De novo RNA-Seq Assembly and Transcriptome Studies Using Trinity

with Applications towards Non-model Organism Studies

Brian Haas

Broad Institute

Workshop on Genomics, Cesky Krumlov, Jan 2017

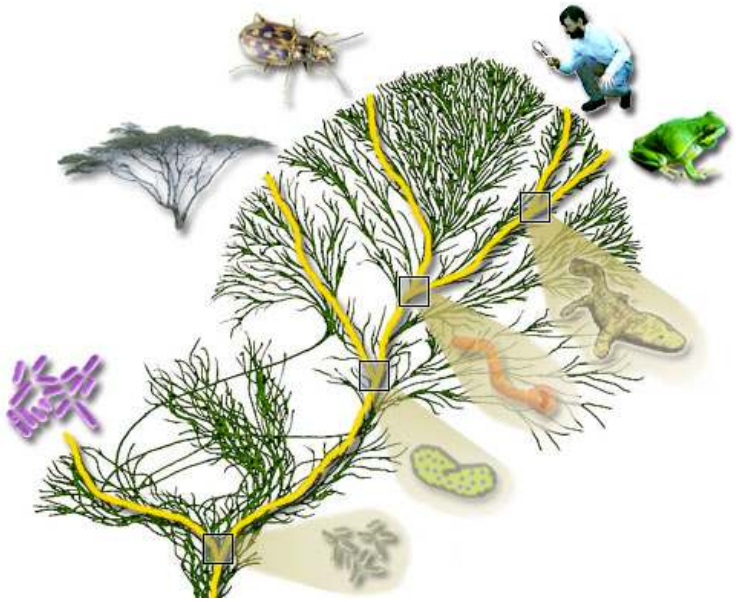
Transcriptomics Lecture Overview

- Overview of RNA-Seq
- Transcript reconstruction methods
- Trinity de novo assembly
- Transcriptome quality assessment

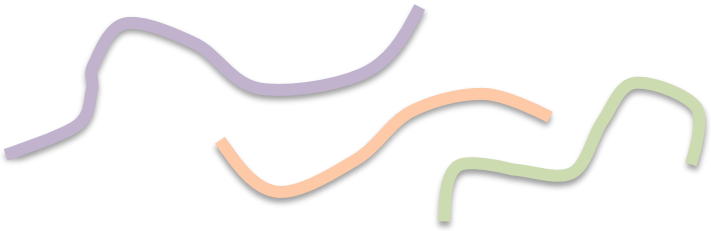
  (coffee break)
- Expression quantitation
- Differential expression analysis
- Functional annotation

  (stretch legs break)
- Case study: salamander transcriptome

RNA-Seq Empowers Transcriptome Studies



Extract RNA, convert to cDNA



Next-gen Sequencer
(pick your favorite)

Generating RNA-Seq: *How to Choose?*

Many different instruments hit the scene in the last decade

Illumina

454

SOLiD

Helicos

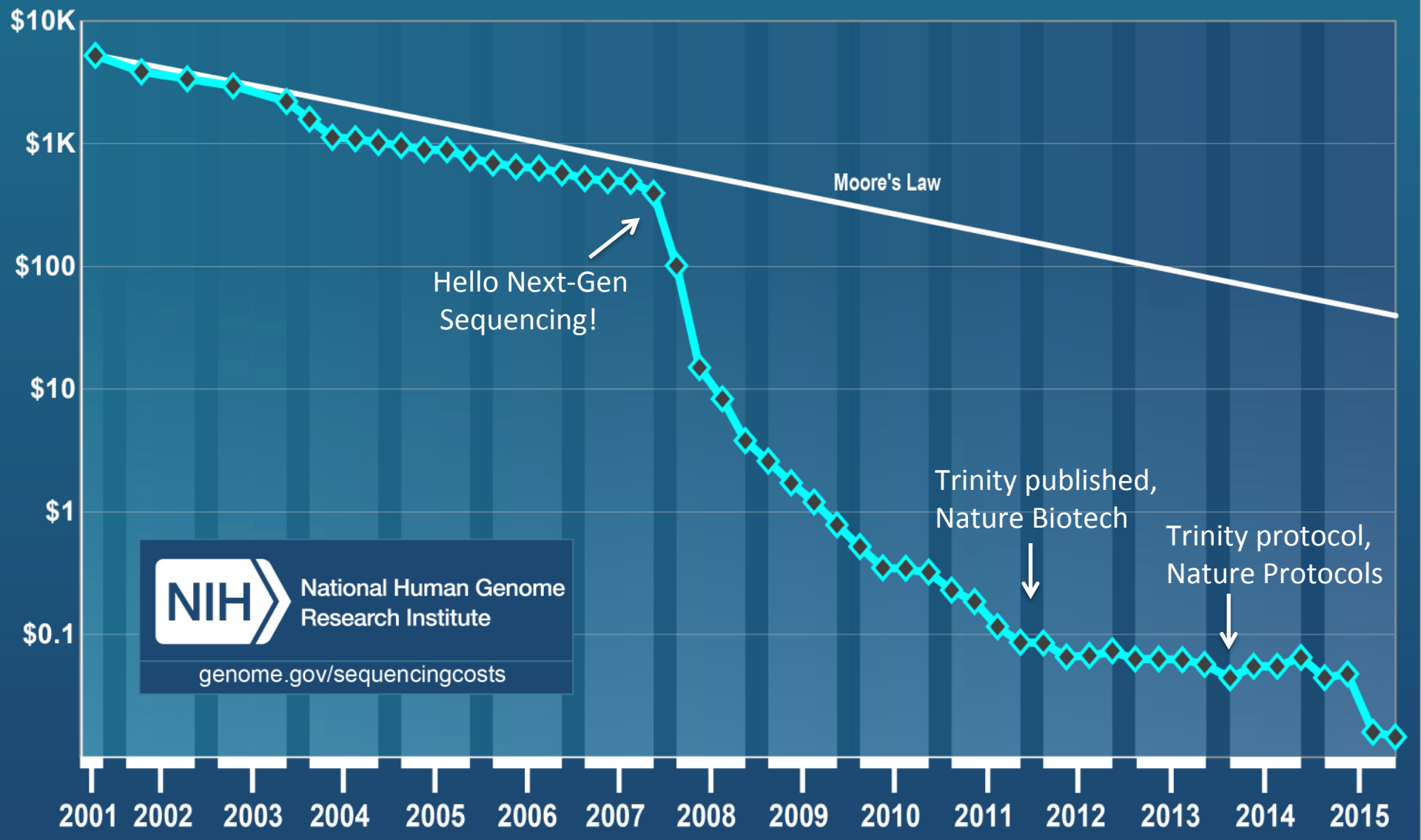
Ion Torrent

Pacific Biosciences

Oxford Nanopore

Slide courtesy of Joshua Levin, Broad Institute.

Cost per Raw Megabase of DNA Sequence



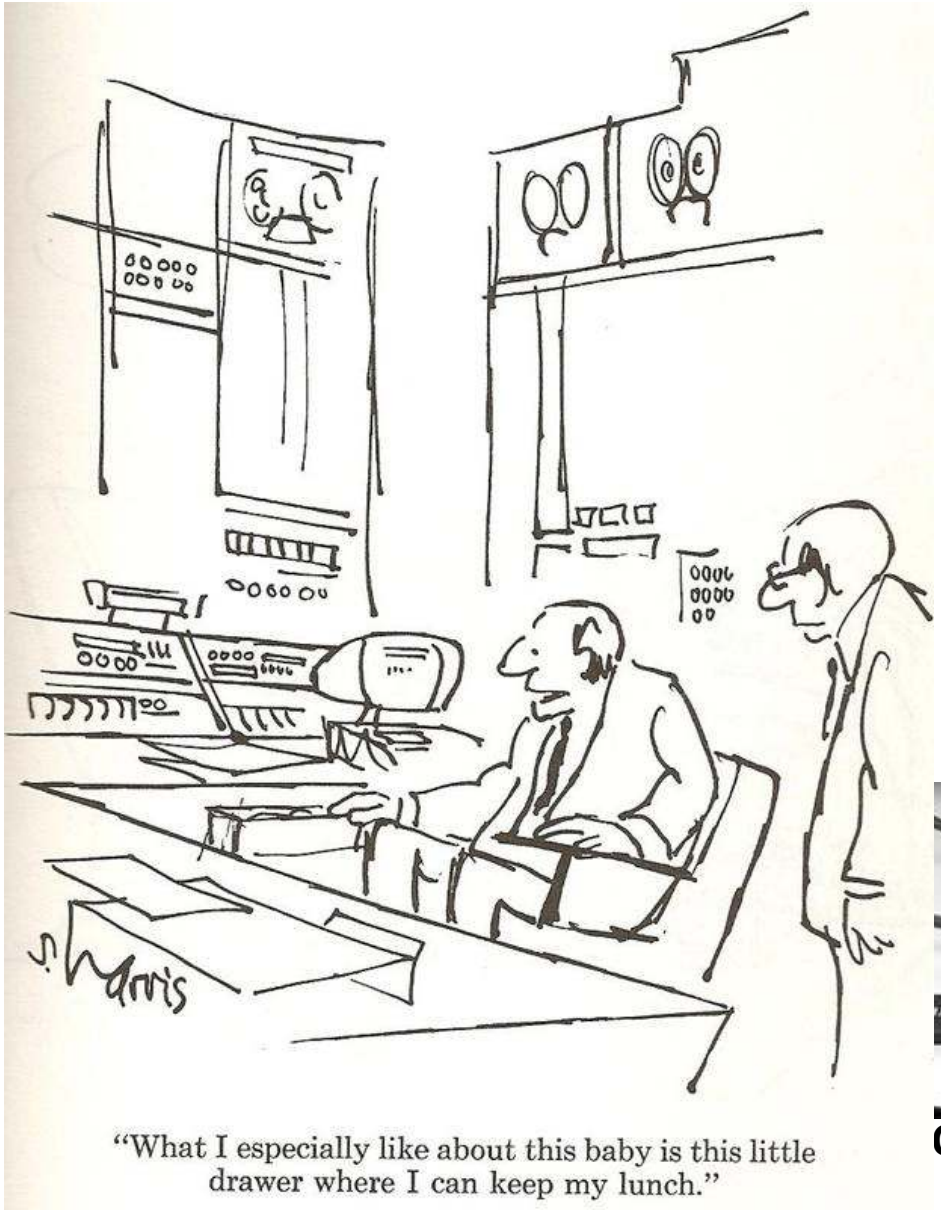
From https://www.genome.gov/sequencingcostsdata/

RNA-Seq: *How to Choose?*

Illumina

Helicos

Ion Torrent



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

Oxford Nanopore

Slide courtesy of Joshua Levin, Broad Institute.

Generating RNA-Seq: *How to Choose?*

Popular choices for RNA-Seq today

Illumina

454

SOLiD

Helicos

Ion Torrent

Pacific Biosciences

Oxford Nanopore

Generating RNA-Seq: *How to Choose?*

Popular choices for RNA-Seq today

[Current RNA-Seq workhorse]

Illumina

Ion Torrent

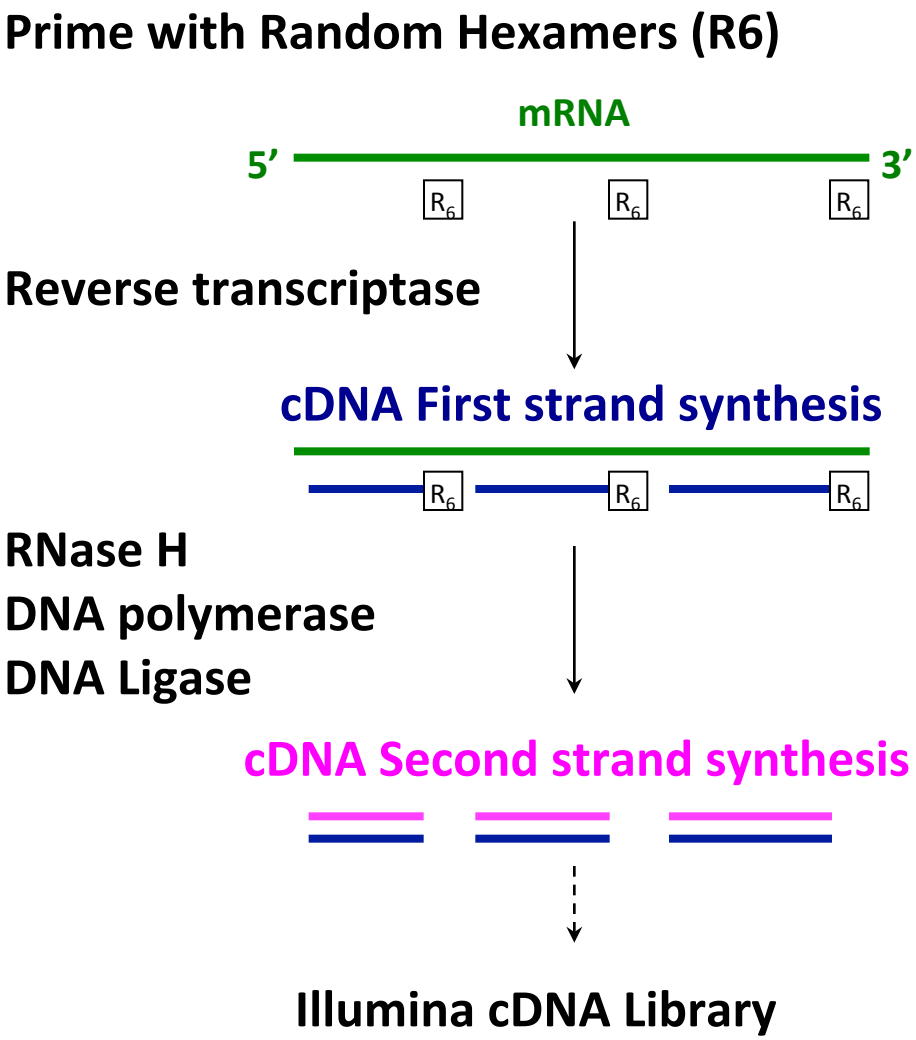
[Full-length single molecule sequencing]

Pacific Biosciences

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

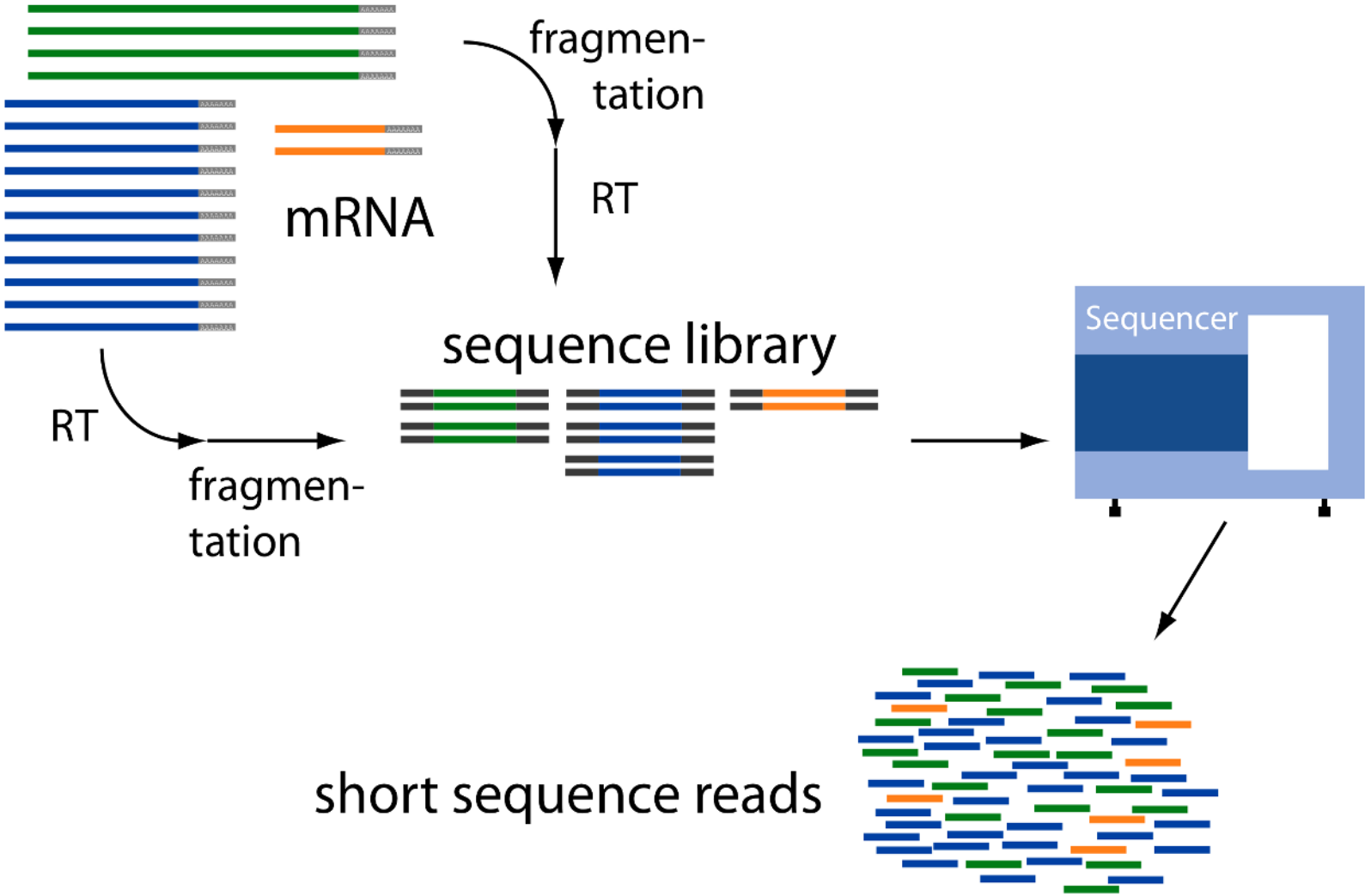
Oxford Nanopore

RNA-Seq: How do we make cDNA?



Slide courtesy of Joshua Levin, Broad Institute.

Overview of RNA-Seq



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

Common Data Formats for RNA-Seq

FASTA format:

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

FASTQ format:

```
@61DFRAAXX100204:1:100:10494:3070/1
AAACAACAGGGCACATTGTCACTCTTGTATTTGAAAAACACTTTCCGGCCAT
+
ACCCCCCCCCCCCCCCCCCCCCCCCCCCCCCBC?CCCCCCCCC@@CACCCCCA
```

Read

Quality values

AsciiEncodedQual(x) = -10 * log10(Pwrong(x)) + 33

AsciiEncodedQual (‘C’) = **64**

So, Pwrong(‘C’) = 10^((**64**-33/ (-10)) = 10^-3.4 = **0.0004**

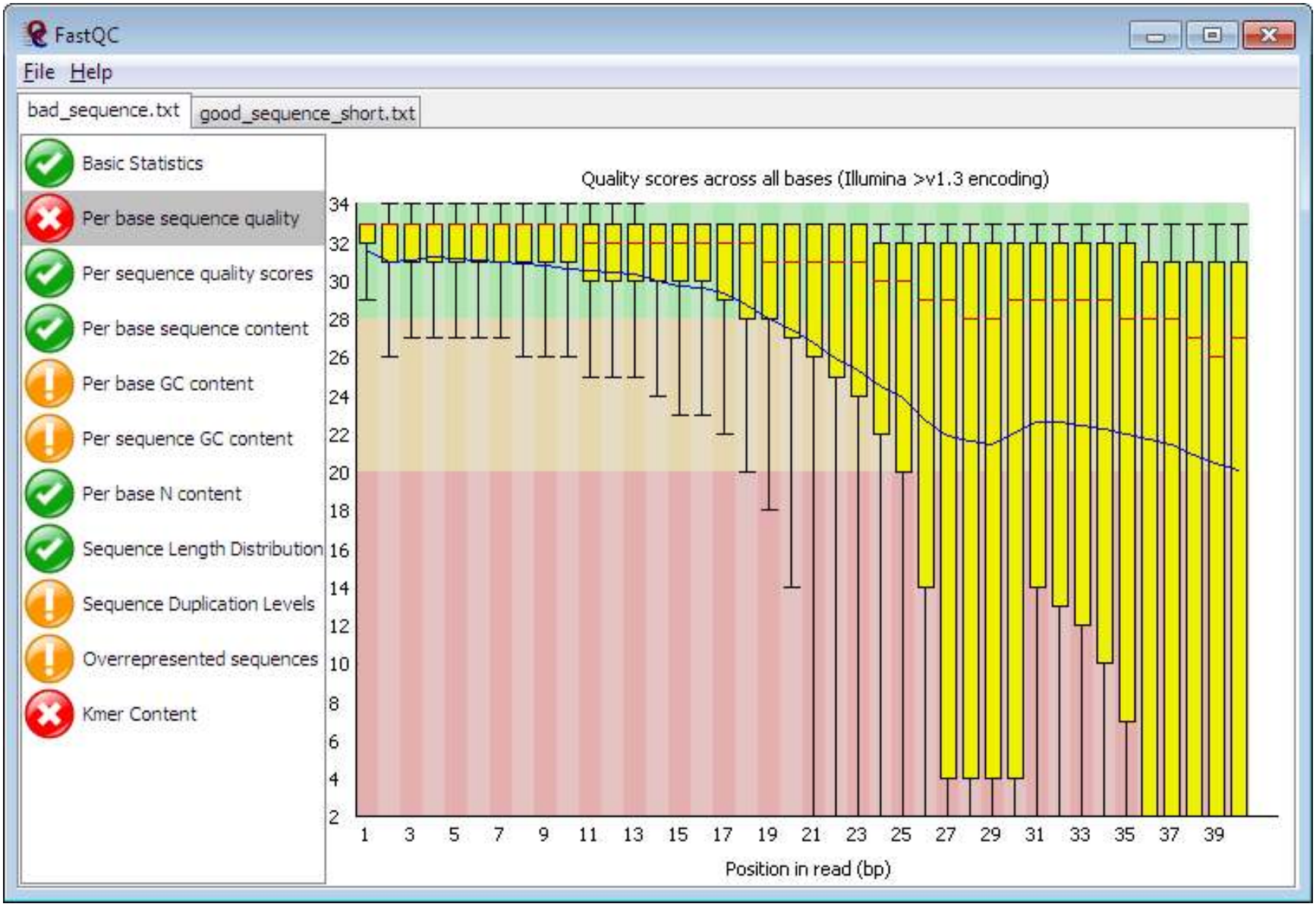
Paired-end Sequences

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

```
@61DFRAAXX100204:1:100:10494:3070/1
AAACAACAGGGCACATTGTCACTCTTGTATTTGAAAAACACTTTCCGGCCAT
+
ACCCCCCCCCCCCCCCCCCCCCCCCCCCCCCBC?CCCCCCCCC@@CACCCCCA
```

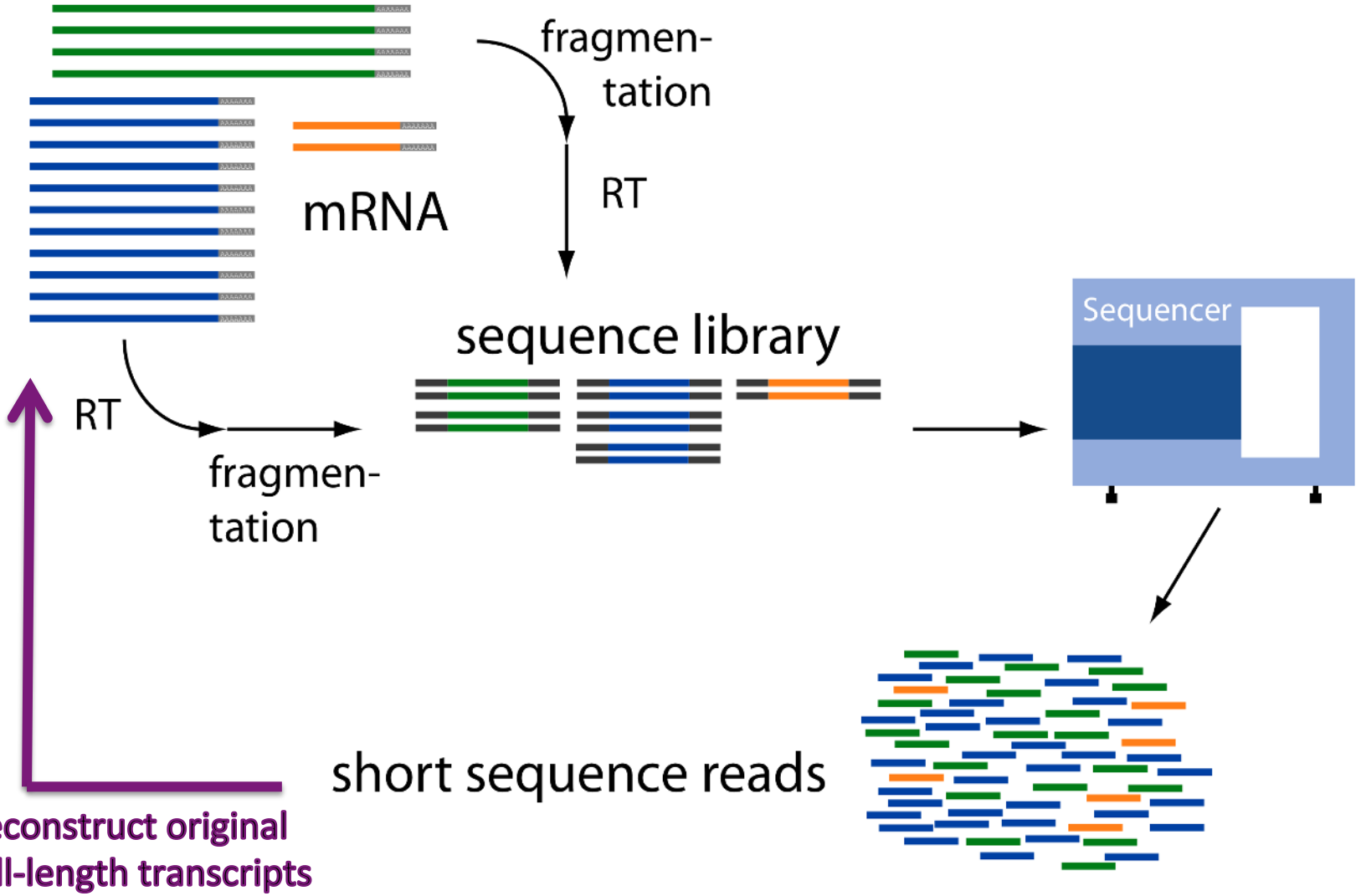
```
@61DFRAAXX100204:1:100:10494:3070/2
CTCAAATGGTTAATTCTCAGGCTGCAAATATTCGTTCAGGATGGAAGAACA
+
C<CCCCCCCACCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCBCCCC
```

Read Quality Assessment



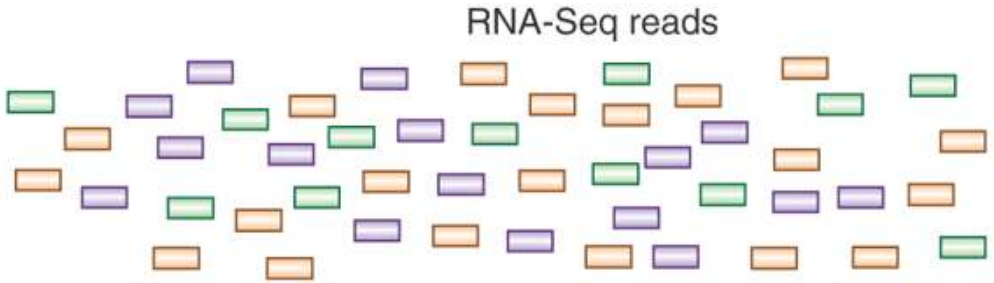
From: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Overview of RNA-Seq



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

Transcript Reconstruction from RNA-Seq Reads



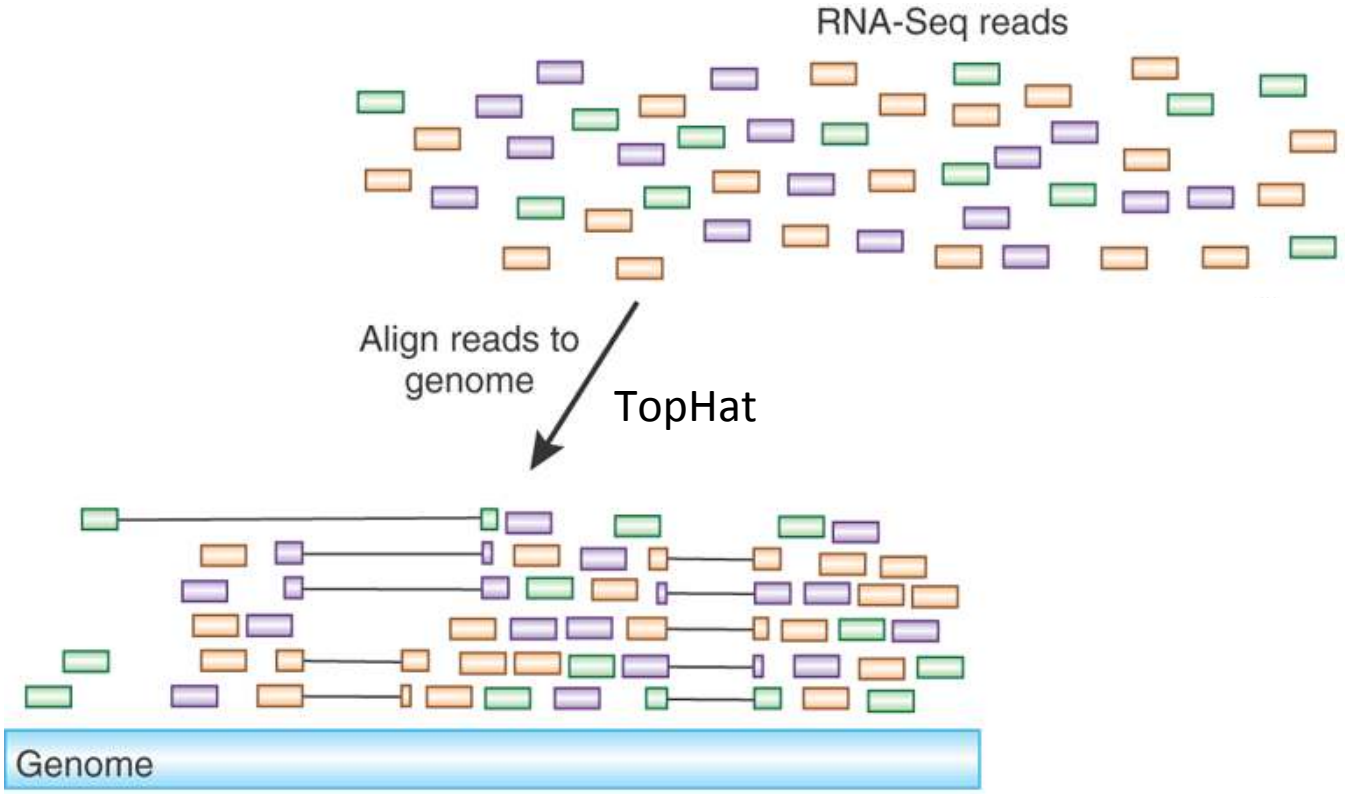
Advancing RNA-Seq analysis

Brian J Haas & Michael C Zody

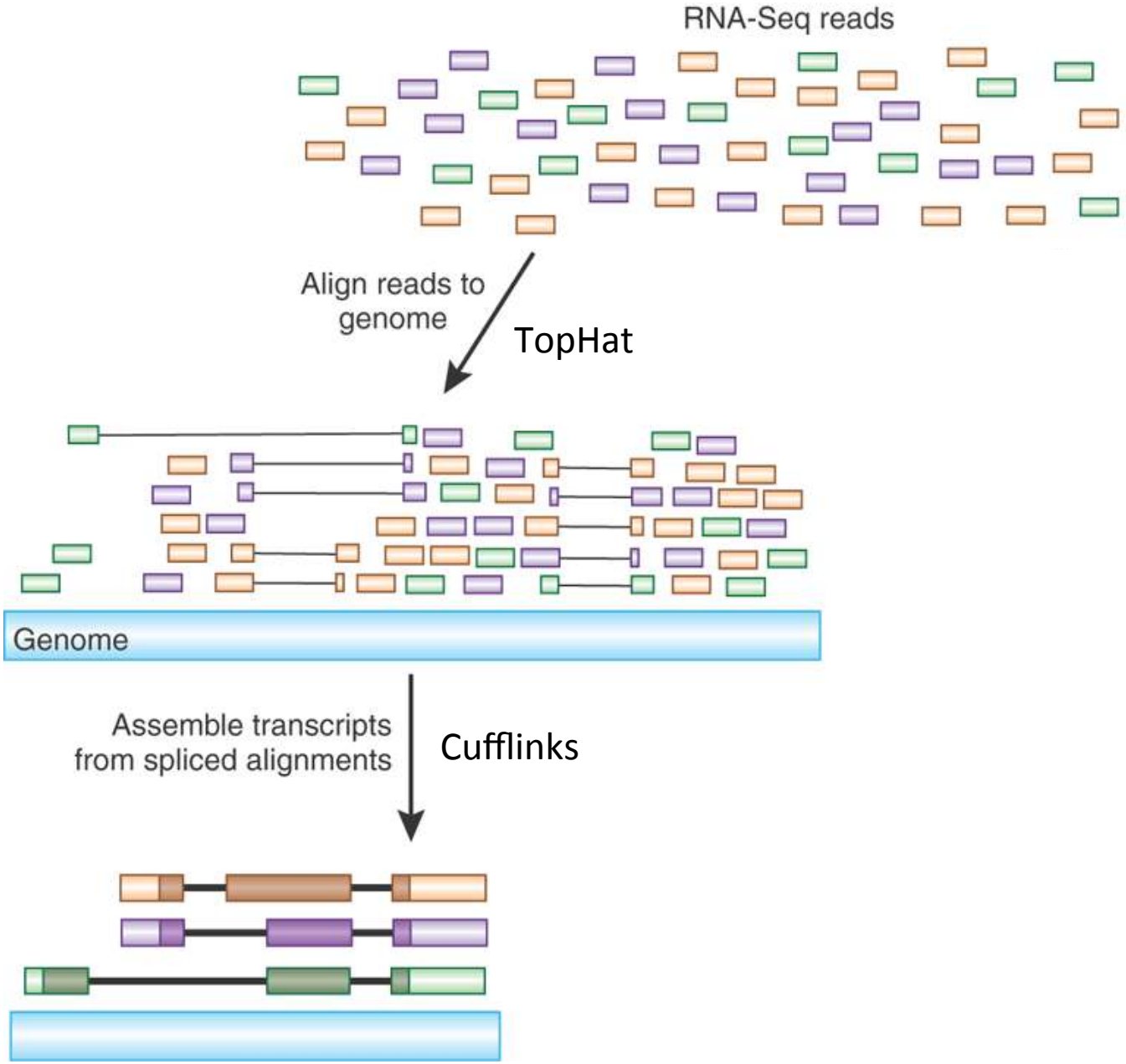
Nature Biotech, 2010

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

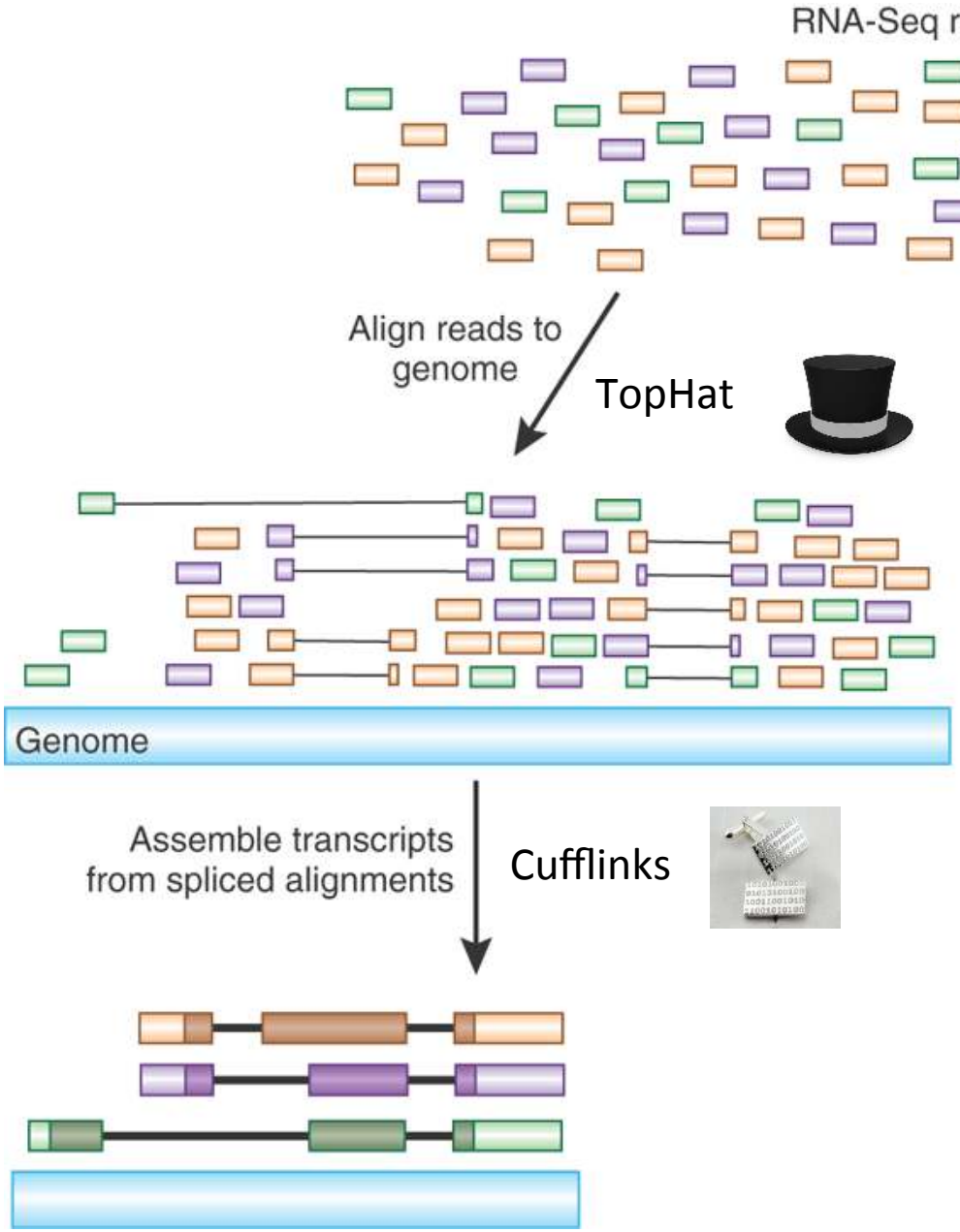
Transcript Reconstruction from RNA-Seq Reads



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

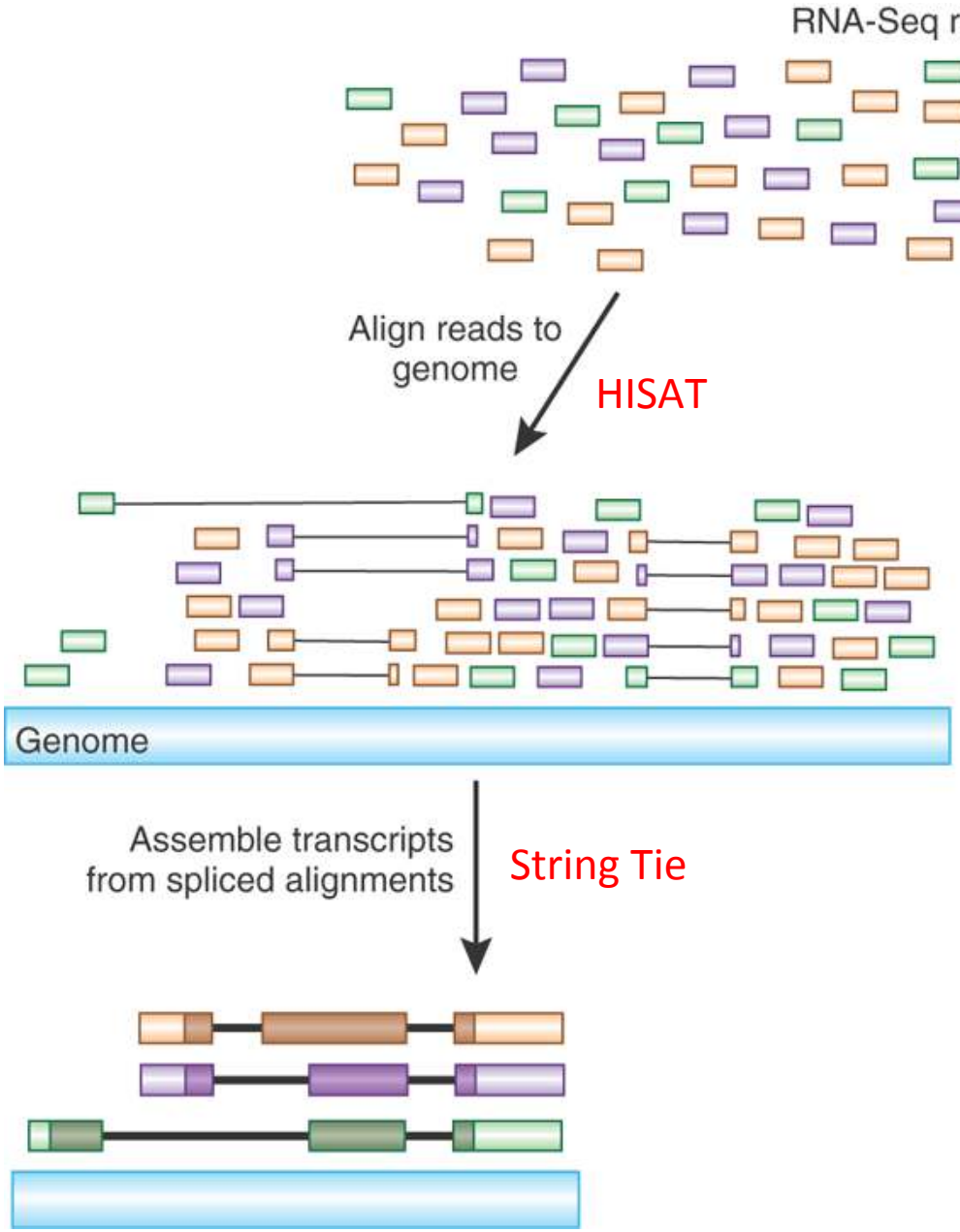
Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks

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

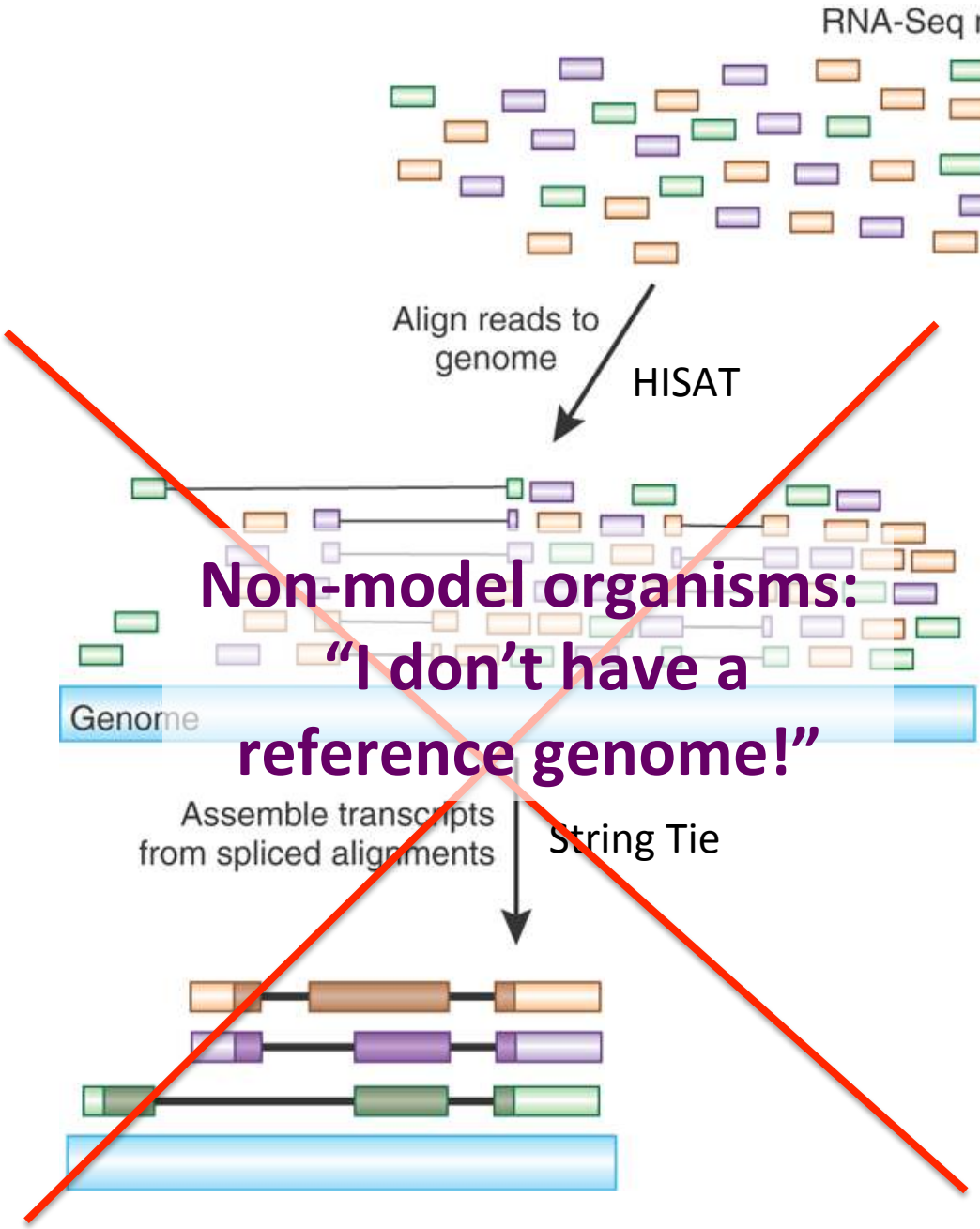
Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown

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

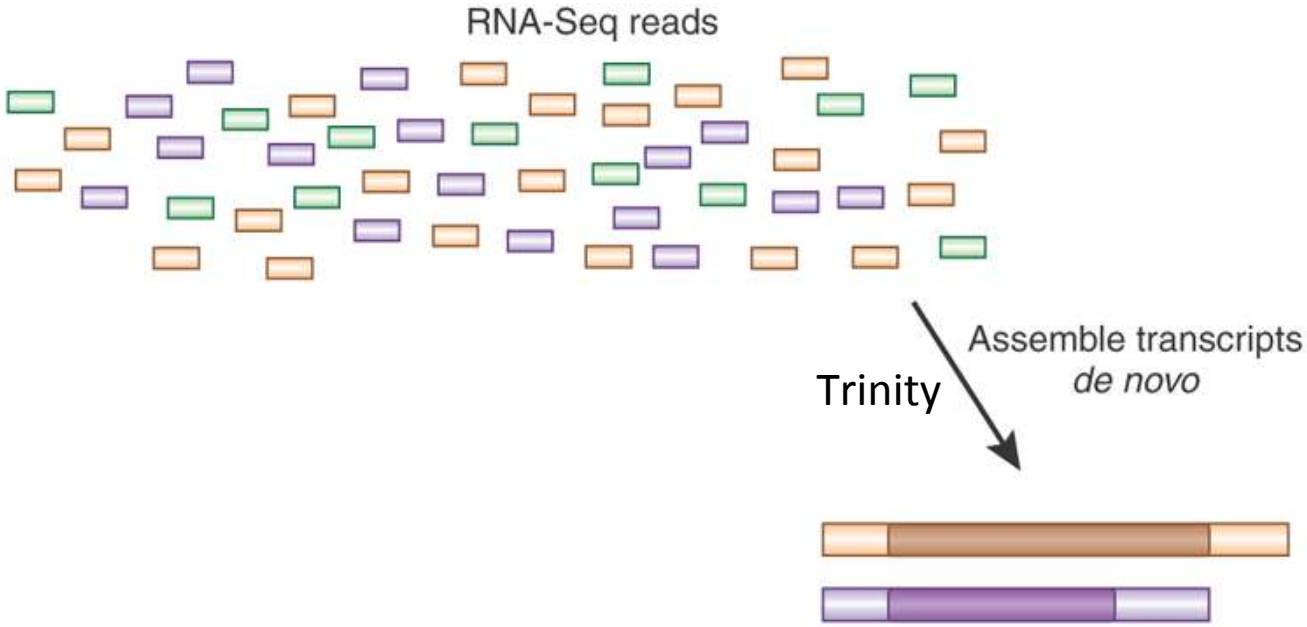
Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown

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

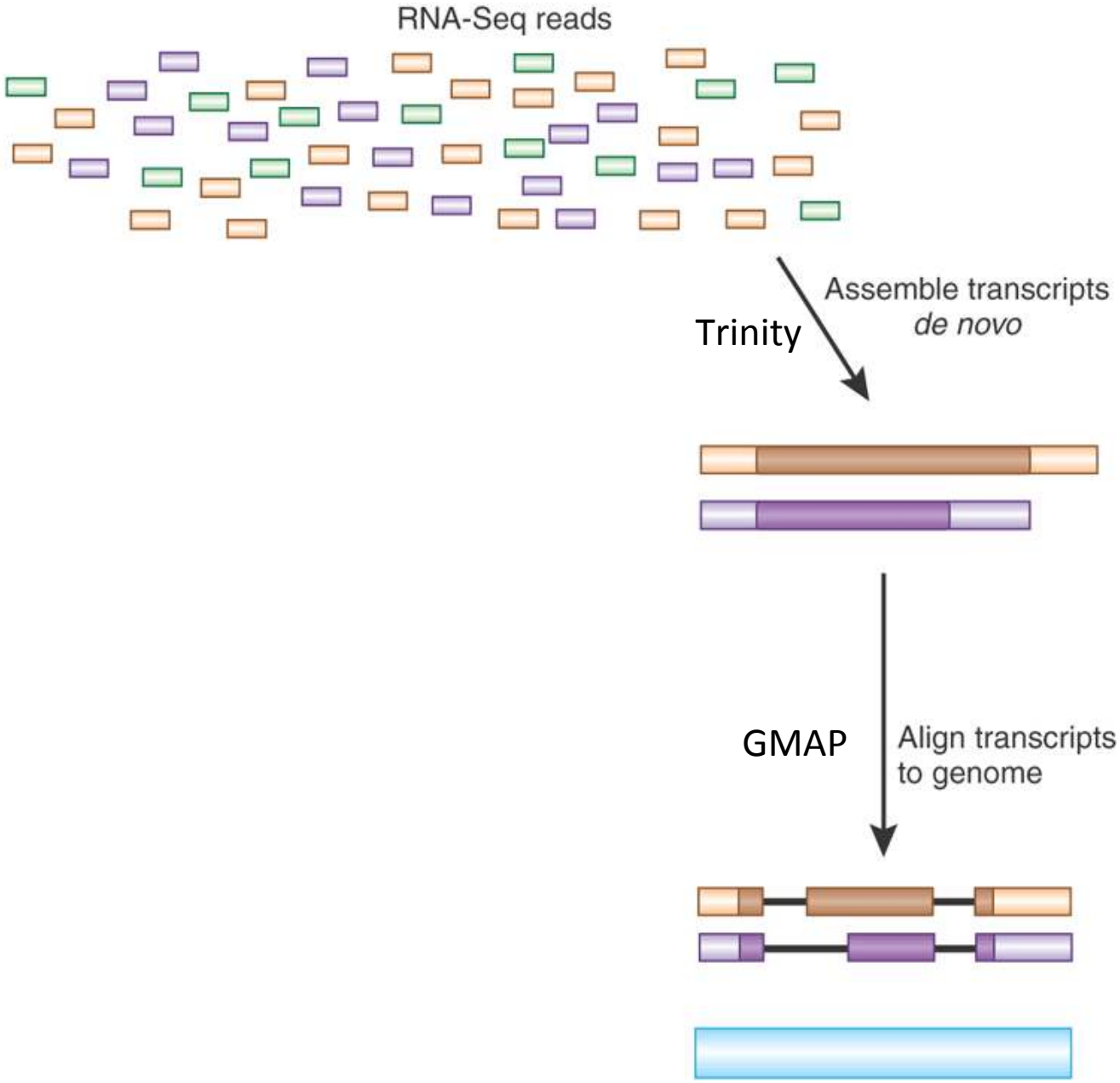
Affiliations | Contributions | Corresponding author

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

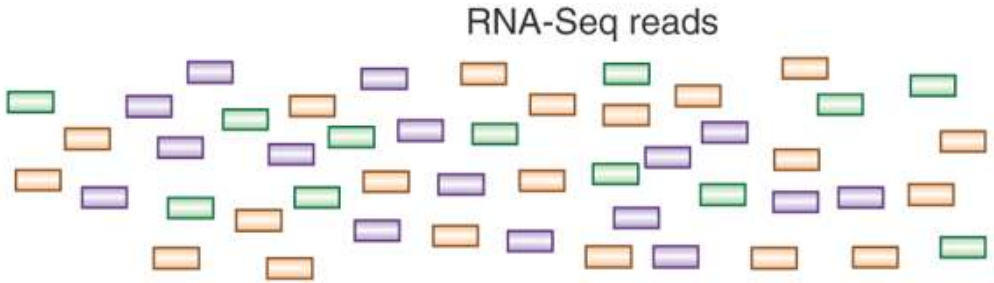
Transcript Reconstruction from RNA-Seq Reads



Transcript Reconstruction from RNA-Seq Reads



Transcript Reconstruction from RNA-Seq Reads



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

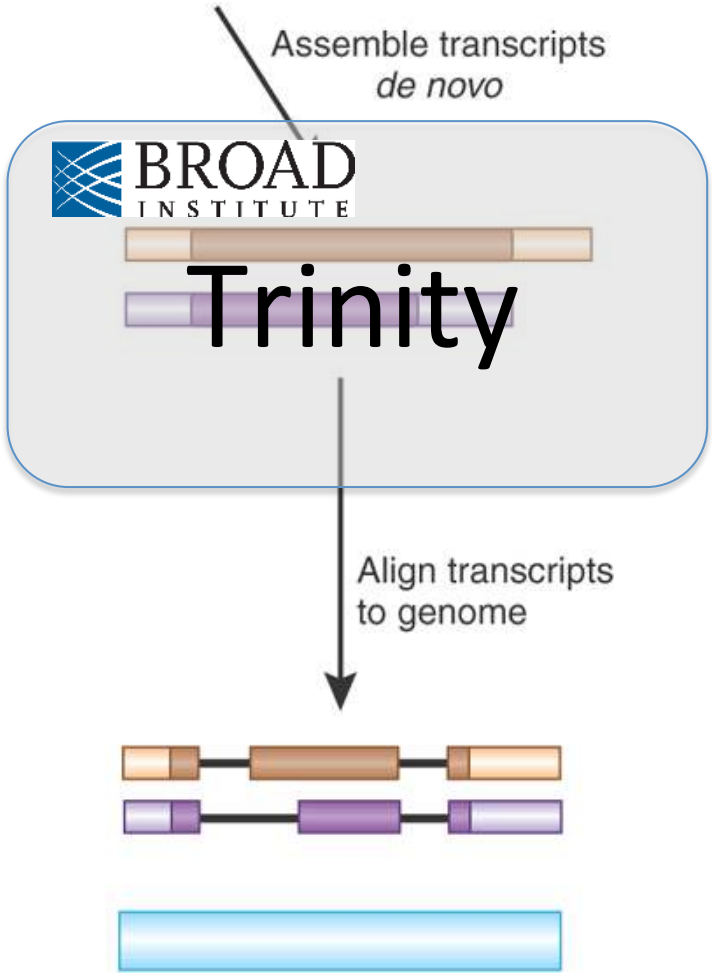
NATURE PROTOCOLS | PROTOCOL

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

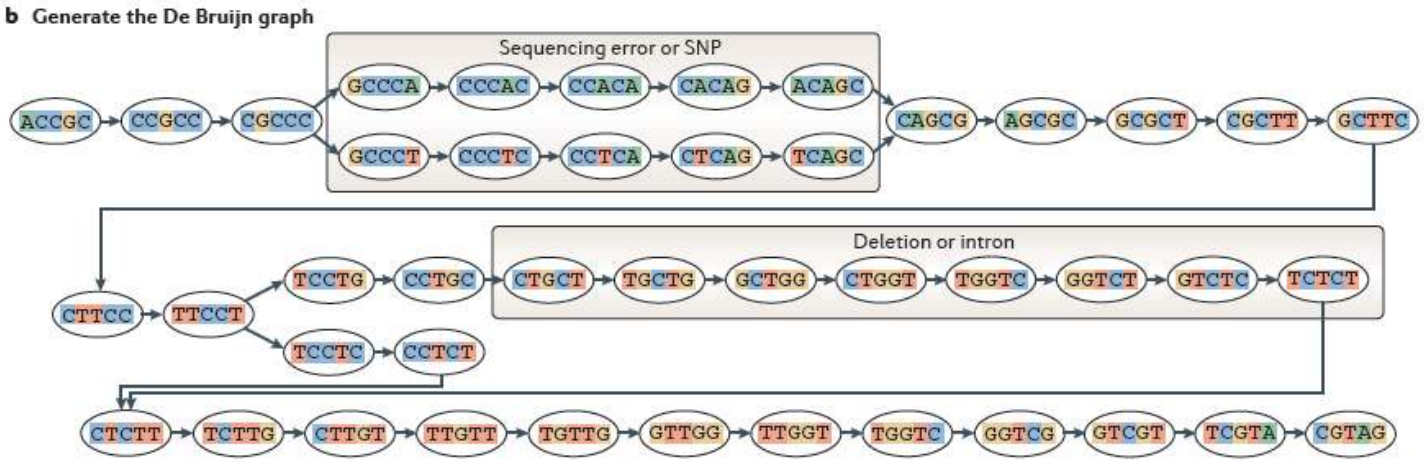
Affiliations | Contributions | Corresponding authors

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

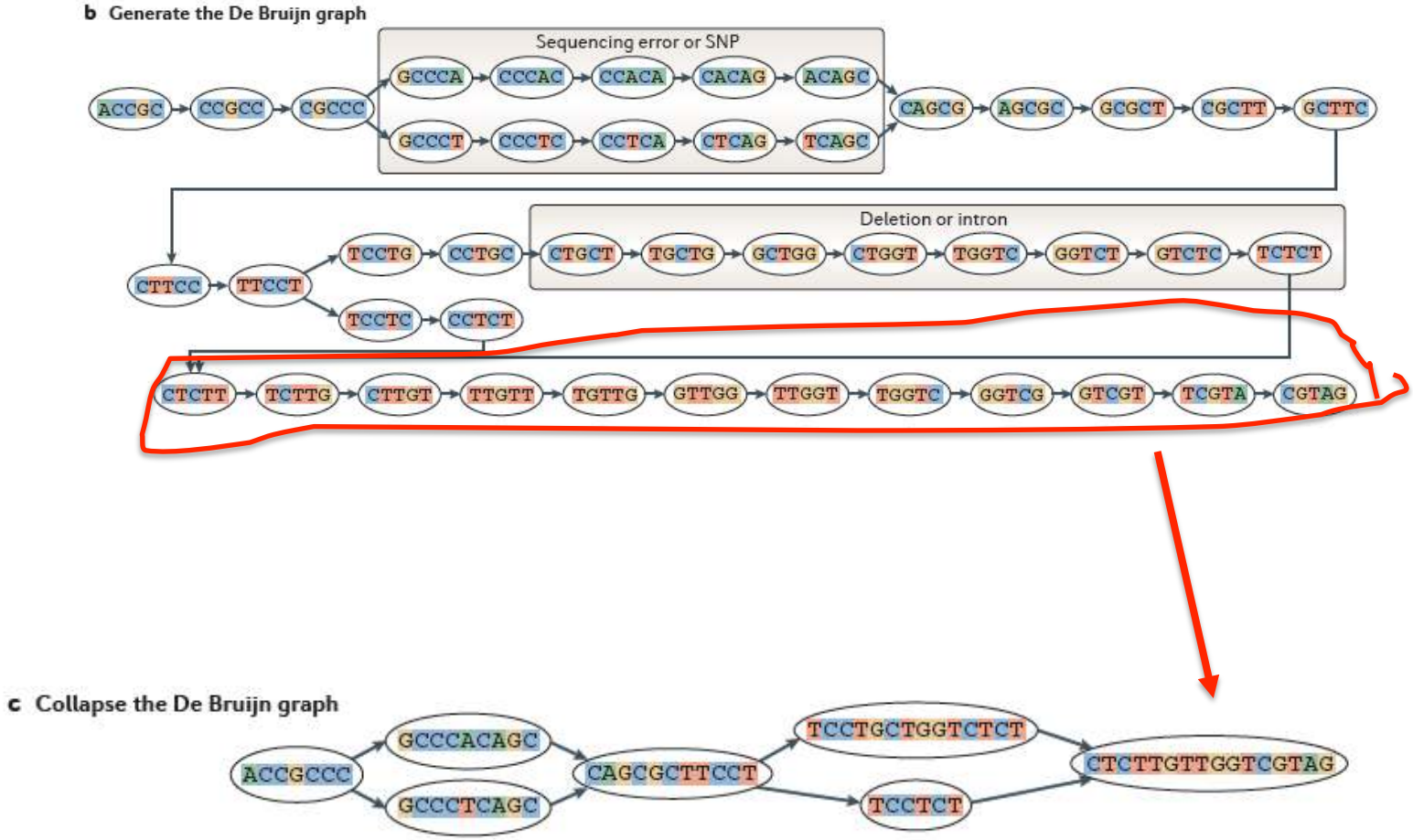


The General Approach to *De novo* RNA-Seq Assembly Using De Bruijn Graphs

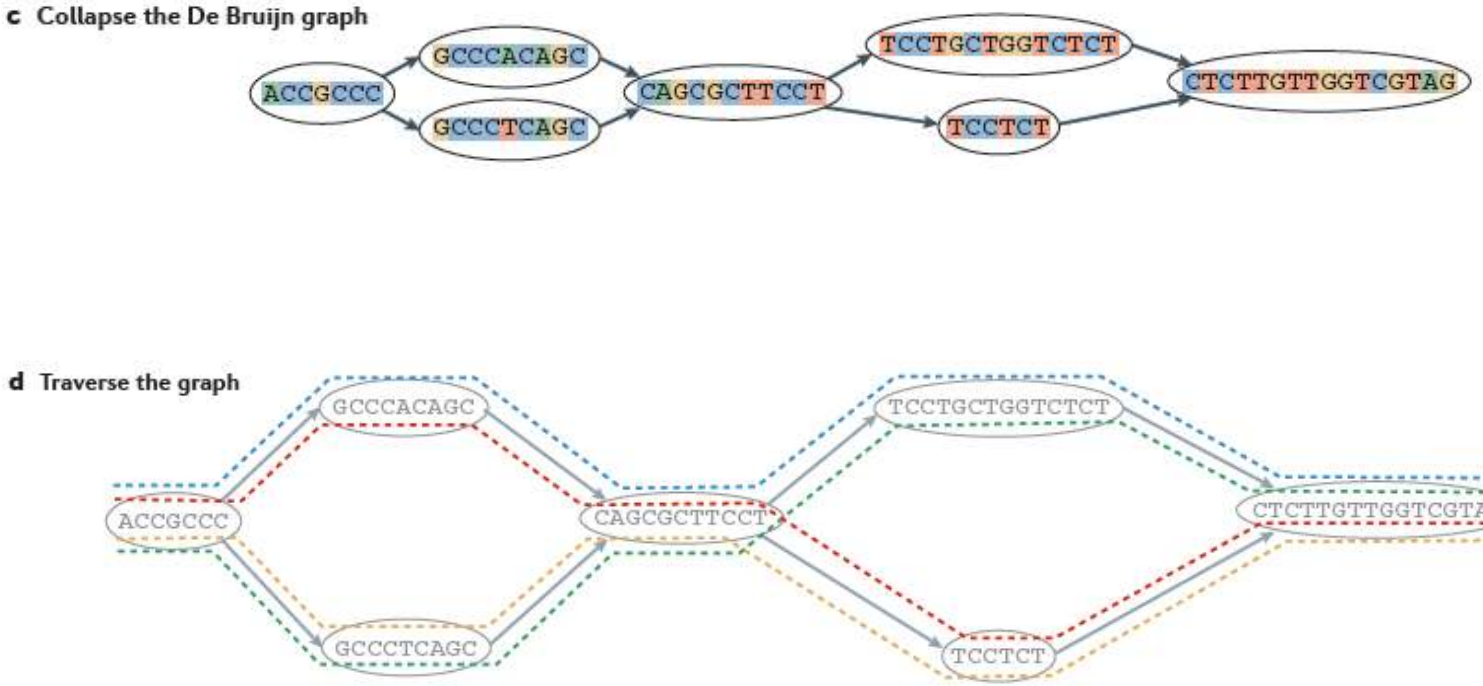
Sequence Assembly via De Bruijn Graphs



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



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



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

Contrasting Genome and Transcriptome Assembly

Genome Assembly

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

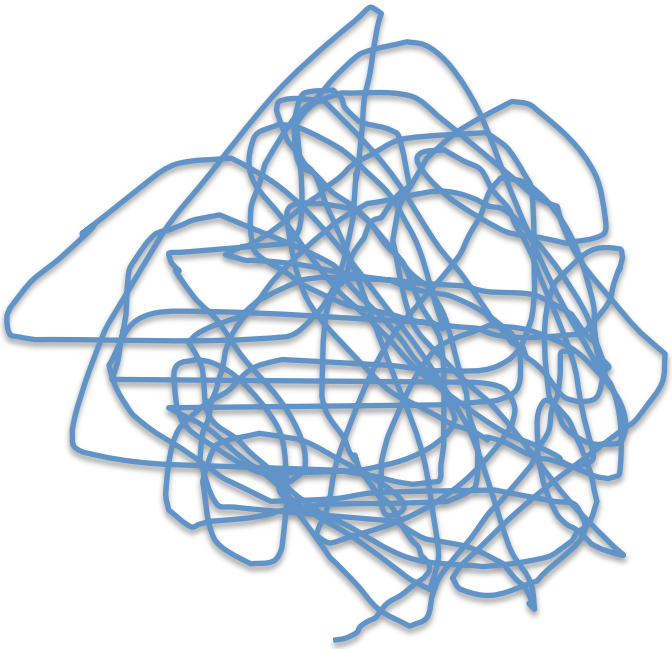
Transcriptome Assembly

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

Trinity Aggregates Isolated Transcript Graphs

Genome Assembly

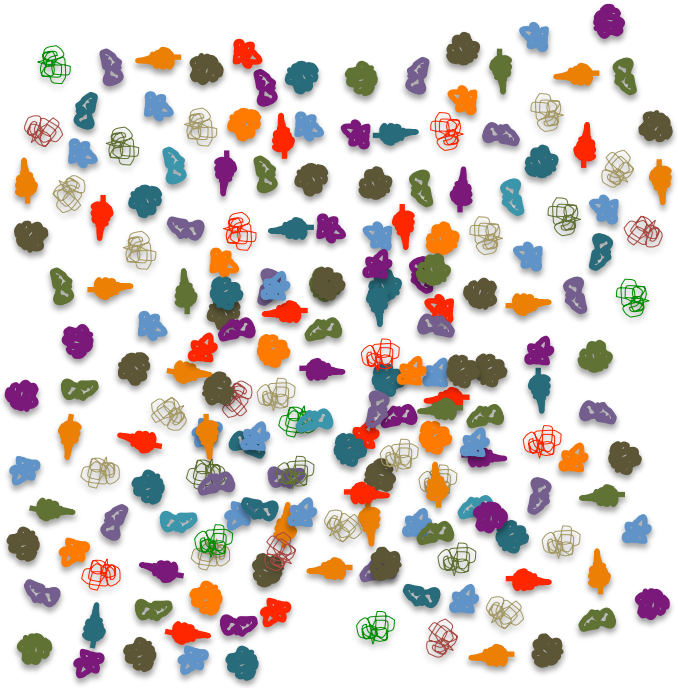
Single Massive Graph



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**

>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

TGTGAAA

…

Overlapping kmers of length (k)

Kmer Catalog (hashtable)

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

Inchworm Algorithm

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

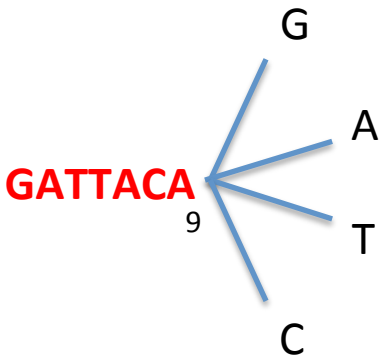
GATTACA
9

Kmer Catalog (hashtable)

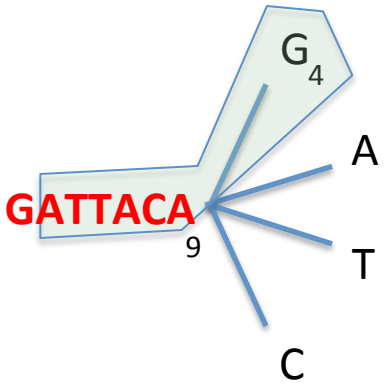
Kmer	Count among all reads
AATGTGA	4
ATGTGAA	2
TGTGAAA	1
GATTACA	9

Inchworm Algorithm

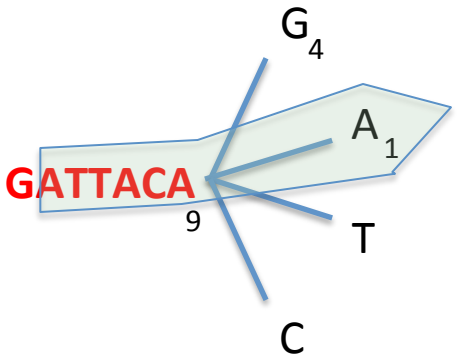
- Decompose all reads into overlapping Kmers => hashtable(kmer, count)
- Identify seed kmer as most abundant Kmer, ignoring low-complexity kmers.
- Extend kmer at 3' end, guided by coverage.



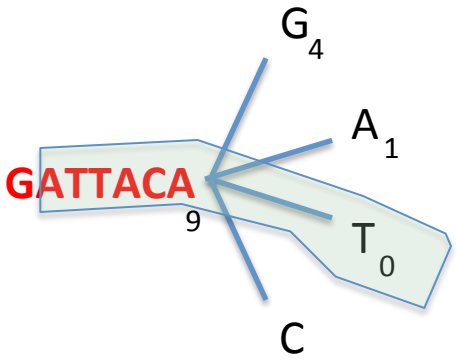
Inchworm Algorithm



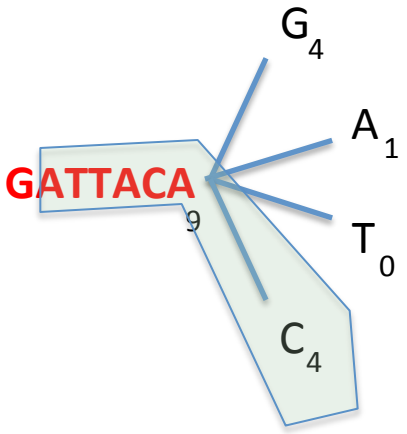
Inchworm Algorithm



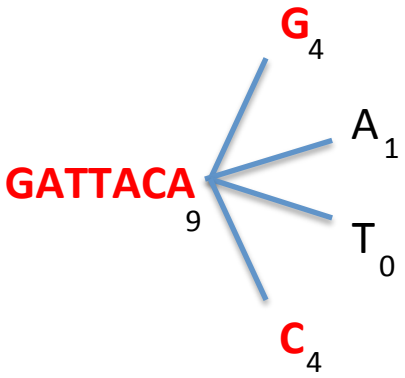
Inchworm Algorithm



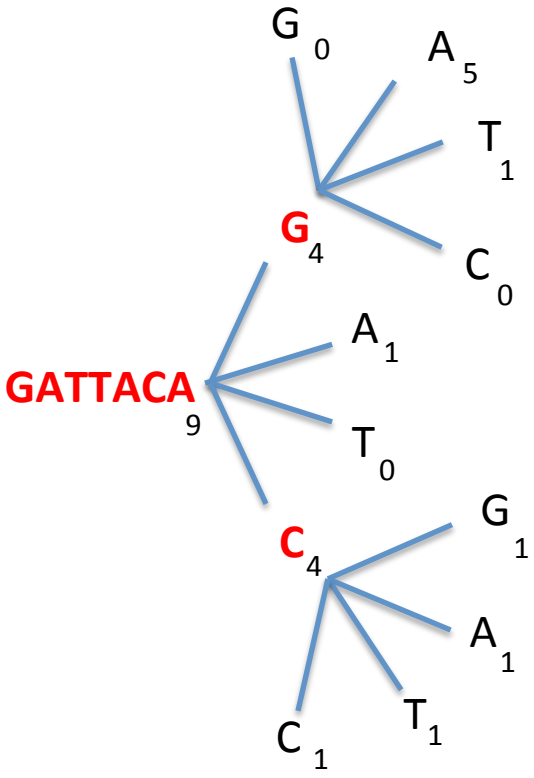
Inchworm Algorithm



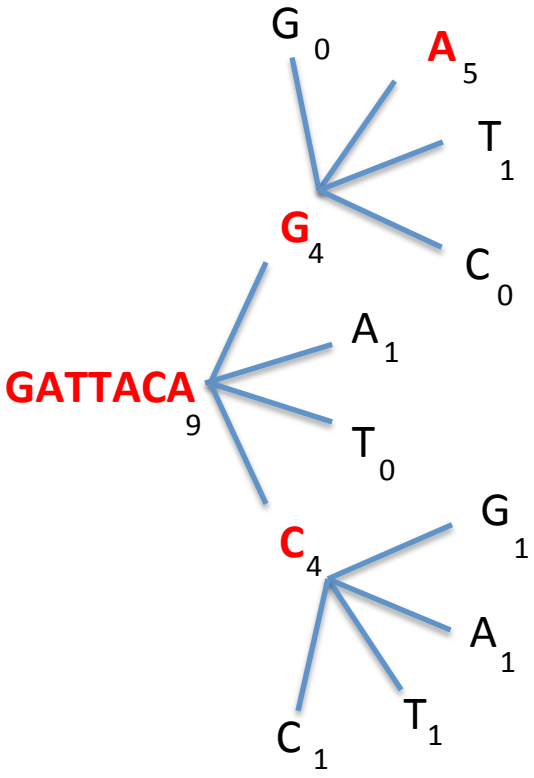
Inchworm Algorithm



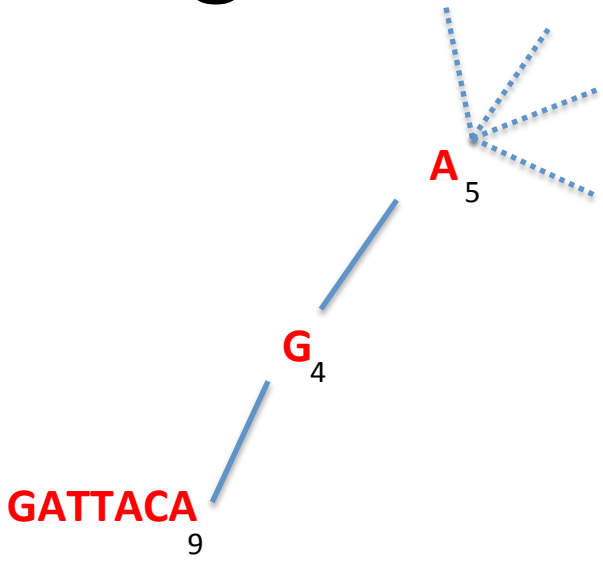
Inchworm Algorithm



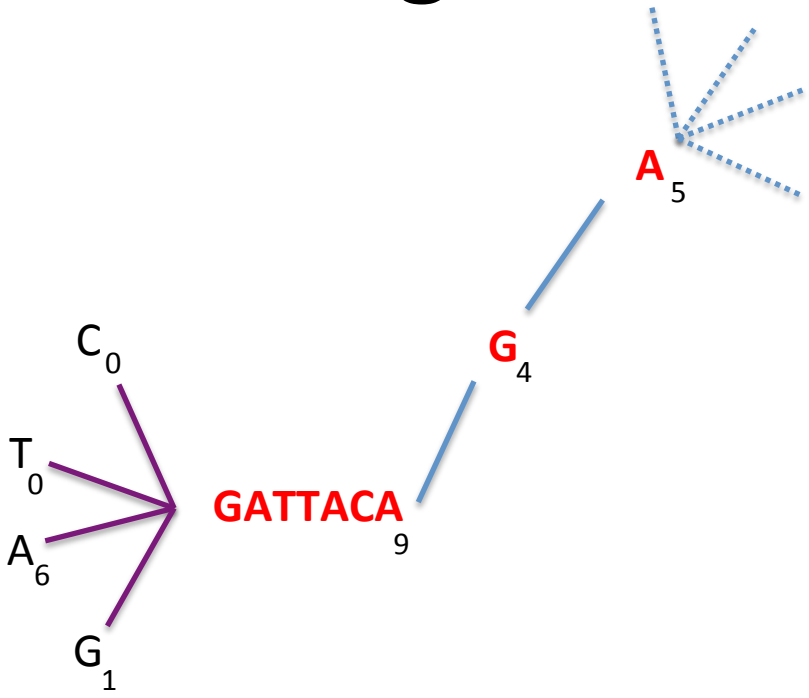
Inchworm Algorithm



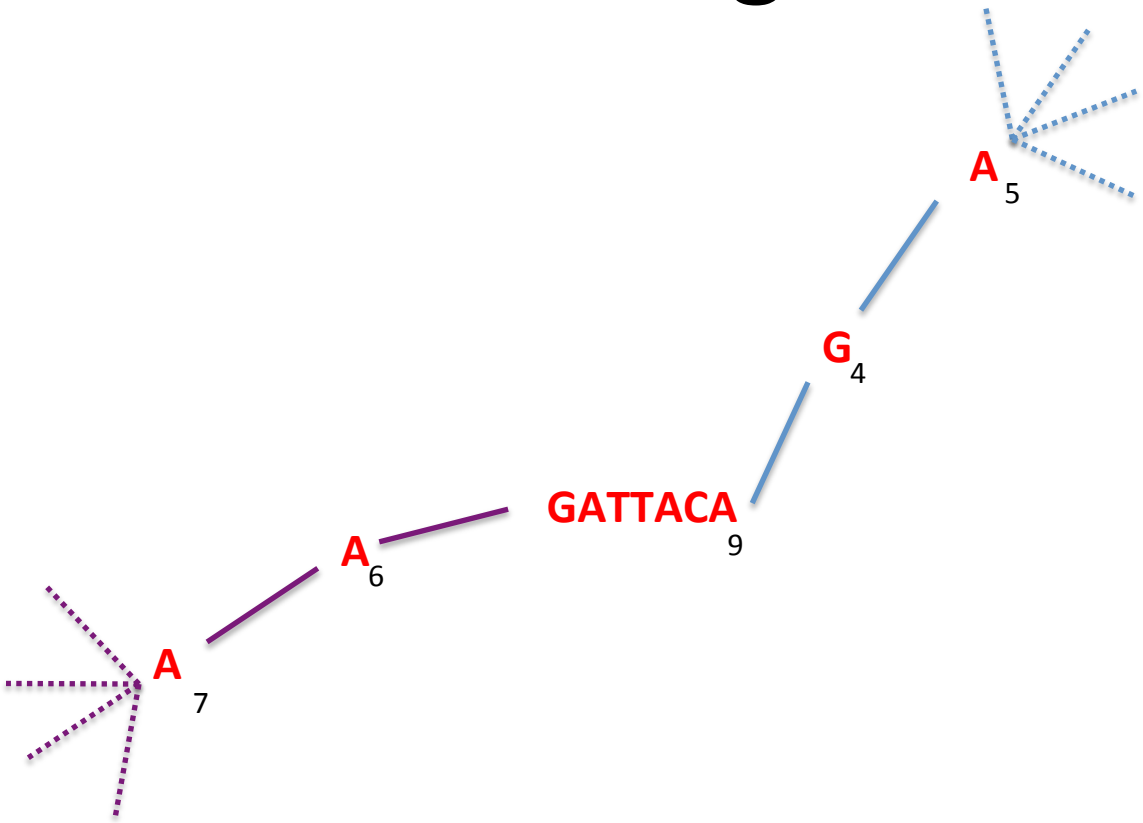
Inchworm Algorithm



Inchworm Algorithm



Inchworm Algorithm



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

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

Inchworm Contigs from Alt-Spliced Transcripts

Expressed isoforms

Isoform A

Isoform B

Inchworm Contigs from Alt-Spliced Transcripts

Expressed isoforms

Isoform A

Isoform B

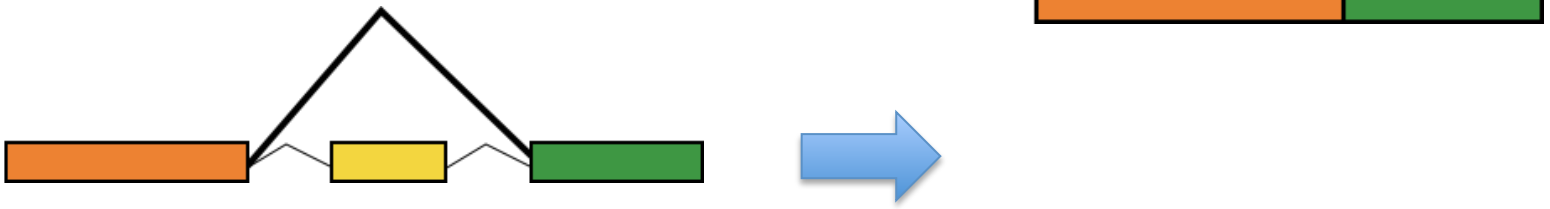
Expression

(low)

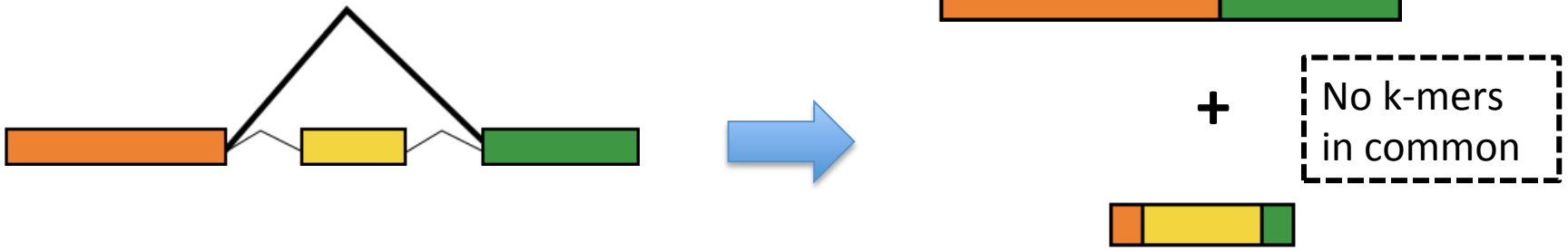
(high)

Graphical representation

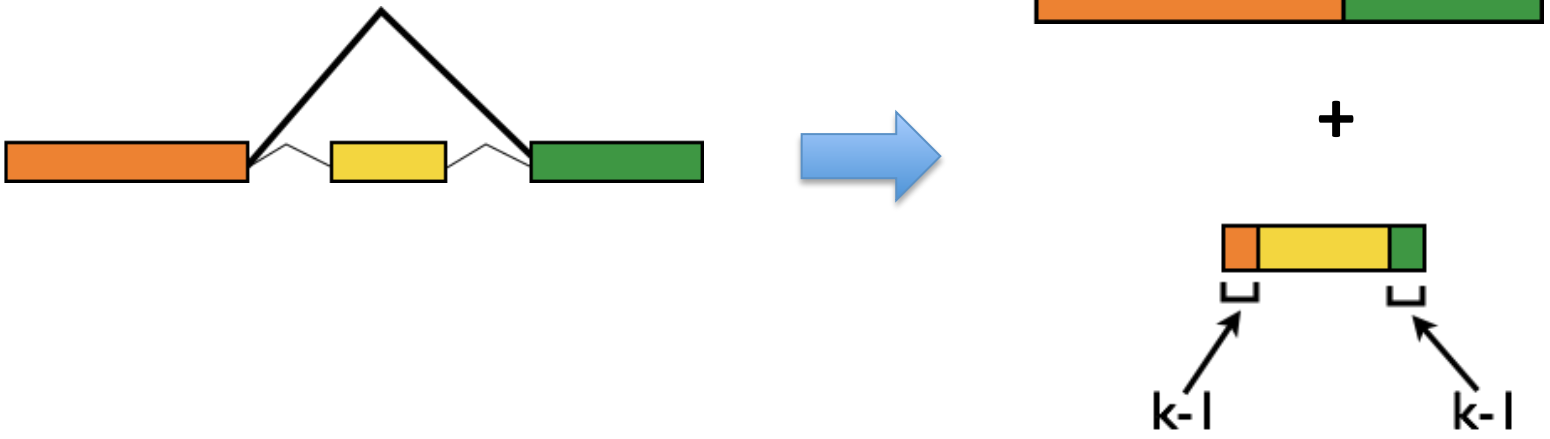
Inchworm Contigs from Alt-Spliced Transcripts



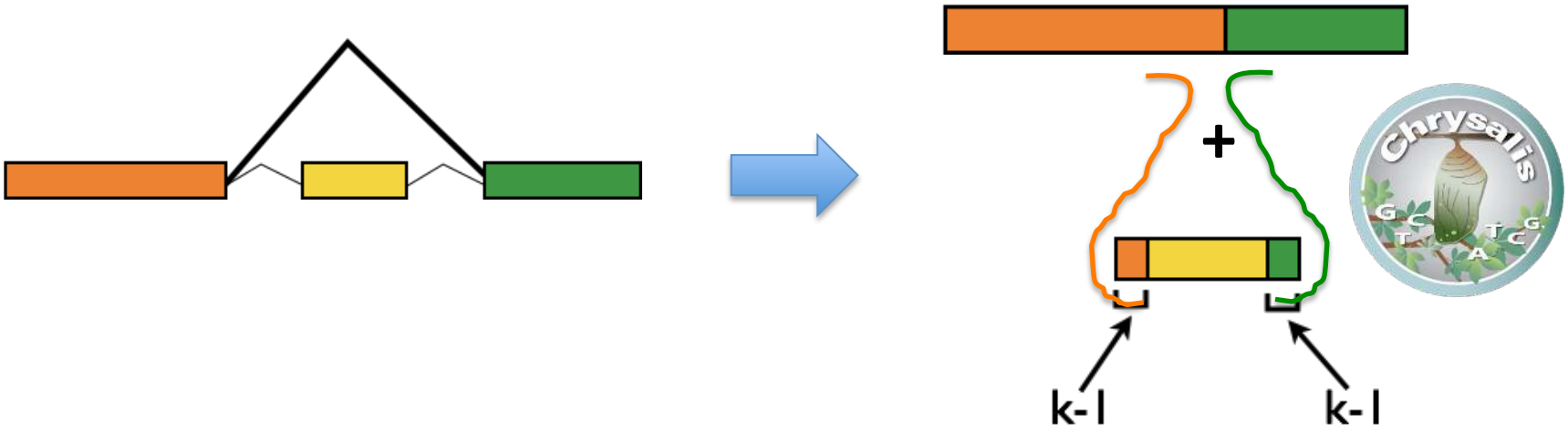
Inchworm Contigs from Alt-Spliced Transcripts



Inchworm Contigs from Alt-Spliced Transcripts



Chrysalis Re-groups Related Inchworm Contigs



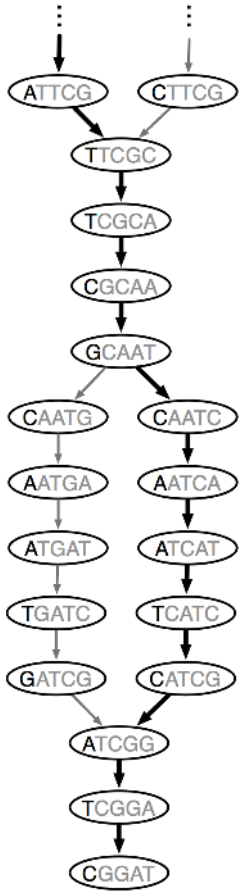
Chrysalis uses (k-1) overlaps and read support to link related Inchworm contigs

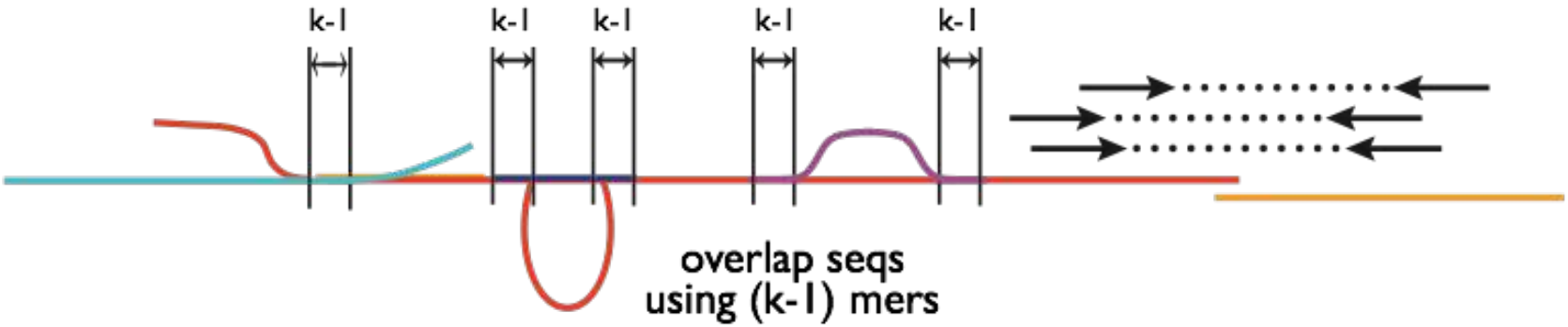
Chrysalis

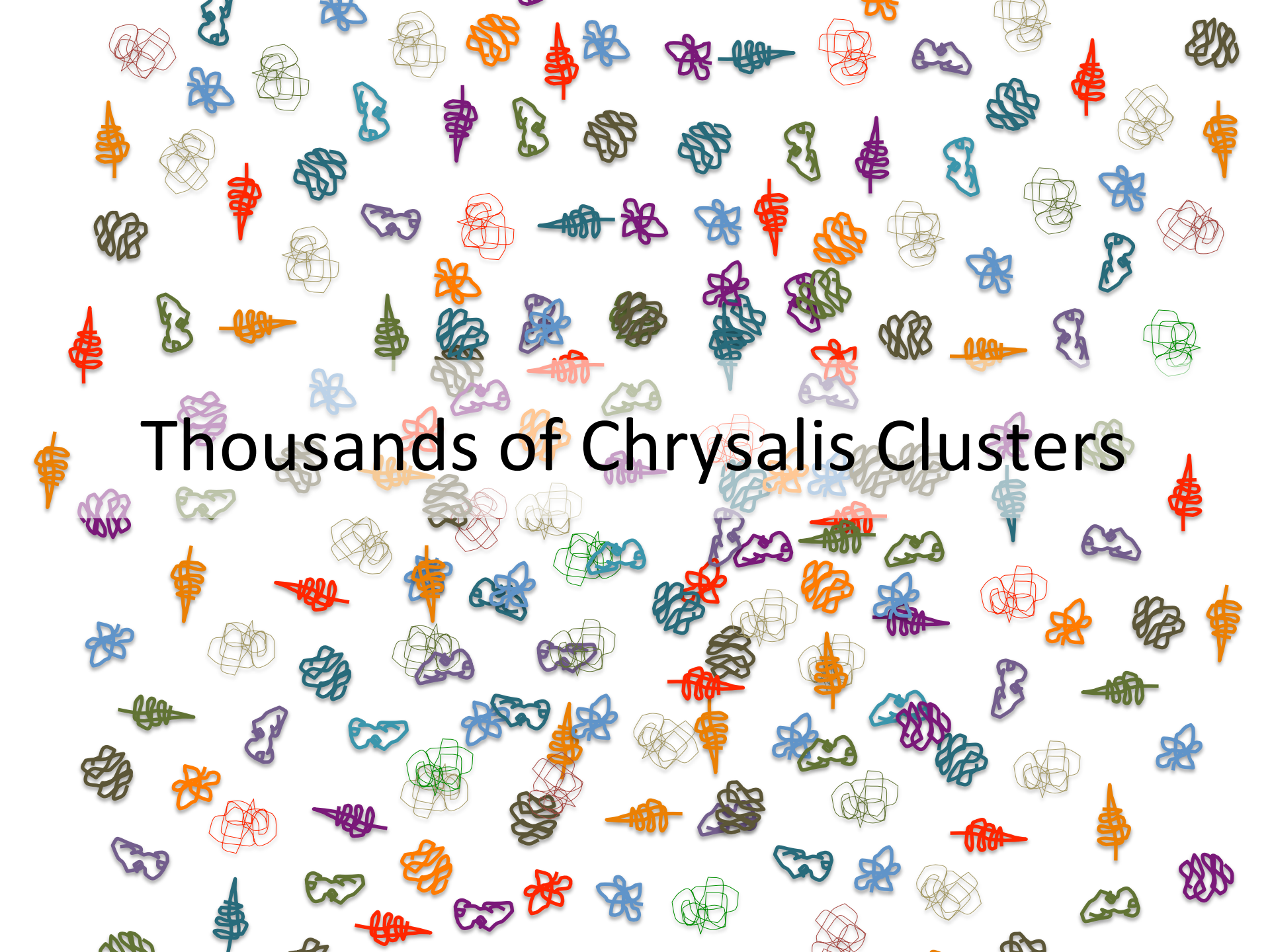
>a121:len=5845
>a122:len=2560
>a123:len=4443
>a124:len=48
>a125:len=8876
>a126:len=66

Integrate isoforms
via k-1 overlaps

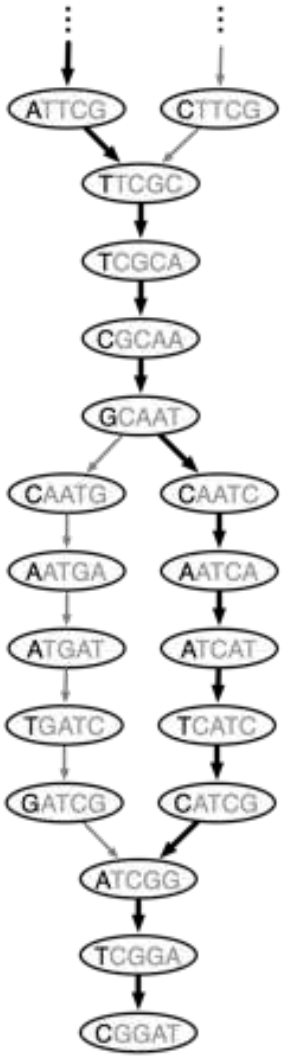
Build de Bruijn Graphs
(ideally, one per gene)





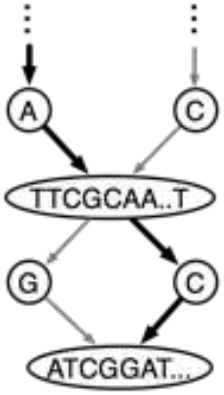

Thousands of Chrysalis Clusters

Butterfly



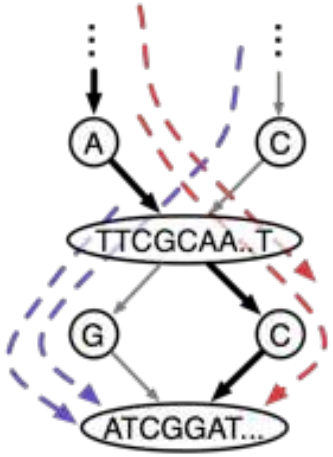
de Bruijn graph

compacting



compact graph

finding paths



compact graph with reads

extracting sequences

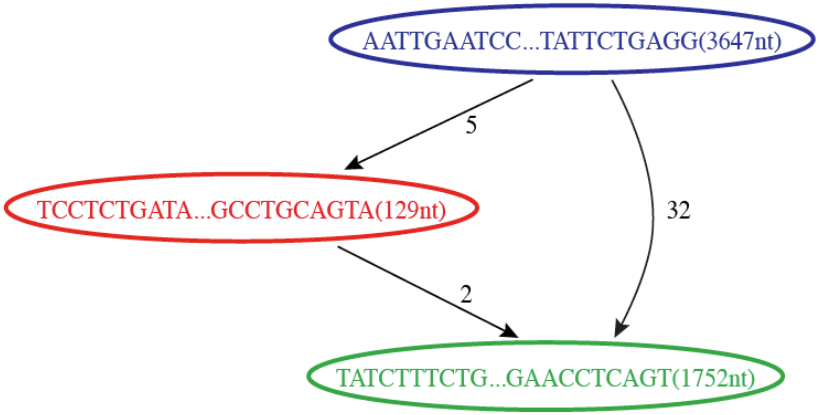
..CTTCGCAA..TGATCGGAT...

..ATTCGCAA..TCATCGGAT...

sequences (isoforms and paralogs)

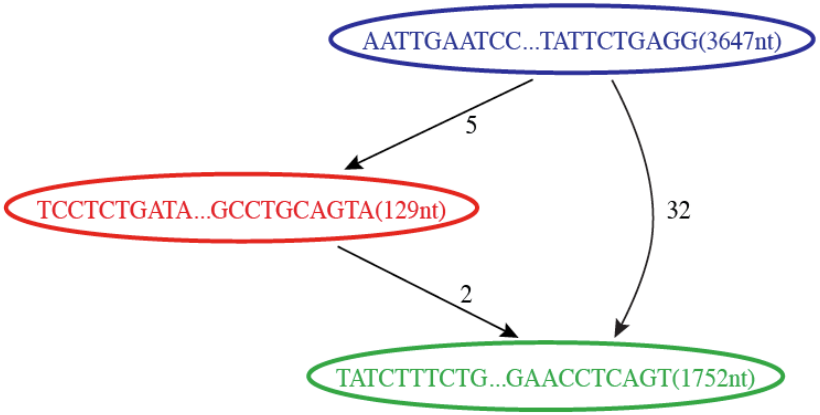
Butterfly Example 1:
Reconstruction of Alternatively Spliced Transcripts

Butterfly's Compacted Sequence Graph



Reconstruction of Alternatively Spliced Transcripts

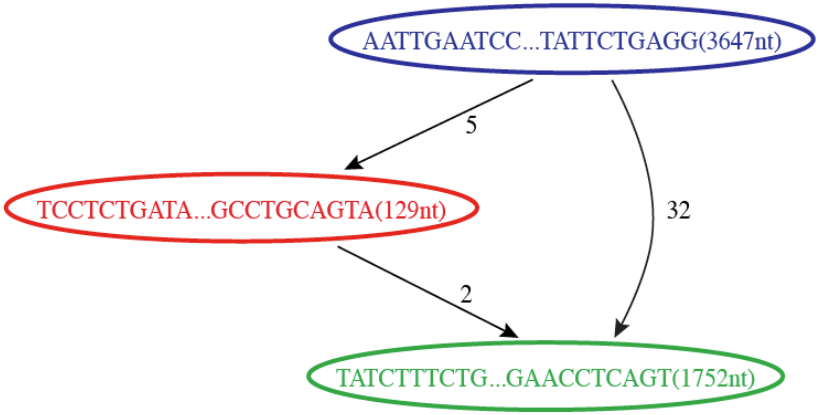
Butterfly's Compacted Sequence Graph



Reconstructed Transcripts

Reconstruction of Alternatively Spliced Transcripts

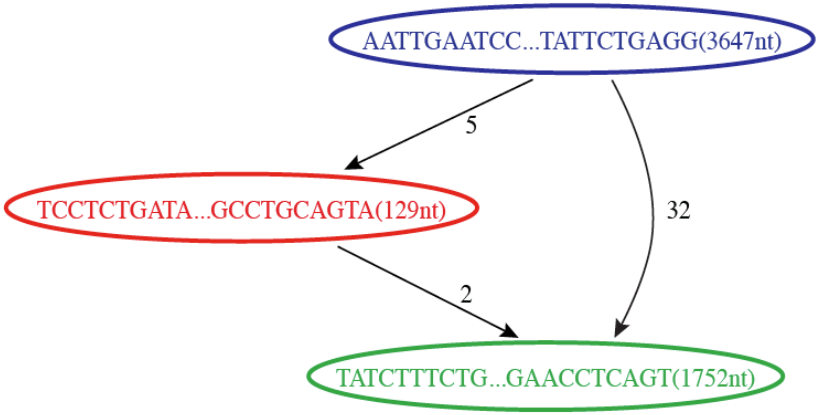
Butterfly's Compacted Sequence Graph



Reconstructed Transcripts

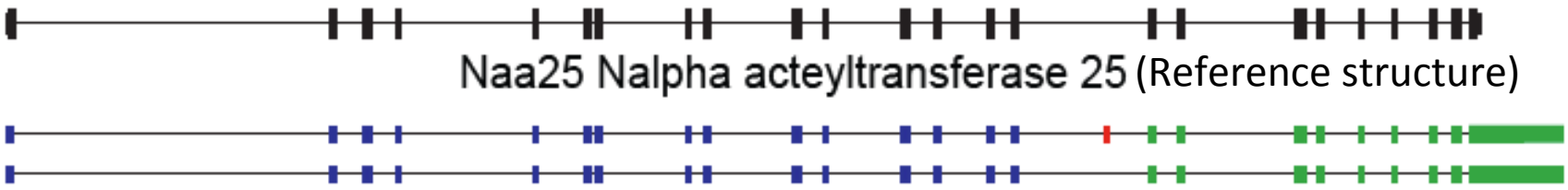
Reconstruction of Alternatively Spliced Transcripts

Butterfly's Compacted Sequence Graph

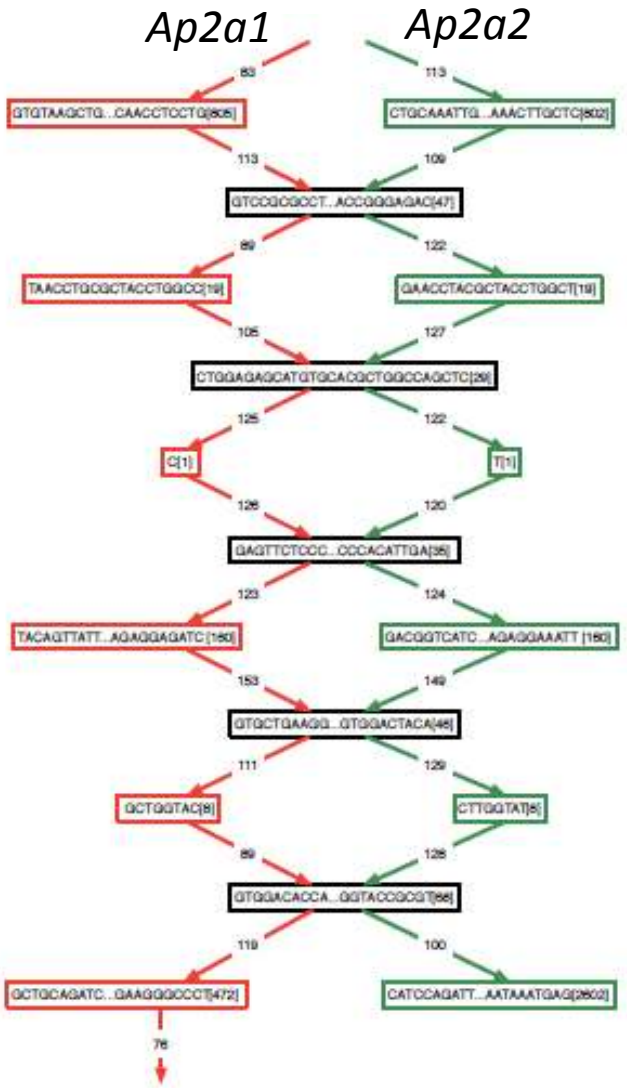


Reconstructed Transcripts

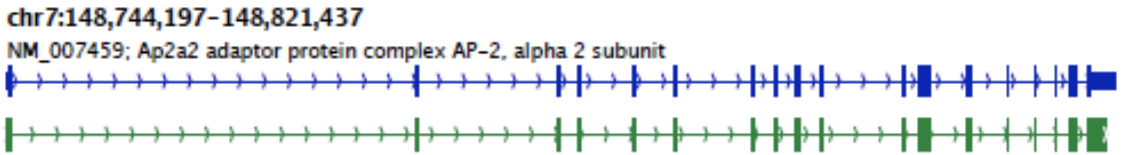
Aligned to Mouse Genome

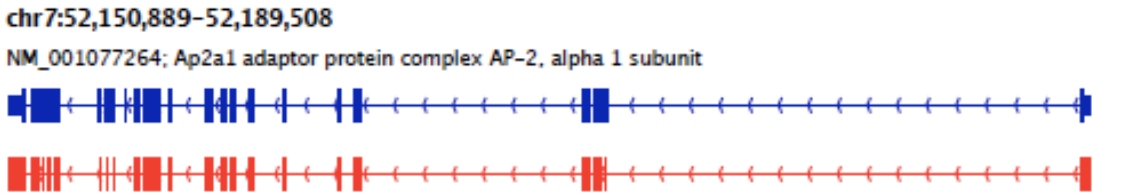


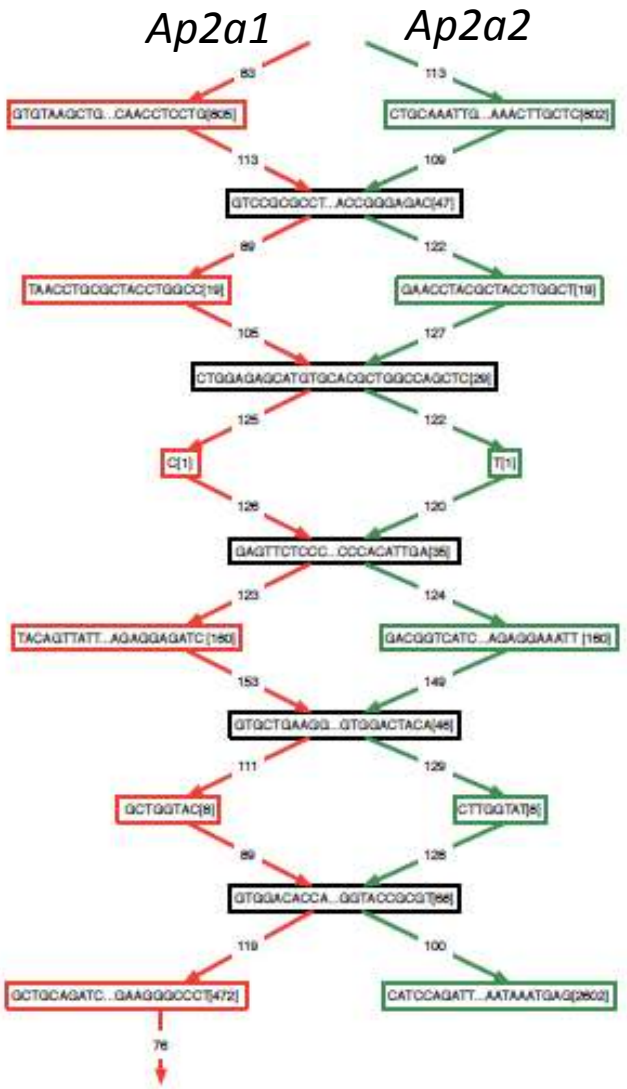
Butterfly Example 2:
Teasing Apart Transcripts of Paralogous Genes



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

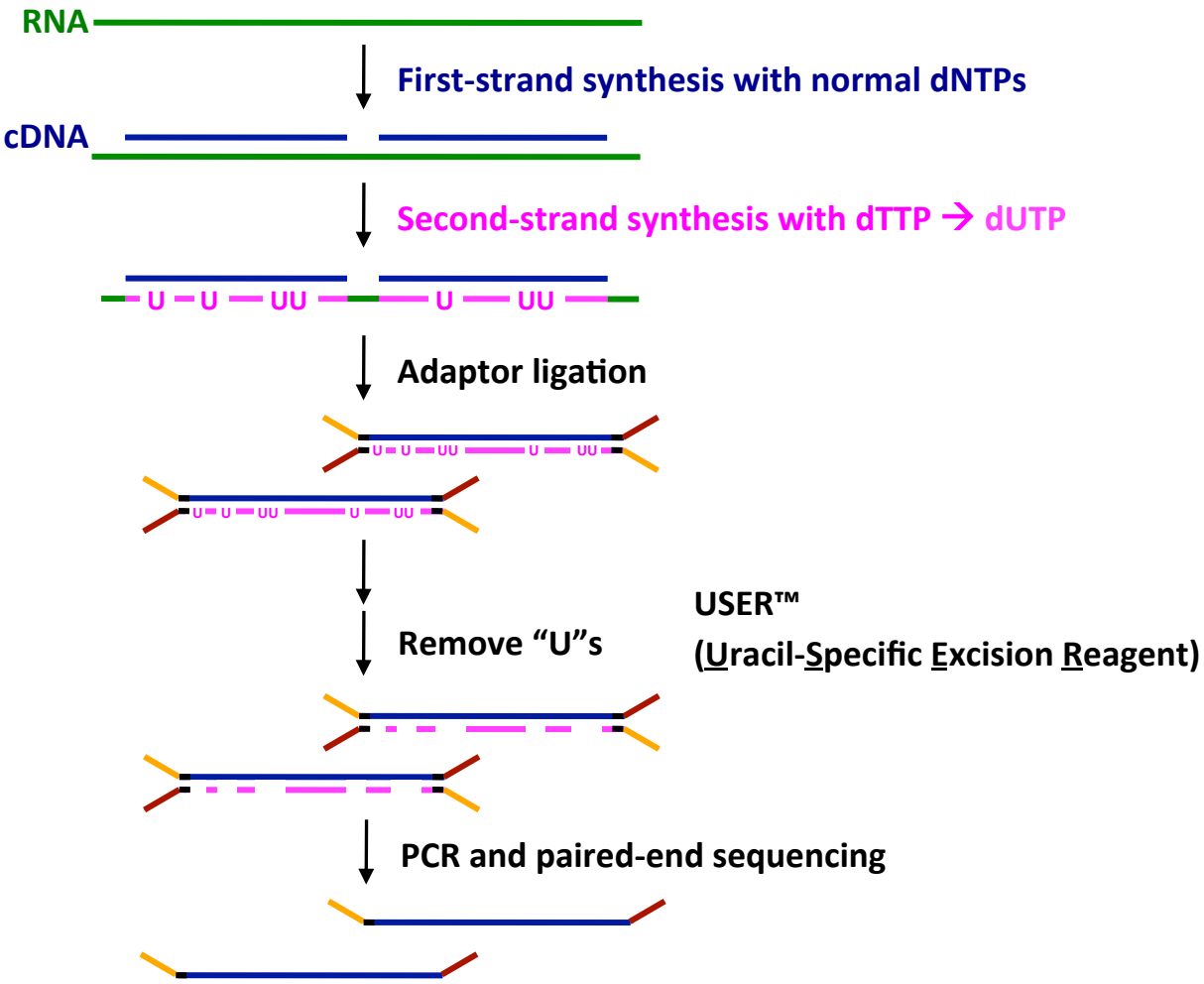
'dUTP second strand marking' identified as the leading protocol

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

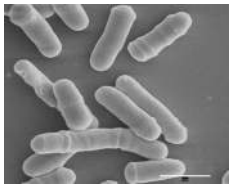
dUTP 2nd Strand Method: Our Favorite



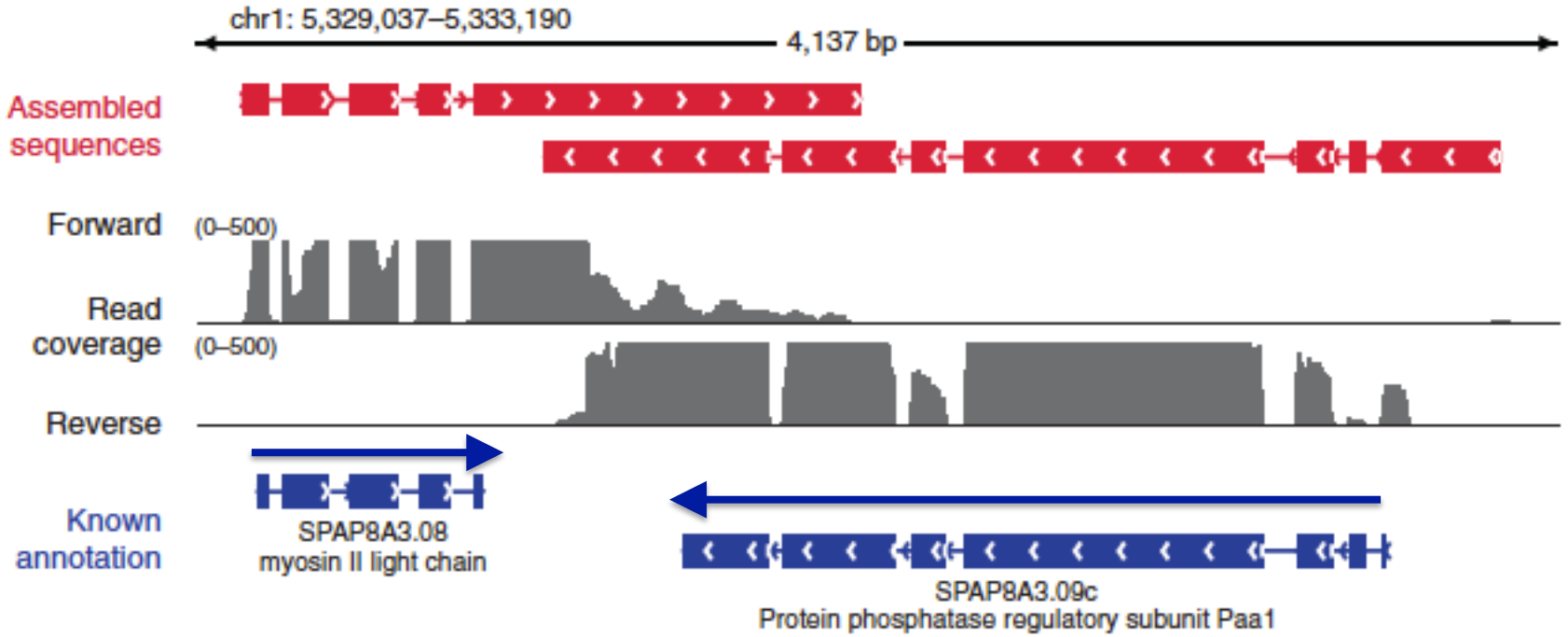
Modified from Parkhomchuk *et al.* (2009) *Nucleic Acids Res.* 37:e123

Slide courtesy of Joshua Levin, Broad Institute.

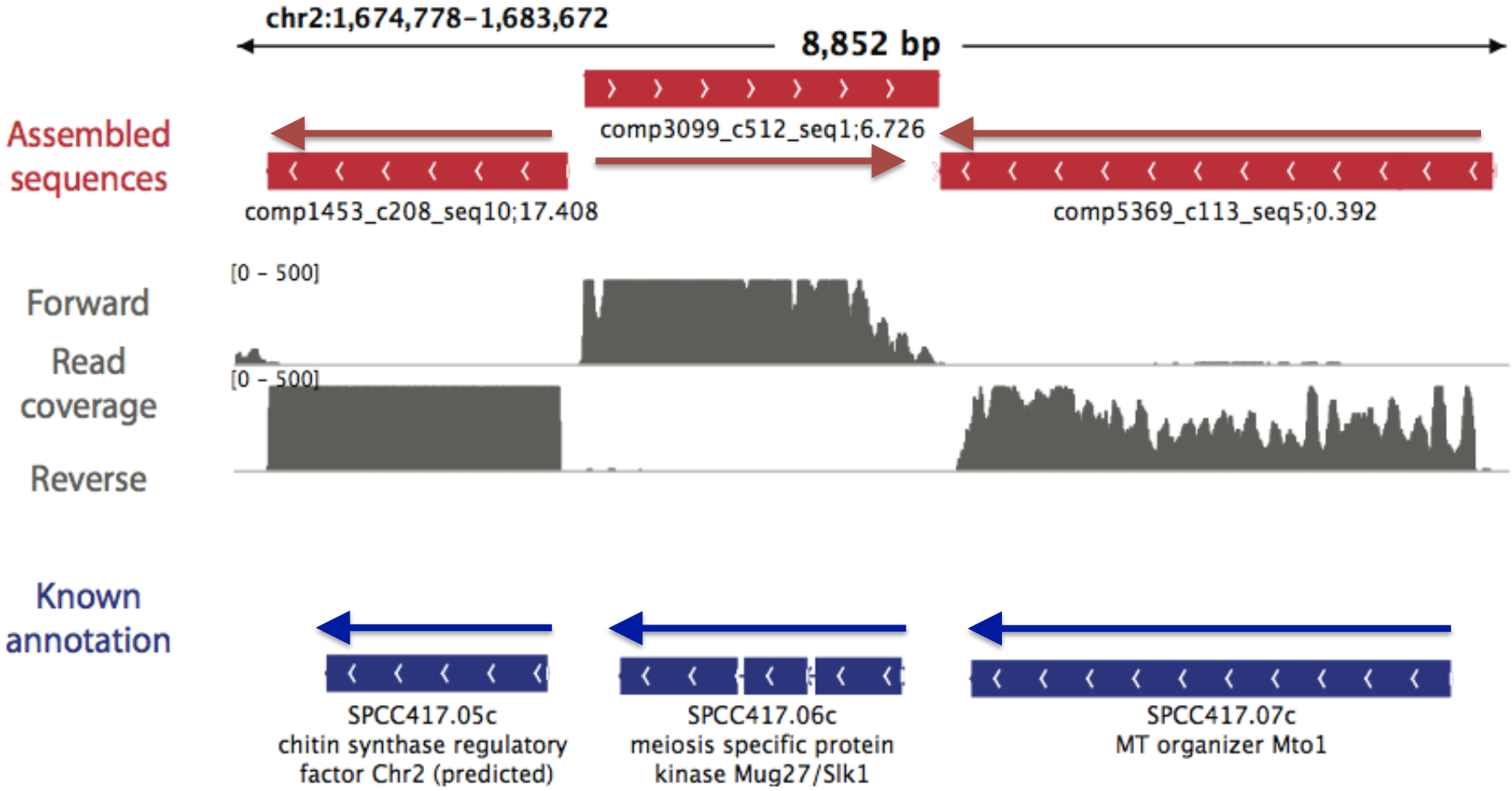
Overlapping UTRs from Opposite Strands



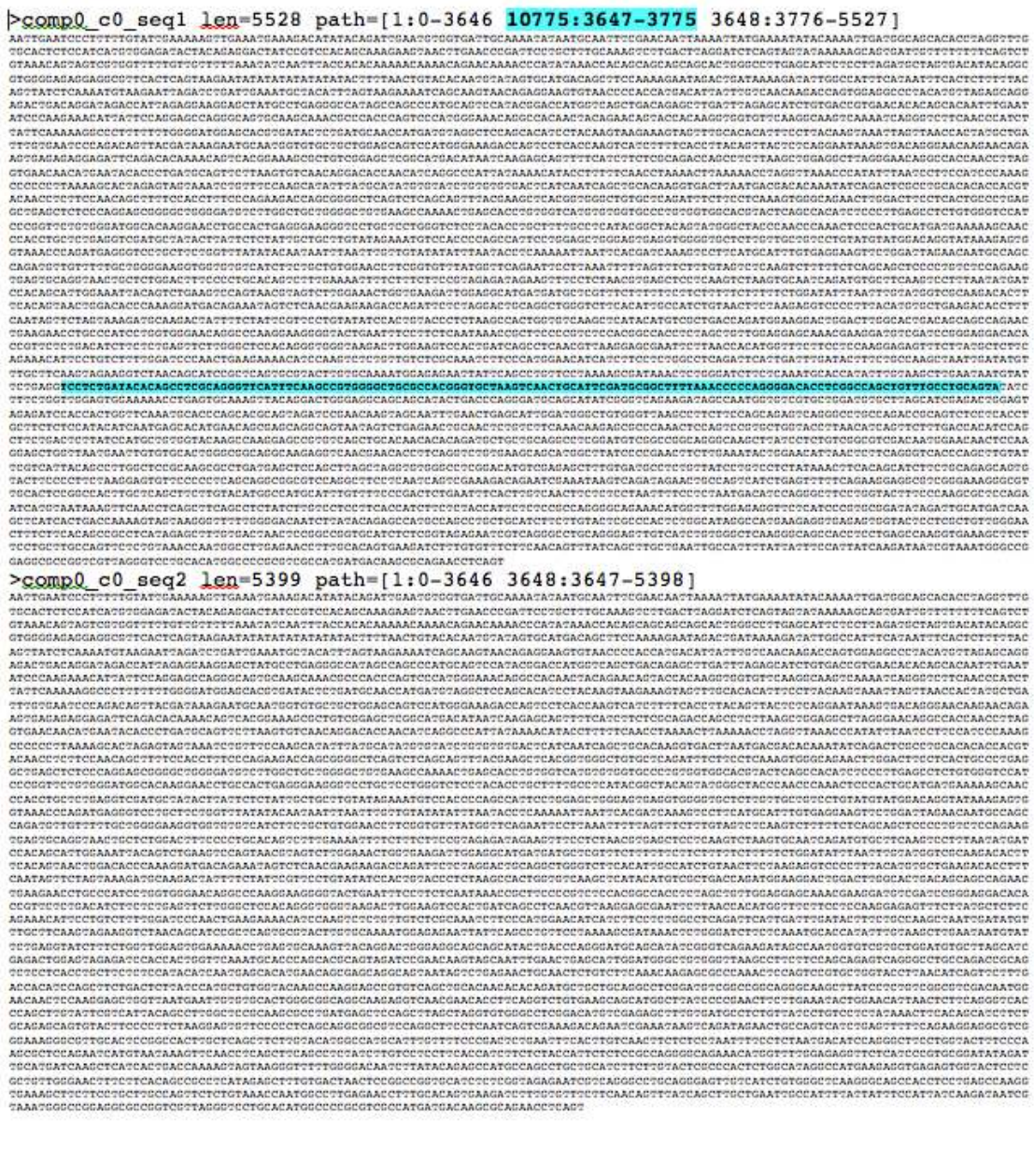
Schizosacharomyces pombe
(fission yeast)



Antisense-dominated Transcription

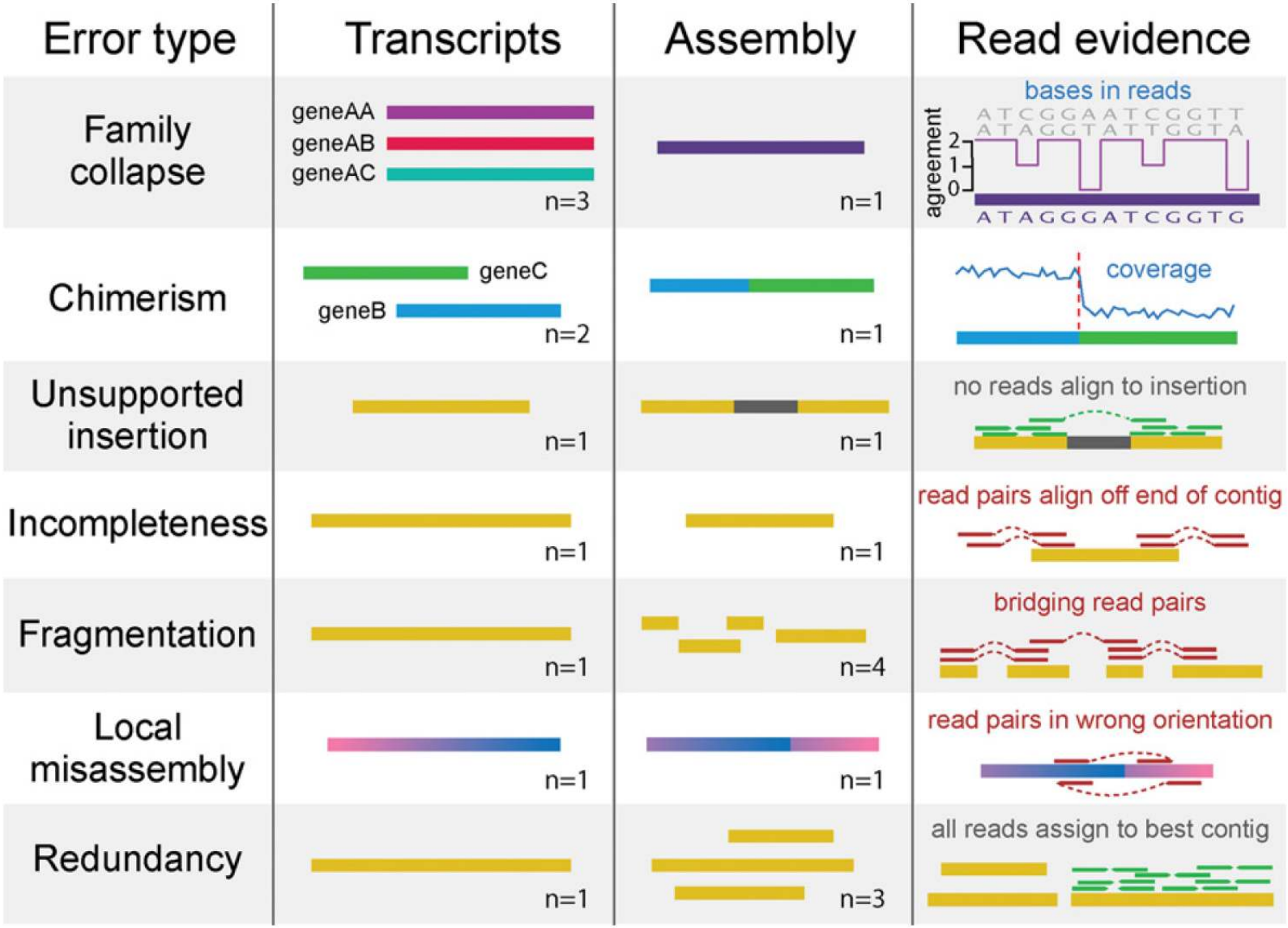


Trinity output: A multi-fasta file



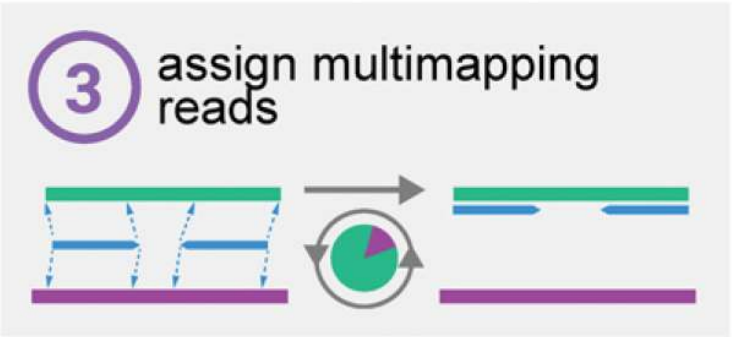
Evaluating the quality of your transcriptome assembly

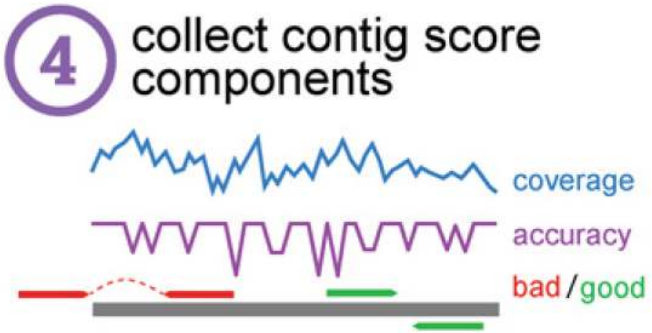
De novo Transcriptome Assembly is Prone to Certain Types of Errors

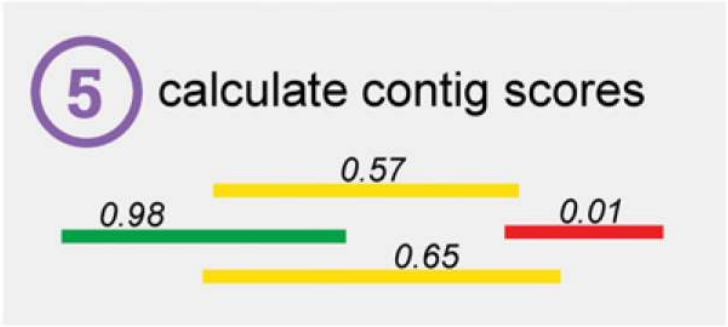


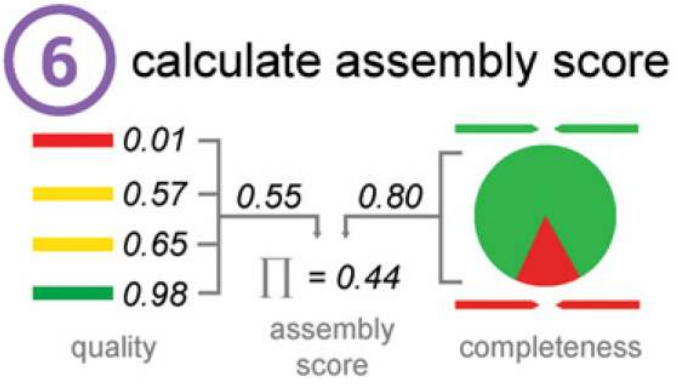
Smith-Unna et al. Genome Research, 2016

TransRate









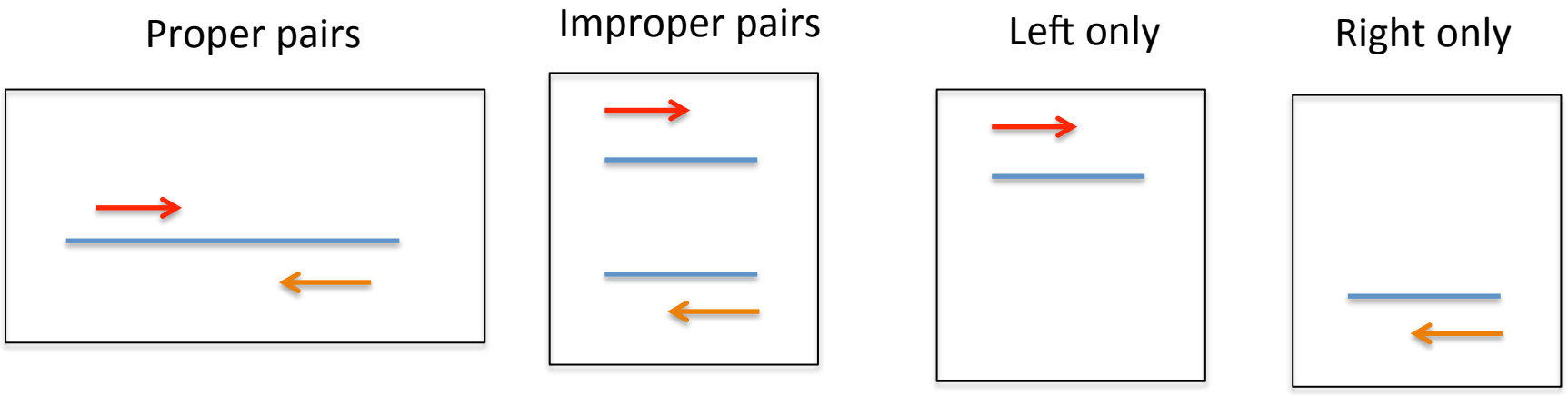
Smith-Unna et al. Genome Research, 2016

Simple Quantitative and Qualitative Assembly Metrics

Read representation by assembly

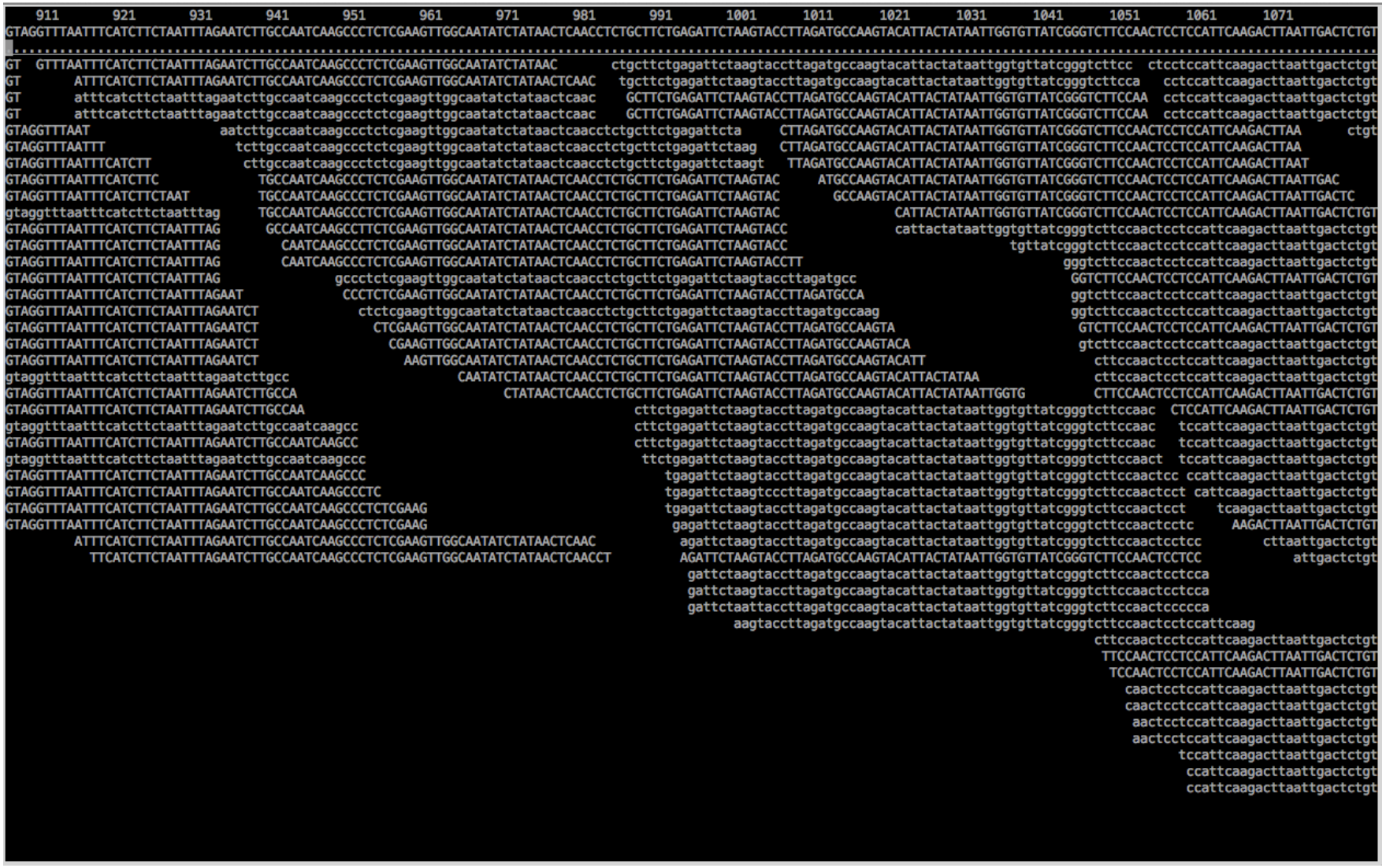
Align reads to the assembled transcripts using Bowtie.
A typical 'good' assembly has ~80 % reads mapping to the assembly and ~80% are properly paired.

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:

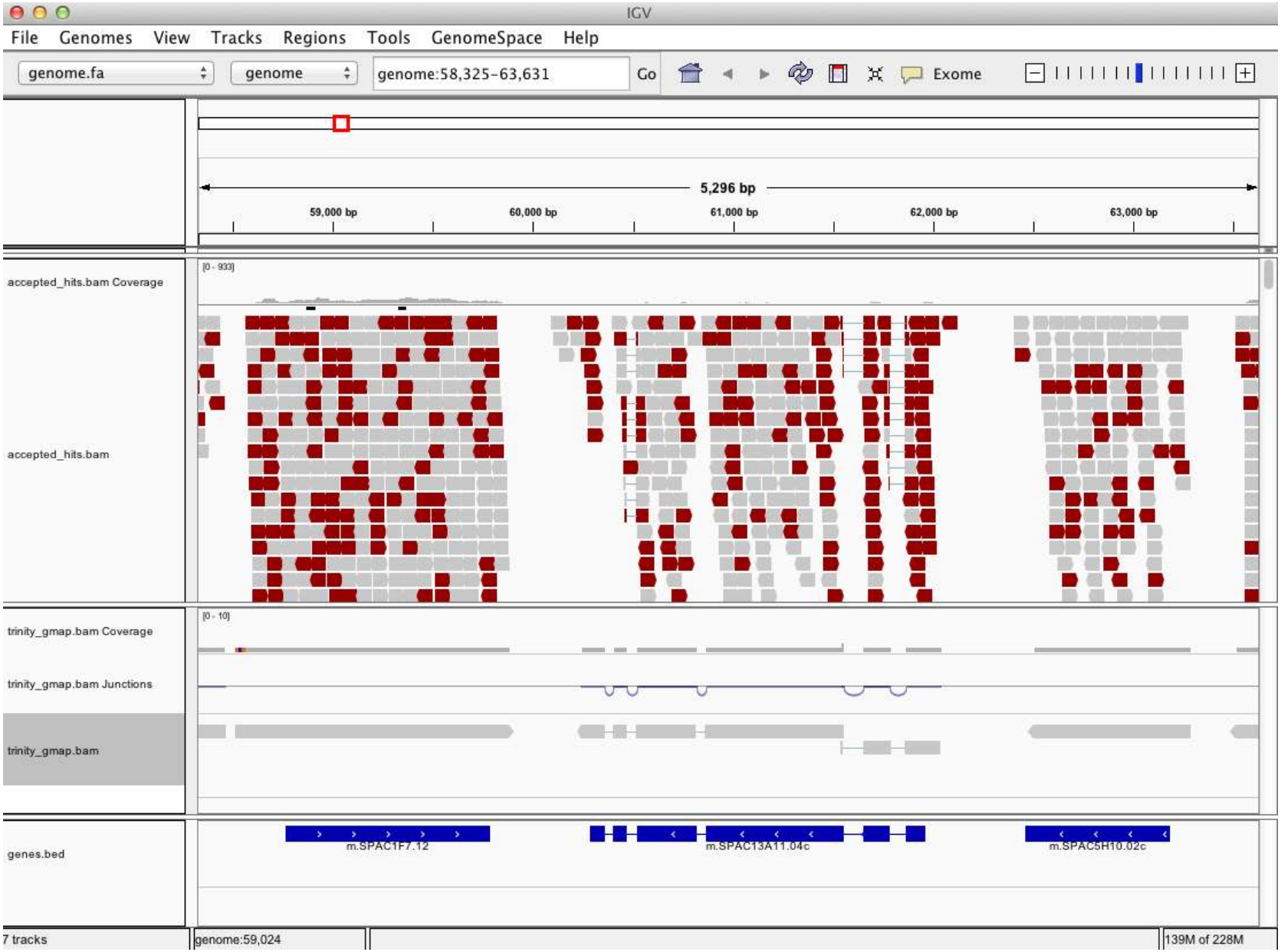
James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, Jill P. Mesirov. **Integrative Genomics Viewer**. Nature Biotechnology 29, 24–26 (2011), or

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

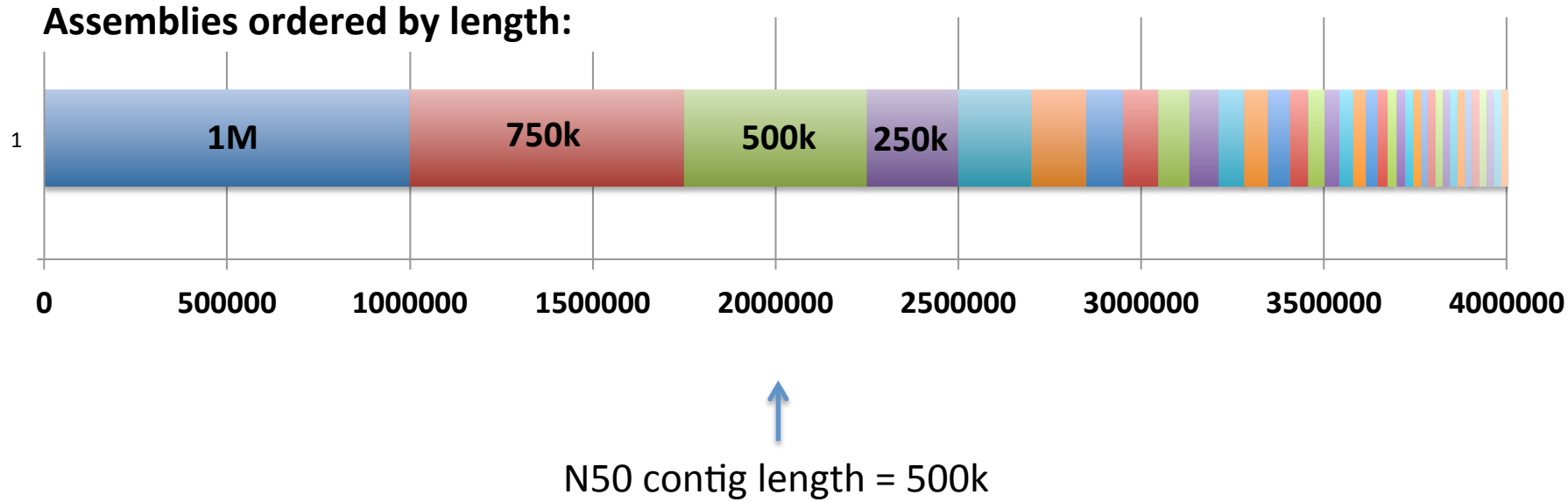
(Trinity transcripts aligned to the genome using GMAP)



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'

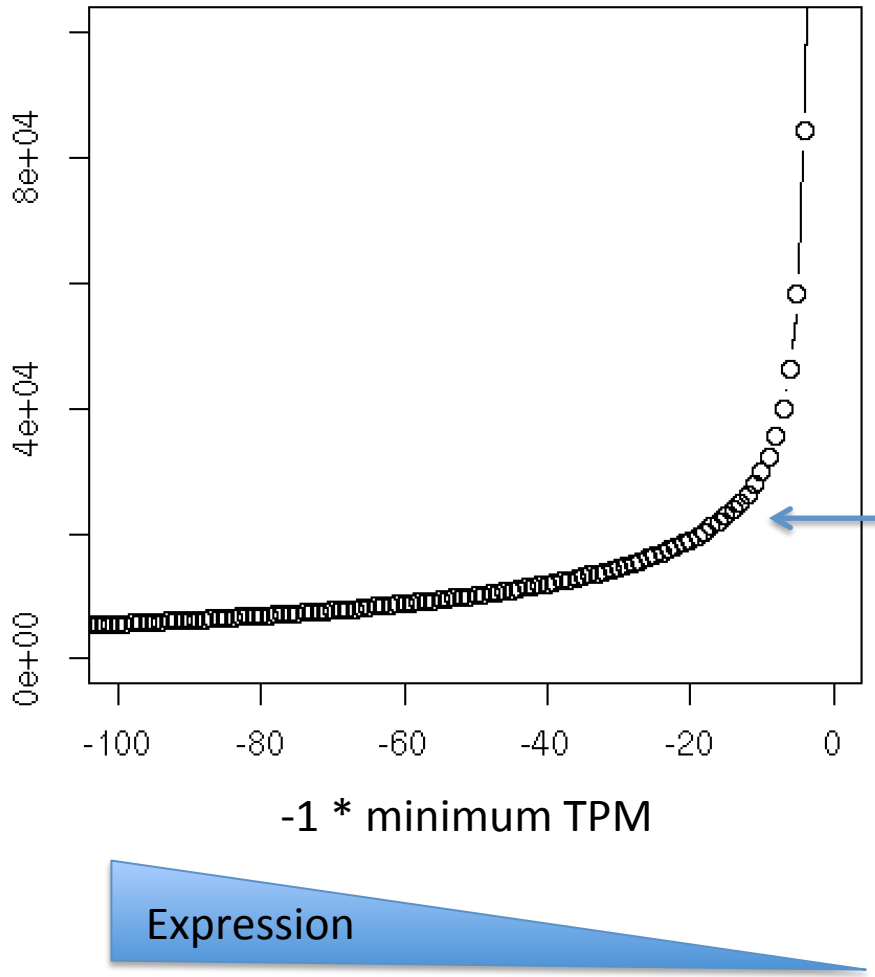


Often, most assembled transcripts are *very* lowly expressed

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

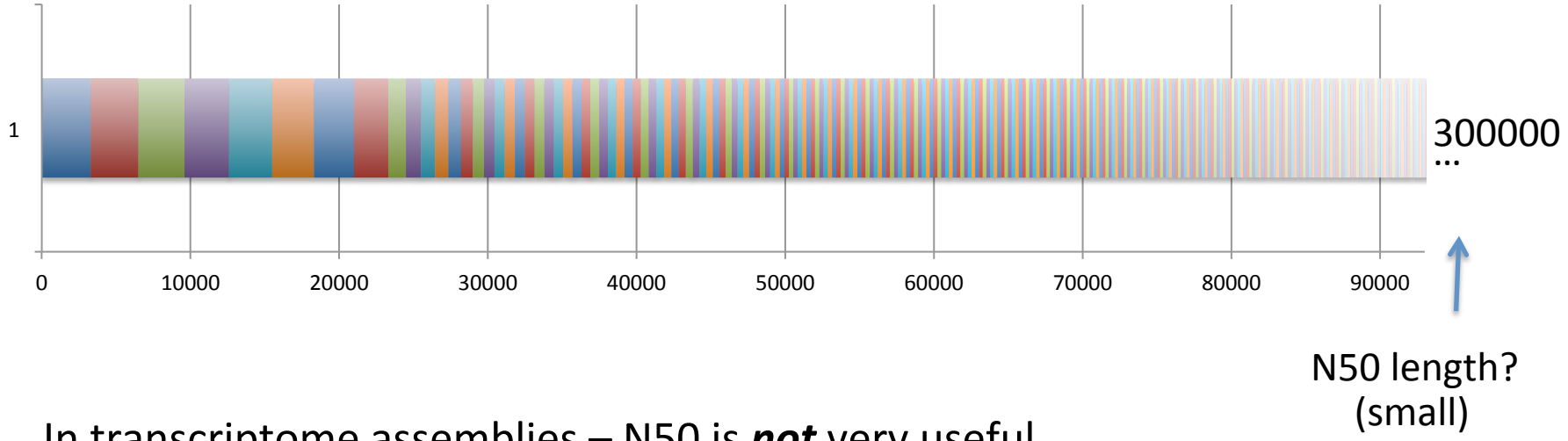
1.4 million Trinity transcript contigs

N50 ~ 500 bases



* Salamander transcriptome

N50 Calculation for *Transcriptome* Assemblies??

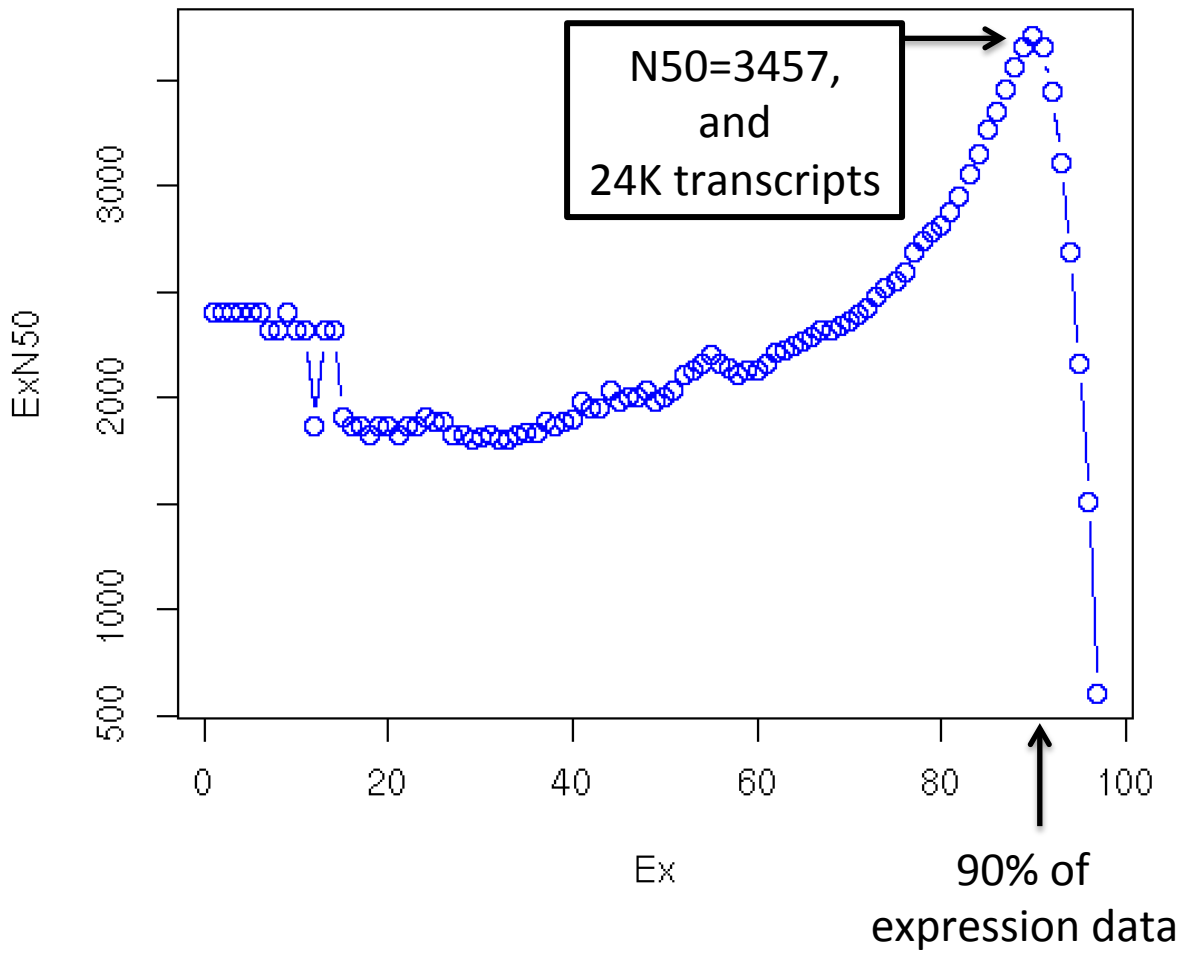


In transcriptome assemblies – N50 is ***not*** very useful.

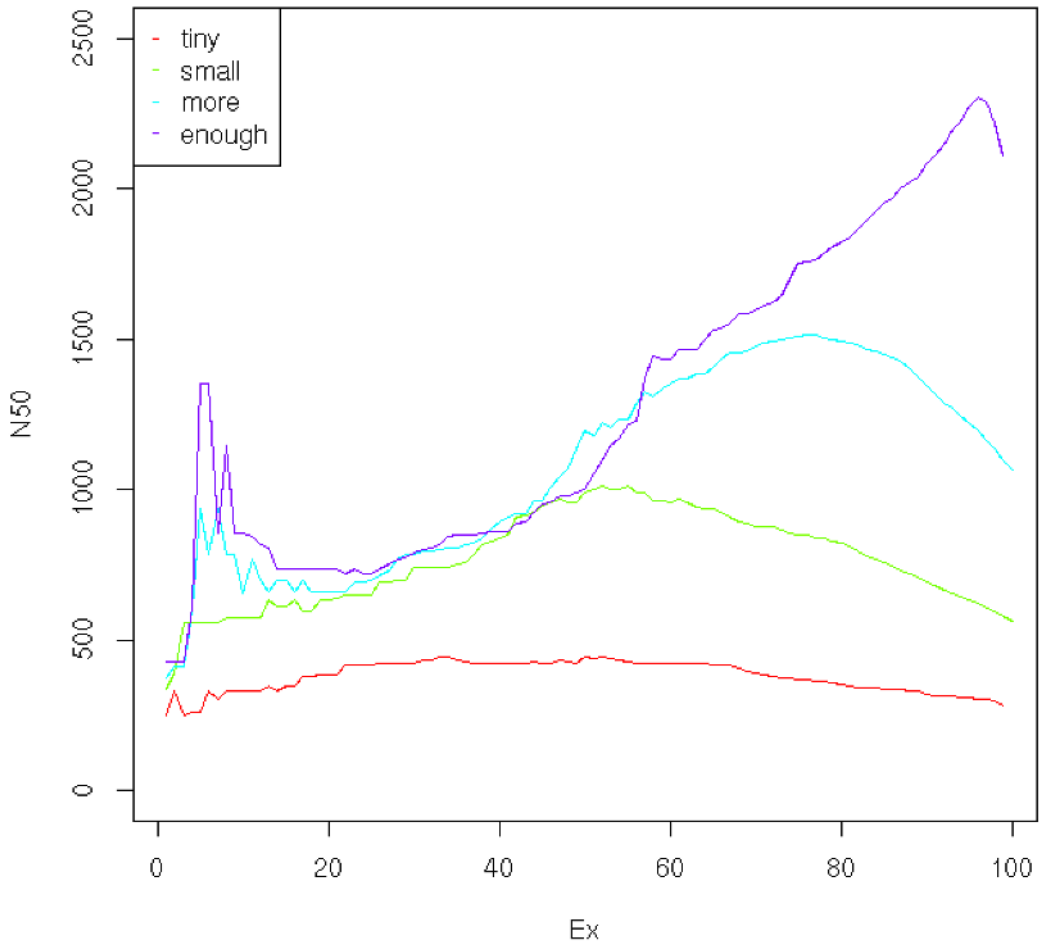
- Overzealous isoform annotation for long transcripts drives higher N50
- Very sensitive reconstruction for short lowly expressed transcripts drives lower N50

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

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



ExN50 Profiles for Different Trinity Assemblies Using Different Read Depths



Millions of Reads

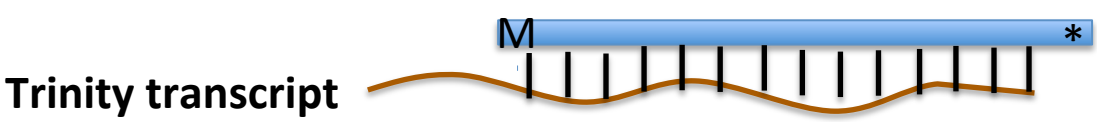
Thousands of Reads

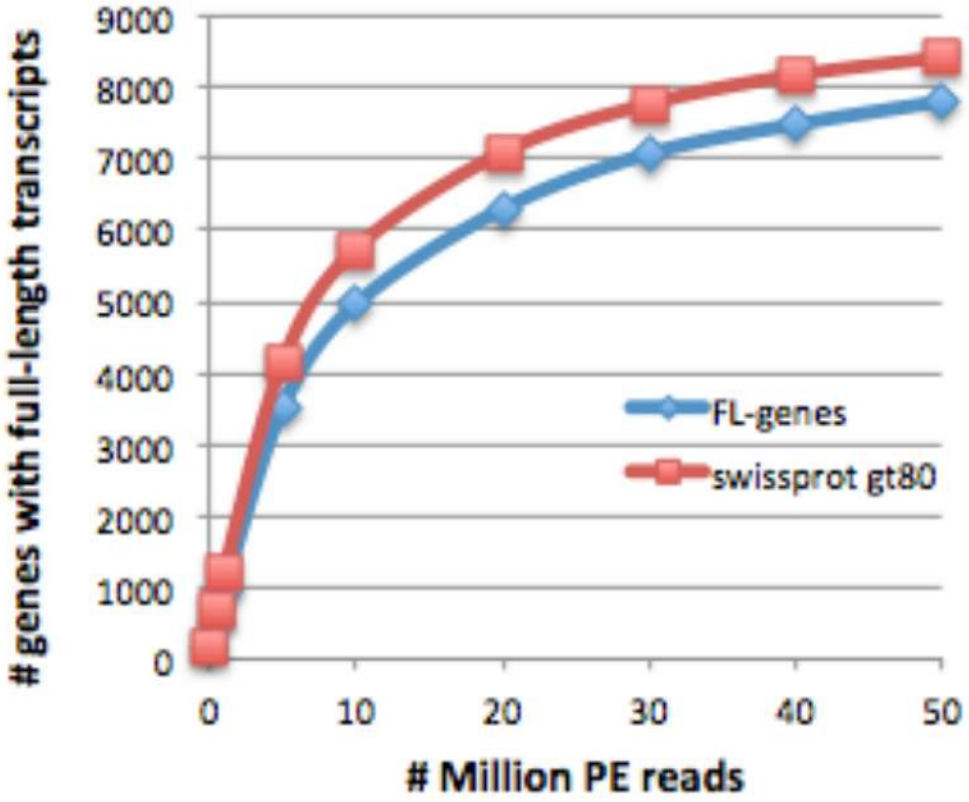
Note shift in ExN50 profiles as you assemble more and more reads.

* 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*

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*

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

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

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

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

Detonate: Which assembly is better?

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

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

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

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

Li et al. **Evaluation of de novo transcriptome assemblies from RNA-Seq data**, Genome Biology 2014

Detonate: Which assembly is better?

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

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

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

$$\log P(A,D) = \log \int_{\Lambda} P(D|A,\Lambda)P(A|\Lambda)P(\Lambda)d\Lambda$$

$$\approx \underbrace{\log P(D|A,\Lambda_{\text{MLE}})}_{\text{likelihood}} + \underbrace{\log P(A|\Lambda_{\text{MLE}})}_{\text{assembly prior}}$$

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

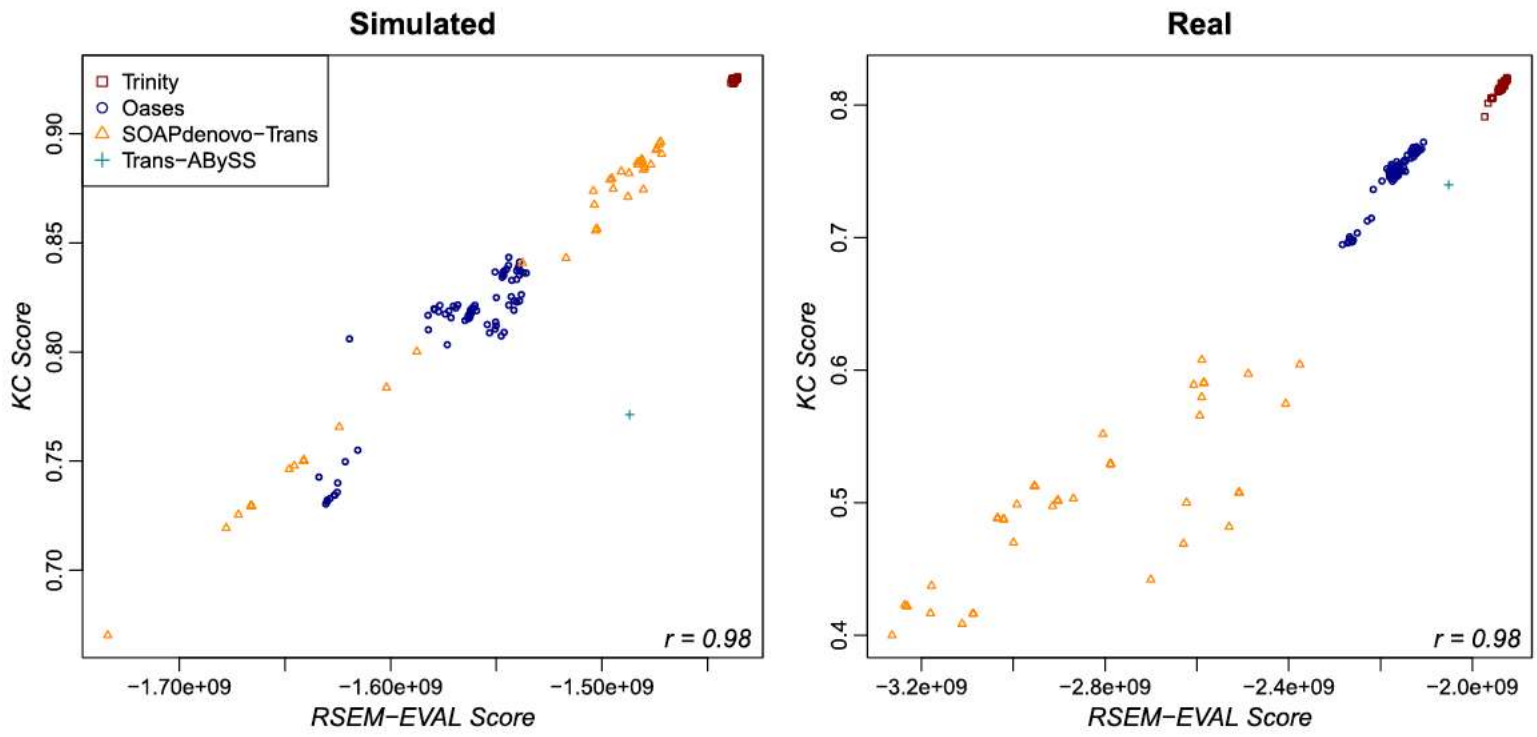
Bigger Score = Better Assembly

Li et al. **Evaluation of de novo transcriptome assemblies from RNA-Seq data**, Genome Biology 2014

Detonate: Which assembly is better?

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

Yay Trinity!!



Li et al. **Evaluation of de novo transcriptome assemblies from RNA-Seq data**, Genome Biology 2014

nature protocols

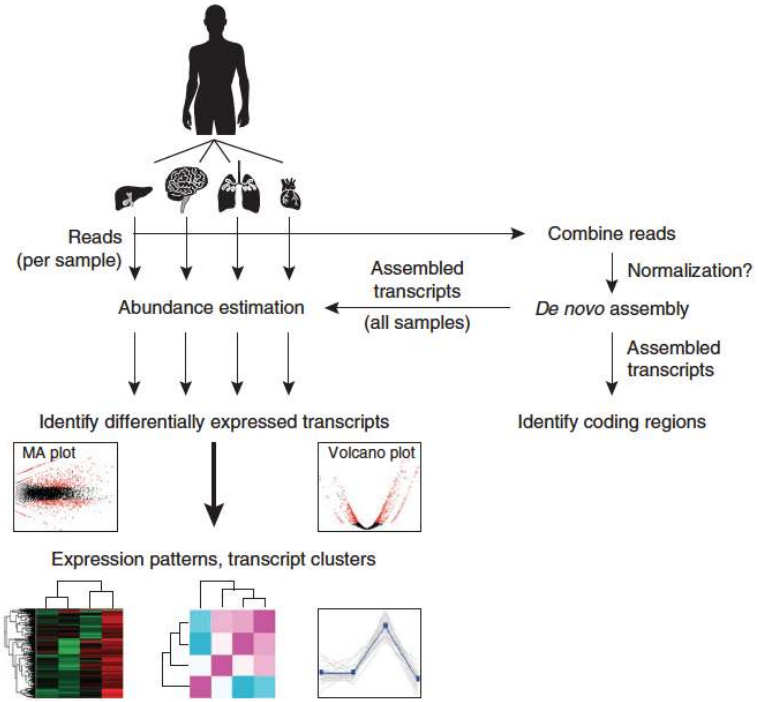
NATURE PROTOCOLS | PROTOCOL

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

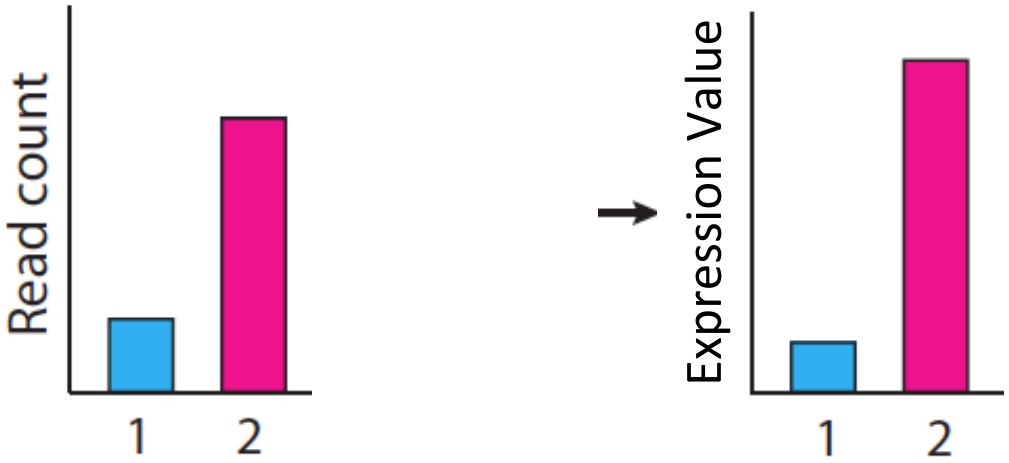
Affiliations | Contributions | Corresponding authors

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



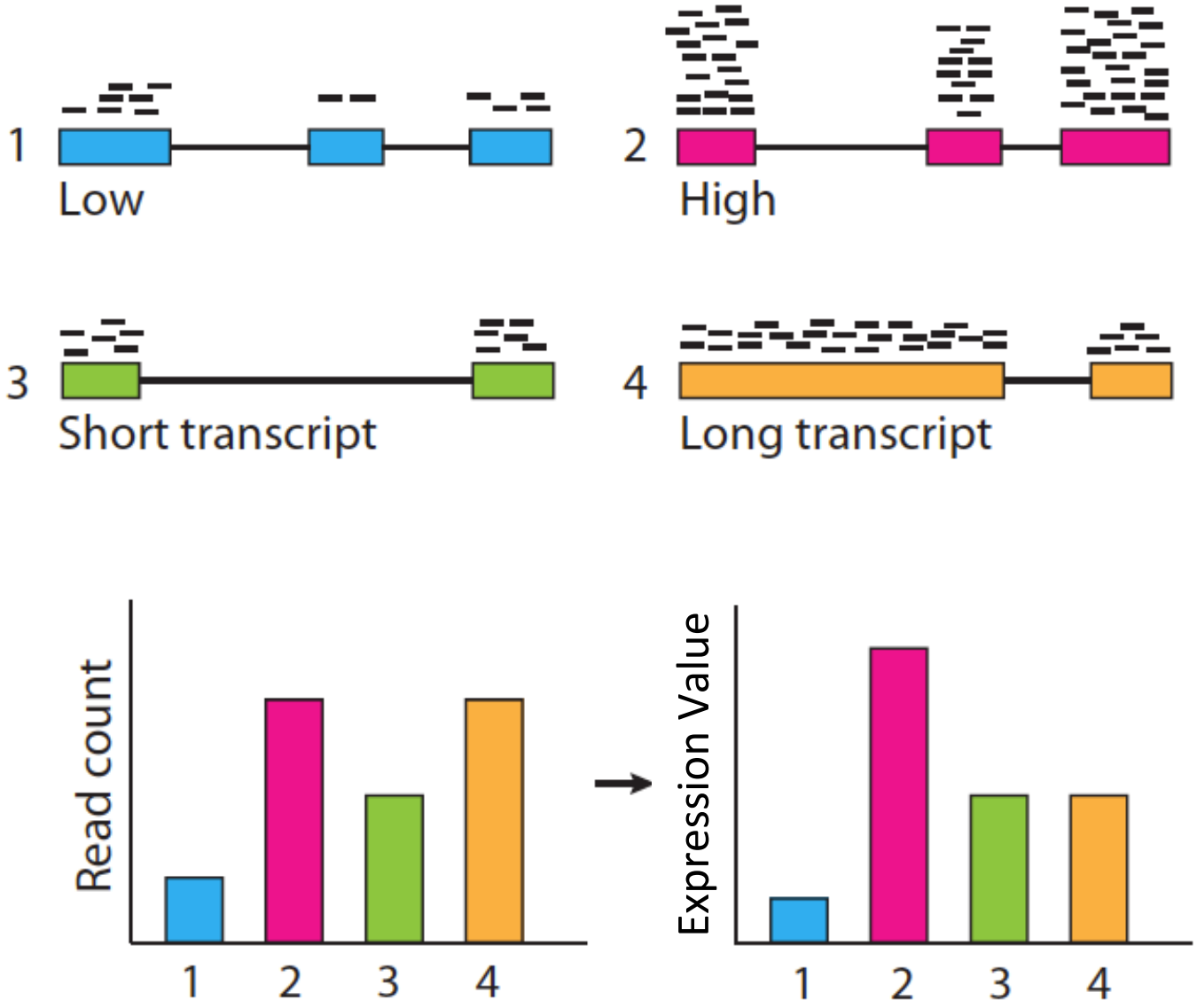
Abundance Estimation
(Aka. Computing Expression Values)

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)

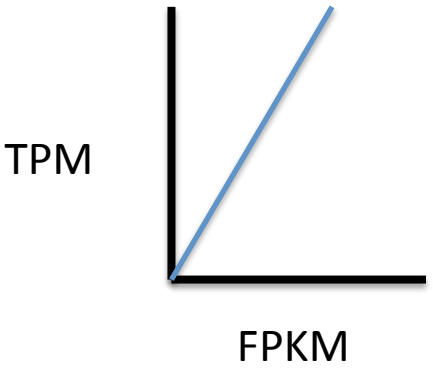
$$TPM_i = \frac{FPKM_i}{\sum_j FPKM} * 1e6$$

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.



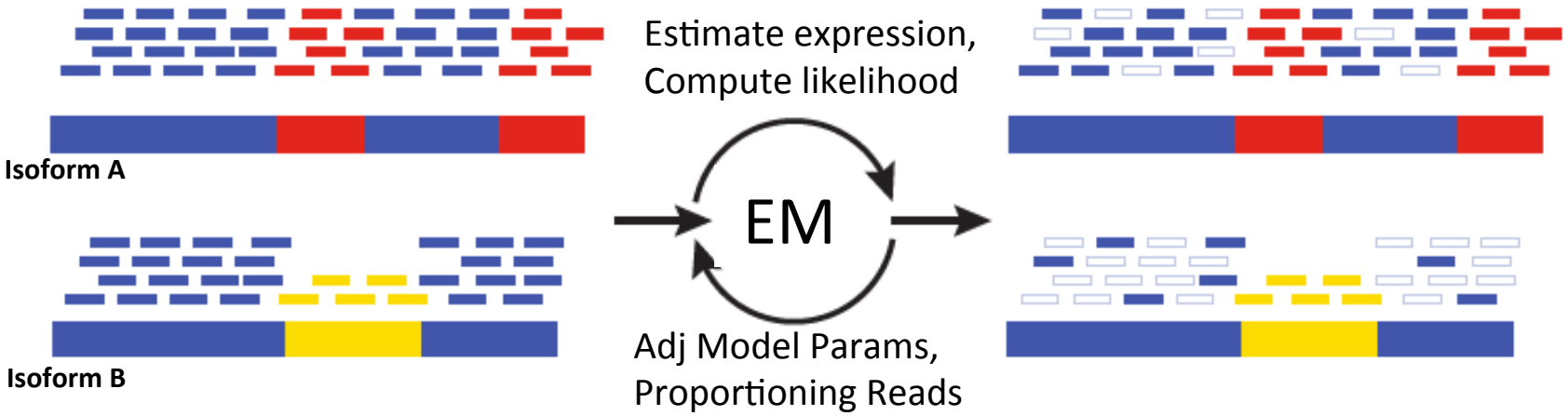
Multiply-mapped Reads Confound Abundance Estimation

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

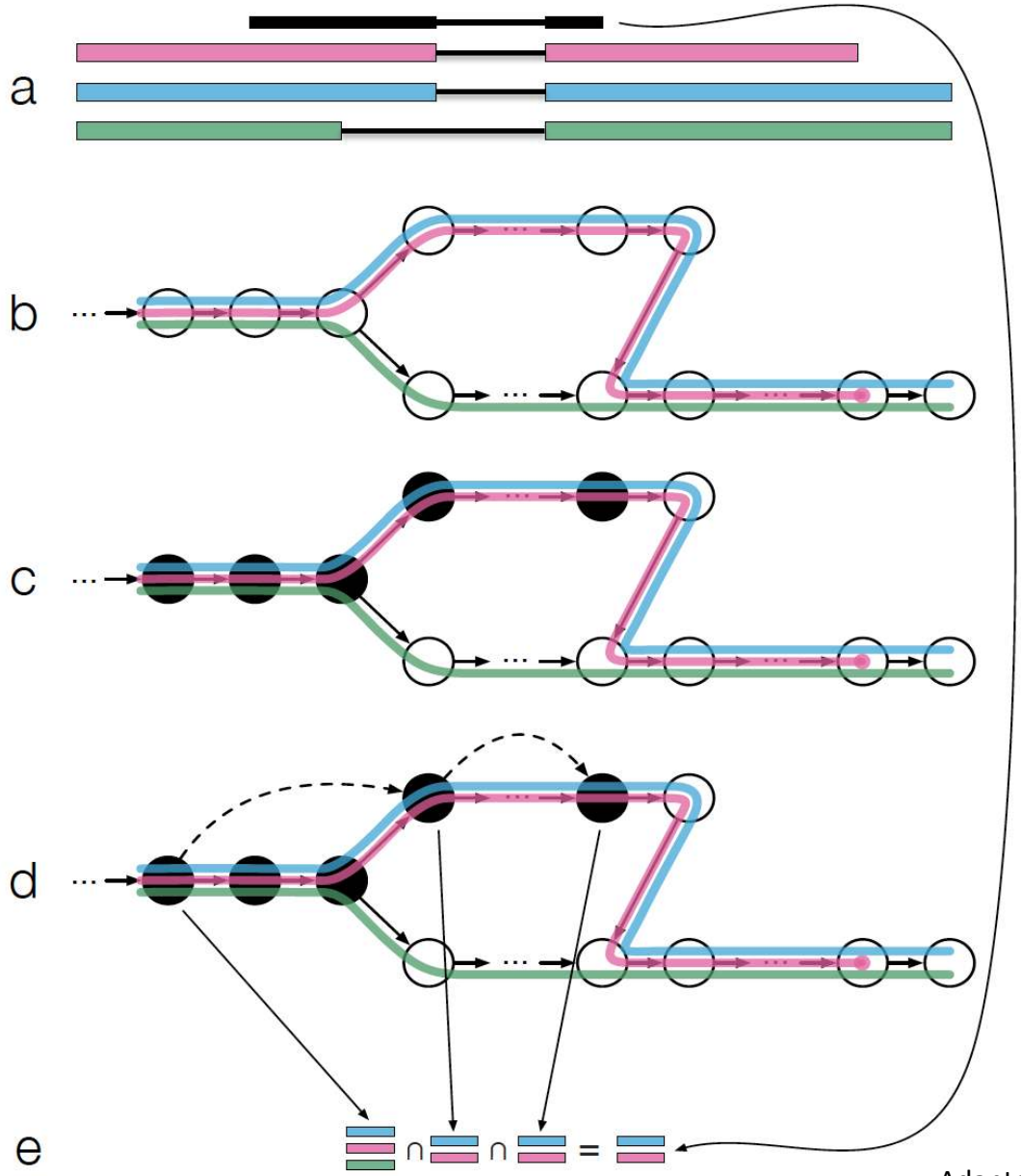
Use Expectation Maximization (EM) to find the most likely assignment of reads to transcripts.

Performed by:

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

Fast Abundance Estimation Using Pseudo-alignments and Equivalence Classes

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



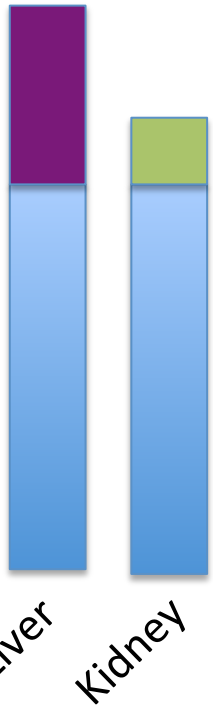
Adapted from Fig 1 from Bray et al.

Comparing RNA-Seq Samples

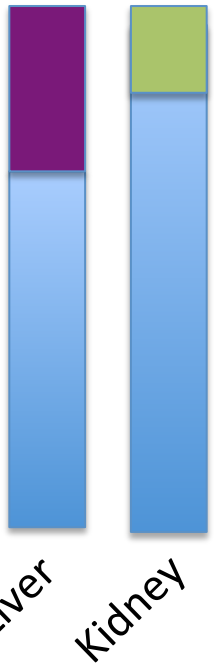
Some Cross-sample Normalization May Be Required

Why cross-sample normalization is important

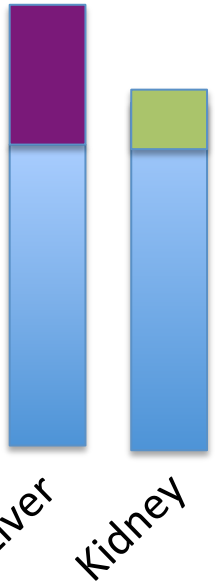
Absolute RNA quantities per cell



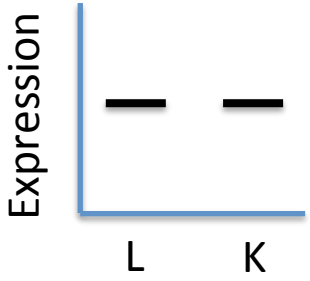
Measured relative abundance via RNA-Seq

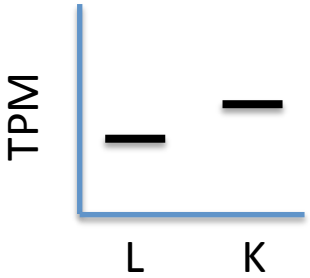


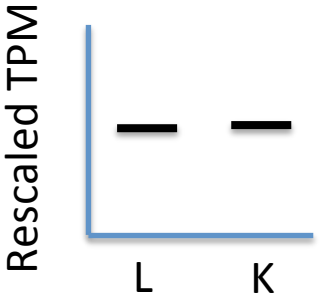
Cross-sample normalized (rescaled) relative abundance



eg. Some housekeeping gene's expression level:







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

Subset of genes highly expressed in liver

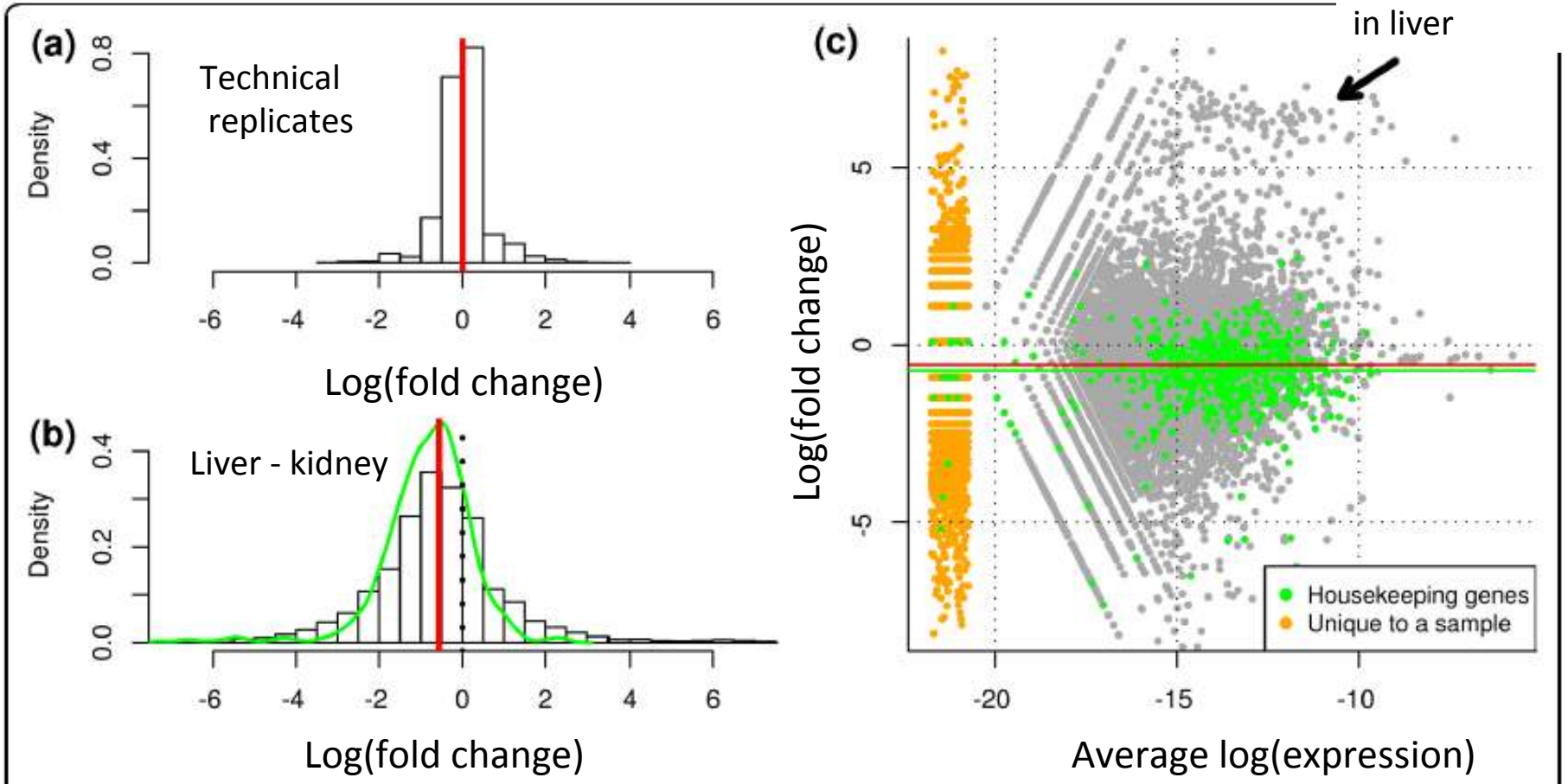
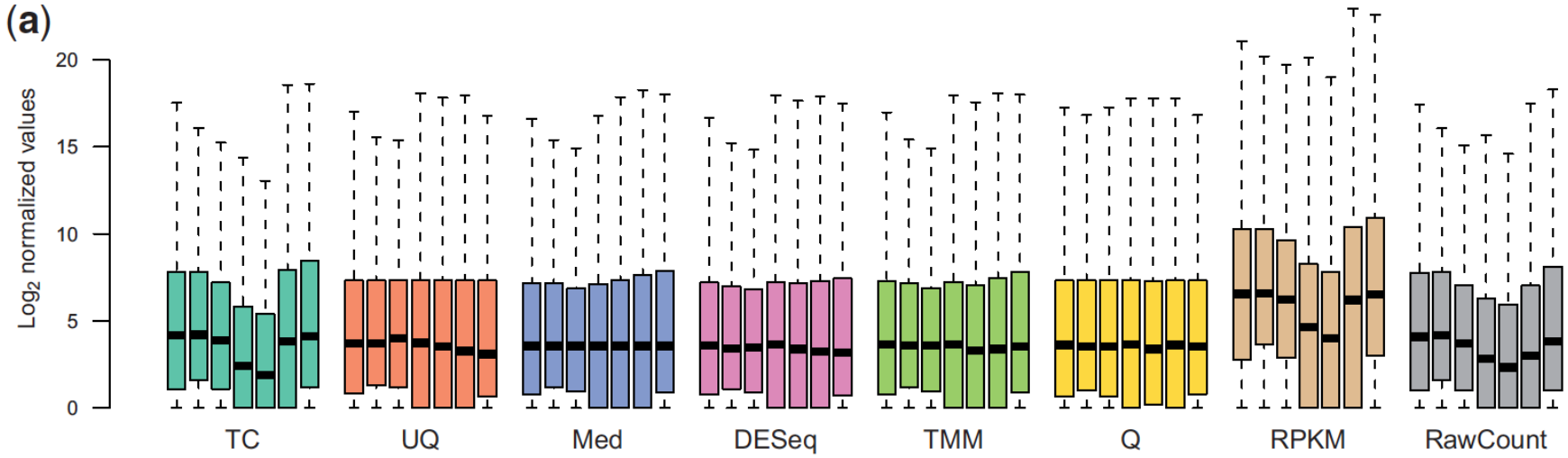


Figure 1 Normalization is required for RNA-seq data. Data from [6] comparing log ratios of **(a)** technical replicates and **(b)** liver versus kidney expression levels, after adjusting for the total number of reads in each sample. The green line shows the smoothed distribution of log-fold-changes of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

the overall bias in log-fold-changes.

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

Differential Expression Analysis

Thx, Charlotte Soneson! ☺

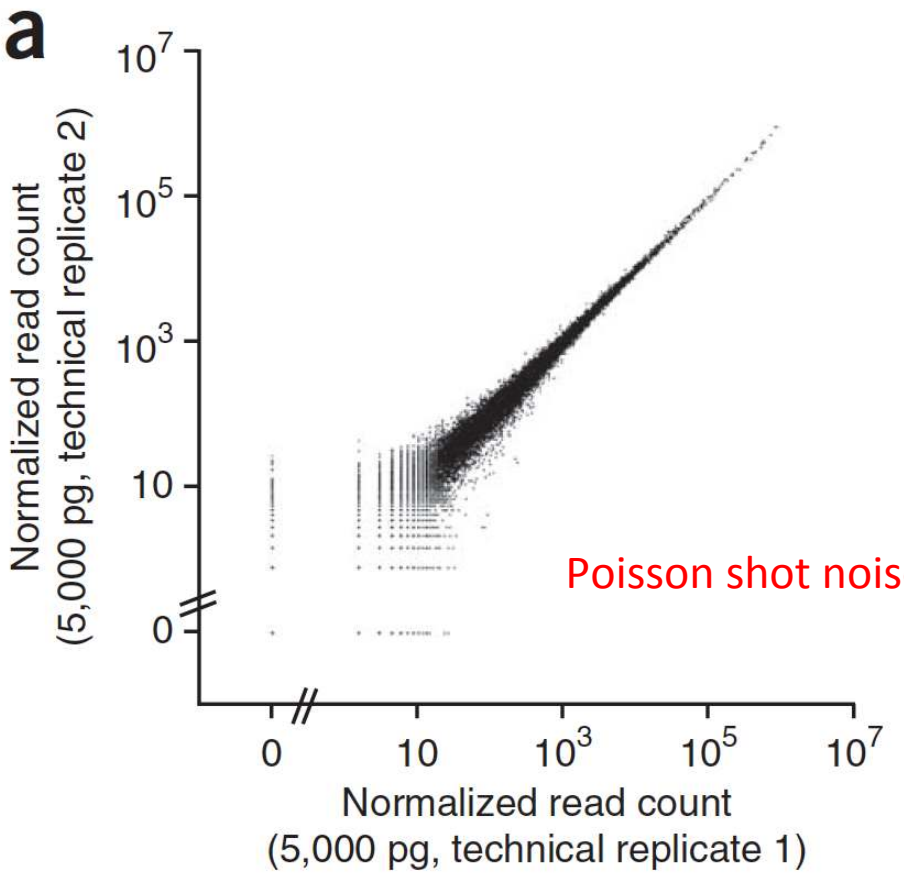
Differential Expression Analysis Involves

- 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



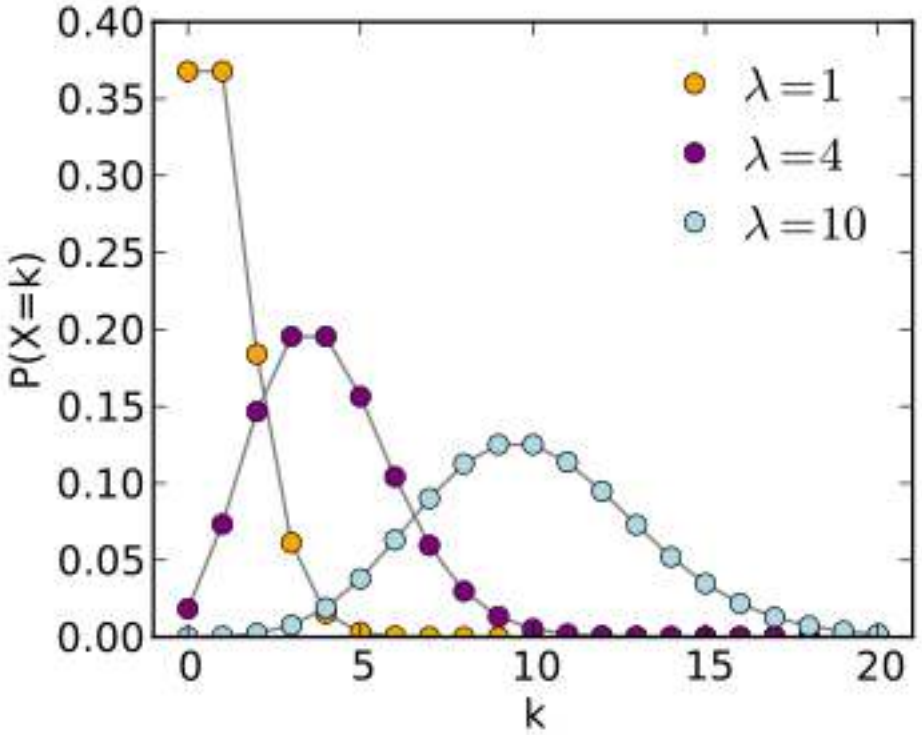
Variation observed is well described by models of random sampling (Poisson Distribution)

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

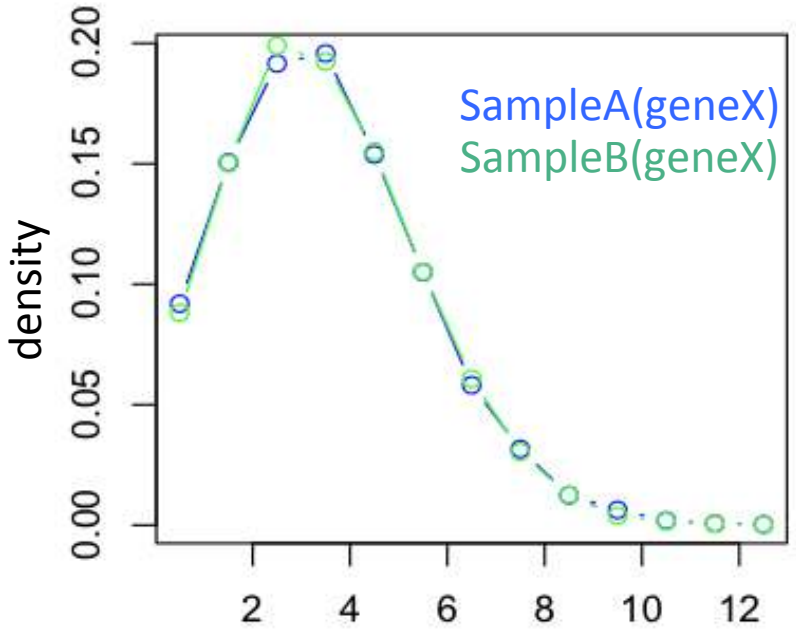
(observed read counts)

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

Example: One gene*not* differentially expressed

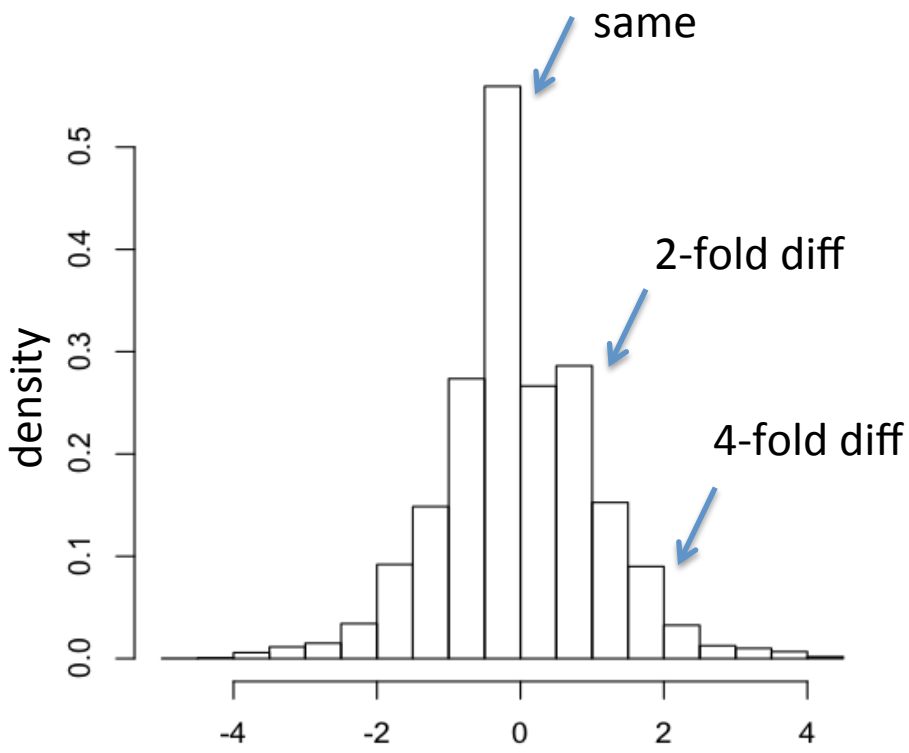
Example: SampleA(gene) = SampleB(gene) = 4 reads

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



(k) number of reads observed

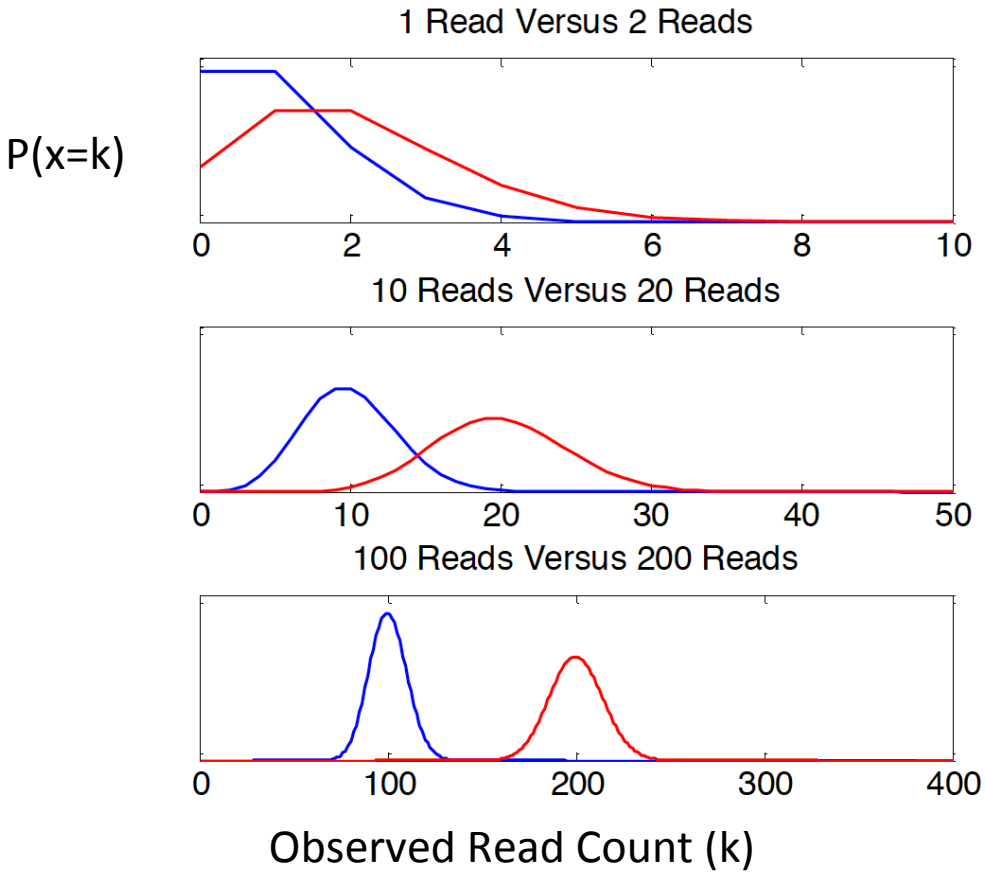
Dist. of $\log_2$(fold change) values



x = $\log_2$(SampleA/SampleB)

Sequencing Depth Matters

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



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*

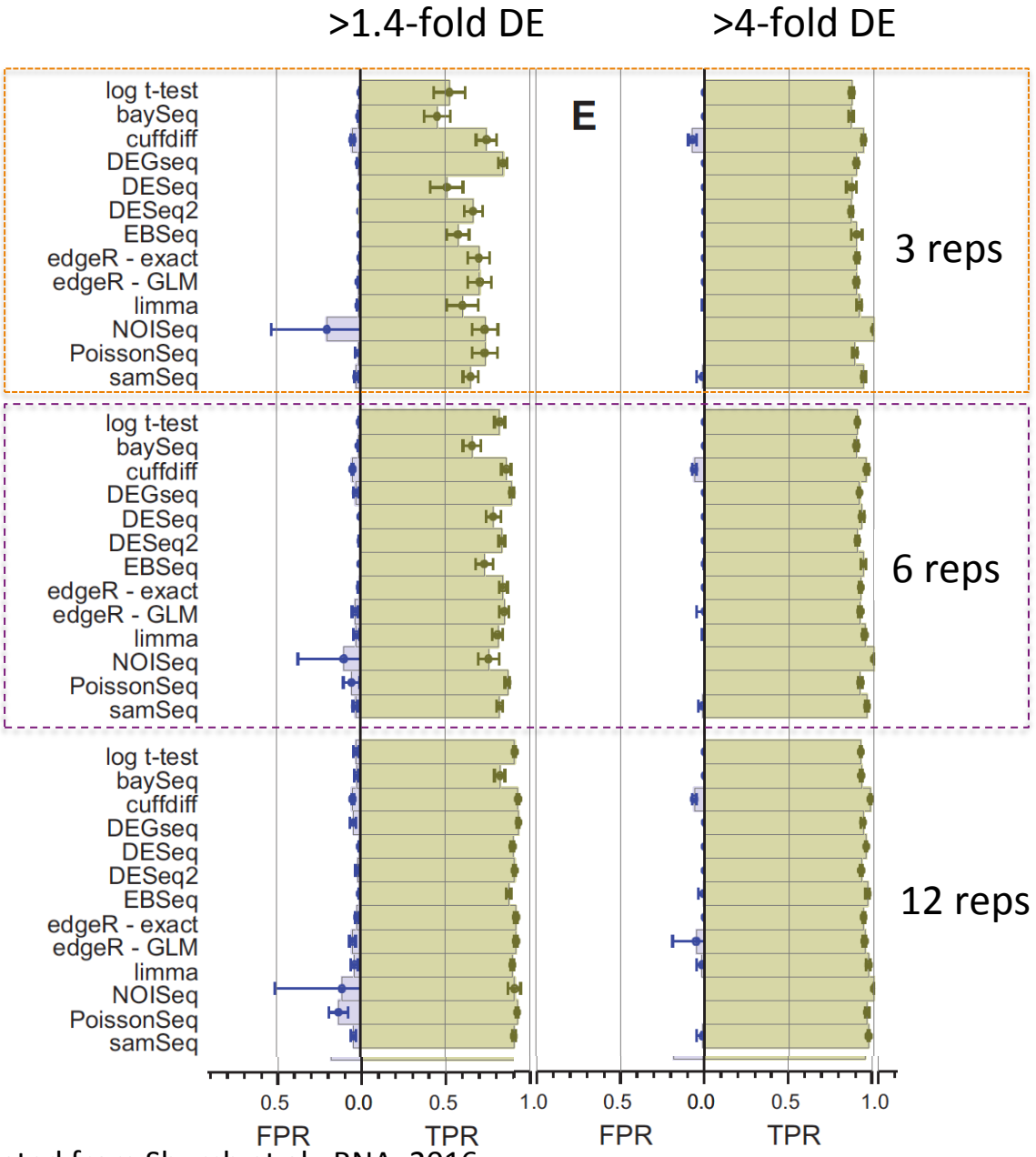
However, biological replicates *ARE* essential

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



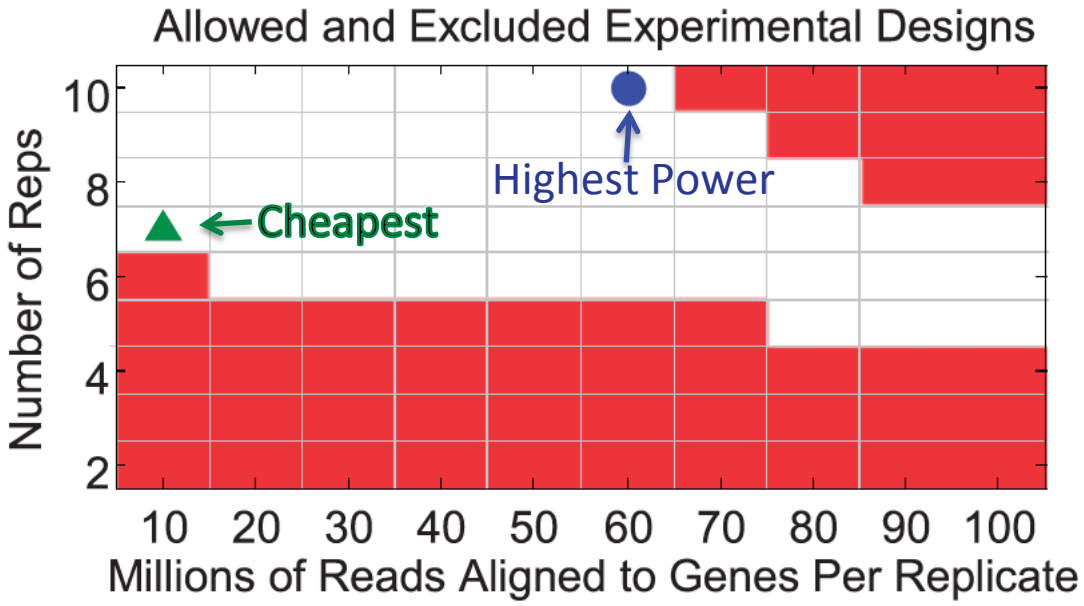
At a minimum, do 3 bio replicates

Respectable goal

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

Planning Experiments:
How many reads and how many replicates?

Input: max total reads, max total replicates, max total $$$



Scotty: http://scotty.genetics.utah.edu/scotty.php

Busby et al., Bioinformatics, 2013

Tools for DE analysis with RNA-Seq

edgeR
ShrinkSeq
DESeq
baySeq
Vsf
Limma/Voom
mmdiff
cuffdiff

ROTS
TSPM
DESeq2
EBSeq
NBPSeq
SAMseq
NoiSeq

(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

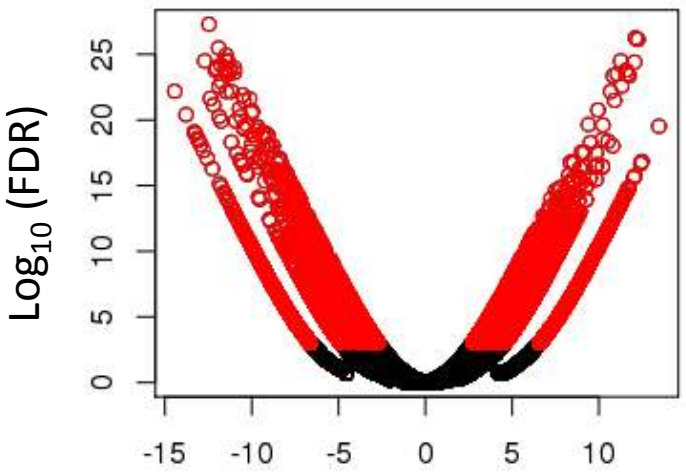
Avg. expression level

Significance

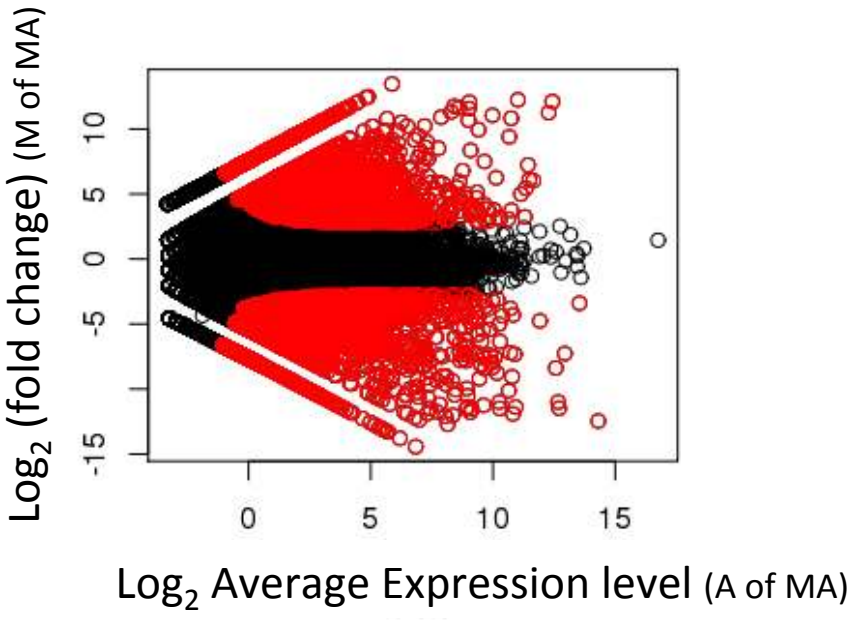
Visualization of DE results and Expression Profiling

Plotting Pairwise Differential Expression Data

Volcano plot
(fold change vs. significance)

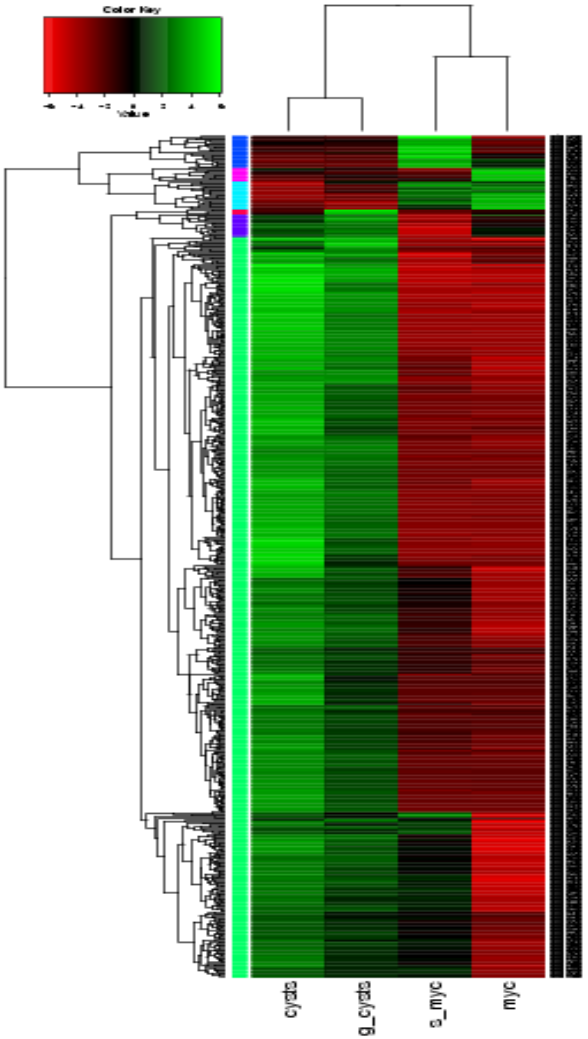


MA plot
(abundance vs. fold change)



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.

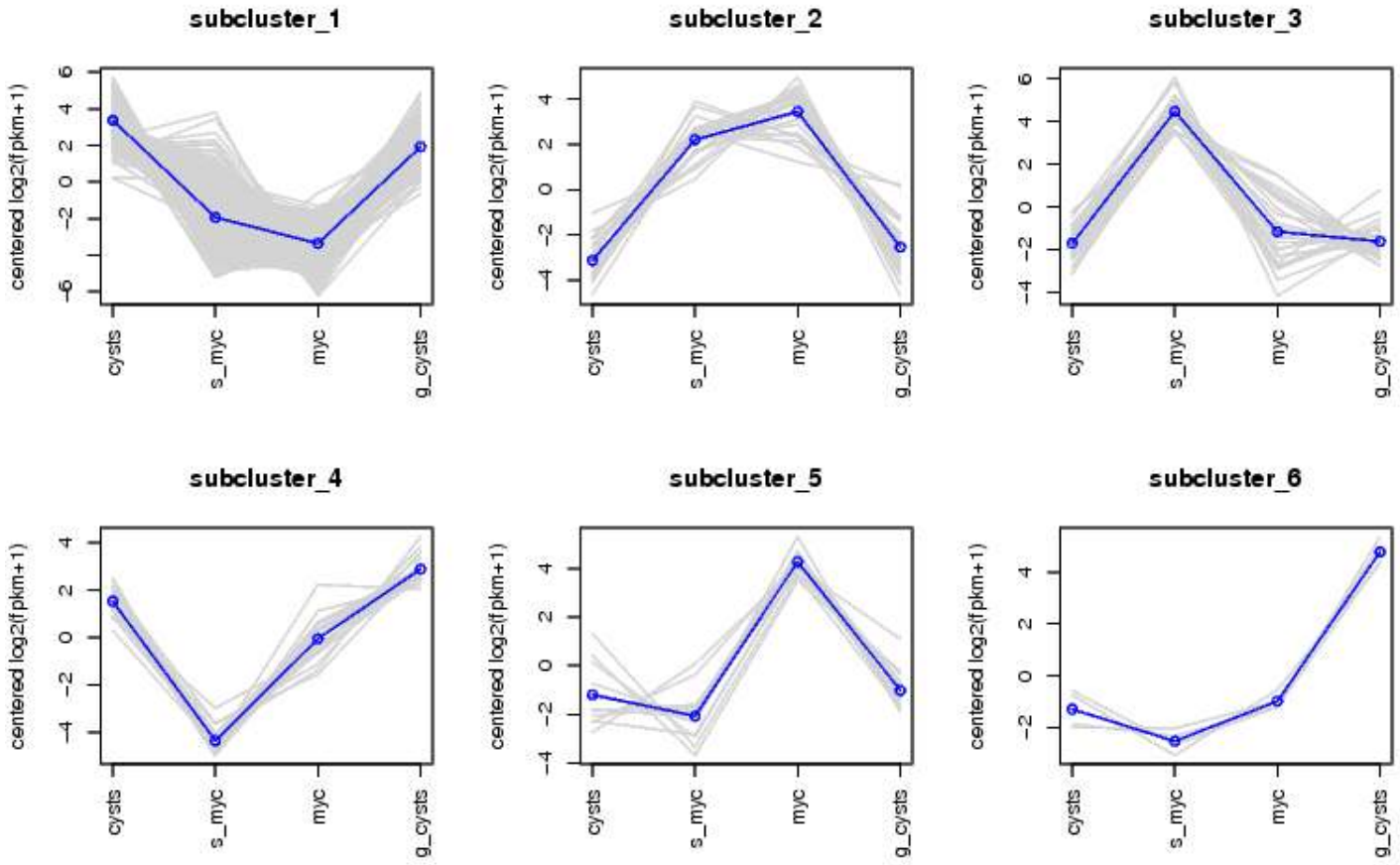
Clustering can be performed across both axes:

-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.



Functional Annotation of Transcripts

Trinotate

NCBI BLAST

HMMER Pfam

UniProt

eggNOG version 3.0

SQLite

GO-Seq

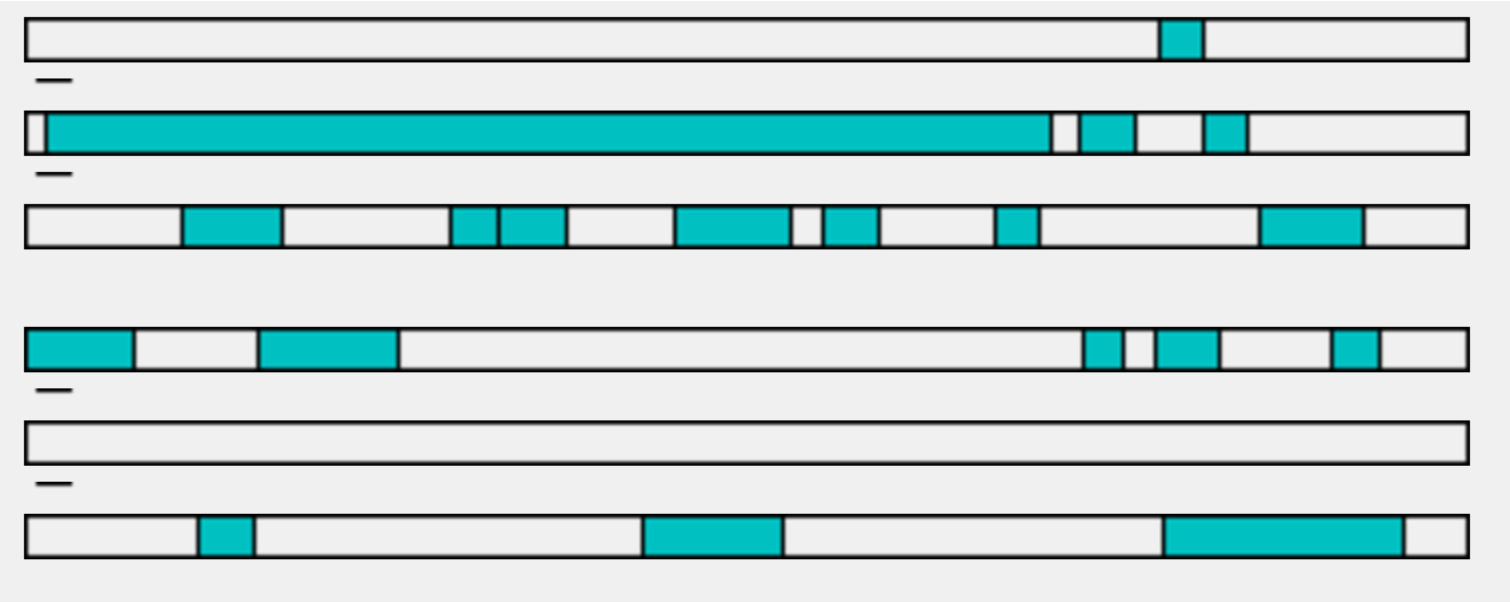
the Gene Ontology

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

Automated Higher Order Biological Analysis

http://trinotate.sf.net

Find Likely Coding Regions
(using TransDecoder)



- Find all ORFs
- Score each ORF according to likely coding potential (Markov model)
- Report highest scoring ORFs

http://transdecoder.github.io

BLAST SwissProt

RecName: Full=Nucleosomal histone kinase 1; AltName: Full=Protein baellchen

Sequence ID: gi|75009857|sp|Q7KRY6.1|NHK1_DROME **Length:** 599 **Number of Matches:** 1

Range 1: 40 to 347 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
99.9 bits(228)	4e-20	Compositional matrix adjust.	87/321(27%)	114/321(35%)	41/321(12%)

```
Query  8    SNVVGVHYRVGKKIGEGSFGMLFQGVNL--------INNQP-------IALKFESRKSEV  52
            +    + R+G  IG G FG  +               +  +P       + + F  R
Sbjct  40   TDLAKGQWRIGPSIGVGGFGEIYAACKVGEKNYDAVVKCEPHGNGPLFVEMHFYLRNAKL  99

Query  53   PQLRDEYLTYKLLMGLPGIPSVYYYG----QEGMYNLLVMDLLGPSLEDLFDYCGRRFSP  108
                 +++    L  L G P +   G            VM   G  L    +  G R
Sbjct  100  EDIK-QFMQKHGLKSL-GMPYYILANGSVEVNGEKHRFIVMPRYGSDLTKFLEQNGKRLPE  157

Query  109  KTVAMIAKQMITRIQSVHERHFIYRDIKPDNFLIGFPGSKTENVIYAVDFGMAKQYRDPK  168
             TV   A QM    Q  H   ++  D K  N L G          Y VDFG+A ++
Sbjct  158  GTVYRLAIQMLDVYQYMHSNGYVHADLKAANILLGLEKGGAAQA-YLVDFGLASHFV---  213

Query  169  THVHRPYNEHKSLSGTARYMSINTHLGREQSRRDDLESMGHVFMYFLRGSLPW--QGLKA  226
            T    P +  K   GT  Y S + HLG    RR DLE +G      L   LPW  Q L A
Sbjct  214  TGDFKP-DPKKMHNGTIEYTSRDAHLG-VPTRRADLEILGYNLIEWLGAELPWVTQKLLA  271

Query  227  ATNK-QKY-----EKIGEKKQVTPLKEL-CEGYPKEFLQYMIYARNLGYEEAPDYDYLRS  279
               K QK      + IGE      LK L   G P     +M Y   L   + PDYD  RS
Sbjct  272  VPPKVQKAKEAFMDNIGE-----SLKTLFPKGVPPPIGDFMKYVSKLTHNQEPDYDKCRS  326

Query  280  LFDSLLLRINETDDGKYDWTL  300
             F S L      ++G  D  +
Sbjct  327  WFSSALKQLKIPNNGDLDFKM  347
```

BLASTX and BLASTP

Pfam Search for Conserved Protein Domains

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 **2** 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.

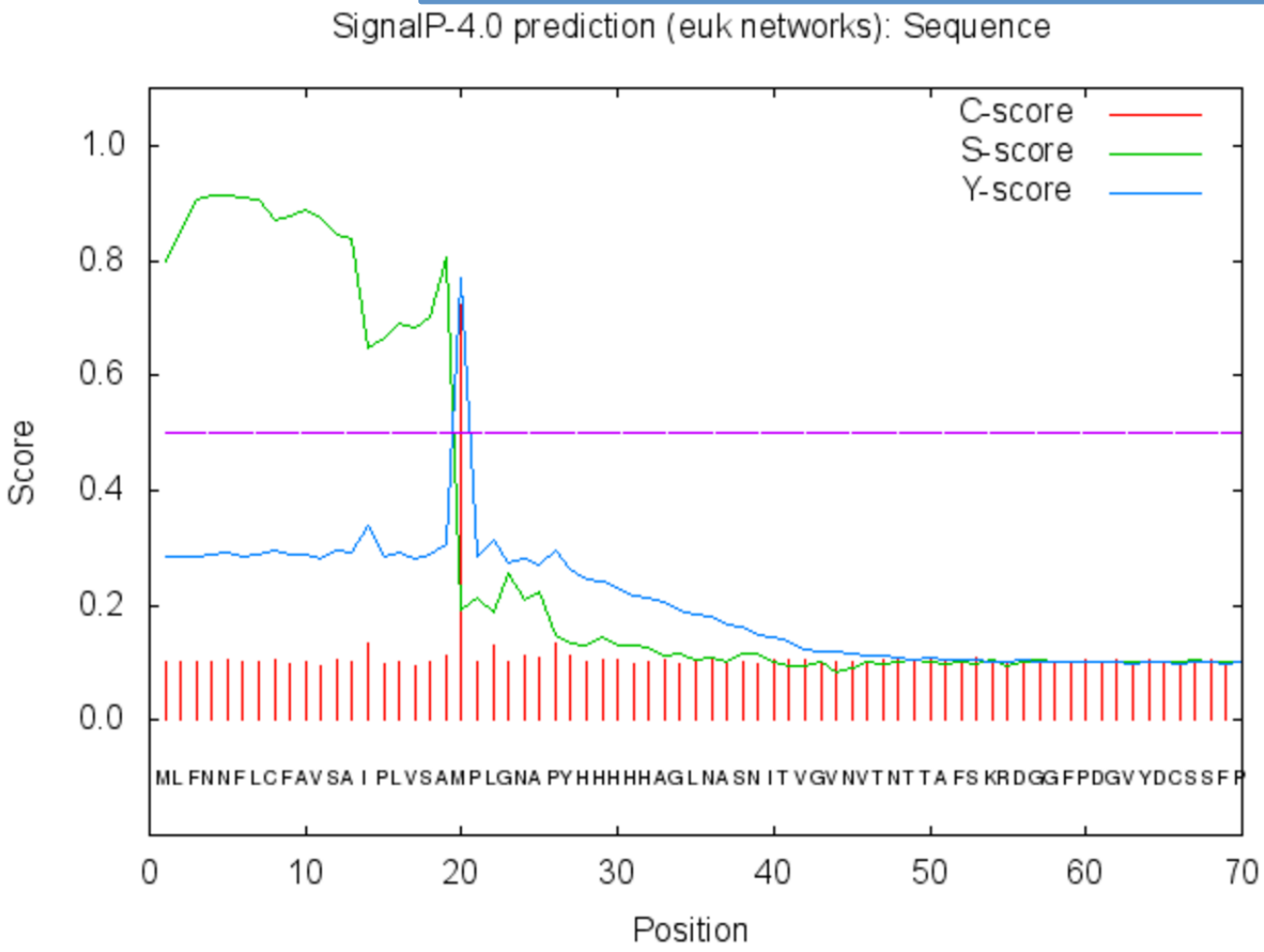
Family	Description	Entry type	Clan	Envelope		Alignment		HMM		HMM length	Bit score	E-value	Predicted active sites	Show/hide alignment
				Start	End	Start	End	From	To					
Glyco_hydro_63N	Glycosyl hydrolase family 63 N-terminal ...	Domain	n/a	41	261	41	258	1	**225**	228	202.9	6.7e-60	n/a	Show
Glyco_hydro_63	Glycosyl hydrolase family 63 C-terminal ...	Domain	CL0059	297	806	298	806	**2**	491	491	622.6	4.4e-187	n/a	Show

Comments or questions on the site? Send a mail to **pfam-help@ebi.ac.uk**.

European Molecular Biology Laboratory

Signal Peptides via SignalP

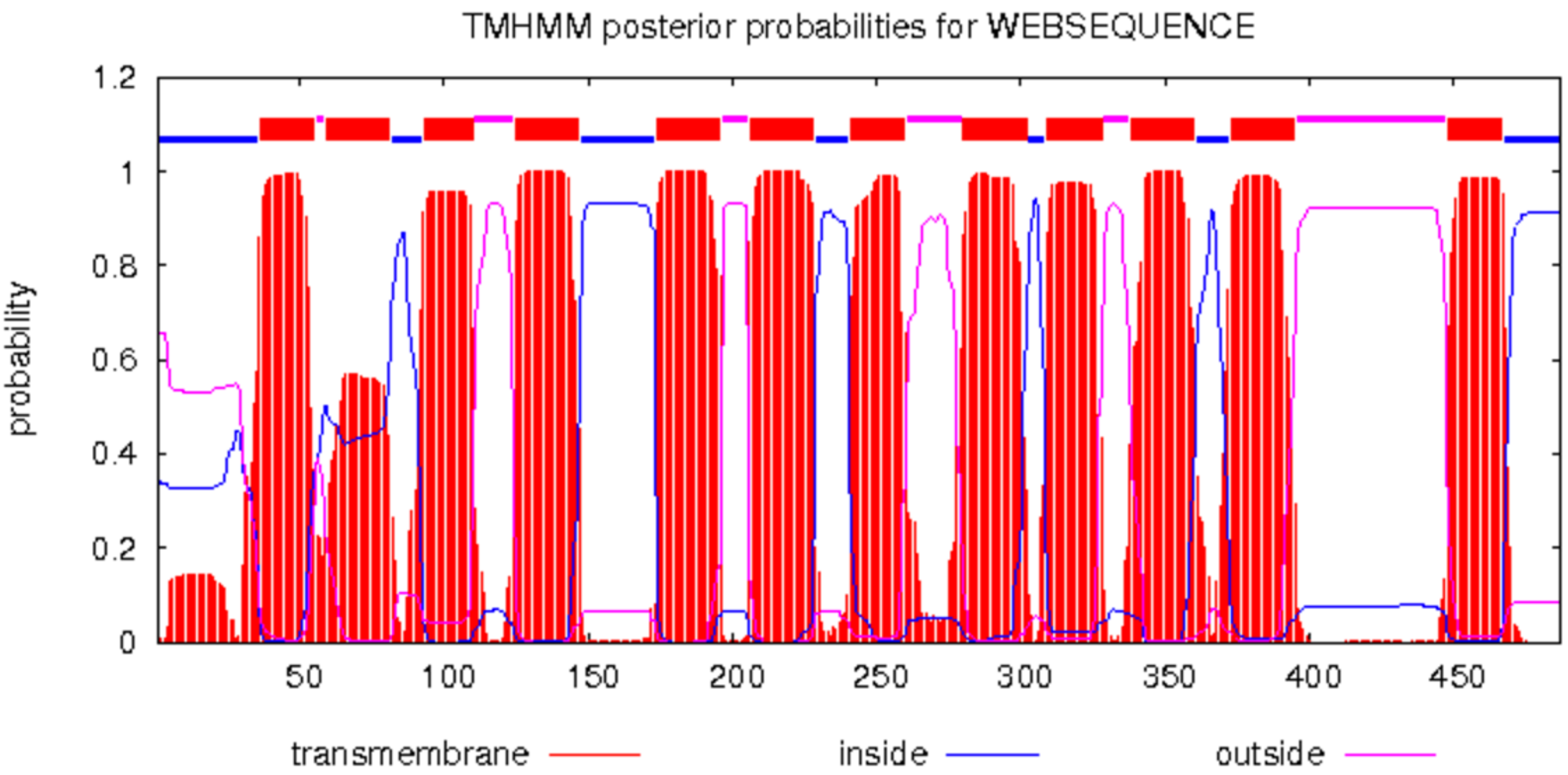
```
# SignalP-4.0 euk prediction
>Sequence
```




```
# Measure  Position  Value    Cutoff   signal peptide?
  max. C    20        0.724
  max. Y    20        0.769
  max. S     5        0.915
  mean S     1-19     0.820
       D     1-19     0.797   0.450    YES
Name=Sequence   SP='YES' Cleavage site between pos. 19 and 20: VSA-MP D=0.797 D-cutoff=0.450 Networks=SignalP-noTM
```

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

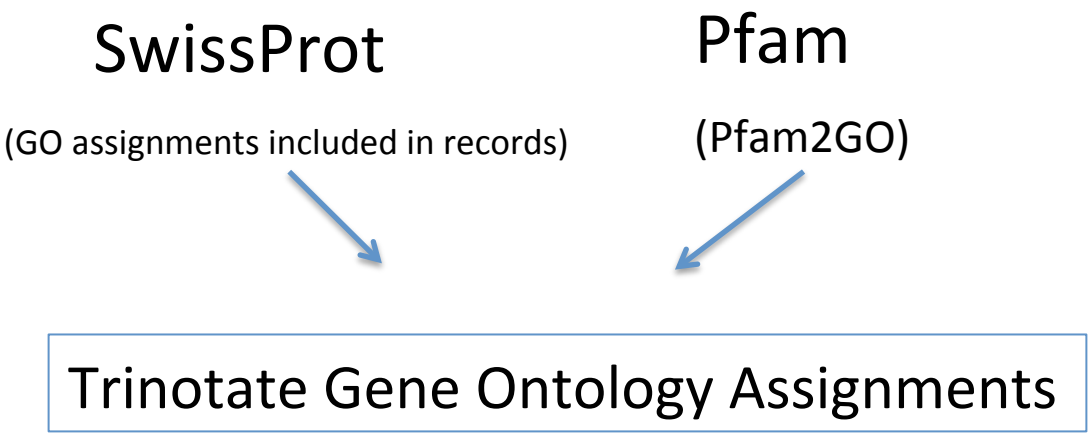
Trans-membrane Domains via TmHMM



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

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

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

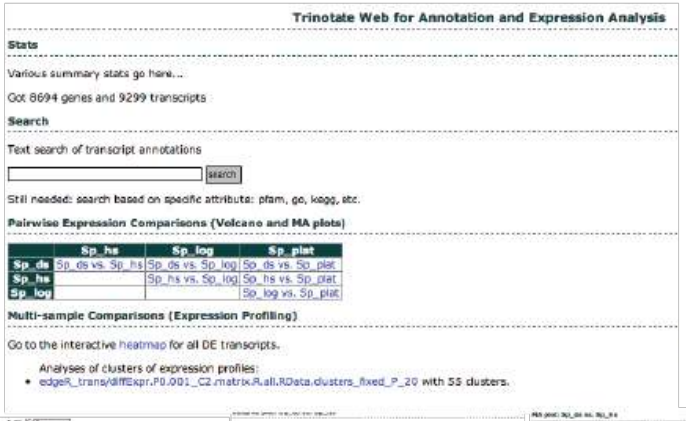
	drawn	not drawn	total
green marbles	k	$K - k$	K
red marbles	$n - k$	$N + k - n - K$	$N - K$
total	n	$N - n$	N

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

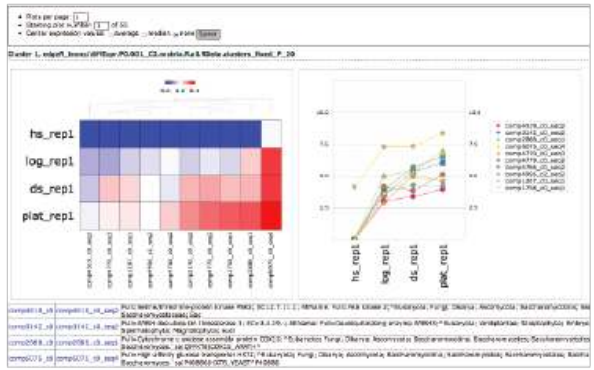
$$P(X = k) = f(k; N, K, n) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}.$$

Trinotate Web for Interactive Analysis

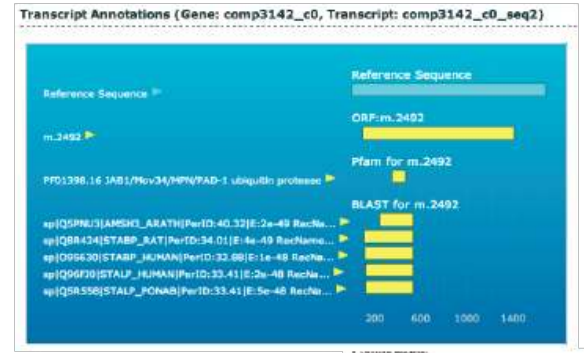
TrinotateWeb Entry Point

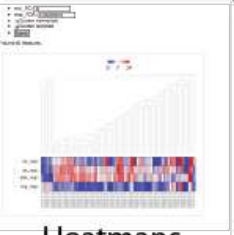


Clustered Expression Profiles

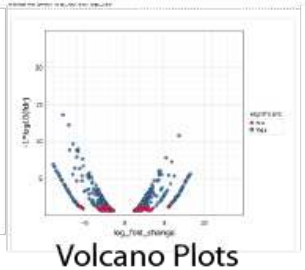


Transcript/Protein Annotation Report
Blast Hits, Pfam Domains, etc.

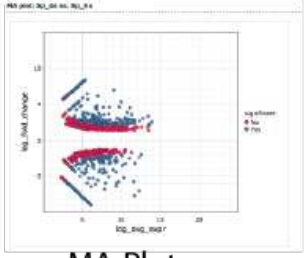




Heatmaps

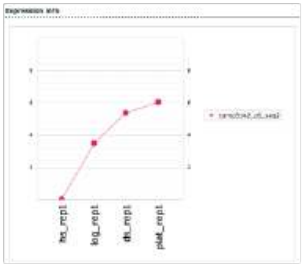


Volcano Plots



MA-Plots

Very Early Release and
Just Scratching the Surface



Individual Transcript
Expression Profiles

Transcript and
Protein Sequences

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

Work done in collaboration with Jessica Whited's lab

BWH

BRMC

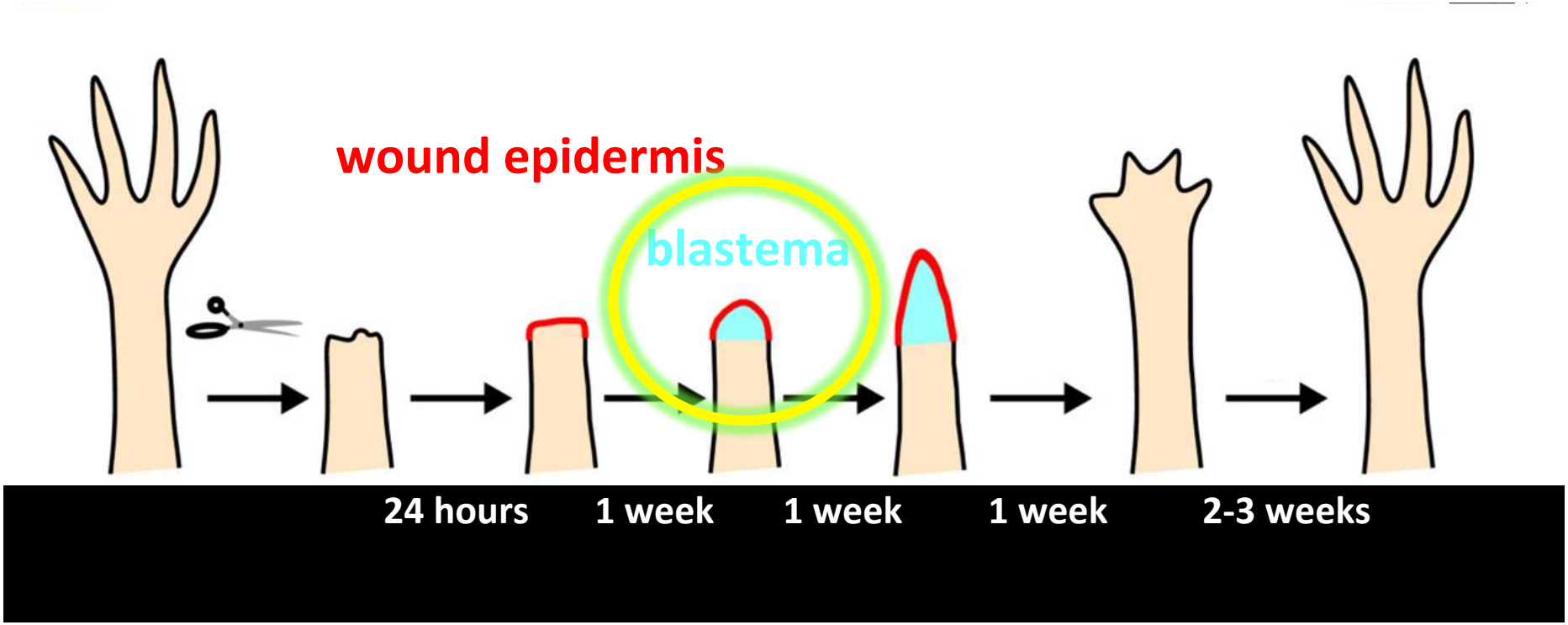
Brigham Regenerative Medicine Center

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)

Key morphological steps during limb regeneration



BWH BRMC Brigham Regenerative Medicine Center

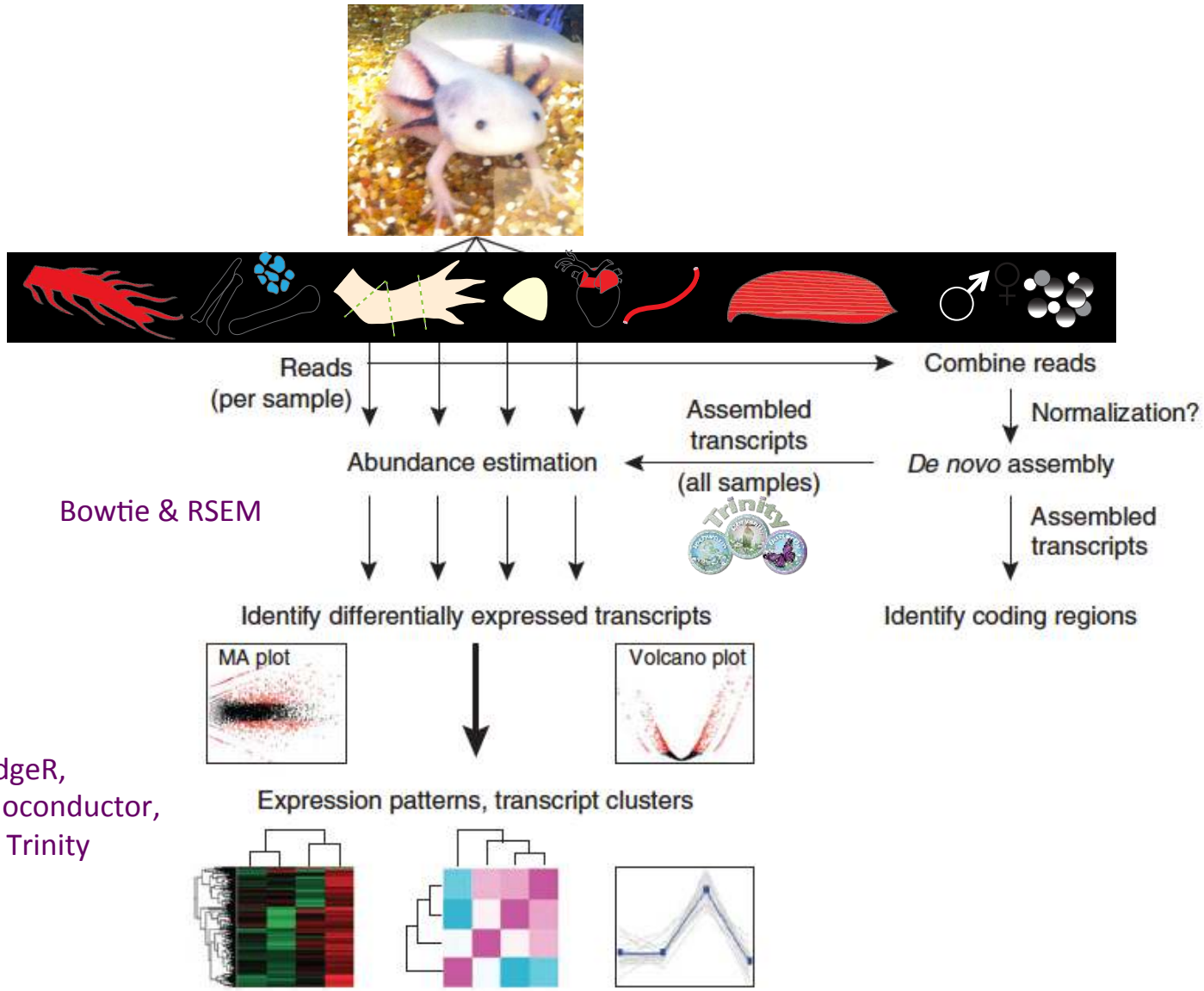
Jessica Whited, Mark Mannucci, Ari Haberberg

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



1.3 Billion
Total Reads

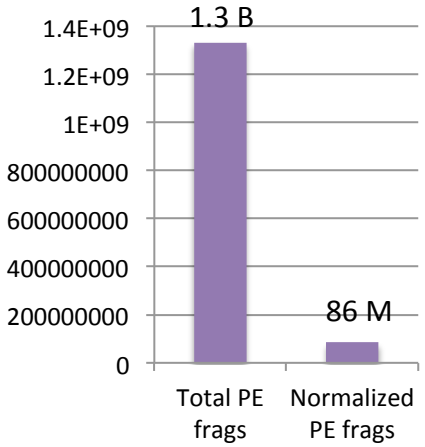
86 Million
Normalized Reads

Bowtie & RSEM

EdgeR,
Bioconductor,
& Trinity

Axolotl Transcriptome De novo Assembly Statistics And Quality Assessment

In silico Normalization

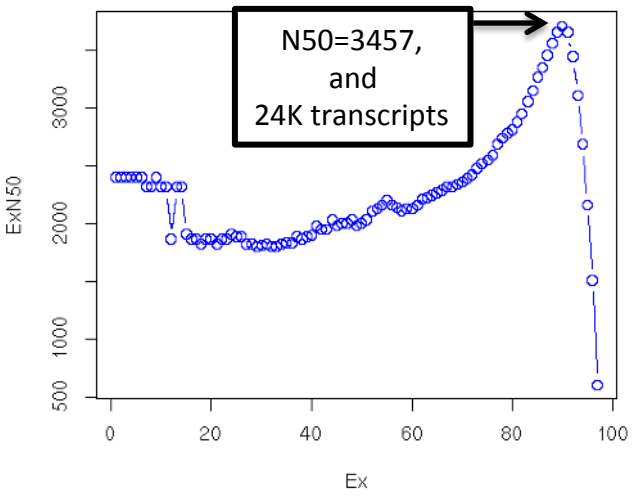


Counts of Transcripts

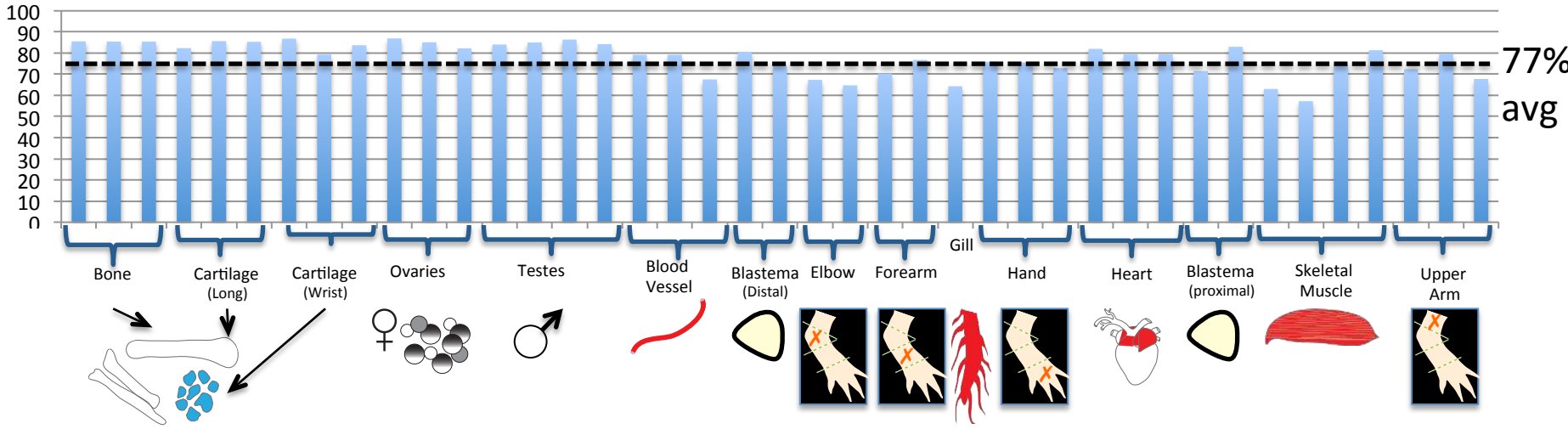
Trinity contigs (transcripts)	**1,554,055**
Trinity components (genes)	**1,388,798**

Min. length 200 bases

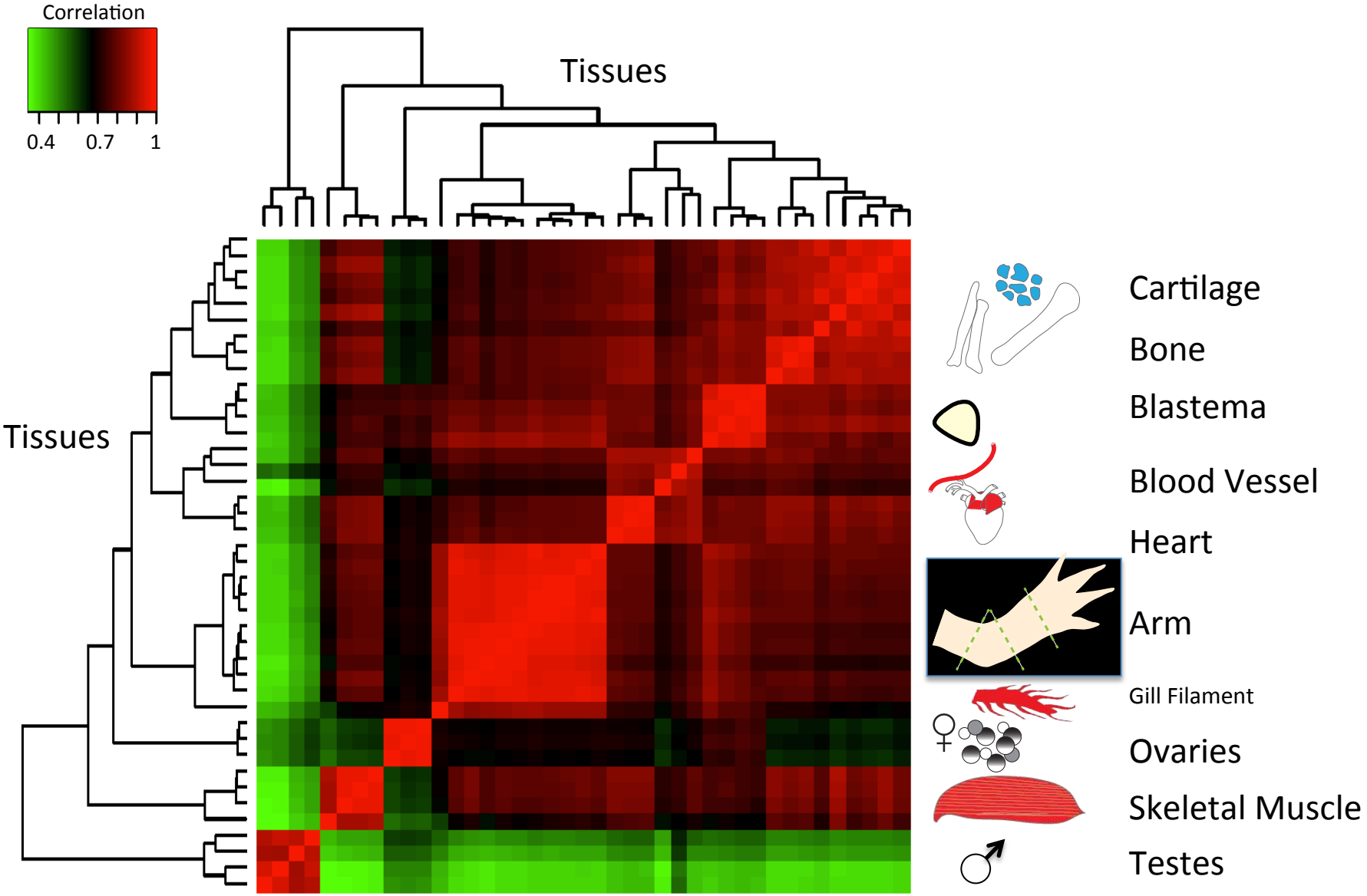
ExN50 looks good!



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

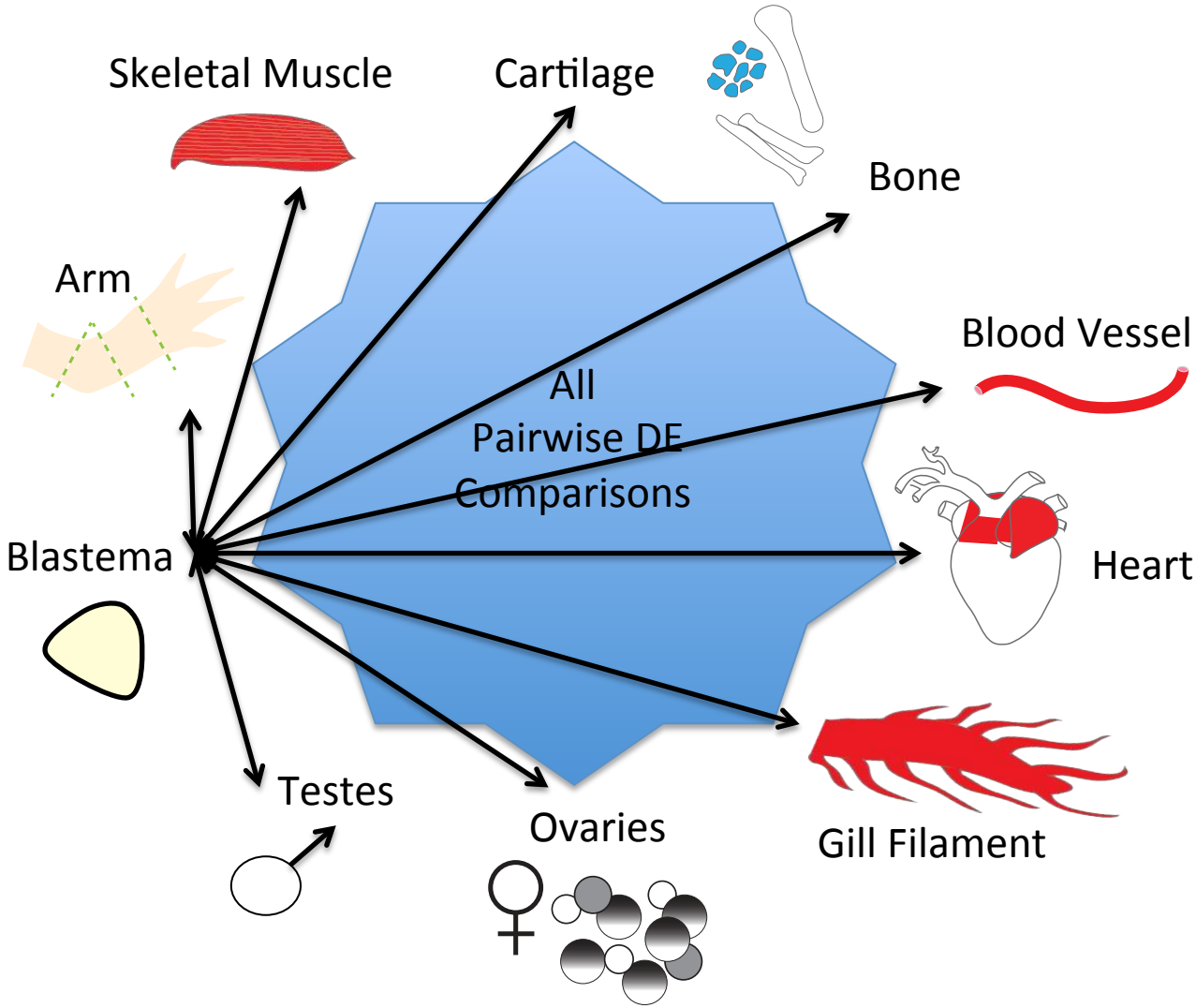


Biological Replicates Cluster According to Sample



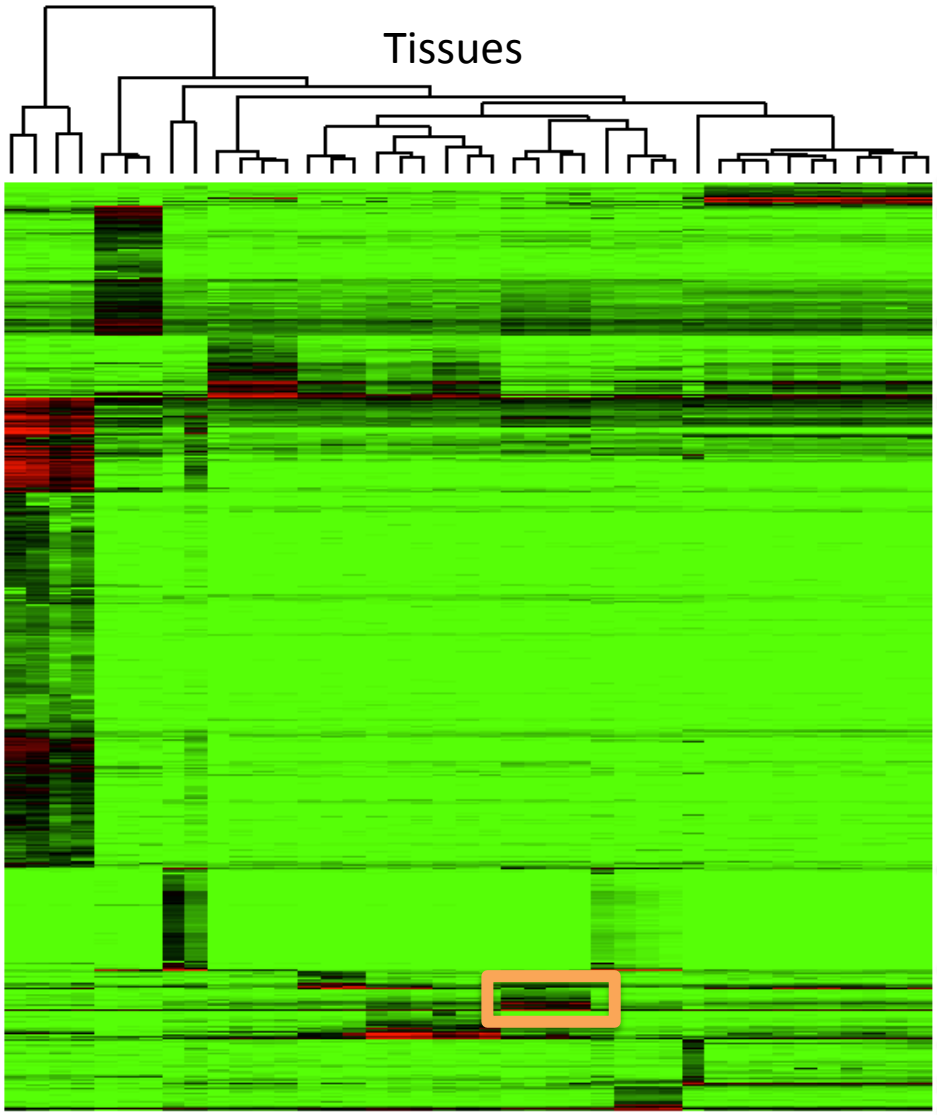
Pearson Correlation Matrix for Tissue Replicates

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]

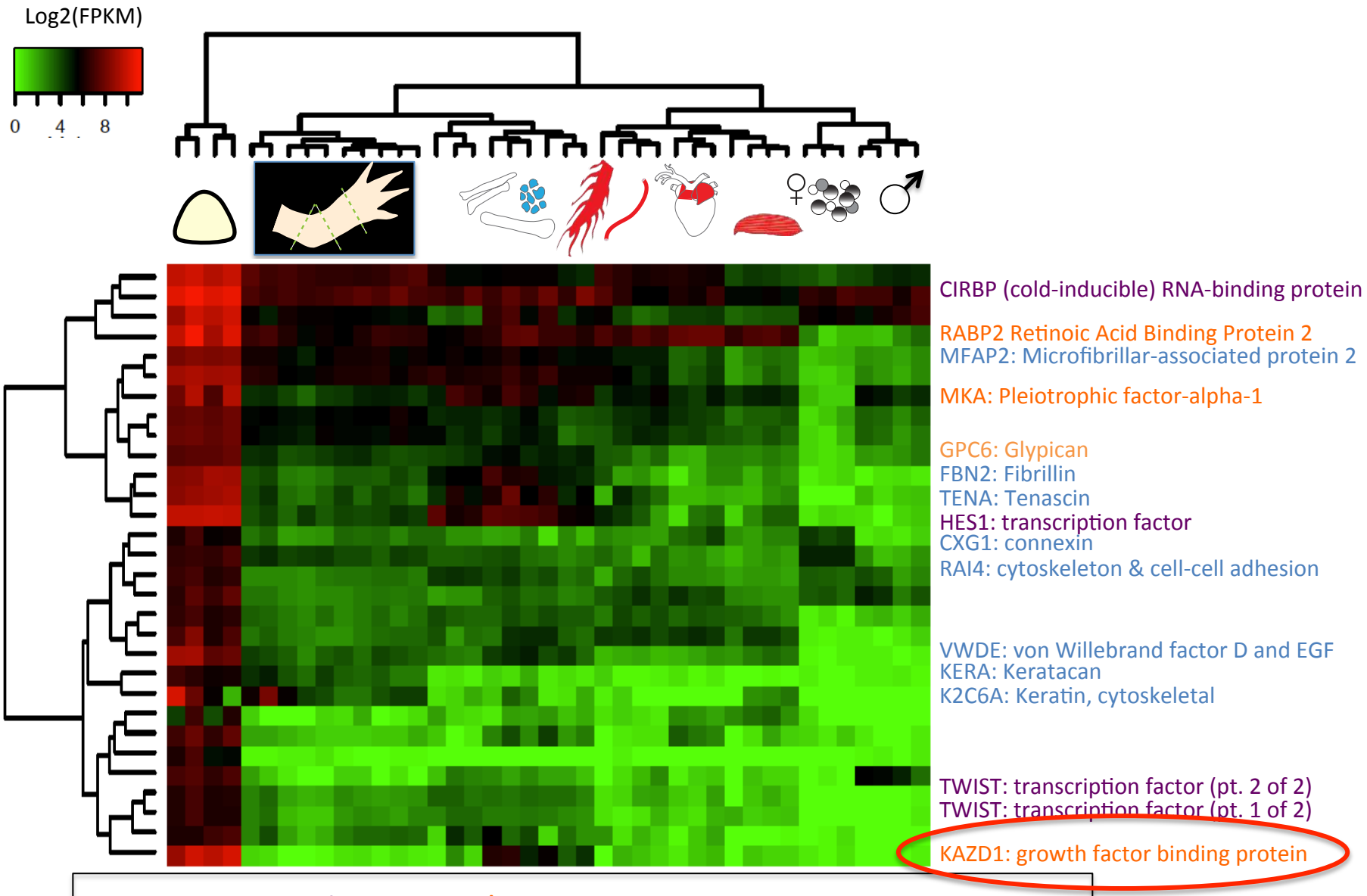
Gill Filament (765)

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

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

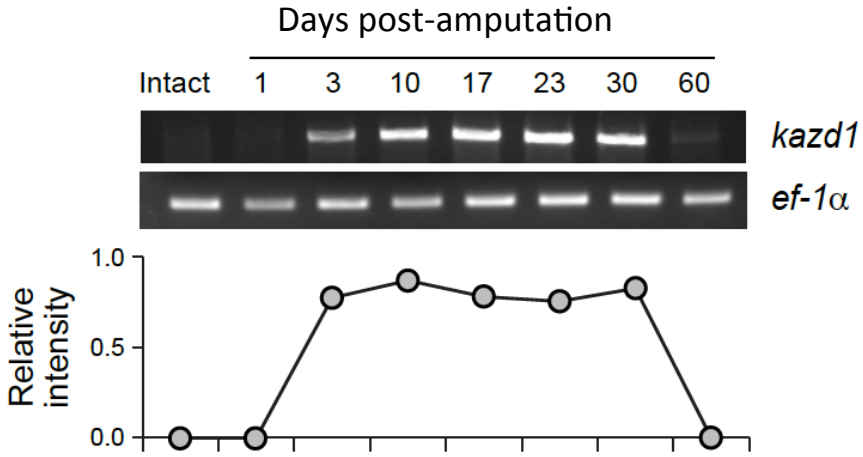
Functional enrichment using GO-Seq

Most Highly Expressed Blastema-enriched Genes

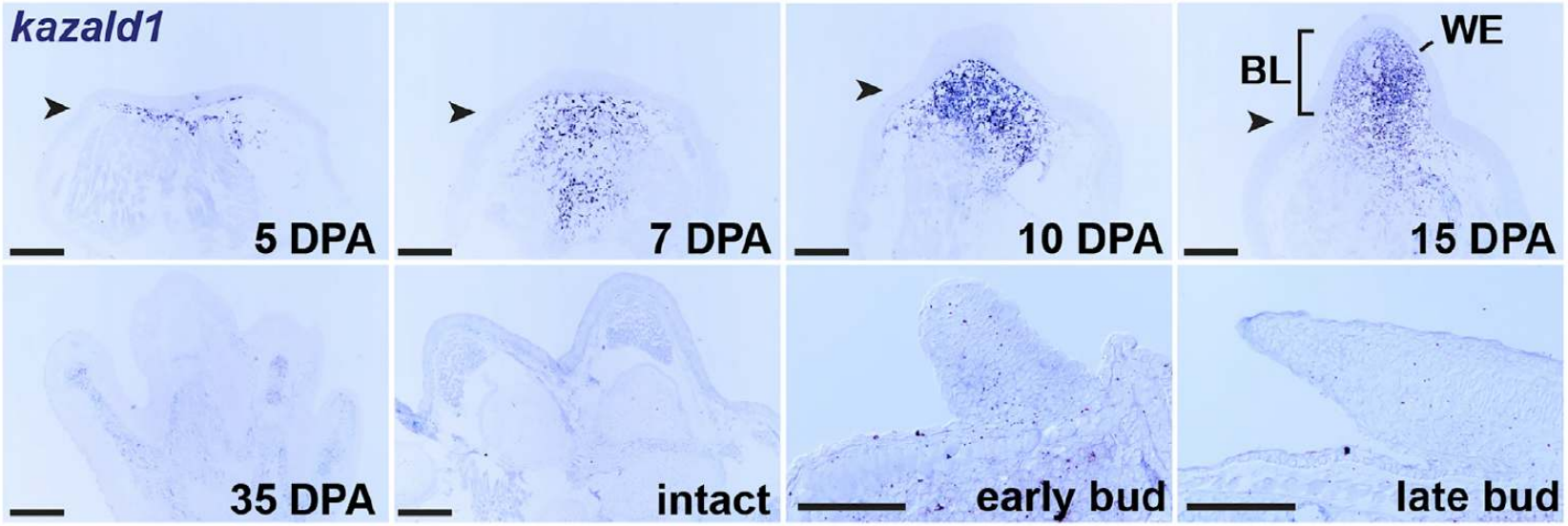


Functional Characterization of Blastema-enriched KAZD1

RT-PCR Timecourse of Kazald1 Expression

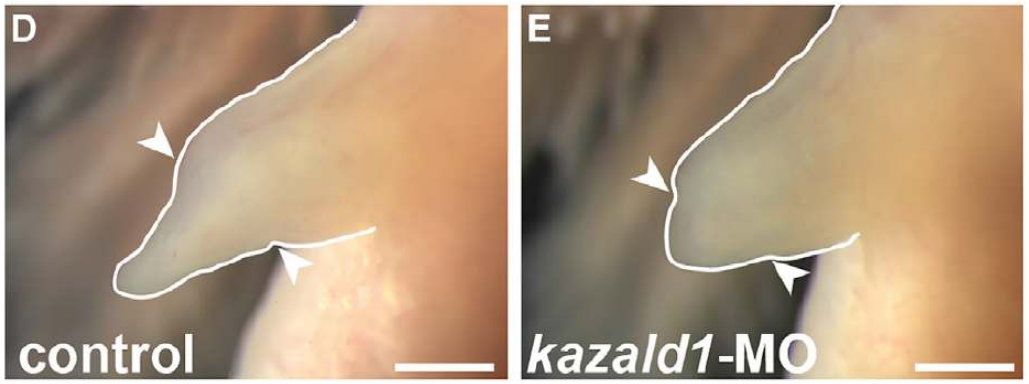


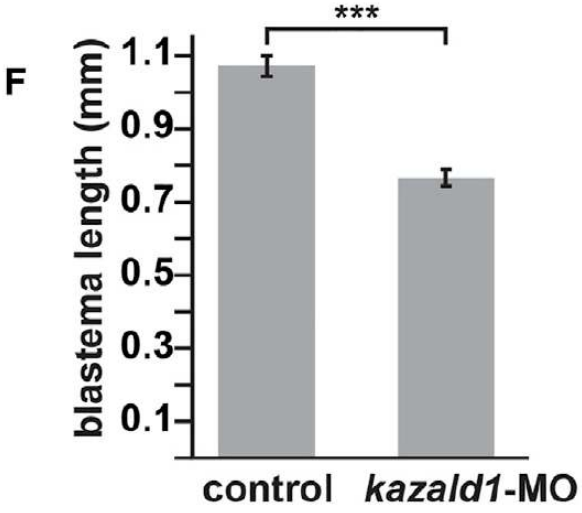
In situ hybridization of kazald1 over course of regeneration



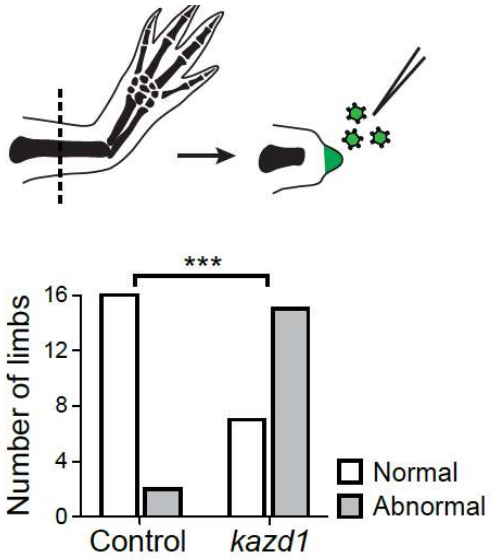
Work by Jessica Whited's group, Cell Reports, 2017

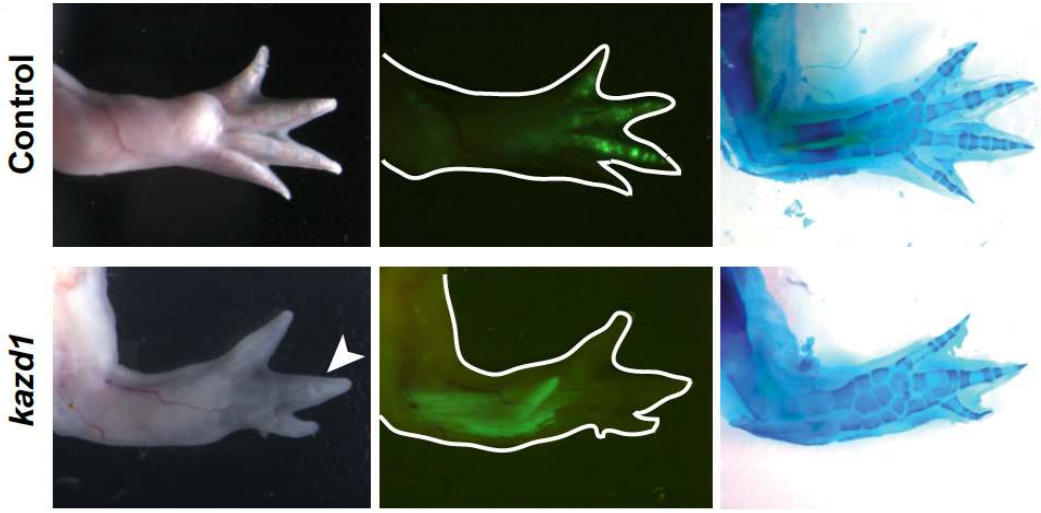
Morpholino Knockdown of Kazald1 Expression



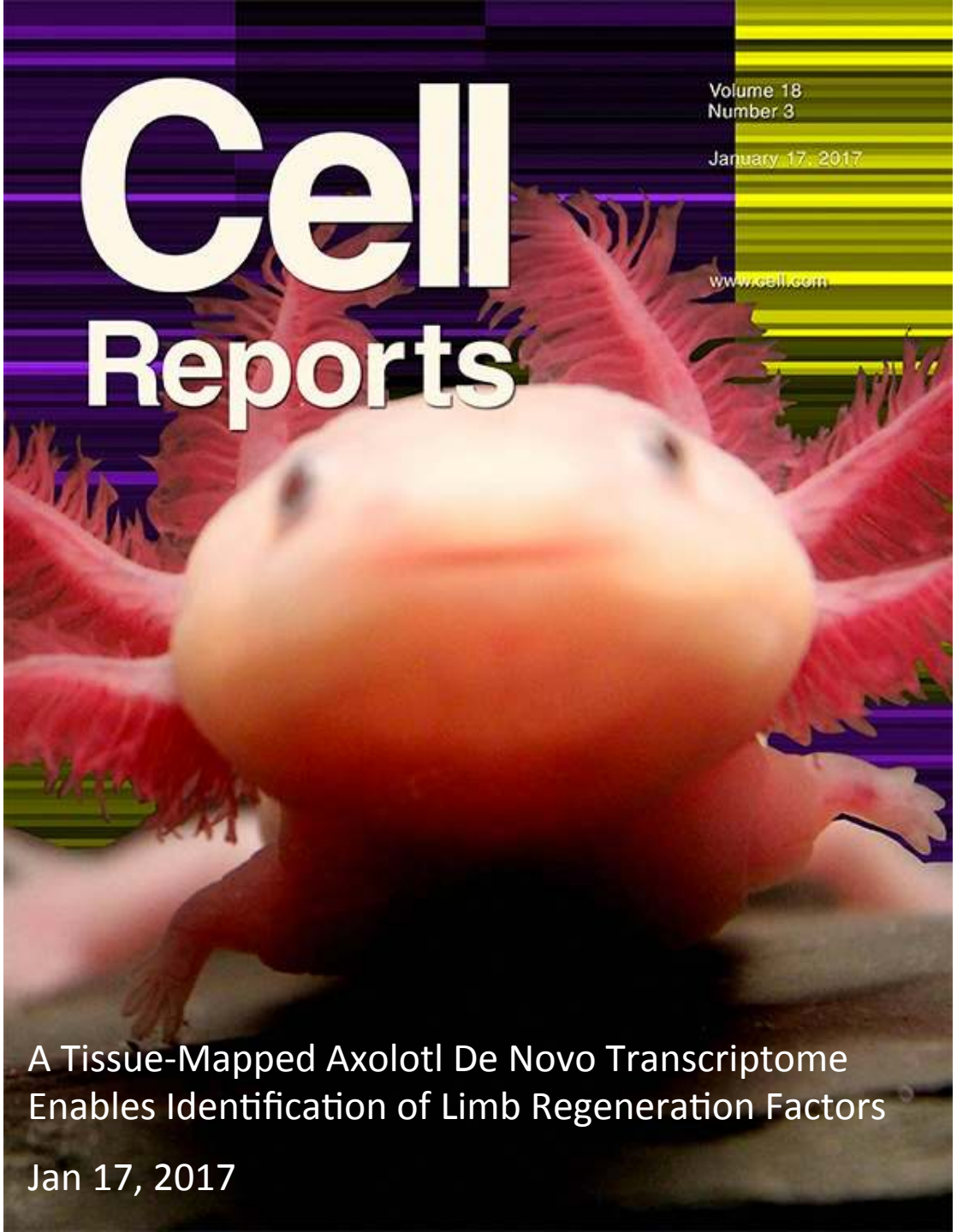


Viral-based Delivered Over-expression of KAZD1 Leads to Regeneration Defects





Work by Jessica Whited's group, Cell Reports, 2017



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

Jan 17, 2017

Summary of Key Points

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

Acknowledgements

BROAD INSTITUTE

Aviv Regev
Brian Haas
Timothy Tickle
Asma Bankapur

BWH BRIGHAM AND WOMEN'S HOSPITAL

Nathalie Pochet

Thomas Doak
Carrie Ganote
Robert Henschel
Ben Fulton

BWH BRMC Brigham Regenerative Medicine Center

Salamander limb regeneration
Jessica Whited
Tia DiTommaso
Tae Lee
Anna Guzikowski
Donald Bryant

igv

Jill Mesirov
James Robinson

Trinotate & TrinoateWeb
Brian Couger
Leonardo Gonzalez

ITCR Informatics Technology for Cancer Research