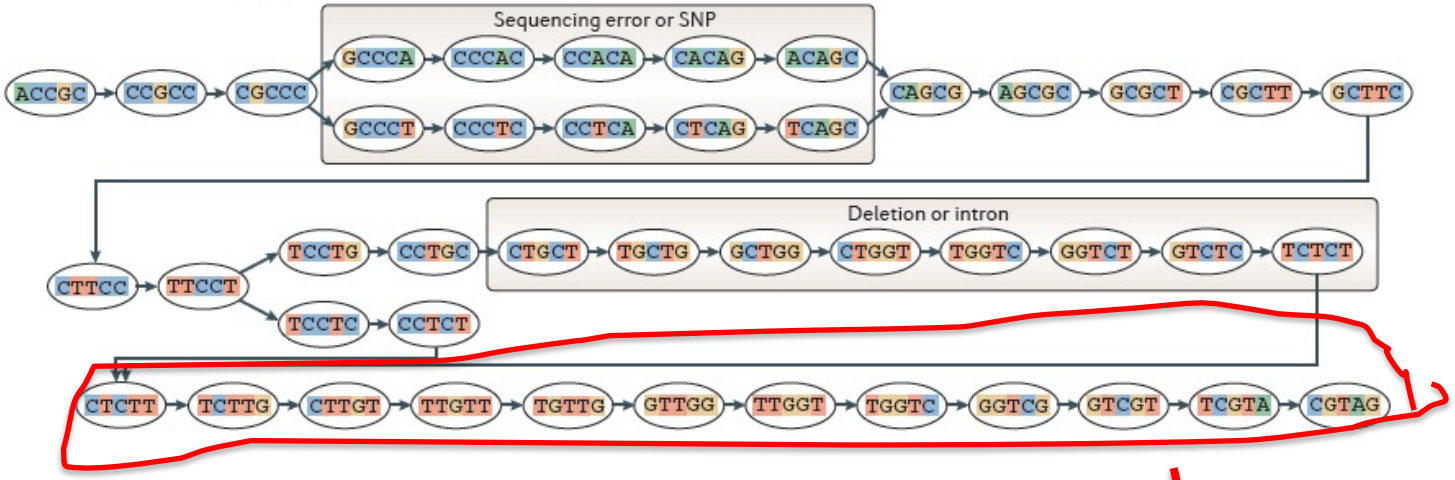
Construct the de Bruijn graph



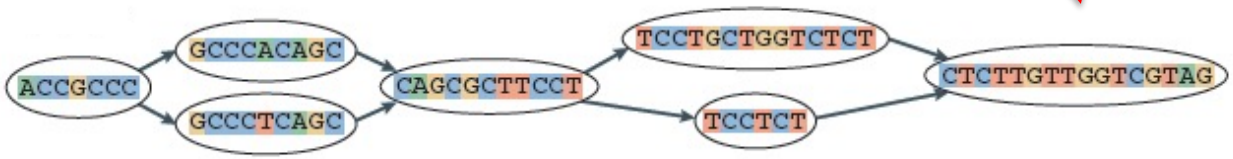
Nodes = unique k-mers
Edges = overlap by (k-1)

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

Construct the de Bruijn graph

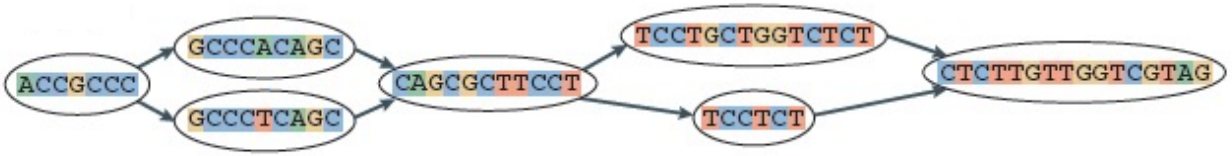


Collapse the de Bruijn graph

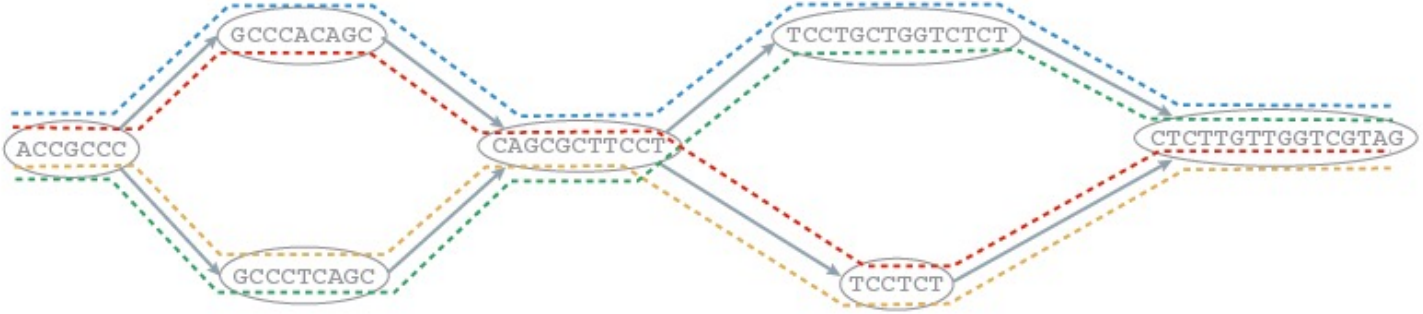


From Martin & Wang, Nat. Rev. Genet. 2011

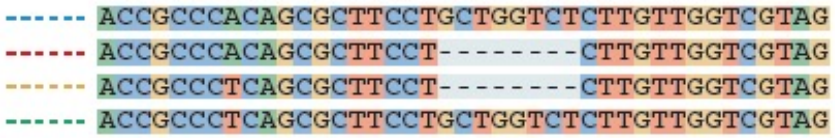
Collapse the de Bruijn graph



Traverse the graph

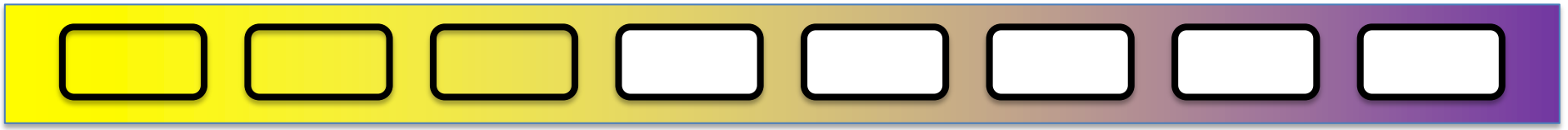


Assemble Transcript Isoforms



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

Part 3. Trinity De novo Assembly



Contrasting Genome and Transcriptome Assembly

Genome Assembly

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

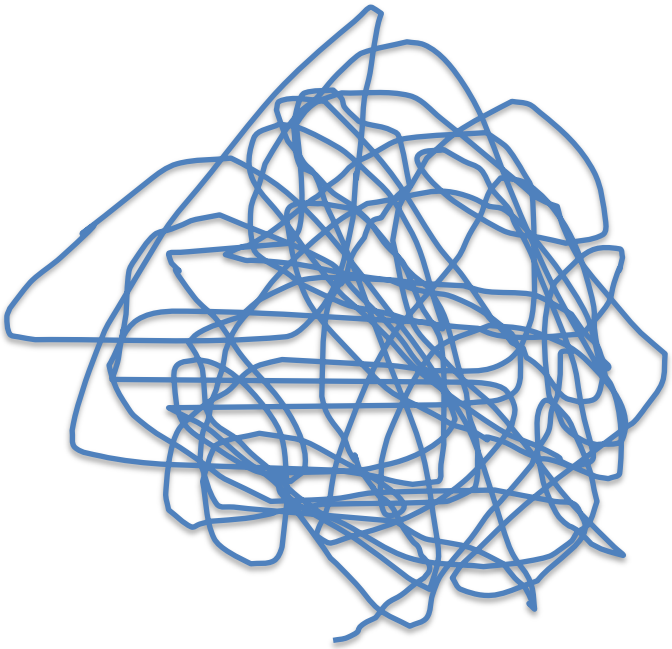
Transcriptome Assembly

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

Trinity Aggregates Isolated Transcript Graphs

Genome Assembly

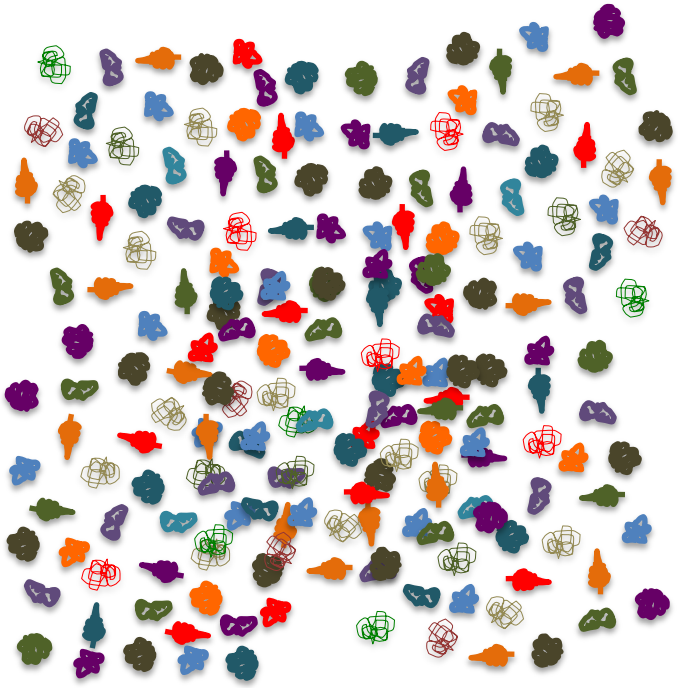
Single Massive Graph



Entire chromosomes represented.

Trinity Transcriptome Assembly

Many Thousands of Small Graphs



Ideally, one graph per expressed gene.

Trinity – How it works:

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

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

Thousands of disjoint graphs

Trinity – How it works:

Younger me

Manfred Grabherr

Moran Yassour

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

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

Thousands of disjoint graphs

Trinity – How it works:

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

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

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

Thousands of disjoint graphs

Inchworm Algorithm

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

Read: **AATGTGAAAACTGGATTACATGCTGGTATGTC**…

AATGTGA

ATGTGAA Overlapping kmers of length (k)

TGTGAAA

…

Kmer Catalog (hashtable)

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

Inchworm Algorithm

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

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

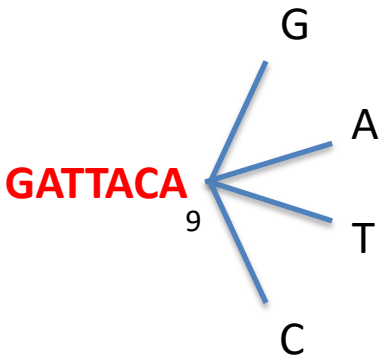
GATTACA
9

Kmer Catalog (hashtable)

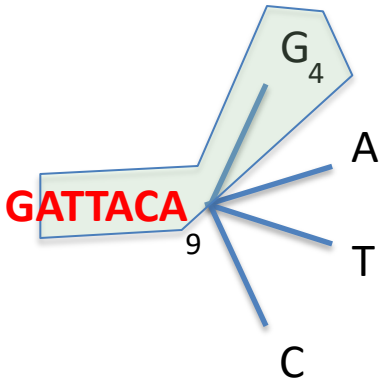
Kmer	Count among all reads
AATGTGA	4
ATGTGAA	2
TGTGAAA	1
GATTACA	9

Inchworm Algorithm

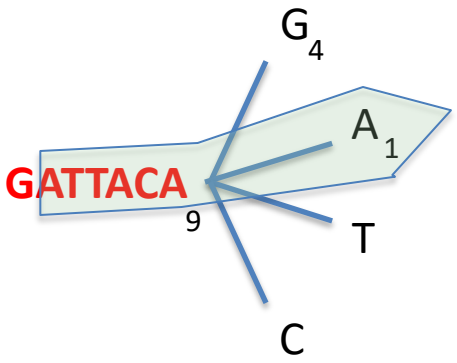
- Decompose all reads into overlapping Kmers => hashtable(kmer, count)
- Identify seed kmer as most abundant Kmer, ignoring low-complexity kmers.
- Extend kmer at 3’ end, guided by coverage.



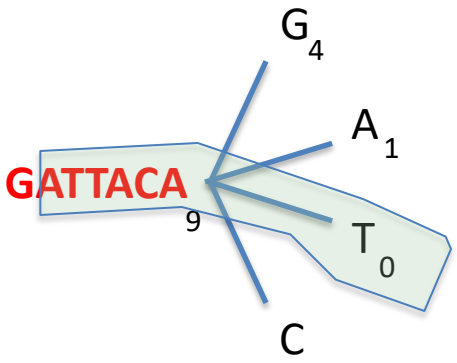
Inchworm Algorithm



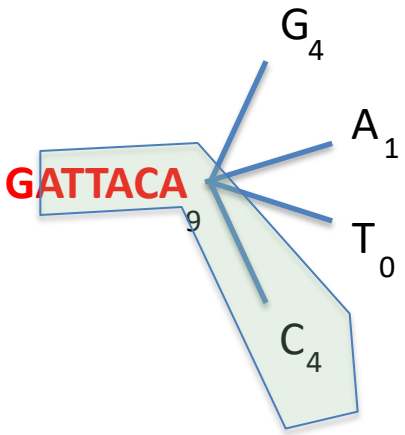
Inchworm Algorithm



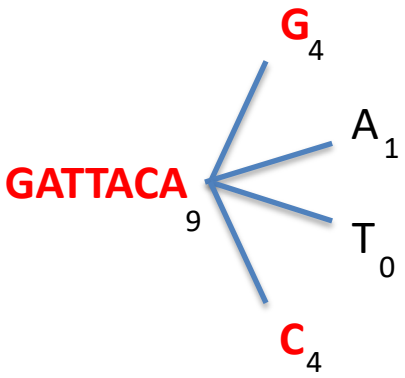
Inchworm Algorithm



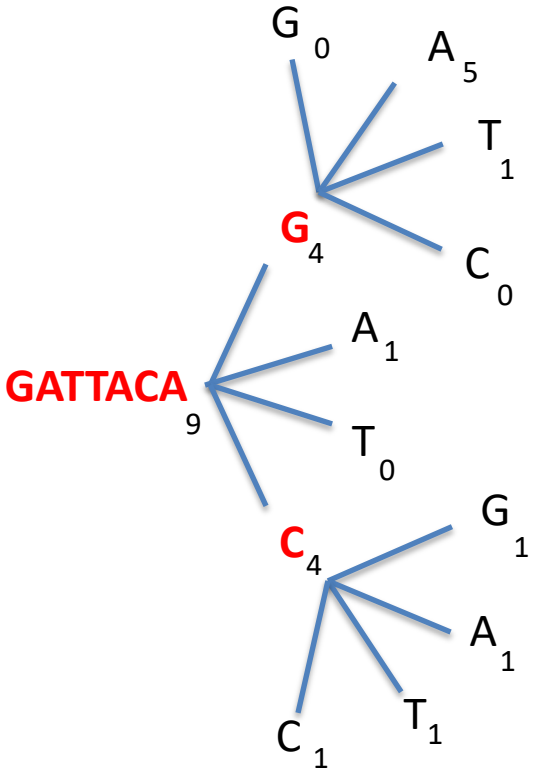
Inchworm Algorithm



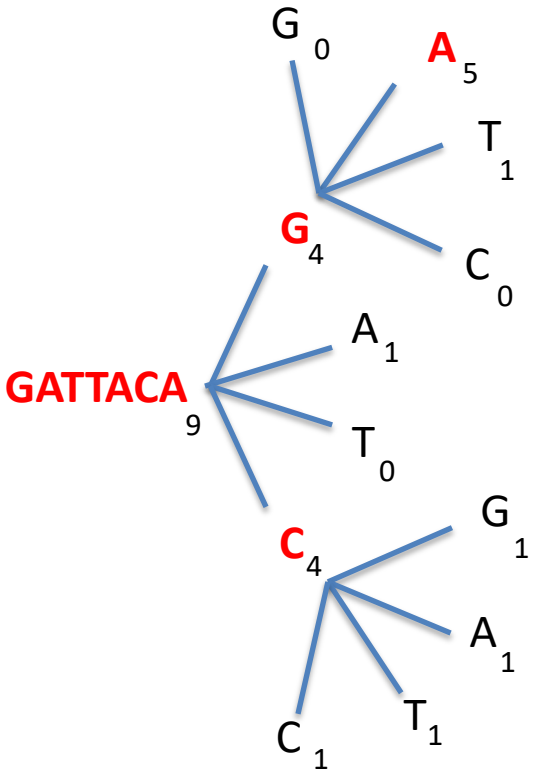
Inchworm Algorithm



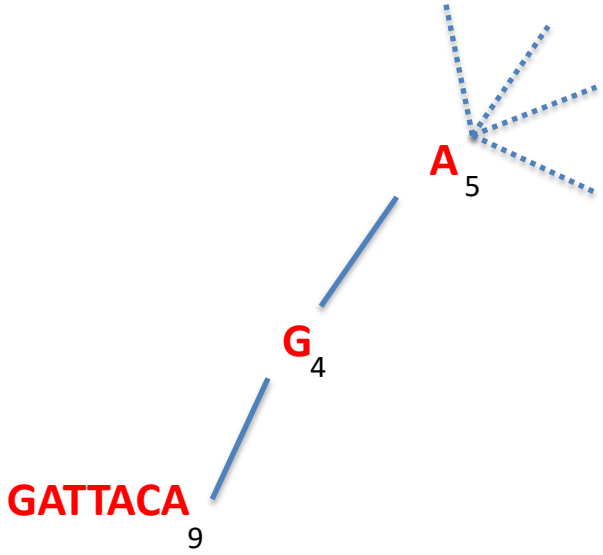
Inchworm Algorithm



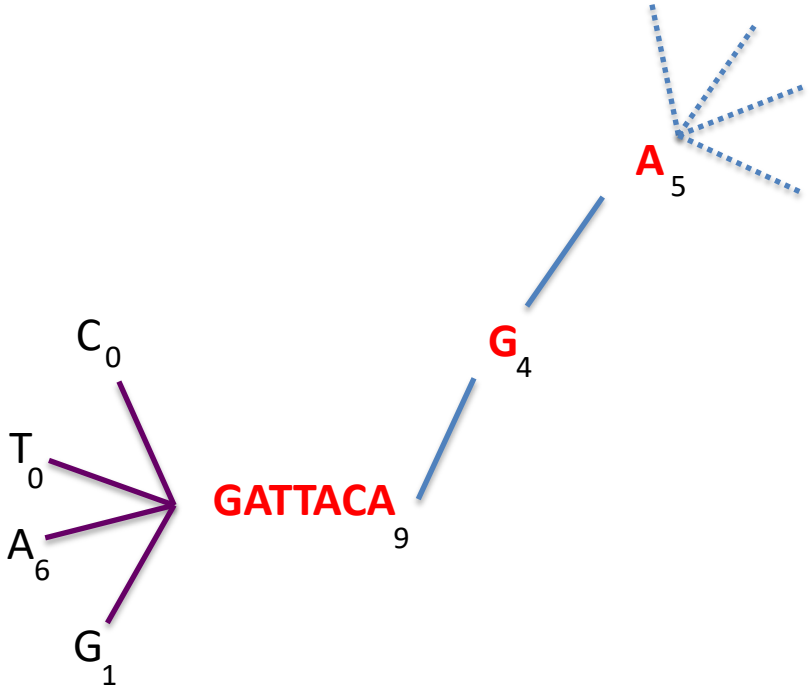
Inchworm Algorithm



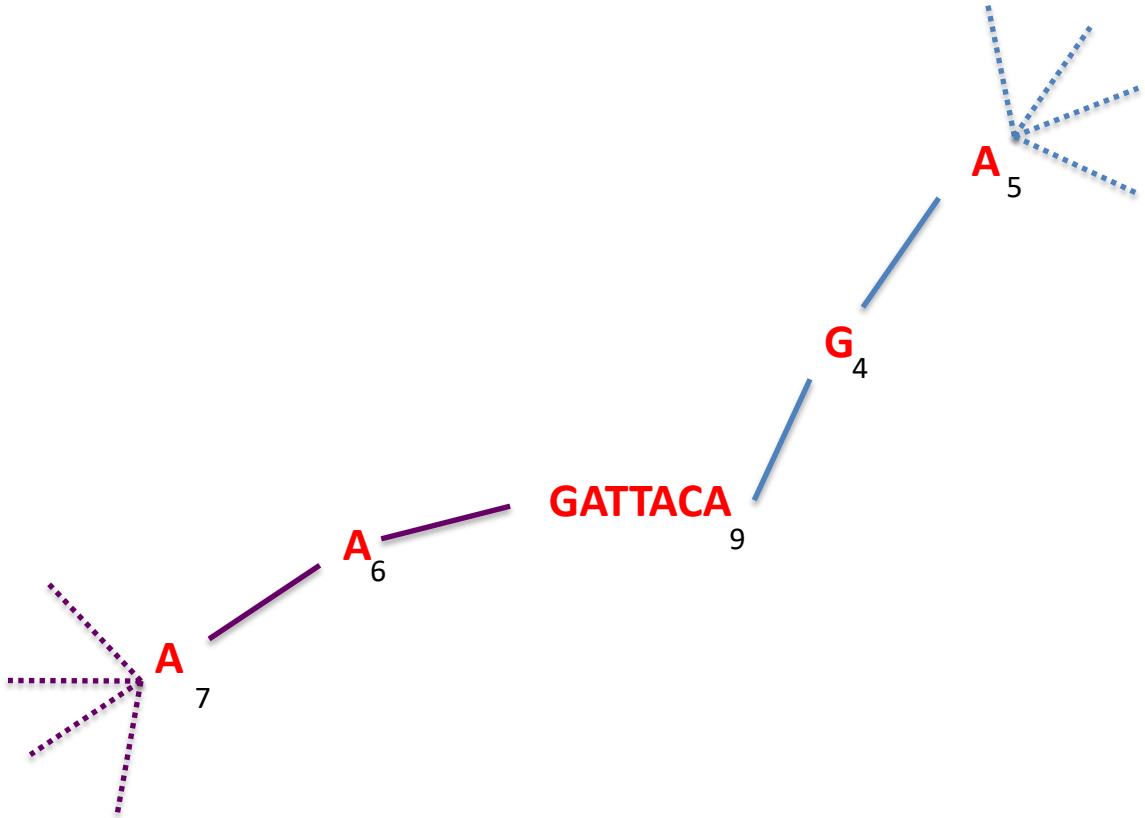
Inchworm Algorithm



Inchworm Algorithm



Inchworm Algorithm



Report contig: ….AAGATTACAGA….

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

Trinity – How it works:

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

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

Thousands of disjoint graphs

Inchworm Contigs from Alt-Spliced Transcripts

Expressed isoforms

Isoform A

Isoform B

Inchworm Contigs from Alt-Spliced Transcripts

Expressed isoforms

Isoform A

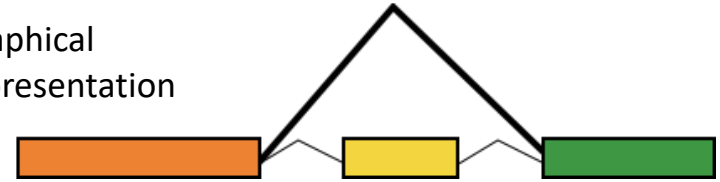
Isoform B

Expression

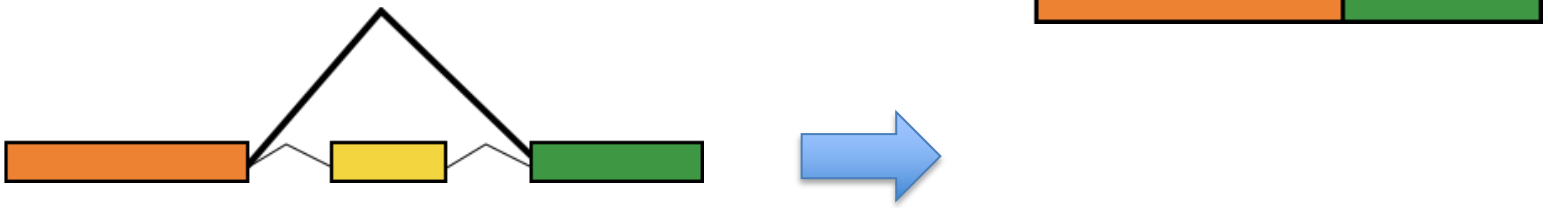
(low)

(high)

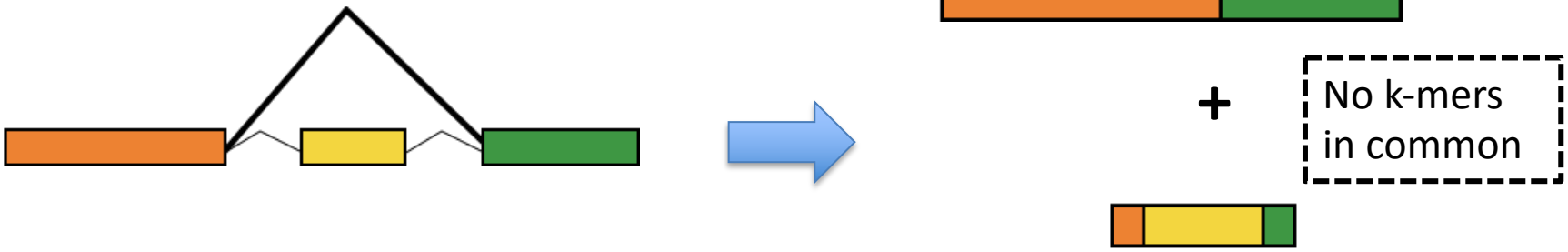
Graphical representation



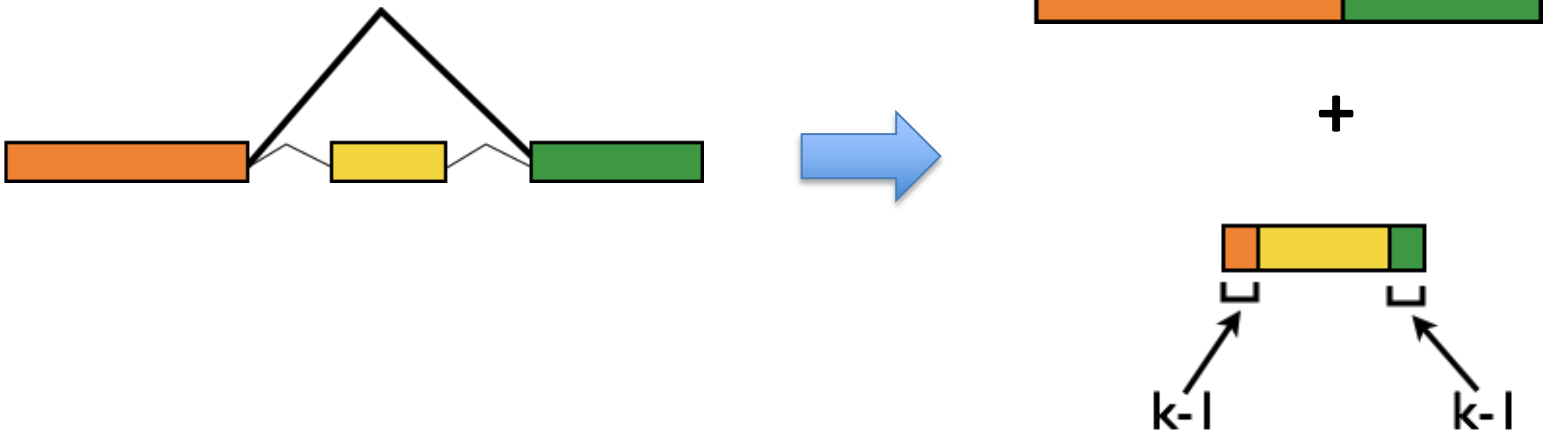
Inchworm Contigs from Alt-Spliced Transcripts



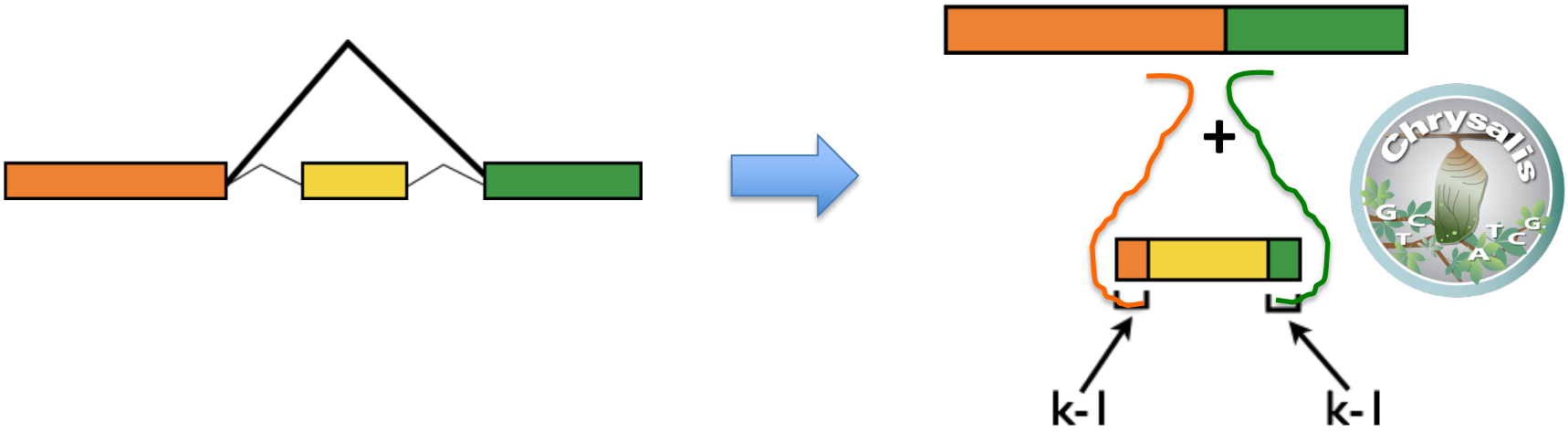
Inchworm Contigs from Alt-Spliced Transcripts



Inchworm Contigs from Alt-Spliced Transcripts

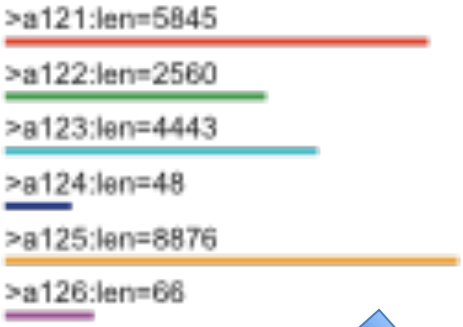


Chrysalis Re-groups Related Inchworm Contigs



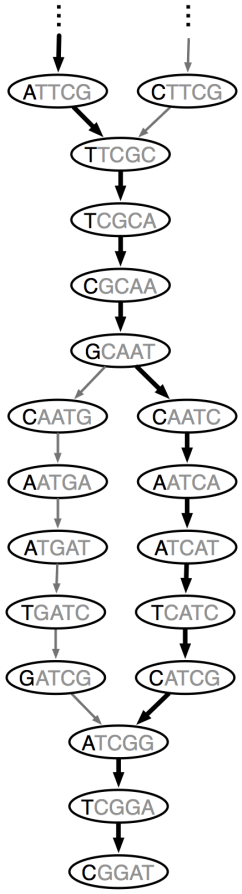
Chrysalis uses (k-1) overlaps and read support to link related Inchworm contigs

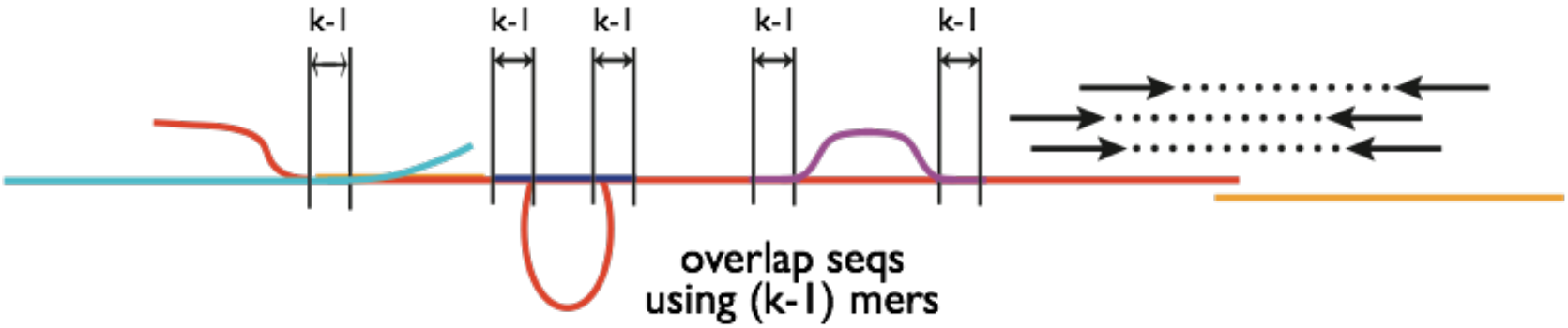
Chrysalis



Integrate isoforms
via k-1 overlaps

Build de Bruijn Graphs
(ideally, one per gene)






Thousands of Chrysalis Clusters

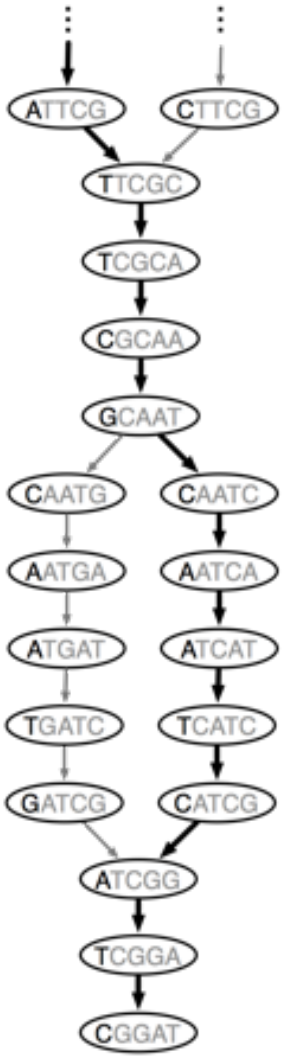
Trinity – How it works:

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

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

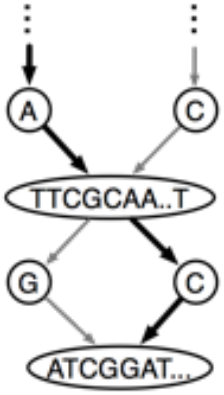
Thousands of disjoint graphs

Butterfly



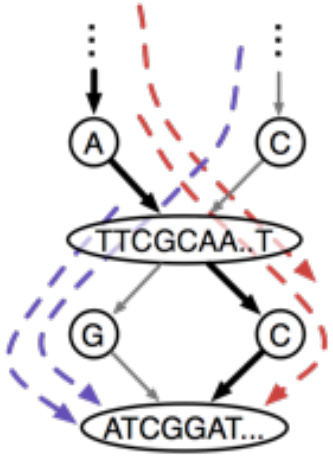
de Bruijn graph

compacting



compact graph

finding paths



compact graph with reads

extracting sequences

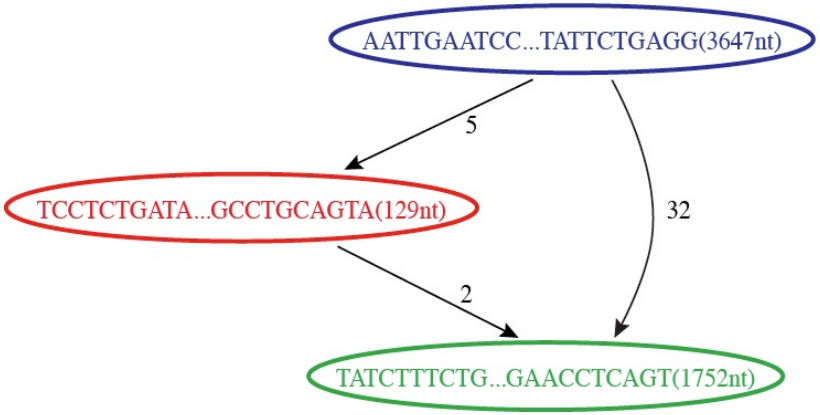
..CTTCGCAA..TGATCGGAT...

..ATTCGCAA..TCATCGGAT...

sequences (isoforms and paralogs)

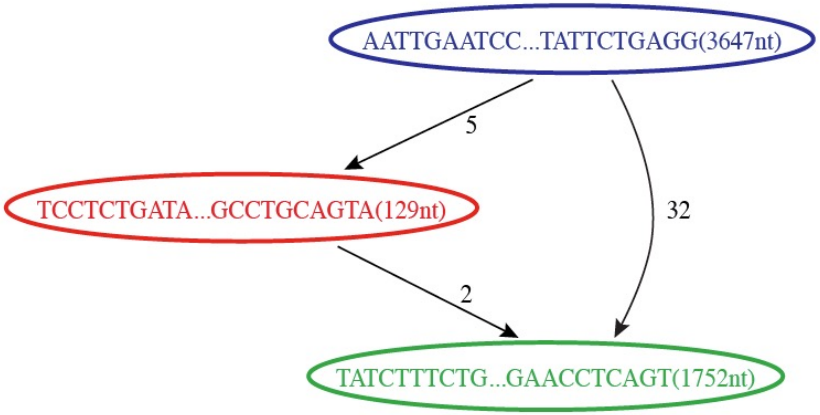
Butterfly Example 1:
Reconstruction of Alternatively Spliced Transcripts

Butterfly's Compacted Sequence Graph



Reconstruction of Alternatively Spliced Transcripts

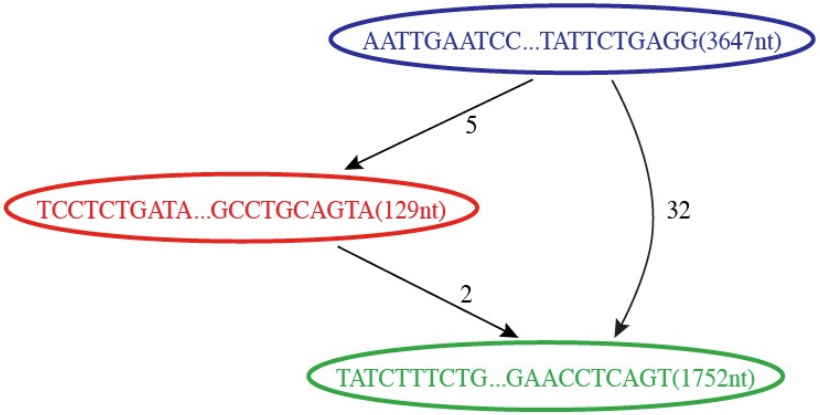
Butterfly's Compacted Sequence Graph



Reconstructed Transcripts

Reconstruction of Alternatively Spliced Transcripts

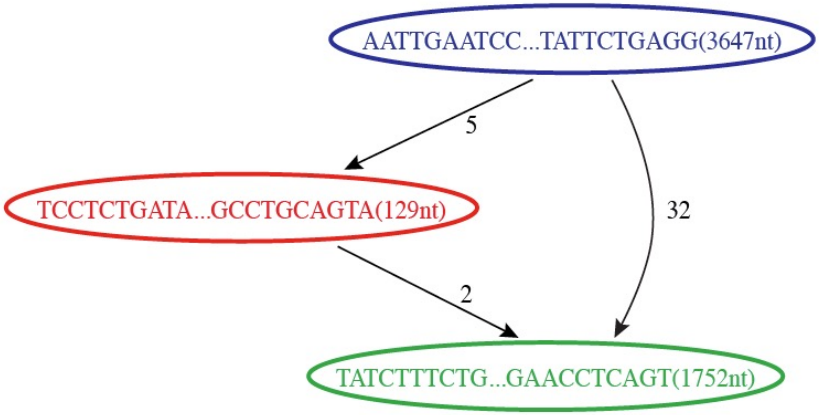
Butterfly's Compacted Sequence Graph



Reconstructed Transcripts

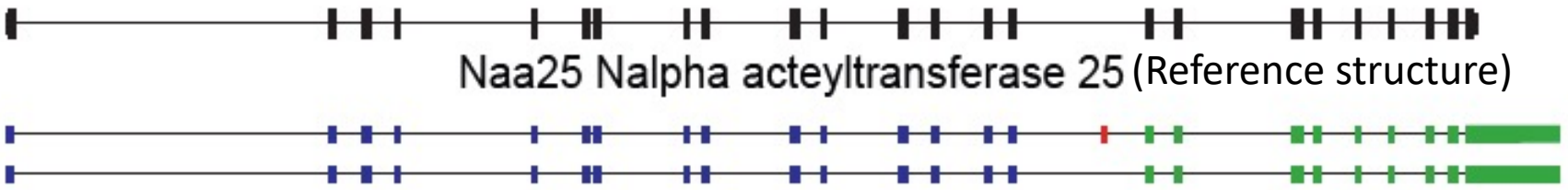
Reconstruction of Alternatively Spliced Transcripts

Butterfly's Compacted Sequence Graph



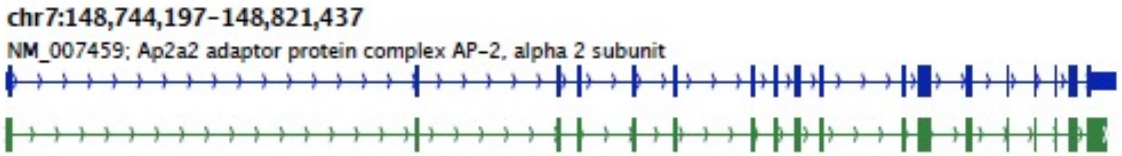
Reconstructed Transcripts

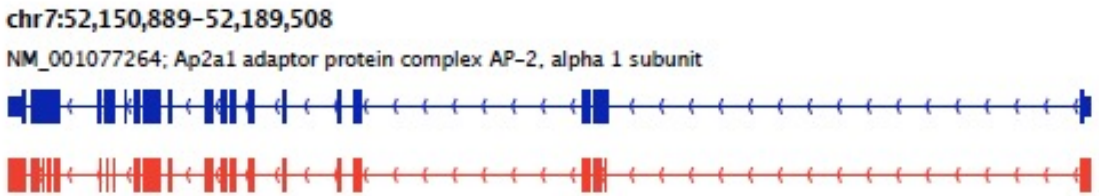
Aligned to Mouse Genome



Butterfly Example 2:
Teasing Apart Transcripts of Paralogous Genes

Teasing Apart Transcripts of Paralogous Genes





Strand-specific RNA-Seq is Preferred

Computationally: fewer confounding graph structures in de novo assembly:

ex. Forward != reverse complement

(GGAA != TTCC)

Biologically: separate sense vs. antisense transcription

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

BROAD INSTITUTE

Comprehensive comparative analysis of strand-specific RNA sequencing methods

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

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

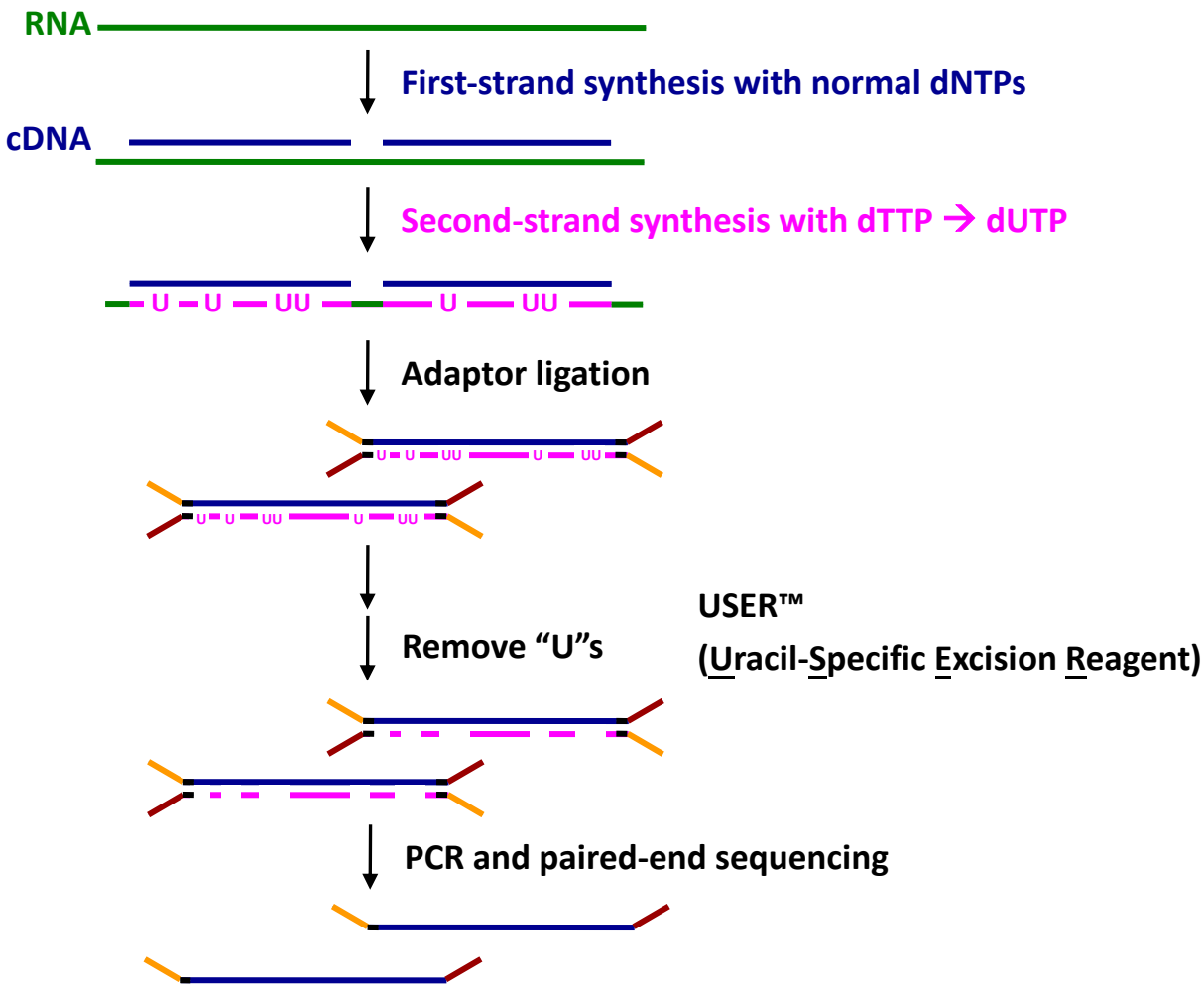
'dUTP second strand marking' identified as the leading protocol

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

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

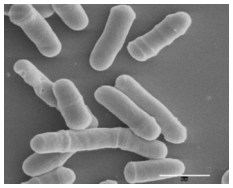
dUTP 2nd Strand Method: Our Favorite



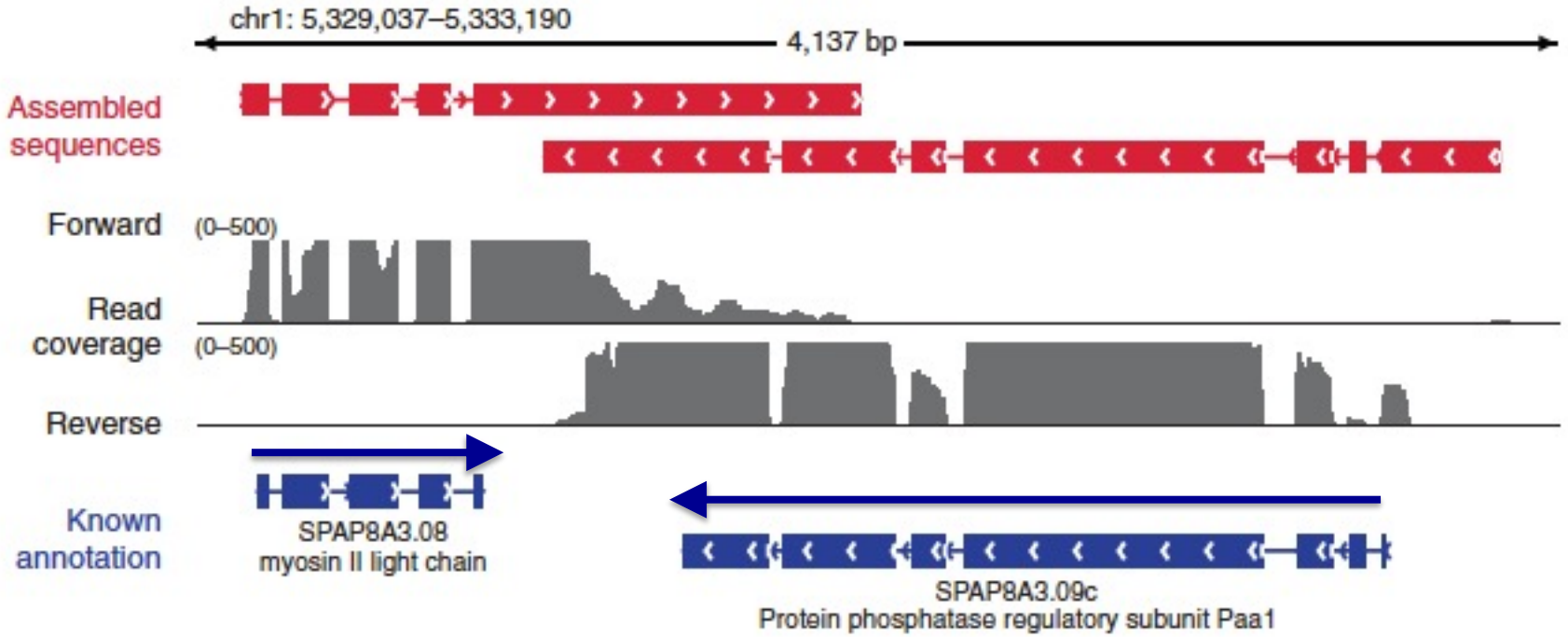
Modified from Parkhomchuk *et al.* (2009) *Nucleic Acids Res.* 37:e123

Slide courtesy of Joshua Levin, Broad Institute.

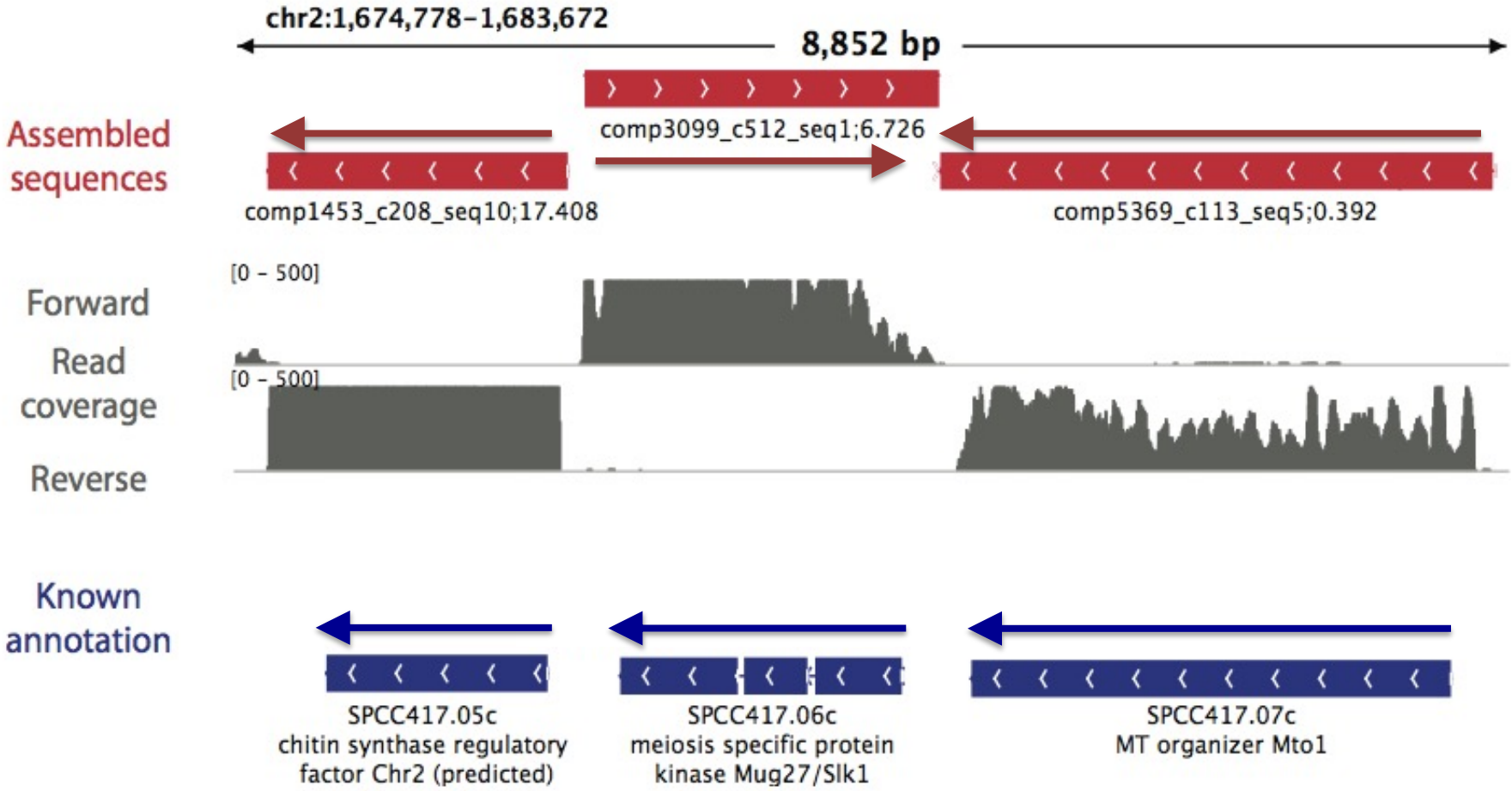
Overlapping UTRs from Opposite Strands

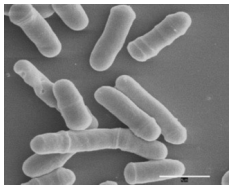


Schizosacharomyces pombe
(fission yeast)



Antisense-dominated Transcription





Trinity is a Highly Effective and Highly Popular RNA-Seq Assembler

Nature Biotechnology, 2011

Thousands of routine users.

>13k literature citations

Freely available, well-supported, open source software

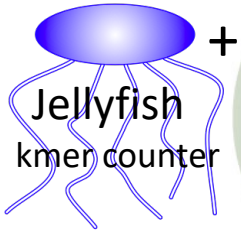
GitHub

http://trinityrnaseq.github.io

Trinity – Today, Many More Components

(off-the-shelf and into the Trinity ecosystem)

Rob Patro



+

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

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

+

BOW TIE

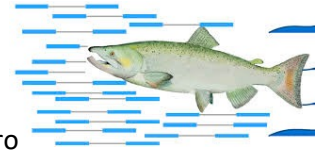
(Capture paired-end links between inchworm contigs)

Ben Langmead

...CTTCGCAA...TGATCGGAT...

...ATTCGCAA...TCATCGGAT...

+



Rob Patro

Salmon expression quantification (eliminate assembly artifacts)

nature protocols

Transcriptome Assembly is Just the End of the Beginning…

NATURE PROTOCOLS | PROTOCOL

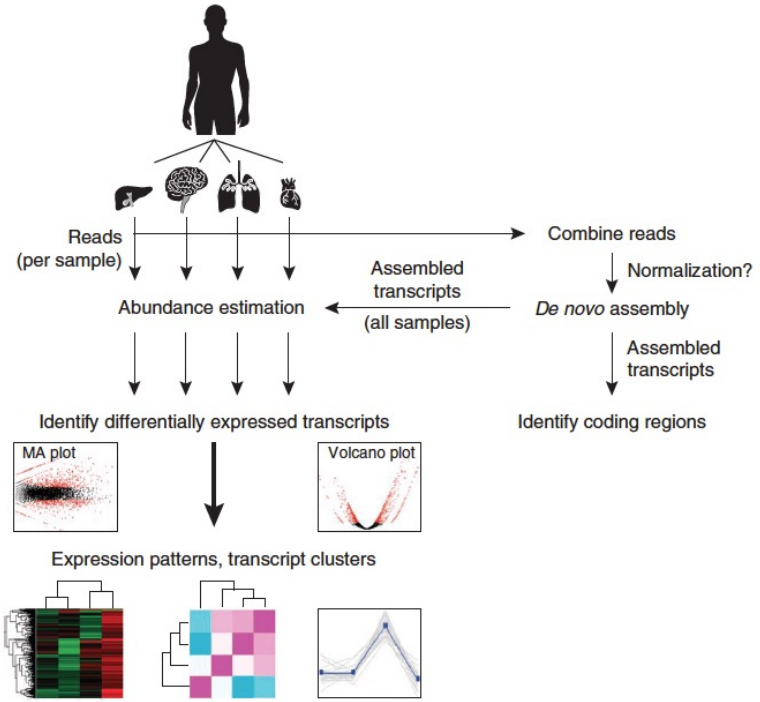
De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis

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

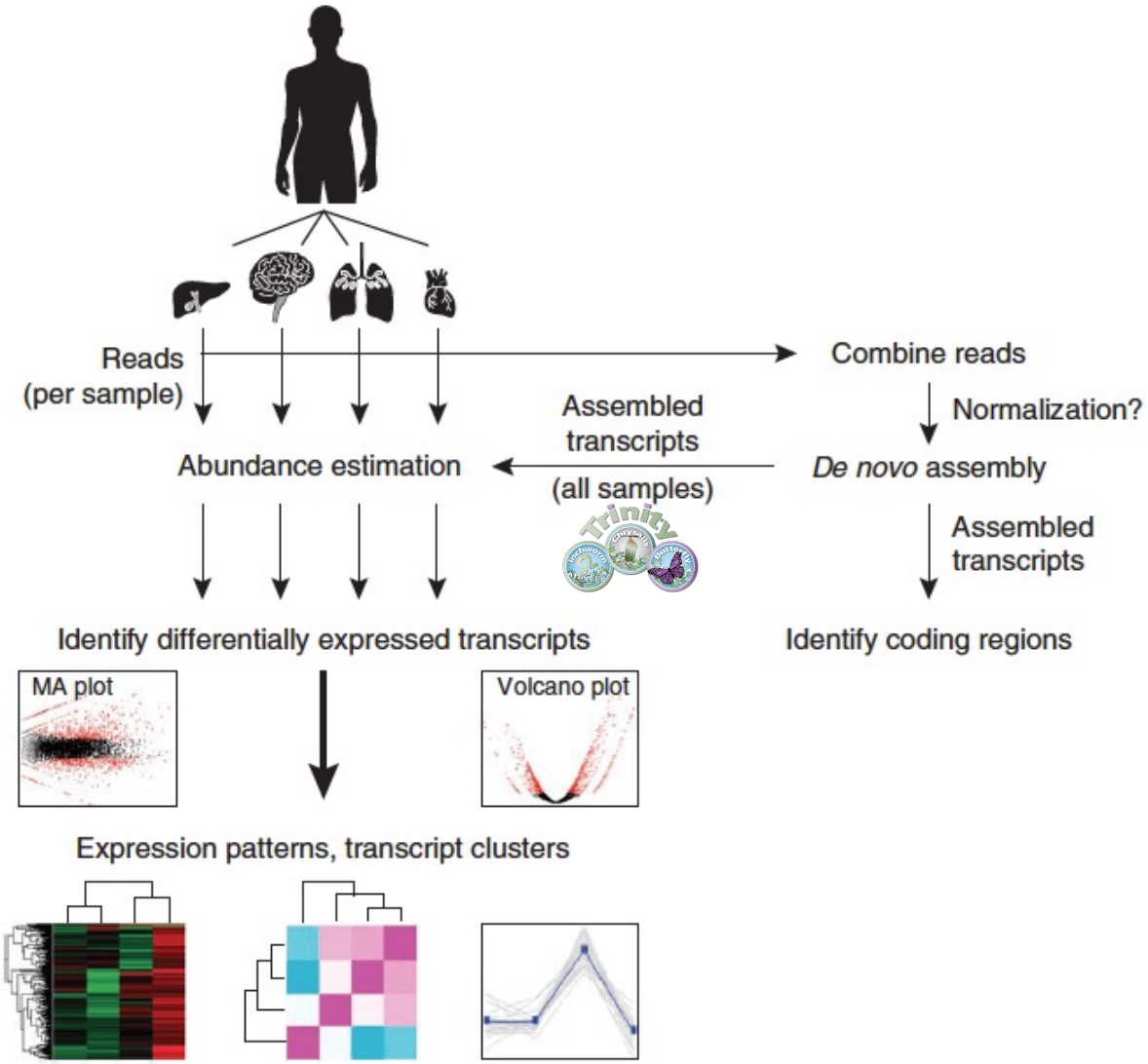
Nature Protocols **8**, 1494–1512 (2013) | doi:10.1038/nprot.2013.084

Published online 11 July 2013



Trinity Framework for De novo Transcriptome Assembly and Analysis

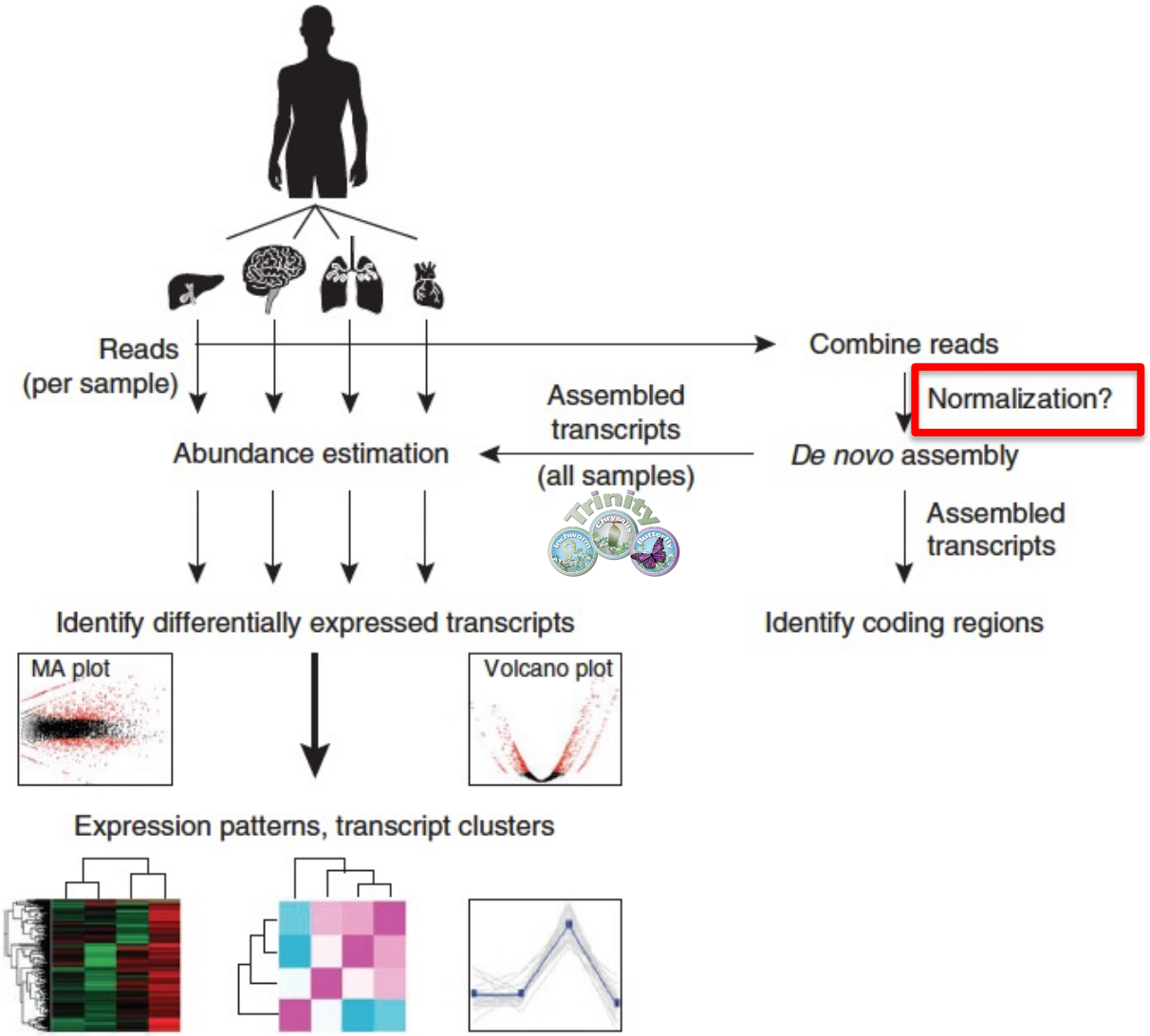
(focus of the transcriptomics lab)



Bioconductor,
& Trinity

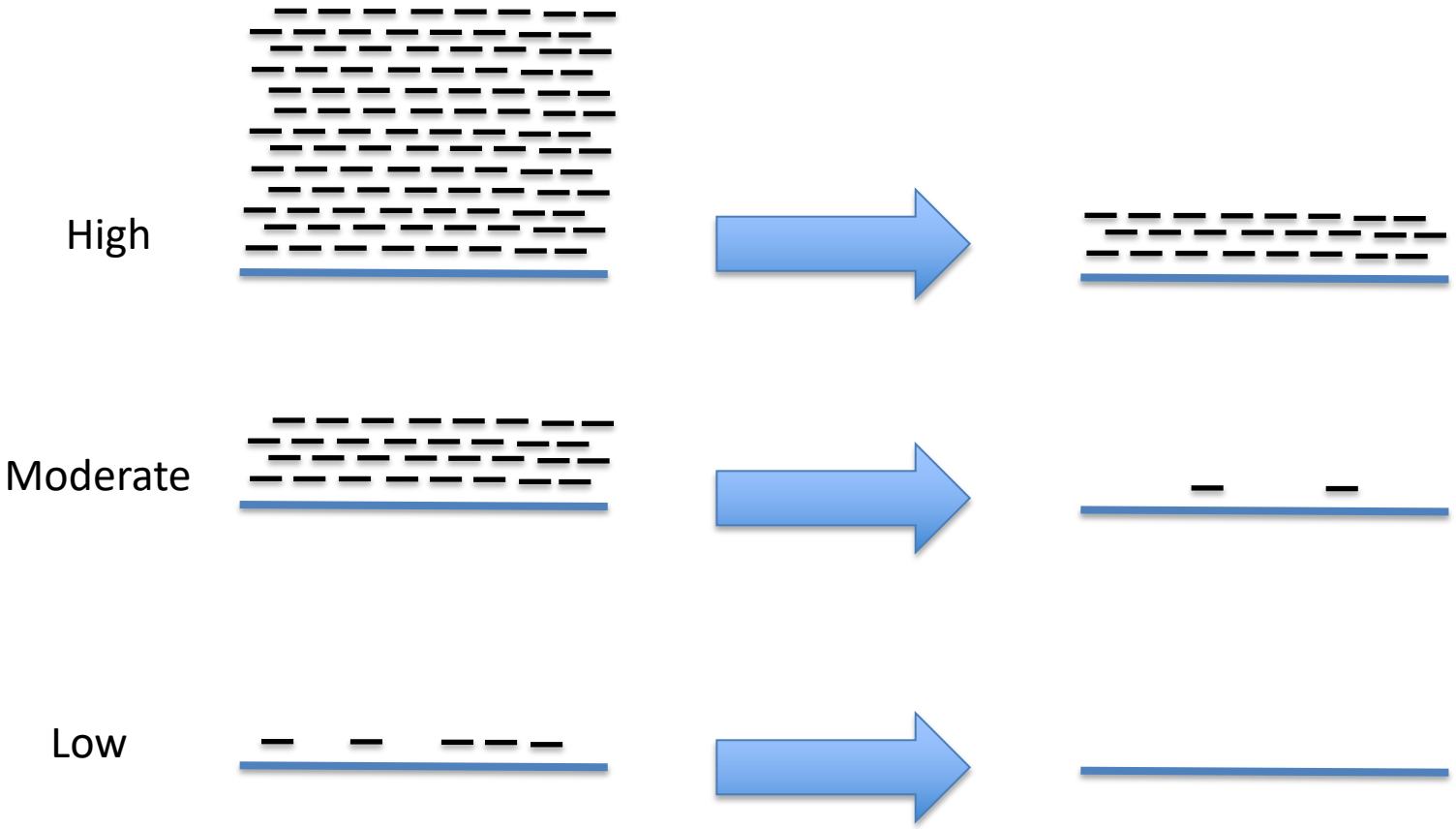
Trinity Framework for De novo Transcriptome Assembly and Analysis

(focus of the transcriptomics lab)

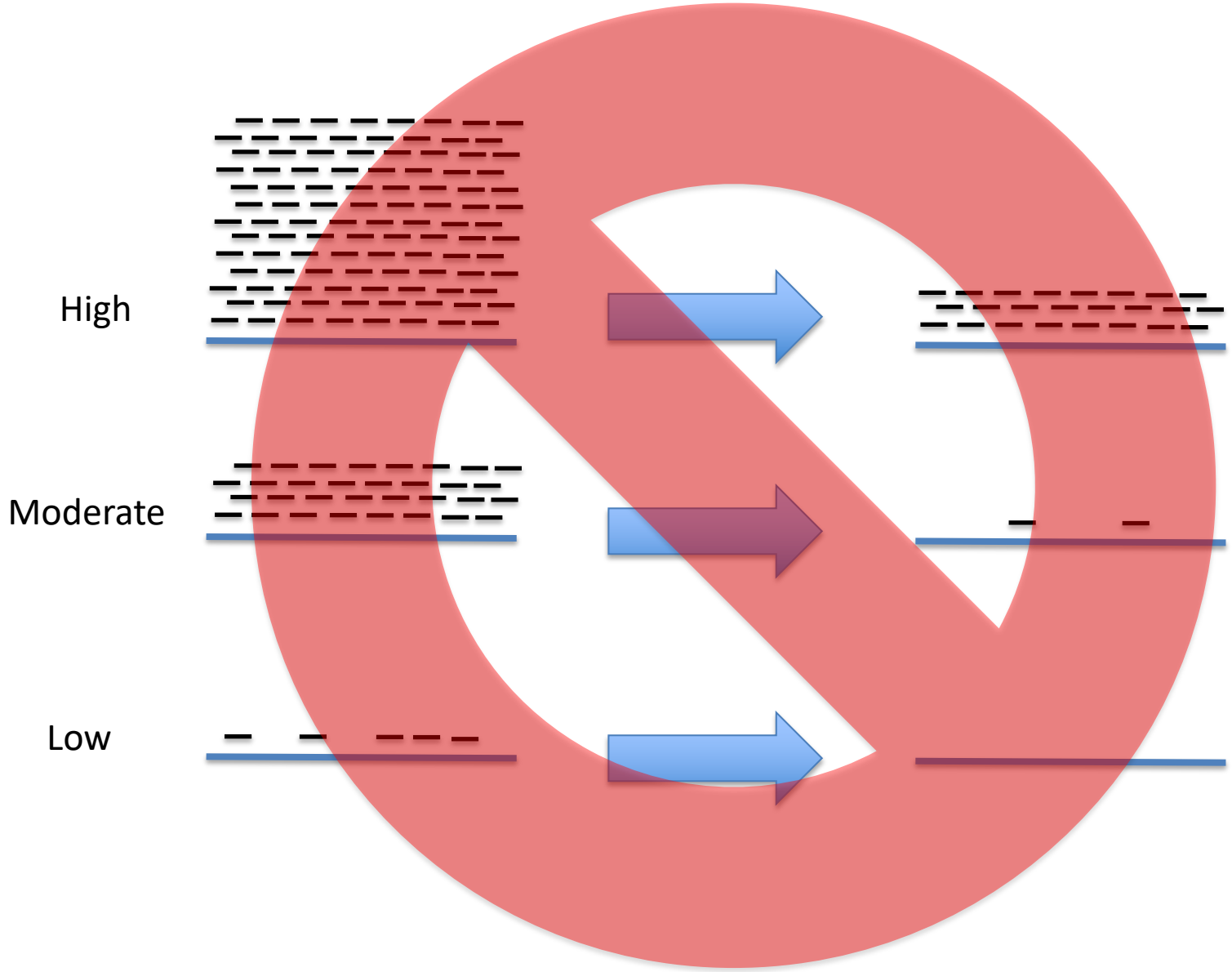


Bioconductor,
& Trinity

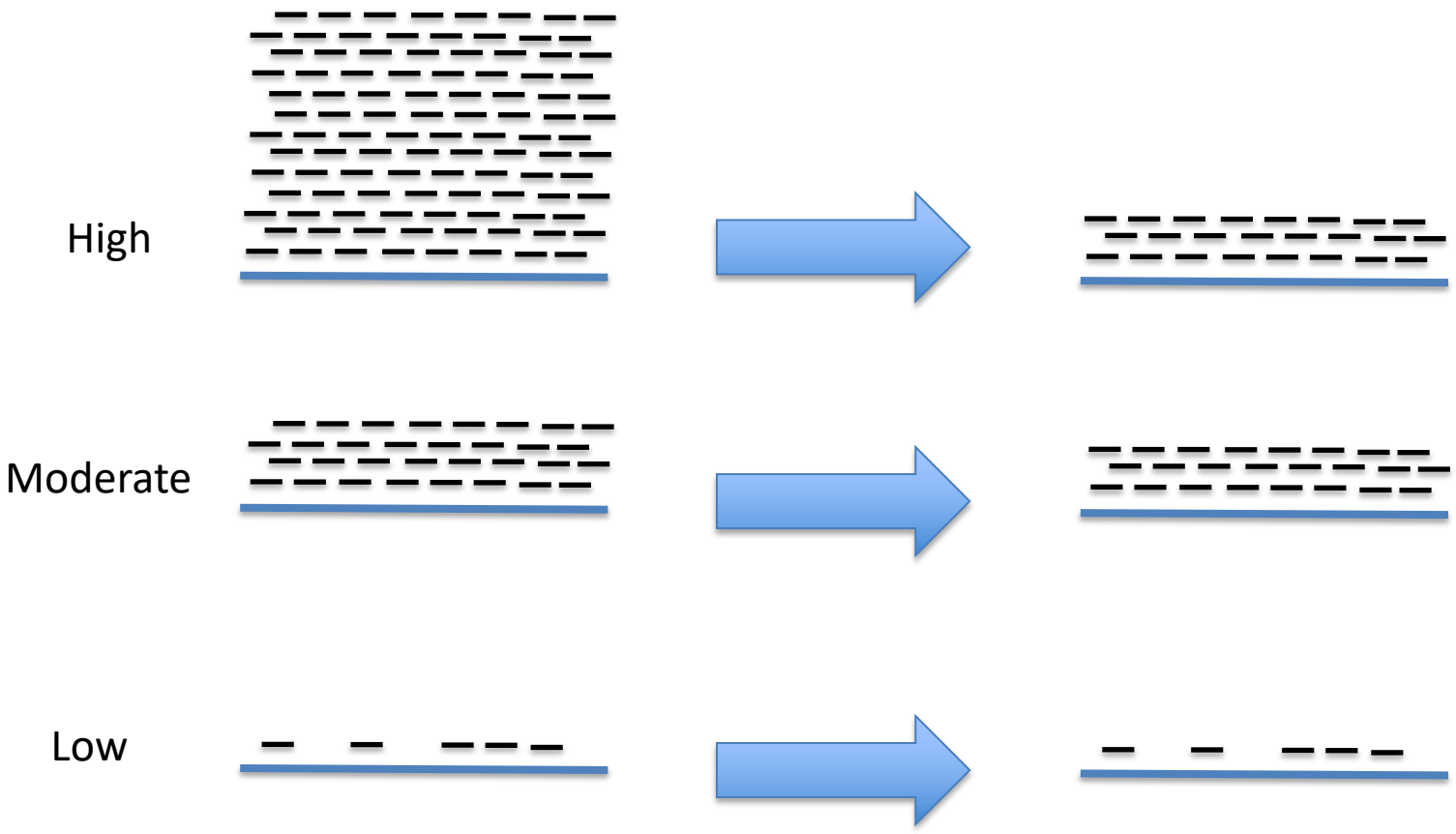
Could sub-sample the reads



Could sub-sample the reads



In silico normalization of reads

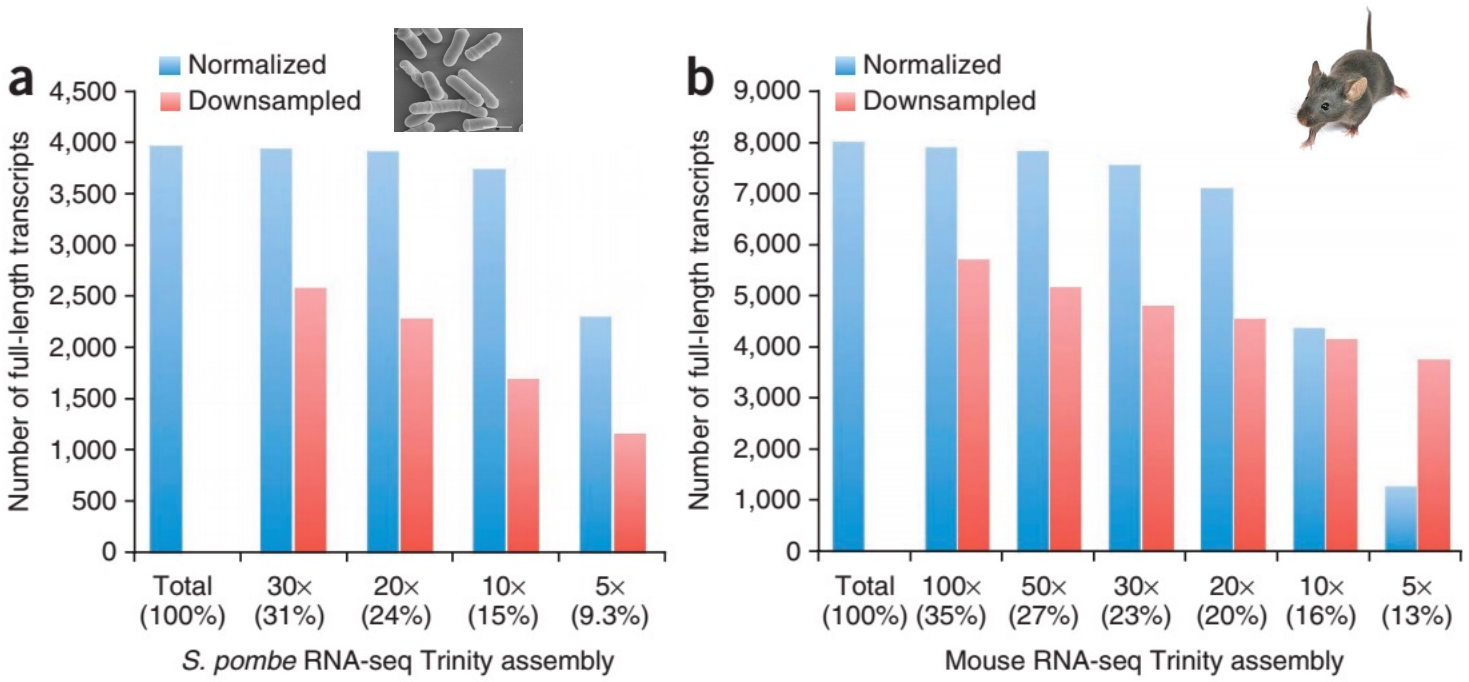


Select reads according to the probability:

$$P(\text{select read}) = \text{Min}\left(\frac{\text{target\_coverage(read)}}{\text{observed\_coverage(read)}}, 1\right)$$

Inspired by C. Titus Brown's Diginorm

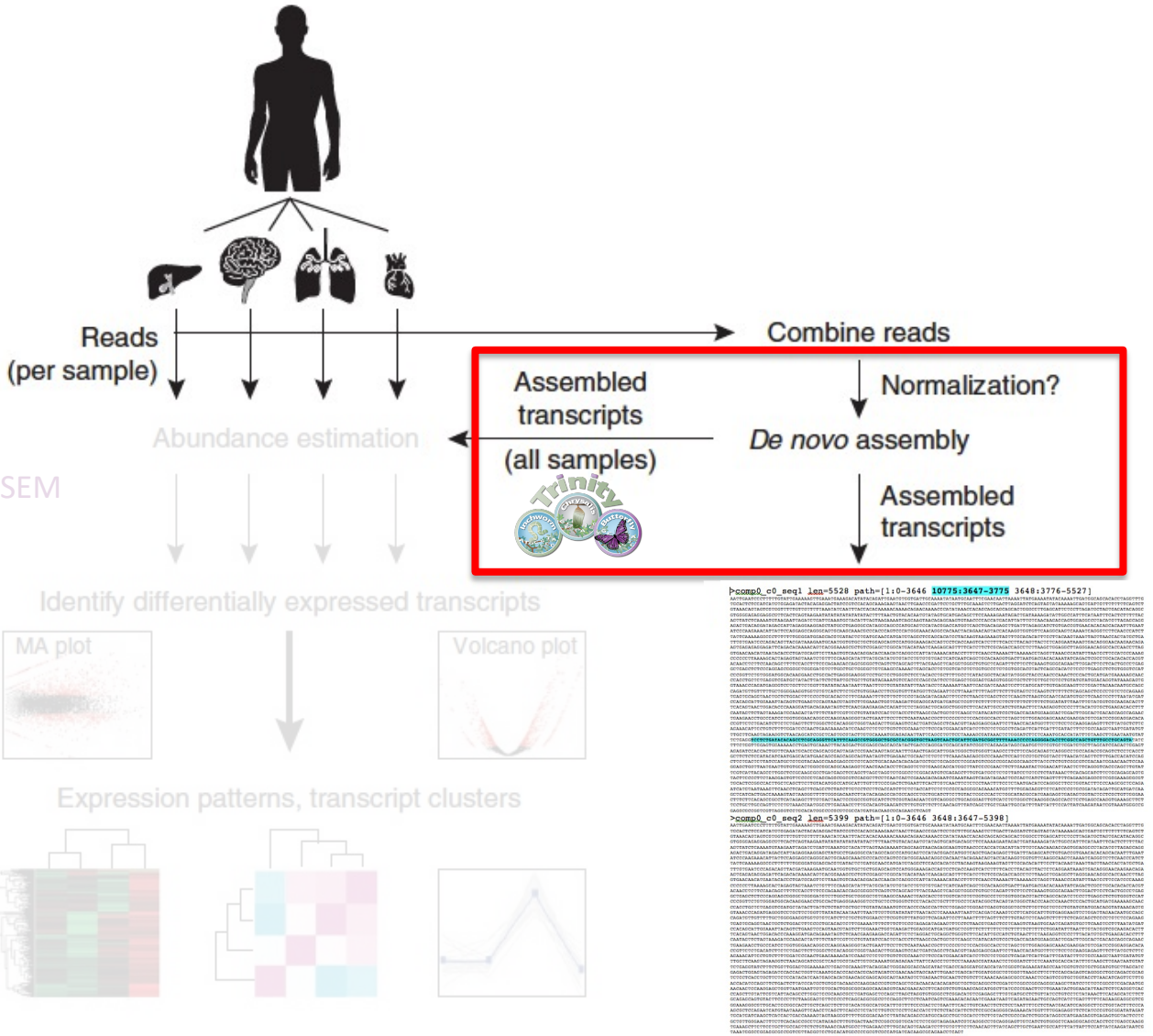
Impact of Normalization on *De novo* Full-length Transcript Reconstruction



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

Haas et al., 2013

The product of Trinity: a Fasta file of assembled transcripts



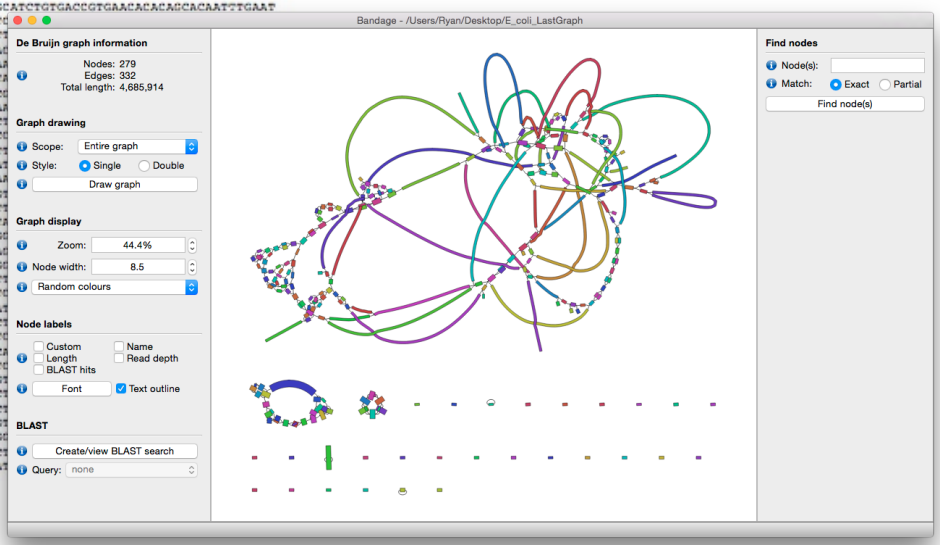
Trinity output: A multi-fasta file

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

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

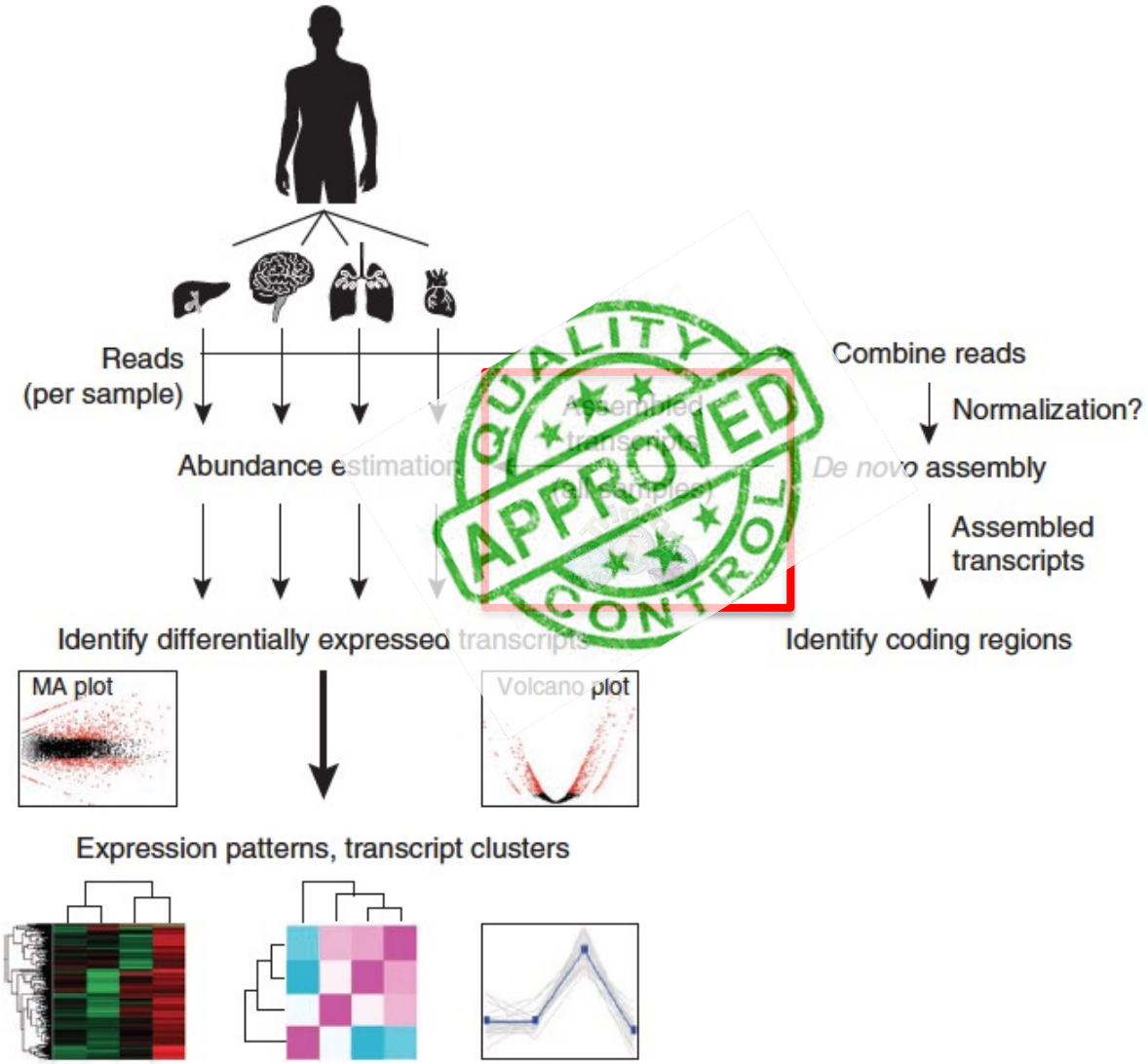
Can visualize using Bandage

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



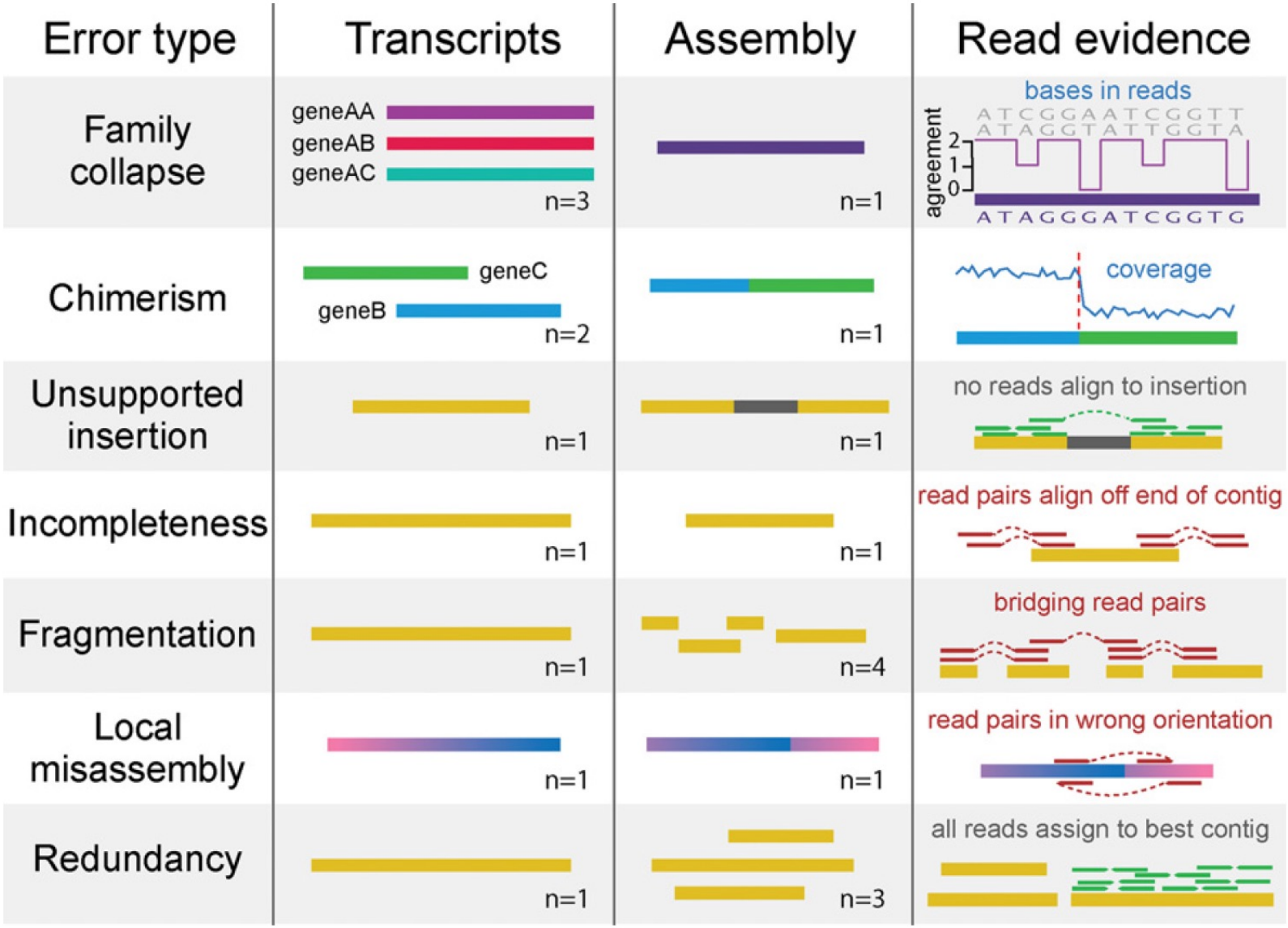
Part 4. Transcriptome Quality Assessment

Evaluating the quality of your transcriptome assembly



Bioconductor,
& Trinity

De novo Transcriptome Assembly is Prone to Certain Types of Errors



Smith-Unna et al. Genome Research, 2016

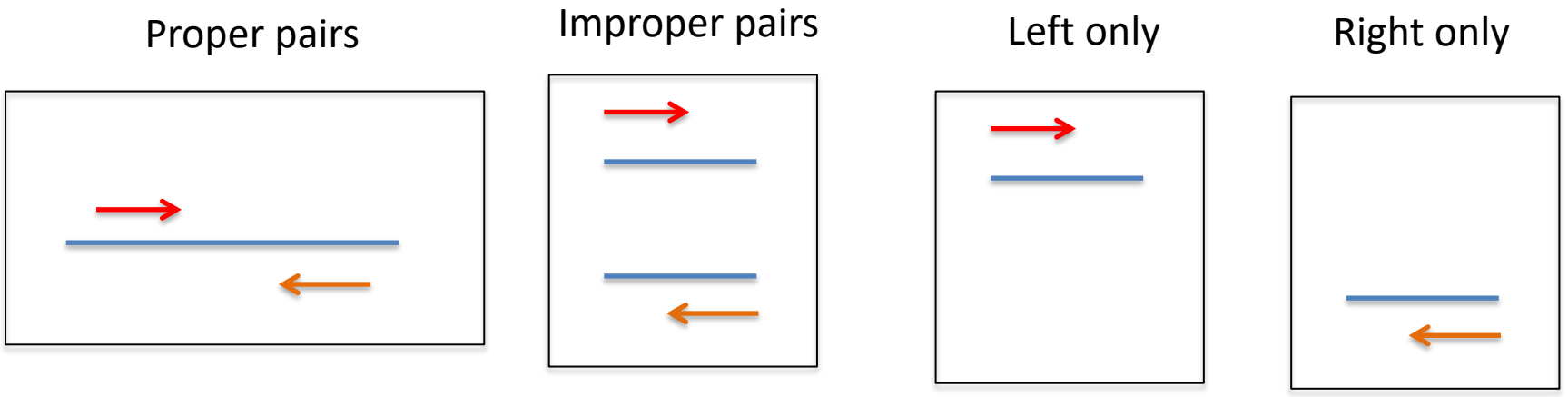
Simple Quantitative and Qualitative Assembly Metrics

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.

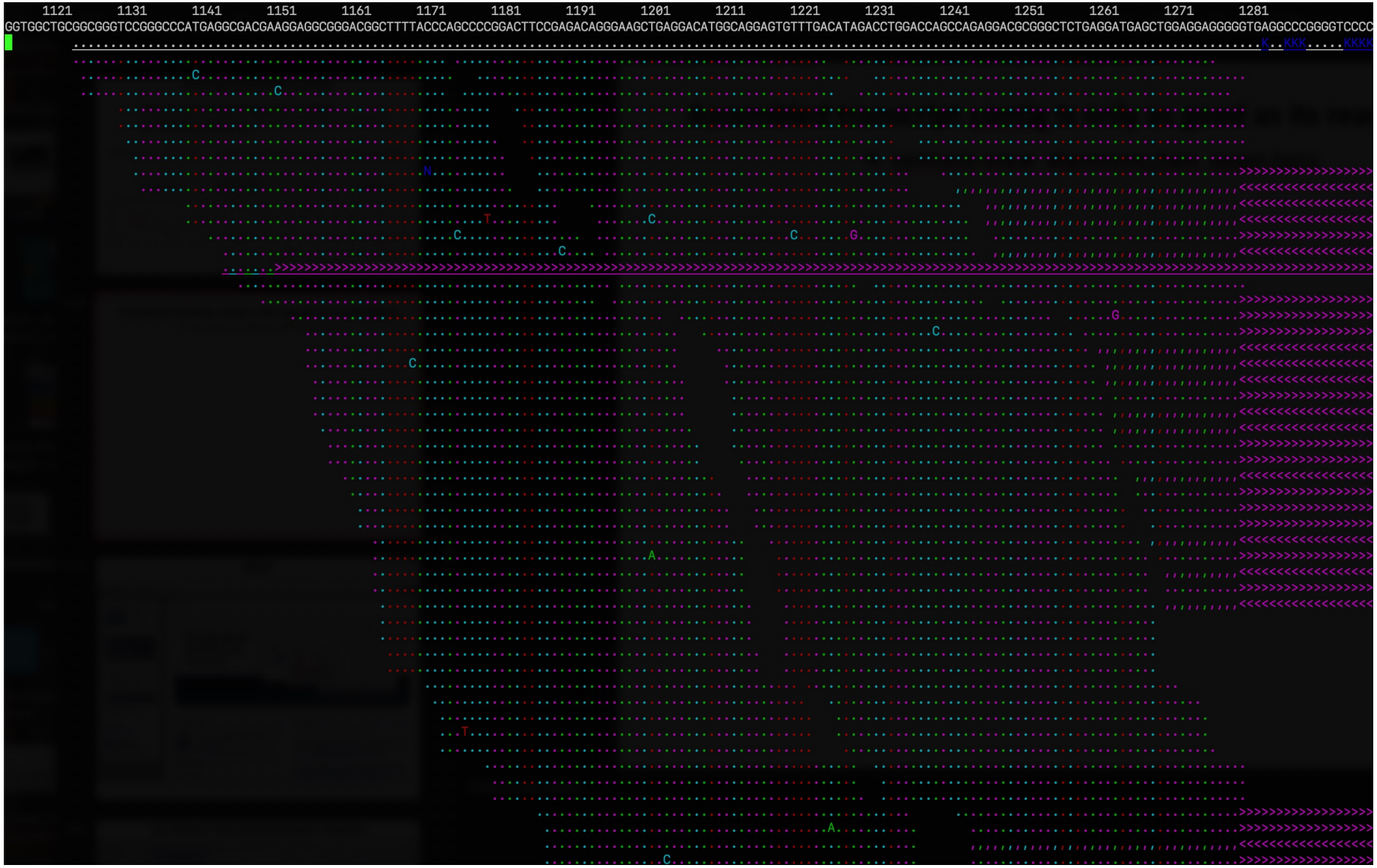
Given read pair: 

Possible mapping contexts in the Trinity assembly are reported:



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

% samtools tview alignments.bam target.fasta



IGV

www.broadinstitute.org/igv/

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

BROAD INSTITUTE

Home

What's New

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

April 20, 2012. IGV 2.1 has been released. See the release notes for more details.

April 19, 2012. See our new IGV paper in Briefings in Bioinformatics.

Overview

Citing IGV

To cite your use of IGV in your publication:

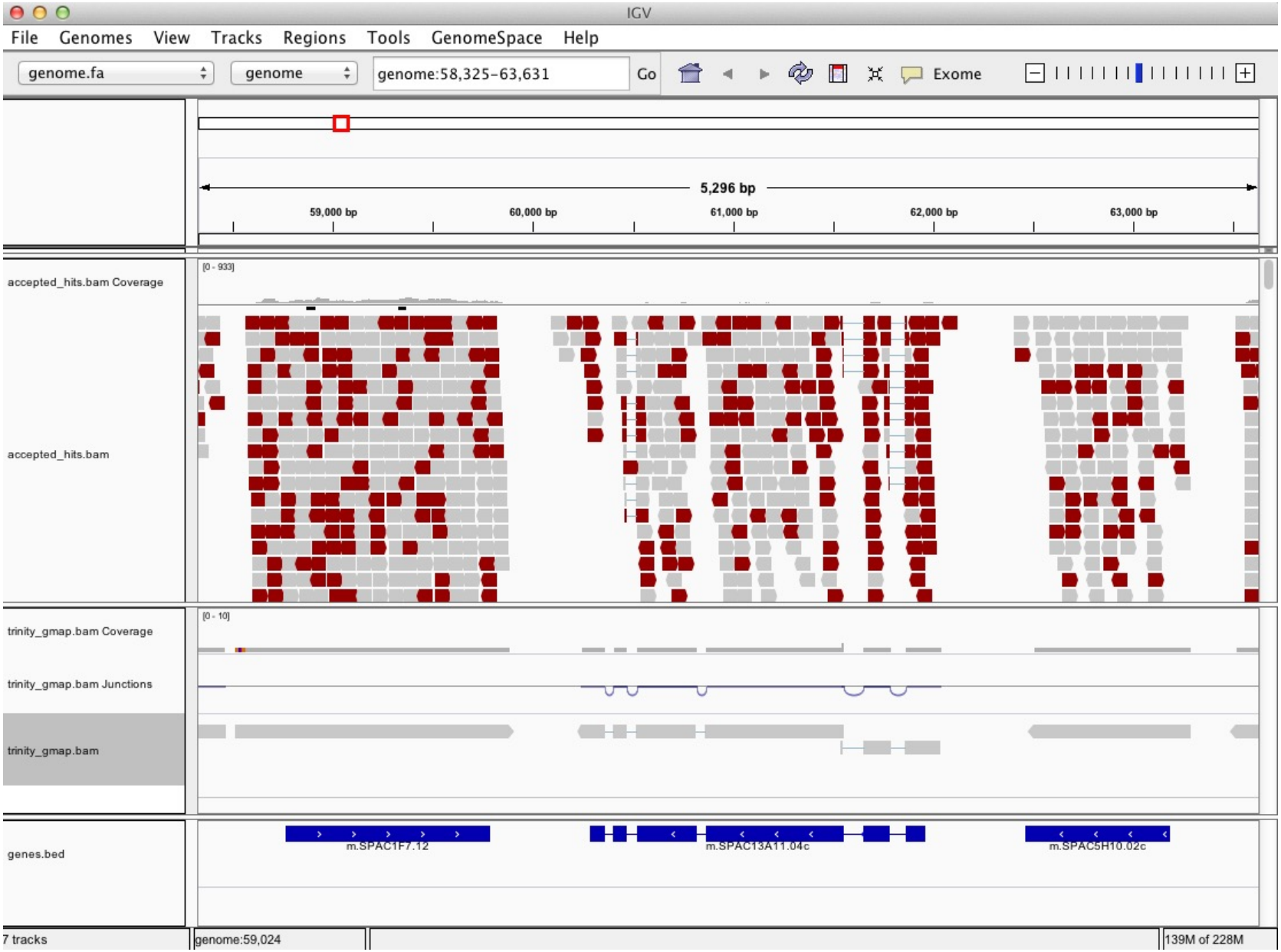
James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, Jill P. Mesirov. **Integrative Genomics Viewer**. Nature Biotechnology 29, 24–26 (2011), or

Helga Thorvaldsdottir, James T. Robinson, Jill P. Mesirov. **Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration.**

Can Examine Transcript Read Support Using IGV

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

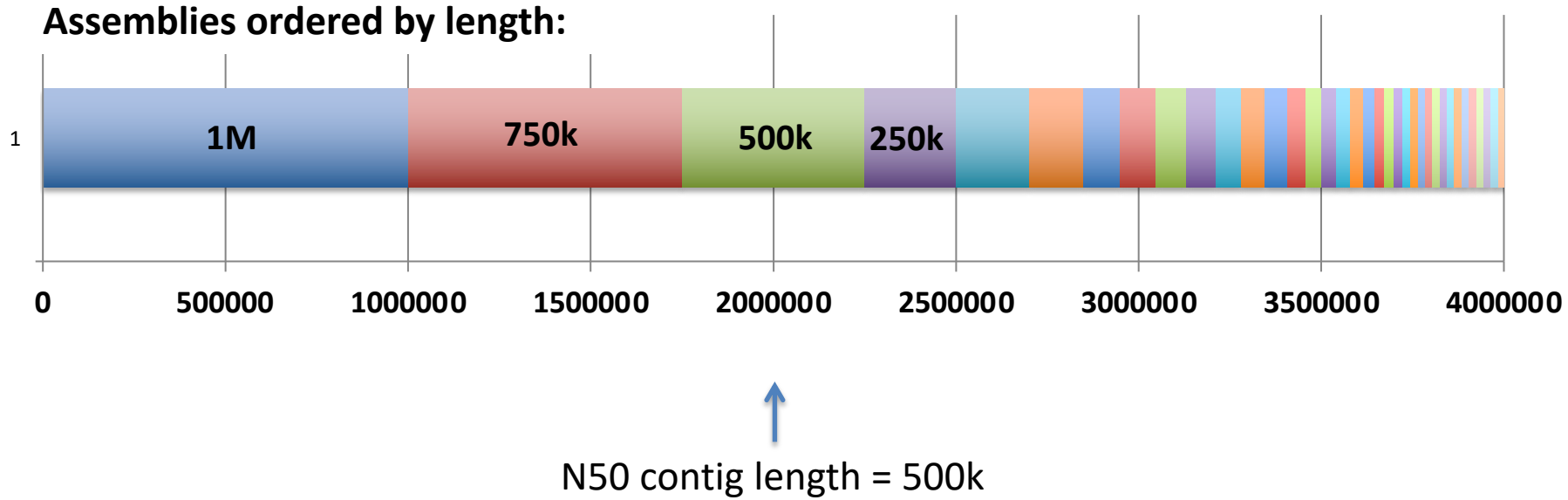
(Trinity transcripts aligned to the genome using GMAP)



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'



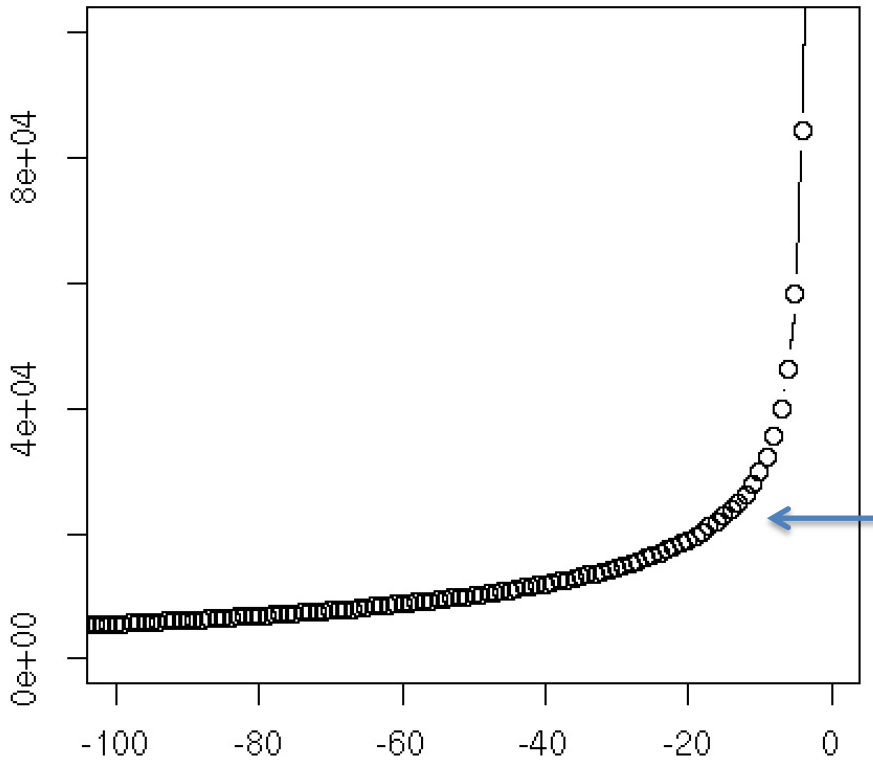
Often, most assembled transcripts are *very* lowly expressed

(How many 'transcripts & genes' are there really?)

1.4 million Trinity transcript contigs

N50 ~ 500 bases

Cumulative # of Transcripts

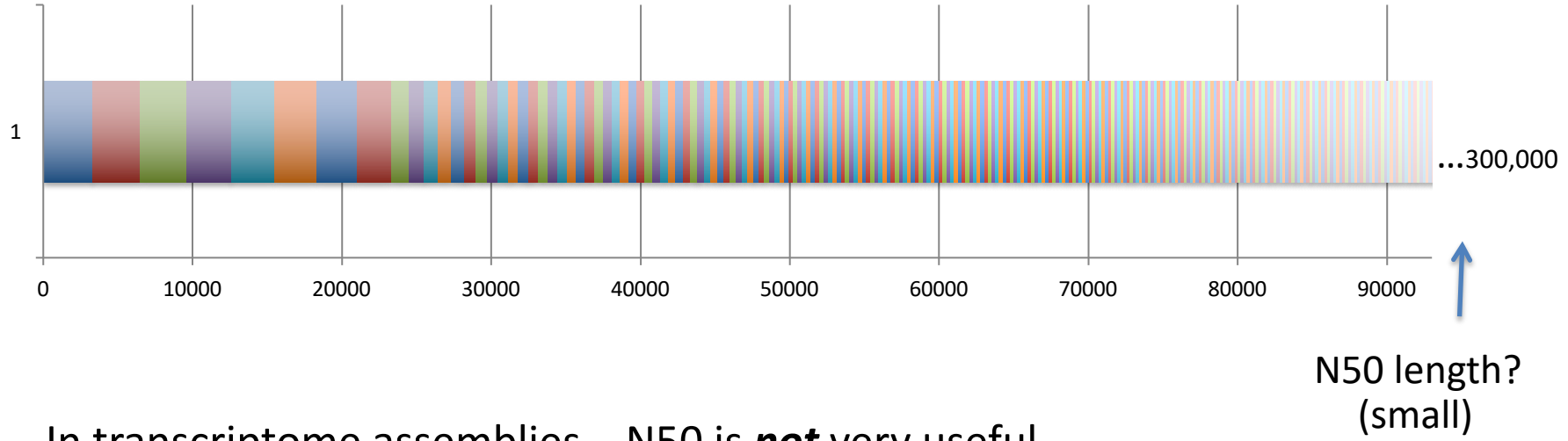


20k transcripts

Expression

* Salamander transcriptome

N50 Calculation for *Transcriptome* Assemblies??



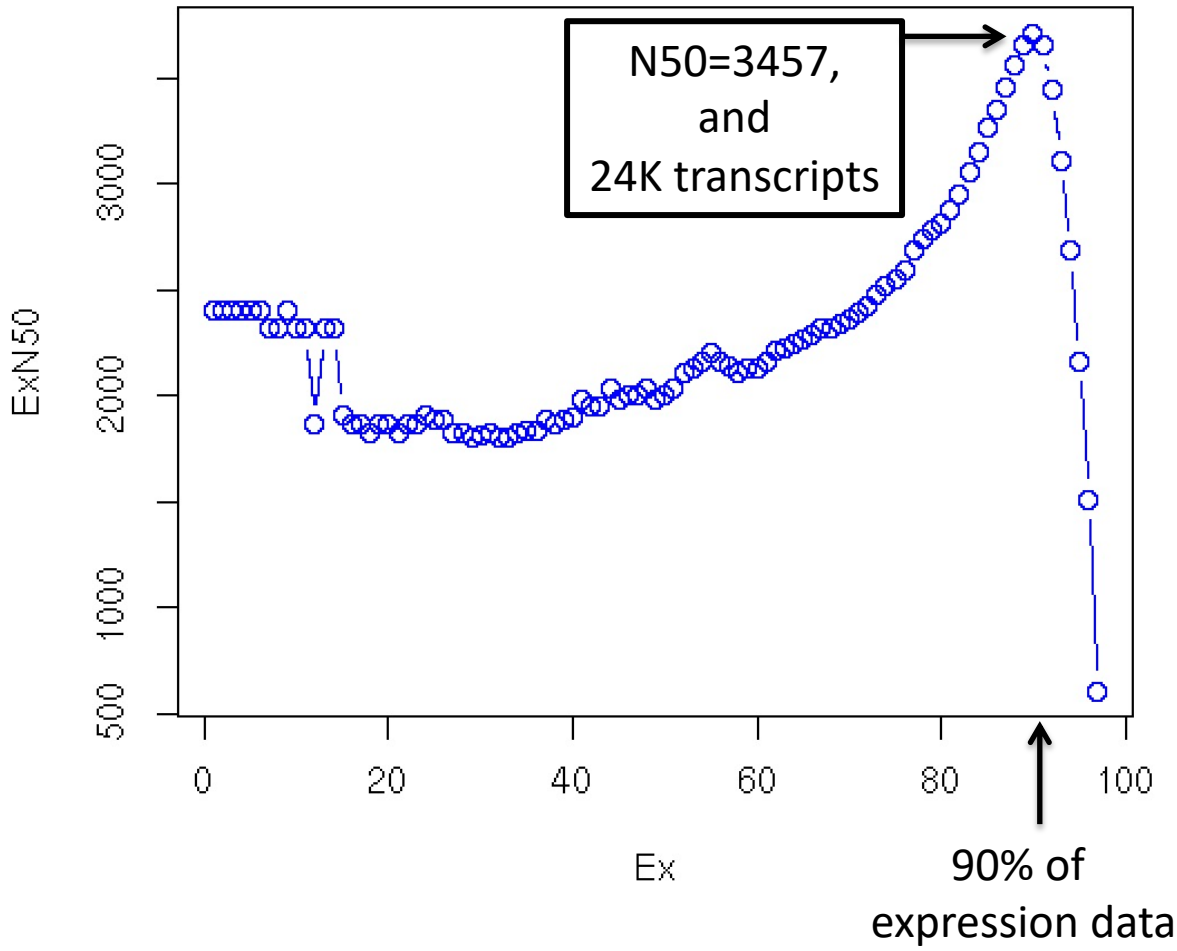
In transcriptome assemblies – N50 is ***not*** very useful.

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

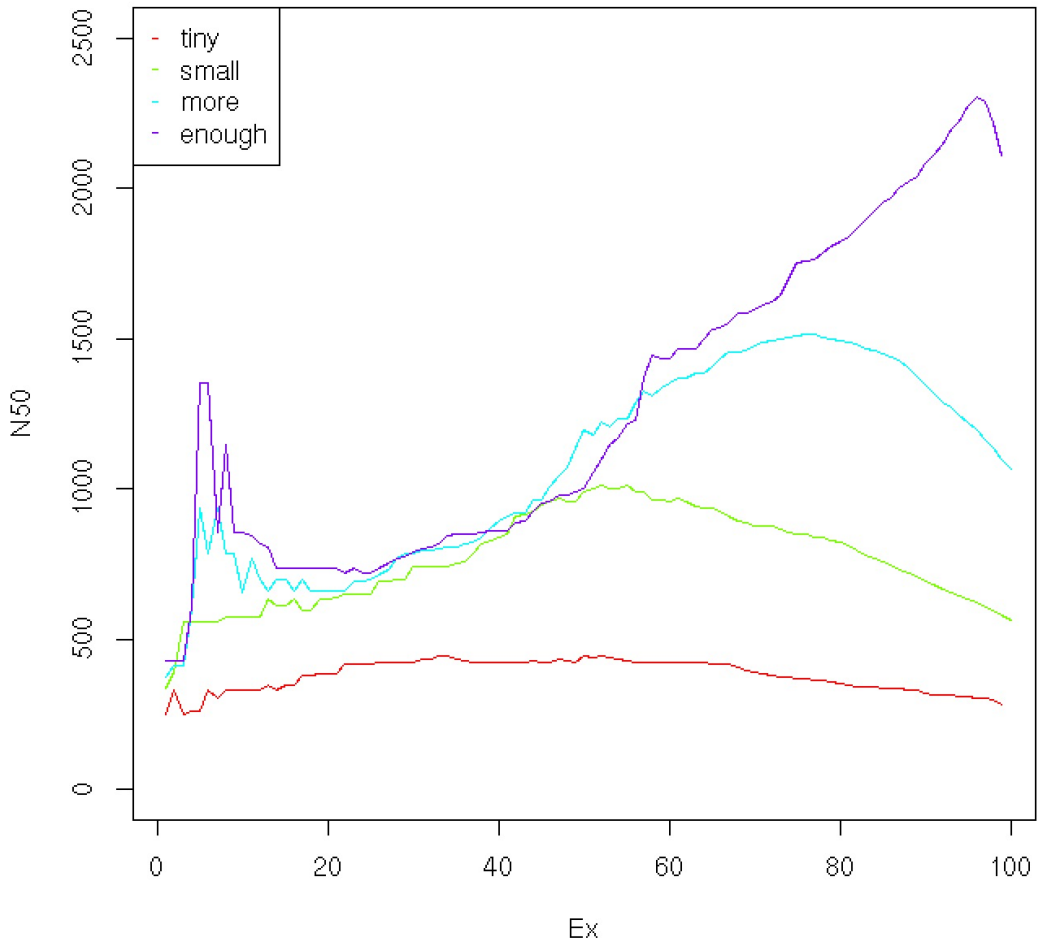
Expression-informed N50 Calculation for Transcriptome Assemblies (ExN50)

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



Millions of Reads

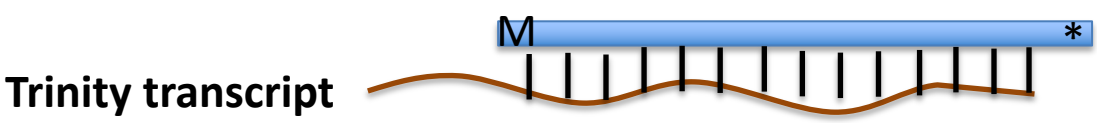
Thousands of Reads

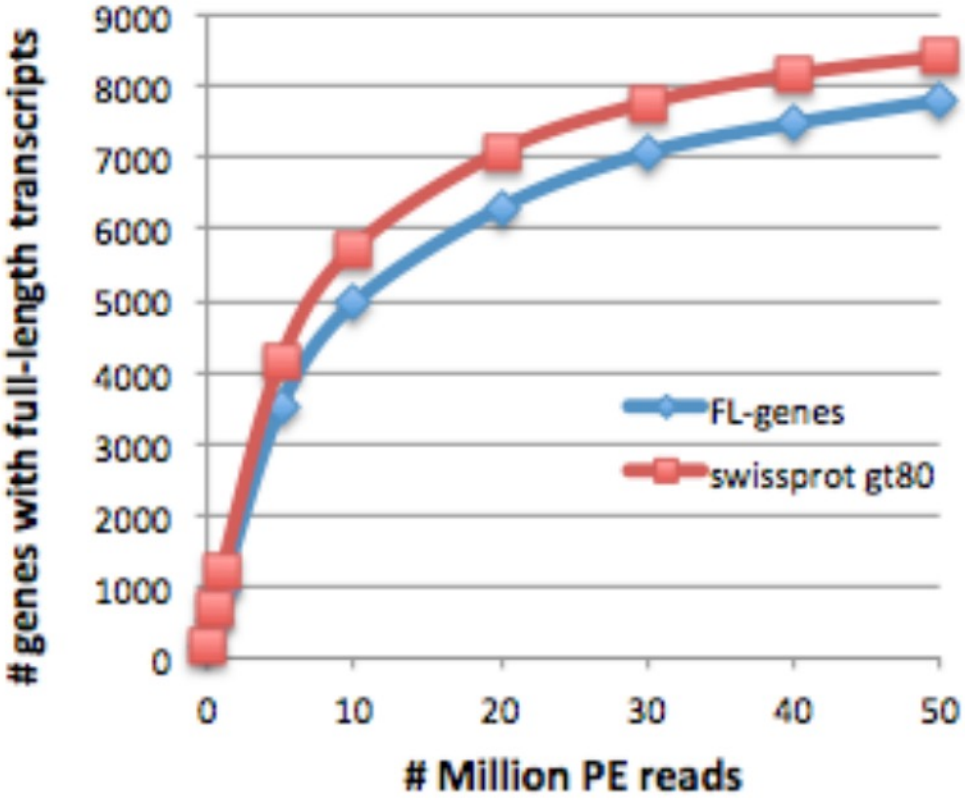
Note shift in ExN50 profiles as you assemble more and more reads.

* Candida transcriptome

Evaluating the quality of your transcriptome assembly

Full-length Transcript Detection via BLASTX





Have you sequenced deeply enough?

* Mouse transcriptome

Haas et al. Nat. Protoc. 2013

busco.ezlab.org

UNIVERSITÉ DE GENÈVE
FACULTÉ DE MÉDECINE

Zdobnov's Computational Evolutionary Genomics group

CEGG Home | OrthoDB *v9* | BUSCO *v2*

Latest is v5.3.2

BUSCO v2

Assessing genome assembly and annotation completeness with **B**enchmarking **U**niversal **S**ingle-**C**opy **O**rthologs

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.

busco.ezlab.org

UNIVERSITÉ DE GENÈVE
FACULTÉ DE MÉDECINE

Zdobnov's Computational Evolutionary Genomics group

CEGG Home | OrthoDB *v9* | BUSCO *v2*

Latest is v5.3.2

BUSCO v2

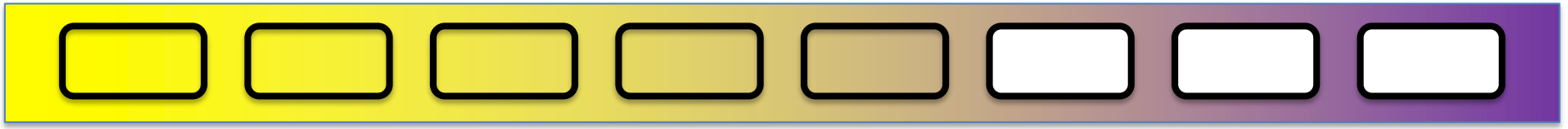
Assessing genome assembly and annotation completeness with Benchmarking Universal Single-Copy Orthologs

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

Summarized benchmarks in BUSCO notation:
C:88%[D:53%],F:4.5%,M:7.3%,n:3023

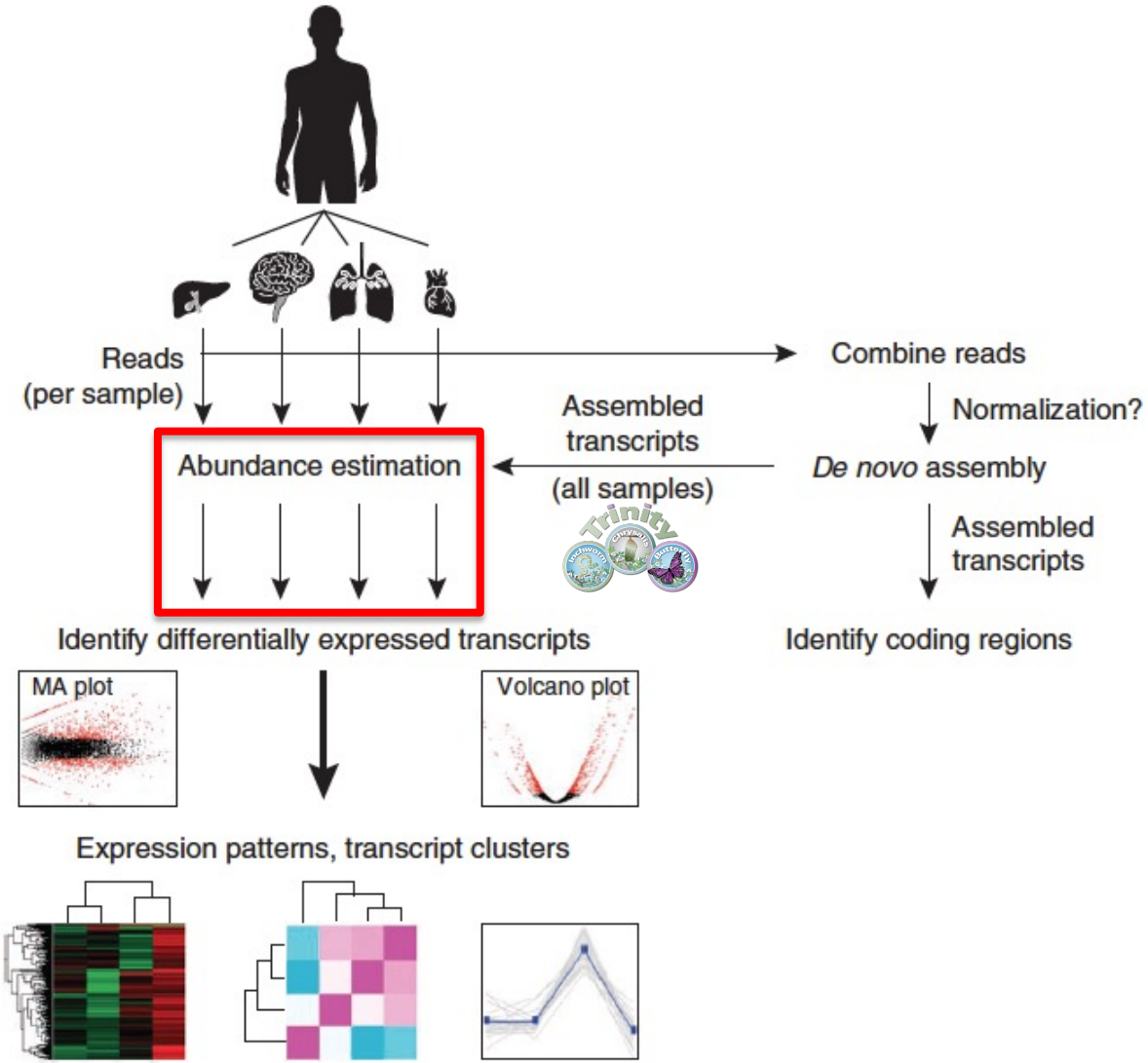
Representing:
1045 Complete Single-copy BUSCOs
1617 Complete Duplicated BUSCOs
139 Fragmented BUSCOs
222 Missing BUSCOs
3023 Total BUSCO groups searched

Part 5. Expression Quantification



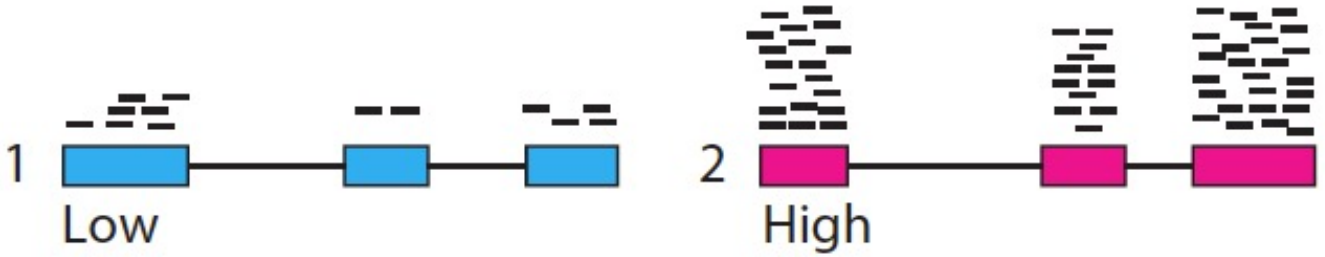
Abundance Estimation

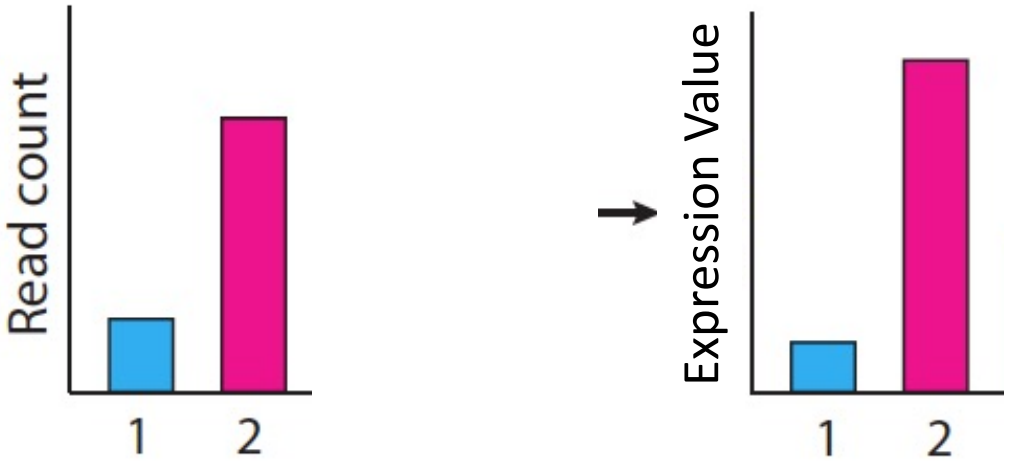
(Aka. Computing Expression Values)



Bioconductor,
& Trinity

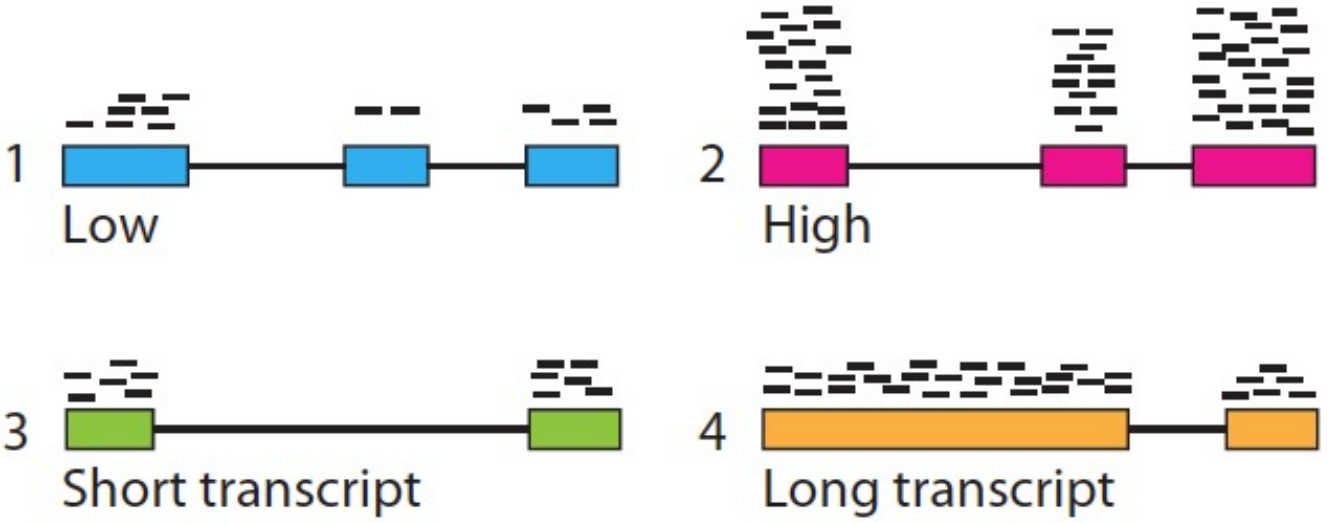
Calculating expression of genes and transcripts

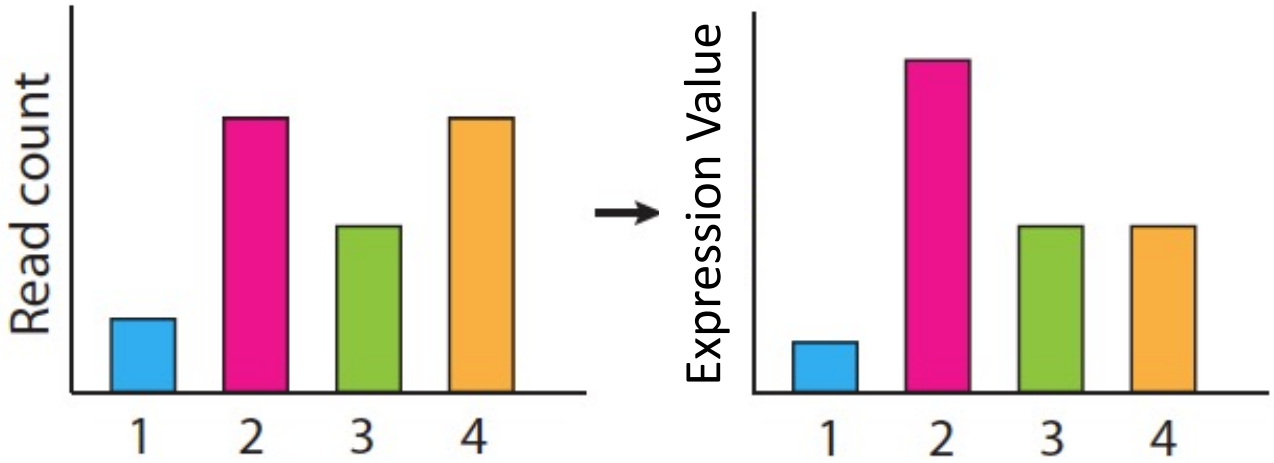




Slide courtesy of Cole Trapnell

Calculating expression of genes and transcripts





Slide courtesy of Cole Trapnell

Normalized Expression Values

- Transcript-mapped read counts are normalized for both length of the transcript and total depth of sequencing.

- Reported as: Number of RNA-Seq **F**ragments
  Per **K**ilobase of transcript
  per total **M**illion fragments mapped

FPKM

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

Transcripts per Million (TPM)

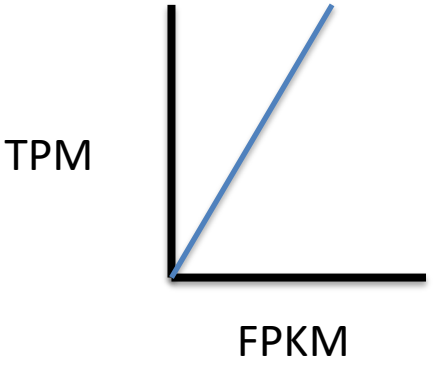
$$TPM_i = \frac{FPKM_i}{\sum_j FPKM} * 1e6$$

Preferred metric for measuring expression

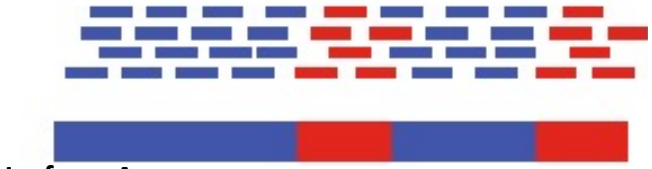
- Better reflects transcript concentration in the sample.
- Nicely sums to 1 million

Linear relationship between TPM and FPKM values.

Both are valid metrics, but best to be consistent.



Multiply-mapped Reads Confound Abundance Estimation

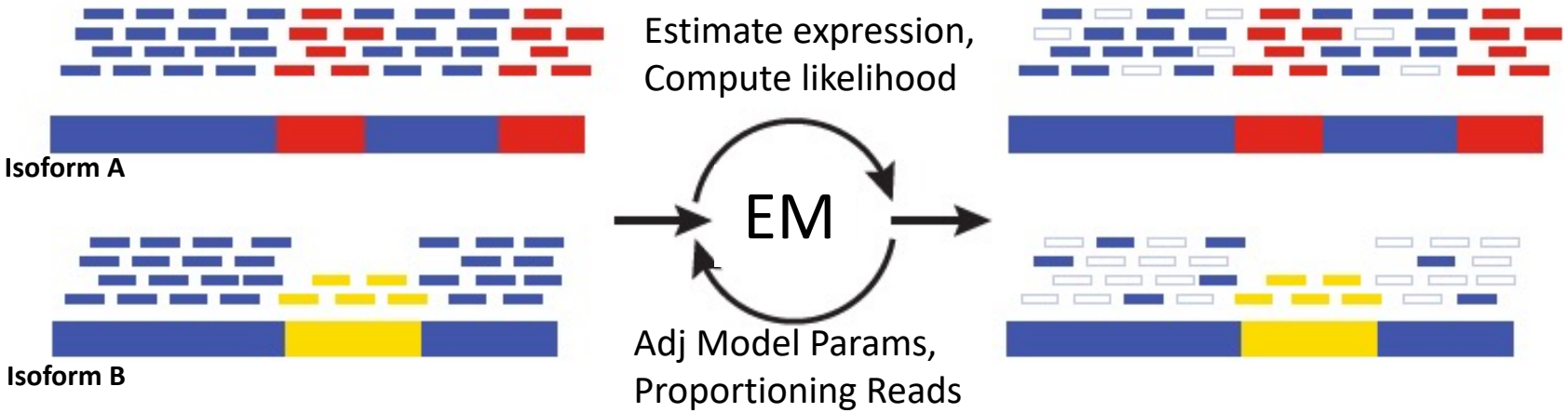


Isoform A

Isoform B

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:

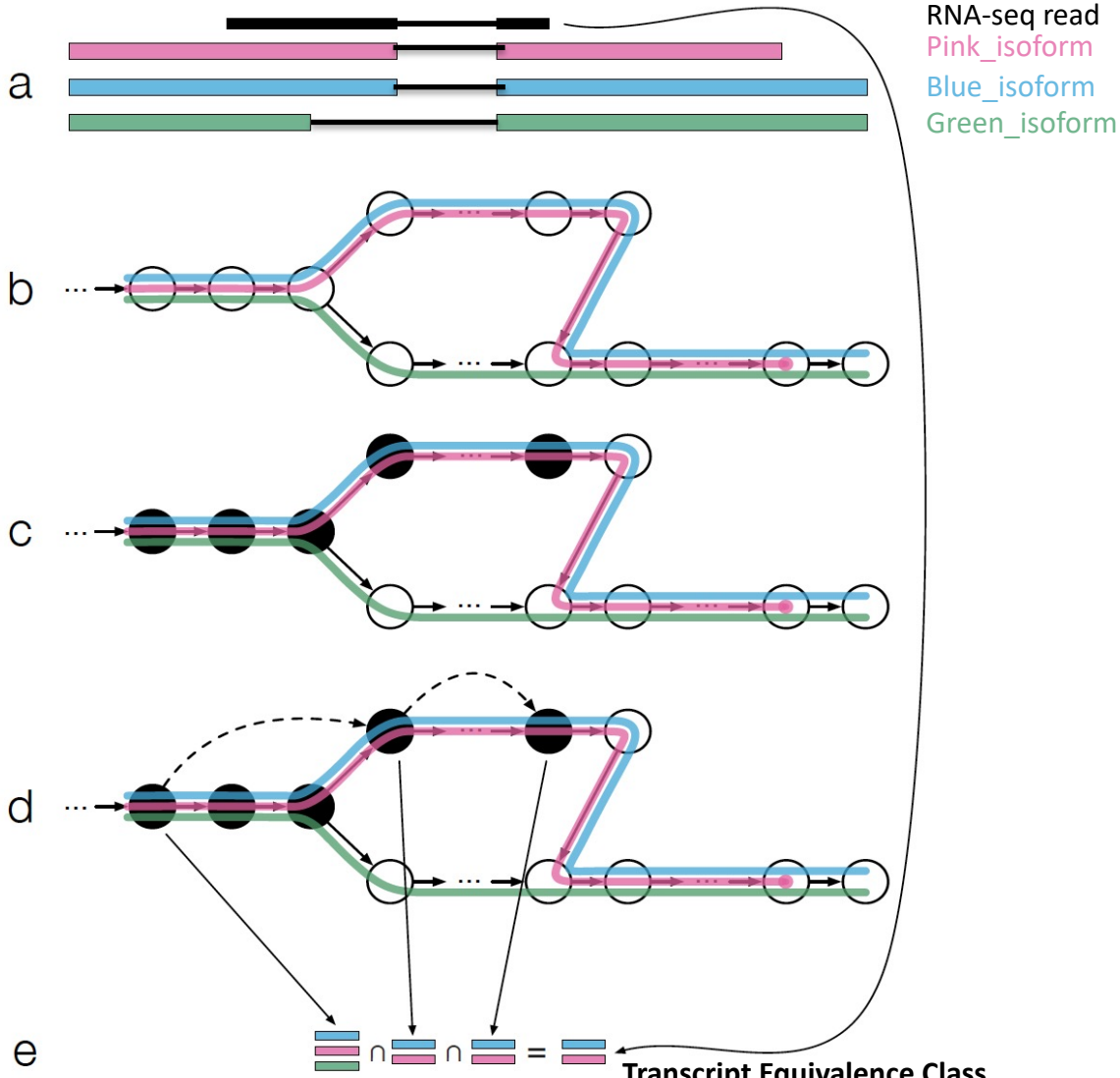
- RSEM (genome-free)
- Kallisto, Salmon (alignment-free)

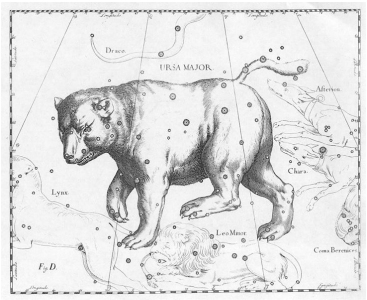
Fast Abundance Estimation Using Pseudo-alignments and Equivalence Classes

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

Alignment-Free!

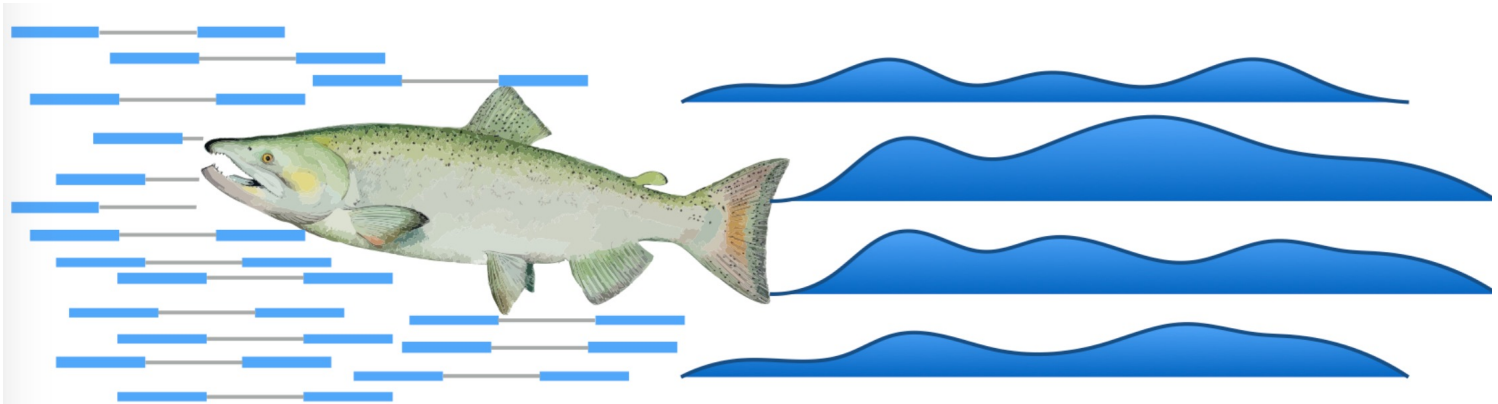
De bruijn graph for isoforms, not reads





https://pachterlab.github.io/kallisto/

Adapted from Fig 1 from Bray et al.



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

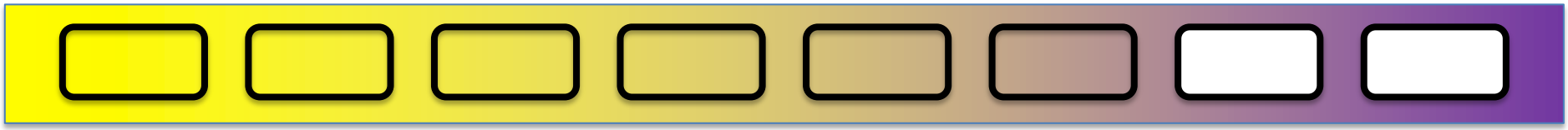
Rob Patro, Geet Duggal, Michael I Love, Rafael A Irizarry & Carl Kingsford

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/

Part 6. Differential Expression



Differential Expression Analysis

Thx, Charlotte Soneson! ☺

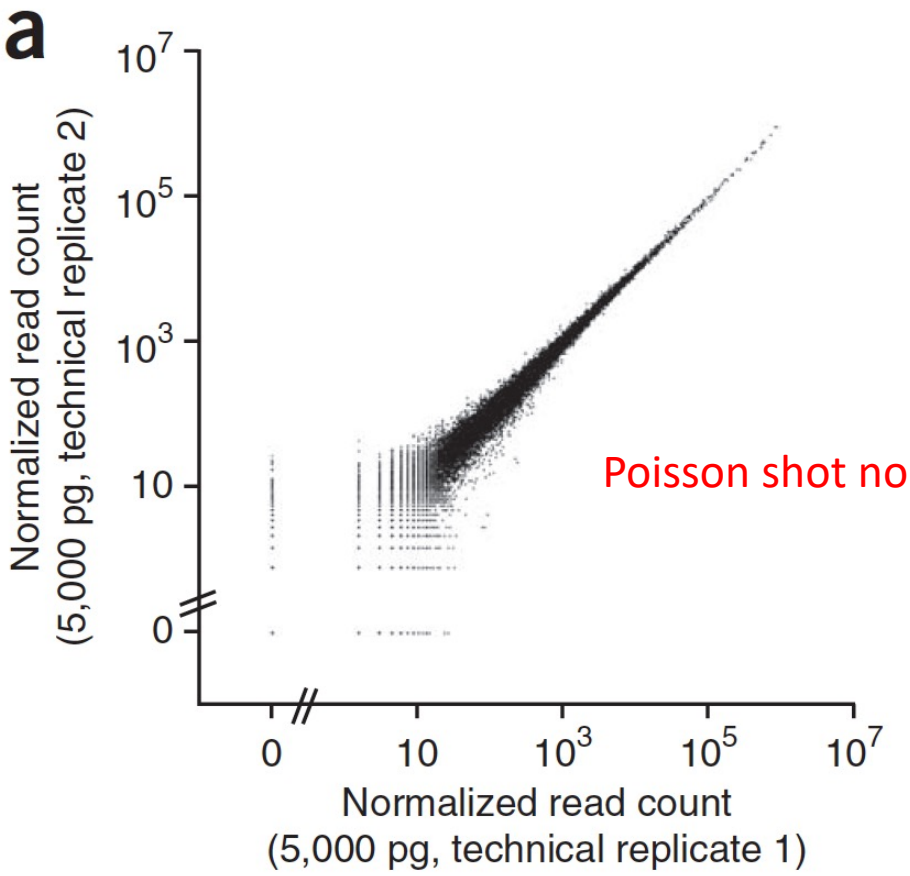
Differential Expression Analysis Involves

- Counting reads mapped to features
- Statistical significance testing

Beware of small counts leading to notable fold changes

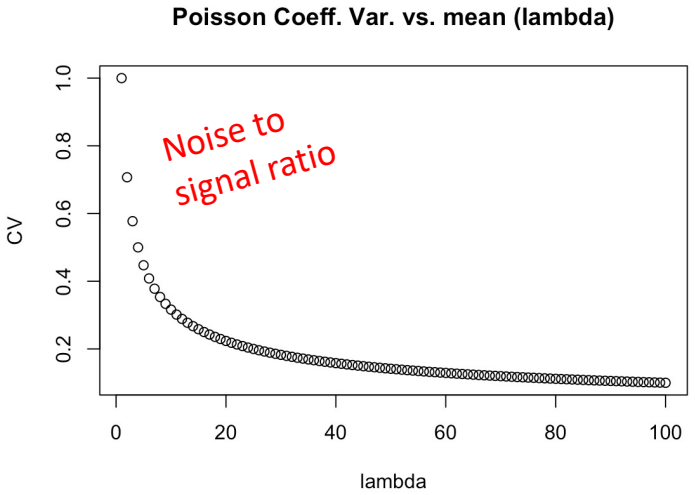
	Sample_A	**Sample_B**	**Fold_Change**	**Significant?**
Gene A	1	2	2-fold	No
Gene B	100	200	2-fold	Yes

Variation Observed Between Technical Replicates



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

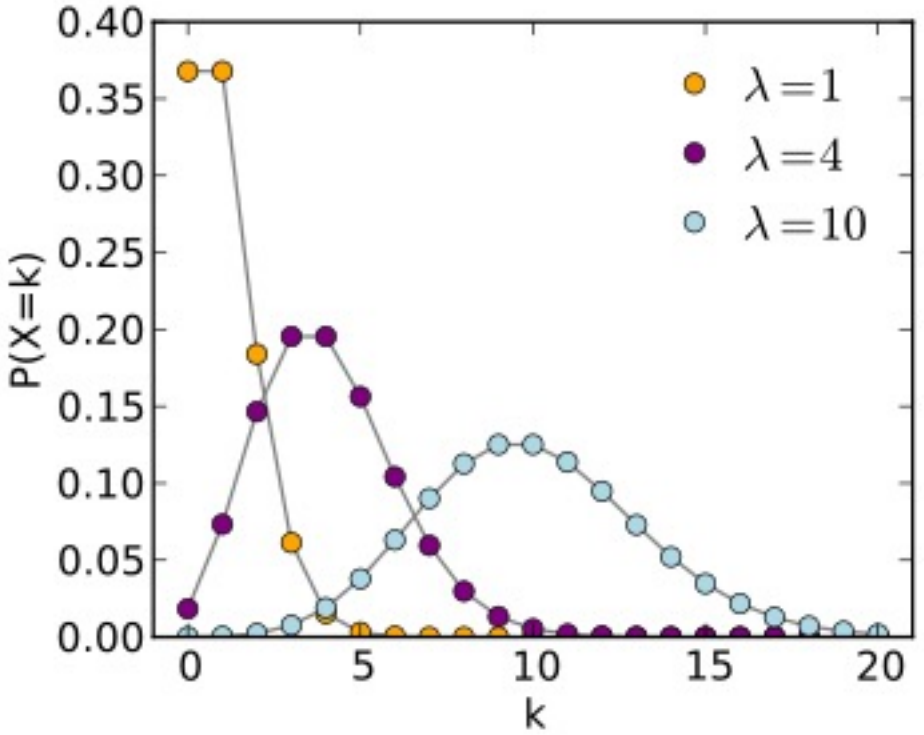
Poisson shot noise is high for small counts.



* 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



Mean # fragments

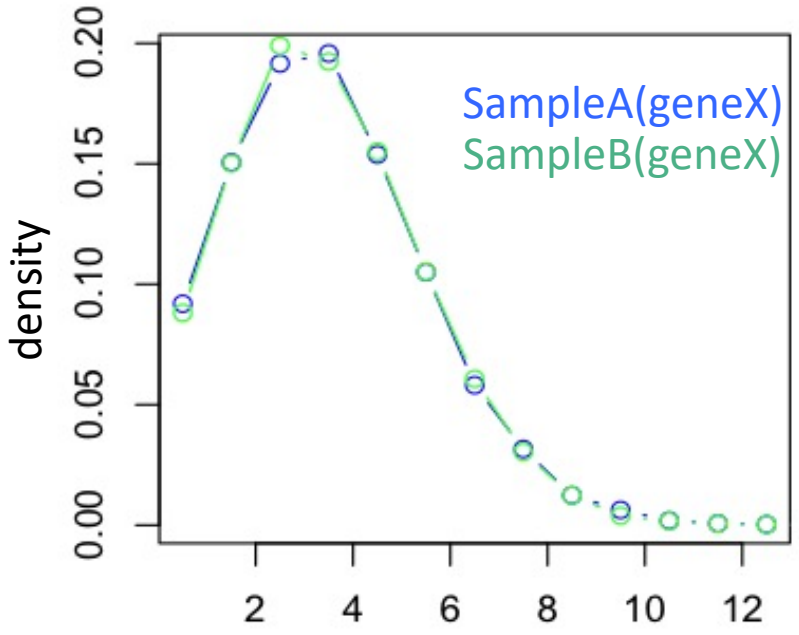
(observed read counts)

See: http://en.wikipedia.org/wiki/Poisson_distribution

Example: One gene*not* differentially expressed

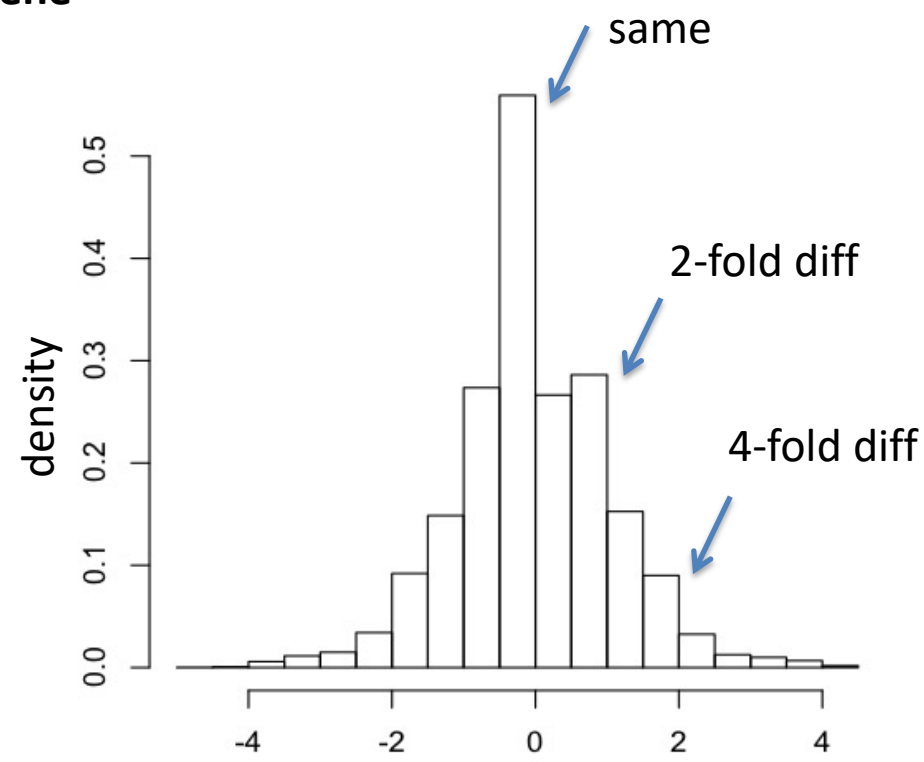
Example: SampleA(gene) = SampleB(gene) = 4 reads

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



(k) number of reads observed

Dist. of $\log_2$(fold change) values

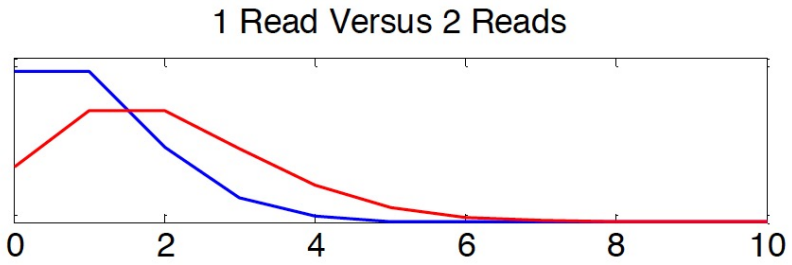


x = $\log_2$(SampleA/SampleB)

Sequencing Depth Matters

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

P(x=k)

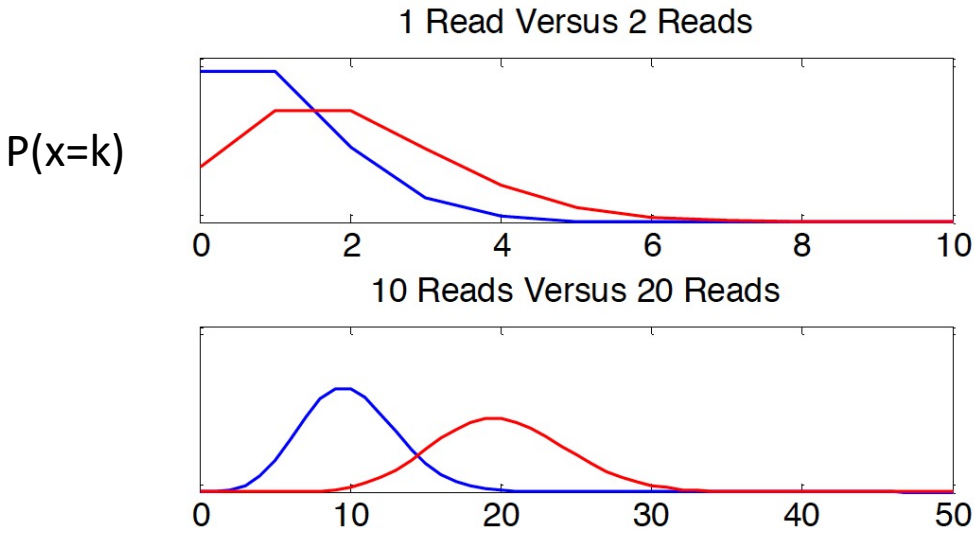


No confidence in 2-fold difference. Likely observed by chance.

From: http://gkno2.tumblr.com/post/24629975632/thinking-about-rna-seq-experimental-design-for
and from supplementary text of Busby et al., Bioinformatics, 2013

Sequencing Depth Matters

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

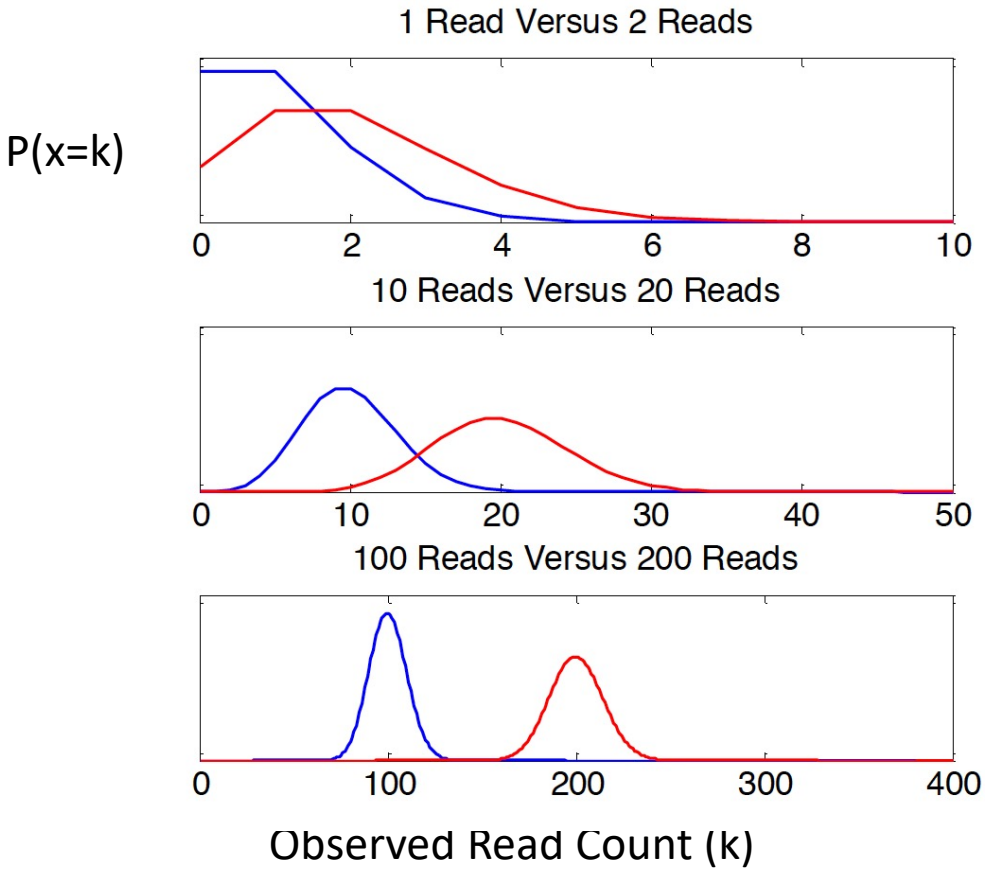


No confidence in 2-fold difference. Likely observed by chance.

From: http://gkno2.tumblr.com/post/24629975632/thinking-about-rna-seq-experimental-design-for
and from supplementary text of Busby et al., Bioinformatics, 2013

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.

From: http://gkno2.tumblr.com/post/24629975632/thinking-about-rna-seq-experimental-design-for
and from supplementary text of Busby et al., Bioinformatics, 2013

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*

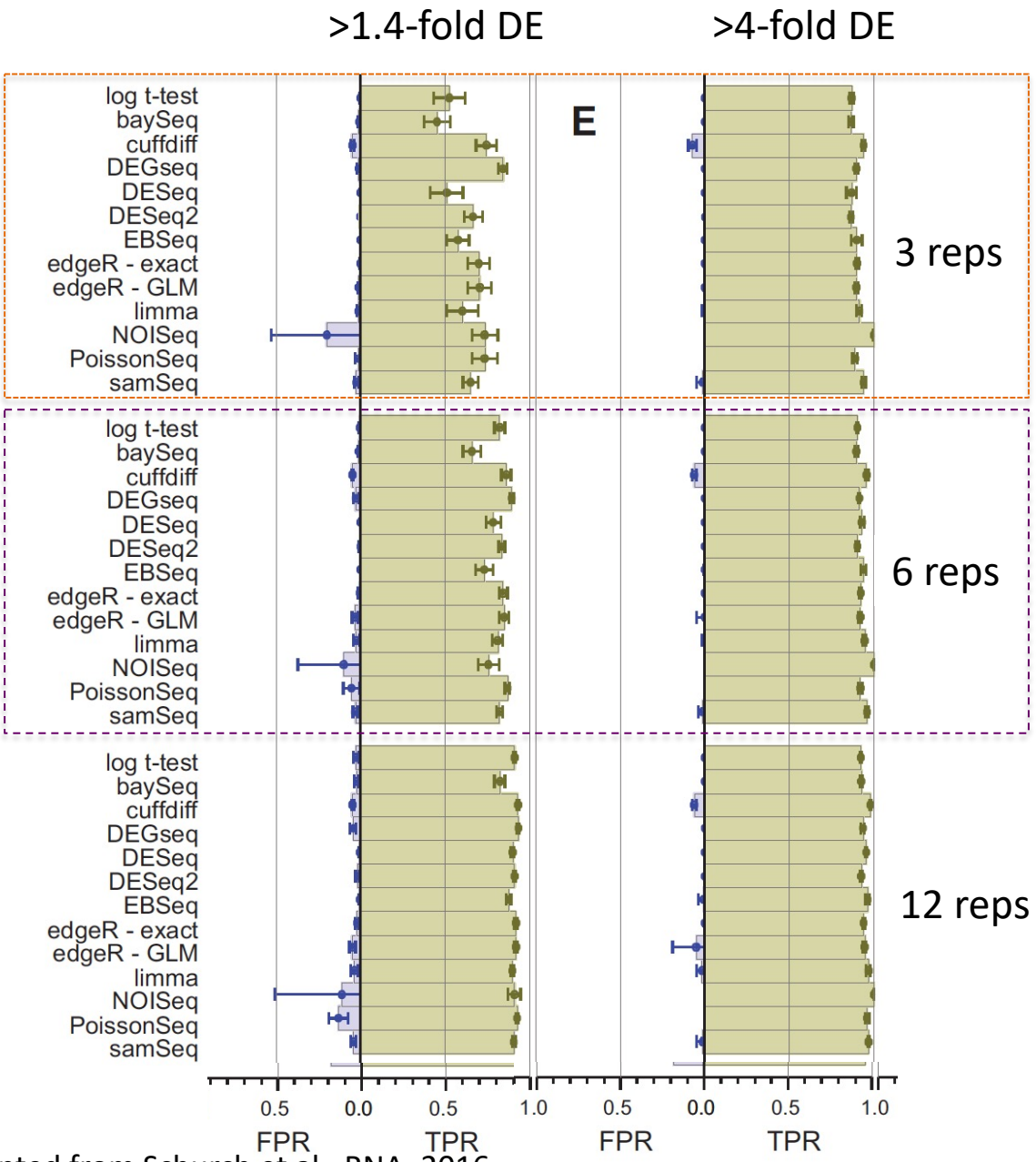
However, biological replicates *ARE* essential

total_variance = technical_variance + 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



At a minimum, do 3 bio replicates

Respectable goal

*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

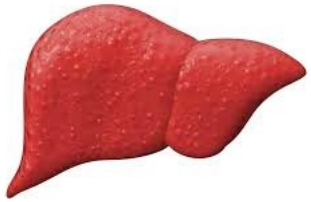
Avg. expression level

Significance

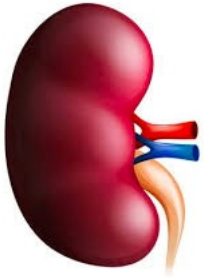
-- Before Comparing RNA-Seq Samples --

Some Cross-sample Normalization May Be Required

eg.

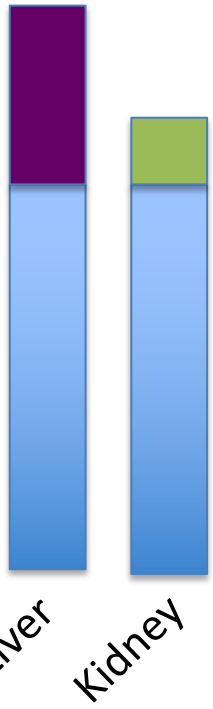


Vs.

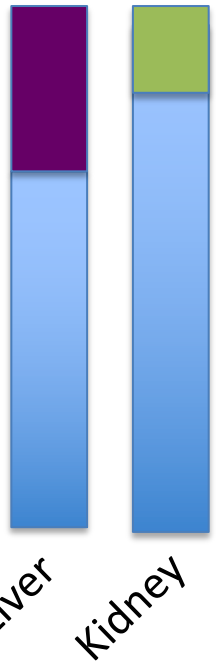


Why cross-sample normalization is important

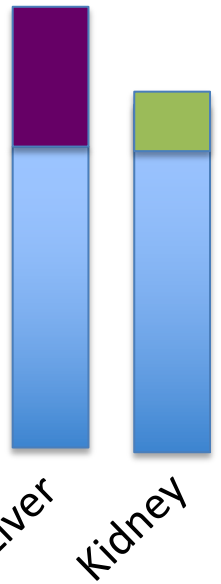
Absolute RNA quantities per cell



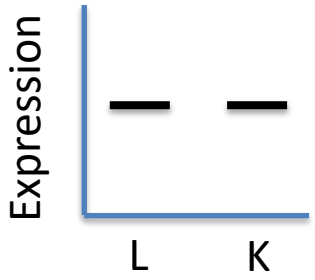
Measured relative abundance via RNA-Seq

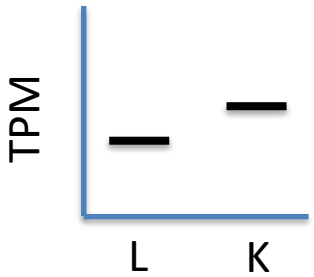


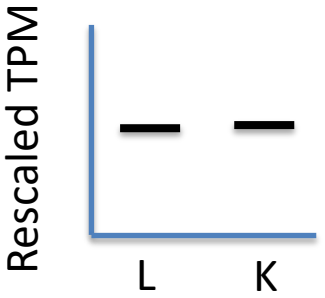
Cross-sample normalized (rescaled) relative abundance



eg. Some housekeeping gene's expression level:







Cross-sample Normalization Required
Otherwise, housekeeping genes look diff expressed due to sample composition differences

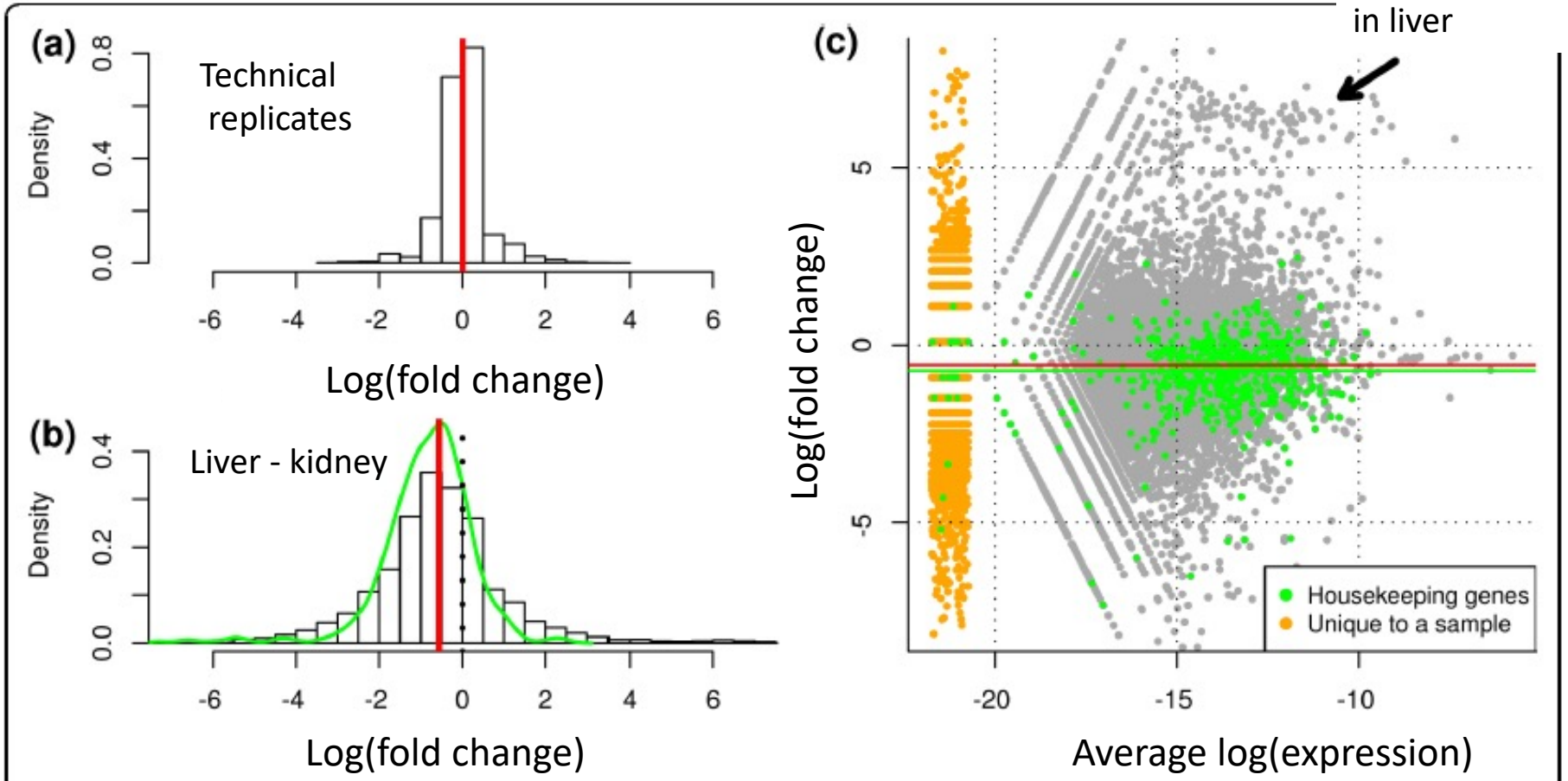
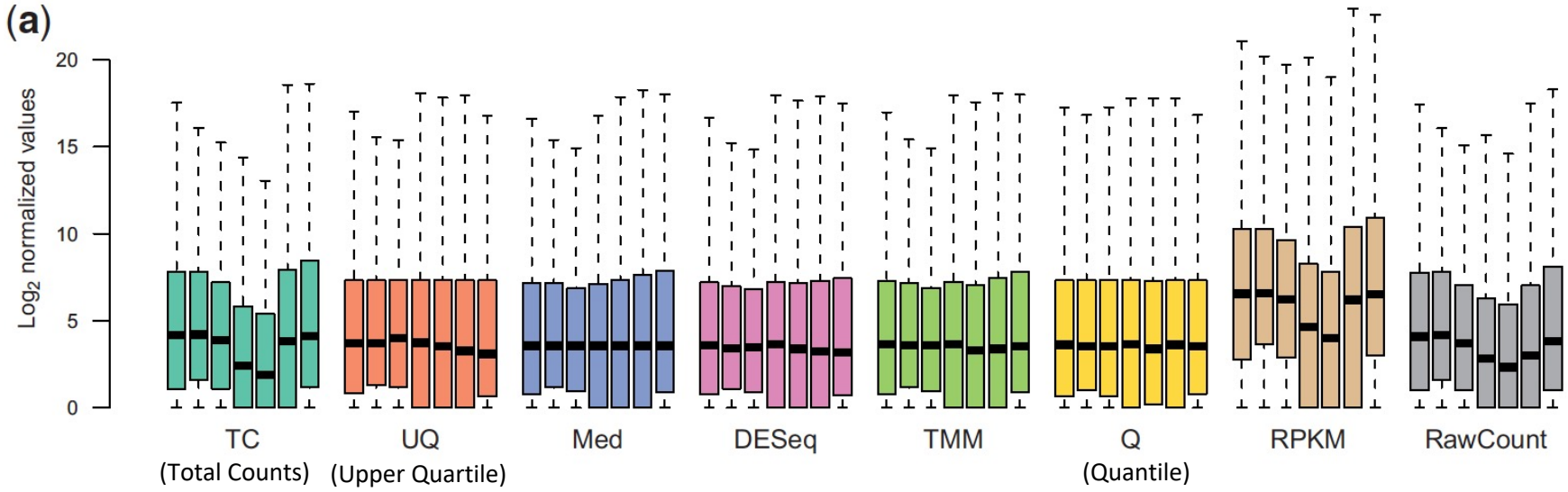


Figure 1 Normalization is required for RNA-seq data. Data from [6] comparing log ratios of **(a)** technical replicates and **(b)** liver versus kidney expression levels, after adjusting for the total number of reads in each sample. The green line shows the smoothed distribution of log-fold-changes of 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.

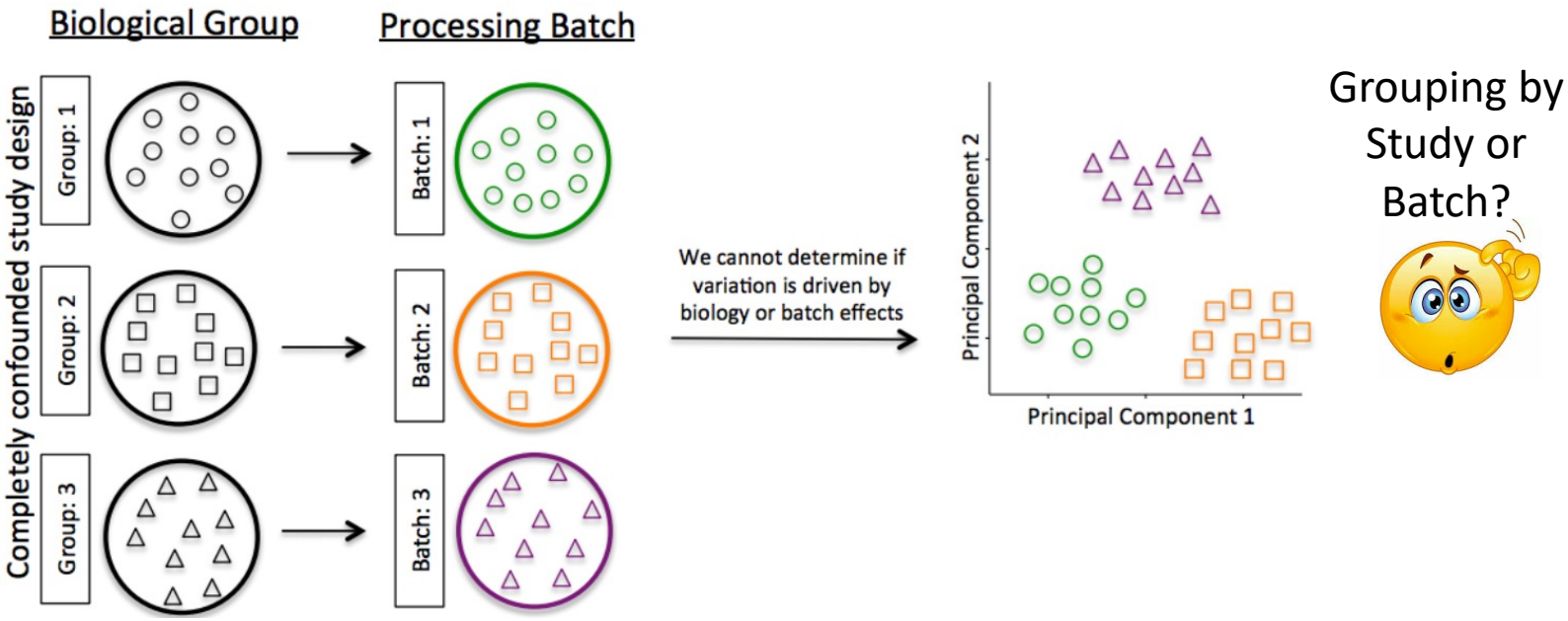
Adapted from: Robinson and Oshlack, Genome Biology, 2010

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

Avoid Batch Effects

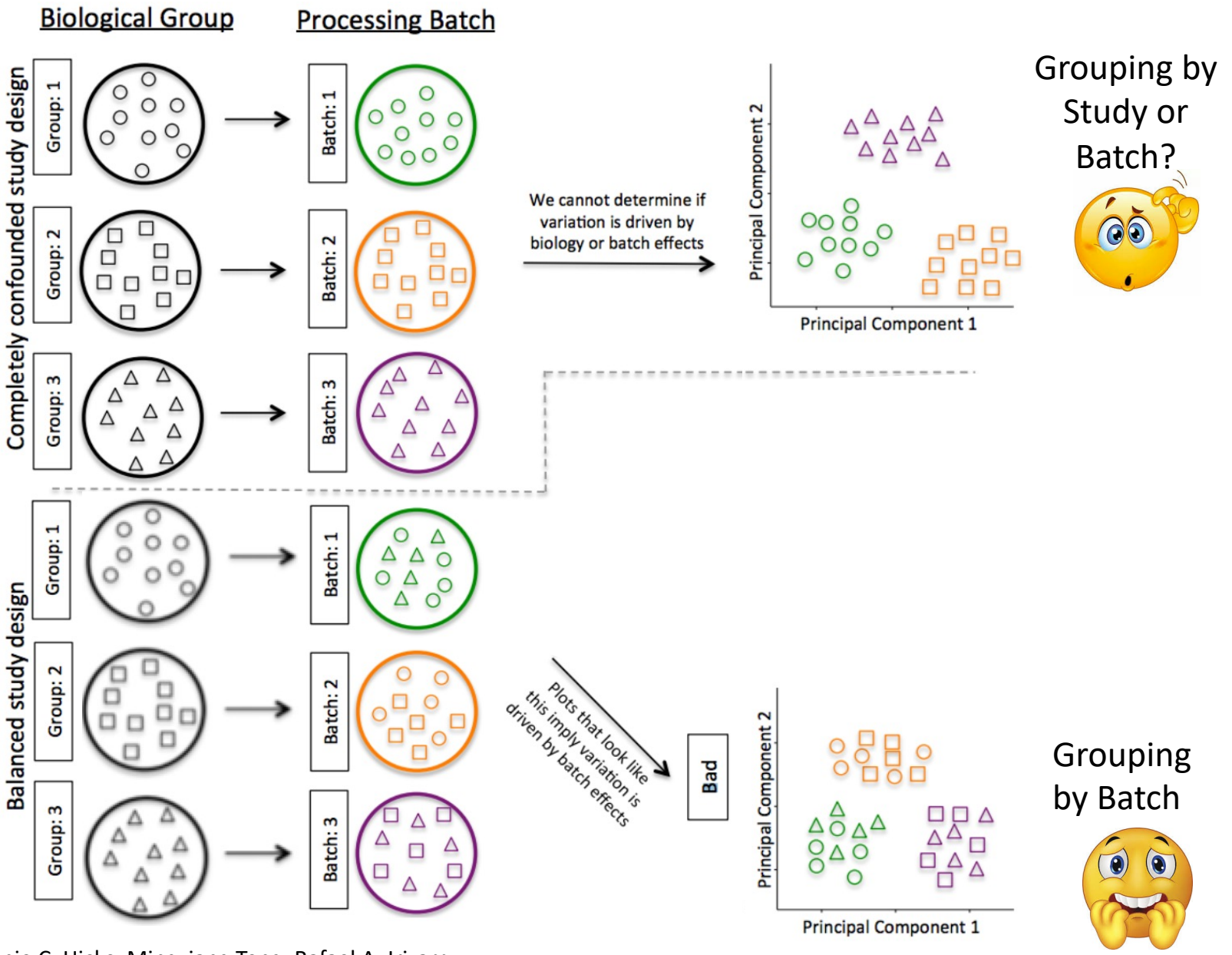


Batch variable types:

- 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.

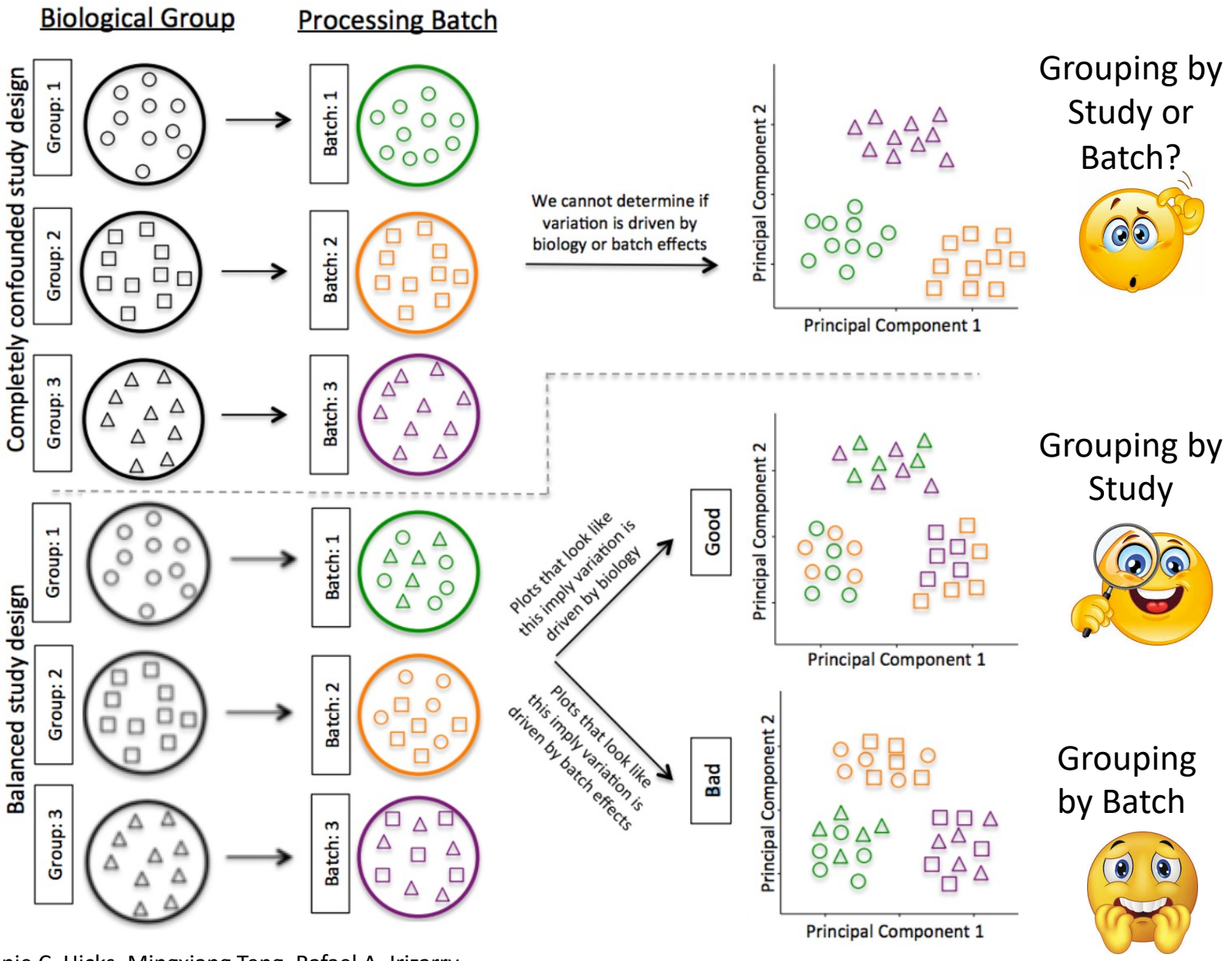
Avoid Batch Effects



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



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. *?!?!?*

This Article | Info for Authors | Subscribe | About

PNAS

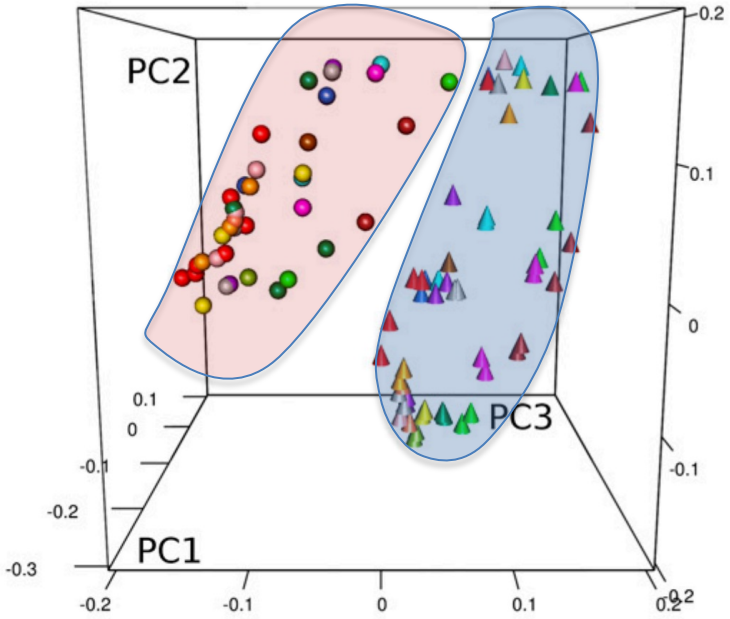
Proceedings of the National Academy of Sciences of the United States of America

Proc Natl Acad Sci U S A. 2014 Dec 2; 111(48): 17224–17229.
Published online 2014 Nov 20. doi: 10.1073/pnas.1413624111
Genetics

PMCID: PMC4260565
PMID: 25413365

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]

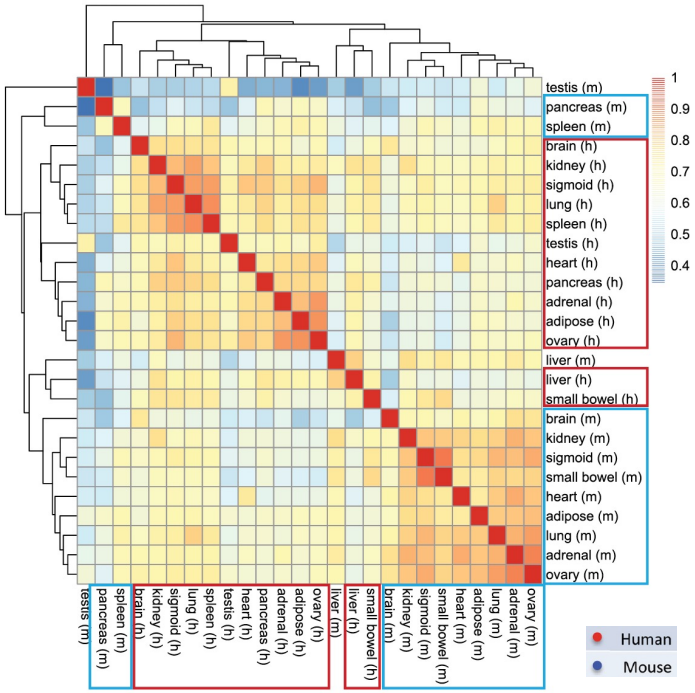
Yoav Gilad, Orna Mizrahi-Man

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

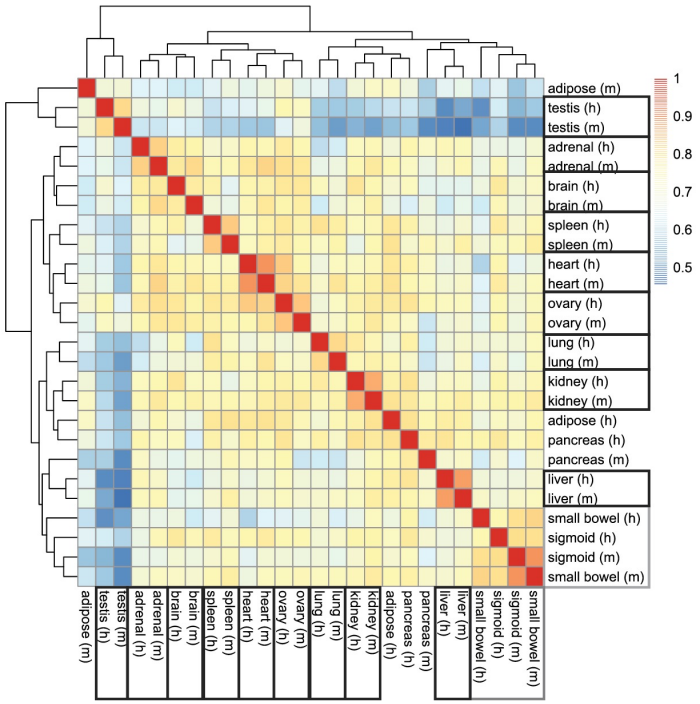
Grouping of samples by Sequencing Batch

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

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



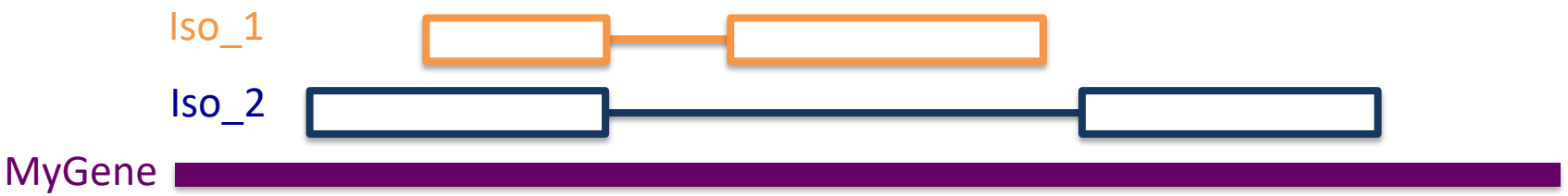
Post Batch Correction:
Tissue patterns more similar than by species



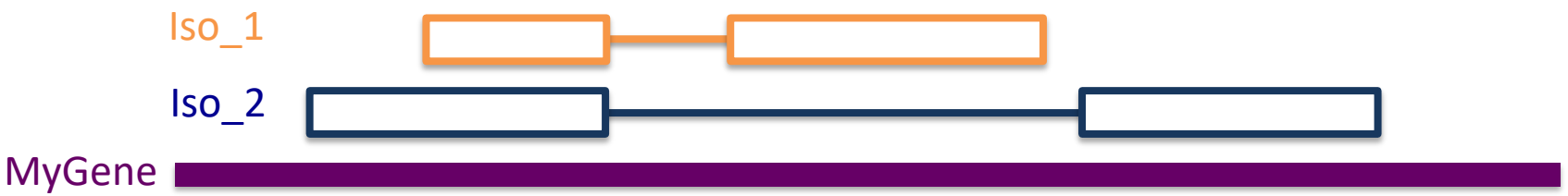
Flavors of Differential Expression Analyses

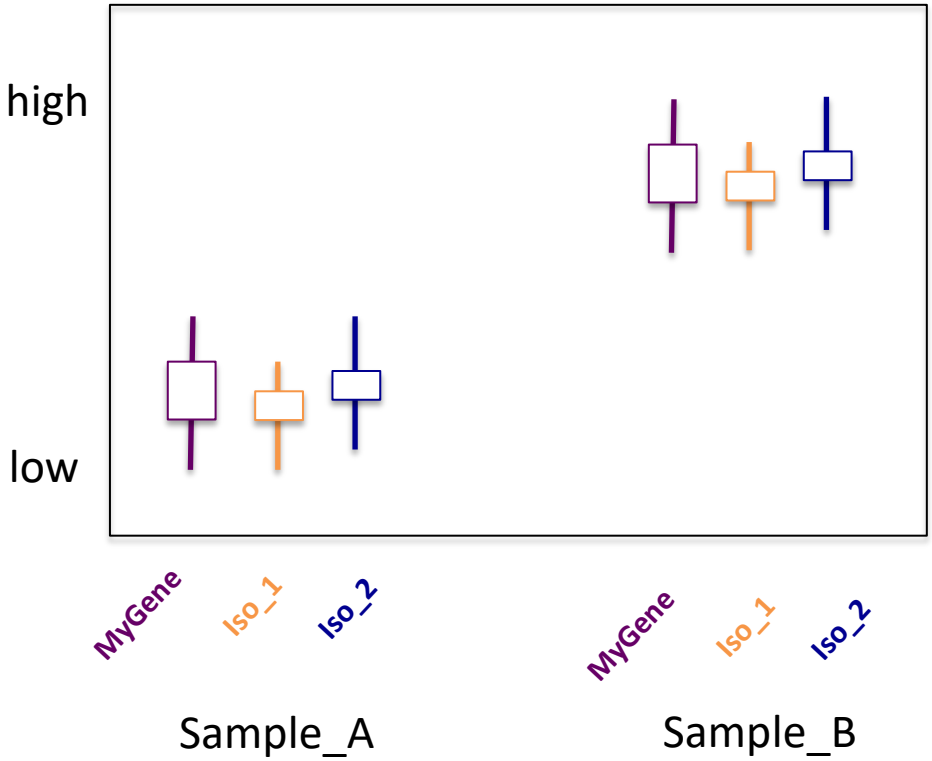
- Transcripts:
  - Differential Transcript Expression (DTE)
  - Differential Transcript Usage (DTU)
  - Differential Exon Usage (DEU)
- Gene:
  - Differential Gene Expression (DGE)

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



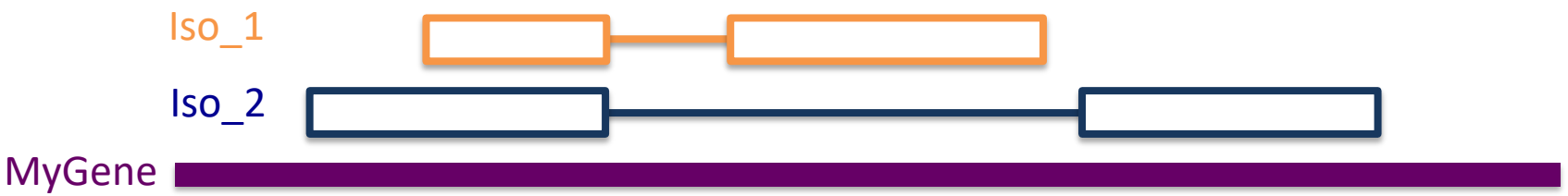
Differential Gene Expression (DGE) and Differential Transcript Expression (DTE) (Example 1)

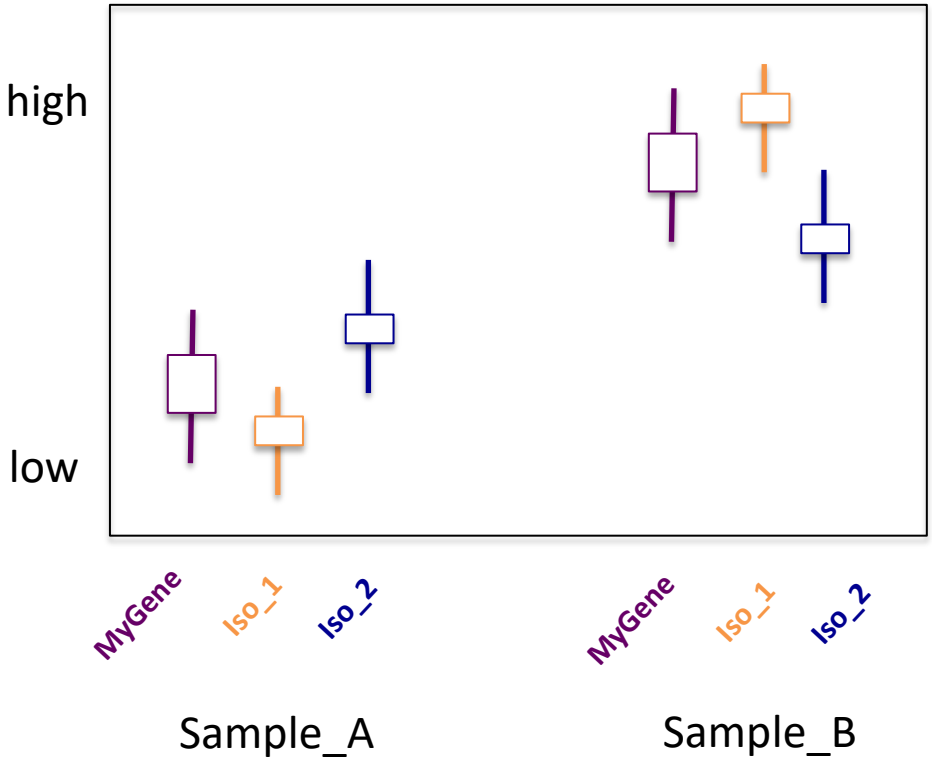




Feature	Diff Expressed?
MyGene	Yes
Iso_1	Yes
Iso_2	Yes
Diff. Transcript Usage ? (eg. Isoform switching)	No

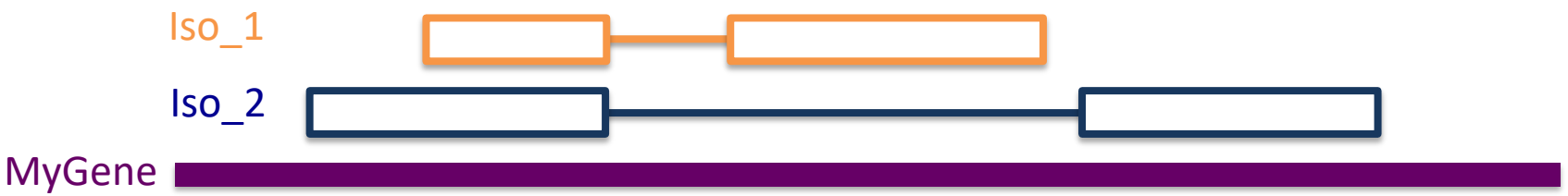
Differential Gene Expression (DGE) and Differential Transcript Expression (DTE) (Example 2)

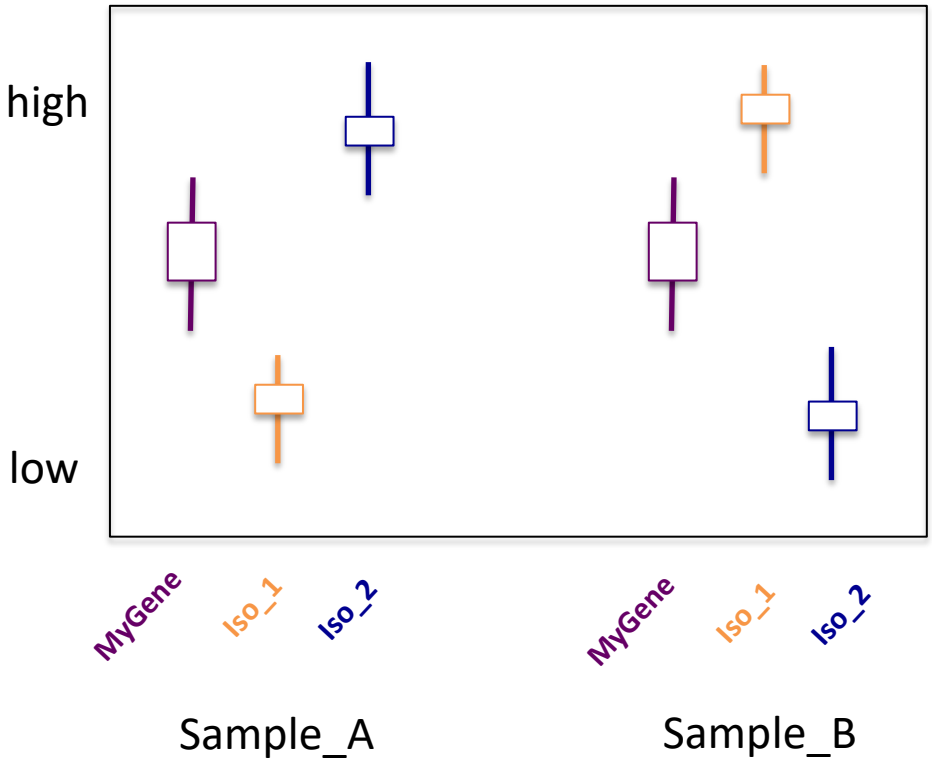




Feature	Diff Expressed?
MyGene	Yes
Iso_1	Yes
Iso_2	Yes
Diff. Transcript Usage ? (eg. Isoform switching)	Yes

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



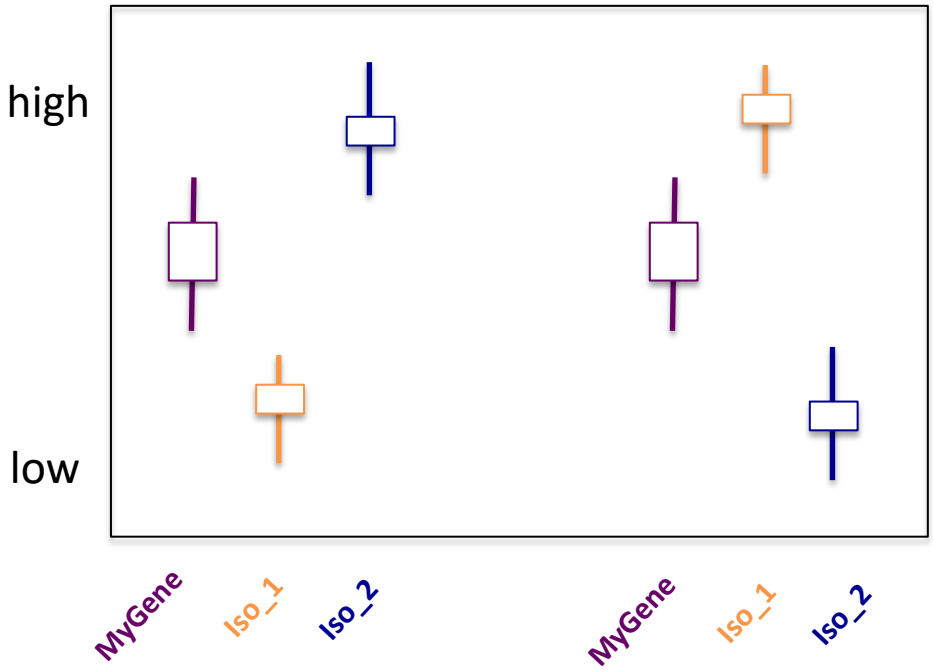


Feature	Diff Expressed?
MyGene	No
Iso_1	Yes
Iso_2	Yes
Diff. Transcript Usage ? (eg. Isoform switching)	Yes

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

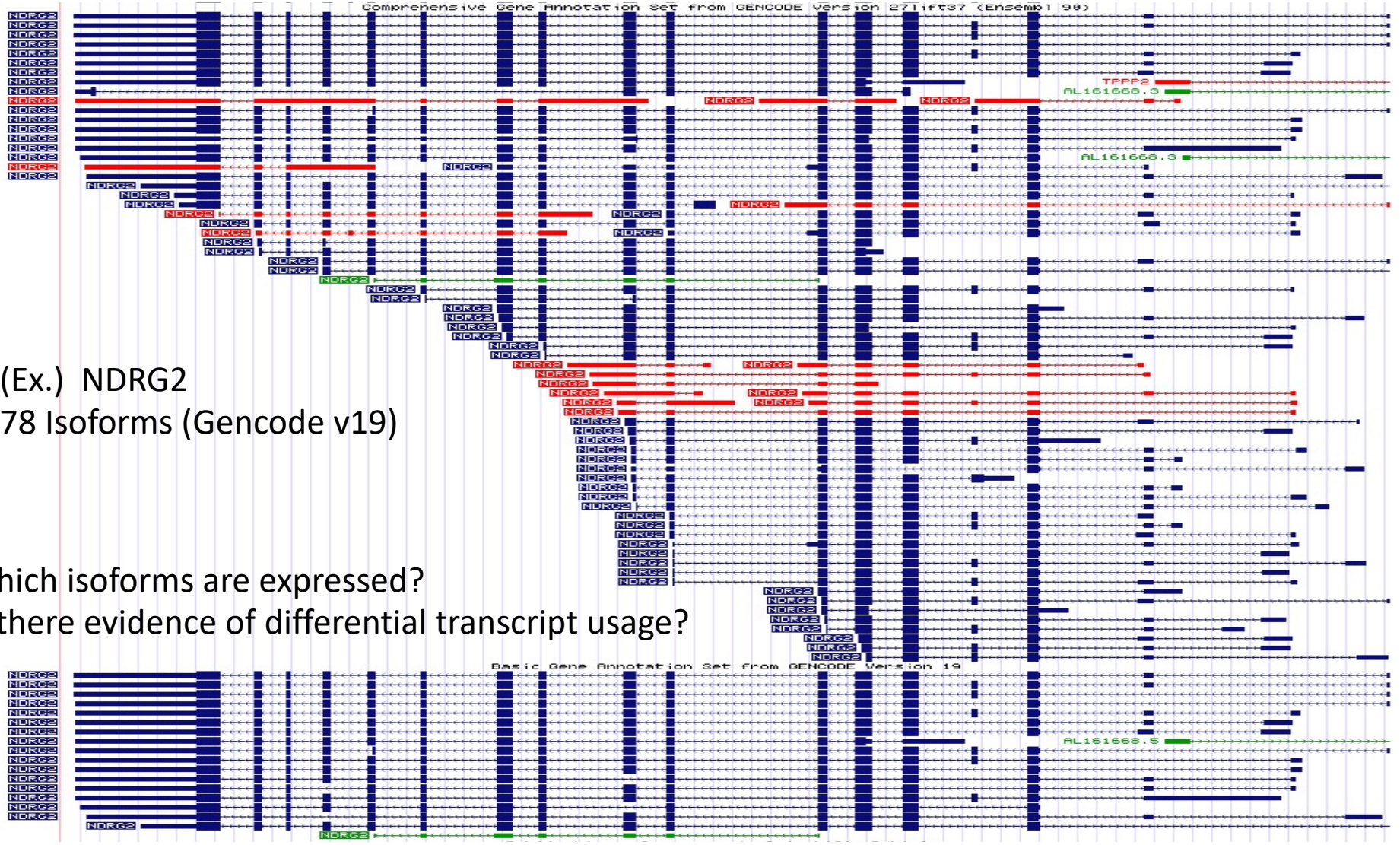
From Gene-level view (DGE): not apparent
From Transcript-level view: Yes, gene should be acknowledged as having changed.

Prevailing viewpoint:
DTE or DTU -> Gene is Differentially Expressed



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

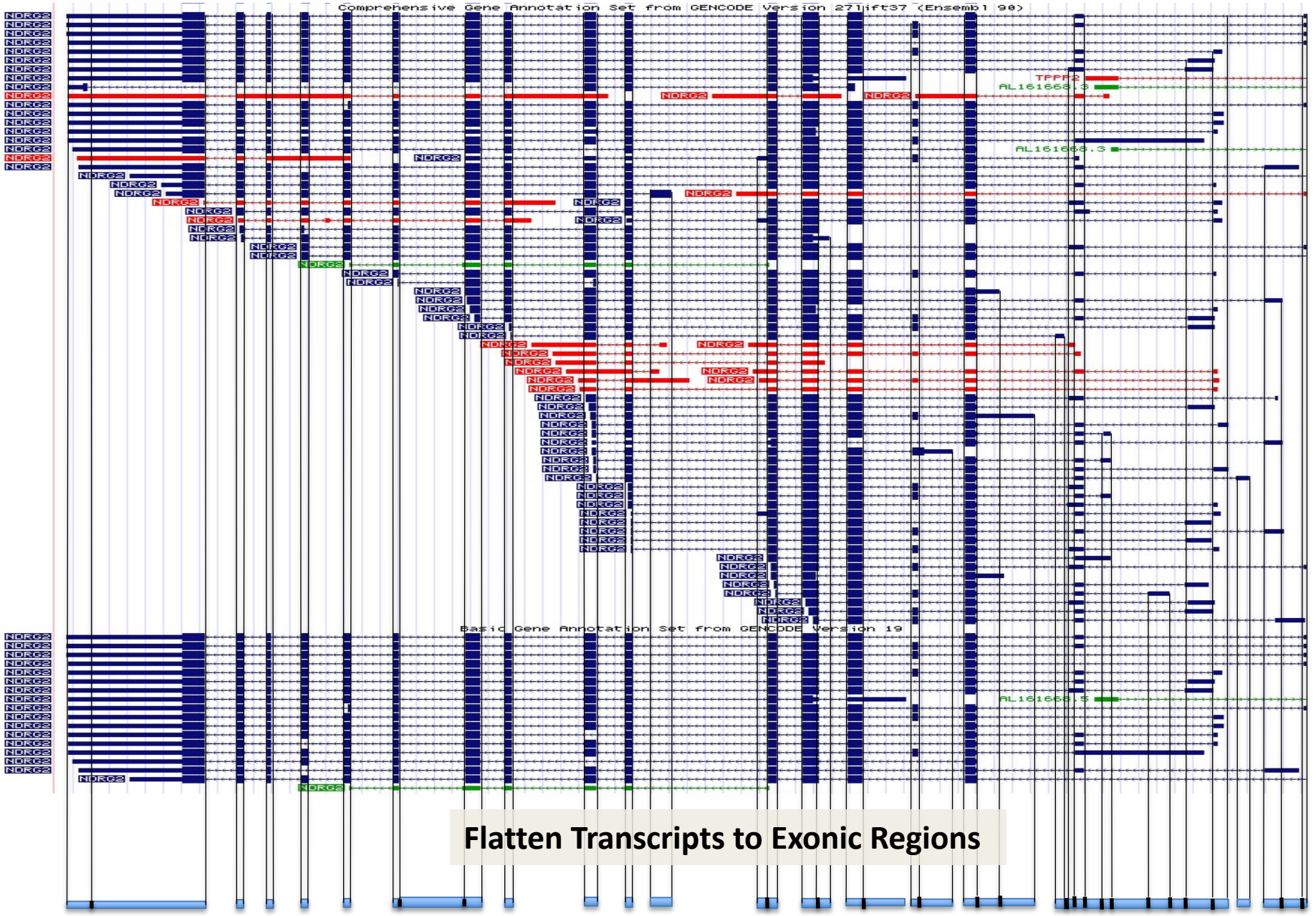
High Confidence Differential Transcript Expression is Difficult to Attain With Many Candidate Isoforms



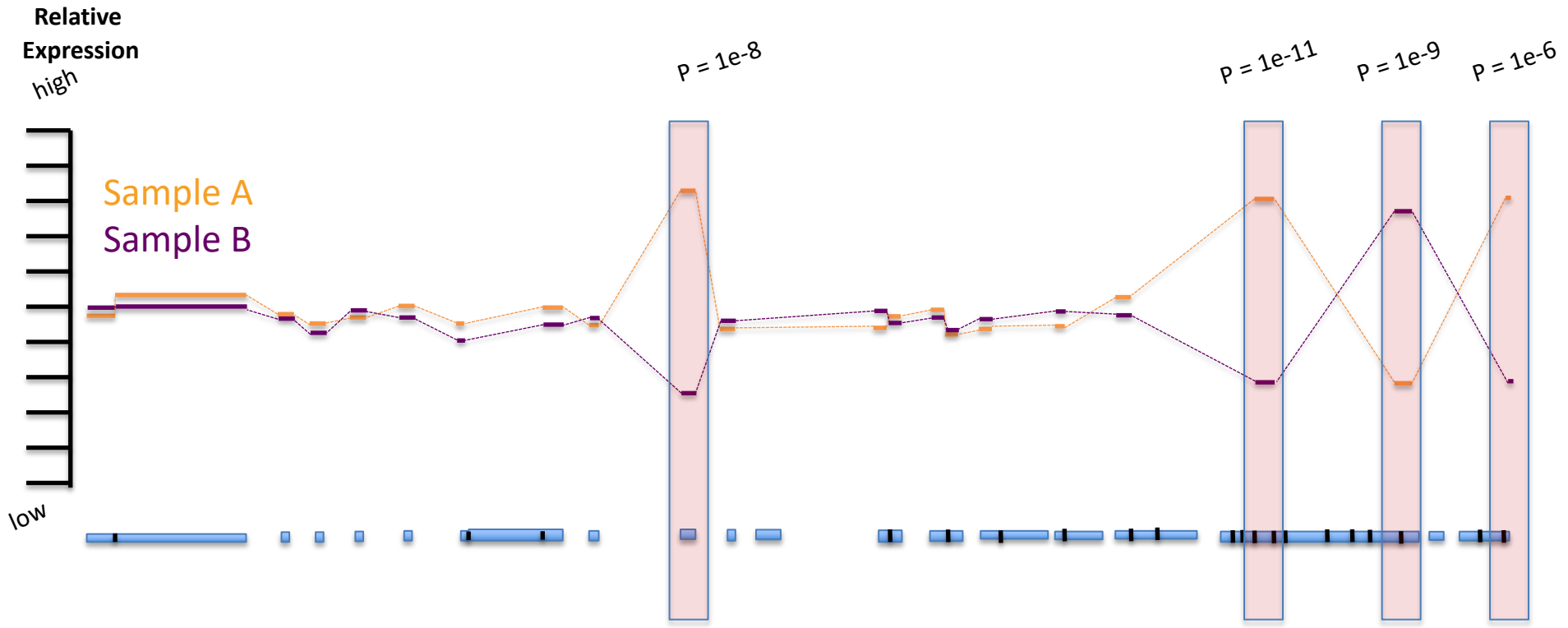
(Ex.) NDRG2
78 Isoforms (Gencode v19)

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

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



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



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

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

Detecting differential usage of exons from RNA-seq data

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

Averaged Replicates

Each Replicate

Flattened gene structure:

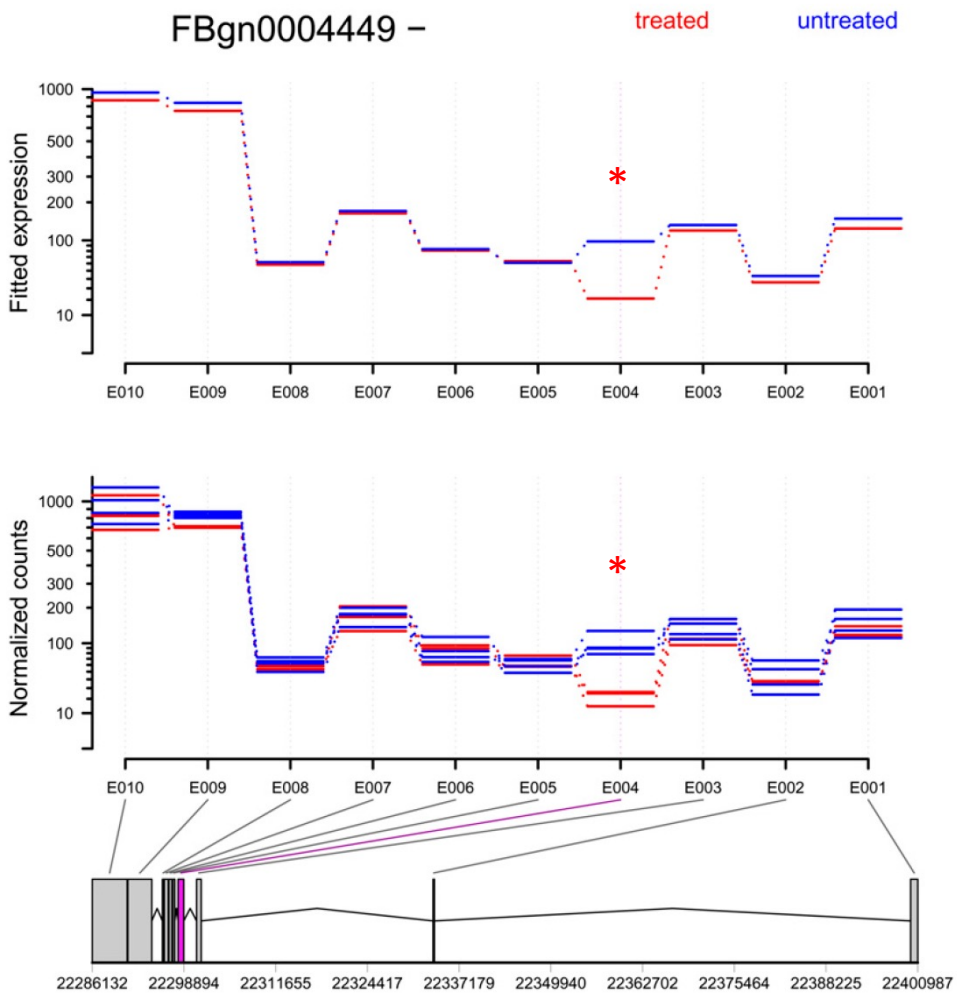


Figure 3. The treatment of knocking down the splicing factor *pasilla* affects the fourth exon (counting bin E004) of the gene *Ten-m* (CG5723). (*Top* panel) Fitted values according to the linear model; (*middle* panel) normalized counts for each sample; (*bottom* panel) flattened gene model. (Red)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

Enabling Differential Transcript Usage Analysis for De novo Transcriptome Assemblies

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

Genome Biology

METHOD Open Access

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

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

Enabling Differential Transcript Usage Analysis for De novo Transcriptome Assemblies

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

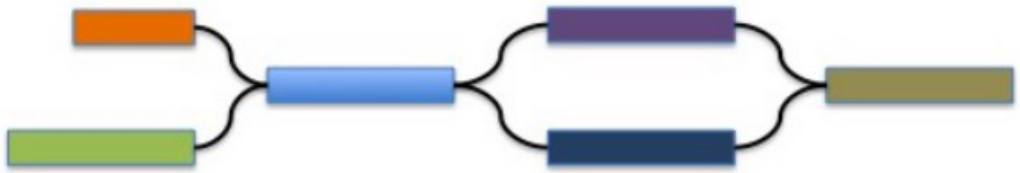
Genome Biology

METHOD Open Access

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

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

Transcript splice graph:



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

Enabling Differential Transcript Usage Analysis for De novo Transcriptome Assemblies

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

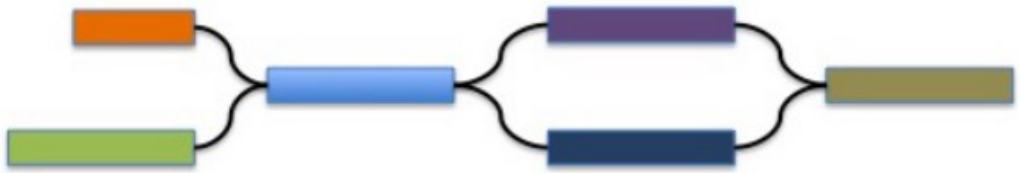
Genome Biology

METHOD Open Access

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

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

Transcript splice graph:



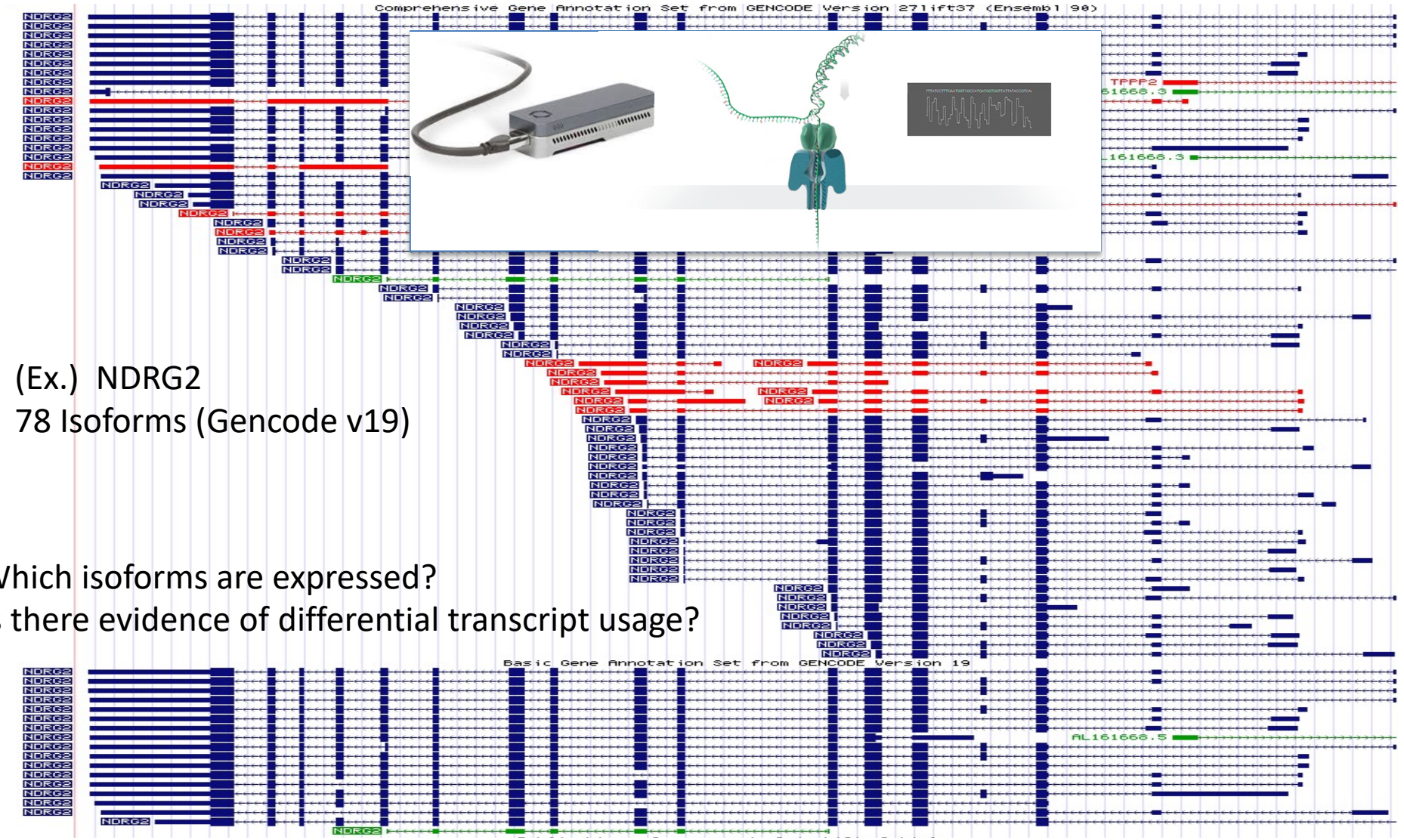
Linearize graph via topological sorting
or graph multiple alignment

SuperTranscript:

DEXseq for DTU,
GATK for Variant Detection

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

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



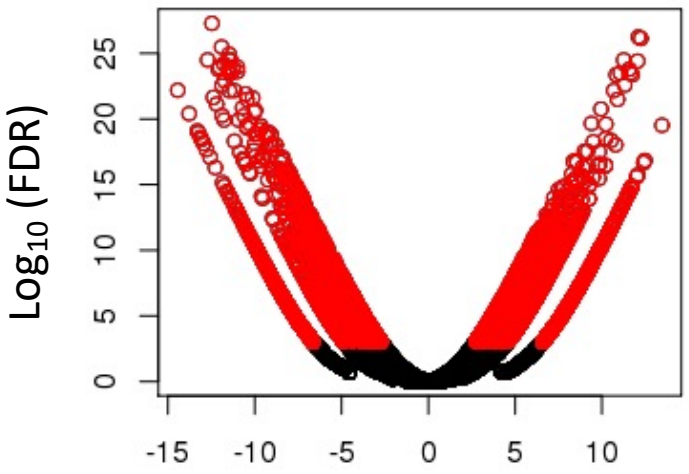
(Ex.) NDRG2
78 Isoforms (Gencode v19)

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

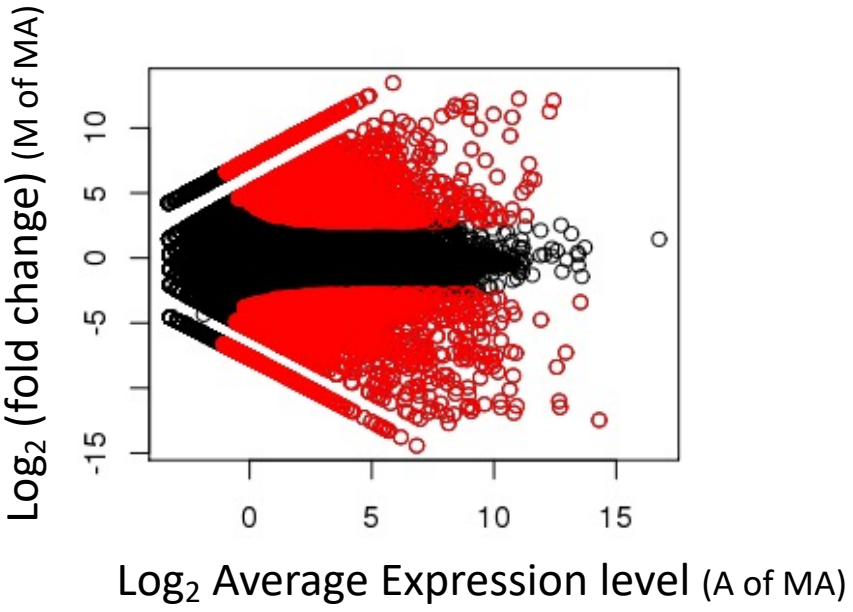
Visualization of DE results and Expression Profiling

Plotting Pairwise Differential Expression Data

Volcano plot
(fold change vs. significance)

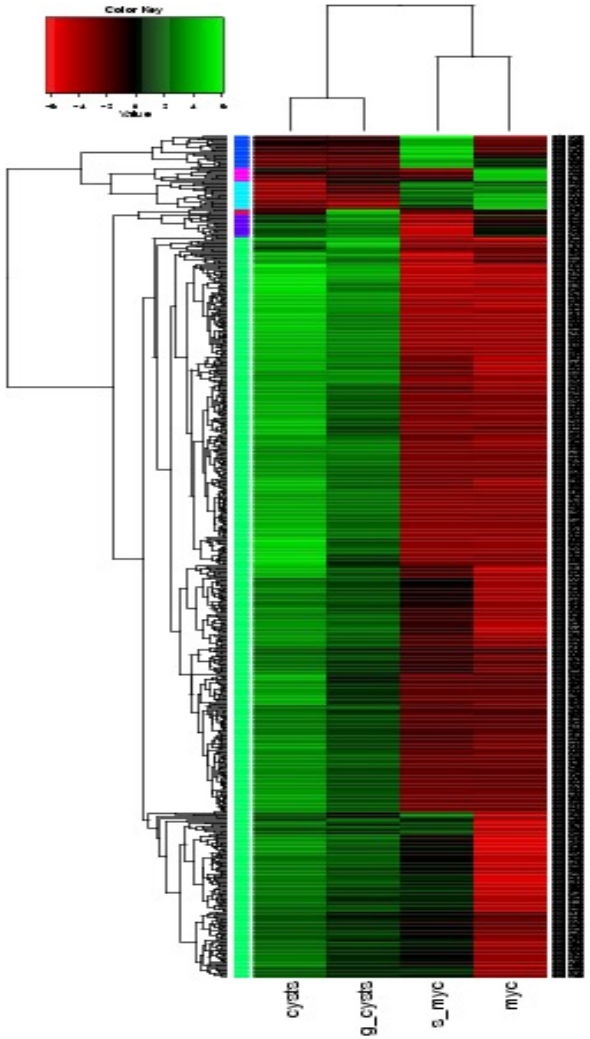


MA plot
(abundance vs. fold change)



Significantly differently expressed transcripts have FDR <= 0.001
(shown in red)

Comparing Multiple Samples



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

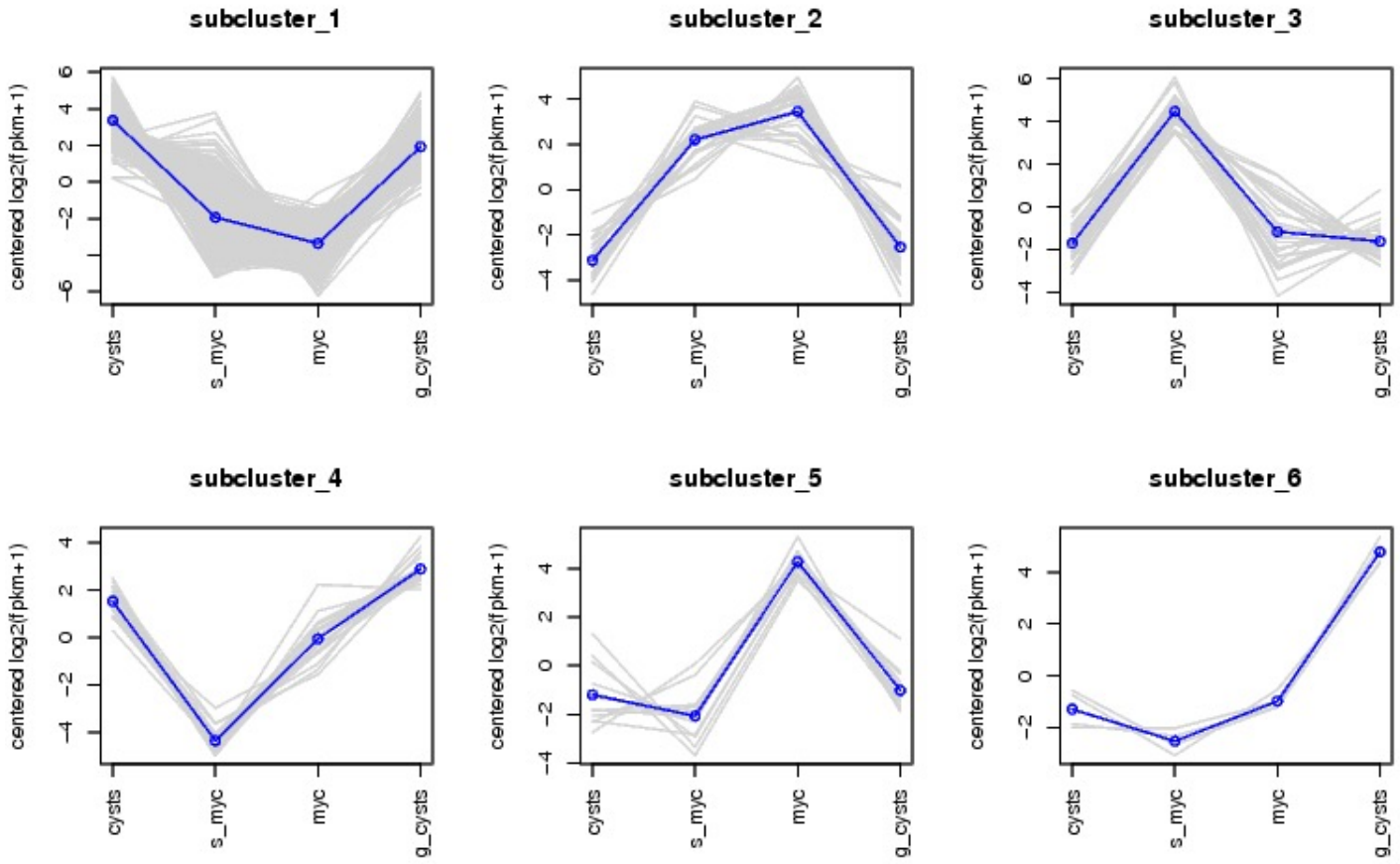
Clustering can be performed across both axes:

-cluster transcripts with similar expression patters.

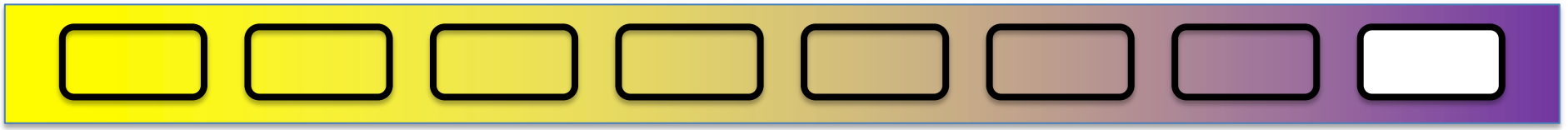
-cluster samples according to similar expression values among transcripts.

Examining Patterns of Expression Across Samples

Can extract clusters of transcripts and examine them separately.



Part 7. Functional Annotation



Transcript Functional Annotation

GGAGCTGGAGGCCCCAGGCAACTACACCGTCCACGTACCCAGAGGGGCTGGGCCCTCCC
ACCAGAGACCACGCCCTGGTGTGCCTTAGGGGCCCTGGTTTGTTAGTCTCTGAGTGTGCA
GTTGCTGCACATGGGGCCCTGGCGCTTGCTGCACCAACTTCCTGTTGGGCCCGTGGTCCT
TGGAGGCATGCAGTTCAGCAGACAGTGACTCAGCCATCCACCCAACATGCGGAACGTGTC
TCTTCTGCAGGTCCCGGTCCACAGCAGGATTCCCCCTCTGTGAAAAGGCACGCTGATCTG
TCTGGA … TCGAC
TCTCCC … TCCCA
AAAGAC … CCTGG
GGCTTC … CCTAA
TGACCT … TGCTG
GAAAAC … CAGCC
TTGTCA … TTCCA
GGAAGCACATAATTGAAGGACTGAAAGCGTCCCTGGAGCGGCTGCAGCTGGAGTACGTGG
ATGTGGTTTTTGCCAACCGCCCAGACCCCAACACGCCCATGGAAGAGACCGTGCGGCCA
TGACCCATGTCATCAACCAGGGGATGGCCATGTACTGGGGCACATCACGCTGGAGCTCCA
TGGAGATCATGGAGGCCTACTCGGTGGCTCGGCAGTTCAACCTGATCCCGCCCATCTGCG
AGCAAGCGGAATATCACATGTTCCAGAGGGAGAAGGTGGAGGTCCAGCTGCCAGAGCTGT
TCCACAAGATAGGAGTAGGTGCCATGACCTGGTCCCCTCTGGCGTGCGGCATCGTCTCAG
GGAAGTATGACAGCGGGATCCCACCCTACTCCAGAGCCTCCCTGAAGGGCTACCAGTGGT
TGAAGGACAAGATCCTGAGTGAGGAGGGTCGCCGCCAGCAGGCCAAGCTGAAGGAACTGC
AGGCCATTGCCGAACGCCTGGGCTGCACCCTACCCCAGCTGGCCATAGCCTGGTGCCTGA
GGAATGAGGGTGTCAGCTCCGTGCTTCTGGGTGCTTCCAATGCAGAACAACTTATGGAGA

Can we gather hints of biological function from sequence?

Methods used to predict function from sequence

- Sequence homology

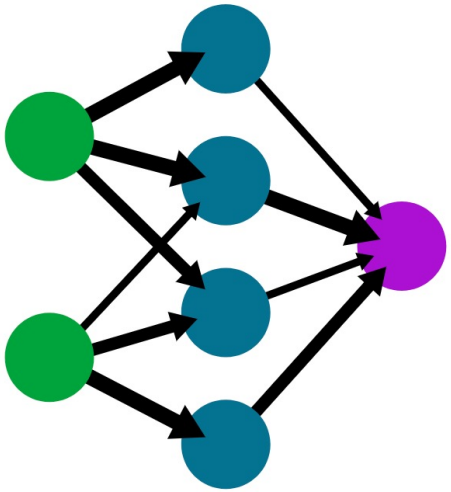
Searching protein database for sequence similarity

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

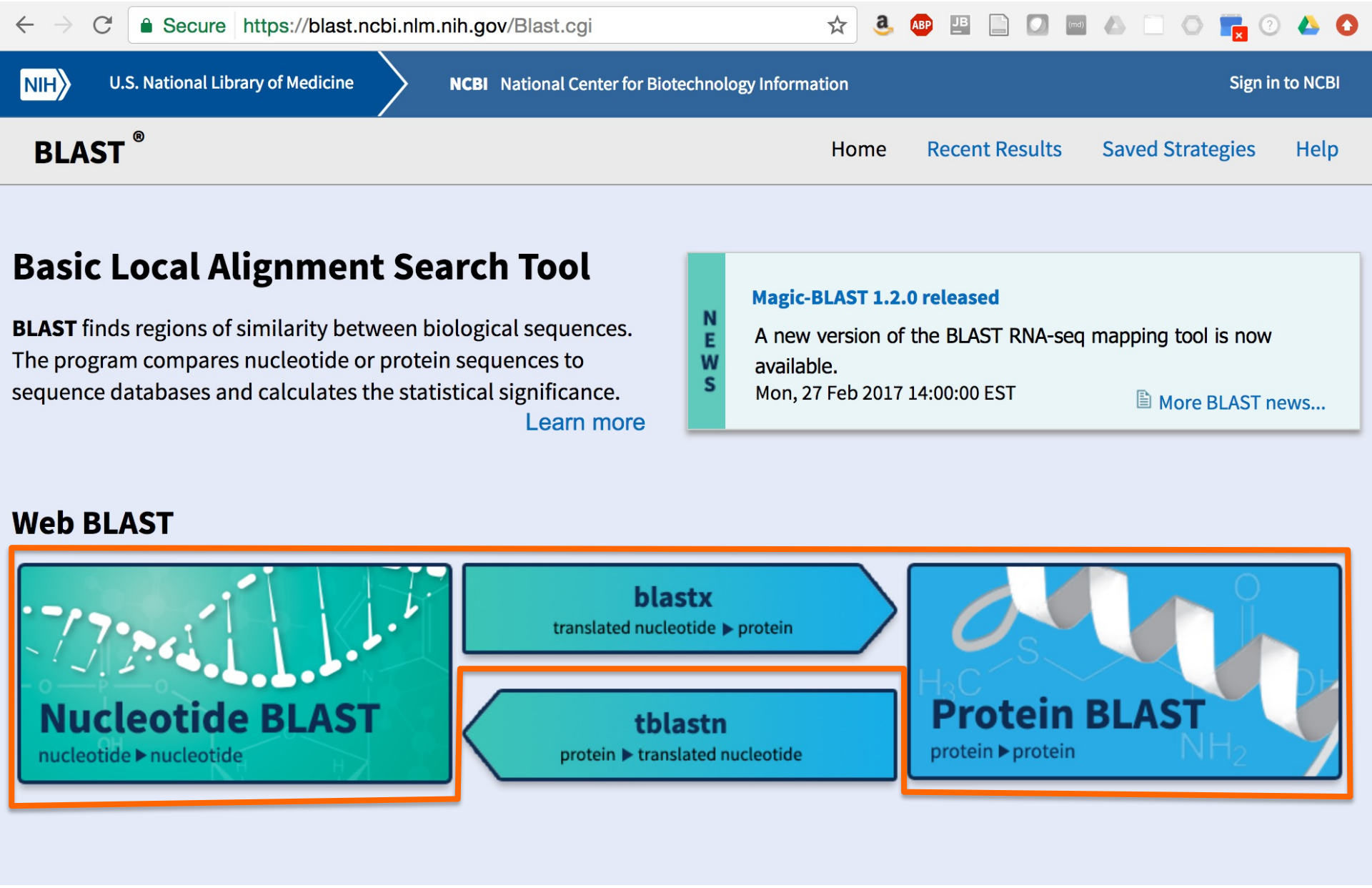
- Sequence composition

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

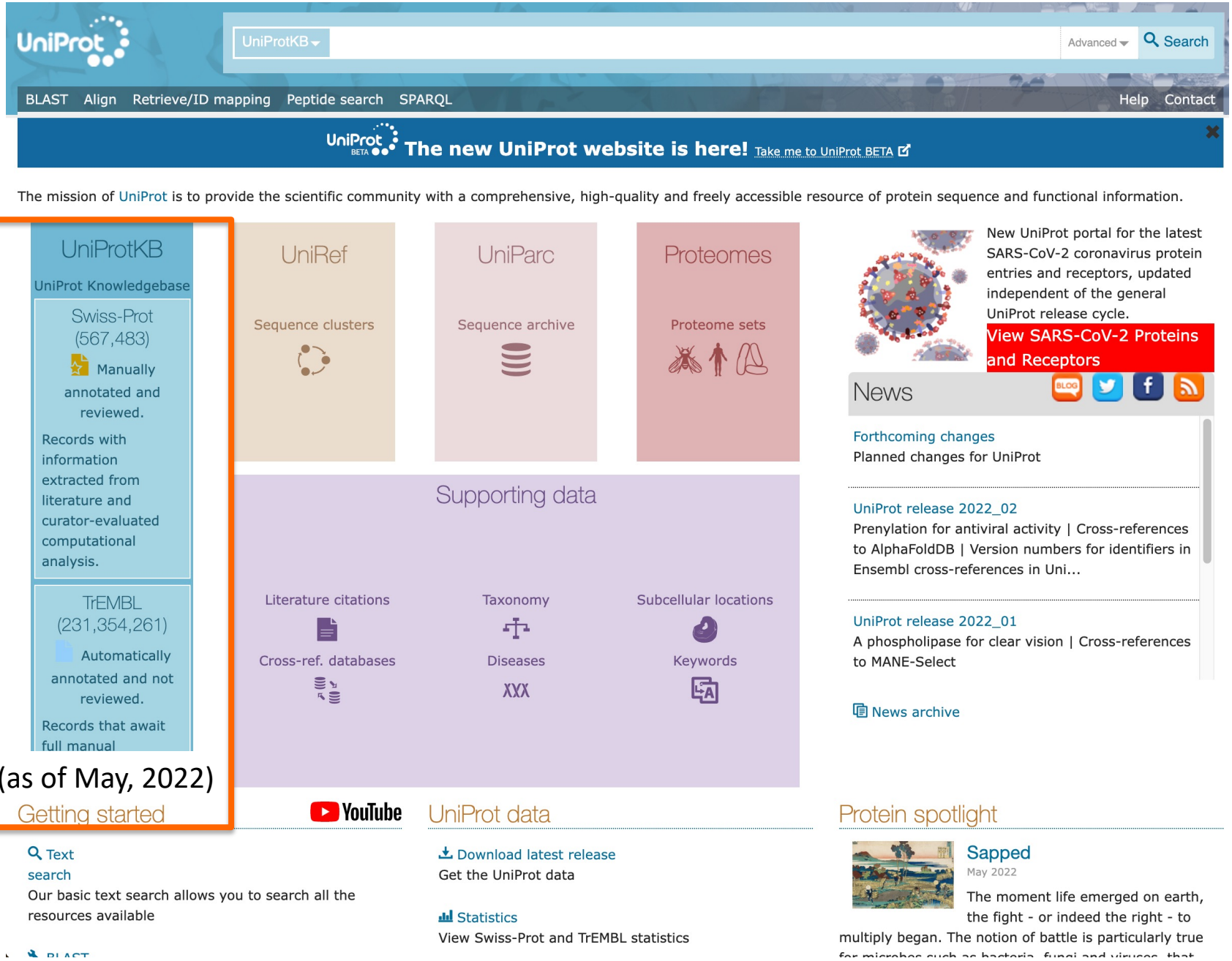
- Neural Networks
- Hidden Markov Models



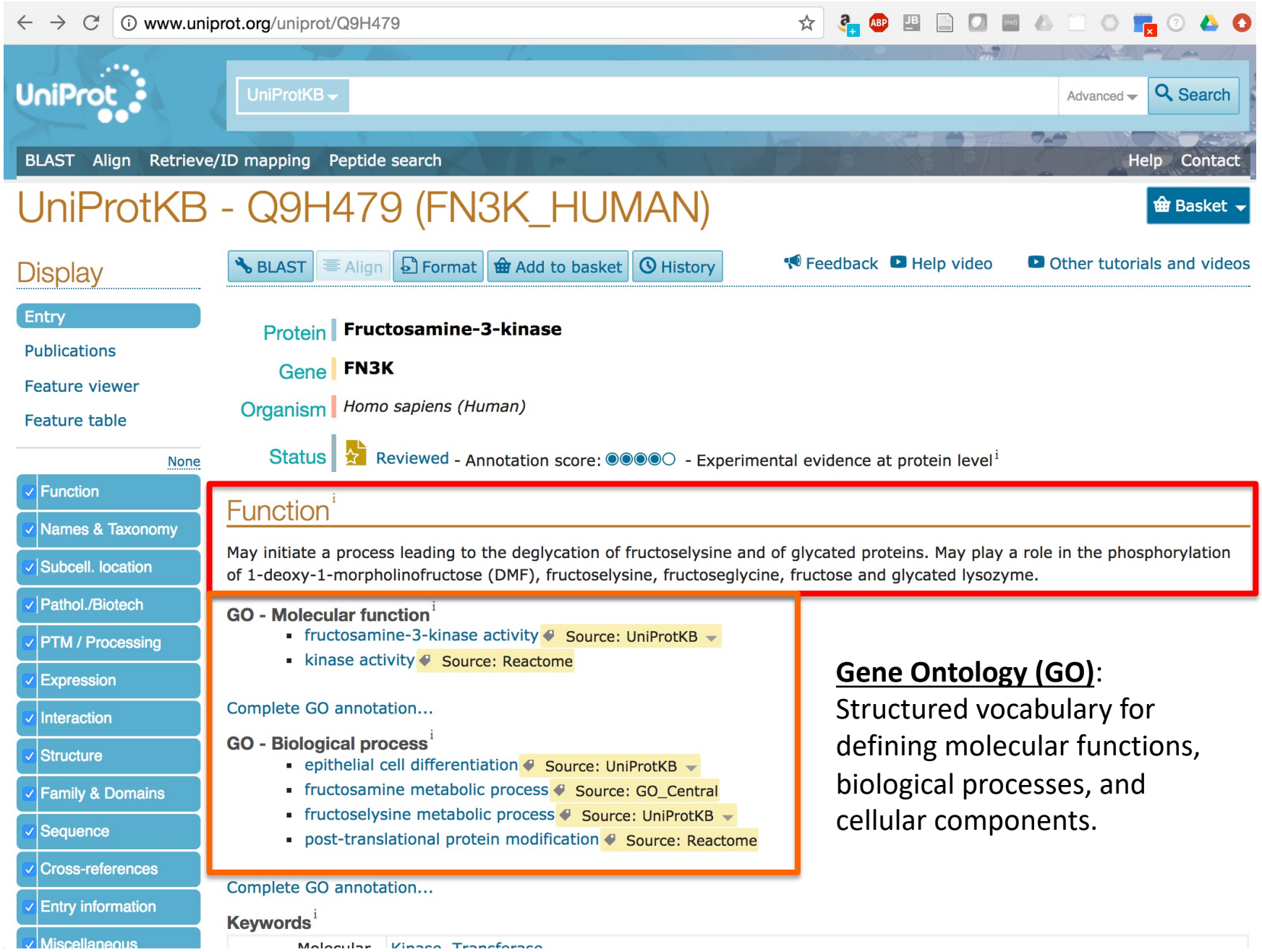
Use BLAST to search for sequence similarity to known proteins



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



Example of a Swiss-Prot Record



Gene Ontology (GO):
Structured vocabulary for defining molecular functions, biological processes, and cellular components.

No significant sequence similarity… What else?

GGAGCTGGAGGCCCCCAGGCAACTACACCGTCCACGTACCCAGAGGGGCTGGGCCCTCCC
ACCAGAGACCACGCCCTGGTGTGCCTTAGGGGCCCTGGTTTGTTAGTCTCTGAGTGTGCA
GTTGCTGCACATGGGGCCCTGGCGCTTGCTGCACCAACTTCCTGTTGGGCCCGTGGTCCT
TGGAGGCATGCAGTTCAGCAGACAGTGACTCAGCCATCCACCCAACATGCGGAACGTGTC
TCTTCTGCAGGTCCCGGTCCACAGCAGGATTCCCCCTCTGTGAAAAGGCACGCTGATCTG
TCTGGATAAGTGTGGCCGGCCCCATGTATCCGGAATCAACCACGGGGTCCCCAGCTCGAC
TCTCCCTGCGGCAGACAGGCTCCCCCGGGATGATCTACAGTACTCGTTATGGGAGTCCCA
AAAGACAGCTCCAGTTTTACAGGAATCTGGGCAAATCTGGCCTTCGGGTCTCCTGCCTGG
GGCTTGGAACATGGGTGACCTTCGGGGGCCAGATCACGGATGAGATGGCAGAGCACCTAA
TGACCTTGGCCTACGATAATGGCATCAACCTGTTCGATACGGCGGAGGTCTACGCTGCTG
GAAAAGCTGAAGTGGTATTAGGGAACATCATTAAGAAGAAGGGATGGAGACGGTCCAGCC
TTGTCATCACCACCAAGATCTTCTGGGGTGGAAAAGCGGAGACTGAGAGAGGCCTTTCCA
GGAAGCACATAATTGAAGGACTGAAAGCGTCCCTGGAGCGGCTGCAGCTGGAGTACGTGG
ATGTGGTTTTTGCCAACCGCCCAGACCCCAACACGCCCATGGAAGAGACCGTGCGGGCCA
TGACCCATGTCATCAACCAGGGGATGGCCATGTACTGGGGCACATCACGCTGGAGCTCCA
TGGAGATCATGGAGGCCTACTCGGTGGCTCGGCAGTTCAACCTGATCCCGCCCATCTGCG
AGCAAGCGGAATATCACATGTTCCAGAGGGAGAAGGTGGAGGTCCAGCTGCCAGAGCTGT
TCCACAAGATAGGAGTAGGTGCCATGACCTGGTCCCCTCTGGCGTGCGGCATCGTCTCAG
GGAAGTATGACAGCGGGATCCCACCCTACTCCAGAGCCTCCCTGAAGGGCTACCAGTGGT
TGAAGGACAAGATCCTGAGTGAGGAGGGTCGCCGCCAGCAGGCCAAGCTGAAGGAACTGC
AGGCCATTGCCGAACGCCTGGGCTGCACCCTACCCCAGCTGGCCATAGCCTGGTGCCTGA
GGAATGAGGGTGTCAGCTCCGTGCTTCTGGGTGCTTCCAATGCAGAACAACTTATGGAGA

Is there an ORF for a potential Coding Region?

GGAGCTGGAGGCCCCCAGGCAACTACACCGTCCACGTACCCAGAGGGGCTGGGCCCTCCC
ACCAGAGACCACGCCCTGGTGTGCCTTAGGGGCCCTGGTTTGTTAGTCTCTGAGTGTGCA
GTTGCTGCACATGGGGCCCTGGCGCTTGCTGCACCAACTTCCTGTTGGGCCCGTGGTCCT
TGGAGGCATGCAGTTCAGCAGACAGTGACTCAGCCATCCACCCAACATGCGGAACGTGTC
TCTTCTGCAGGTCCCGGTCCACAGCAGGATTCCCCCTCTGTGAAAAGGCACGCTGATCTG
TCTGGATAAGTGTGGCCGGCCCCATGTATCCGGAATCAACCACGGGGTCCCCAGCTCGAC
TCTCCCTGCGGCAGACAGGCTCCCCCGGGATGATCTACAGTACTCGTTATGGGAGTCCCA
AAAGACAGCTCCAGTTTTACAGGAATCTGGGCAAATCTGGCCTTCGGGTCTCCTGCCTGG
GGCTTGGAACATGGGTGACCTTCGGGGGCCAGATCACGGATGAGATGGCAGAGCACCTAA
TGACCTTGGCCTACGATAATGGCATCAACCTGTTCGATACGGCGGAGGTCTACGCTGCTG
GAAAAGCTGAAGTGGTATTAGGGAACATCATTAAGAAGAAGGGATGGAGACGGTCCAGCC
TTGTCATCACCACCAAGATCTTCTGGGGTGGAAAAGCGGAGACTGAGAGAGGCCTTTCCA
GGAAGCACATAATTGAAGGACTGAAAGCGTCCCTGGAGCGGCTGCAGCTGGAGTACGTGG
ATGTGGTTTTTGCCAACCGCCCAGACCCCAACACGCCCATGGAAGAGACCGTGCGGGCCA
TGACCCATGTCATCAACCAGGGGATGGCCATGTACTGGGGCACATCACGCTGGAGCTCCA
TGGAGATCATGGAGGCCTACTCGGTGGCTCGGCAGTTCAACCTGATCCCGCCCATCTGCG
AGCAAGCGGAATATCACATGTTCCAGAGGGAGAAGGTGGAGGTCCAGCTGCCAGAGCTGT
TCCACAAGATAGGAGTAGGTGCCATGACCTGGTCCCCTCTGGCGTGCGGCATCGTCTCAG
GGAAGTATGACAGCGGGATCCCACCCTACTCCAGAGCCTCCCTGAAGGGCTACCAGTGGT
TGAAGGACAAGATCCTGAGTGAGGAGGGTCGCCGCCAGCAGGCCAAGCTGAAGGAACTGC
AGGCCATTGCCGAACGCCTGGGCTGCACCCTACCCCAGCTGGCCATAGCCTGGTGCCTGA
GGAATGAGGGTGTCAGCTCCGTGCTTCTGGGTGCTTCCAATGCAGAACAACTTATGGAGA

Is there an ORF for a potential Coding Region?

GGAGCTGGAGGCCCCCAGGCAACTACACCGTCCACGTACCCAGAGGGGCTGGGCCCTCCC
ACCAGAGACCACGCCCTGGTGTGCCTTAGGGGCCCTGGTTTGTTAGTCTCTGAGTGTGCA
GTTGCTGCAC**ATGGGGCCCTGGCGCTTGCTGCACCAACTTCCTGTTGGGCCCGTGGTCCT**
TGGAGGCATGCAGTTCAGCAGACAGTGACTCAGCCATCCACCCAACATGCGGAACGTGTC
TCTTCTGCAGGTCCCGGTCCACAGCAGGATTCCCCCTCTGTGAAAAGGCACGCTGATCTG
TCTGGATAAGTGTGGCCGGCCCCATGTATCCGGAATCAACCACGGGGTCCCCAGCTCGAC
TCTCCCTGCGGCAGACAGGCTCCCCCGGGATGATCTACAGTACTCGTTATGGGAGTCCCA
AAAGACAGCTCCAGTTTTACAGGAATCTGGGCAAATCTGGCCTTCGGGTCTCCTGCCTGG
GGCTTGGAACATGGGTGACCTTCGGGGGCCAGATCACGGATGAGATGGCAGAGCACCTAA
TGACCTTGGCCTACGATAATGGCATCAACCTGTTCGATACGGCGGAGGTCTACGCTGCTG
GAAAAGCTGAAGTGGTATTAGGGAACATCATTAAGAAGAAGGGATGGAGACGGTCCAGCC
TTGTCATCACCACCAAGATCTTCTGGGGTGGAAAAGCGGAGACTGAGAGAGGCCTTTCCA
GGAAGCACATAATTGAAGGACTGAAAGCGTCCCTGGAGCGGCTGCAGCTGGAGTACGTGG
ATGTGGTTTTTGCCAACCGCCCAGACCCCAACACGCCCATGGAAGAGACCGTGCGGGCCA
TGACCCATGTCATCAACCAGGGGATGGCCATGTACTGGGGCACATCACGCTGGAGCTCCA
TGGAGATCATGGAGGCCTACTCGGTGGCTCGGCAGTTCAACCTGATCCCGCCCATCTGCG
AGCAAGCGGAATATCACATGTTCCAGAGGGAGAAGGTGGAGGTCCAGCTGCCAGAGCTGT
TCCACAAGATAGGAGTAGGTGCCATGACCTGGTCCCCTCTGGCGTGCGGCATCGTCTCAG
GGAAGTATGACAGCGGGATCCCACCCTACTCCAGAGCCTCCCTGAAGGGCTACCAGTGGT
TGAAGGACAAGATCCTGAGTGAGGAGGGTCGCCGCCAGCAGGCCAAGCTGAAGGAACTGC
AGGCCATTGCCGAACGCCTGGGCTGCACCCTACCCCAGCTGGCCATAGCCTGGTGCCTGA
GGAATGAGGGTGTCAGCTCCGTGCTTCTGGGTGCTTCCAATGCAGAACAACTTATGGAGA

Find all ORFs using ORFfinder

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

NCBI Resources How To Sign in to NCBI

ORFfinder PubMed Search

Open Reading Frame Finder

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.

This web version of the ORF finder is limited to the subrange of the query sequence up to 50 kb long. Stand-alone version, which doesn't have query sequence length limitation, is available for Linux x64.

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

- NC_011604 Salmonella enterica plasmid pWES-1; genetic code: 11; 'ATG' and alternative initiation codons; minimal ORF length: 300 nt
- NM_000059; genetic code: 1; start codon: 'ATG only'; minimal ORF length: 150 nt

Enter Query Sequence

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

```
GGAGCTGGAGGCCCCCAGGCAACTACACCGTCCACGTACCCAGAGGGGCTGGGCCCTCCC
ACCAGAGACCACGCCCTGGTGTGCCTTAGGGGCCCTGGTTTGTTAGTCTCTGAGTGTGCA
GTTGCTGCACATGGGGCCCTGGCGCTTGCTGCACCAACTTCCTGTTGGGCCCGTGGTCCT
TGGAGGCATGCAGTTCAGCAGACAGTGACTCAGCCATCCACCCAACATGCGGAACGTGTC
TCTTCTGCAGGTCCCGGTCCACAGCAGGATTCCCCCTCTGTGAAAAGGCACGCTGATCTG
TCTGGATAAGTGTGGCCGGCCCCATGTATCCGGAATCAACCACGGGGTCCCCAGCTCGAC
TCTCCCTGCGGCAGACAGGCTCCCCCGGGATGATCTACAGTACTCGTTATGGGAGTCCCA
AAAGACAGCTCCAGTTTTACAGGAATCTGGGCAAATCTGGCCTTCGGGTCTCCTGCCTGG
GGCTTGGAACATGGGTGACCTTCGGGGGCCAGATCACGGATGAGATGGCAGAGCACCTAA
TGACCTTGGCCTACGATAATGGCATCAACCTGTTCGATACGGCGGAGGTCTACGCTGCTG
```

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

ORFs found: 12 Genetic code: 1 Start codon: 'ATG' only

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

Add six-frame translation track

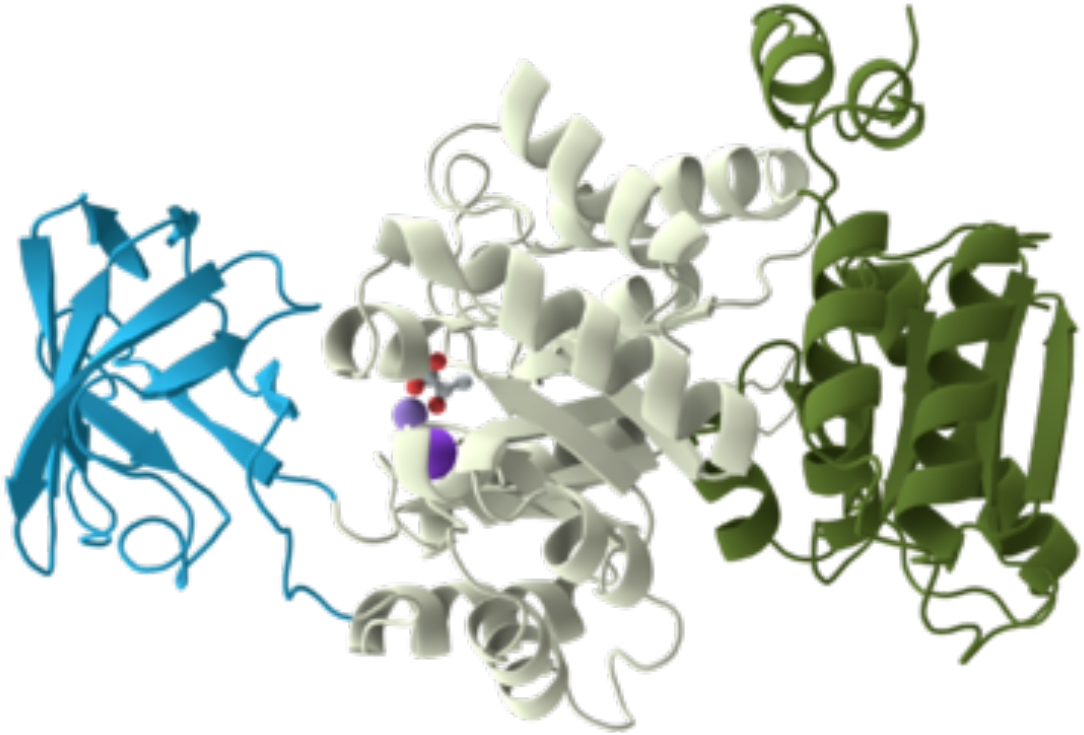
ORF5 (367 aa) Display ORF as... Mark

```
>lcl|ORF5
MYPESTTGSPARLSLRQTGSPGMIYSTRYGSPKRQLQFYR
NLGKSGLRVSCLGLGTWVTFGGQITDEMAEHLMTLAYDNG
INLFDTAEVYAAGKAEVVLGNIIKKKGWRRSSLVITTKIF
WGGKAETERGLSRKHIIEGLKASLERLQLEYVDVVFANRP
DPNTPMEETVRAMTHVINQGMAMYWGTSRWSSMEIMEAYS
VARQFNLIPPICEQAEYHMFQREKVEVQLPELFHKIGVGA
MTWSPLACGIVSGKYDSGIPPYSRASLKGYQWLKDKILSE
EGRRQQAKLKELQAIAERLGCTLPQLAIAWCLRNEGVSSV
LLGASNAEQLMENIGAIQVLPKLSSSIVHEIDSILGNKPY
SKKDYRS
```

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?

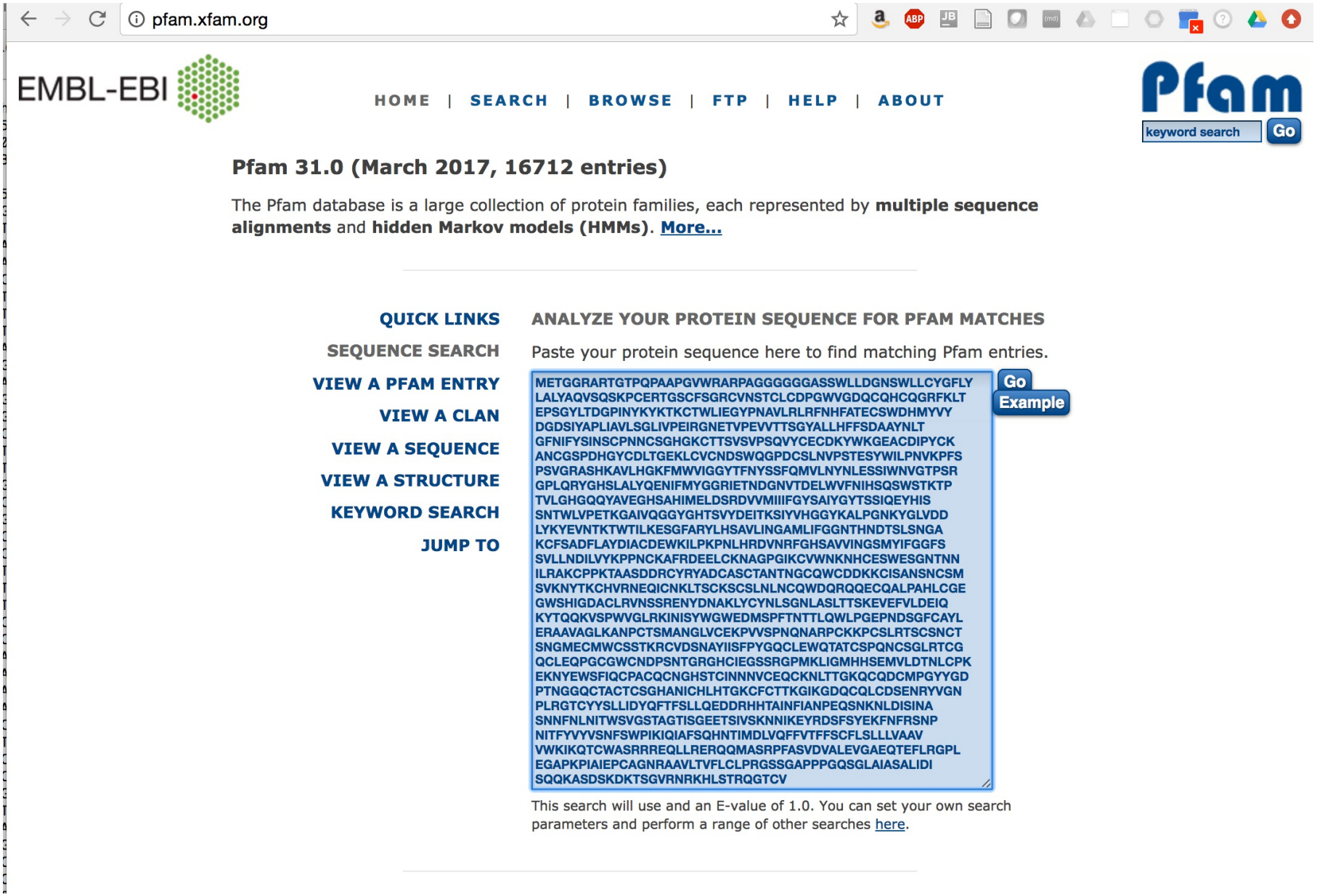


Hints at substrate binding or catalytic activity

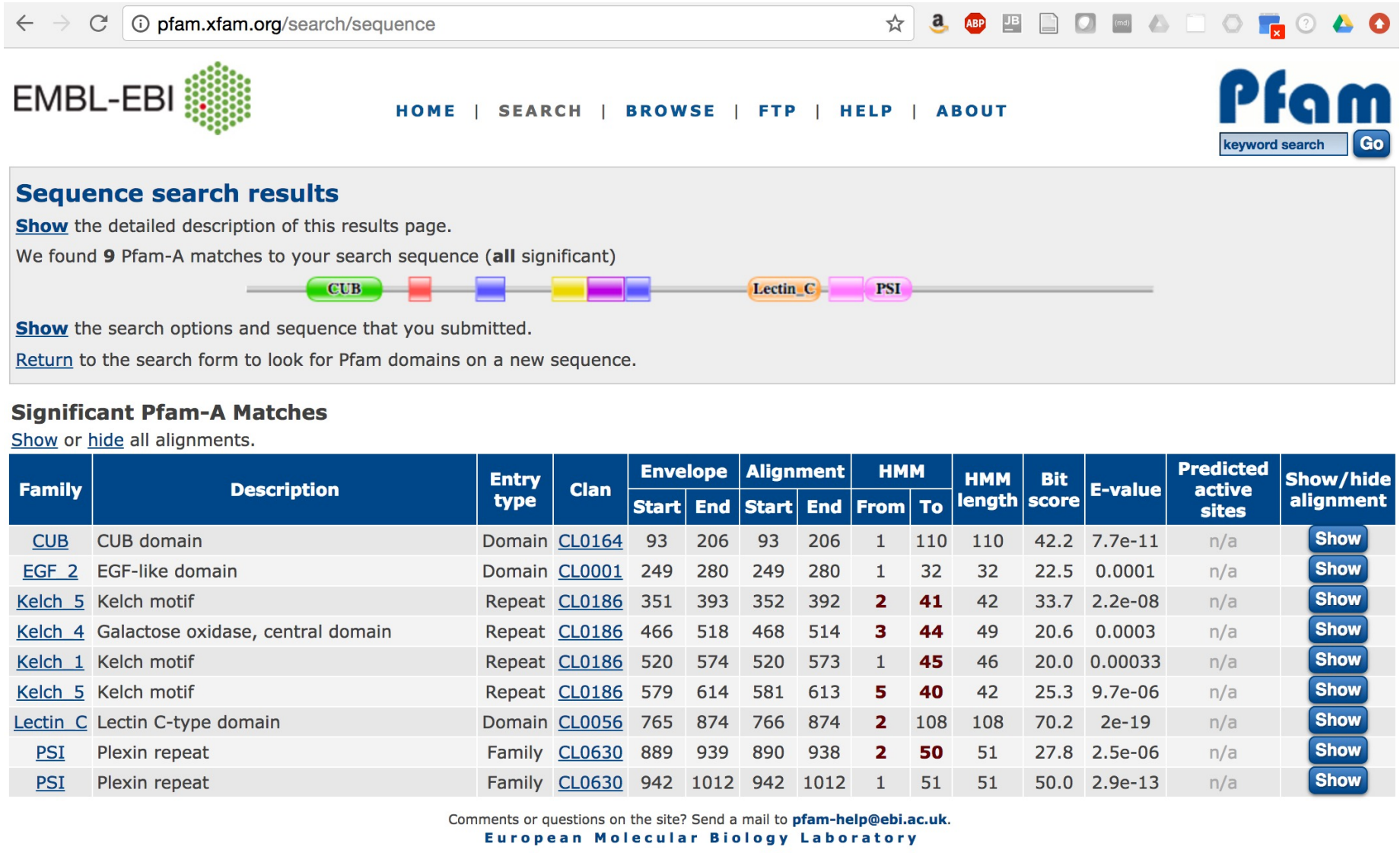
DNA, RNA, calcium, phoshate, etc.

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

Search the Pfam library of HMMs to identify potential functional domains



Example Pfam report illustrating modular domain architecture



pfam.xfam.org/search/sequence

EMBL-EBI

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Pfam

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

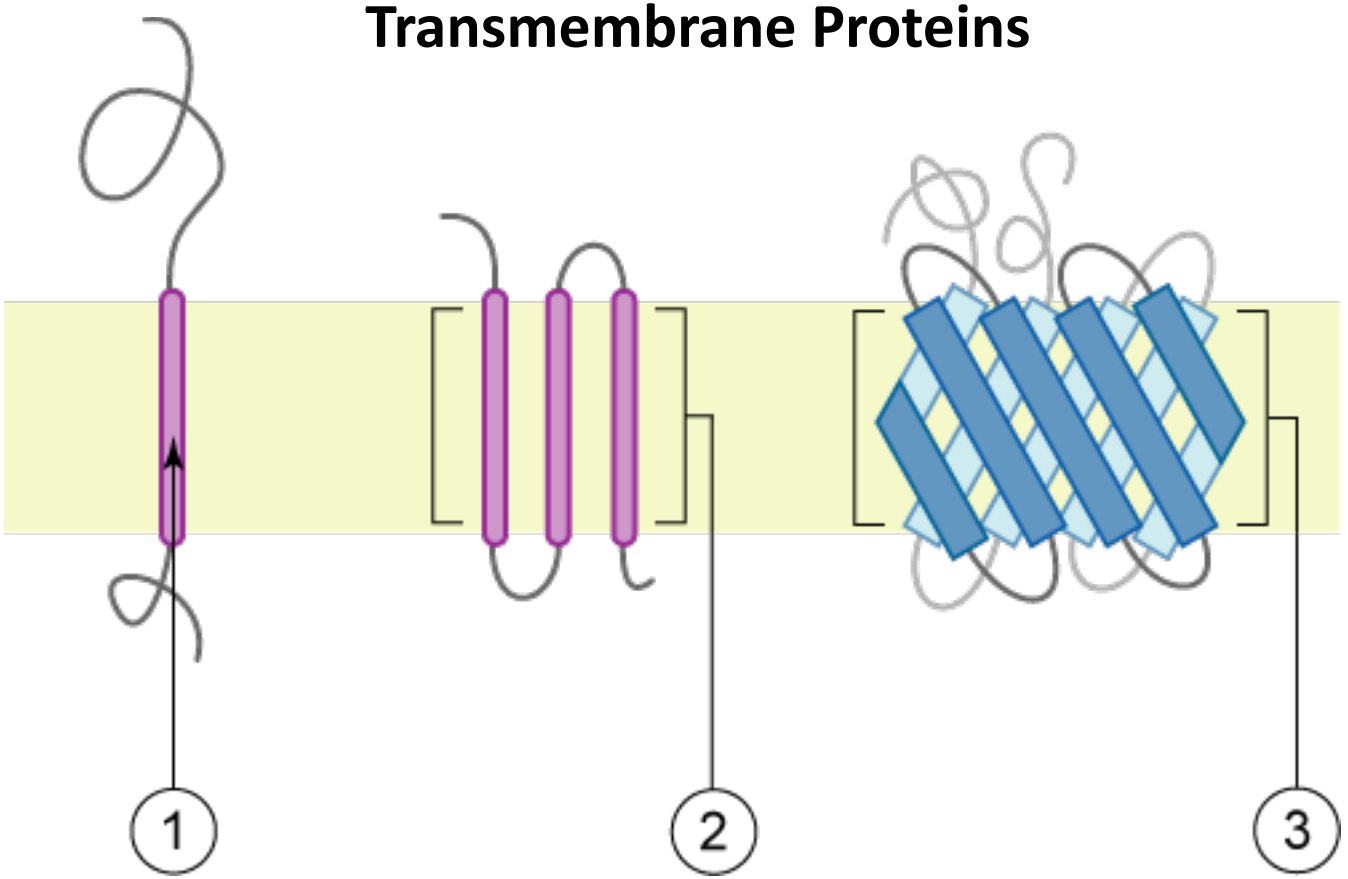
Show or hide all alignments.

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



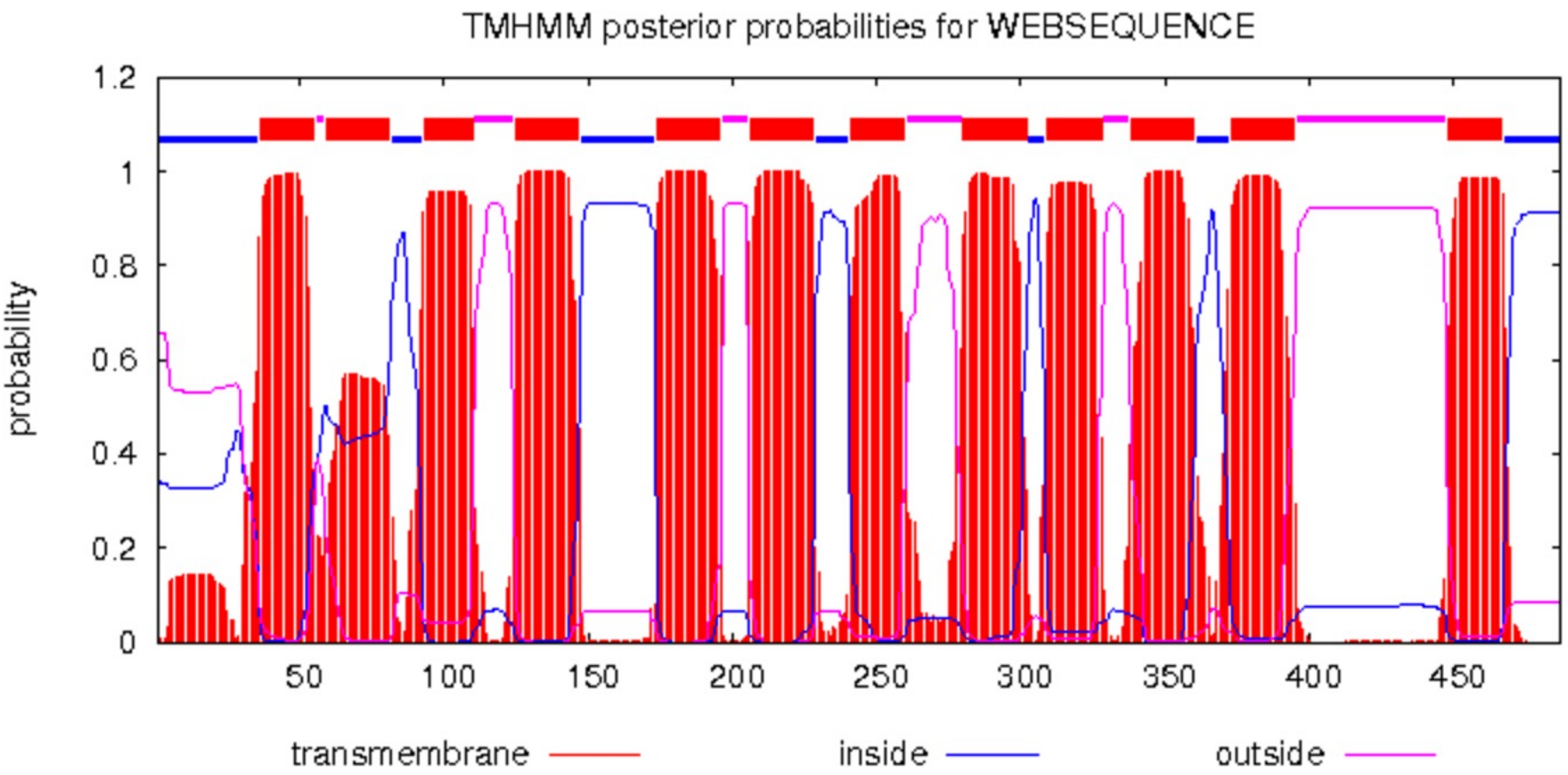
Single transmembrane α-helix (bitopic membrane protein)

Polytopic transmembrane α-helical protein

Polytopic transmembrane β-sheet protein

From: https://en.wikipedia.org/wiki/Transmembrane_protein

Trans-membrane Domains via TmHMM

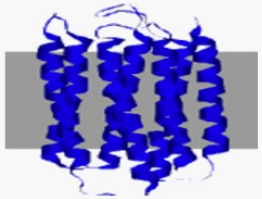


Topology=i36-55o59-81i93-110o125-147i174-196o206-228i241-260o280-302i309-328o338-360i373-395o448-467i

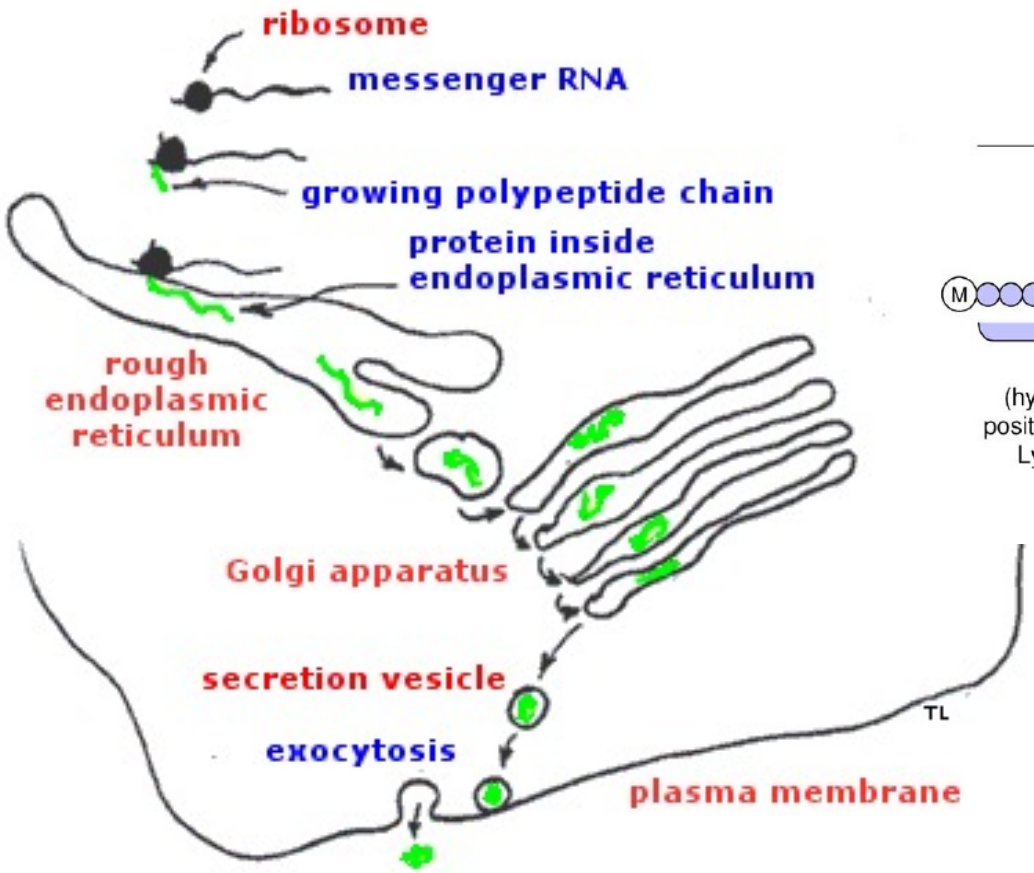
TMHMM Server v. 2.0

Prediction of transmembrane helices in proteins

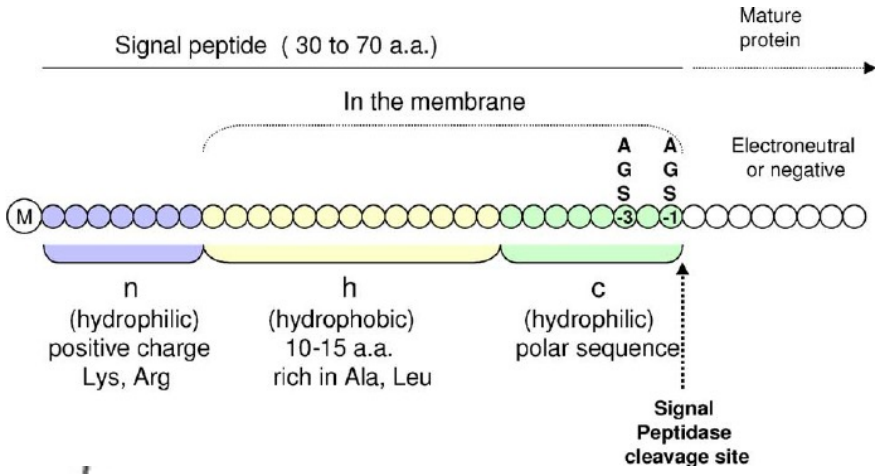
http://www.cbs.dtu.dk/services/TMHMM/



Predicting Secreted Proteins



(from: https://courses.washington.edu/conj/cell/secretion.htm)



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

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

SignalP 4.1 Server - prediction results

Technical University of Denmark

Transcriptome-scale functional annotation using Trinotate

Trinotate: Transcriptome Functional Annotation and Analysis

Trinotate

TransDecoder

Trinity

Inchworm

Chrysalis

Butterfly

TMHMM

SignalP

NCBI BLAST

HMMER

Pfam

UniProt

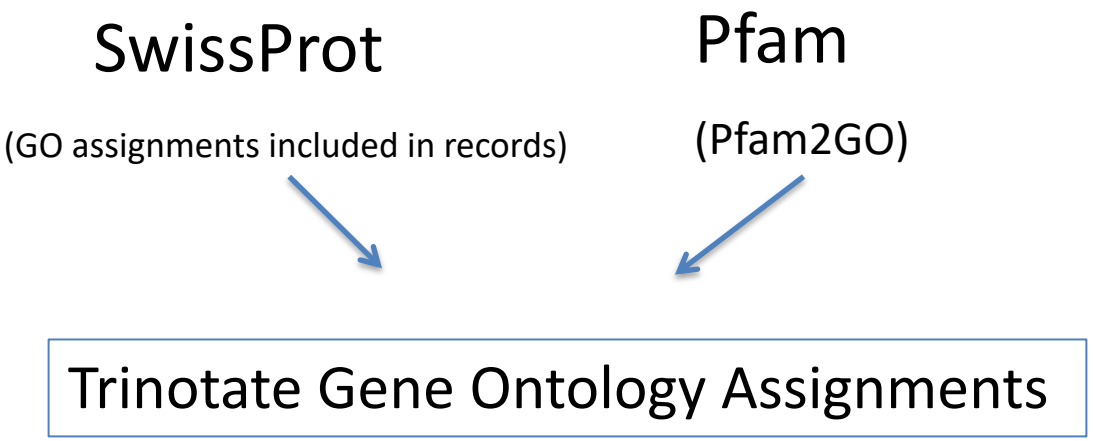
eggNOG version 3.0

SQLite

the Gene Ontology

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

Gene ontology functional enrichment

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

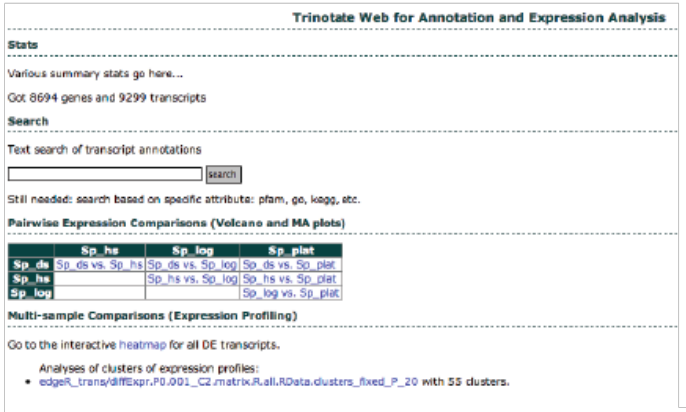
	drawn	not drawn	total
green marbles	k	$K - k$	K
red marbles	$n - k$	$N + k - n - K$	$N - K$
total	n	$N - n$	N

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

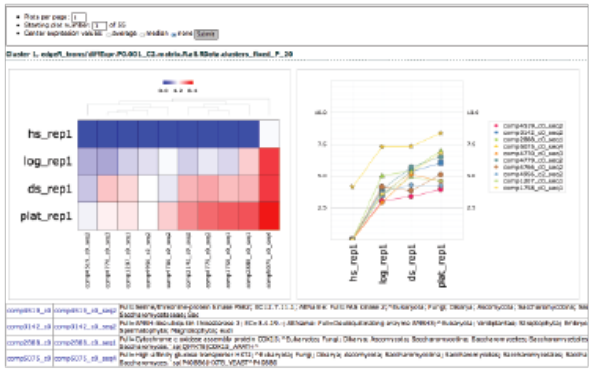
$$P(X = k) = f(k; N, K, n) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}.$$

Trinotate Web for Interactive Analysis

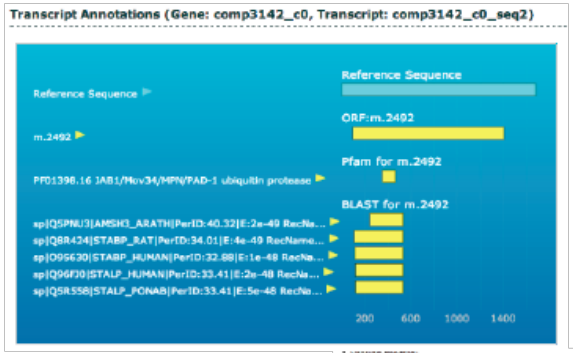
TrinotateWeb Entry Point

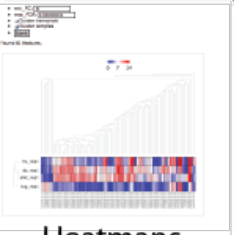


Clustered Expression Profiles

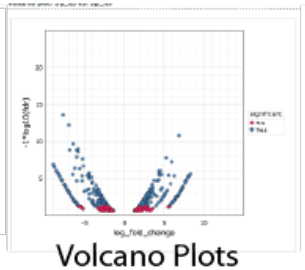


Transcript/Protein Annotation Report
Blast Hits, Pfam Domains, etc.

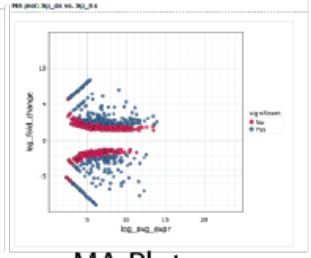




Heatmaps

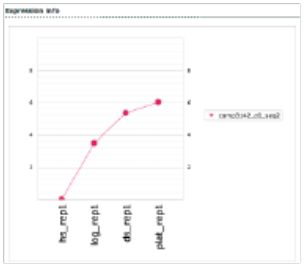


Volcano Plots



MA-Plots

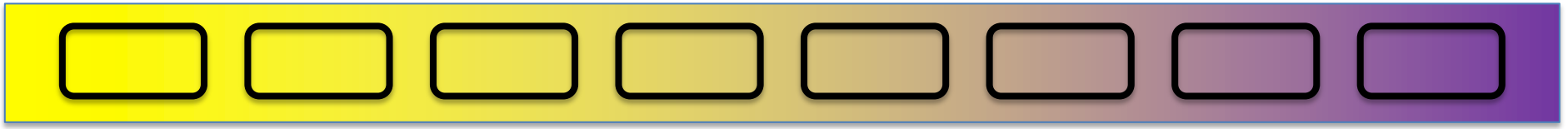
Very Early Release and
Just Scratching the Surface



Individual Transcript
Expression Profiles

Transcript and
Protein Sequences

Part 8. Case study: salamander transcriptome



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

Work done in collaboration with Jessica Whited's lab

BWH

BRMC

Brigham Regenerative Medicine Center

Axolotl (*Ambystoma mexicanum*) Transcriptomics

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

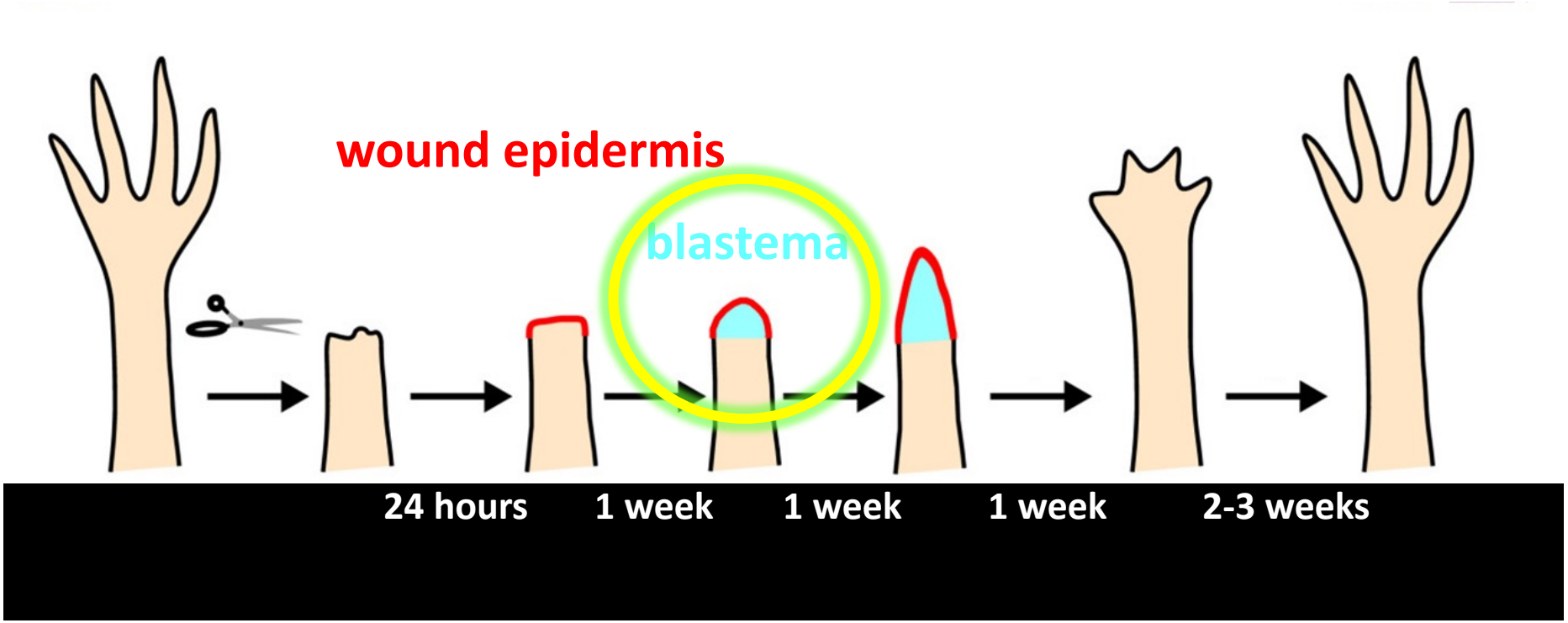
- Model for vertebrate studies of tissue regeneration
- Short generation time
- Can fully regenerate a severed limb in just weeks.
- Genome estimated at ~30 Gb (not yet sequenced)

Google Anonymous Axolotl Icon

Lovable Pets, Too!

Rayan Chikhi's
pet axolotls

Key morphological steps during limb regeneration



BWH BRMC Brigham Regenerative Medicine Center

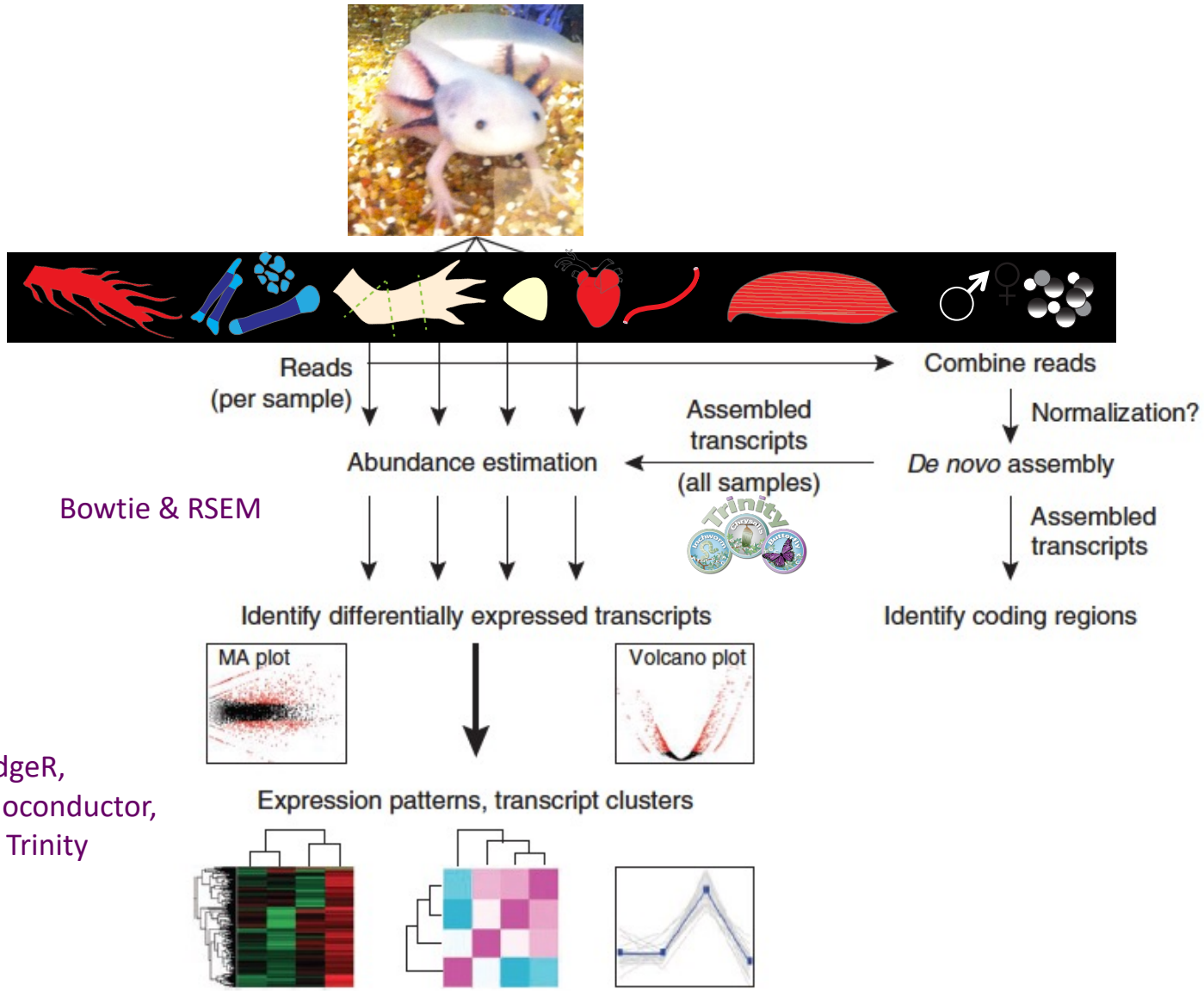
Jessica Whited, Mark Mannucci, Ari Haberberg

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



Bowtie & RSEM

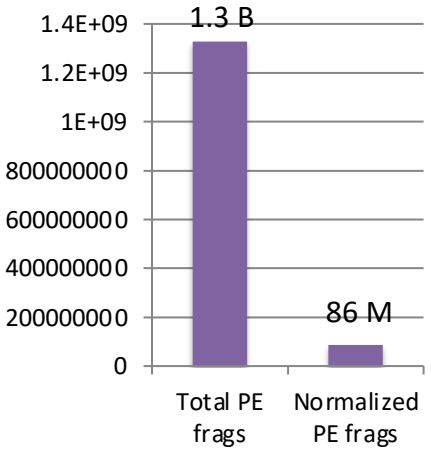
EdgeR,
Bioconductor,
& Trinity

1.3 Billion
Total Reads

86 Million
Normalized Reads

Axolotl Transcriptome De novo Assembly Statistics And Quality Assessment

In silico Normalization

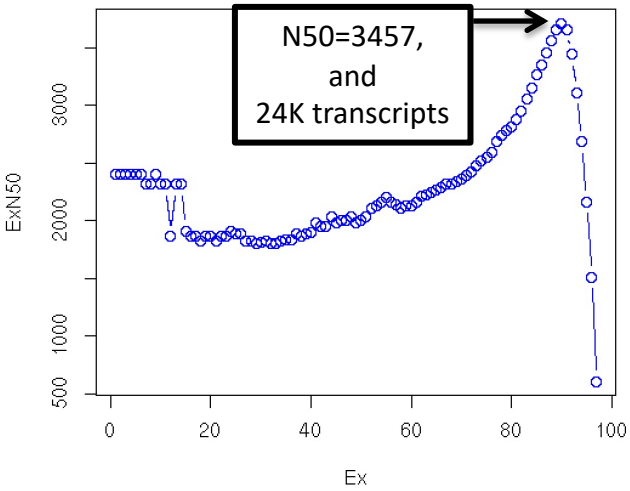


Counts of Transcripts

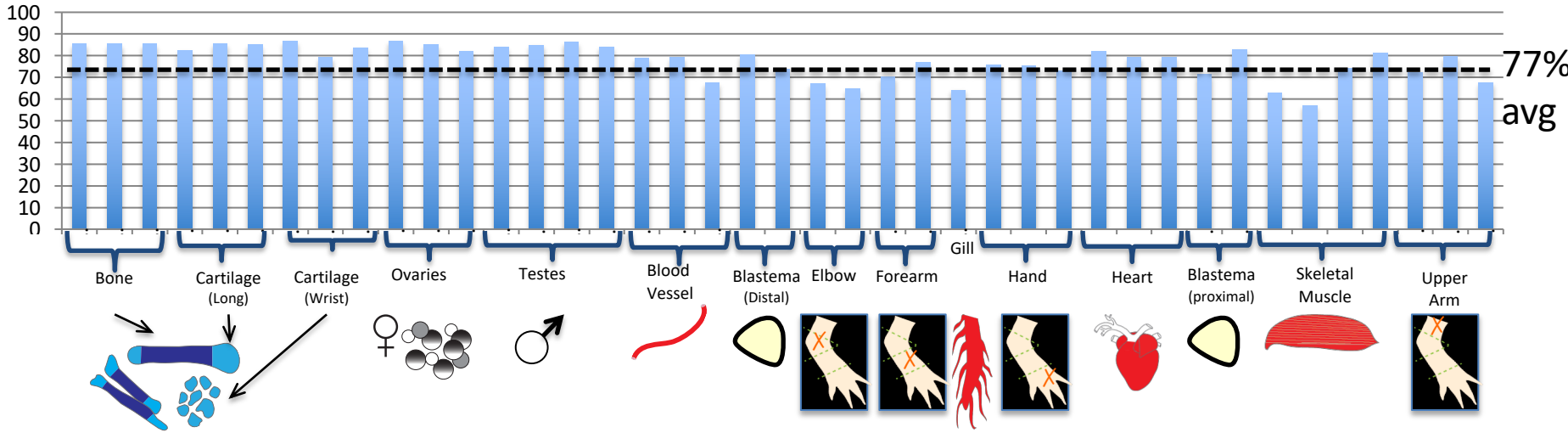
Trinity contigs (transcripts)	**1,554,055**
Trinity components (genes)	**1,388,798**

Min. length 200 bases

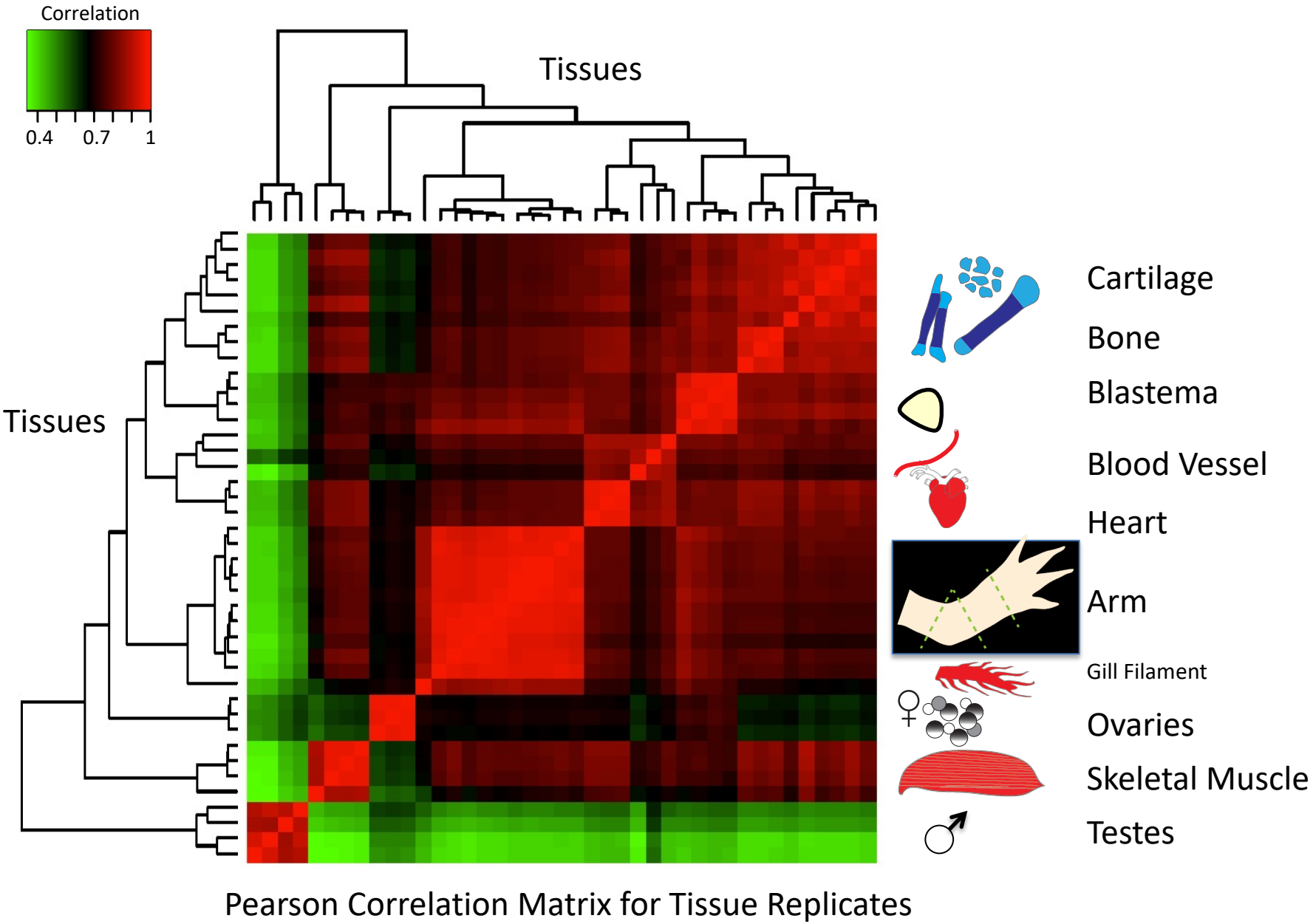
ExN50 looks good!



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

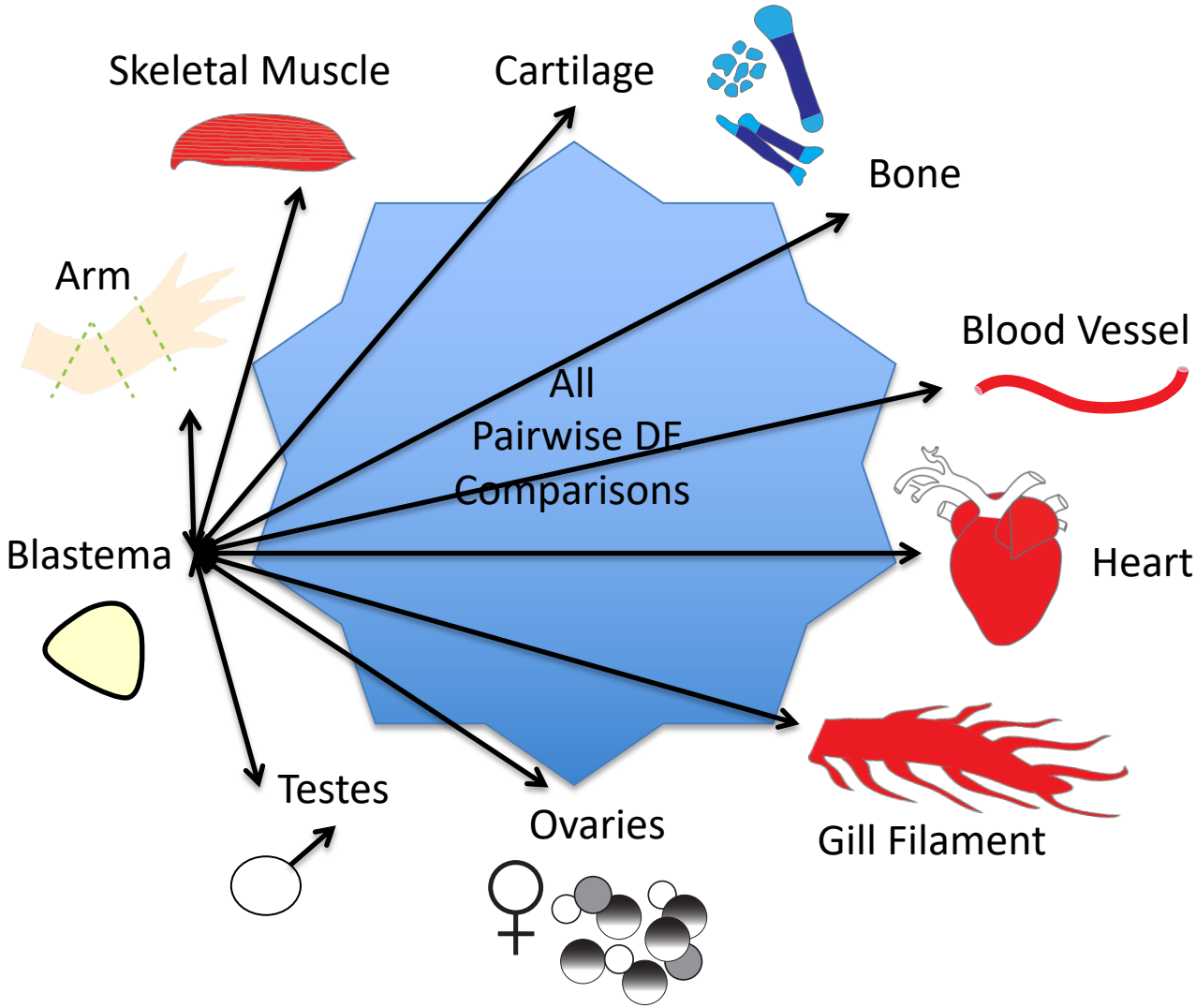


Biological Replicates Cluster According to Sample



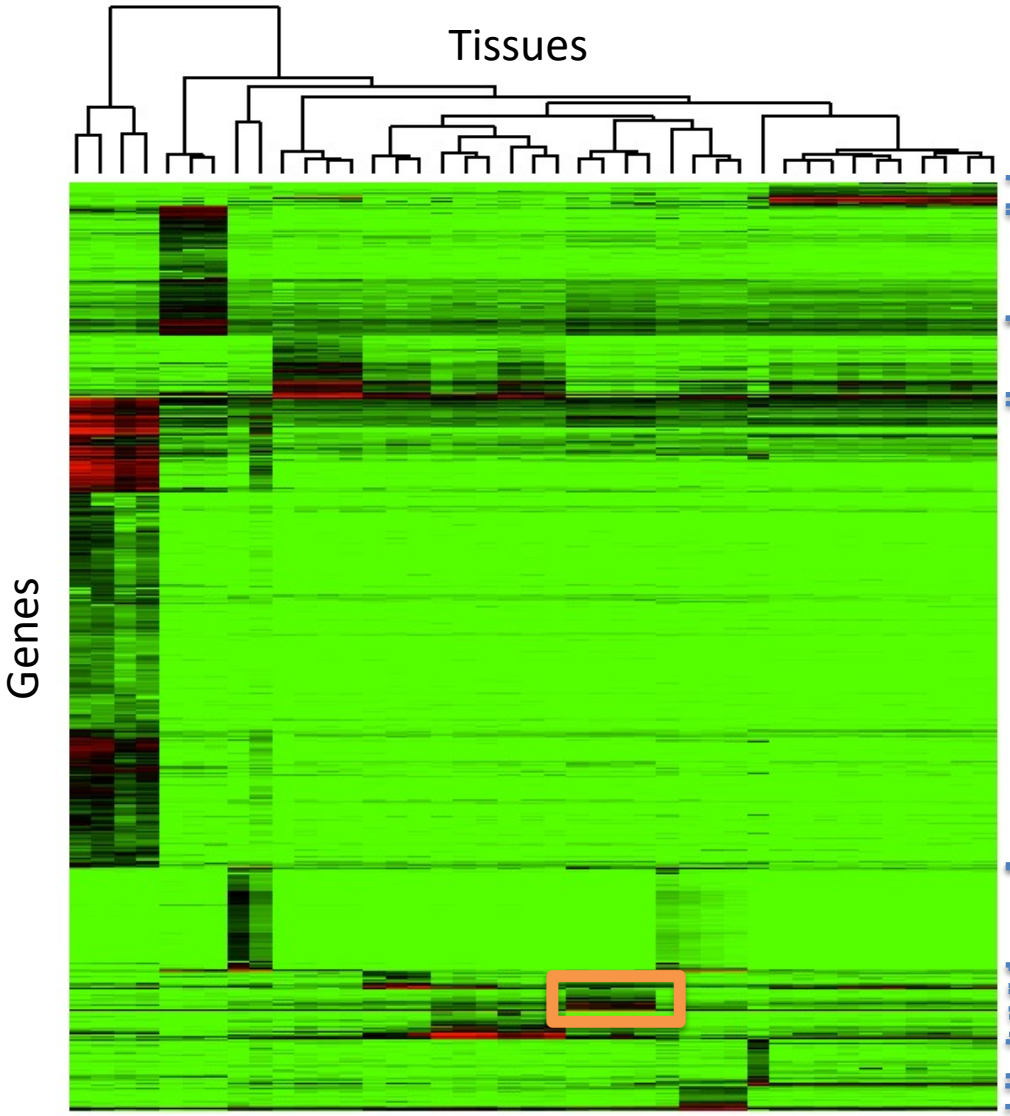
Pearson Correlation Matrix for Tissue Replicates

2. Identification of Tissue-enriched Expression



EdgeR, min 4-fold change, FDR <= 1e-3

Identification of Tissue-enriched Gene Expression



Arm (193), GO: thick ascending limb development [8.8e-5]

Ovaries (1225)

Skeletal Muscle (539)

Testes (4113), GO: spermatogenesis [2.5e-14]

Blood Vessel (939)

Bone (272), GO: myeloid leukocyte differentiation [2.2e-3]

Blastema (202): limb morphogenesis [2.5e-5]

Cartilage (255), GO: collagen fibril organization [4.5e-10]

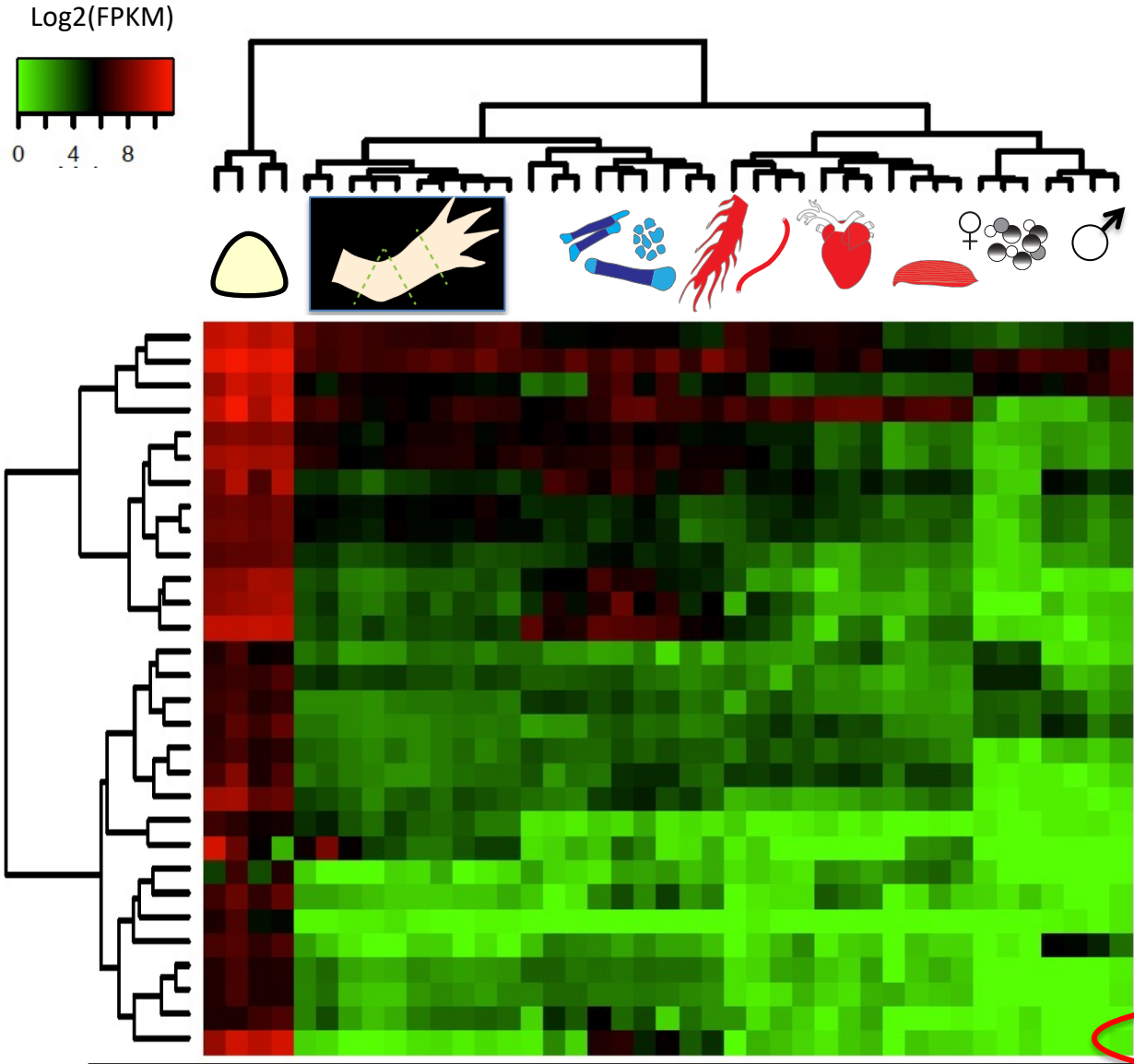
Gill Filament (765)

Heart (238),
GO: vascular process in circulatory system [2.6e-3]

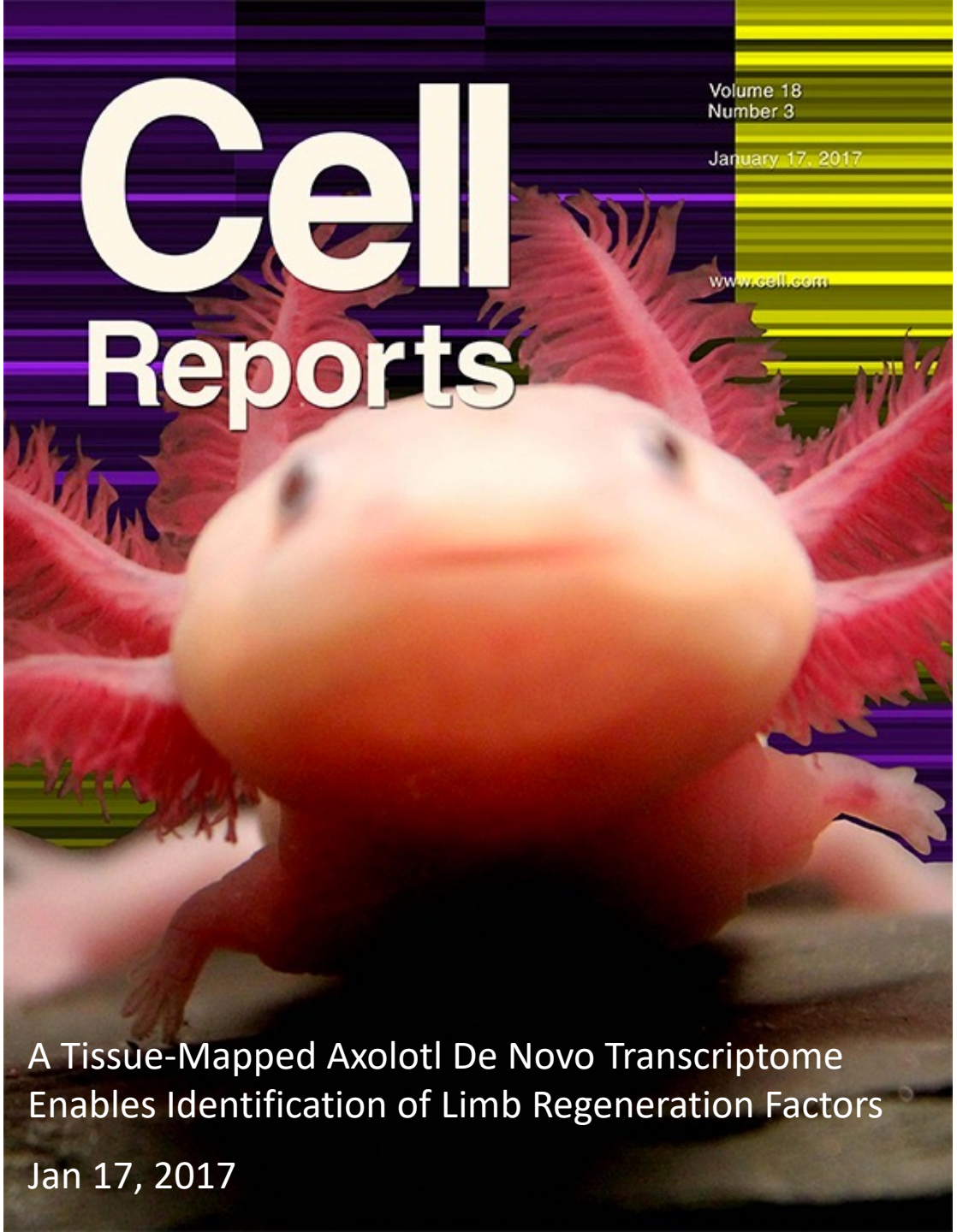
EdgeR, min 4-fold change, FDR <= 1e-3

Functional enrichment using GO-Seq

Most Highly Expressed Blastema-enriched Genes



Color key: Regulator Signaling Structure and Extracellular Matrix



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

Jan 17, 2017

Additional use case for Trinity De novo Assembly with Cancer Transcriptomes: Reconstruct Viruses Evident in Cancer RNA-Seq Samples

Among the 60 cancer cell lines, we discovered:

- Murine type-C retrovirus
- Xenotropic murine leukemia virus
- Squirrel monkey retrovirus
- Bovine polyomavirus

Haas et al. 2019, Genome Biology

nature

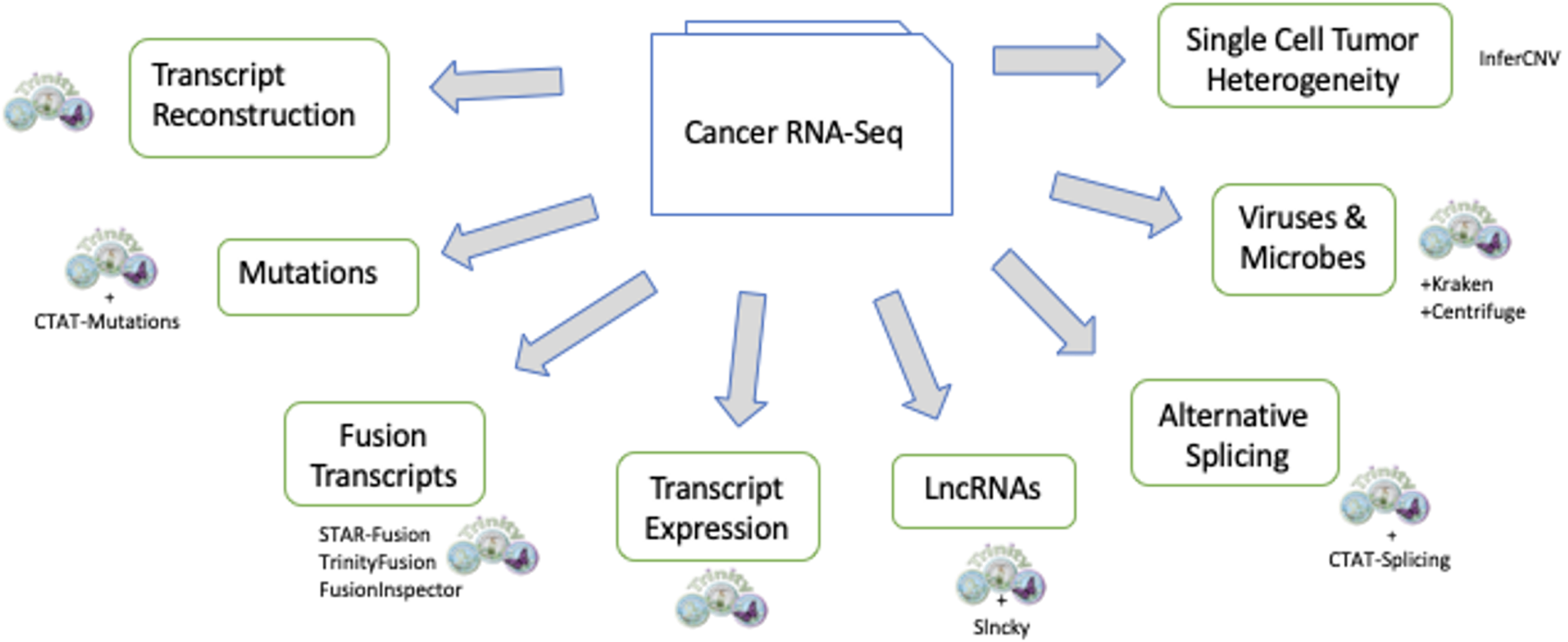
A new coronavirus associated with human respiratory disease in China

Fan Wu, Su Zhao, Bin Yu, Yan-Mei Chen, Wen Wang, Zhi-Gang Song, Yi Hu, Zhao-Wu Tao, Jun-Hua Tian, Yuan-Yuan Pei, Ming-Li Yuan, Yu-Ling Zhang, Fa-Hui Dai, Yi Liu, Qi-Min Wang, Jiao-Jiao Zheng, Lin Xu, Edward C. Holmes & Yong-Zhen Zhang

Nature (2020) | Cite this article

Virus genome discovery aided by Trinity de novo assembly of **RNA-seq from bronchoalveolar lavage fluid** from one of those initially infected in the **seafood market in Wuhan**, Dec. 2019.

Cancer Transcriptome Analysis Toolkit



https://github.com/NCIP/Trinity_CTAT

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 – need deeper sequencing and more replicates.
- 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).
- Leverage longer reads (less guessing about full-length isoform structures)
  (eg. PacBio, ONT)
- Single cell and spatial transcriptomics pushing the envelope.

Acknowledgements

Current and Former Trinity Contributors

Aviv Regev
* Brian Haas
Moran Yassour
Manfred Grabherr
Tim Tickle
Asma Bankapur
Christophe Georgescu
Vrushali Fangal
Maxwell Brown
Joshua Gould

Trinotate & TrinoateWeb
Brian Couger
Leonardo Gonzalez
Garima Lohani

Salamander Transcriptomics
Jessica Whited
Nick Leigh
Donald Bryant
Steven Blair

Trinity is funded by:

Transcriptomics Lab

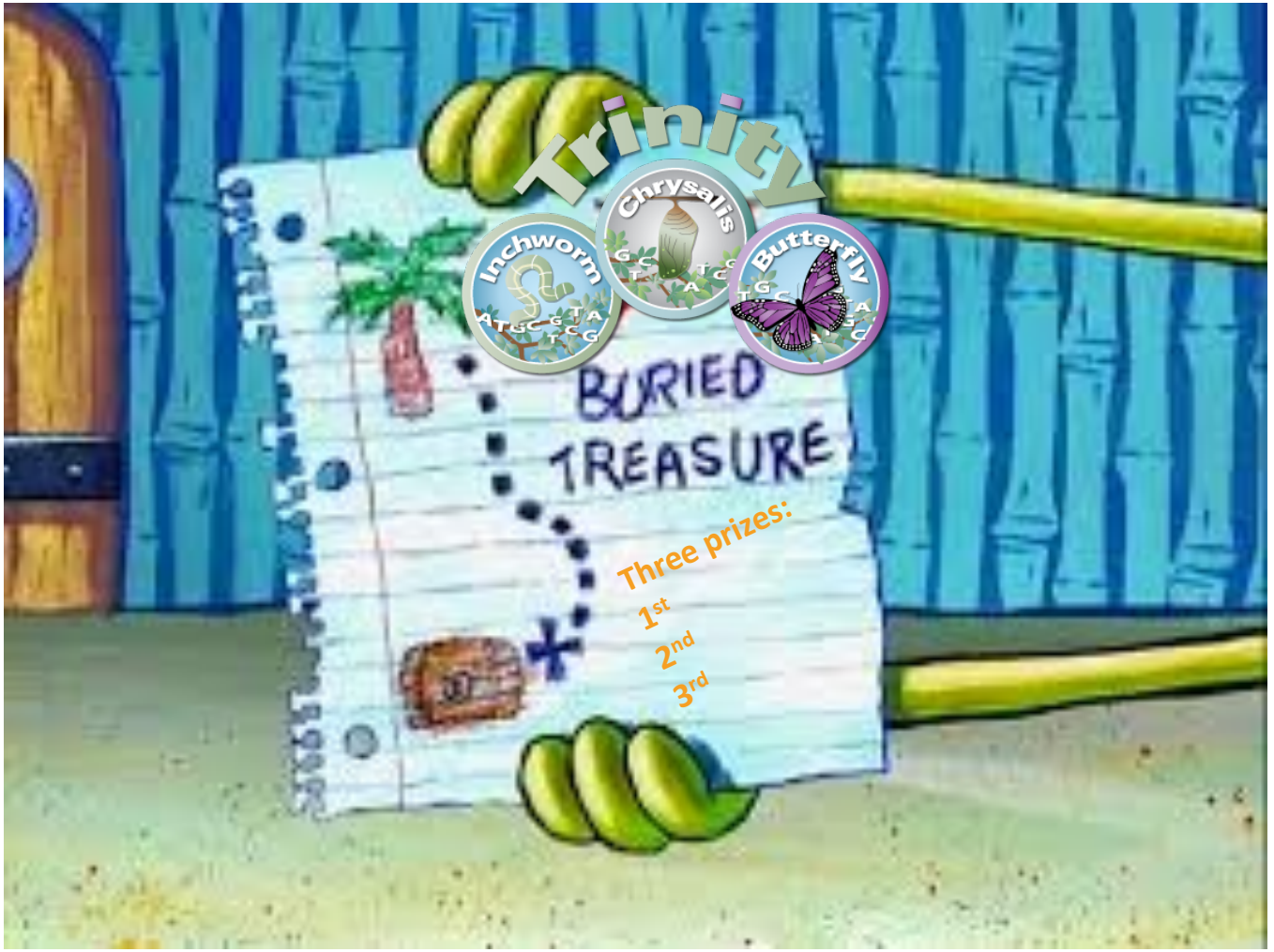
(Krumlov Prelate, 2-5pm and 7-10pm)

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