

RNA-Seq Workshop: Genome-Guided and Genome-Free Transcript Reconstruction and Analysis Using RNA-Seq

Brian Haas

Broad Institute



RNA-Seq Workshop Overview

- Genome-based and genome-free transcript reconstruction from RNA-Seq
- Running the Tuxedo and Trinity software and visualizing the results.
- Basic principles of transcript abundance estimation
- Basic principles of differential expression analysis
- Analysis frameworks included in Tuxedo and Trinity

Trinity RNA-Seq Workshop Website

https://github.com/trinityrnaseq/RNASeq_Trinity_Tuxedo_Workshop/

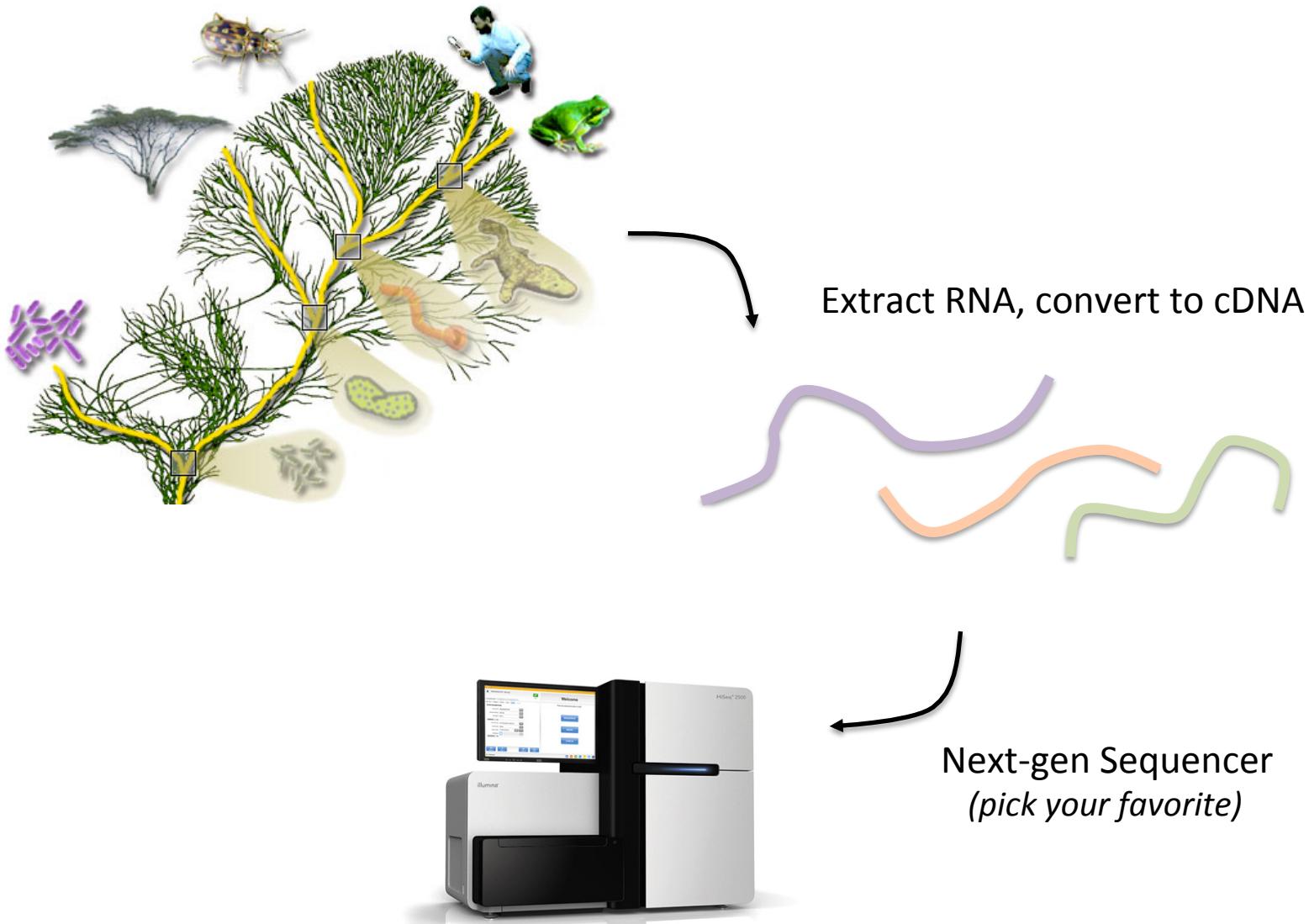
(click the Wiki link at right)



Includes:

- Installation guide for VirtualBox and the workshop VM
 - Ubuntu linux running on Mac or PC via VirtualBox software.
 - Easy installation!
- VM comes installed with workshop data and software.
- Self-guided tutorials for running
 - Genome-guided RNA-Seq analysis using Tuxedo
 - Genome-free RNA-Seq analysis using Trinity

RNA-Seq Empowers Transcriptome Studies



Generating RNA-Seq: *How to Choose?*

Many different instruments hit the scene in the last decade



Illumina



454



SOLiD



Helicos



Ion Torrent

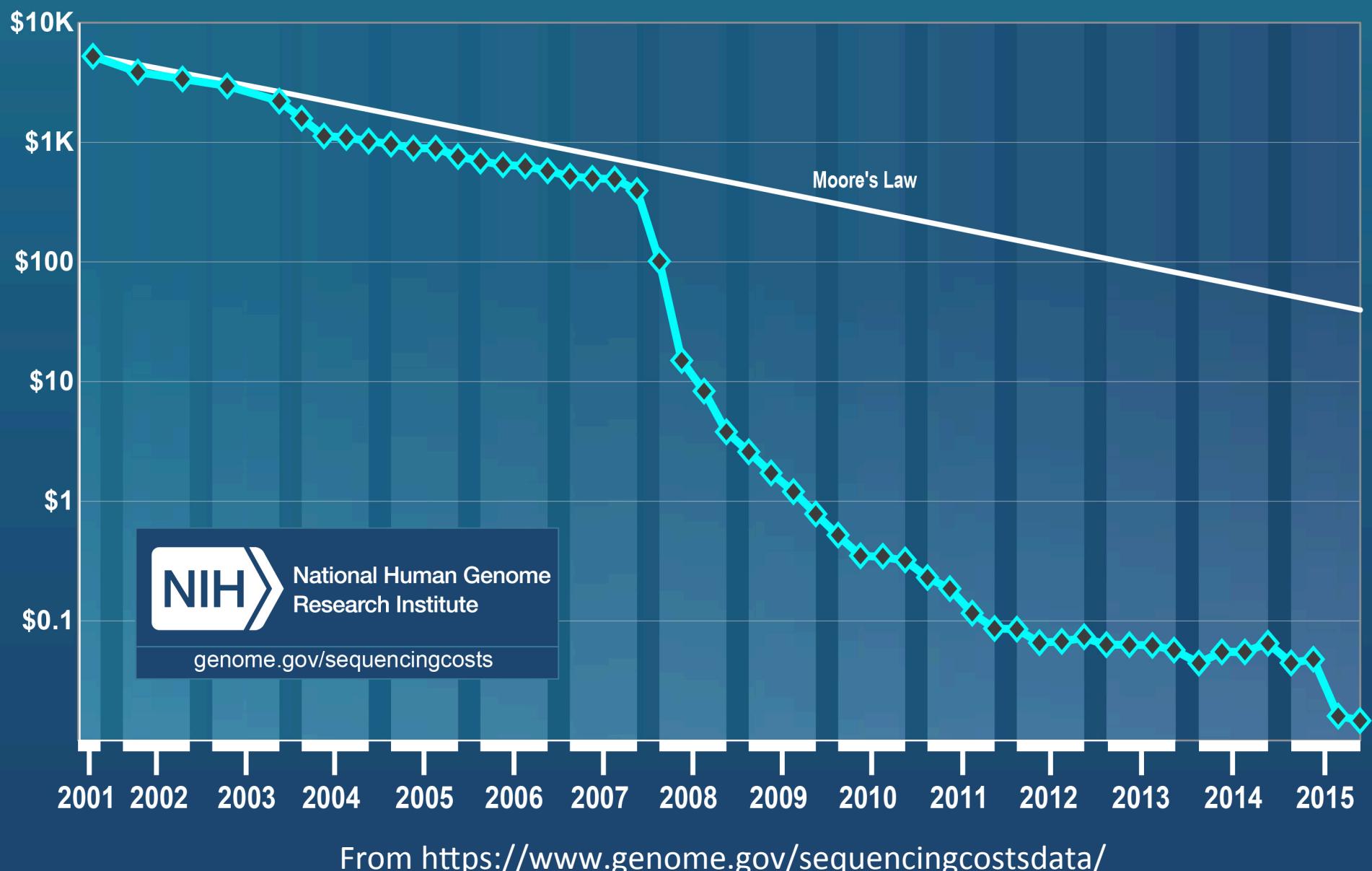


Pacific Biosciences



Oxford Nanopore

Cost per Raw Megabase of DNA Sequence



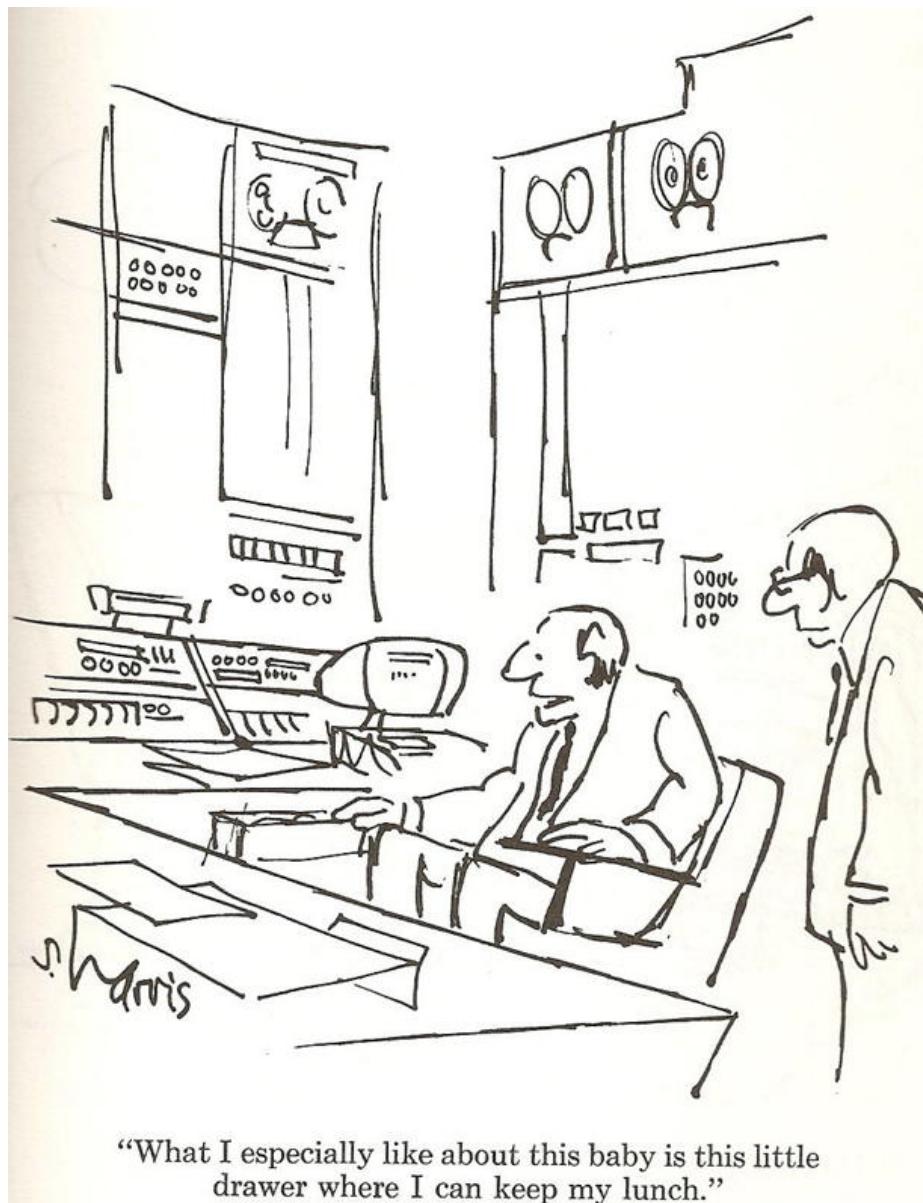
RNA-Seq: *How to Choose?*



Illumina



Ion Torrent



Helicos



Oxford Nanopore

Generating RNA-Seq: *How to Choose?*

Popular choices for RNA-Seq today



Illumina



454



SOLiD



Helicos



Ion Torrent



Pacific Biosciences



Oxford Nanopore

Generating RNA-Seq: *How to Choose?*

Popular choices for RNA-Seq today

[Current RNA-Seq workhorse]



Illumina



Ion Torrent

[Full-length single molecule sequencing]



Pacific Biosciences

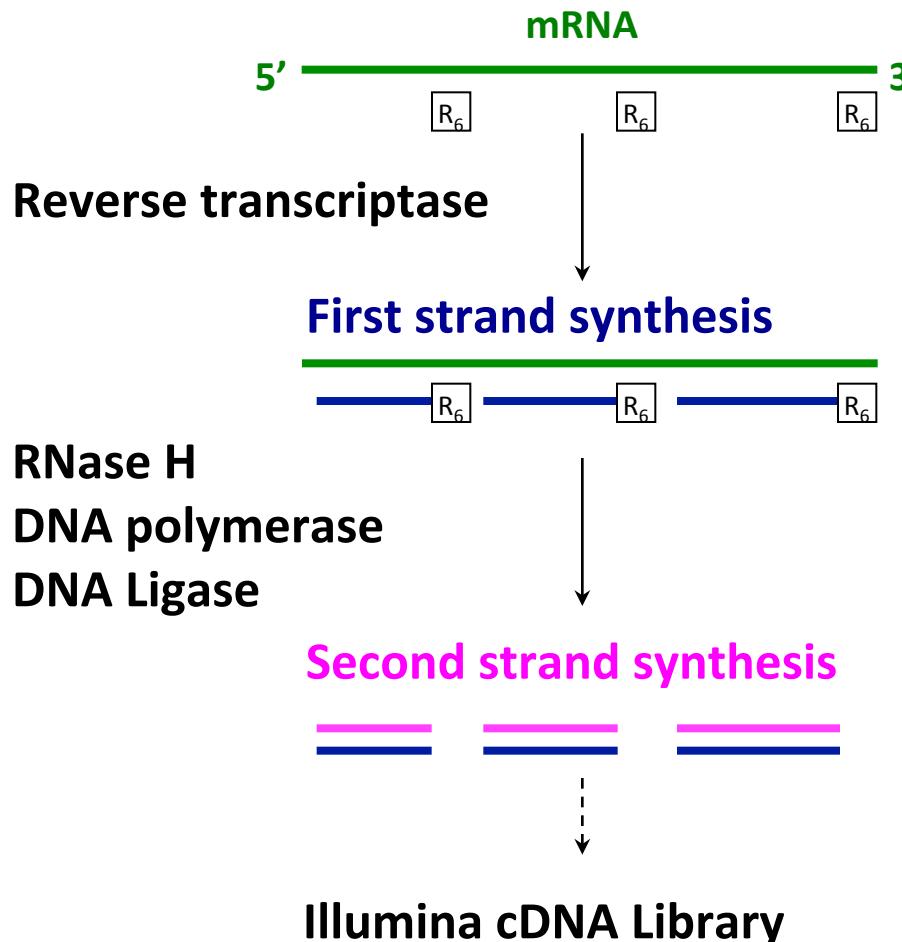
[Newly emerging technology for full-length single molecule sequencing]



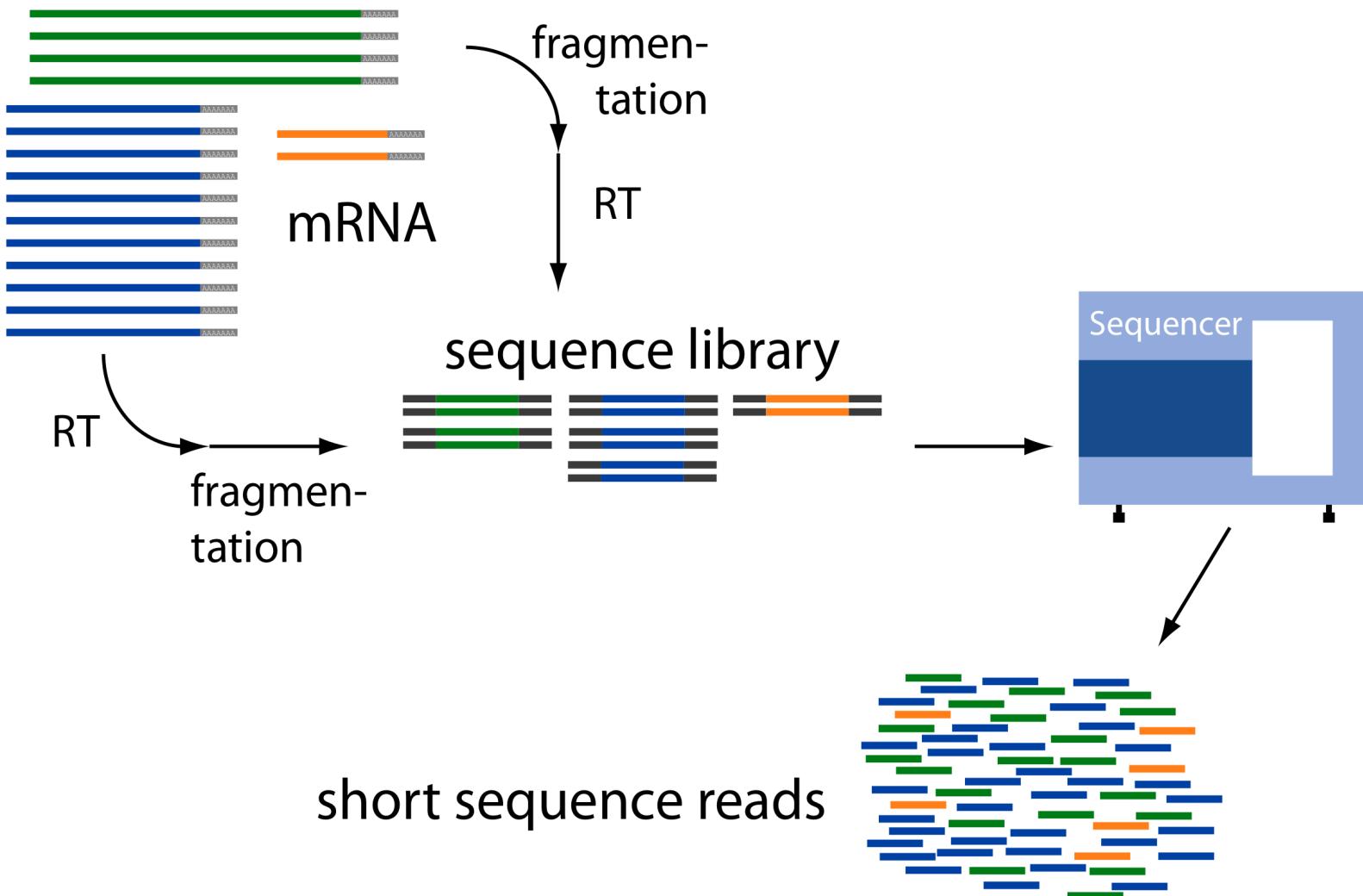
Oxford Nanopore

RNA-Seq: How do we make cDNA?

Prime with Random Hexamers (R6)



Overview of RNA-Seq



Common Data Formats for RNA-Seq

FASTA format:

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

FASTQ format:

```
@61DFRAAXX100204:1:100:10494:3070/1  
AAACAAACAGGGCACATTGTCACTCTTGTATTGAAAAACACTTCCGGCCAT  
+  
ACCCCCCCCCCCCCCCCCCCCCCCCCCCCCBC?CCCCCCCC@CACCCCCA
```

Read

Quality values

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

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

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

Paired-end Sequences

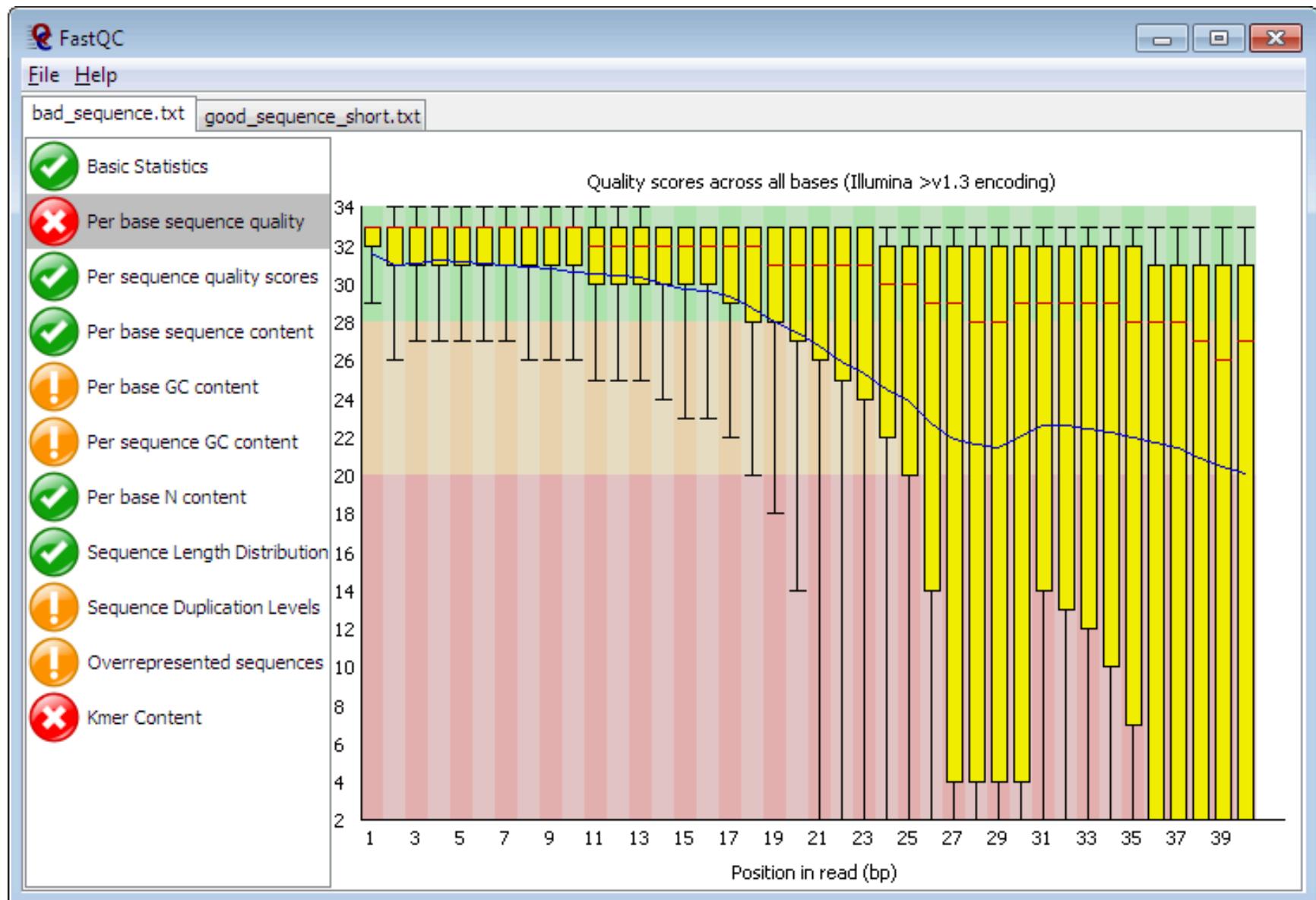


Two FastQ files, read name indicates
left (/1) or right (/2) read of paired-end

```
@61DFRAAXX100204:1:100:10494:3070/1
AAACAAACAGGGCACATTGTCACTCTGTATTTGAAAAACACTTCCGGCCAT
+
ACCCCCCCCCCCCCCCCCCCCCCCCCCCCCBC?CCCCCCCC@ @CACCCCCA
```

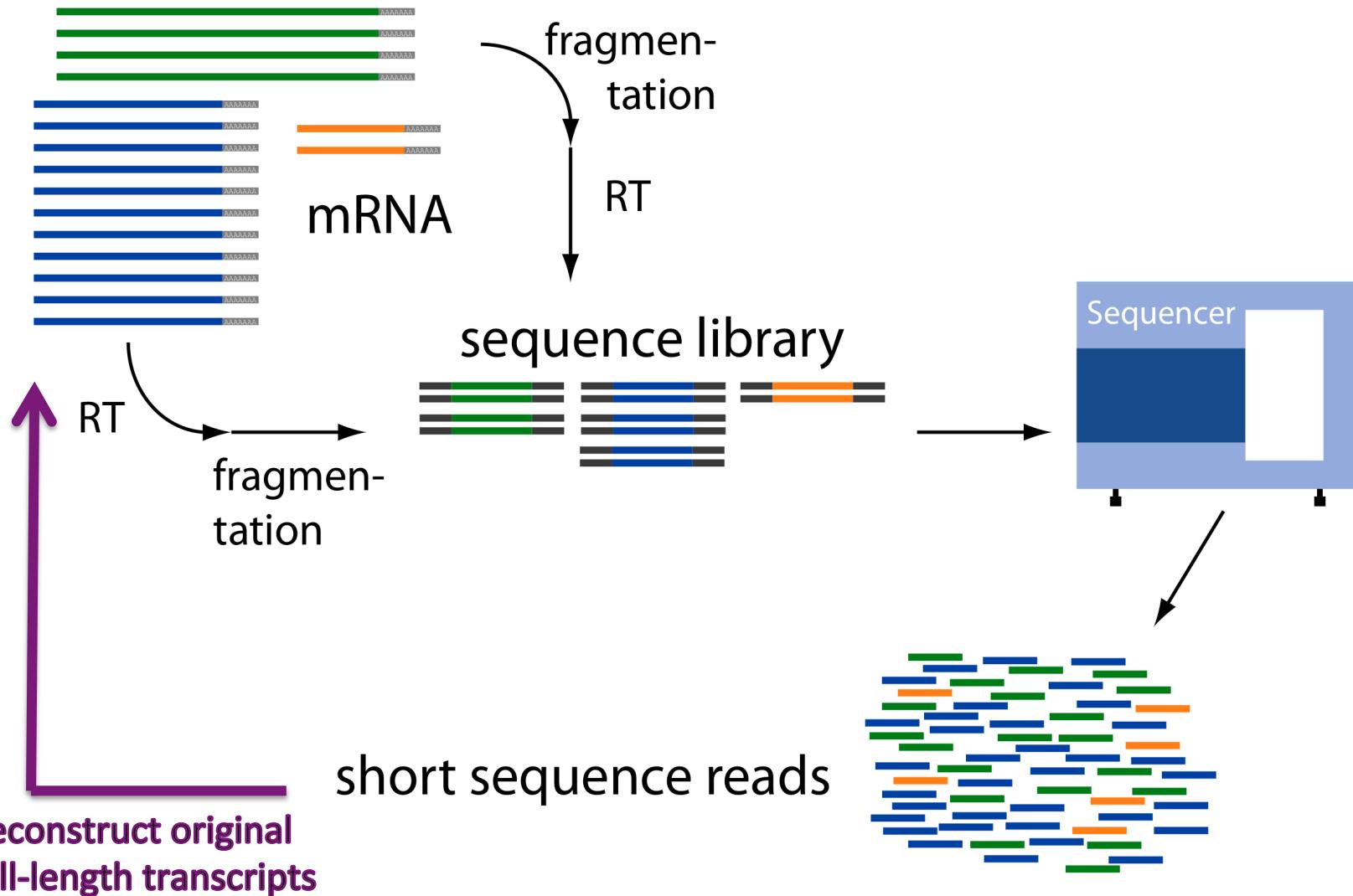
```
@61DFRAAXX100204:1:100:10494:3070/2
CTCAAATGGTTAACATTCTCAGGCTGCAAATATTGTTAGGATGGAAGAAC
+
C<CCCCCCCCACCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCBCCCC
```

Read Quality Assessment

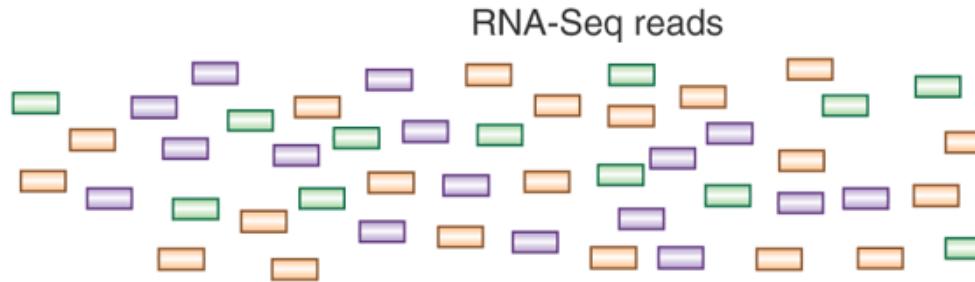


From: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

Overview of RNA-Seq



Transcript Reconstruction from RNA-Seq Reads



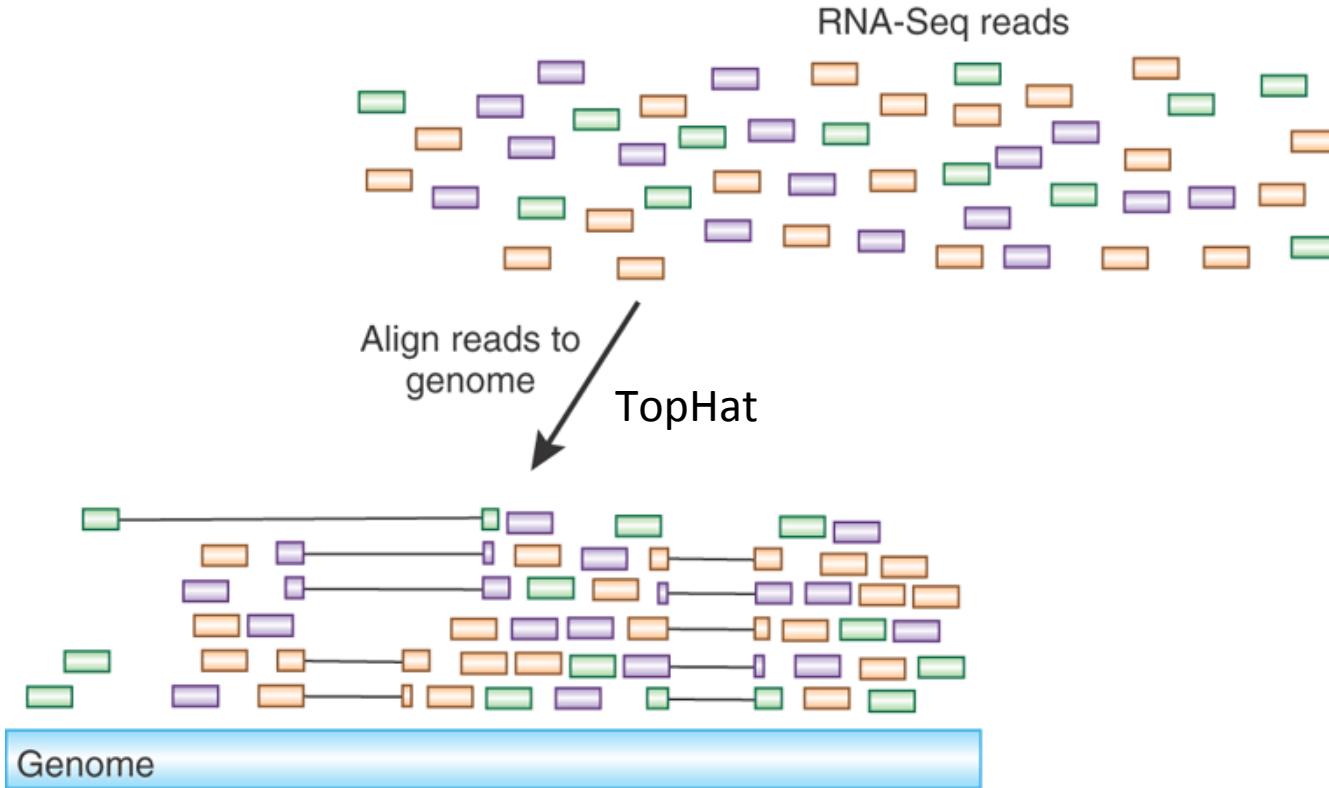
Advancing RNA-Seq analysis

Brian J Haas & Michael C Zody

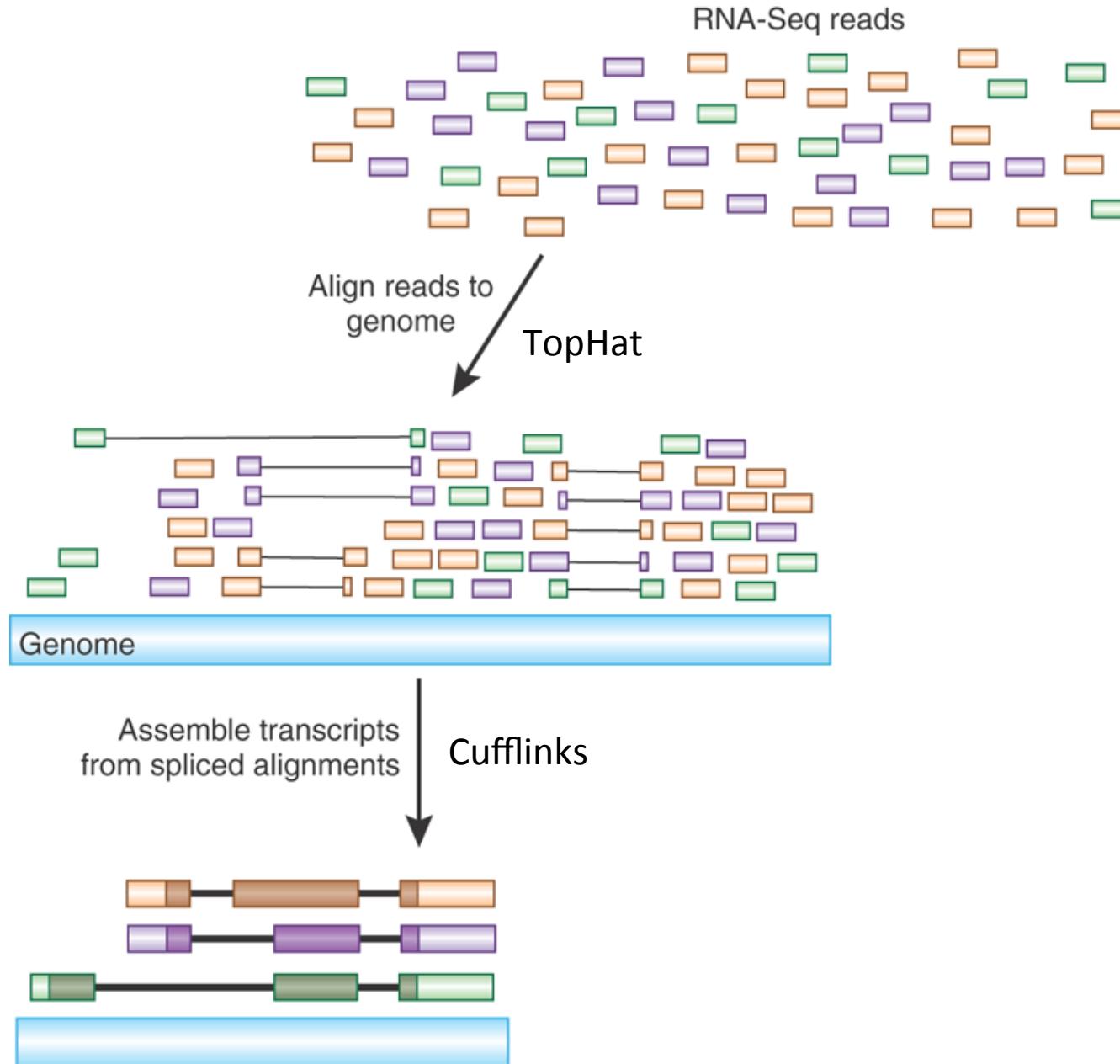
Nature Biotech, 2010

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

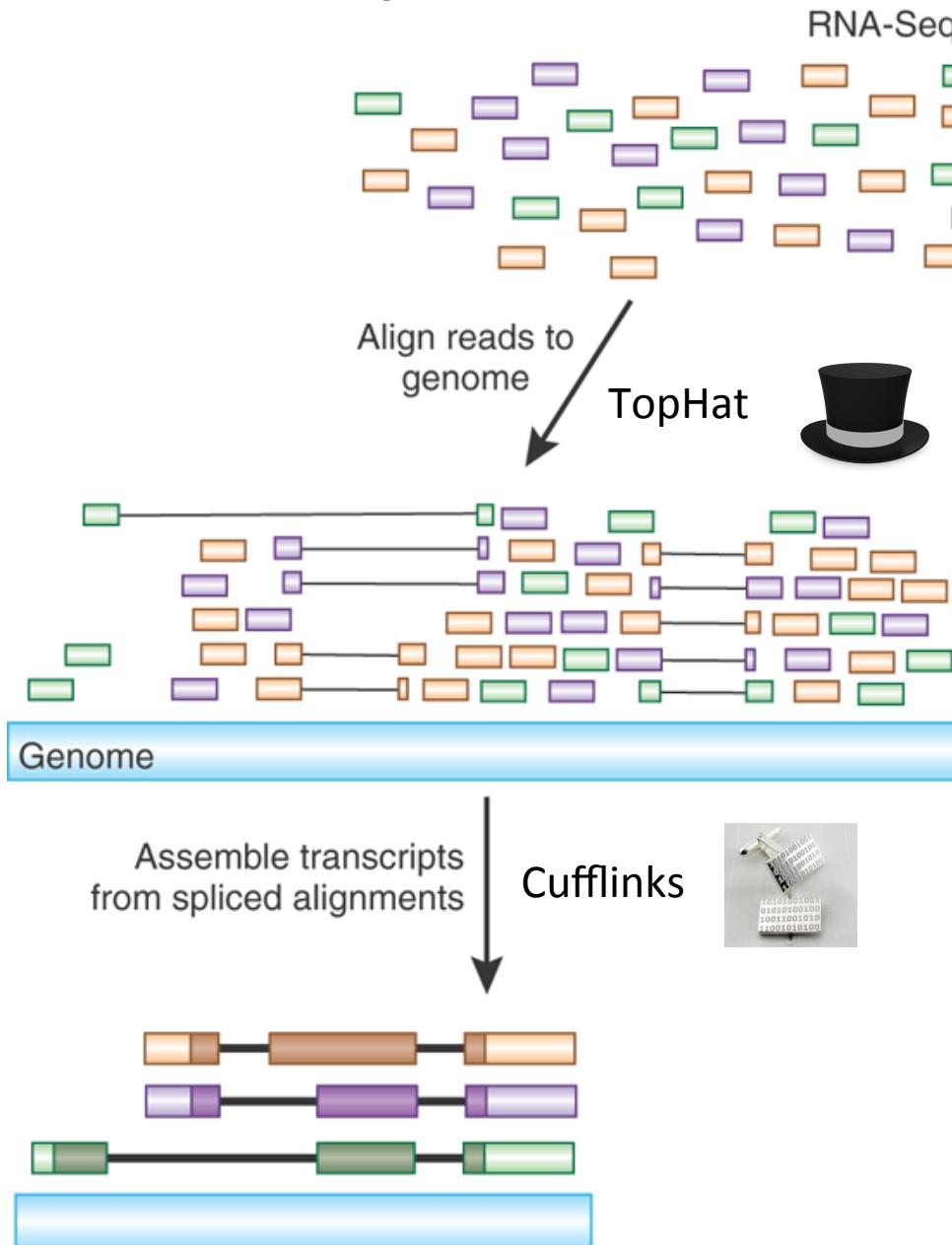
Transcript Reconstruction from RNA-Seq Reads



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

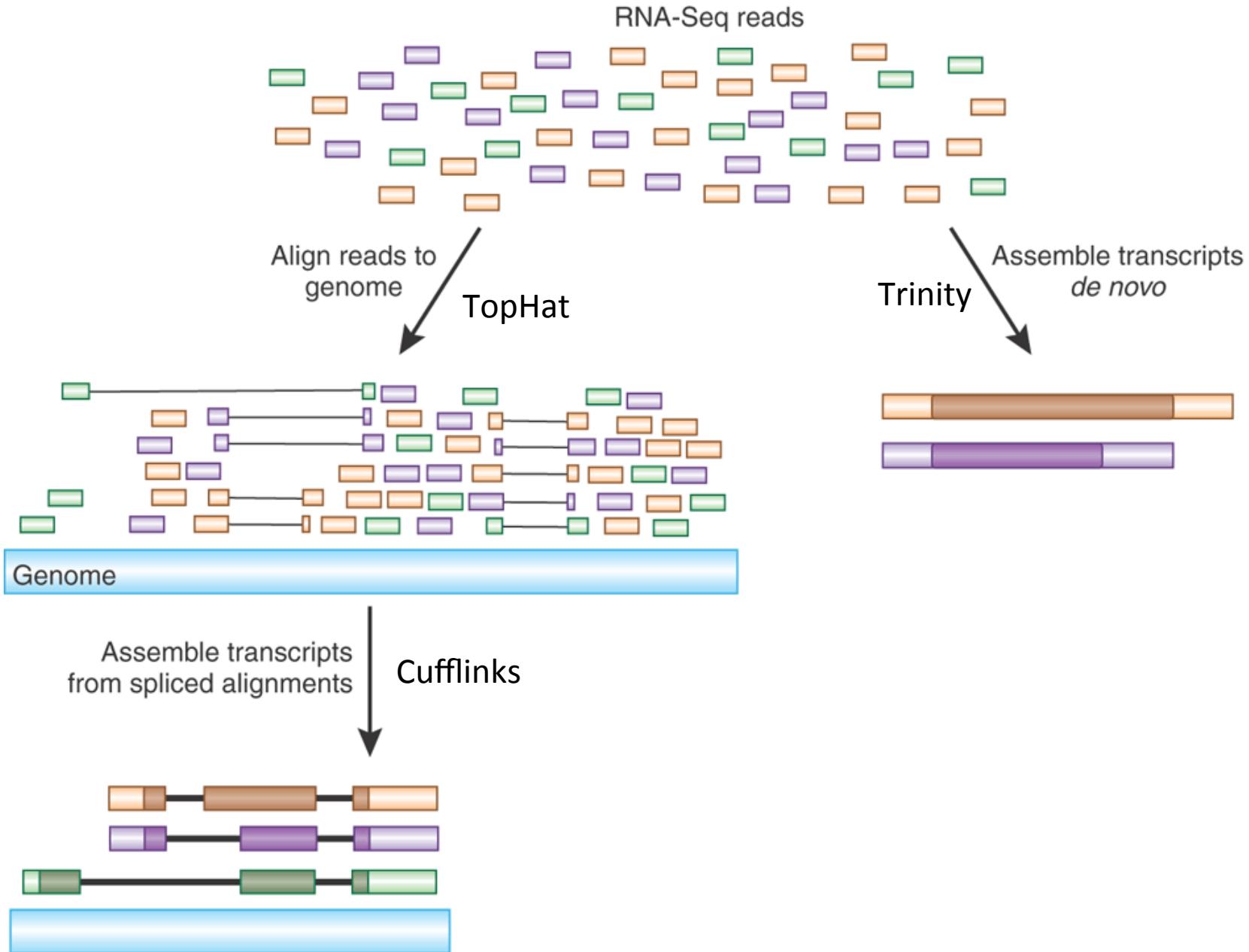
Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn & Lior Pachter

Affiliations | Contributions | Corresponding author

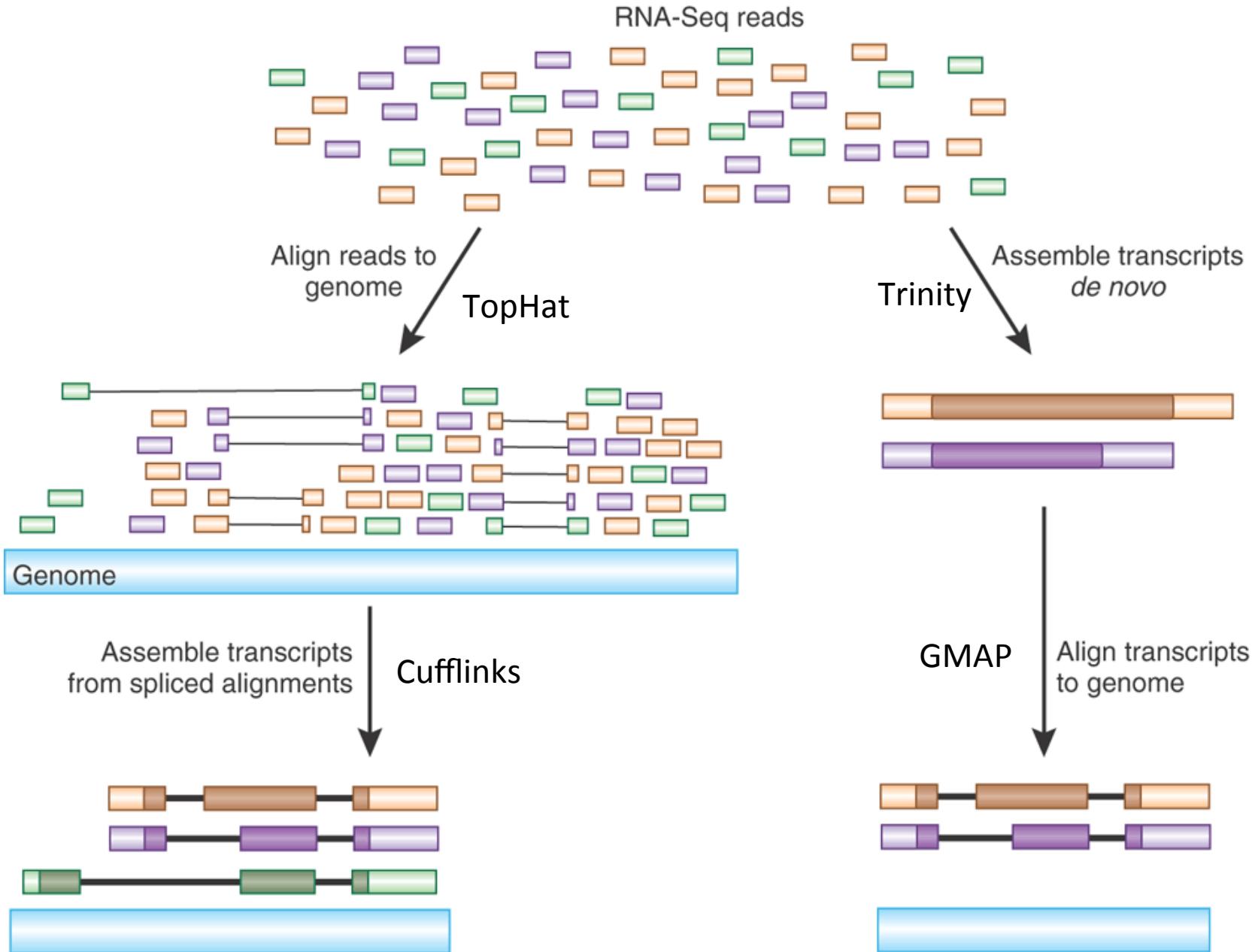
Nature Protocols 7, 562–578 (2012) | doi:10.1038/nprot.2012.016

Published online 01 March 2012

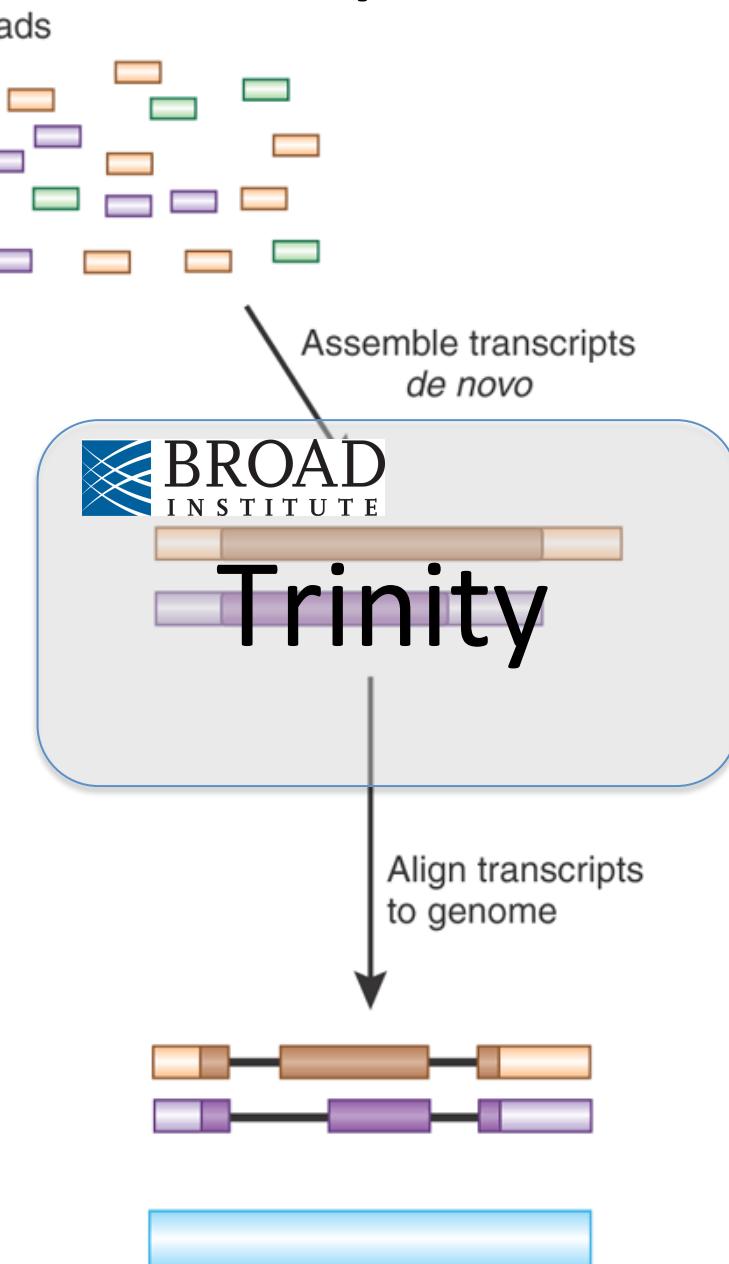
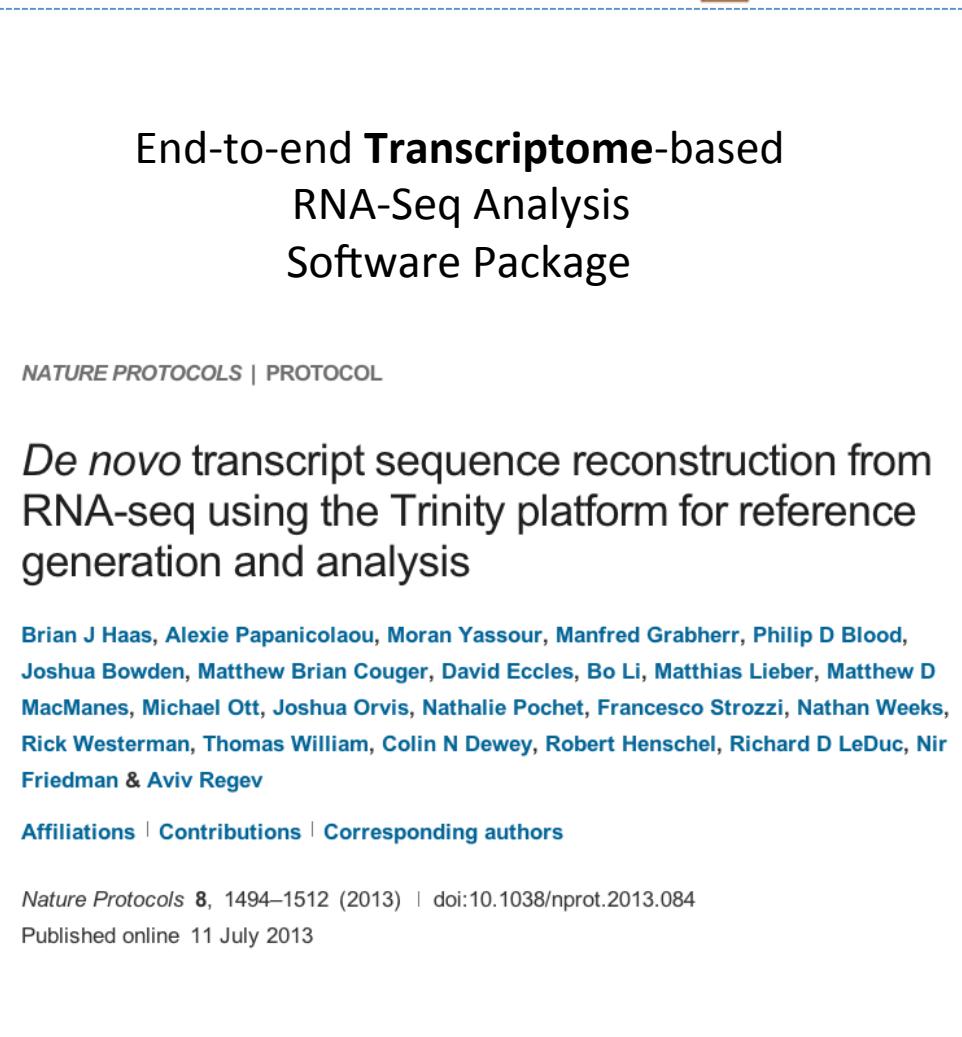
Transcript Reconstruction from RNA-Seq Reads



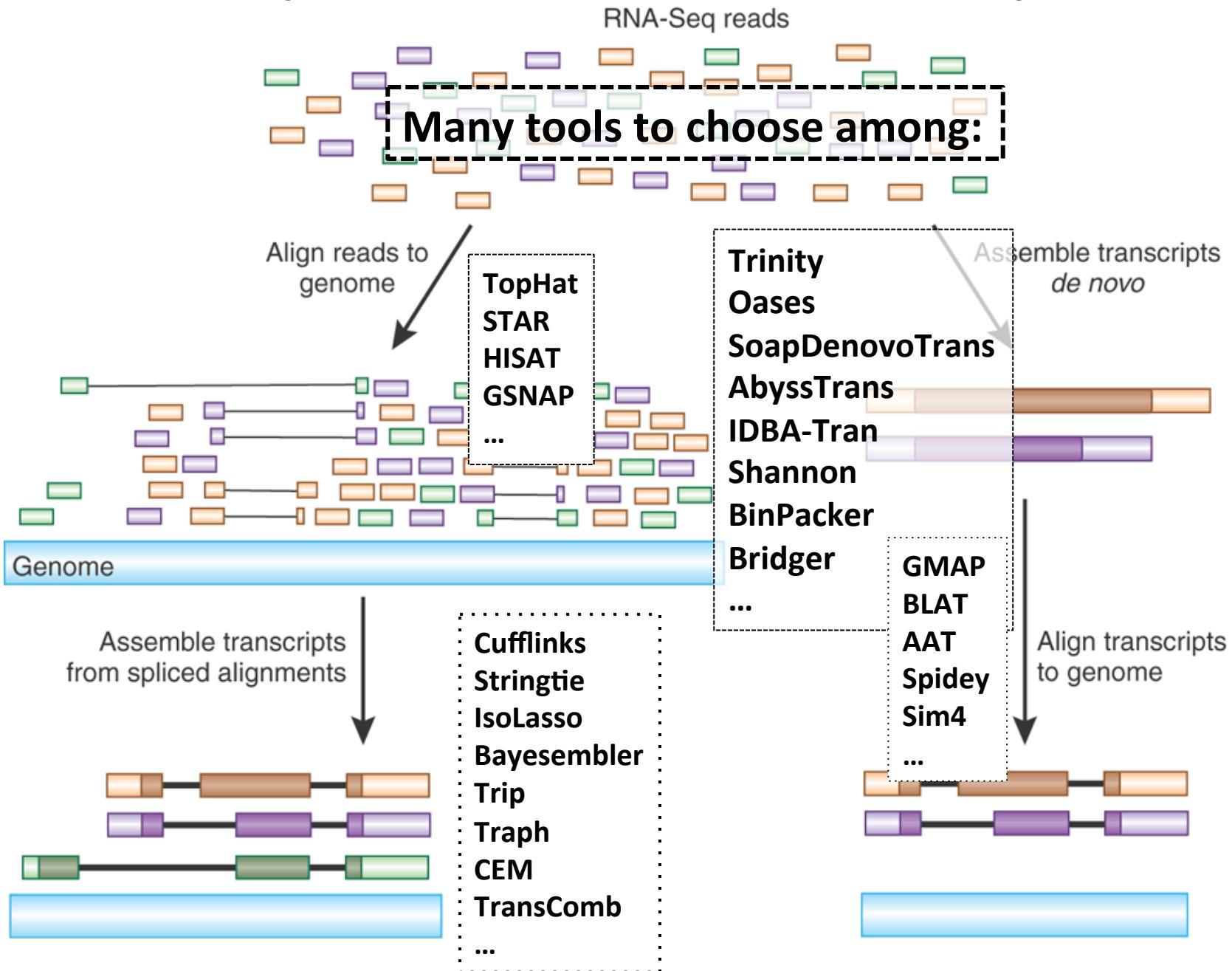
Transcript Reconstruction from RNA-Seq Reads



Transcript Reconstruction from RNA-Seq Reads



Transcript Reconstruction from RNA-Seq Reads



Overview of the Tuxedo Software Suite

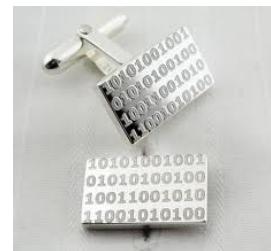
Bowtie (fast short-read alignment)



TopHat (spliced short-read alignment)



Cufflinks (transcript reconstruction from alignments)

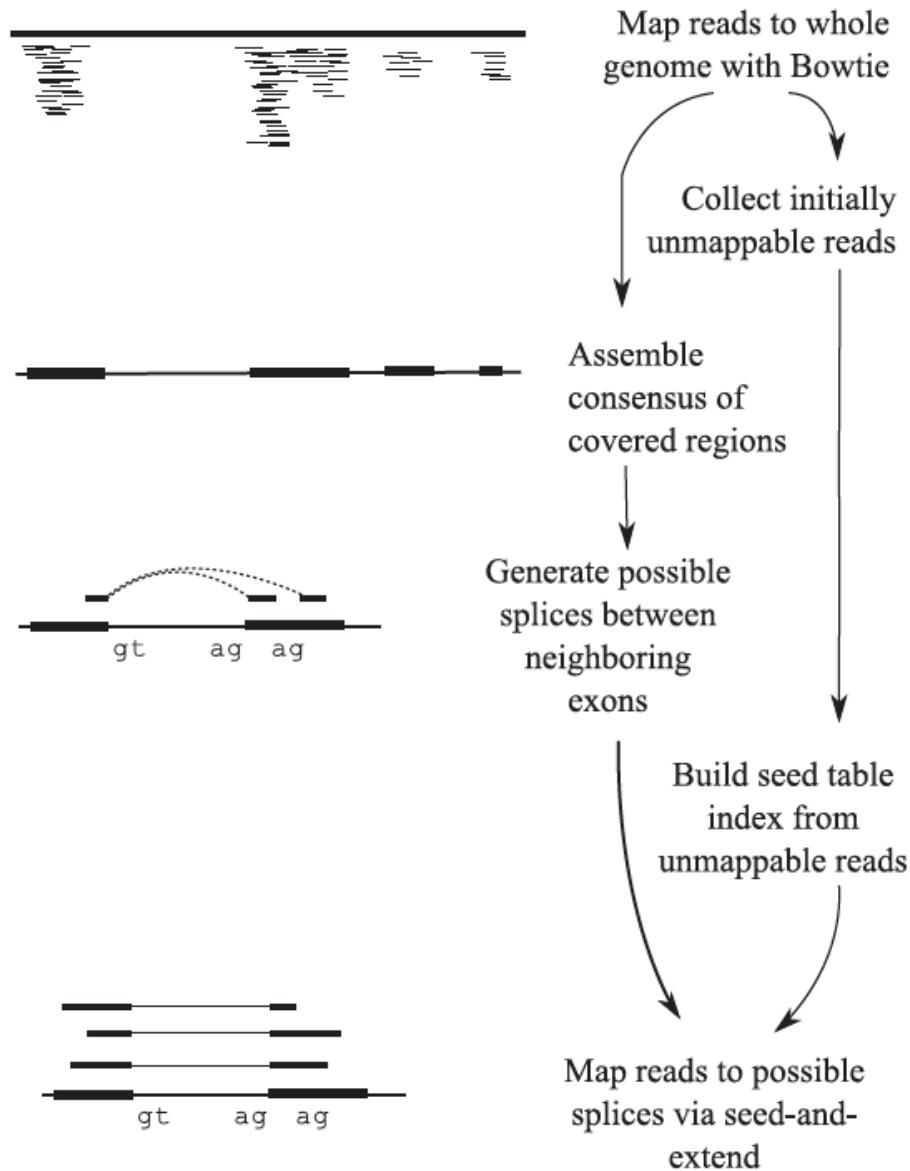


Cuffdiff (differential expression analysis)



CummeRbund (visualization & analysis)

The TopHat Pipeline



Alignments are reported in a compact representation: SAM format

```
0      61G9EAAXX100520:5:100:10095:16477
1      83
2      chr1
3      51986
4      38
5      46M
6      =
7      51789
8      -264
9      CCCAAACAAGCCGAACTAGCTGATTGGCTCGTAAAGACCCGGAAA
10     ## #CB?=ADDDBCBCDEEFFDEFFFDEFFGDBEFGEDGCFGFGGGGG
11     MD:Z:67
12     NH:i:1
13     HI:i:1
14     NM:i:0
15     SM:i:38
16     XQ:i:40
17     X2:i:0
```

SAM format specification: <http://samtools.sourceforge.net/SAM1.pdf>

Alignments are reported in a compact representation: SAM format

```
0      61G9EAAXX100520:5:100:10095:16477 (read name)
1      83  (FLAGS stored as bit fields; 83 = 00001010011 )
2      chr1 (alignment target)
3      51986 (position alignment starts)
4      38
5      46M (Compact description of the alignment in CIGAR format)
6      =
7      51789
8      -264 → (read sequence, oriented according to the forward alignment)
9      CCCAAACAAGCCGAACTAGCTGATTGGCTCGTAAAGACCCGGAAA
10     ## #CB?=ADDBCBCDEEFFDEFFFDEFFGDBEFGEDGCFGFGGGGG
11     MD:Z:67                                     → (base quality values)
12     NH:i:1
13     HI:i:1
14     NM:i:0
15     SM:i:38          (Metadata)
16     XQ:i:40
17     X2:i:0
```

Alignments are reported in a compact representation: SAM format

```
0      61G9EAAXX100520:5:100:10095:16477 (read name)
1      83  (FLAGS stored as bit fields; 83 = 00001010011 )
2      chr1 (alignment target)
```

Still not compact enough...

Millions to billions of reads takes up a lot of space!!

Convert SAM to binary – BAM format.

```
15     SM:i:38      (metadata)
16     XQ:i:40
17     X2:i:0
```

SAM format specification: <http://samtools.sourceforge.net/SAM1.pdf>

Samtools

- Tools for
 - converting SAM <-> BAM
 - Viewing BAM files (eg. samtools view file.bam | less)
 - Sorting BAM files, and lots more:

```
Program: samtools (Tools for alignments in the SAM format)
Version: 0.1.19-44428cd

Usage:   samtools <command> [options]

Command: view      SAM<->BAM conversion
          sort      sort alignment file
          mpileup   multi-way pileup
          depth     compute the depth
          faidx    index/extract FASTA
          index    index alignment
          idxstats  BAM index stats (r595 or later)
          fixmate   fix mate information
          flagstat  simple stats
          calmd     recalculate MD/NM tags and '=' bases
          merge     merge sorted alignments
          rmdup    remove PCR duplicates
          reheader  replace BAM header
          cat       concatenate BAMs
          bedcov    read depth per BED region
          targetcut cut fosmid regions (for fosmid pool only)
          phase    phase heterozygotes
          bamshuf  shuffle and group alignments by name
```

Visualizing Alignments of RNA-Seq reads

Text-based Alignment Viewer

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

IGV

www.broadinstitute.org/igv/

igv Integrative Genomics Viewer ALMEL

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

Search website

search

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

BROAD INSTITUTE
© 2012 Broad Institute

Home

Integrative Genomics Viewer



What's New

NEWS 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.](#)

IGV: Viewing Tophat Alignments

IGV

File Genomes View Tracks Regions Tools GenomeSpace Help

human_cancer_fusi...

NOTCH1-NUP214

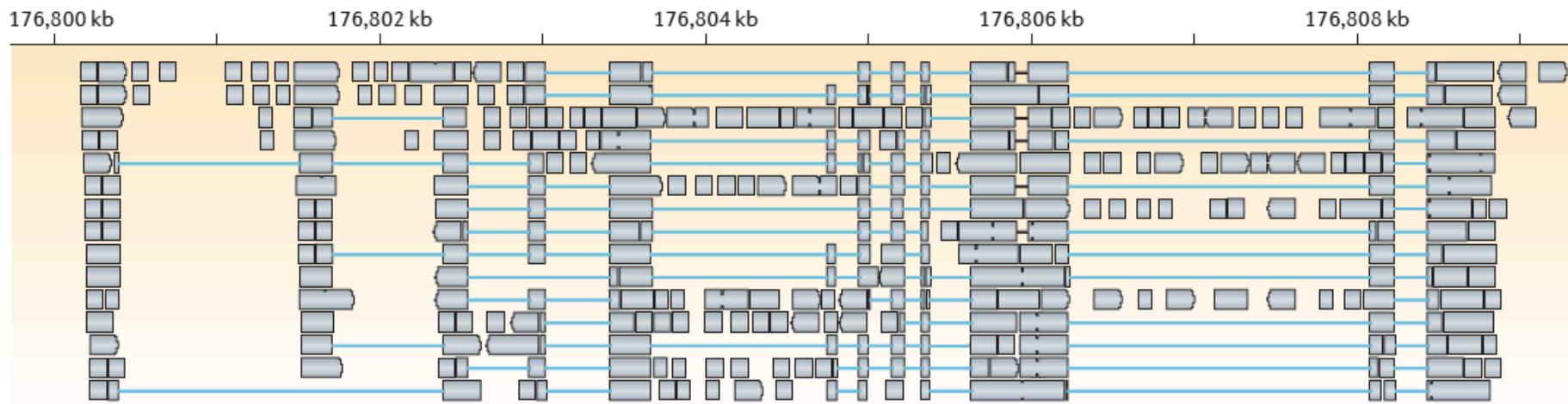
NOTCH1-NUP214:73,649-91,059

Go



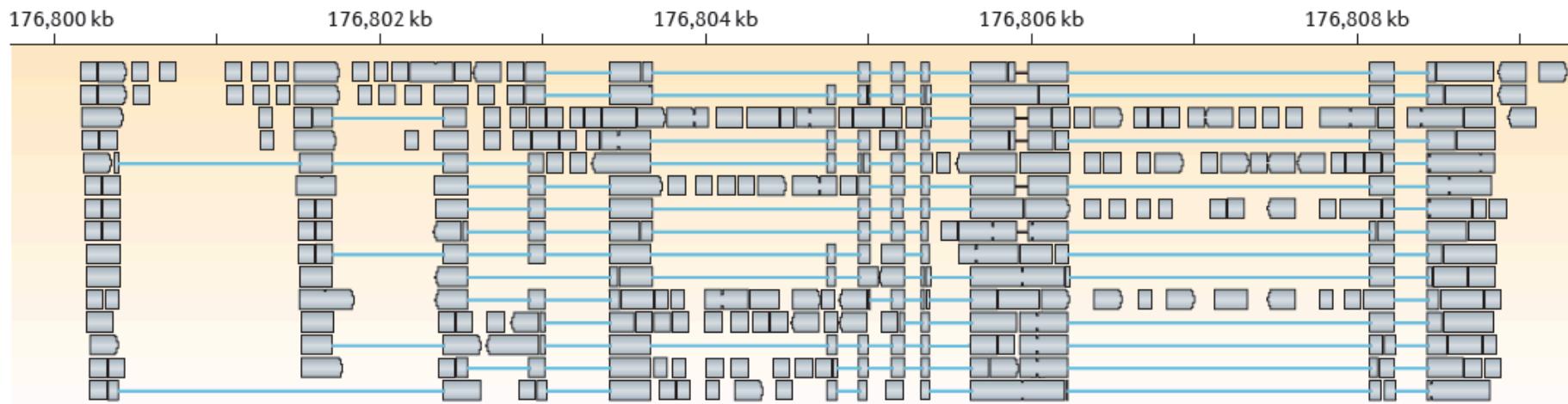
Transcript Reconstruction Using Cufflinks

a Splice-align reads to the genome

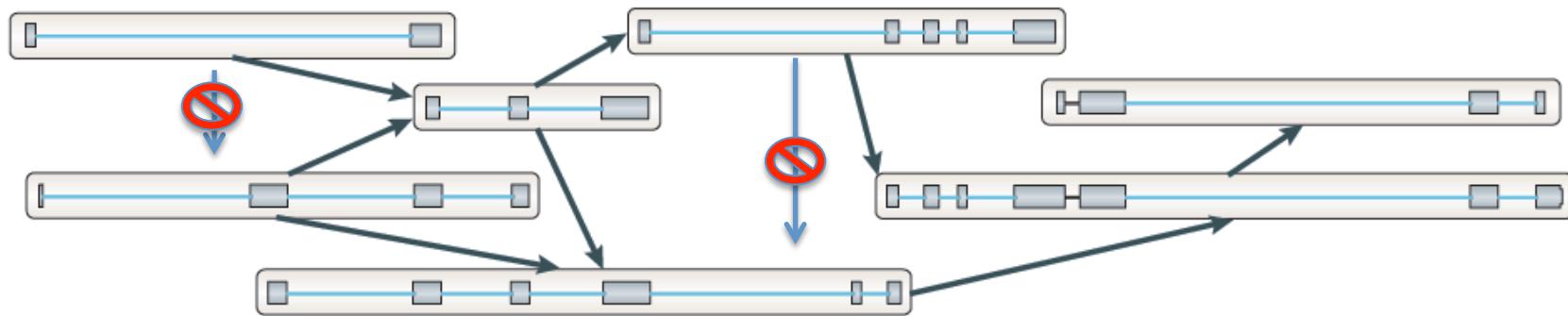


Transcript Reconstruction Using Cufflinks

a Splice-align reads to the genome

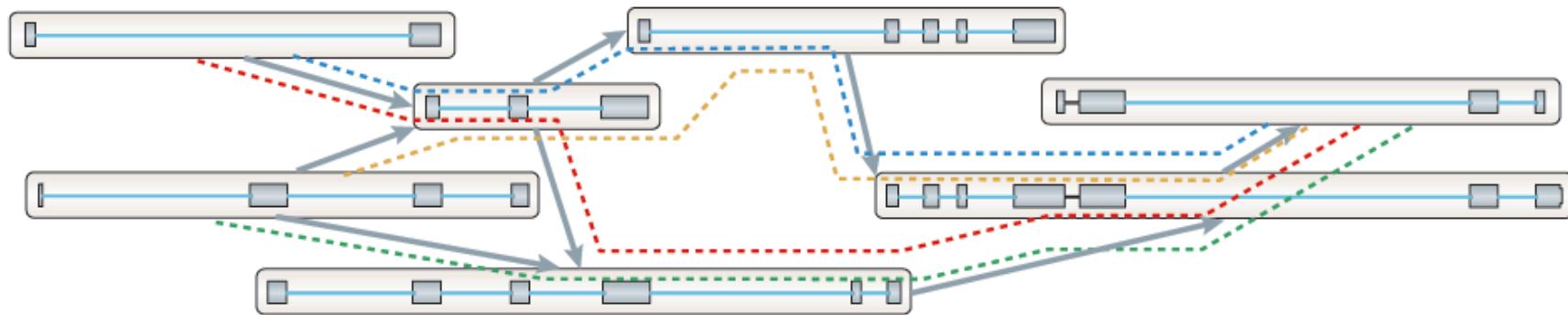


b Build a graph representing alternative splicing events

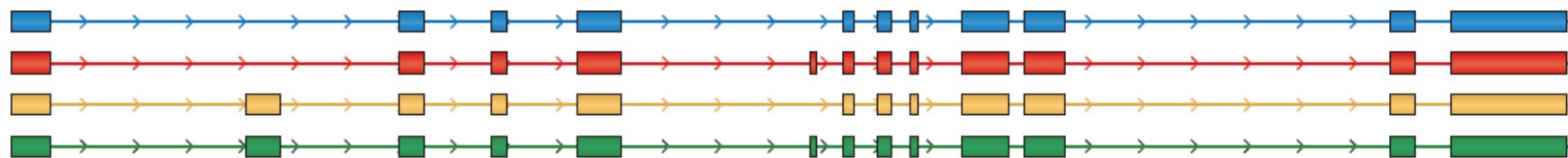


Transcript Reconstruction Using Cufflinks

c Traverse the graph to assemble variants



d Assembled isoforms



Transcript Structures in GTF Format

(tab-delimited fields per line shown transposed to a column format here)

```
0  chr1          (genomic contig identifier)
1  Cufflinks
2  transcript
3  101  (left coordinate)
4  5716 (right coordinate)
5  1000
6  +    (strand)
7  .
8  gene_id "CUFF.1"; transcript_id "CUFF.1.1"; FPKM "378.0239937260"  (annotations)
```

```
0  chr1
1  Cufflinks
2  exon
3  101
4  5716
5  1000
6  +
7  .
8  gene_id "CUFF.1"; transcript_id "CUFF.1.1"; exon_number "1"; FPKM "378.0239937260"
```

Demo: Tuxedo and IGV

https://github.com/trinityrnaseq/RNASeq_Trinity_Tuxedo_Workshop/wiki/Tuxedo-Genome-Guided-Transcriptome-Assembly-Workshop

- Run Tophat to align reads to the genome
- Reconstruct transcripts using cufflinks
- View genome-aligned reads and reconstructed transcripts using IGV

De novo transcriptome assembly

No genome required

Empower studies of non-model organisms

- expressed gene content
- transcript abundance
- differential expression

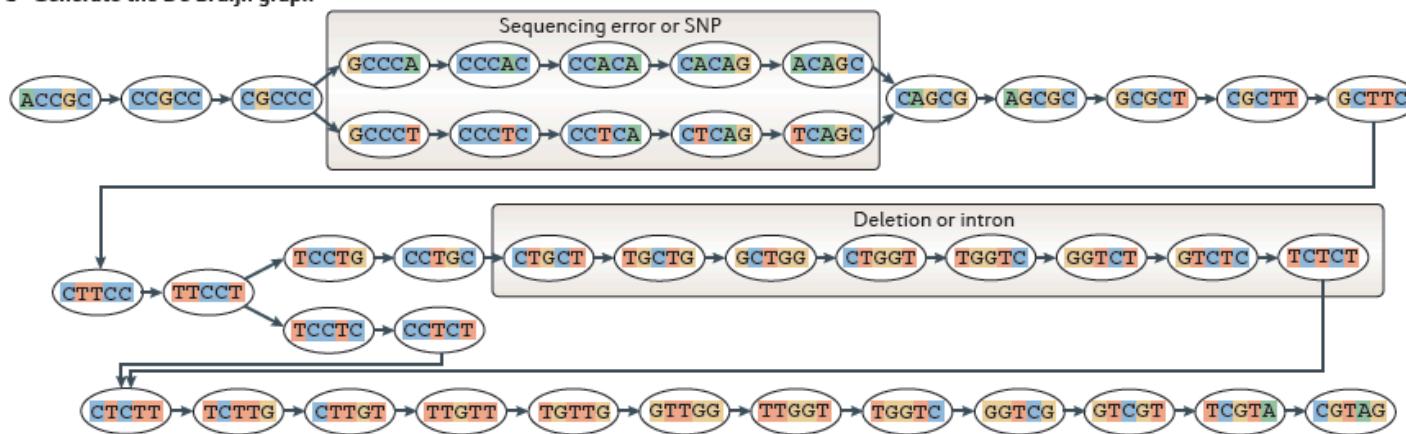
The General Approach to
De novo RNA-Seq Assembly
Using De Bruijn Graphs

Sequence Assembly via De Bruijn Graphs

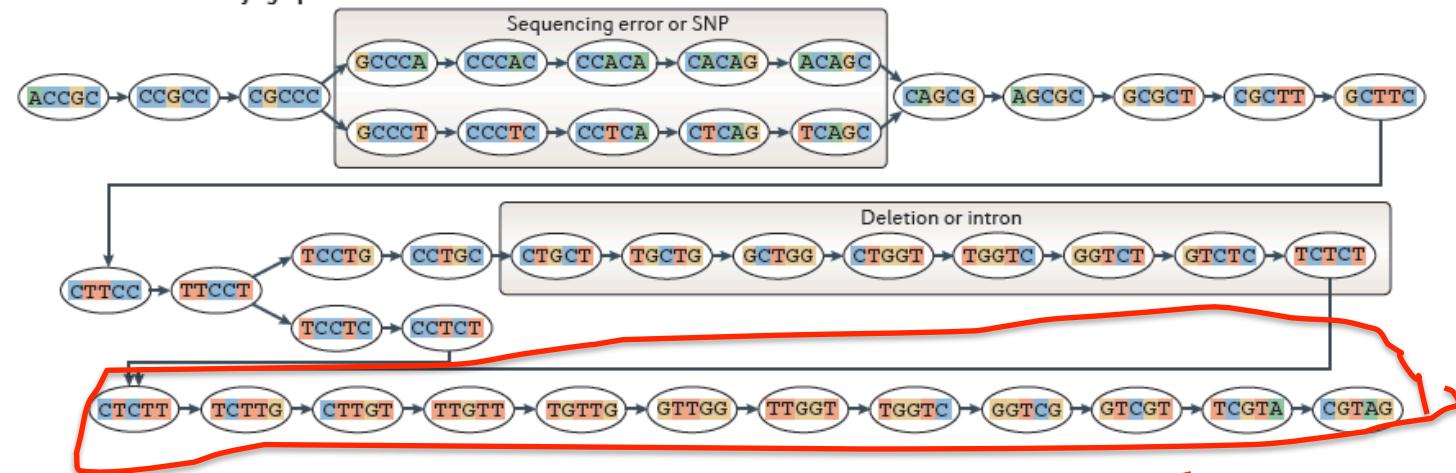
a Generate all substrings of length k from the reads

ACAGC	TCC TG	GT CTC		AGCGC	CT CTT	GG TCG	k-mers (k=5)
CACAG	TTC CCT	GGT CT		CAGCG	CCT CT	TGG TC	
CCACA	CTT CC	TGG TC	TG TTG	TCAGC	TC CTC	TT GGT	
CCCAC	GCT TC	CTGGT	TT GTT	CTC AG	TT CCT	GTT GG	
GCCCA	CG CTT	GCT GG	CTT GT	CCT CA	CTT CC	TG TTG	
CGCCC	GCG CT	TG CTG	TCT TG	CCCTC	GCT TC	TT GTT	
CCGCC	AGCGC	CTG CT	CT CTT	GCC CT	CG CTT	CTT GT	
ACCGC	CAG CG	CCT GC	TCT CT	CGCCC	GCG CT	TCT TG	
ACCGCCCCACAGCGCTTCCTGCTGGTCTCTTGTG				CGCCCTCAGCGCTTCCTCTTGTTGGTCGTAG			Reads

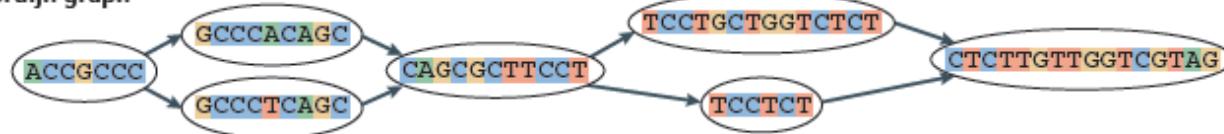
b Generate the De Bruijn graph



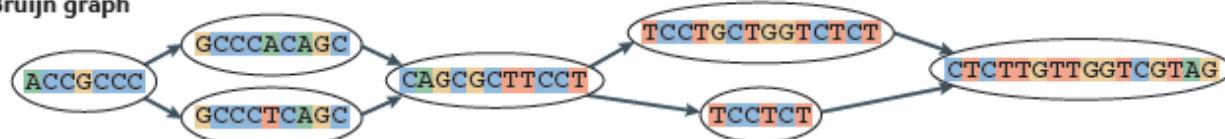
b Generate the De Bruijn graph



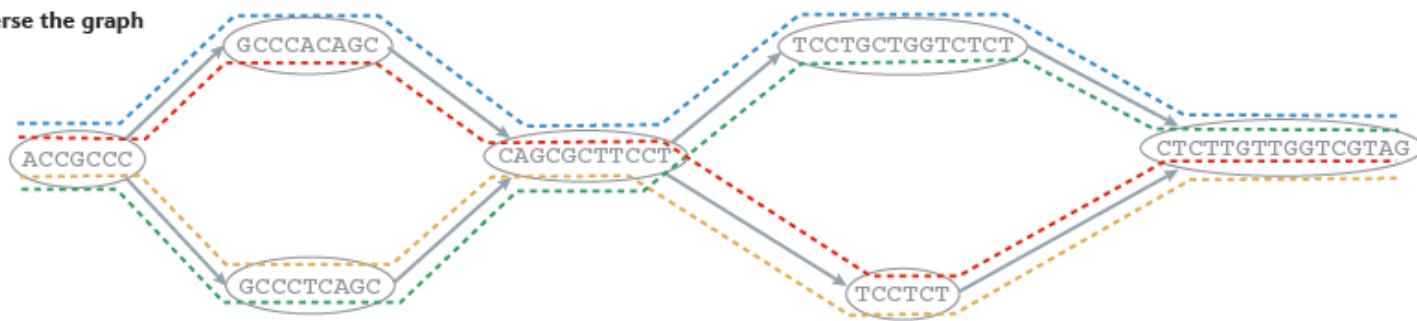
c Collapse the De Bruijn graph



c Collapse the De Bruijn graph



d Traverse the graph



e Assembled isoforms

— ACCGCCACAGCGCTTCCTGCTGGTCTCTTGGTGGT CGTAG
- - - ACCGCCACAGCGCTTCCT - - - CTTGTTGGT CGTAG
--- ACCGCCCTCAGCGCTTCCT - - - CTTGTTGGT CGTAG
- - - ACCGCCCTCAGCGCTTCCTGCTGGTCTCTTGGTGGT CGTAG

Contrasting Genome and Transcriptome Assembly

Genome Assembly

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

Transcriptome Assembly

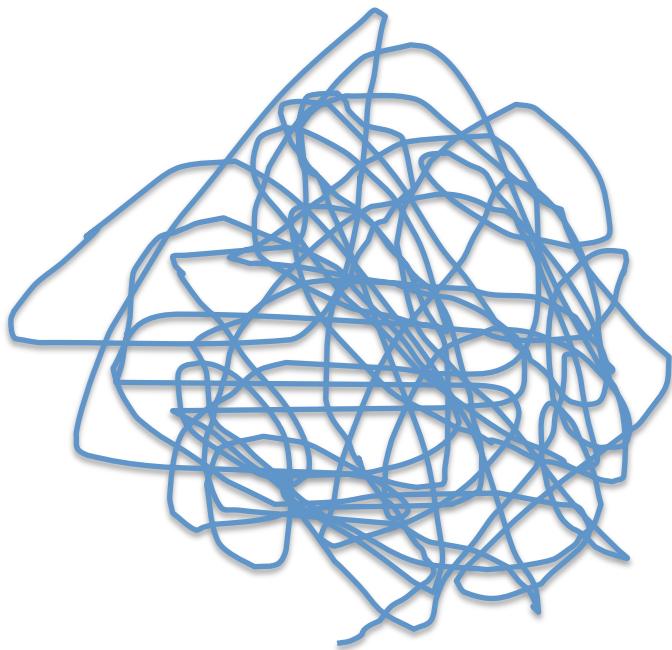
- Exponentially distributed coverage levels
- Multiple contigs per locus (alt splicing)
- Strand-specific



Trinity Aggregates Isolated Transcript Graphs

Genome Assembly

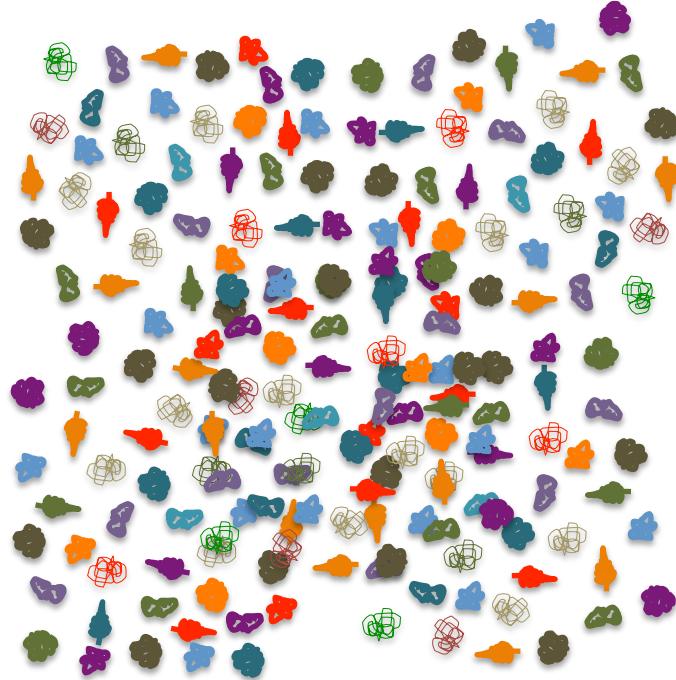
Single Massive Graph



Entire chromosomes represented.

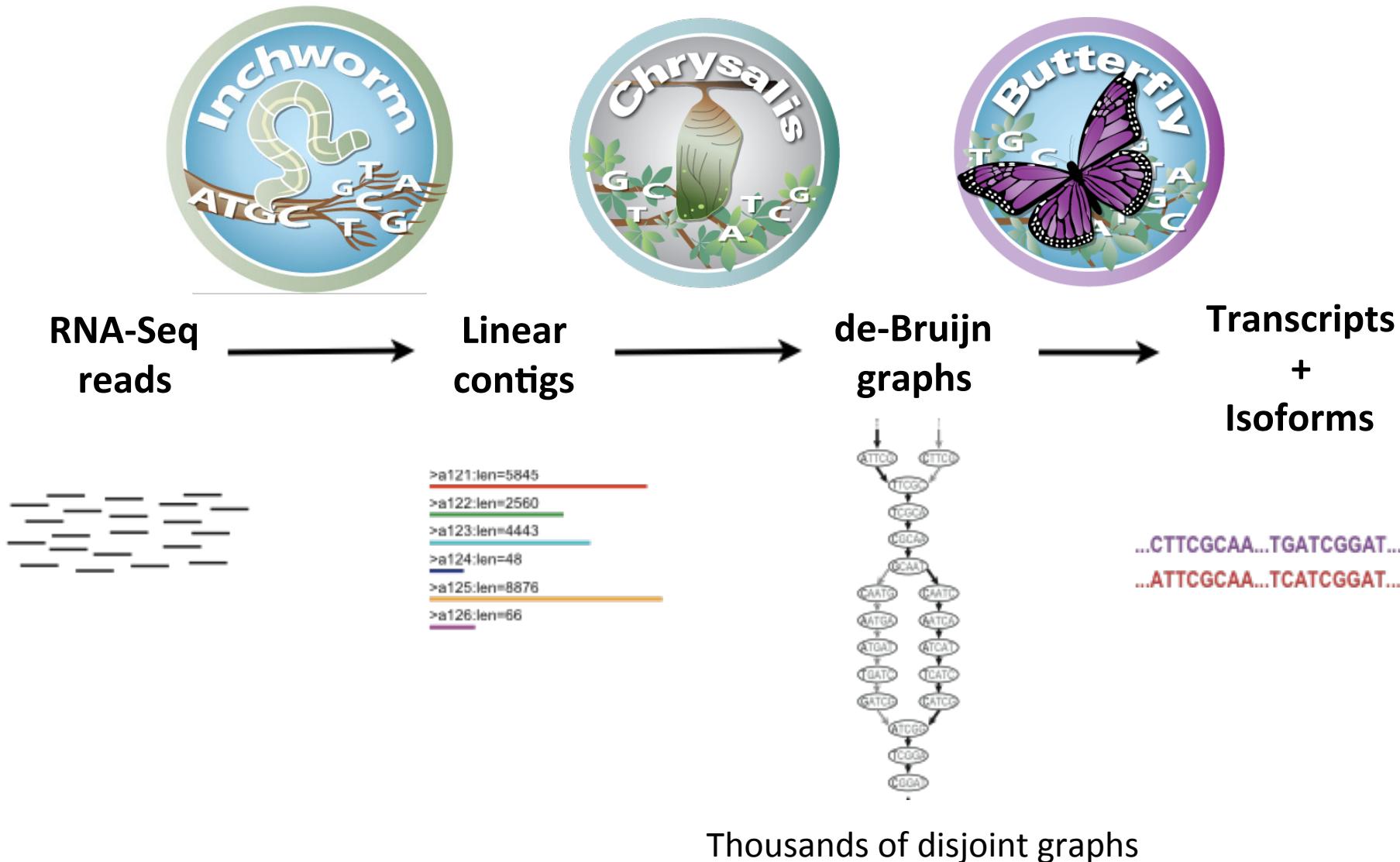
Trinity Transcriptome Assembly

Many Thousands of Small Graphs



Ideally, one graph per expressed gene.

Trinity – How it works:



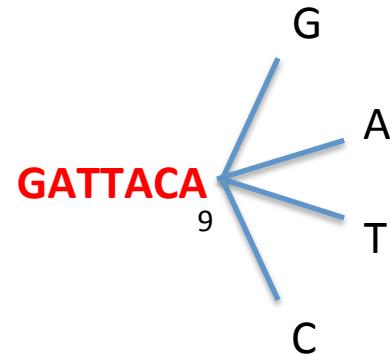


Inchworm Algorithm

Decompose all reads into overlapping Kmers (25-mers)

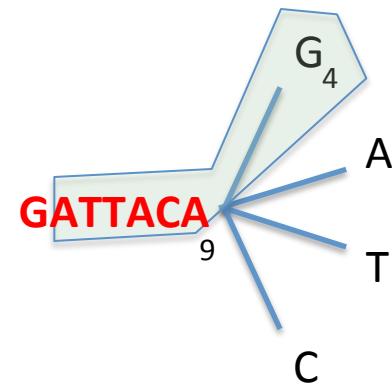
Identify seed kmer as most abundant Kmer, ignoring low-complexity kmers.

Extend kmer at 3' end, guided by coverage.



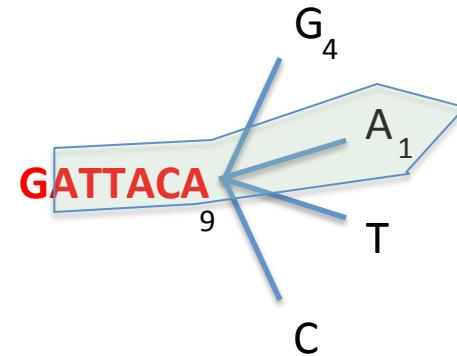


Inchworm Algorithm



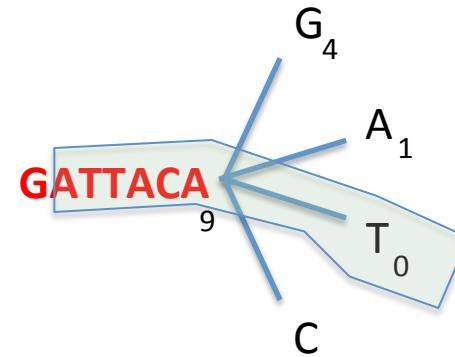


Inchworm Algorithm



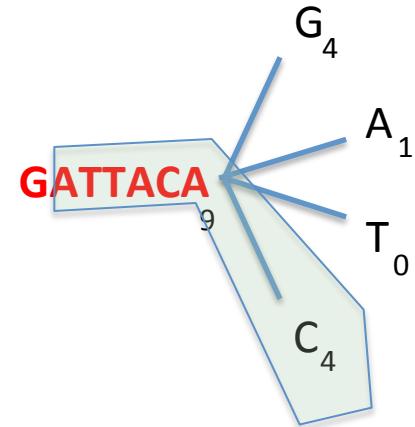


Inchworm Algorithm



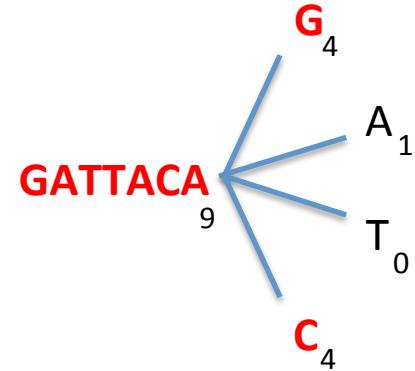


Inchworm Algorithm



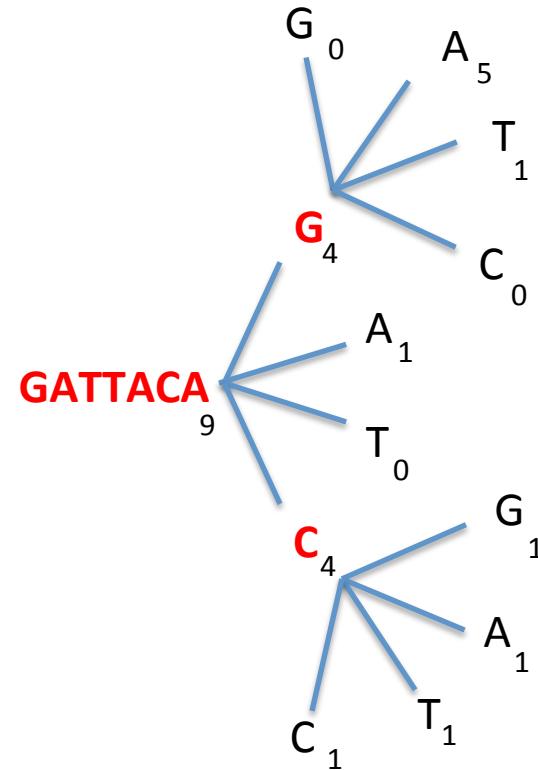


Inchworm Algorithm



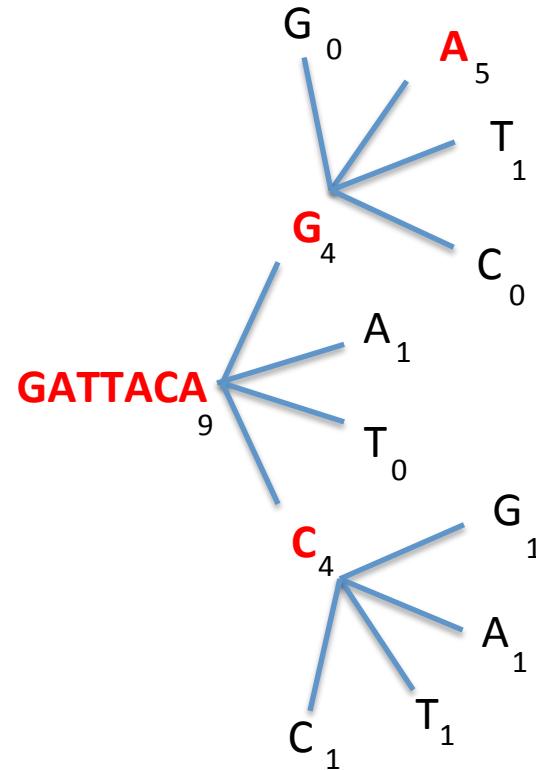


Inchworm Algorithm



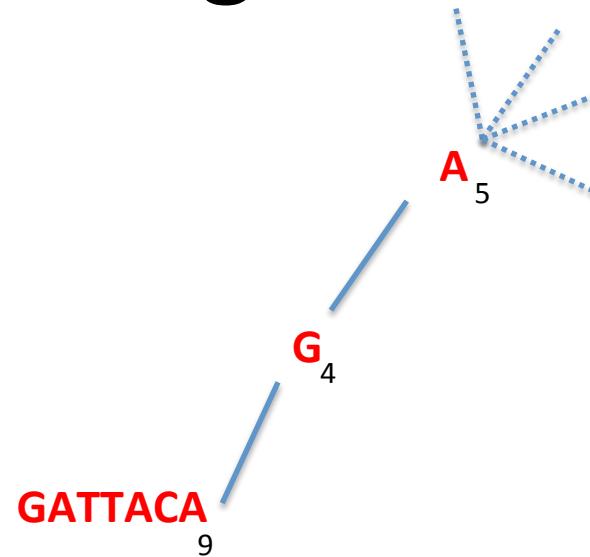


Inchworm Algorithm



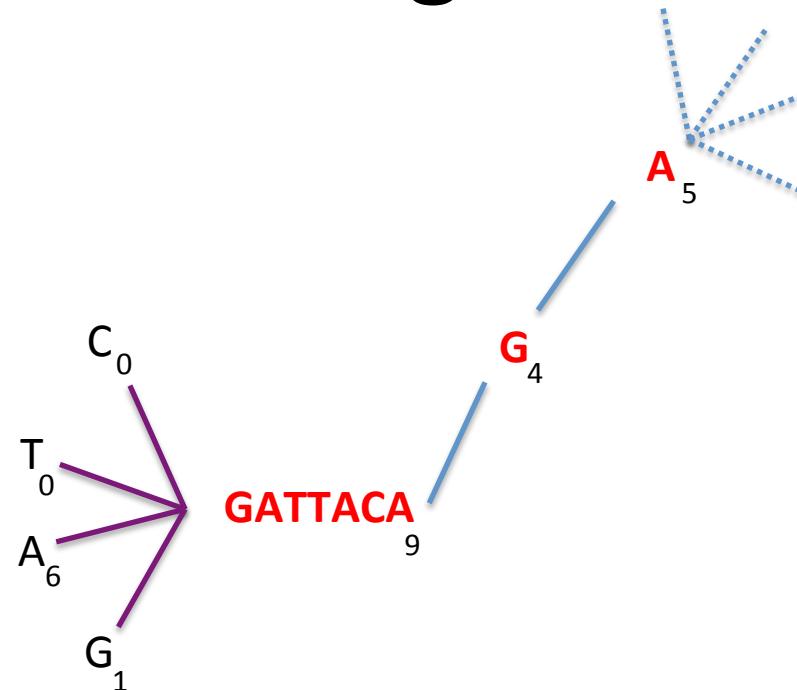


Inchworm Algorithm



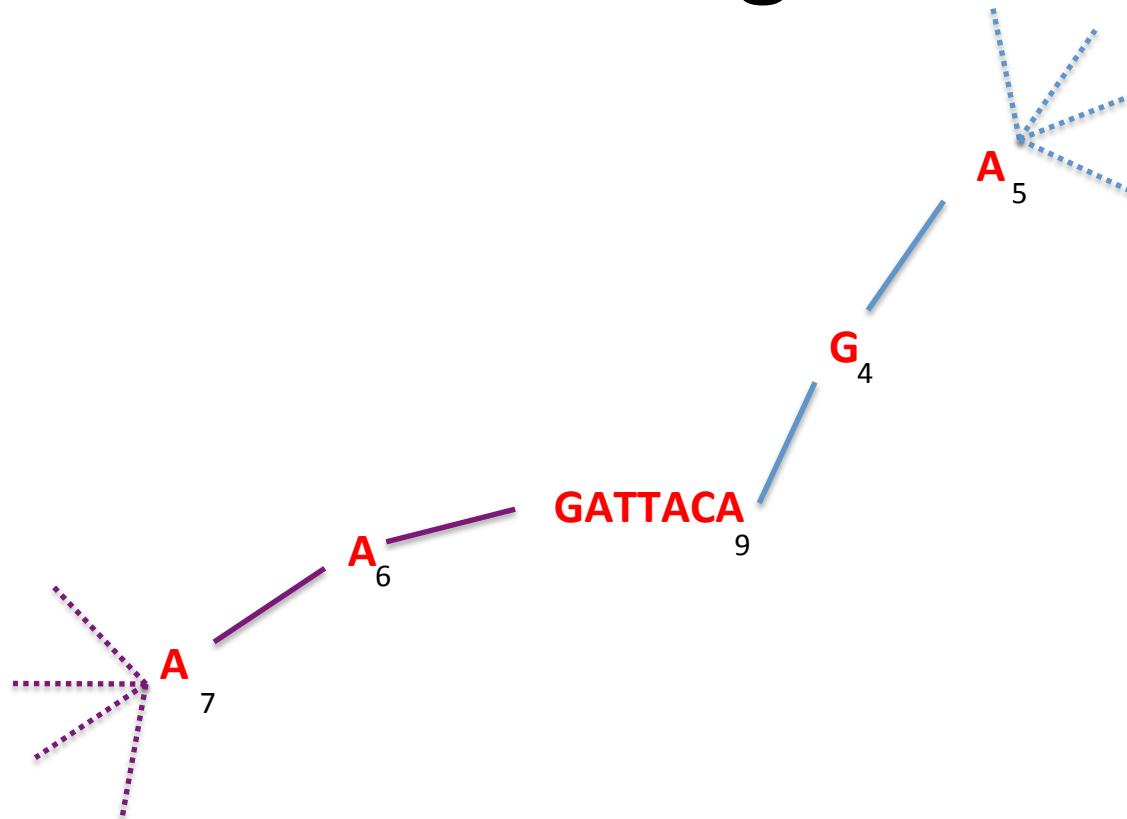


Inchworm Algorithm





Inchworm Algorithm



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

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



Inchworm Contigs from Alt-Spliced Transcripts

Expressed isoforms





Inchworm Contigs from Alt-Spliced Transcripts

Expressed isoforms

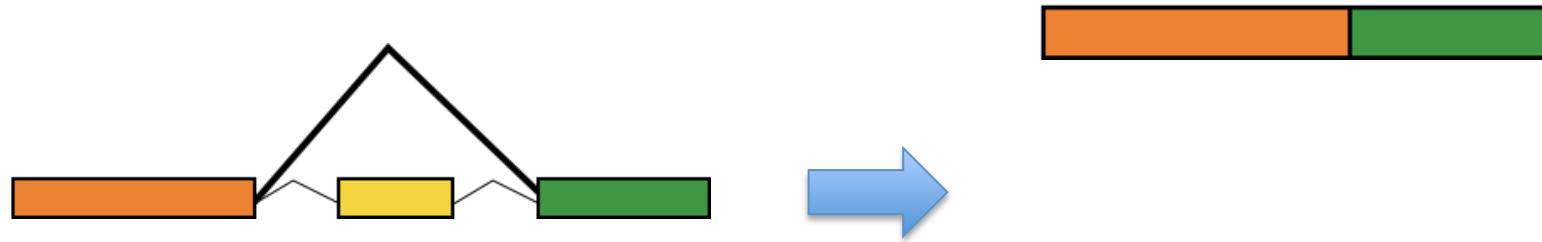


Graphical representation



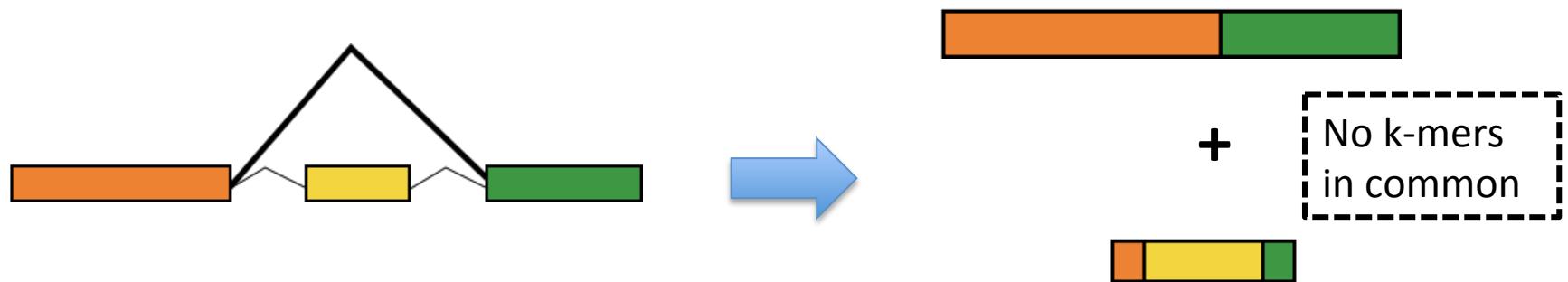


Inchworm Contigs from Alt-SPLICED Transcripts



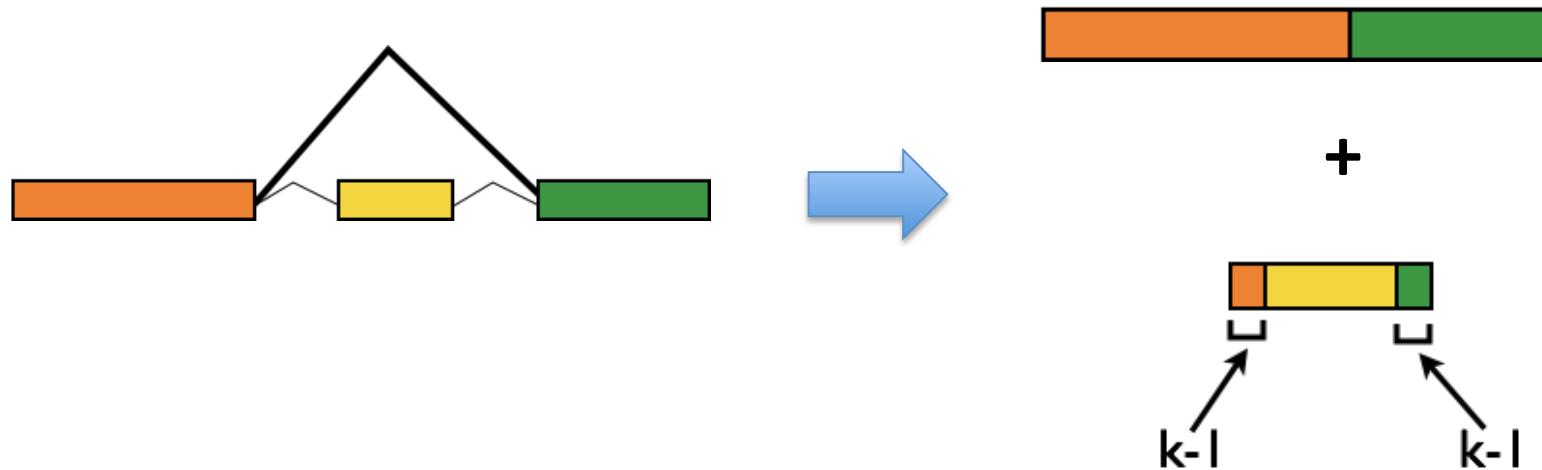


Inchworm Contigs from Alt-Spliced Transcripts

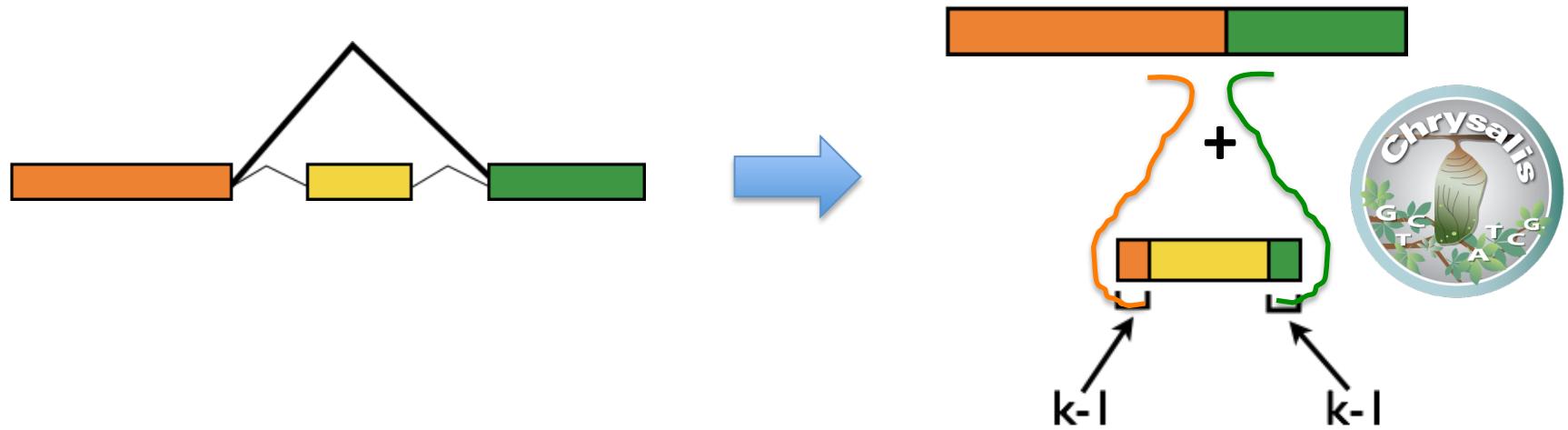




Inchworm Contigs from Alt-Spliced Transcripts



Chrysalis Re-groups Related Inchworm Contigs



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

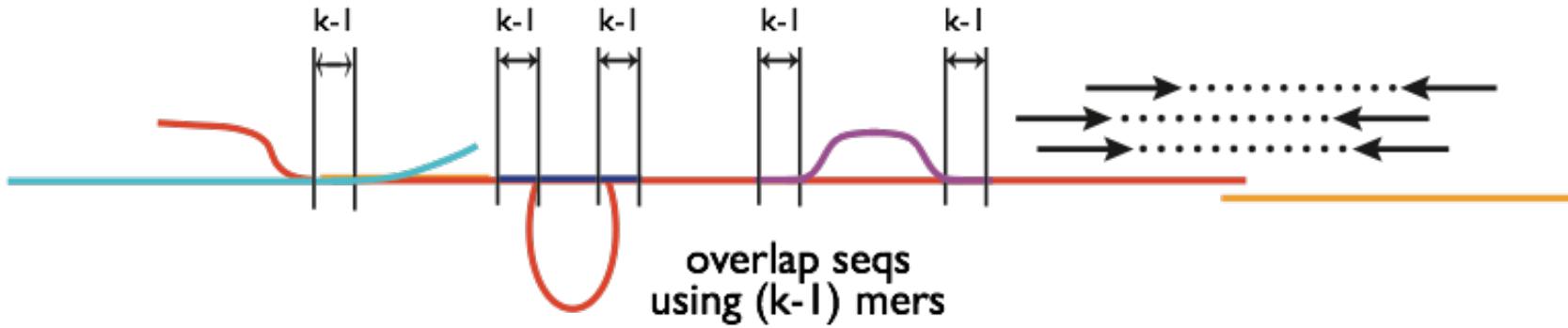
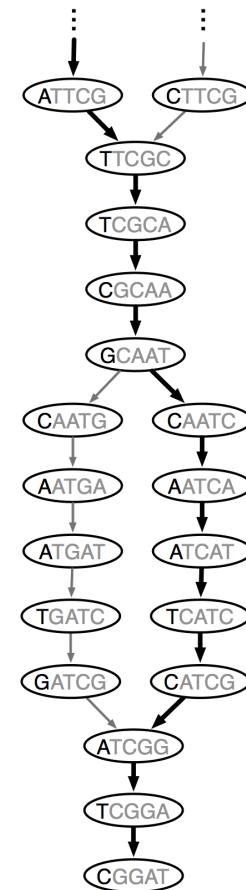
Chrysalis

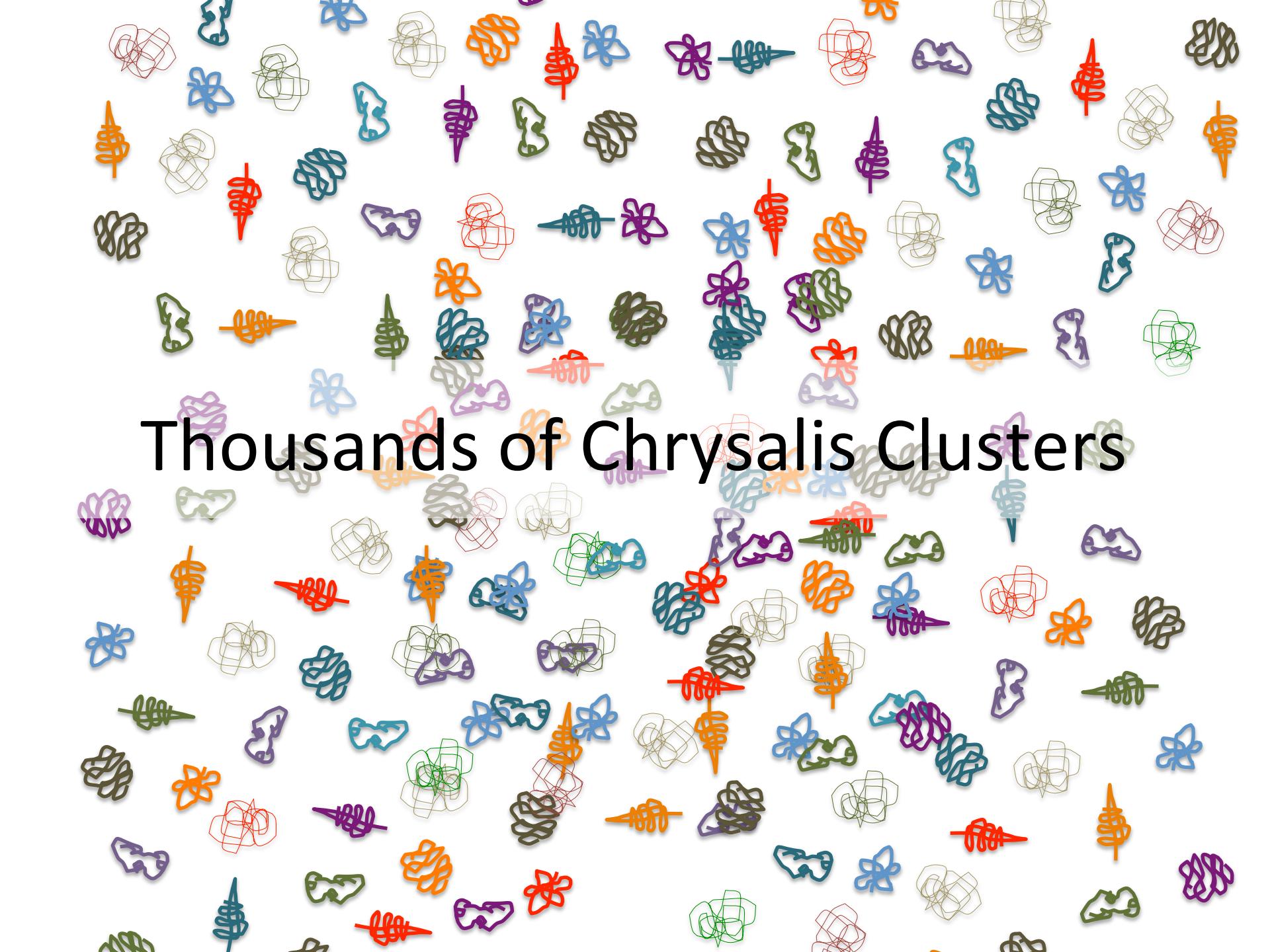
```
>a121:len=5845  
+-----+  
>a122:len=2560  
+-----+  
>a123:len=4443  
+-----+  
>a124:len=48  
+-----+  
>a125:len=8876  
+-----+  
>a126:len=66  
+-----+
```

Integrate isoforms
via k-1 overlaps

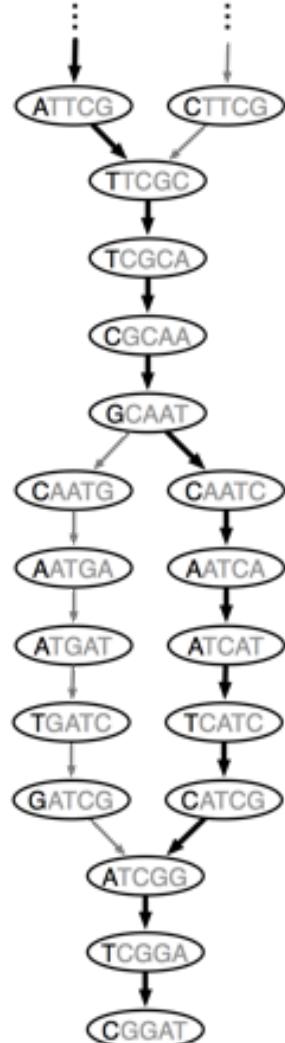


Build de Bruijn Graphs
(ideally, one per gene)



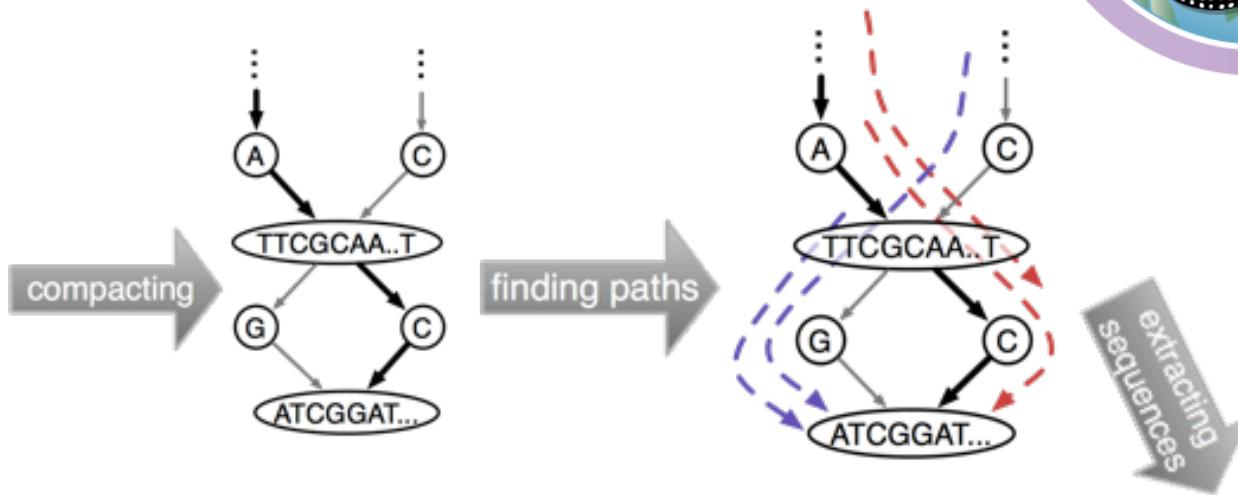


Thousands of Chrysalis Clusters



de Bruijn
graph

Butterfly



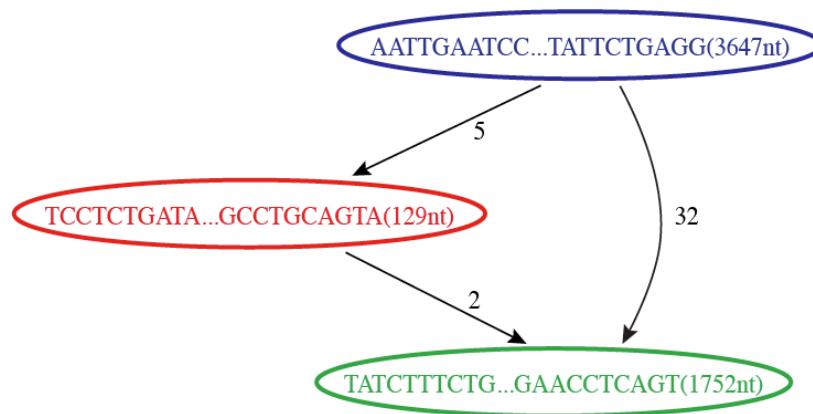
compact
graph

compact
graph with
reads

..CTTCGCAA..TGATCGGAT...
..ATTCGCAA..TCATCGGAT...
sequences
(isoforms and paralogs)

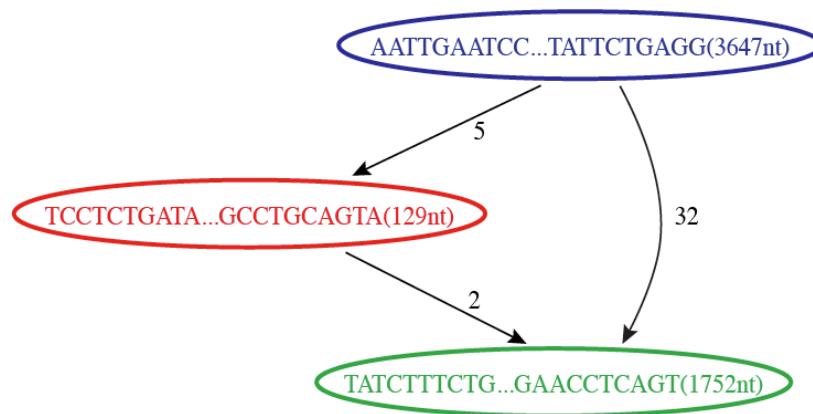
Butterfly Example 1: Reconstruction of Alternatively Spliced Transcripts

Butterfly's Compacted
Sequence Graph



Reconstruction of Alternatively Spliced Transcripts

Butterfly's Compacted Sequence Graph

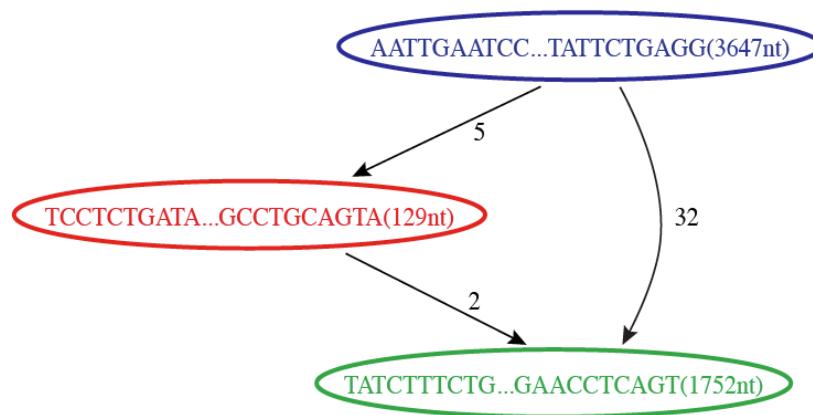


Reconstructed Transcripts



Reconstruction of Alternatively Spliced Transcripts

Butterfly's Compacted Sequence Graph

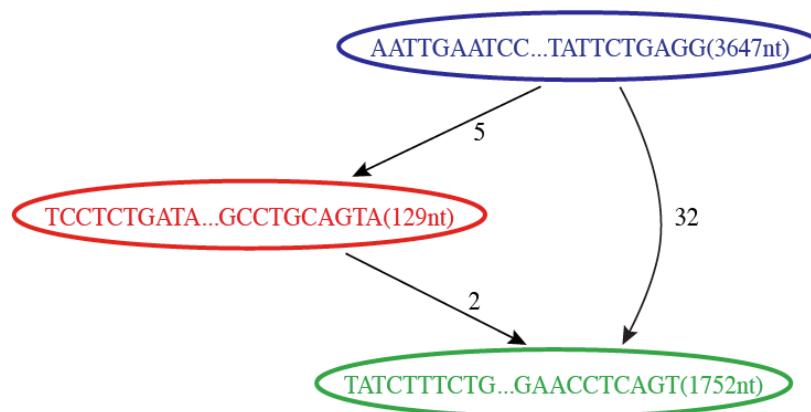


Reconstructed Transcripts



Reconstruction of Alternatively Spliced Transcripts

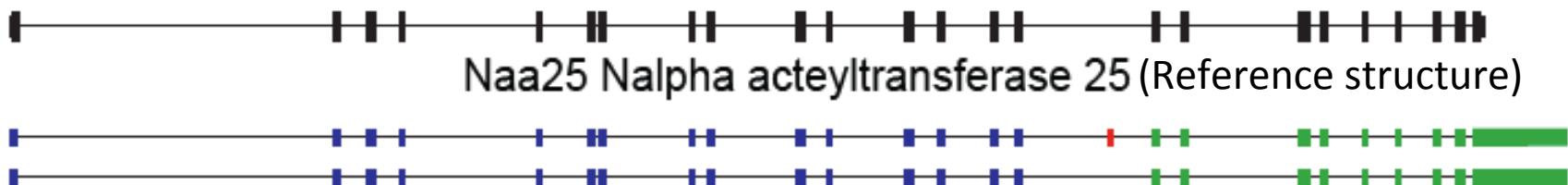
Butterfly's Compacted Sequence Graph



Reconstructed Transcripts



Aligned to Mouse Genome



Butterfly Example 2: Teasing Apart Transcripts of Paralogous Genes



Teasing Apart Transcripts of Paralogous Genes

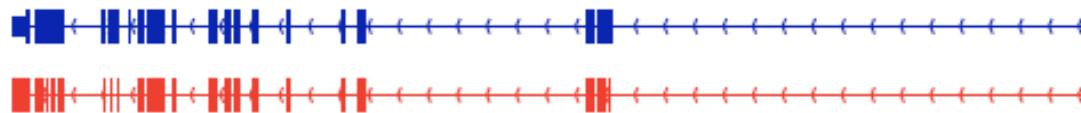
chr7:148,744,197-148,821,437

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



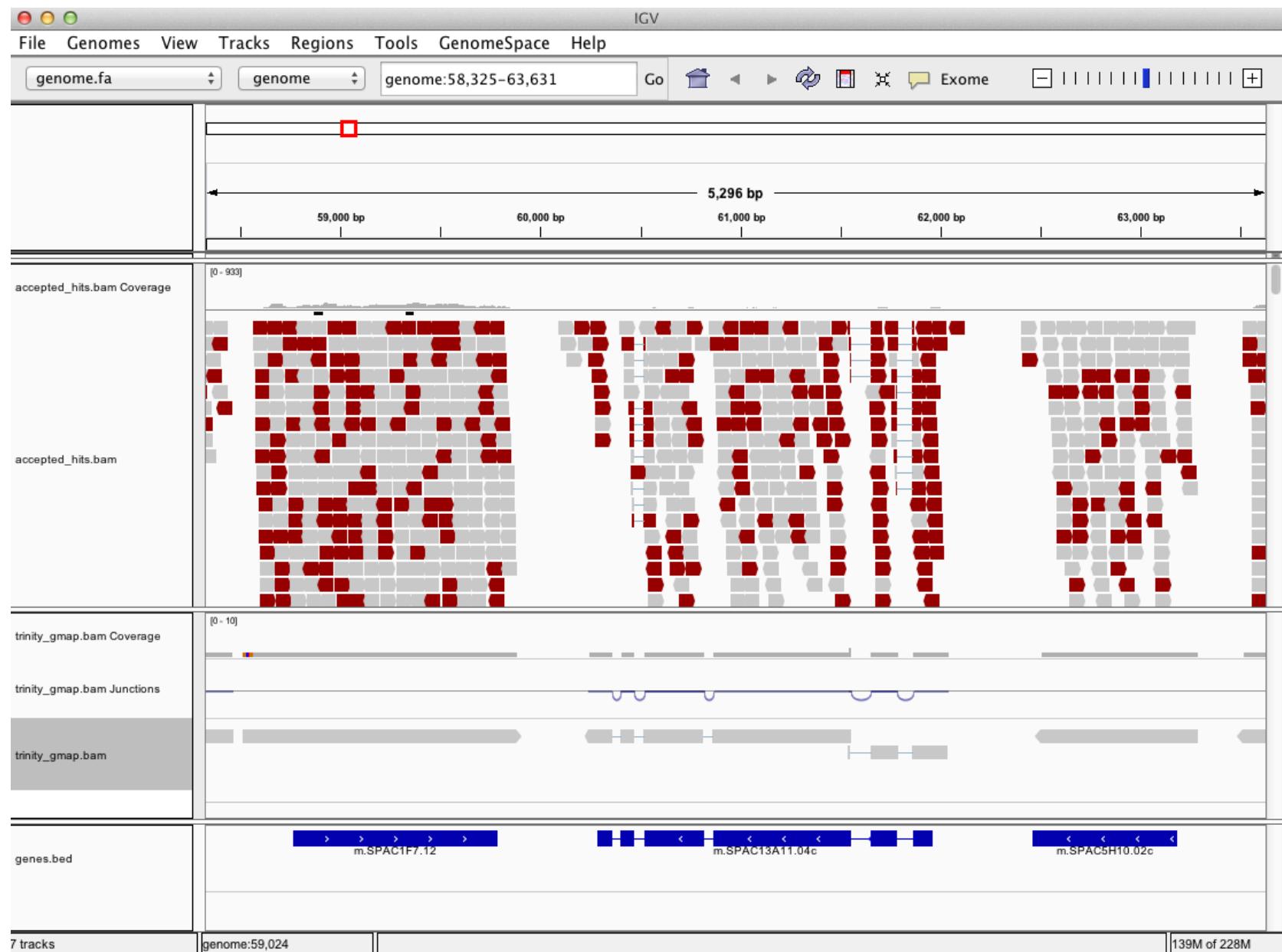
chr7:52,150,889-52,189,508

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



Trinity output: A multi-fasta file

Can align Trinity transcripts to genome scaffolds to examine intron/exon structures (Trinity transcripts aligned using GMAP)



Trinity Demo

https://github.com/trinityrnaseq/RNASeq_Trinity_Tuxedo_Workshop/wiki/Trinity-De-novo-Transcriptome-Assembly-Workshop

- Assemble RNA-Seq using Trinity
- Examine Trinity in context of a genome:
 - Align Trinity transcripts to the genome using GMAP
 - Align rna-seq reads to genome using Tophat
 - Visualize all alignments using IGV

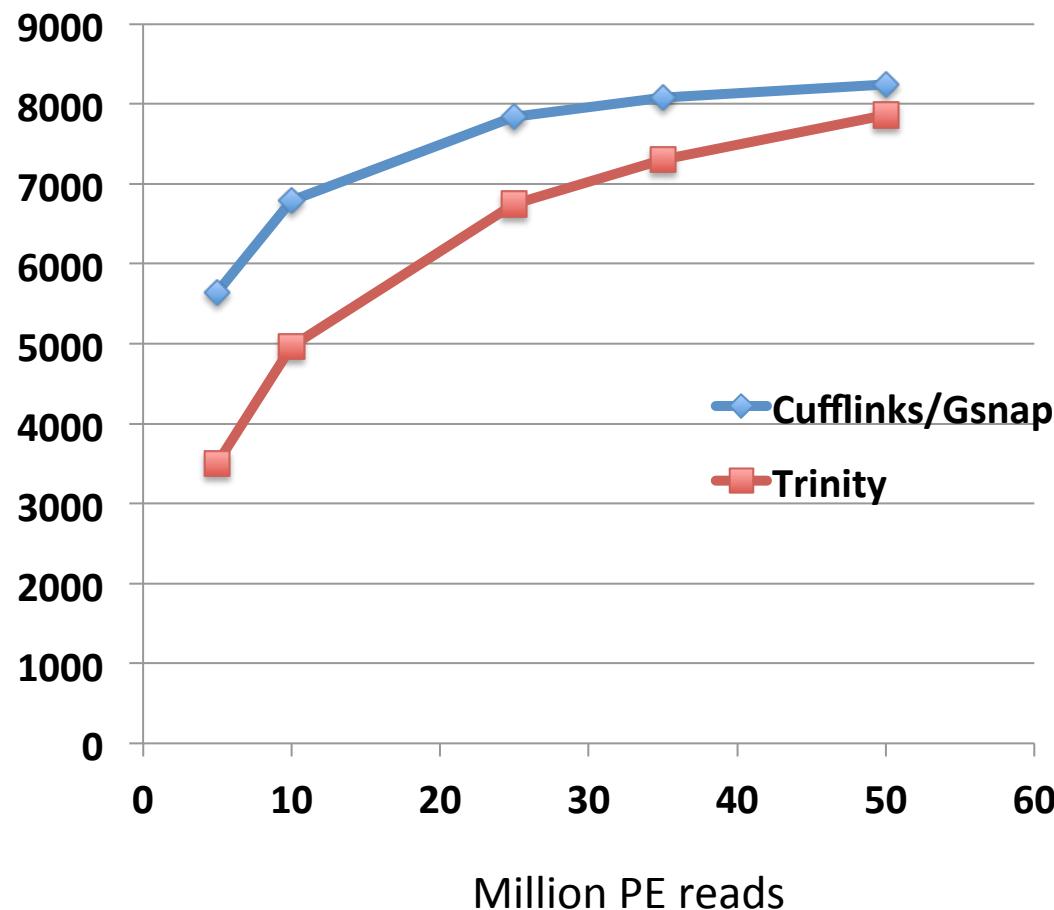
Improved reconstruction with deeper sequencing depth and

Genome-based reconstruction is more sensitive than de novo methods

Genes w/ fully
reconstructed
transcripts



Mouse data



Strand-specific RNA-Seq is Preferred

Computationally: fewer confounding graph structures in de novo assembly:
ex. Forward != reverse complement
(GGAA != TTCC)

Biologically: separate sense vs. antisense transcription

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



Comprehensive comparative analysis of strand-specific RNA sequencing methods

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

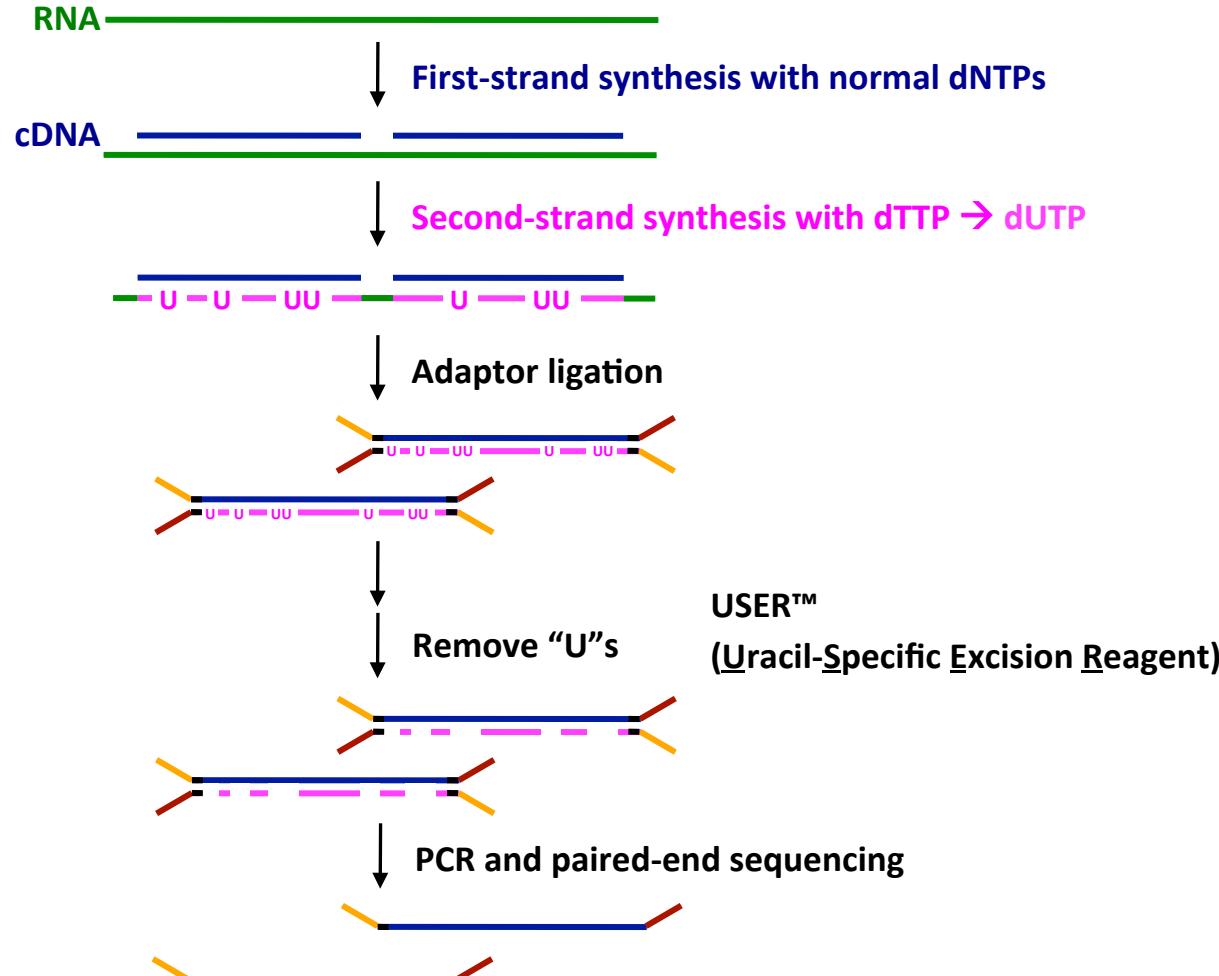
Strand-specific, massively parallel cDNA sequencing (RNA-seq) is a powerful tool for transcript discovery, genome annotation

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

'dUTP second strand marking' identified as the leading protocol

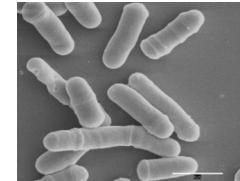
to choose between them. Here we developed a comprehensive computational pipeline to compare library quality metrics from any RNA-seq method. Using the well-annotated *Saccharomyces cerevisiae* transcriptome as a benchmark, we compared seven library-construction protocols, including both published and transcribed strand or other noncoding RNAs, demarcate the exact boundaries of adjacent genes transcribed on opposite strands and resolve the correct expression levels of coding or noncoding overlapping transcripts. These tasks are particularly challenging in small microbial genomes, prokaryotic and eukaryotic, in which

dUTP 2nd Strand Method: Our Favorite

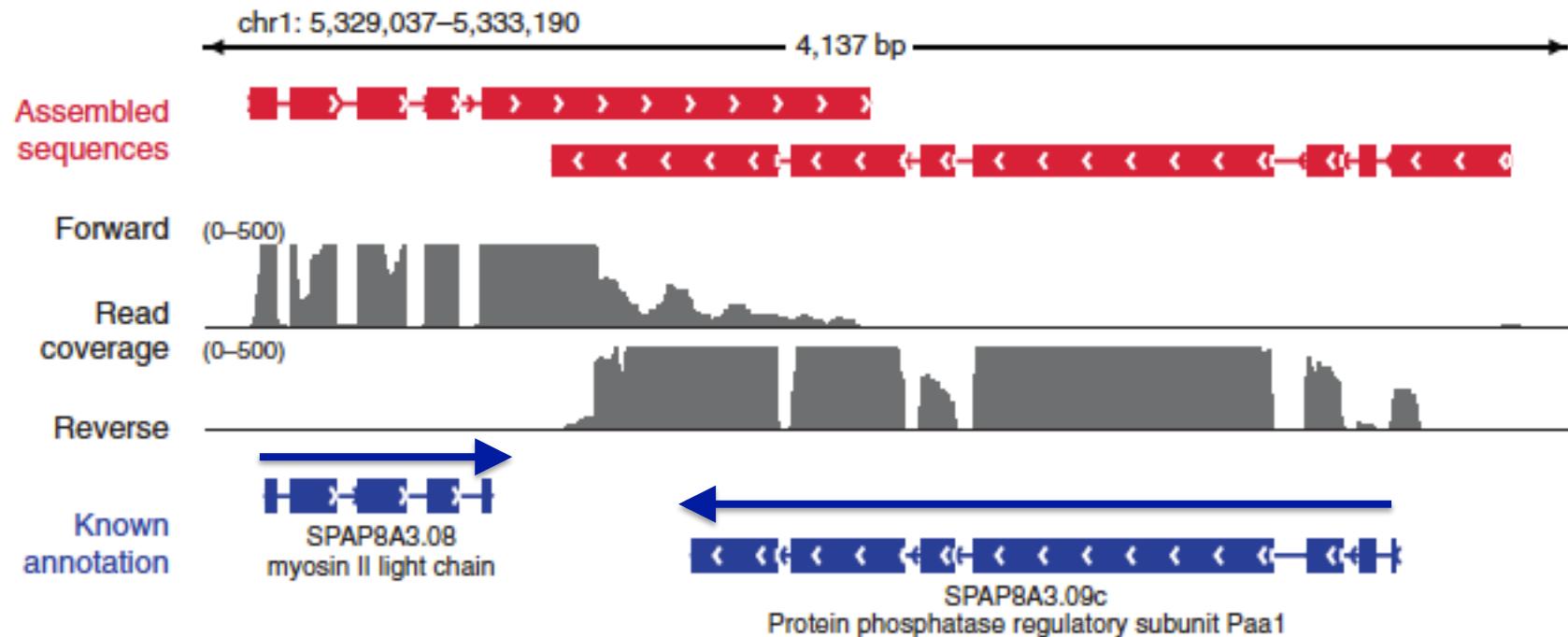


Modified from Parkhomchuk *et al.* (2009) *Nucleic Acids Res.* 37:e123

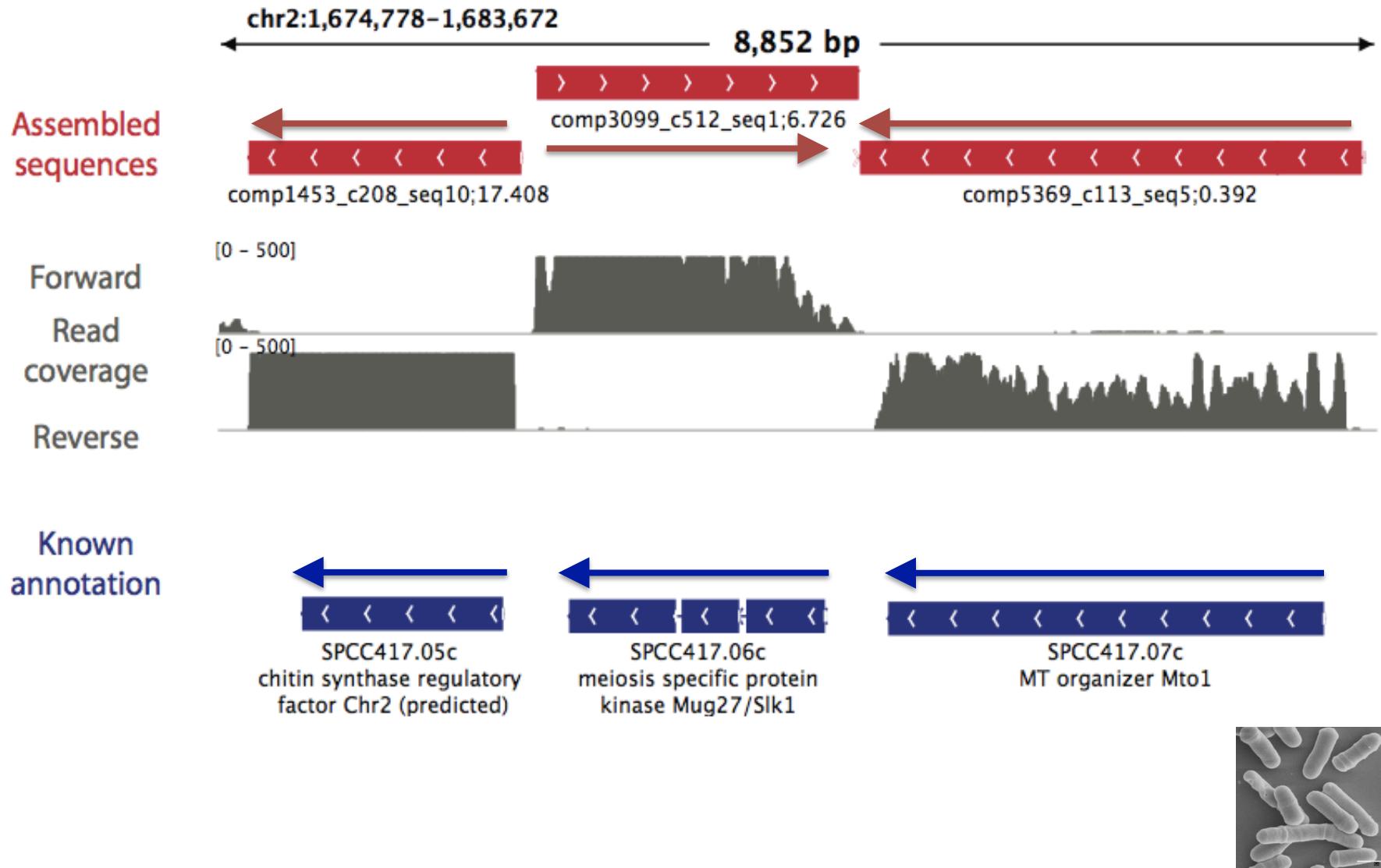
Overlapping UTRs from Opposite Strands



Schizosaccharomyces pombe
(fission yeast)



Antisense-dominated Transcription



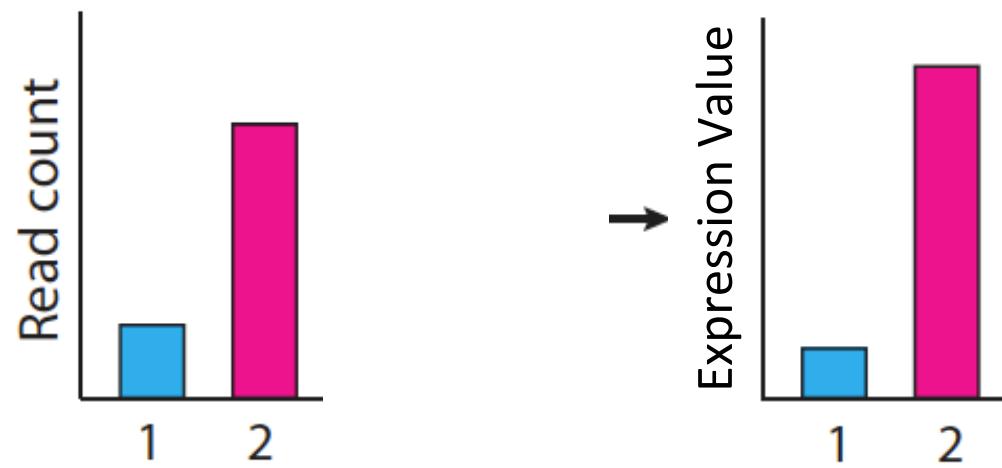
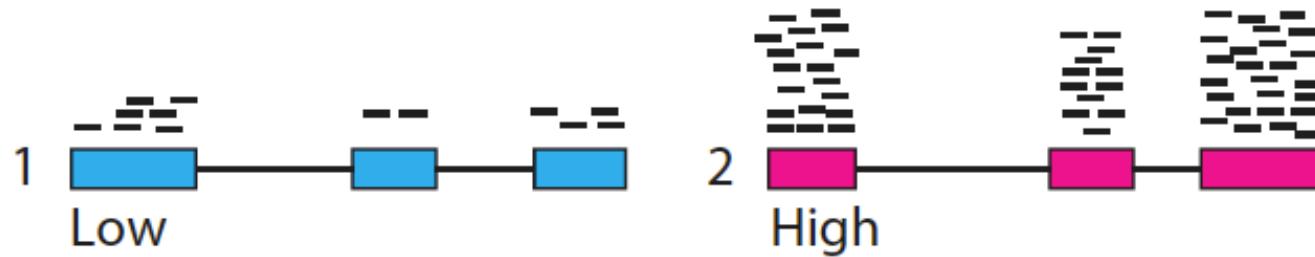
Summary

- Two paradigms for transcript reconstruction
 - Rna-seq alignment assembly
 - Tuxedo (tophat, cufflinks)
 - genome-free de novo read assembly
 - Trinity
- Often best to pursue both strategies
 - Maximize sensitivity for genome-based transcript reconstruction + capture missing or ill-represented transcripts via de novo assembly.

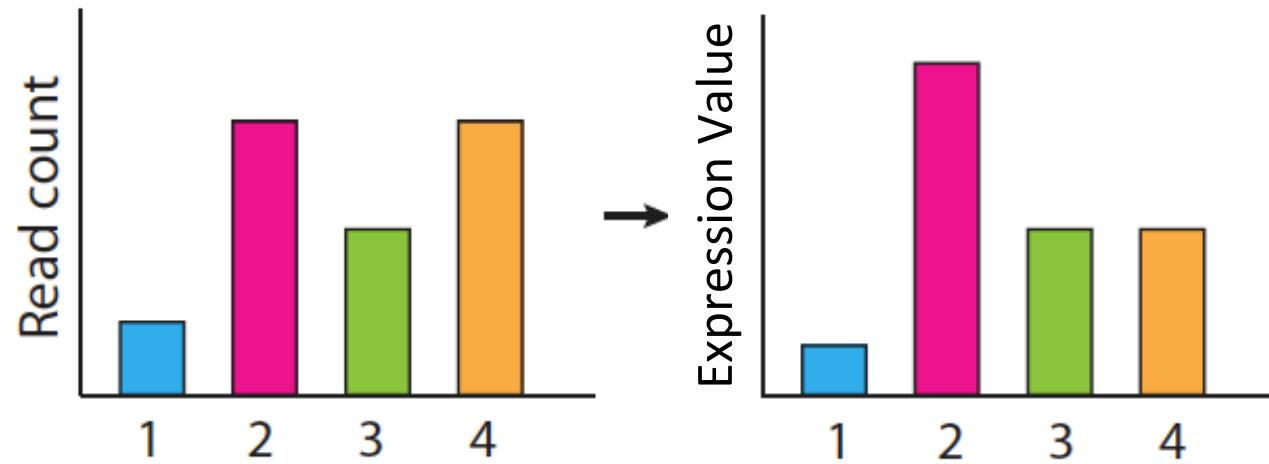
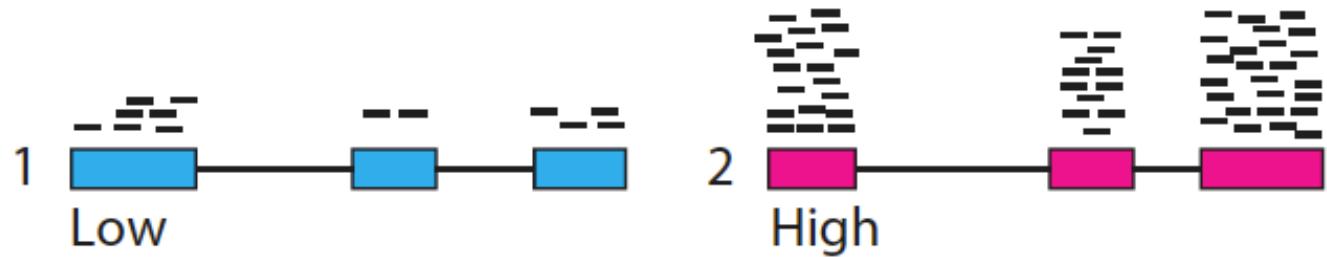
Abundance Estimation

(Aka. Computing Expression Values)

Calculating expression of genes and transcripts



Calculating expression of genes and transcripts



Normalized Expression Values

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

FPKM

Transcripts per Million (TPM)

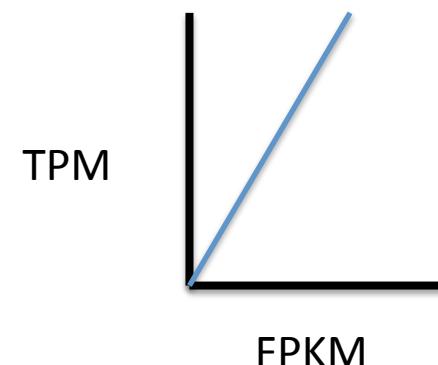
$$TPM_i = \frac{FPKM_i}{\sum_j FPKM} * 1e6$$

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

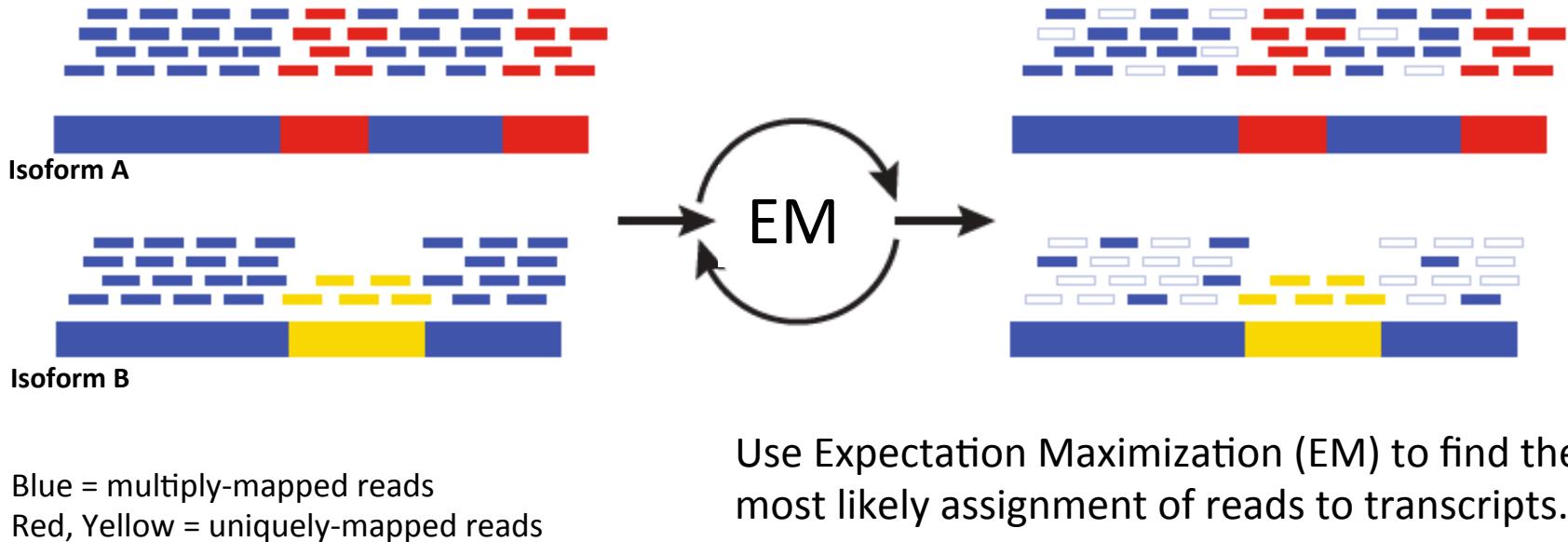


Isoform B

Blue = multiply-mapped reads

Red, Yellow = uniquely-mapped reads

Multiply-mapped Reads Confound Abundance Estimation



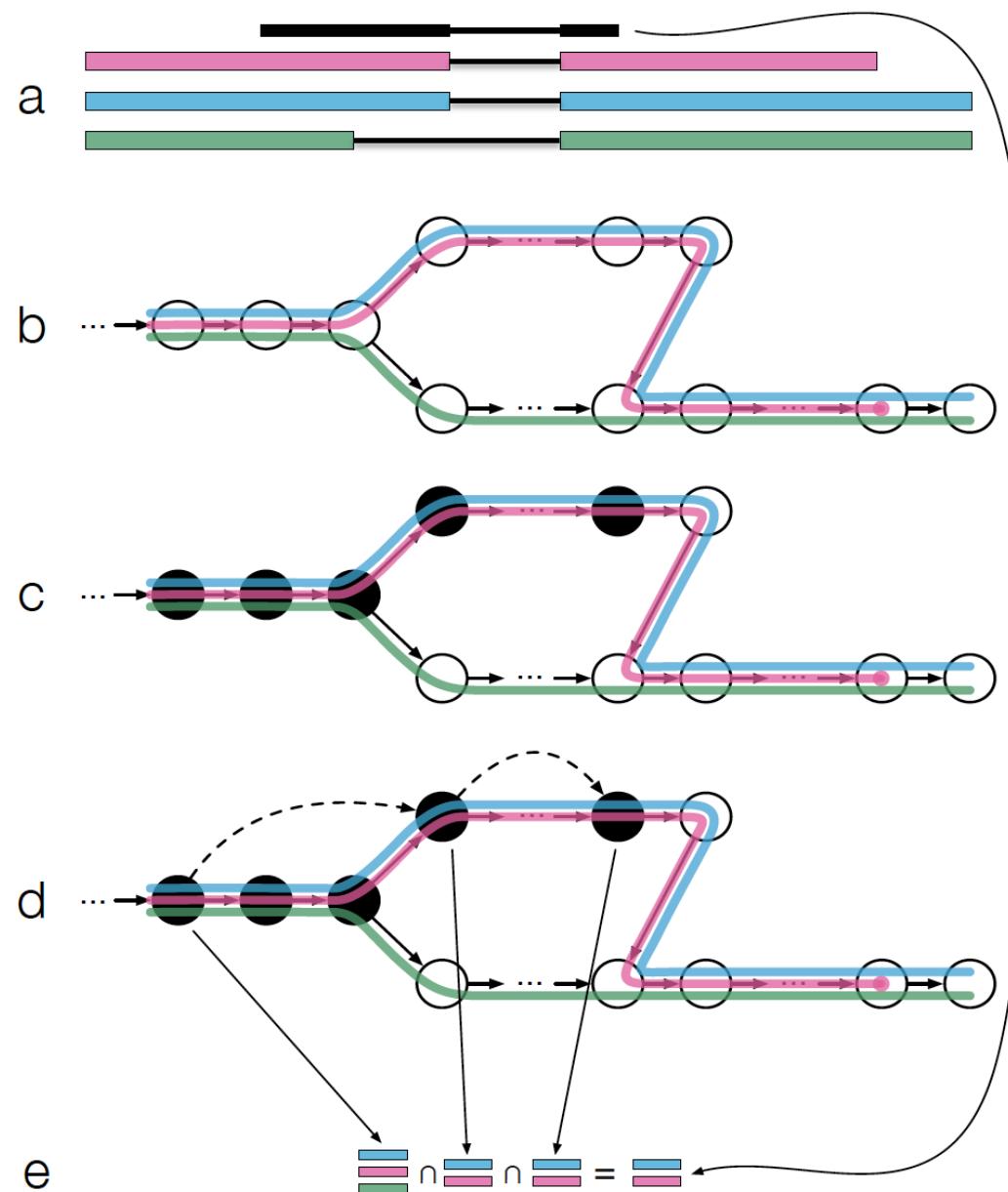
Use Expectation Maximization (EM) to find the most likely assignment of reads to transcripts.

Performed by:

RSEM, eXpress, kallisto, salmon, ...

New fast alignment-free methods
now available! eg. Kallisto

Fast Abundance Estimation Using Pseudo-alignments and Equivalence Classes (Kallisto software, Bray et al., NBT 2016)

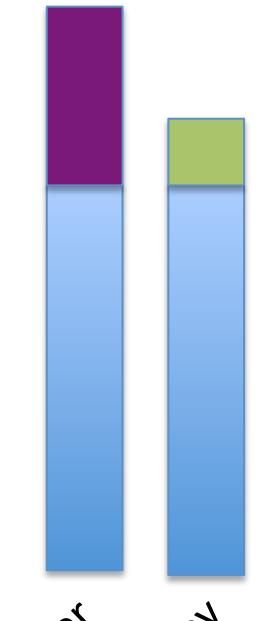


Adapted from Fig 1 from Bray et al.

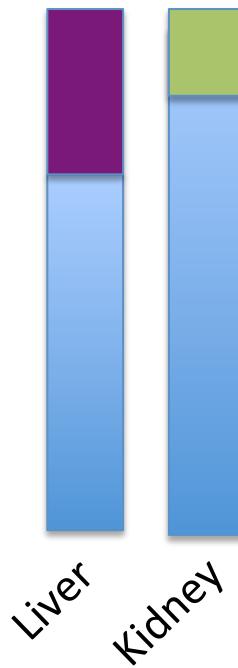
Differential Expression Analysis Using RNA-Seq

Why cross-sample normalization is important

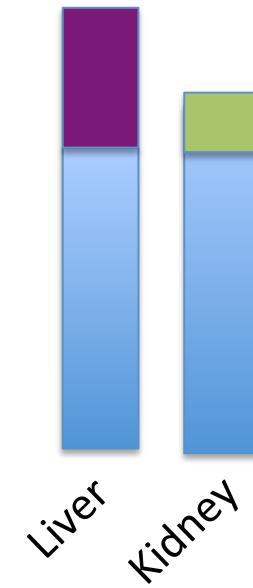
Absolute RNA quantities per cell



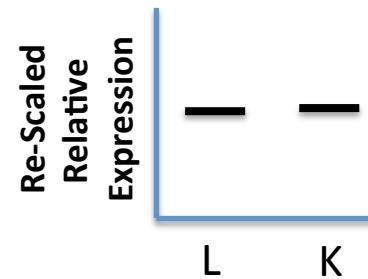
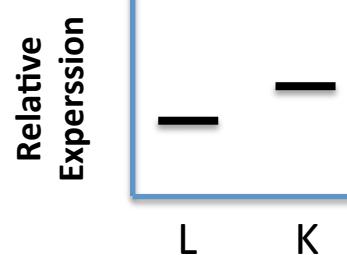
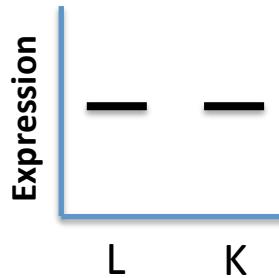
Measured relative abundance via RNA-Seq



Cross-sample normalized (rescaled) relative abundance



eg. Some housekeeping gene's expression level:



Normalization Required

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

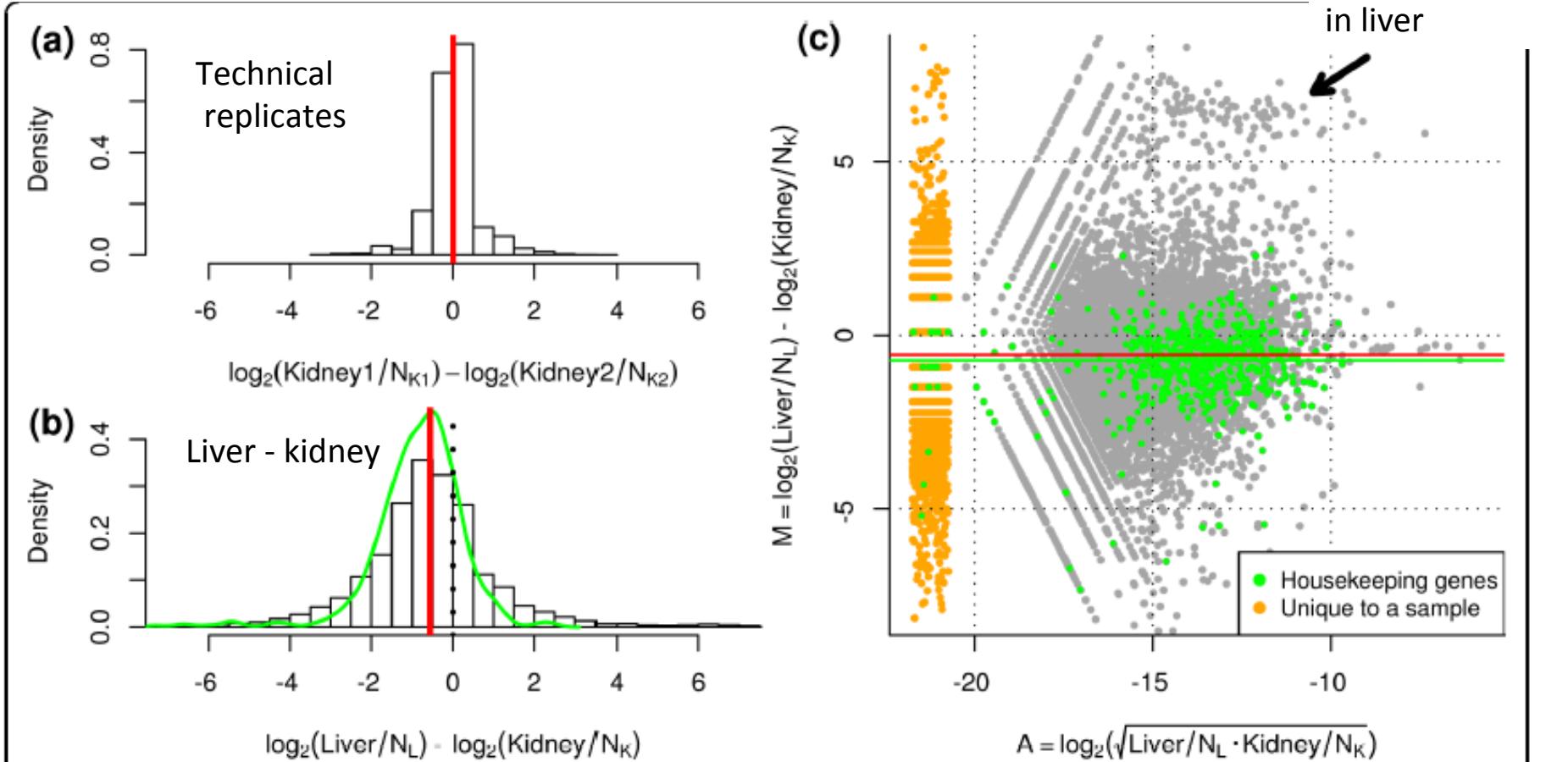
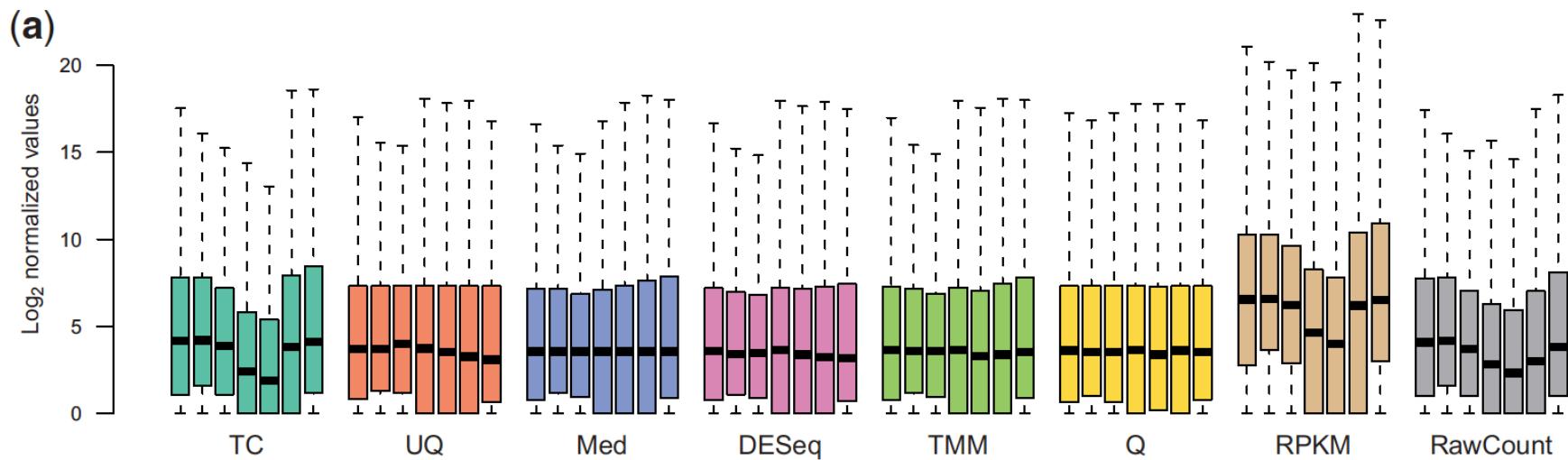


Figure 1 Normalization is required for RNA-seq data. Data from [6] comparing log ratios of (a) technical replicates and (b) liver versus kidney expression levels, after adjusting for the total number of reads in each sample. The green line shows the smoothed distribution of log-fold-changes of the housekeeping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

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



From “A comprehensive evaluation of normalization methods for Illumina high throughput RNA sequencing data analysis” Brief Bioinform. 2013 Nov;14(6):671-83
<http://www.ncbi.nlm.nih.gov/pubmed/22988256>

Differential Expression Analysis



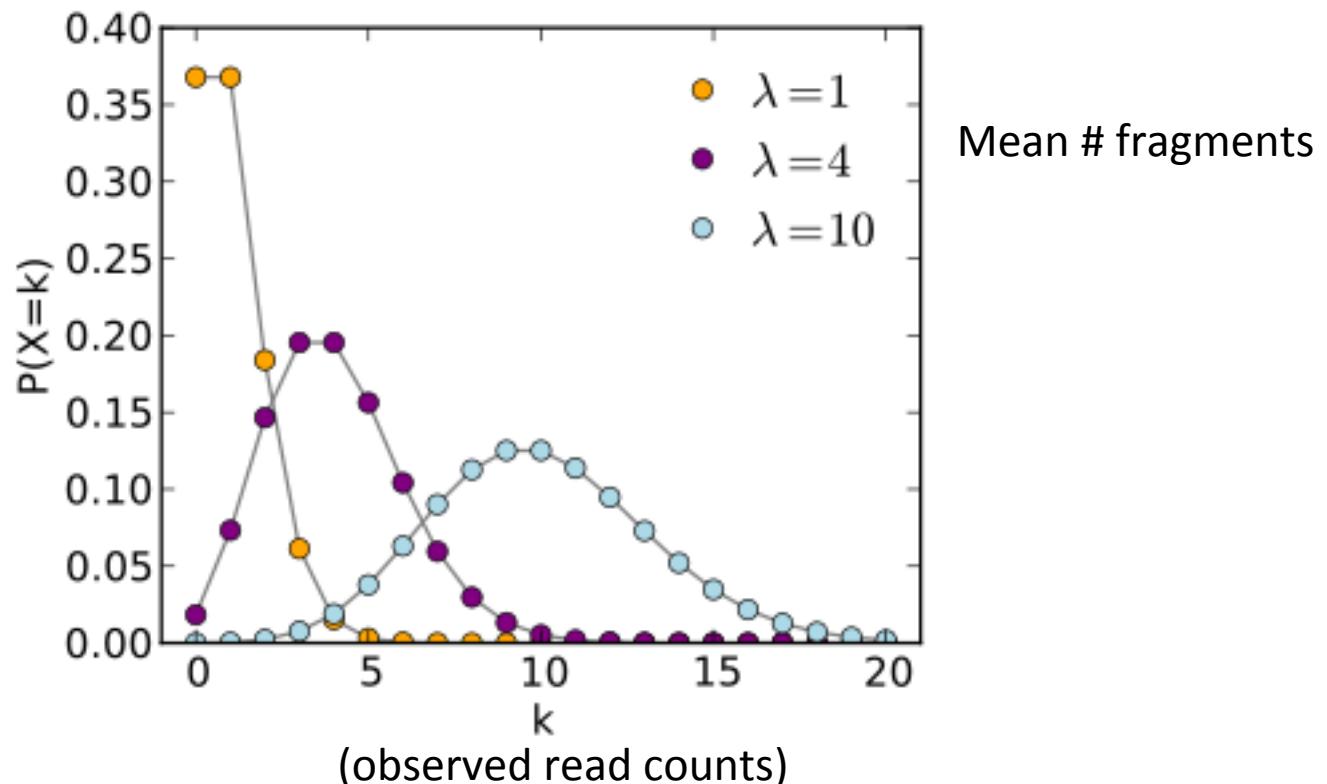
Diff. Expression Analysis Involves

- Counting reads
- Statistical significance testing

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

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

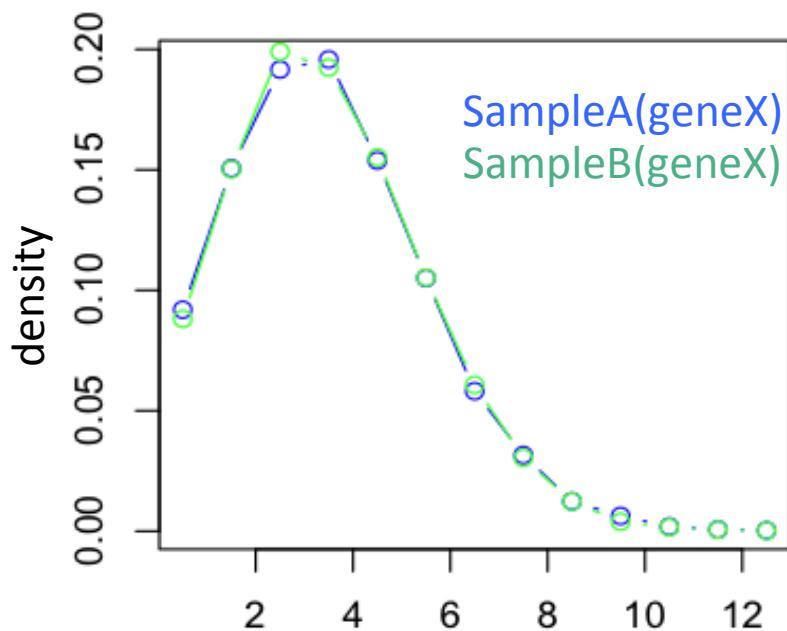
Technical variation in RNA-Seq counts per feature is well modeled by the Poisson distribution



Example: One gene*not* differentially expressed

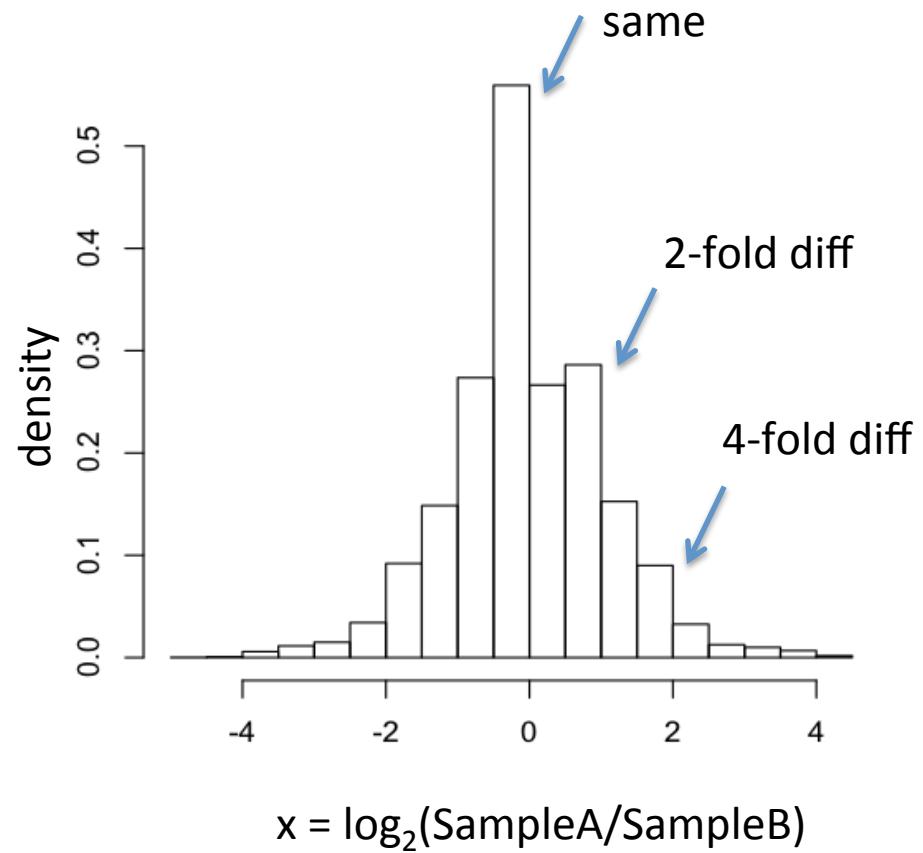
SampleA(gene) = SampleB(gene) = 4 reads

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



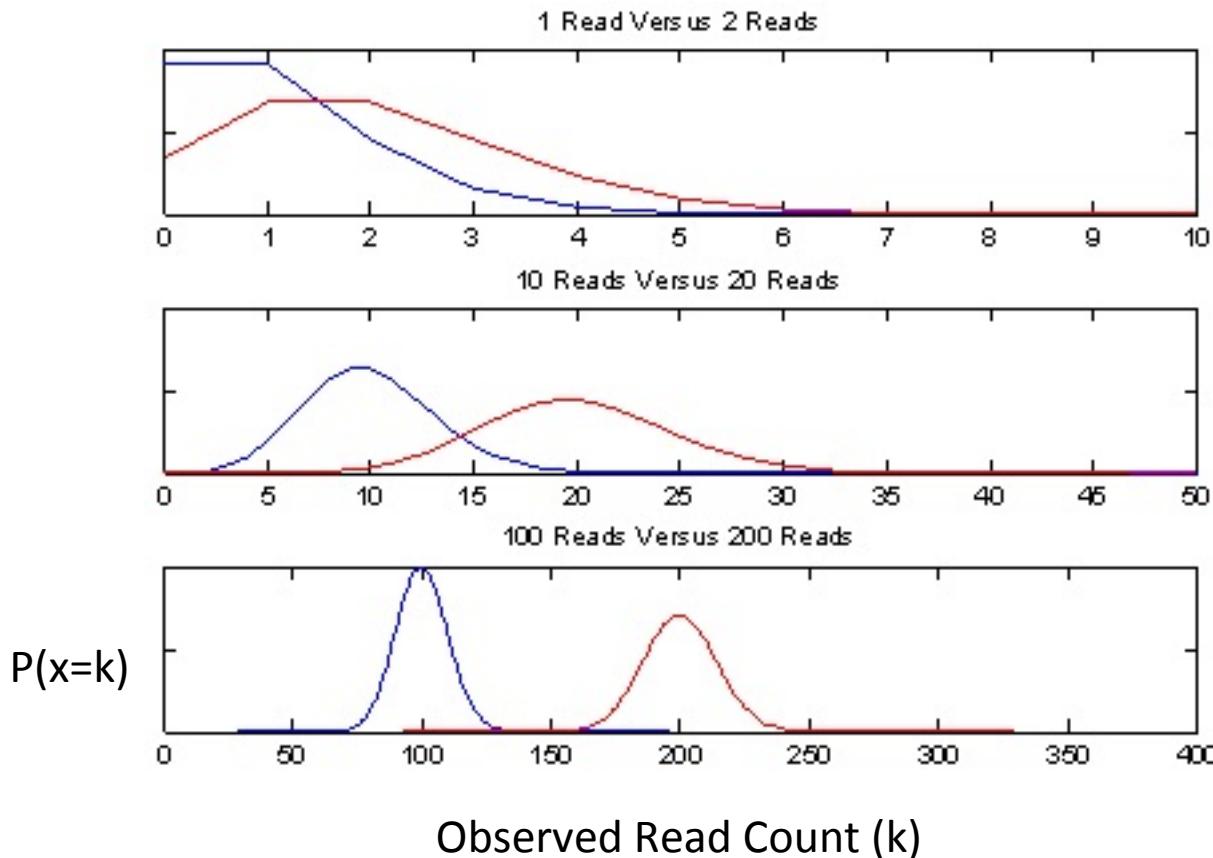
(k) number of reads observed

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



Beware of concluding fold change from small numbers of counts

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.

More Counts = More Statistical Power

Example: 5000 total reads per sample.

Observed 2-fold differences in read counts.

	SampleA	Sample B	Fisher's Exact Test (P-value)
geneA	1	2	1.00
geneB	10	20	0.098
geneC	100	200	< 0.001

Tools for DE analysis with RNA-Seq



edgeR	ROTS
ShrinkSeq	TSPM
DESeq	DESeq2
baySeq	EBSeq
Vsf	NBPSeq
Limma/Voom	SAMseq
<i>mmdiff</i>	NoiSeq
<i>cuffdiff</i>	

*(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

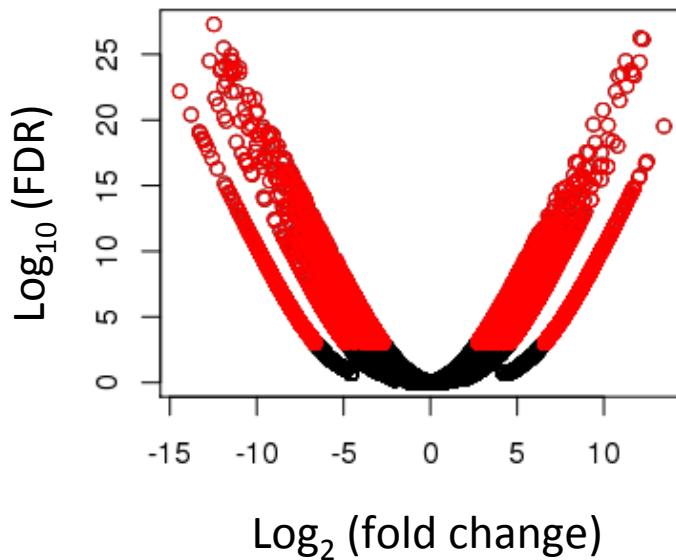


Significance

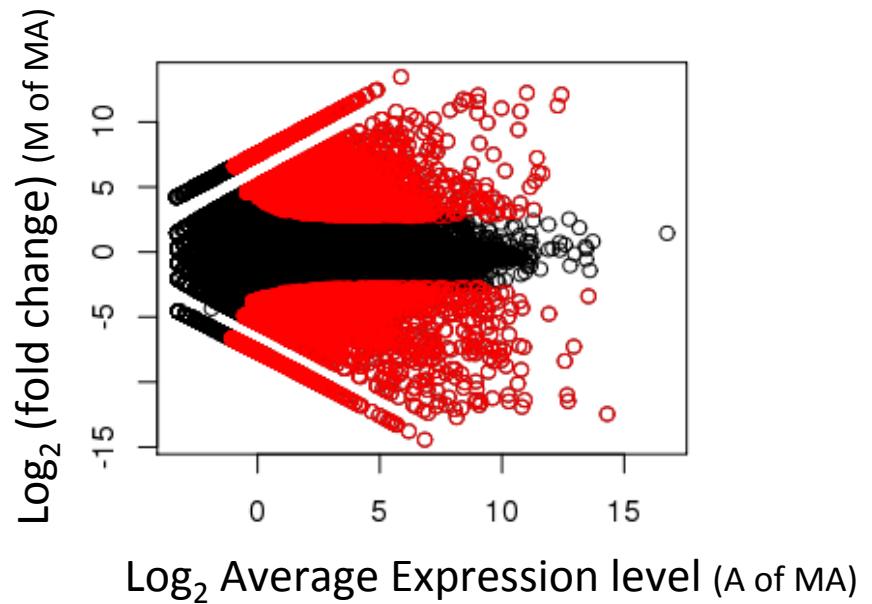
Visualization of DE results and Expression Profiling

Plotting Pairwise Differential Expression Data

Volcano plot
(fold change vs. significance)

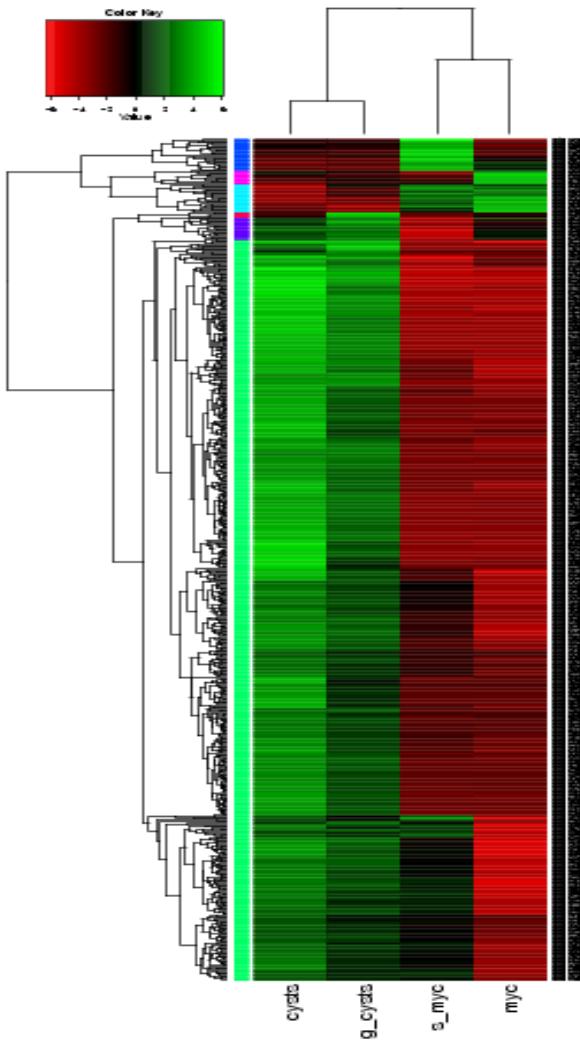


MA plot
(abundance vs. fold change)



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

Comparing Multiple Samples



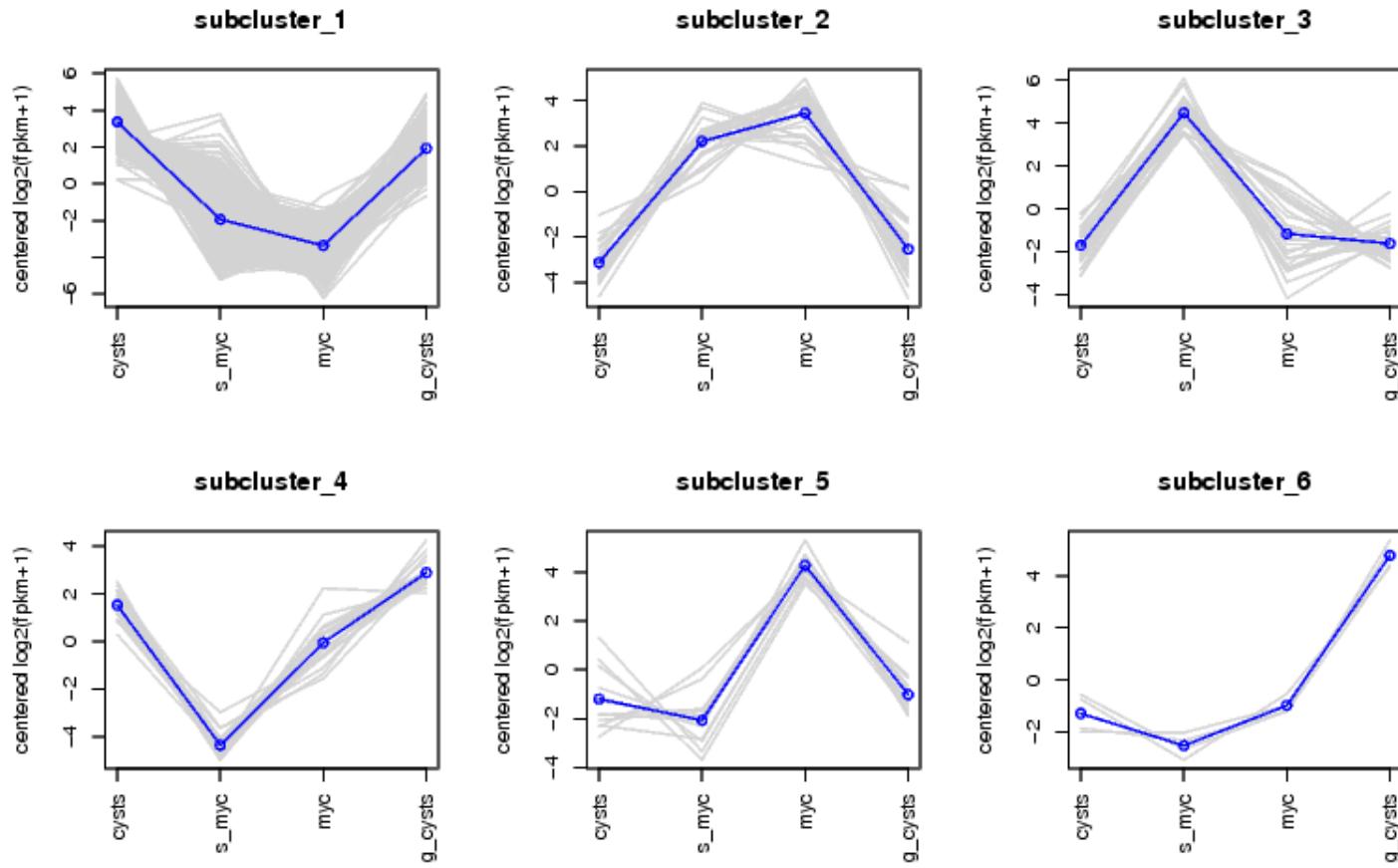
Heatmaps provide an effective tool for navigating differential expression across multiple samples.

Clustering can be performed across both axes:

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

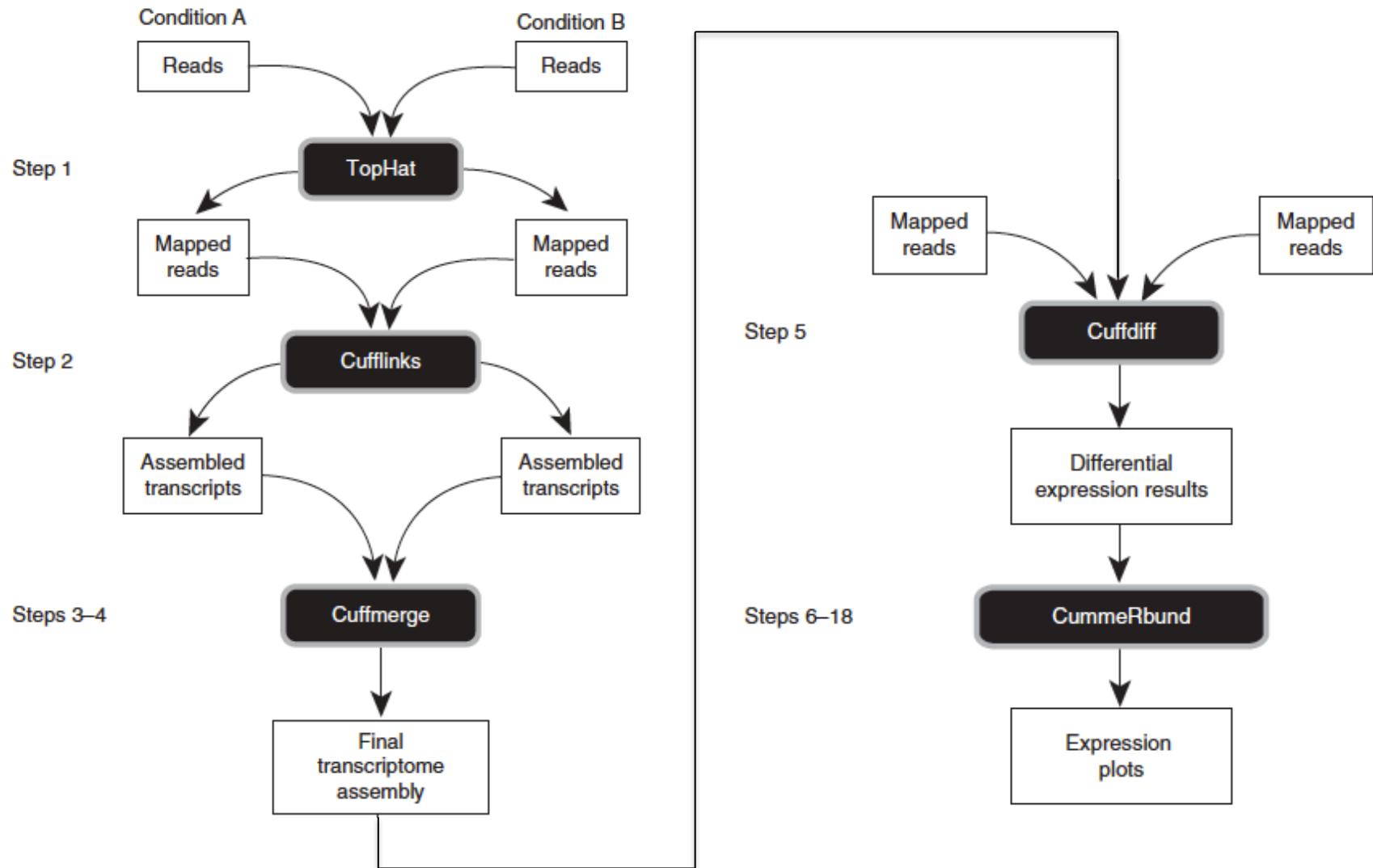
Examining Patterns of Expression Across Samples

Can extract clusters of transcripts and examine them separately.



RNA-Seq Analysis Frameworks

Tuxedo Framework for Transcriptome Analysis



Derived from: Nat Protoc. 2012 Mar 1;7(3):562-78. doi: 10.1038/nprot.2012.016.

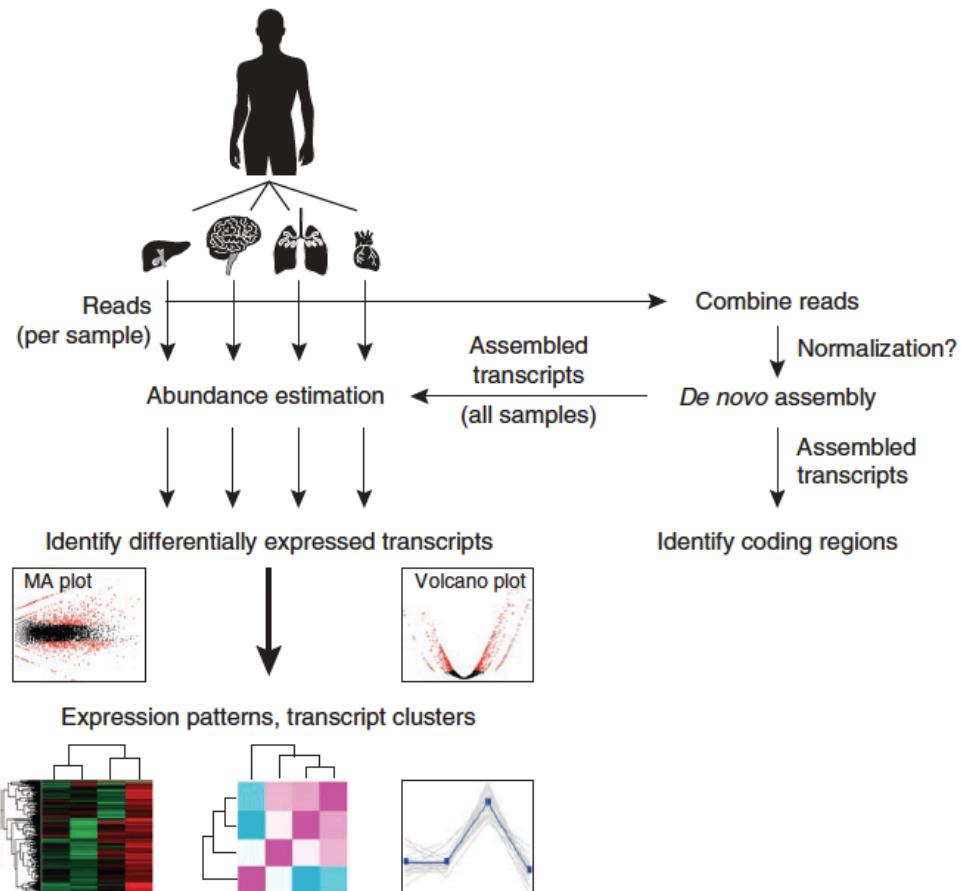
De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis

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



Summary of Key Points

- RNA-Seq is a versatile method for transcriptome analysis enabling quantification and novel transcript discovery.
- Genome-based and genome-free methods exist for transcript reconstruction
- Expression quantification is based on sampling and counting reads derived from transcripts
- Fold changes based on few read counts lack statistical significance.
- Multiple analysis frameworks are available – alternative and often complementary approaches to support biological investigations.

Papers of Interest

- Next generation transcriptome assembly
 - <http://www.nature.com/nrg/journal/v12/n10/full/nrg3068.html>
- Tuxedo protocol
 - <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3334321/>
- Trinity
 - <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3571712/>
 - <http://www.nature.com/nprot/journal/v8/n8/full/nprot.2013.084.html>