



Advanced approaches
for gathering, interpreting
& applying next generation
transcriptome data

18th – 20th June 2013 | Boston, MA

Workshop B: Genome-Based and Genome-Free Transcript Reconstruction and Analysis Using RNA-Seq Data



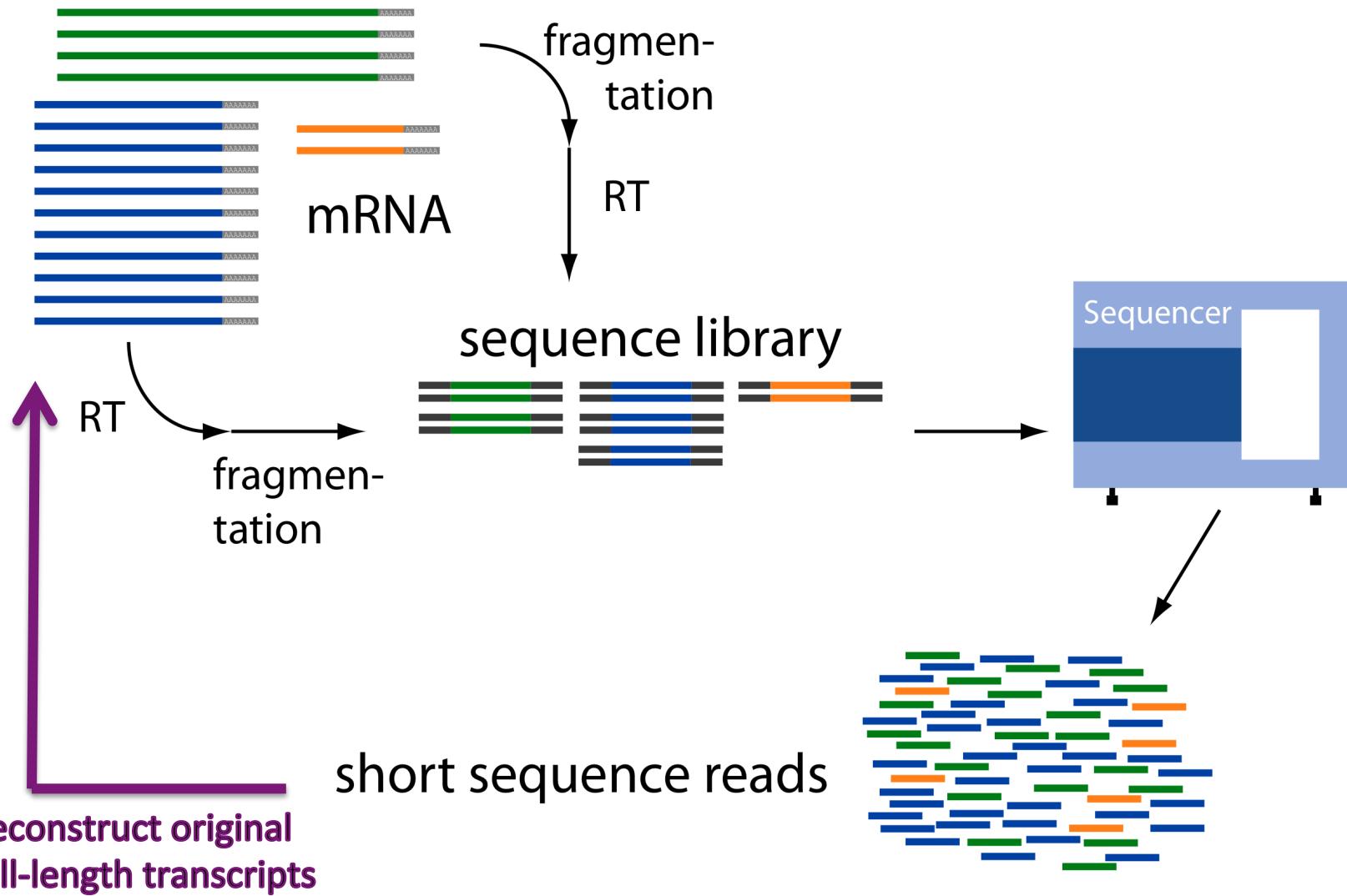
Brian Haas
Broad Institute



Workshop Overview

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

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  
AAACAAACAGGGCACATTGTCACTCTTGTATTGAAAAAACACTTCCGGCCAT
```

FASTQ format:

```
@61DFRAAXX100204:1:100:10494:3070/1  
AAACAAACAGGGCACATTGTCACTCTTGTATTGAAAAAACACTTCCGGCCAT  
+  
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

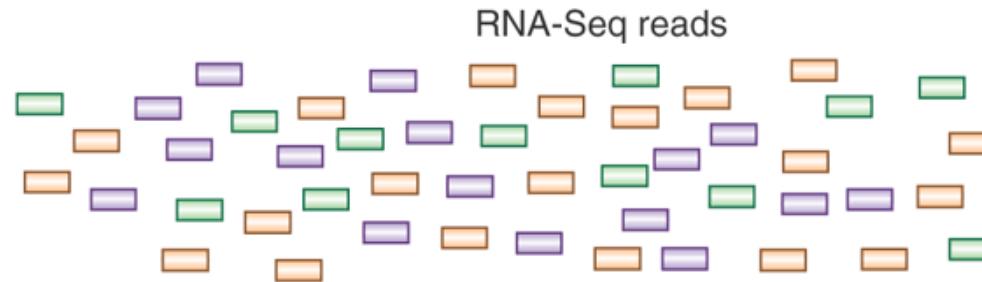


Two FastQ files, read name indicates
left (/1) or right (/2) read of paired-end

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

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

Transcript Reconstruction from RNA-Seq Reads



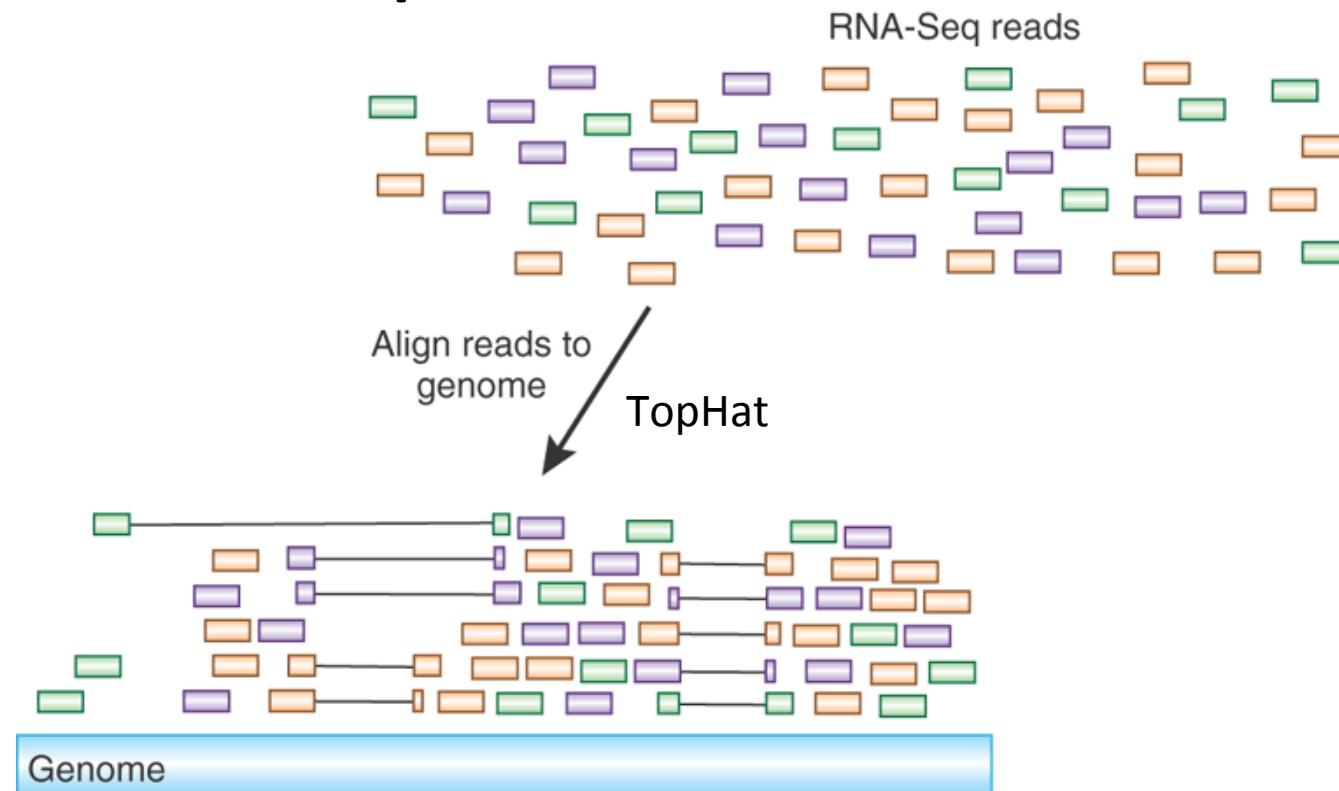
Advancing RNA-Seq analysis

Brian J Haas & Michael C Zody

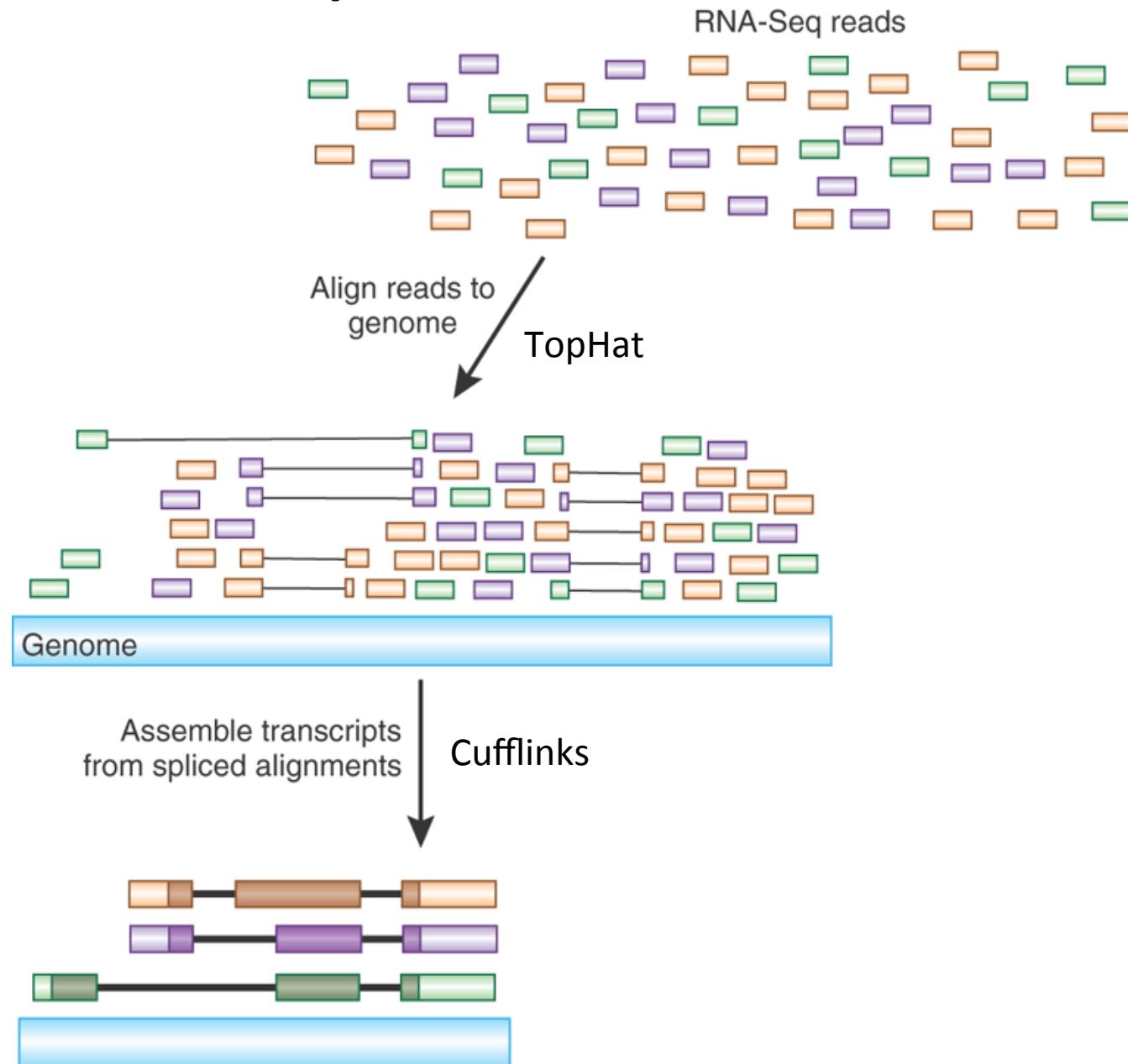
Nature Biotech, 2010

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

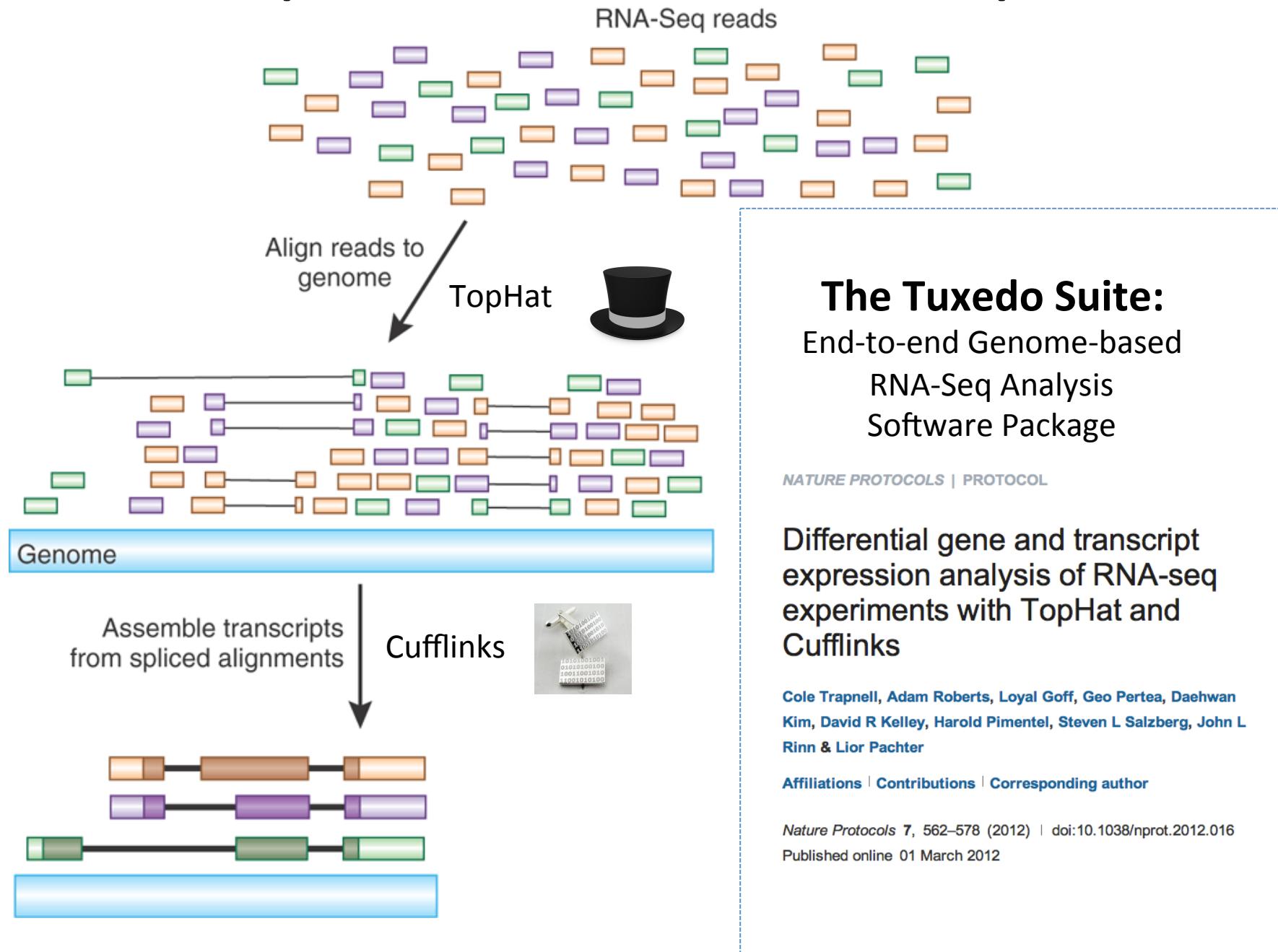
Transcript Reconstruction from RNA-Seq Reads



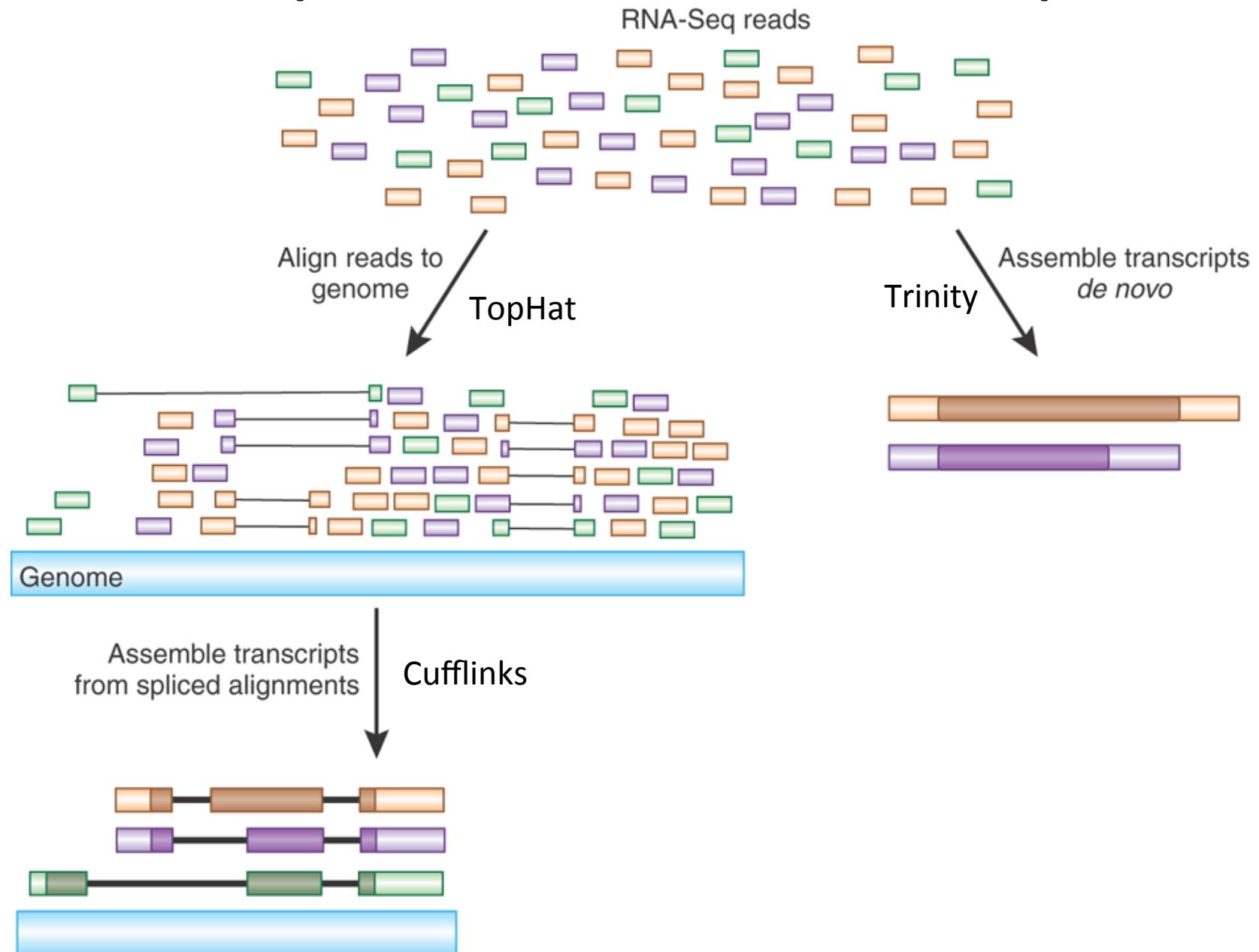
Transcript Reconstruction from RNA-Seq Reads



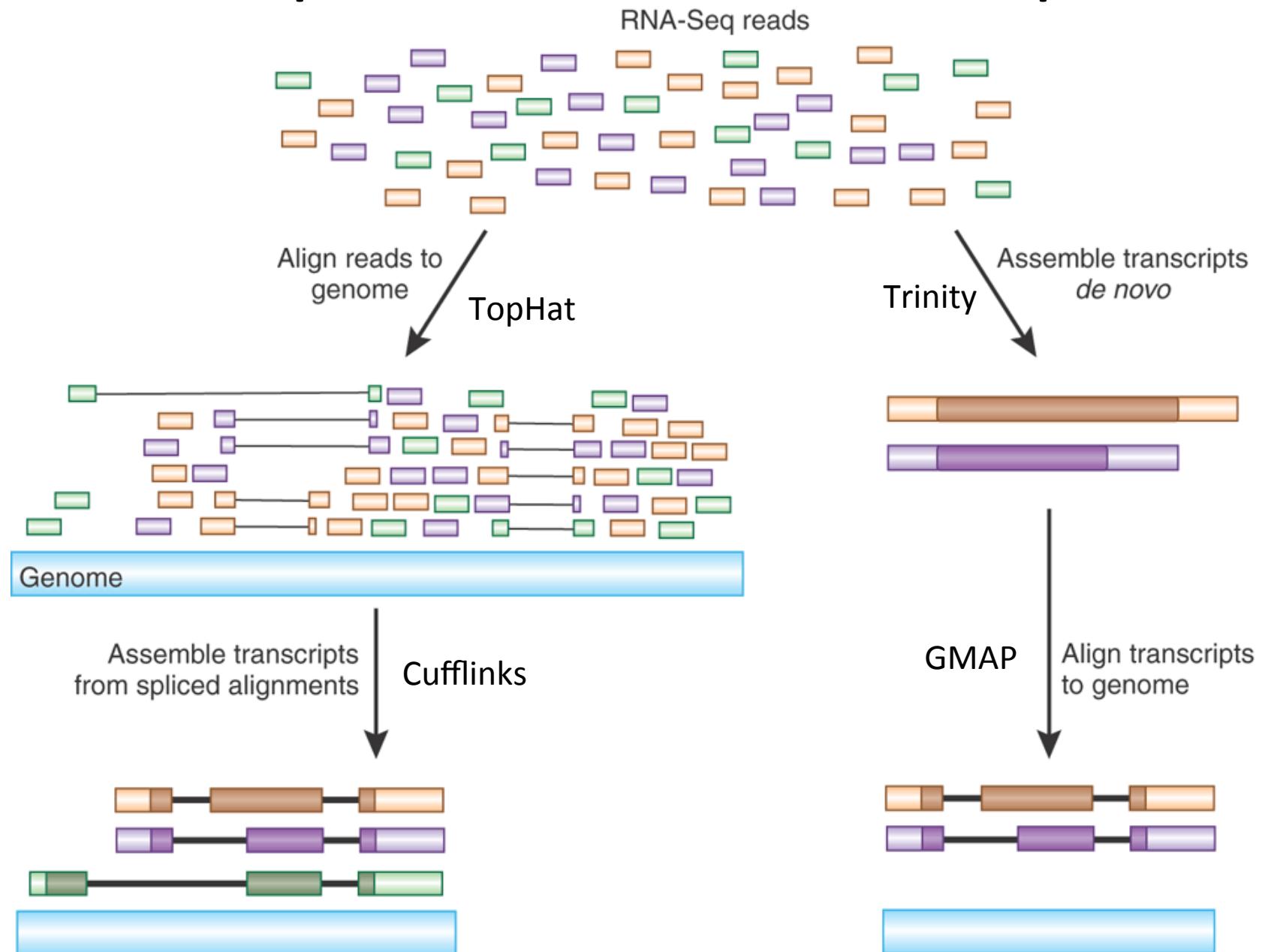
Transcript Reconstruction from RNA-Seq Reads



Transcript Reconstruction from RNA-Seq Reads



Transcript Reconstruction from RNA-Seq Reads



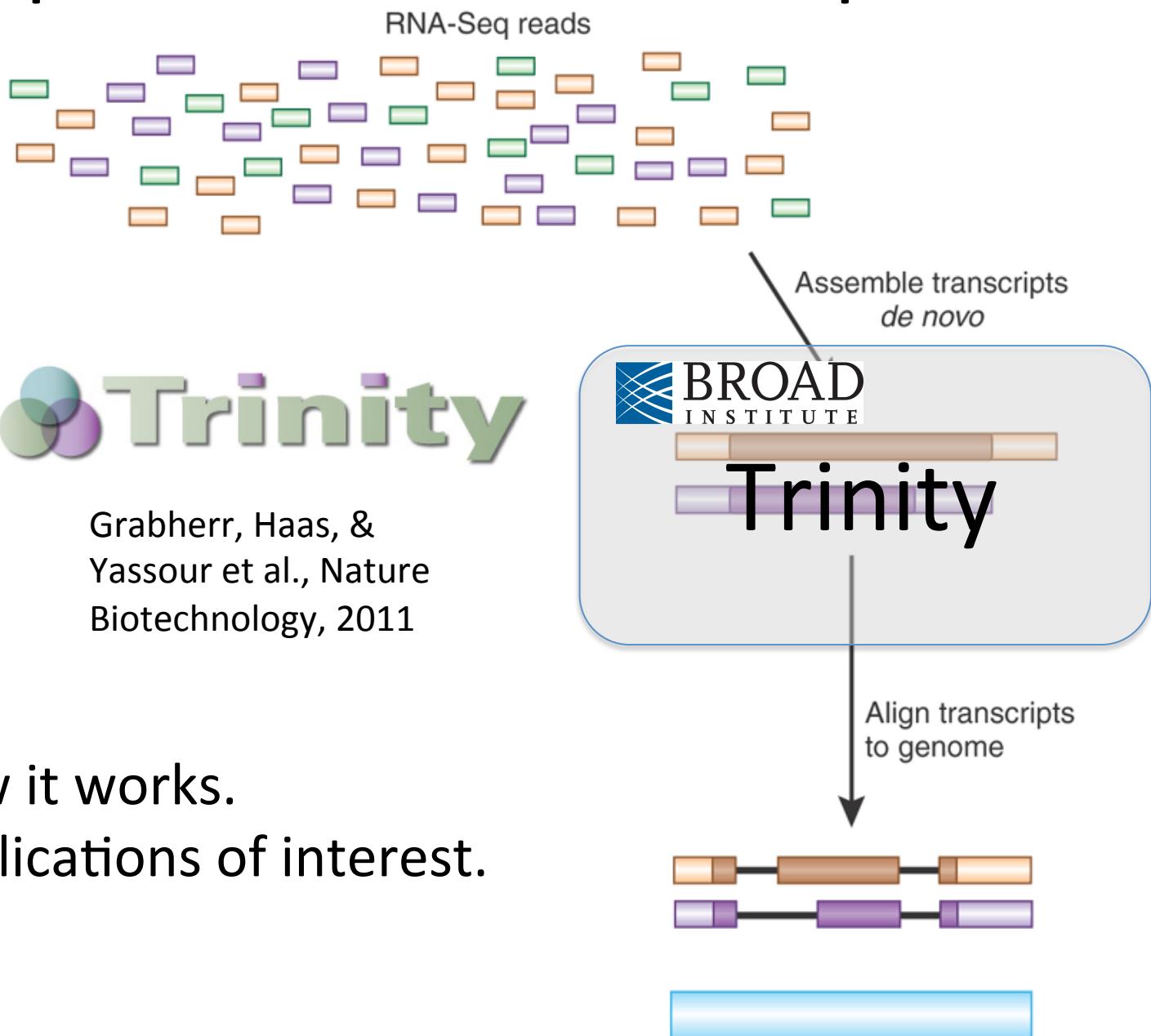
Transcript Reconstruction from RNA-Seq Reads



 **Trinity**

Grabherr, Haas, &
Yassour et al., Nature
Biotechnology, 2011

- How it works.
- Applications of interest.

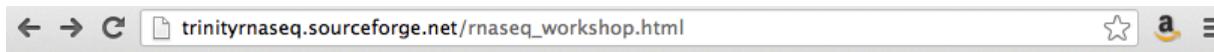


Overview of Companion Materials

- VirtualBox Ubuntu Linux Image
 - Run Linux as a virtual machine in Mac or Windows
 - Trinity and Tuxedo software + more.
 - Sample RNA-Seq data
- Tutorials / Software Demos

Download the RNA-Seq Workshop Virtual Machine (VM) Image and Related Materials

http://trinityrnaseq.sourceforge.net/rnaseq_workshop.html



RNA-Seq Workshop

Do you want to learn the basics of how to analyze RNA-Seq data, including genome-based and genome-free transcript reconstruction and differential expression analysis? Below, we provide materials to help get you started. A VirtualBox virtual machine image is provided that includes several popular tools for analyzing RNA-Seq, including Tuxedo for genome-based and Trinity for genome-free RNA-Seq studies, and small sample (simulated) RNA-Seq data sets. Hands-on tutorials are provided to guide users in a step-by-step analysis of these data. Note, the only purpose of these simulated data is to give the user hands-on experience in using the various analysis tools and be sufficiently small in size such that the data can be processed on an average desktop computer in a small amount of time (minutes rather than hours). Future work may have an increased focus on learning interesting biology from real data sets, in addition to familiarizing users with use of these applications.

To get started, you must first install [VirtualBox](#) and import the [Ubuntu Linux image provided](#) that's preinstalled with the required data and software. A [basic installation guide](#) with screenshots is provided. VirtualBox should allow a user to experiment with this Ubuntu Linux from within their own operating system, including from Mac or Windows operating systems.

The accompanying Trinity and Tuxedo walk-throughs are available from the Trinity ftp site at:
http://sourceforge.net/projects/trinityrnaseq/files/misc/RNASEQ_WORKSHOP/.

In summary, these are the requirements:

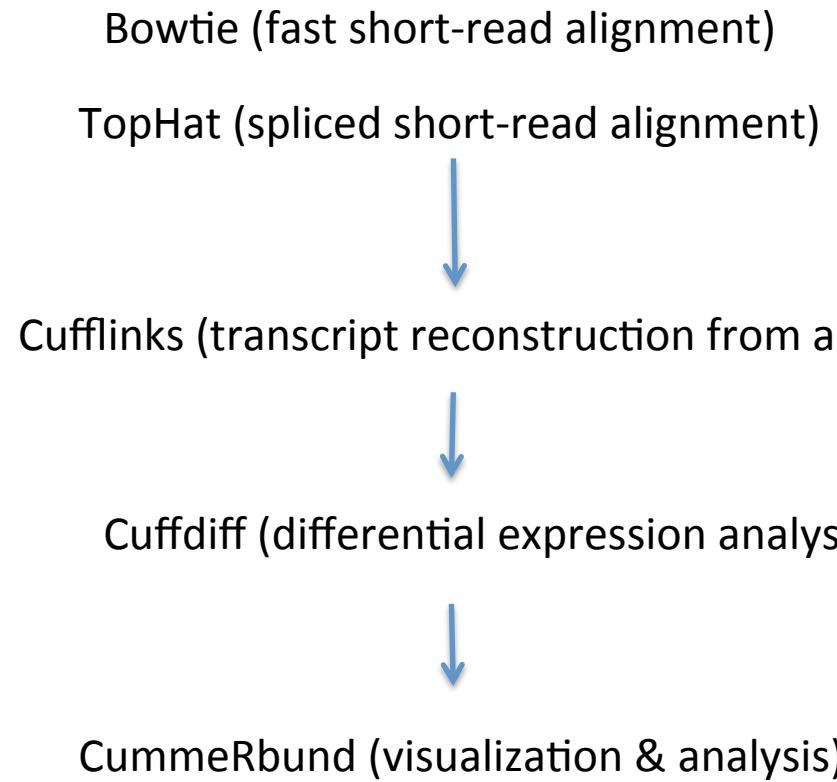
- Install Virtualbox from <http://virtualbox.org>
- Download the [Ubuntu Linux image with pre-installed software and data](#). ([Ubuntu_RNASeqWorkshop.ova](#))
- Install the Ubuntu image following [these instructions](#).
- Run through the Trinity and/or the Tuxedo [tutorials](#).

Download
Materials

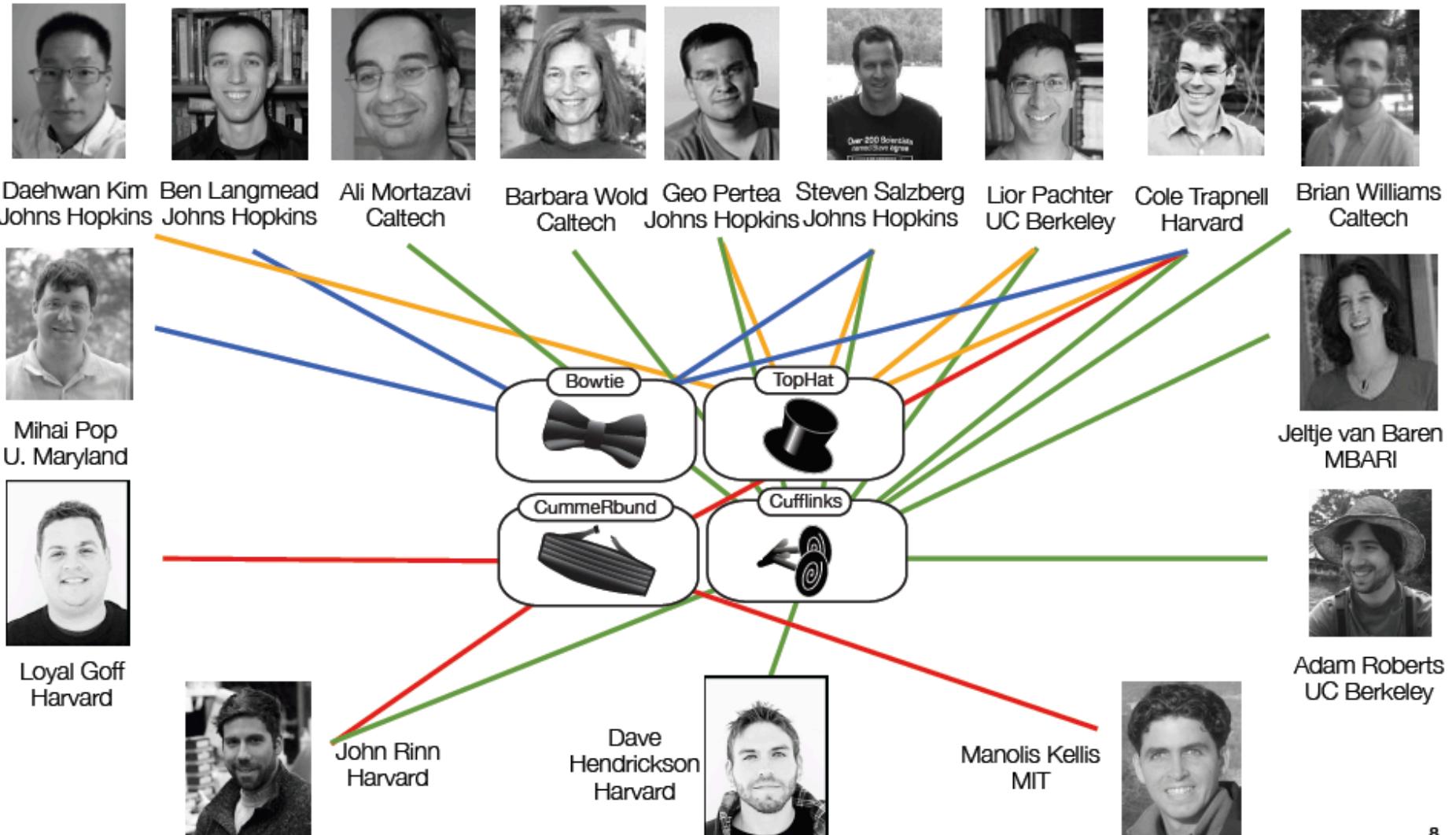
Demo VirtualBox

- Overview of installation
- Run it

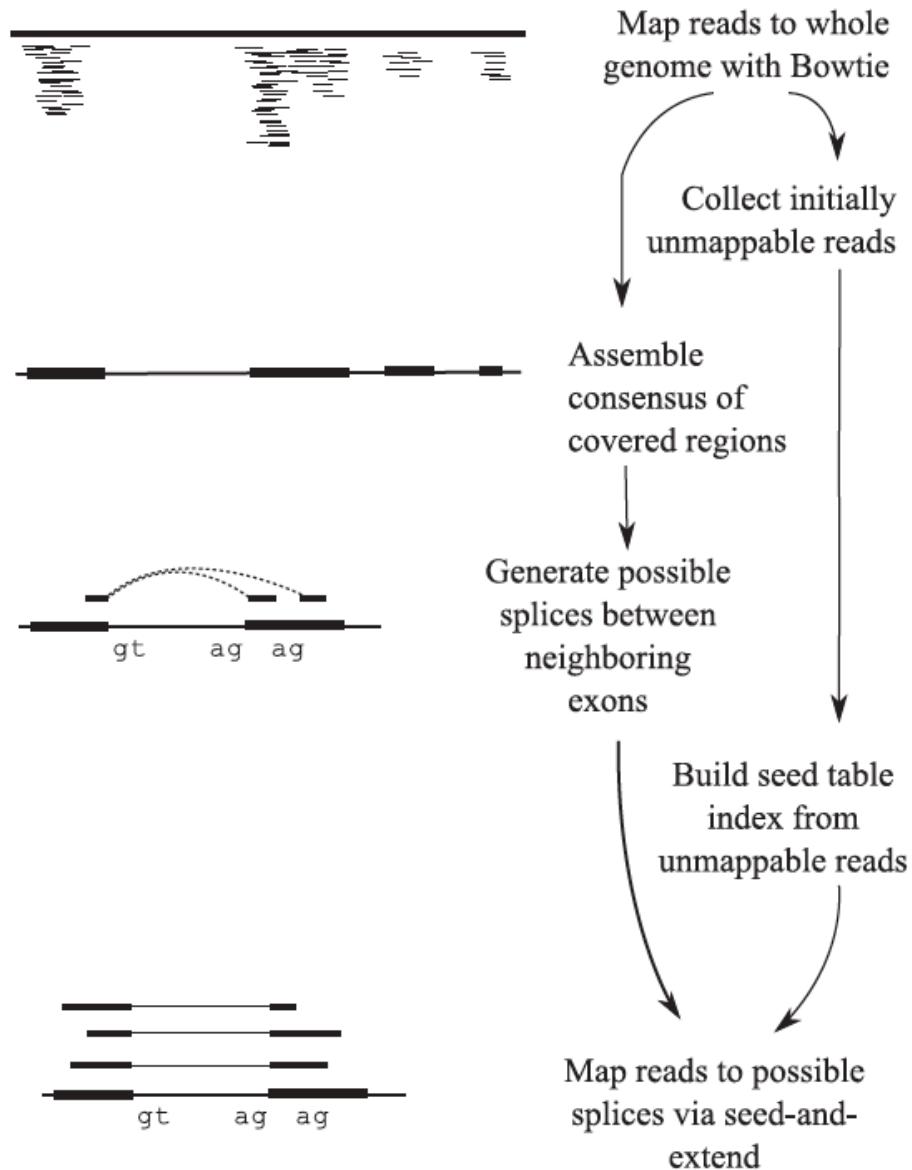
Overview of the Tuxedo Software Suite



Tuxedo development team



The TopHat Pipeline



From Trapnell, Pachter, & Salzberg. Bioinformatics. 2009

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      CCCAAACAAGCCGAACTAGCTGATTTGGCTCGTAAAGACCCGGAAA
10     ###CB?=ADDBCBCDEEFFDEFDEFFGDBEFGEDGCFGFGGGGG
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
```

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)
```

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      (read name)
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: 0.1.18 (r982:295)

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
          tview     text alignment viewer
          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
          targetcut cut fosmid regions (for fosmid pool only)
          phase    phase heterozygotes
```

Visualizing Alignments of RNA-Seq reads

Text-based Alignment Viewer

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

IGV

www.broadinstitute.org/igv/

The screenshot shows the IGV website at www.broadinstitute.org/igv/. The page features a large central image of the IGV software interface, which displays a genomic track with multiple panels showing sequencing data, tracks, and annotations. To the left of the main content area is a sidebar with the IGV logo and a navigation menu. The menu includes links for Home, Downloads, Documents, Hosted Genomes, FAQ, IGV User Guide, File Formats, Release Notes, Credits, Contact, and a search bar. Below the menu is a section for the Broad Institute, including links to Broad Home and Cancer Program, and the text "© 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*.

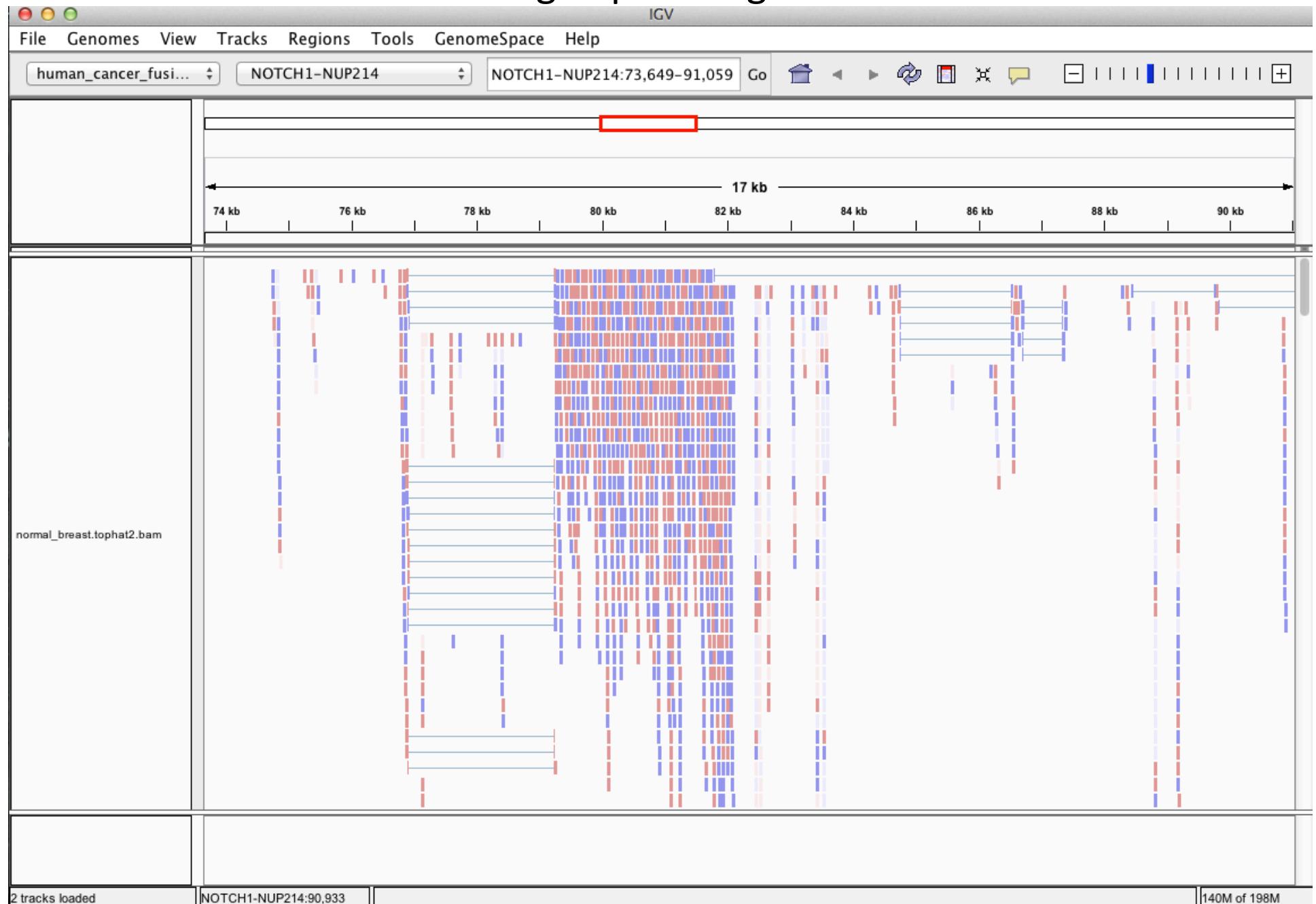
Overview

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 Thorvaldsdóttir, James T. Robinson, Jill P. Mesirov. [Integrative Genomics Viewer \(IGV\): high-performance genomics data visualization and exploration](#).

IGV: Viewing Tophat Alignments



GenomeView

← → C genomeview.org ⭐ a

GenomeView

Demos Plug-ins JAnnot API Join mailing list Support - Frequently asked questions Cite us

Start Now!

Webstart:
[Launch](#)

High-mem webstart

Applet:
[Launch](#)

Documentation

- Quick start guide
- Manual
- Advanced manual
- Tutorials

Navigation

- Download
- Demos
- Plug-ins

GenomeView is a next-generation stand-alone genome browser and editor initiated in the BSB group at VIB and currently developed at Broad Institute. It provides interactive visualization of sequences, annotation, multiple alignments, syntetic mappings, short read alignments and more. Many standard file formats are supported and new functionality can be added using a plugin system.

Getting started



Get started with a five minute quick-start guide that will get up and running in no time

Web start

[Launch](#) Click the launch button to start GenomeView

Download



Download the current release. You can also start GenomeView

Support

If you experience any issues, head over to the [support section](#), we like to help you.

Recent questions

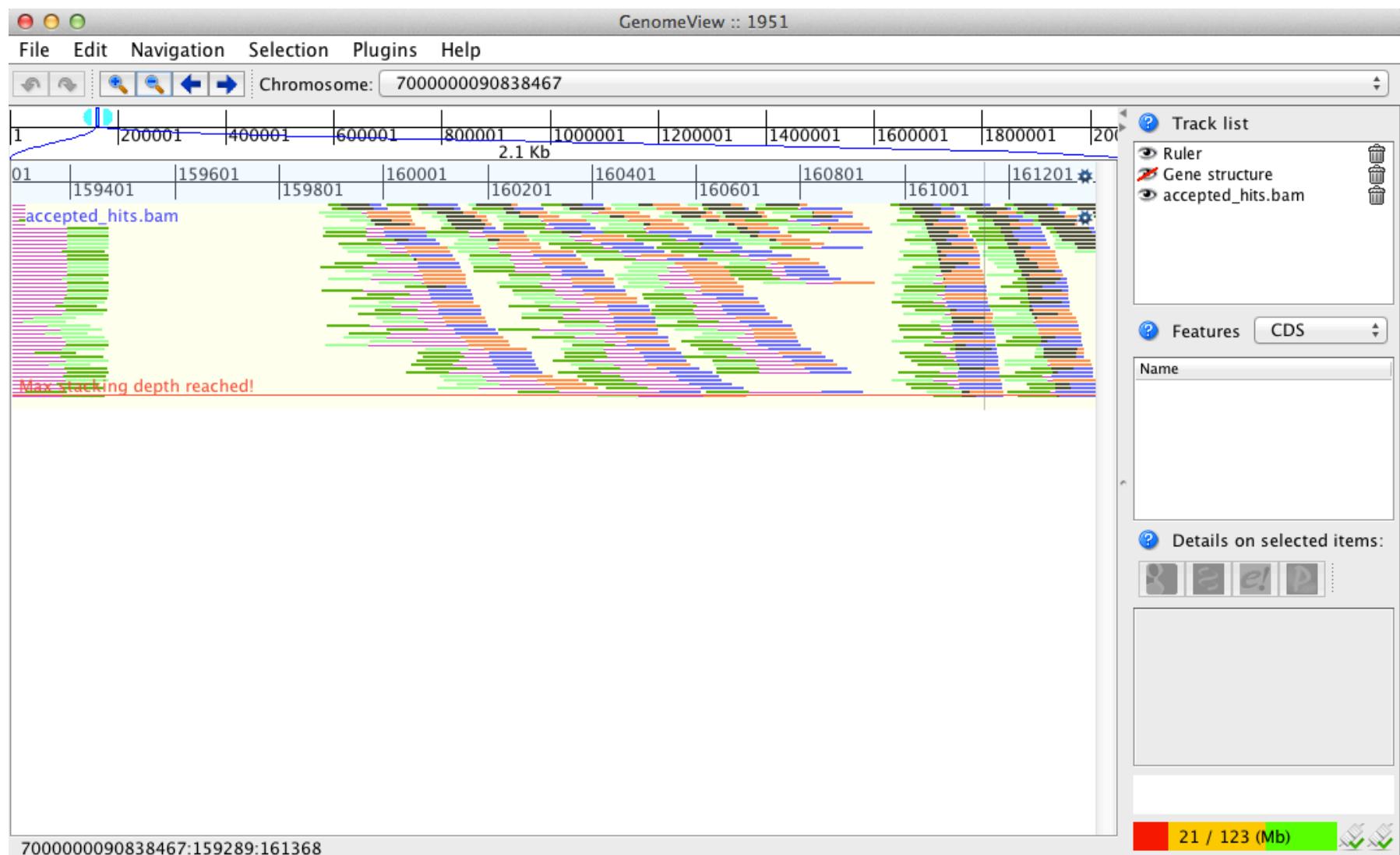
- How do I show annotation on a multiple alignment?
- Why does my multiple alignment load as reference sequences?
- Where do I find documentation?
- Why doesn't GenomeView correctly detect my BED file?
- How do I fix the order of the tracks in an integrated GenomeView instance?
- How do I integrate GenomeView in my



Most creative visualization award @ Illumina iDEA challenge 2011

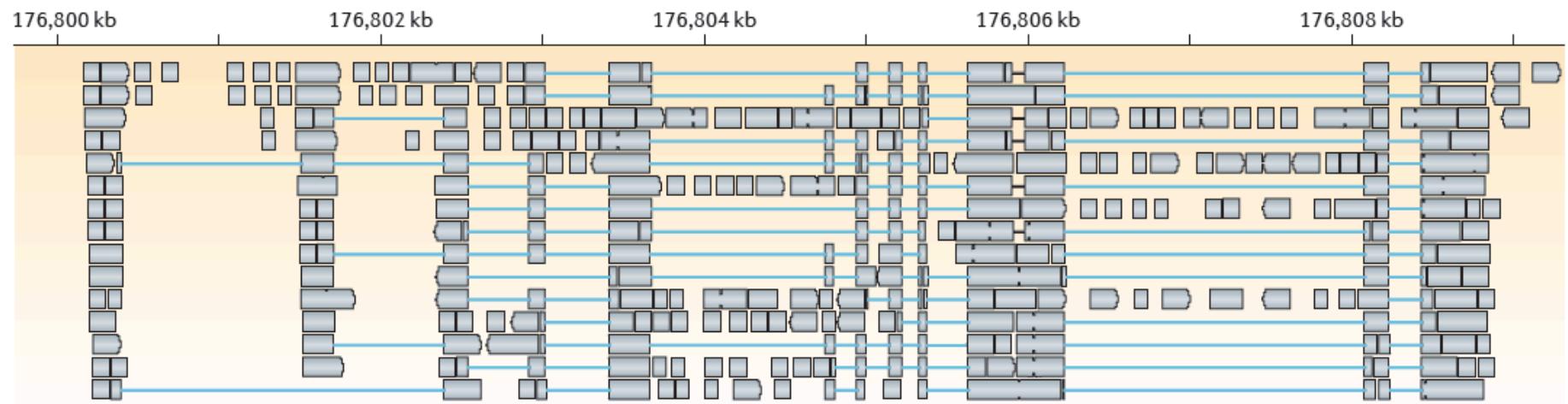


GenomeView: viewing TopHat alignments



Transcript Reconstruction Using Cufflinks

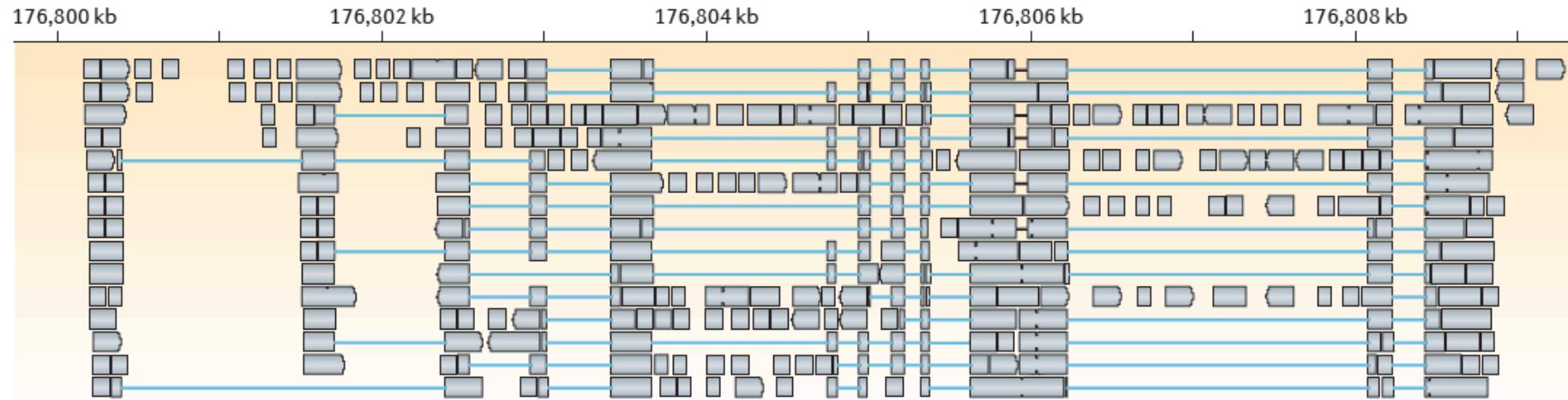
a Splice-align reads to the genome



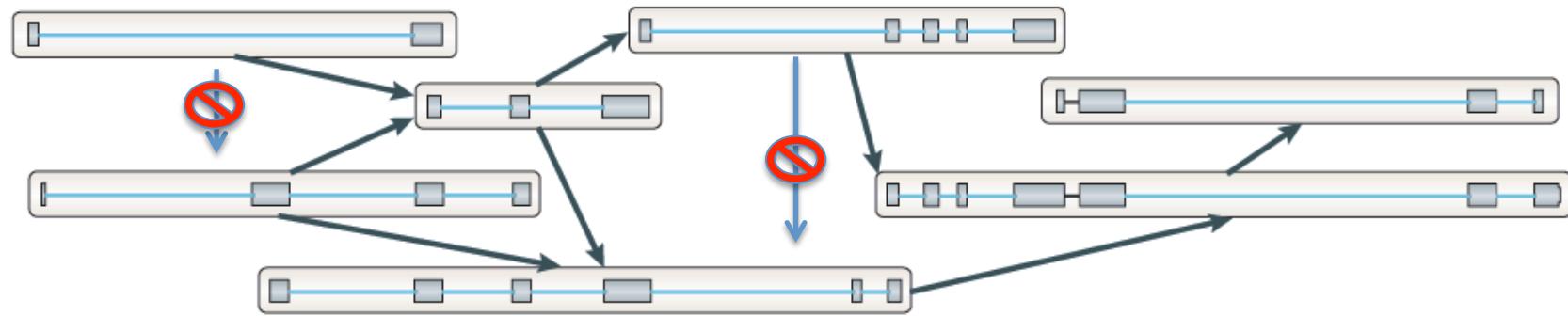
From Martin & Wang. Nature Reviews in Genetics. 2011

Transcript Reconstruction Using Cufflinks

a Splice-align reads to the genome

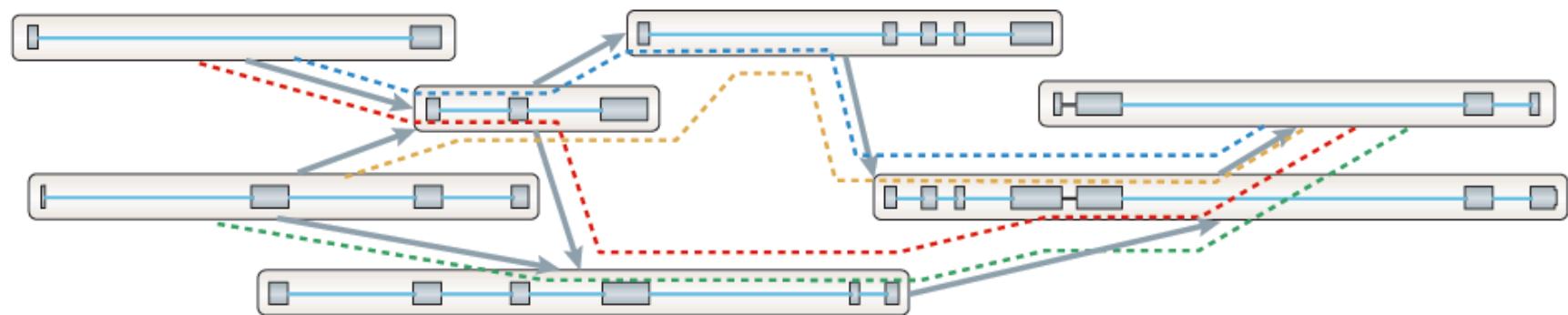


b Build a graph representing alternative splicing events

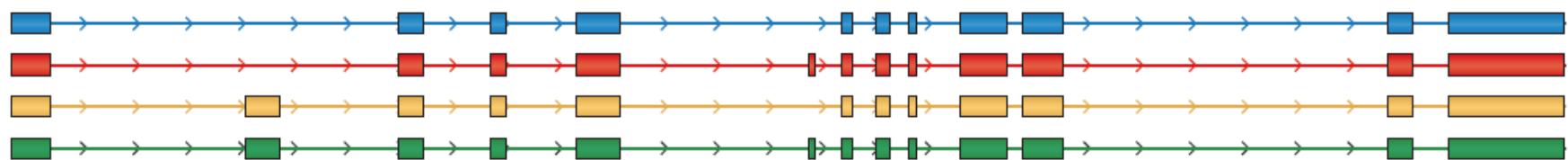


Transcript Reconstruction Using Cufflinks

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 7000000090838467  (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 7000000090838467
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

- 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

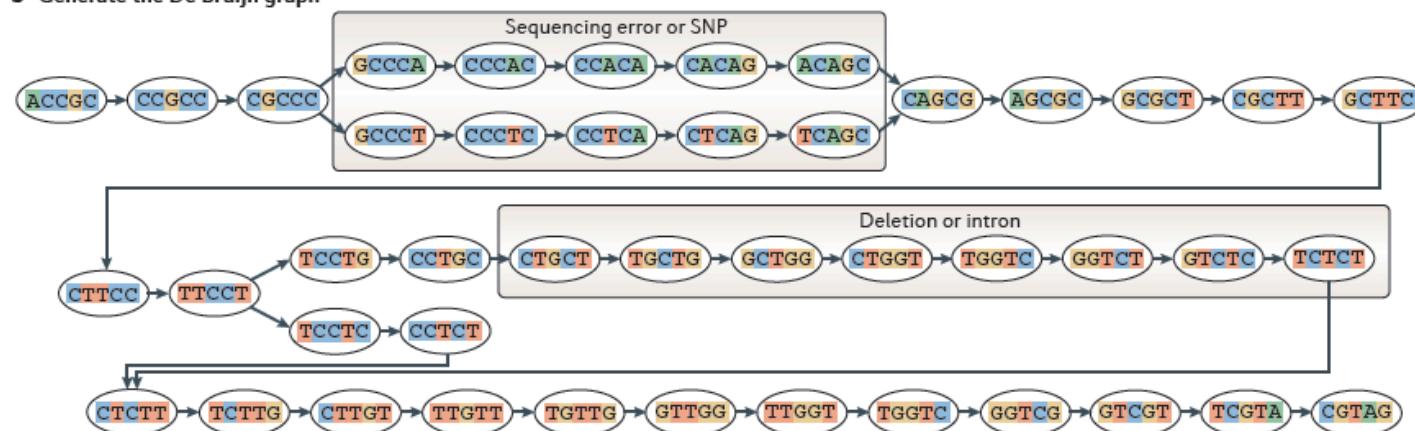
The General Approach to
De novo RNA-Seq Assembly
Using De Bruijn Graphs

Sequence Assembly via De Bruijn Graphs

a Generate all substrings of length k from the reads

ACAGC	TCCTG	GTCTC		AGCGC	CTCTT	GGTCG	k-mers (k=5)
CACAG	TTCCT	GGTCT		CAGCG	CCTCT	TGGTC	
CCACA	CTTCC	TGGTC	TGTTG	TCAGC	TCCTC	TTGGT	
CCCAC	GCTTC	CTGGT	TTGTT	CTCAG	TTCCT	GTTGG	
GCCCC	CGCTT	GCTGG	CTTGT	CCTCA	CTTCC	TGTTG	
CGCCC	GCGCT	TGCTG	TCTTG	CCCTC	GCTTC	TTGTT	
CCGCC	AGCGC	CTGCT	CTCTT	GCCCT	CGCTT	CTTGT	
ACCGC	CAGCG	CCTGC	TCTCT	CGCCC	GCGCT	TCTTG	
ACCGCCACAGCGCTTCCTGCTGGTCTCTGTTG				CGCCCTCAGCGCTTCCTGTTGGTCTGAG			

b Generate the De Bruijn graph

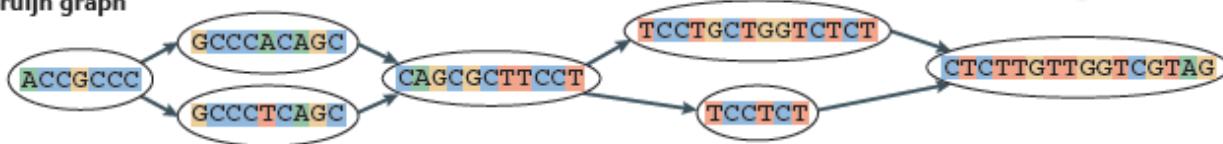


From Martin & Wang, Nat. Rev. Genet. 2011

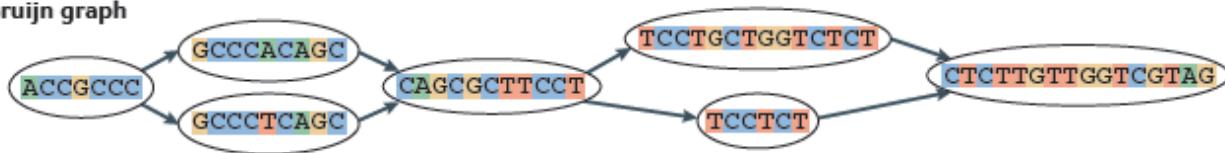
b Generate the De Bruijn graph



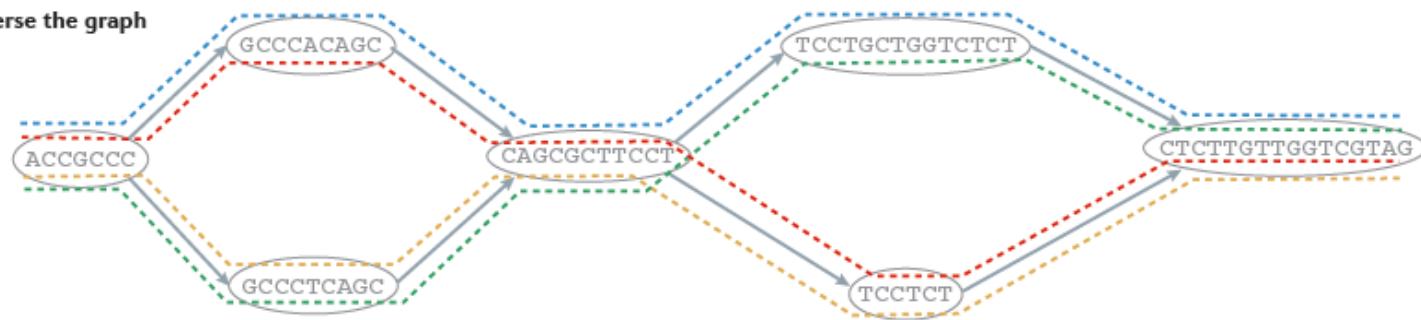
c Collapse the De Bruijn graph



c Collapse the De Bruijn graph



d Traverse the graph



e Assembled isoforms

— ACCGCCACAGCGCTTCCTGCTGGTCTTTGGTGGTCGTAG
- - - ACCGCCACAGCGCTTCCT - - - CTTGGTGGTCGTAG
- - - ACCGCCCTCAGCGCTTCCT - - - CTTGGTGGTCGTAG
- - - ACCGCCCTCAGCGCTTCCTGCTGGTCTTTGGTGGTCGTAG

Contrasting Genome and Transcriptome Assembly

Genome Assembly

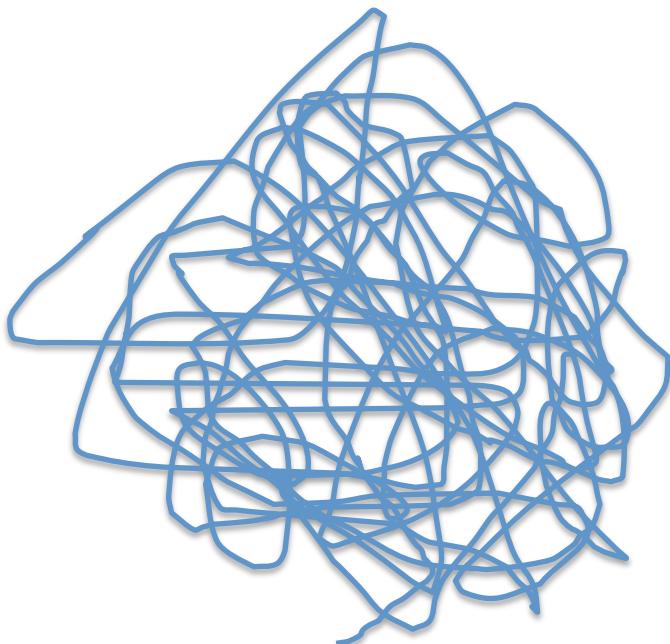
Transcriptome Assembly



Trinity Aggregates Isolated Transcript Graphs

Genome Assembly

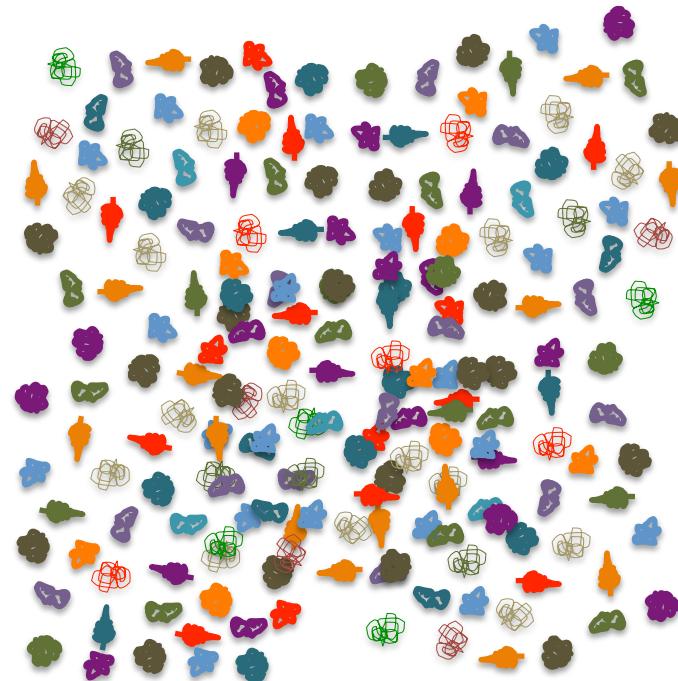
Single Massive Graph



Entire chromosomes represented.

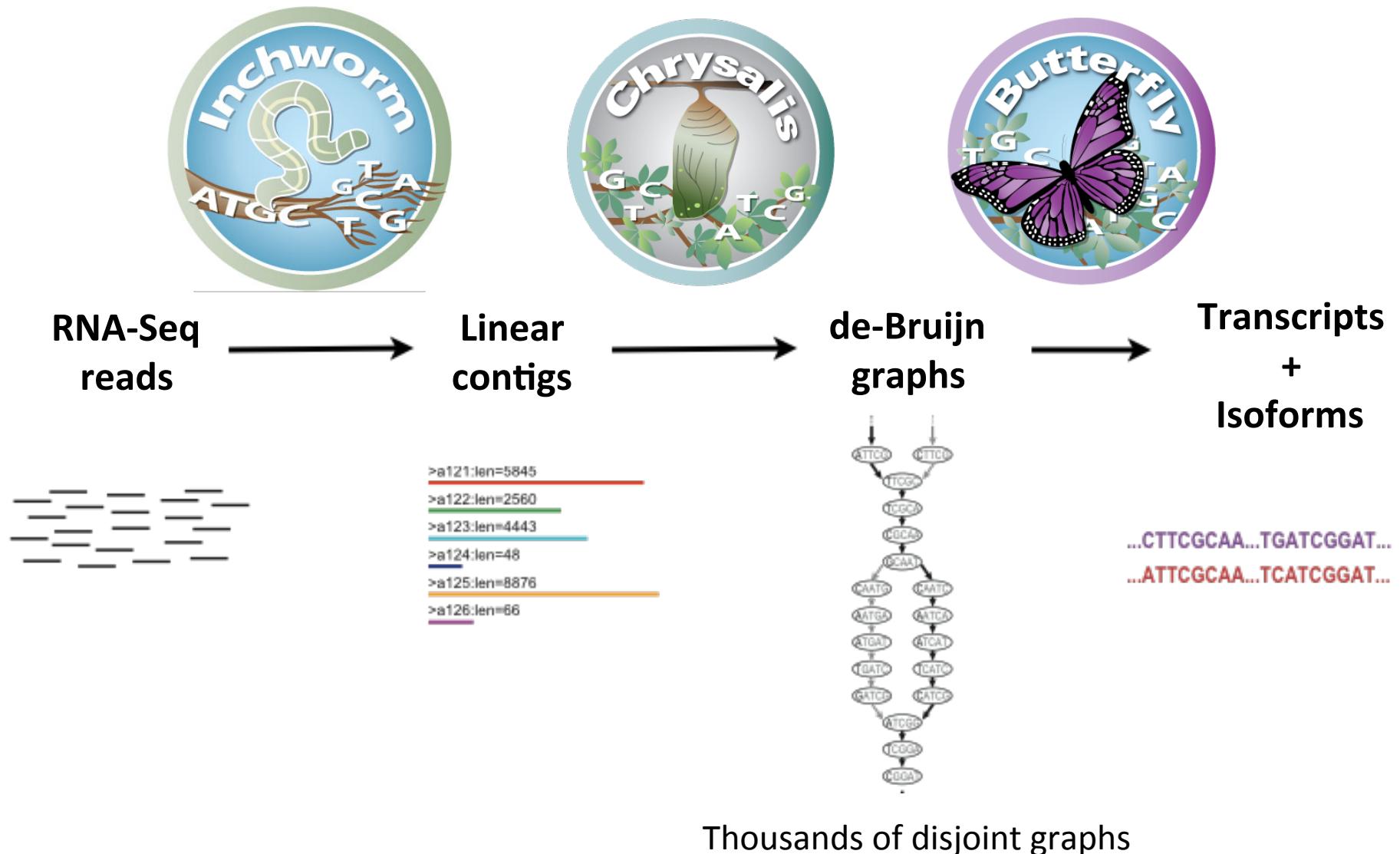
Trinity Transcriptome Assembly

Many Thousands of Small Graphs



Ideally, one graph per expressed gene.

Trinity – How it works:



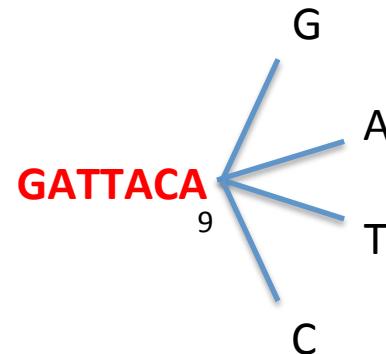


Inchworm Algorithm

Decompose all reads into overlapping Kmers (25-mers)

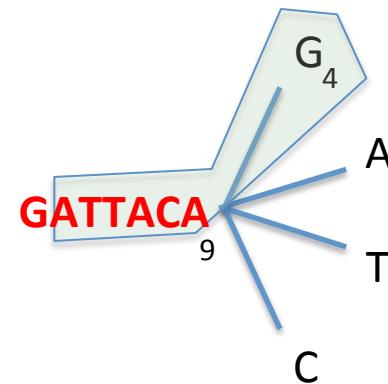
Identify seed kmer as most abundant Kmer, ignoring low-complexity kmers.

Extend kmer at 3' end, guided by coverage.



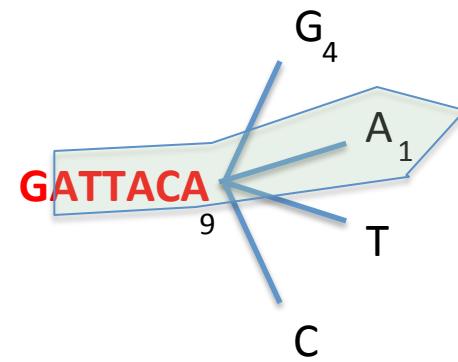


Inchworm Algorithm



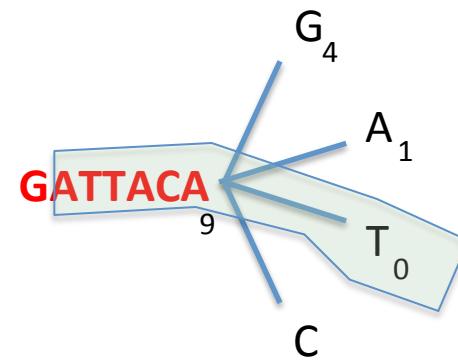


Inchworm Algorithm



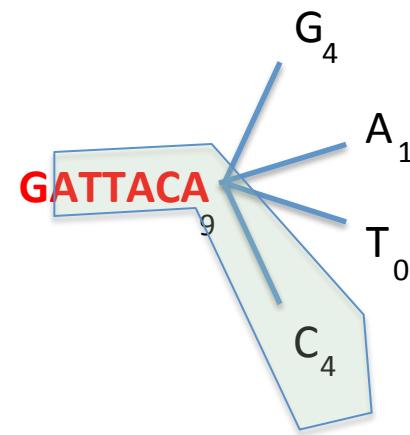


Inchworm Algorithm



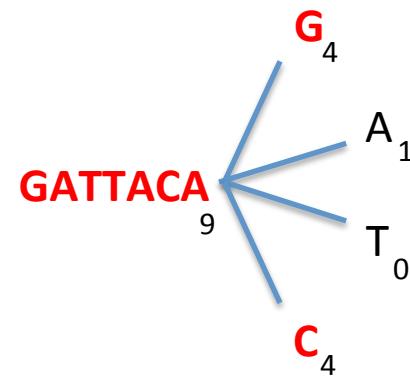


Inchworm Algorithm



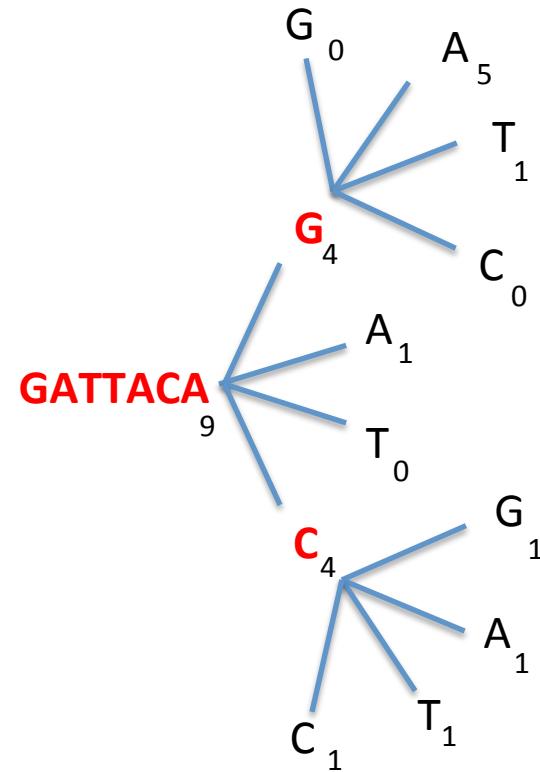


Inchworm Algorithm



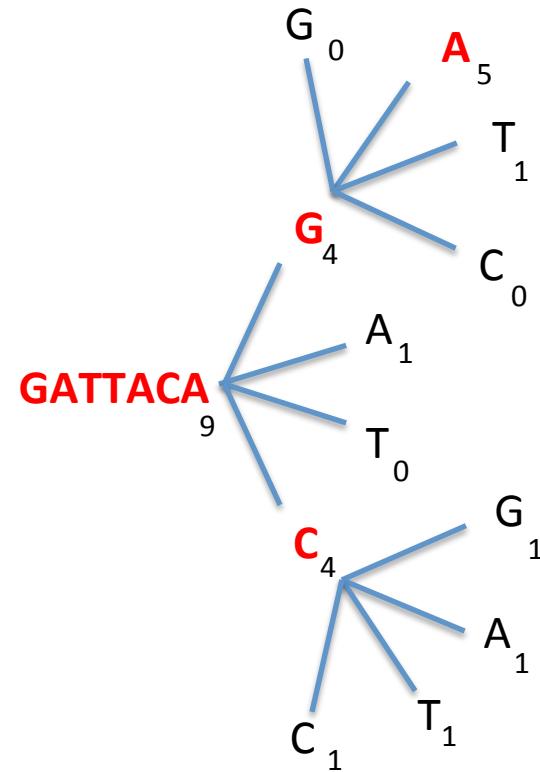


Inchworm Algorithm



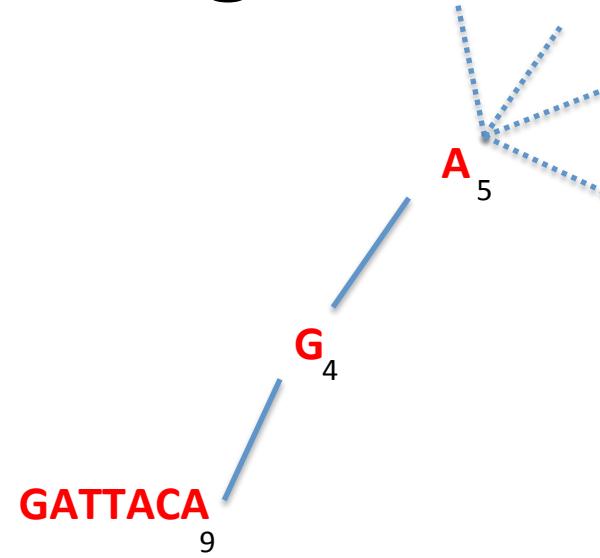


Inchworm Algorithm



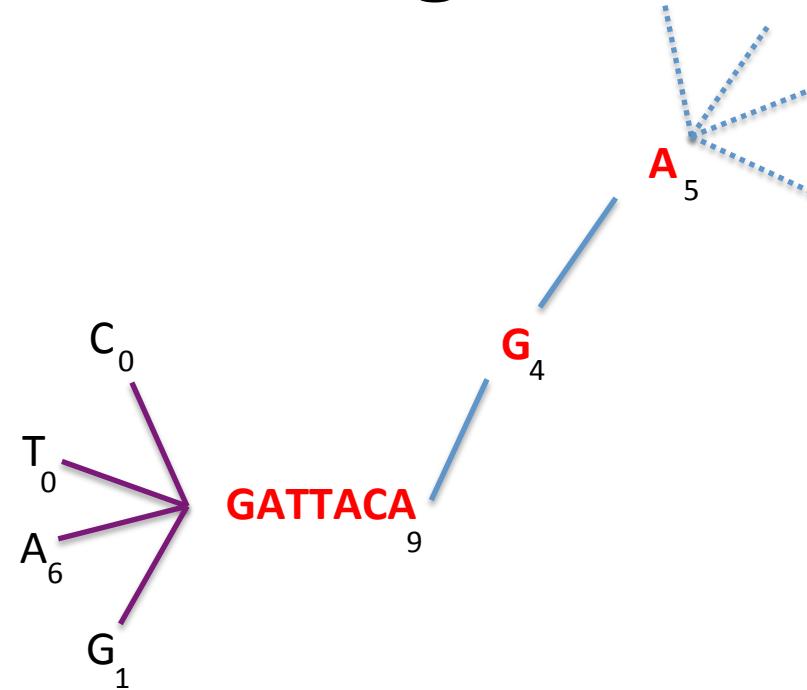


Inchworm Algorithm



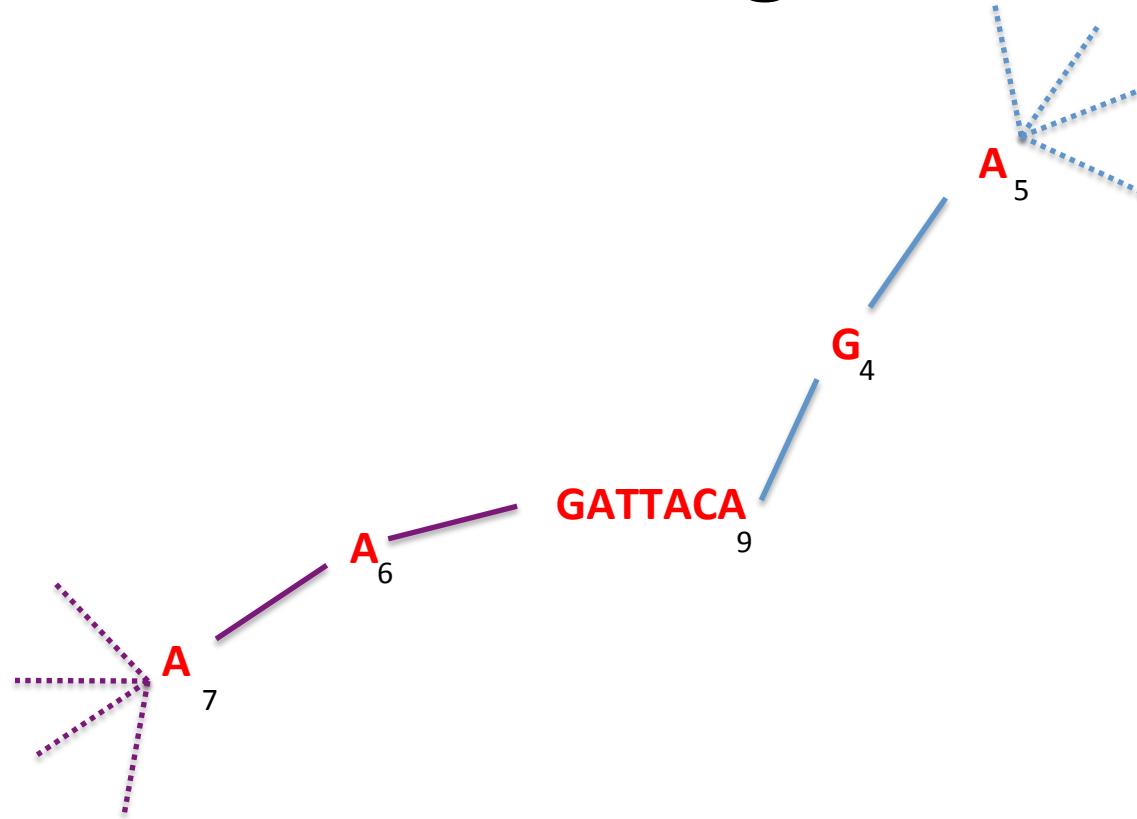


Inchworm Algorithm





Inchworm Algorithm



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

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



Inchworm Contigs from Alt-Spliced Transcripts

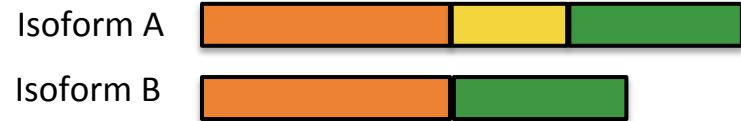
Expressed isoforms





Inchworm Contigs from Alt-Spliced Transcripts

Expressed isoforms

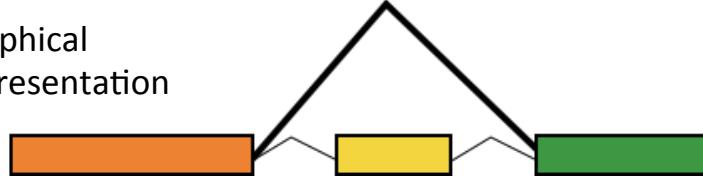


Expression

(low)

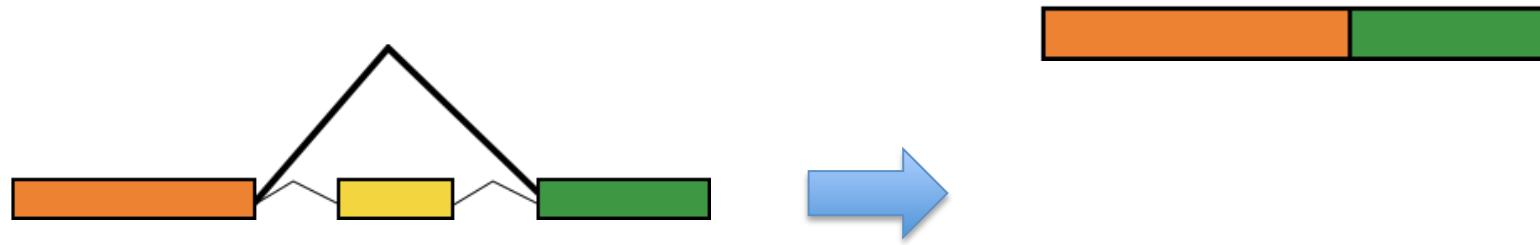
(high)

Graphical representation



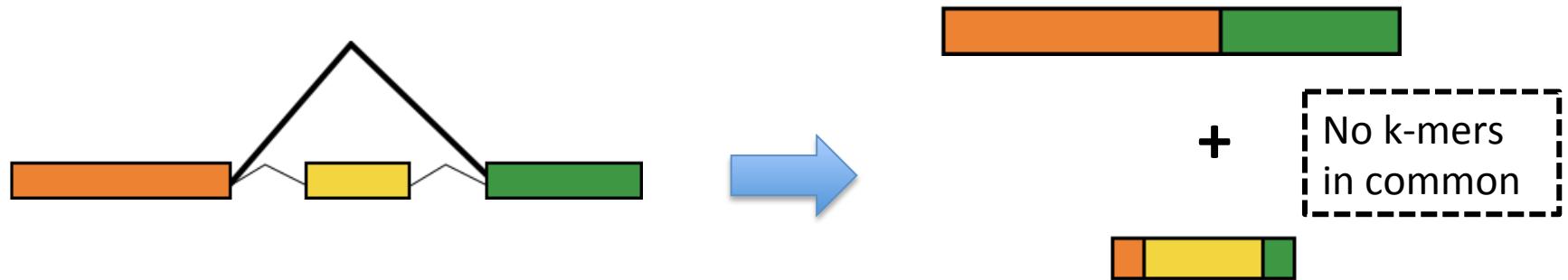


Inchworm Contigs from Alt-Spliced Transcripts



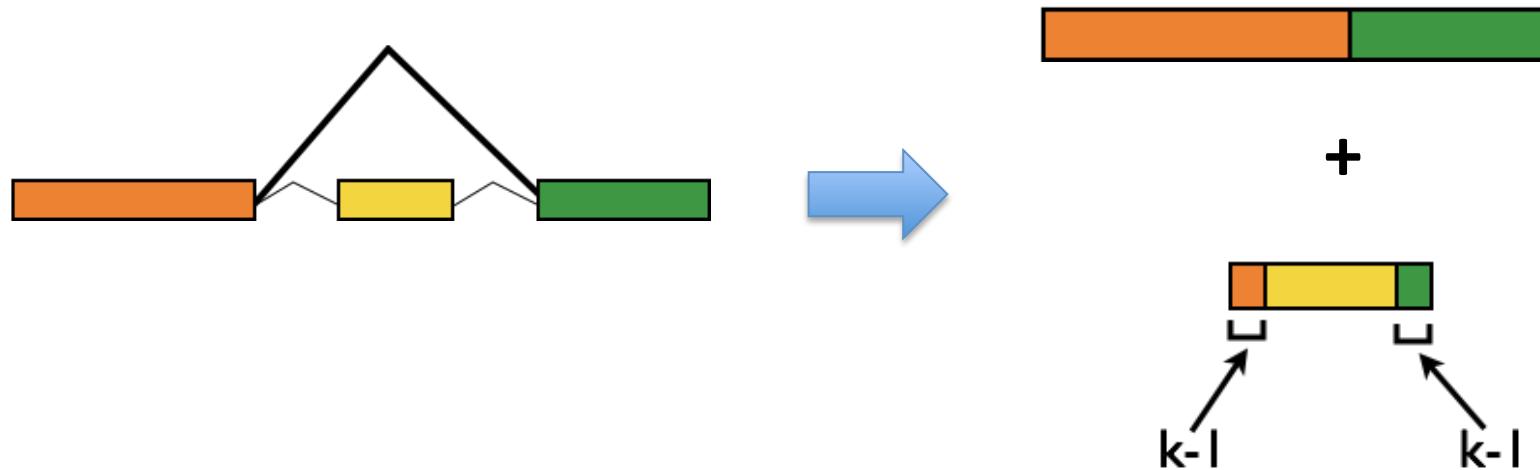


Inchworm Contigs from Alt-Spliced Transcripts

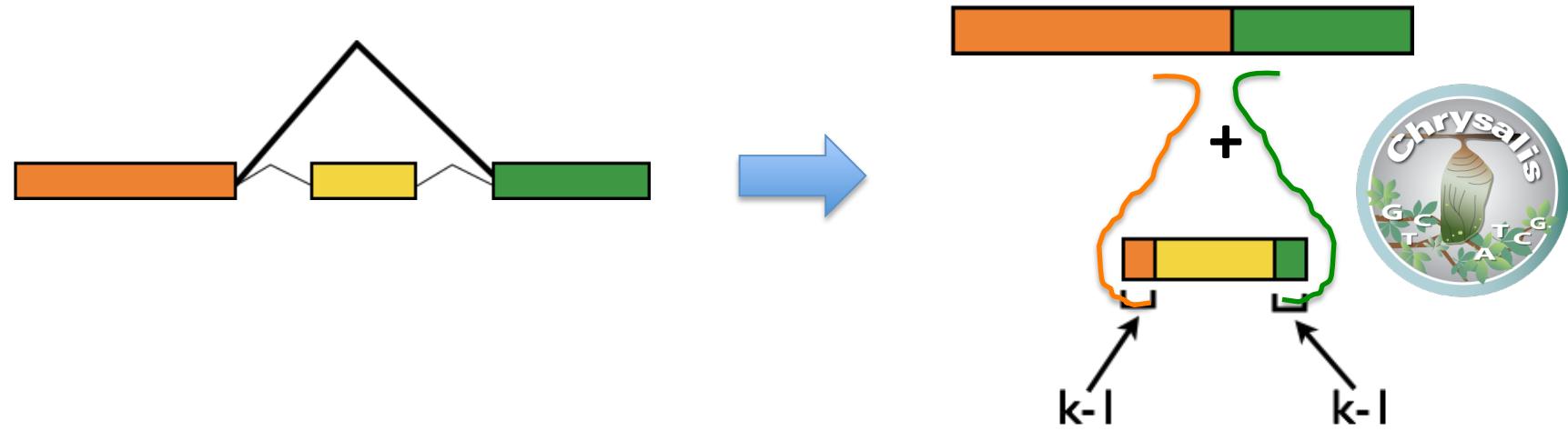




Inchworm Contigs from Alt-Spliced Transcripts



Chrysalis Re-groups Related Inchworm Contigs



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

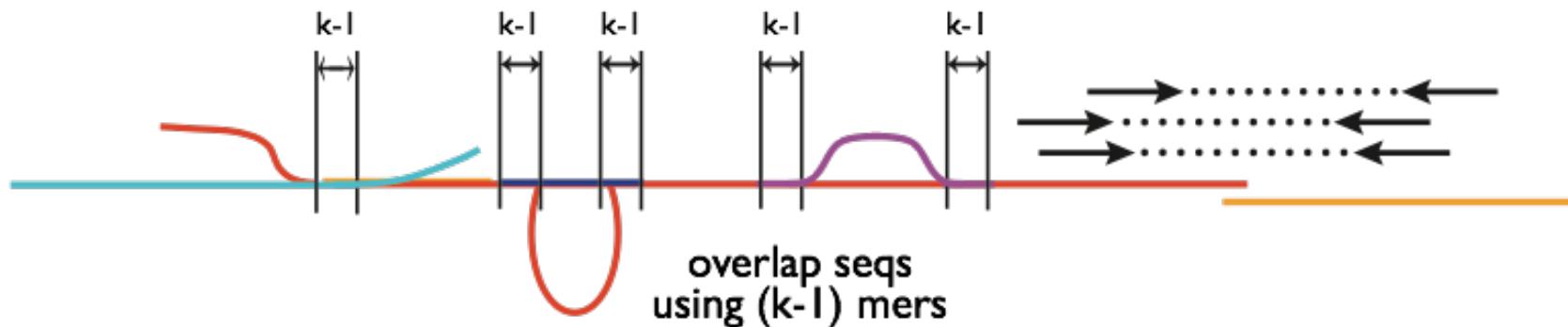
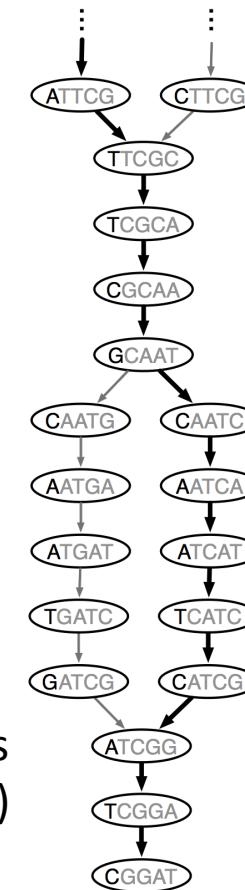
Chrysalis

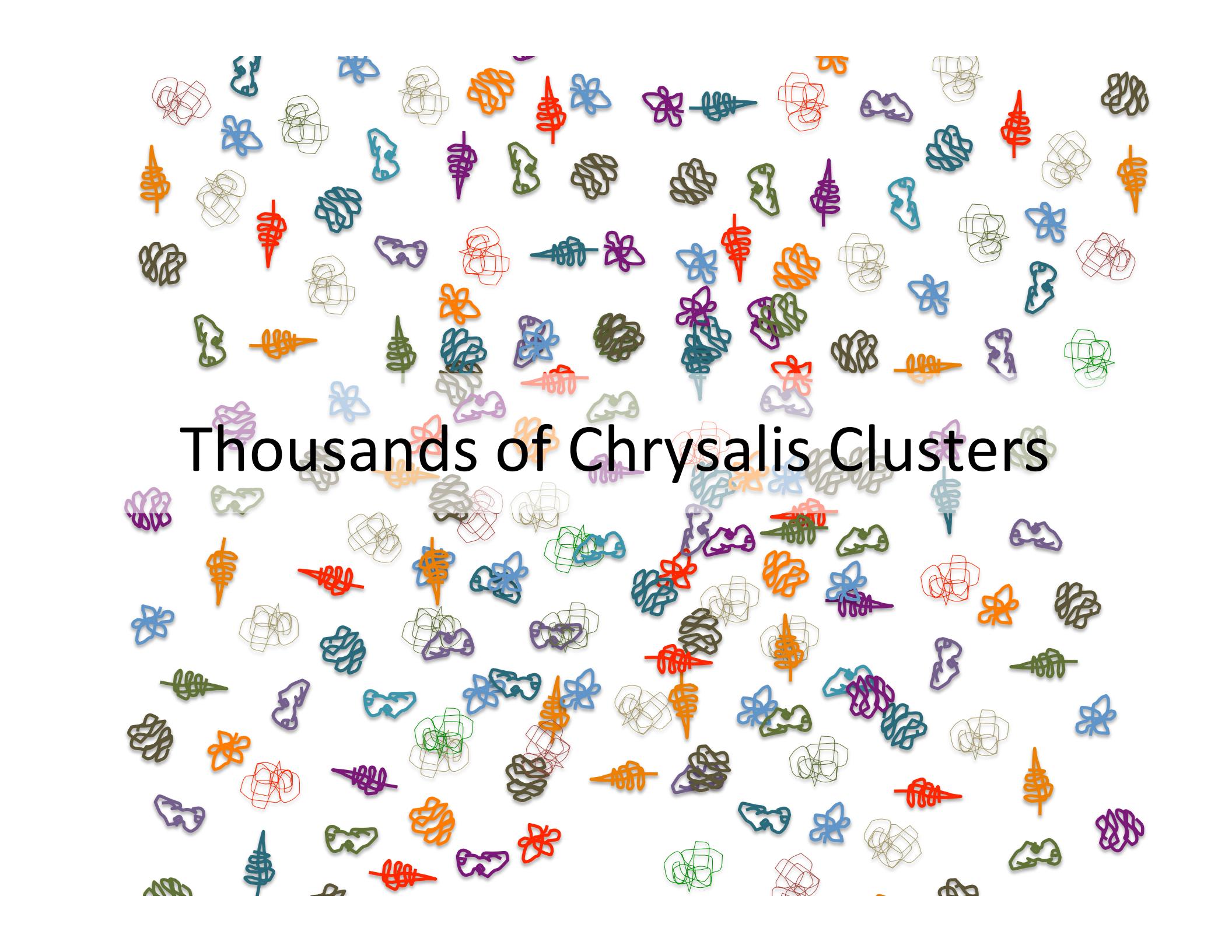
>a121:len=5845
 |
>a122:len=2560
 |
>a123:len=4443
 |
>a124:len=48
 |
>a125:len=8876
 |
>a126:len=66

Integrate isoforms
via k-1 overlaps

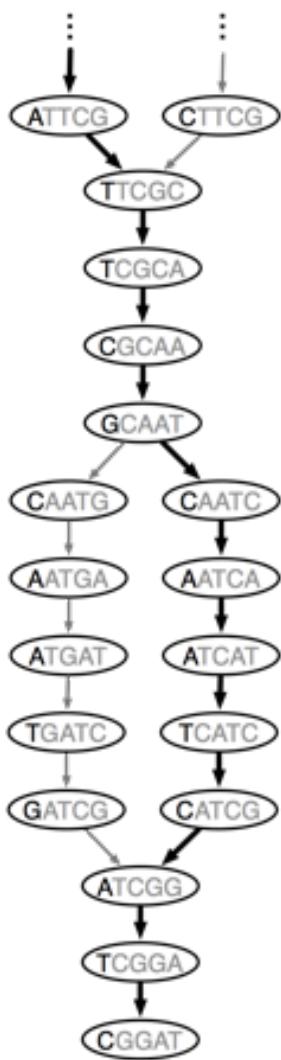


Build de Bruijn Graphs
(ideally, one per gene)



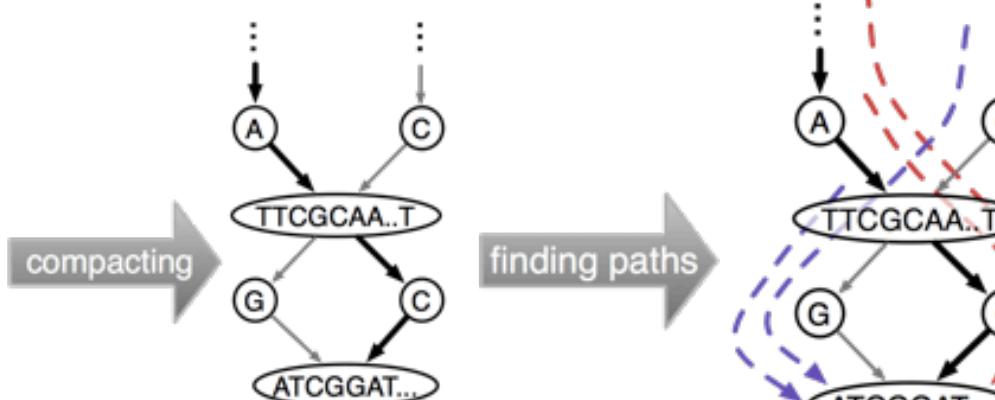


Thousands of Chrysalis Clusters

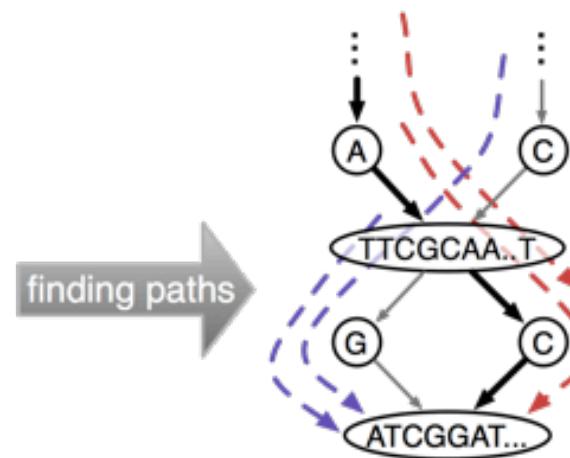


de Bruijn
graph

Butterfly



compact
graph



compact
graph with
reads



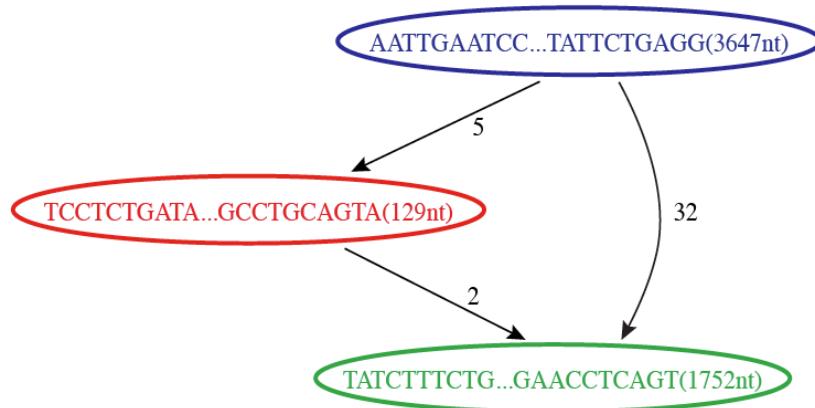
..CTTCGCAA..TGATCGGAT...

..ATTCGCAA..TCATCGGAT...

sequences
(isoforms and paralogs)

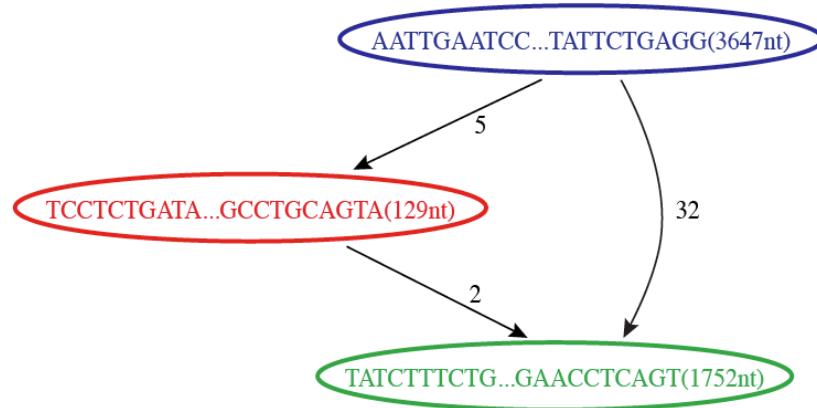
Butterfly Example 1: Reconstruction of Alternatively Spliced Transcripts

Butterfly's Compacted
Sequence Graph



Reconstruction of Alternatively Spliced Transcripts

Butterfly's Compacted Sequence Graph

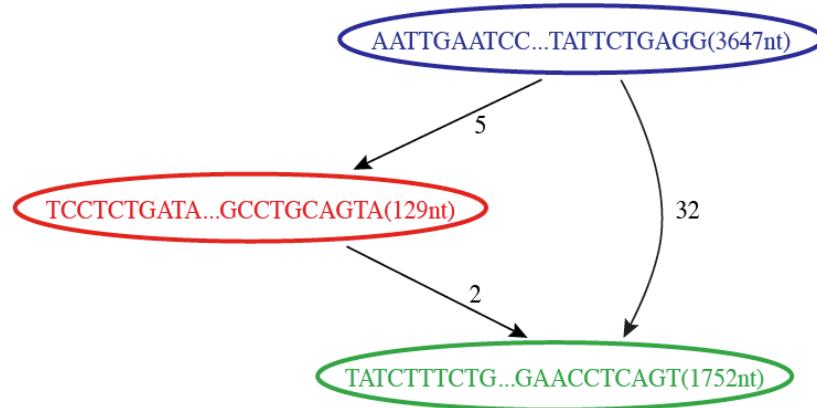


Reconstructed Transcripts



Reconstruction of Alternatively Spliced Transcripts

Butterfly's Compacted Sequence Graph

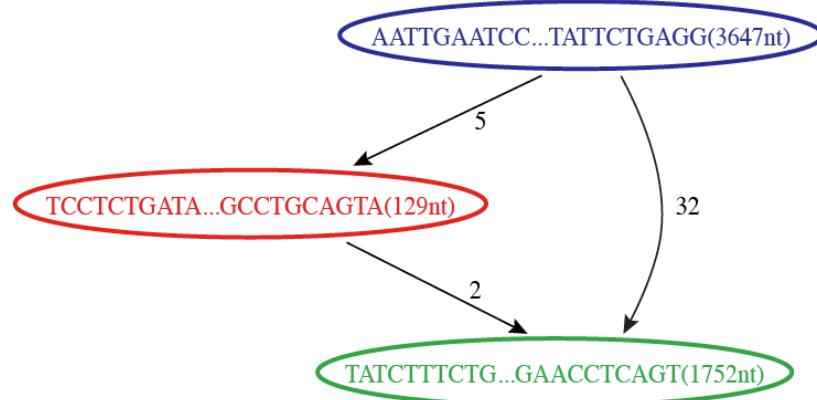


Reconstructed Transcripts



Reconstruction of Alternatively Spliced Transcripts

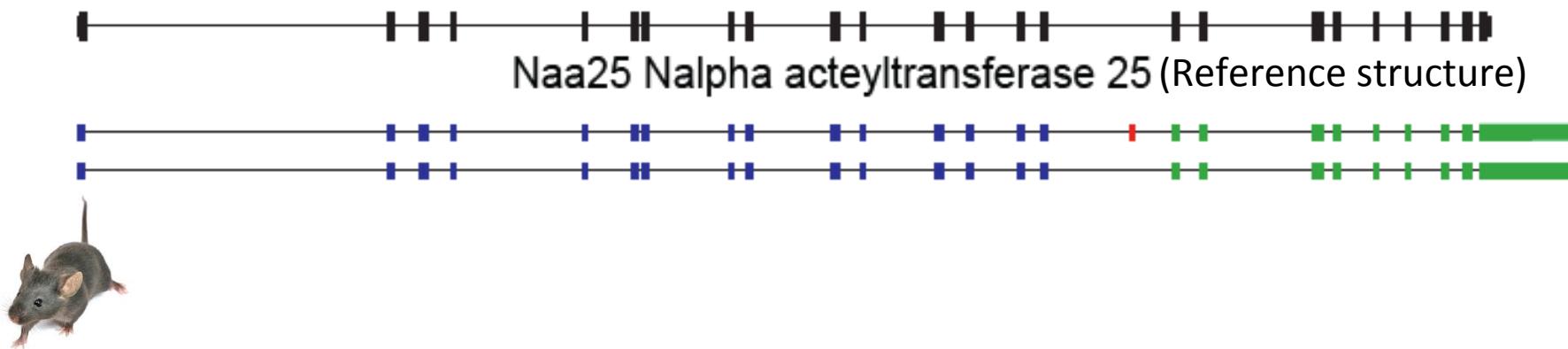
Butterfly's Compacted Sequence Graph



Reconstructed Transcripts



Aligned to Mouse Genome



Butterfly Example 2: Teasing Apart Transcripts of Paralogous Genes



Teasing Apart Transcripts of Paralogous Genes

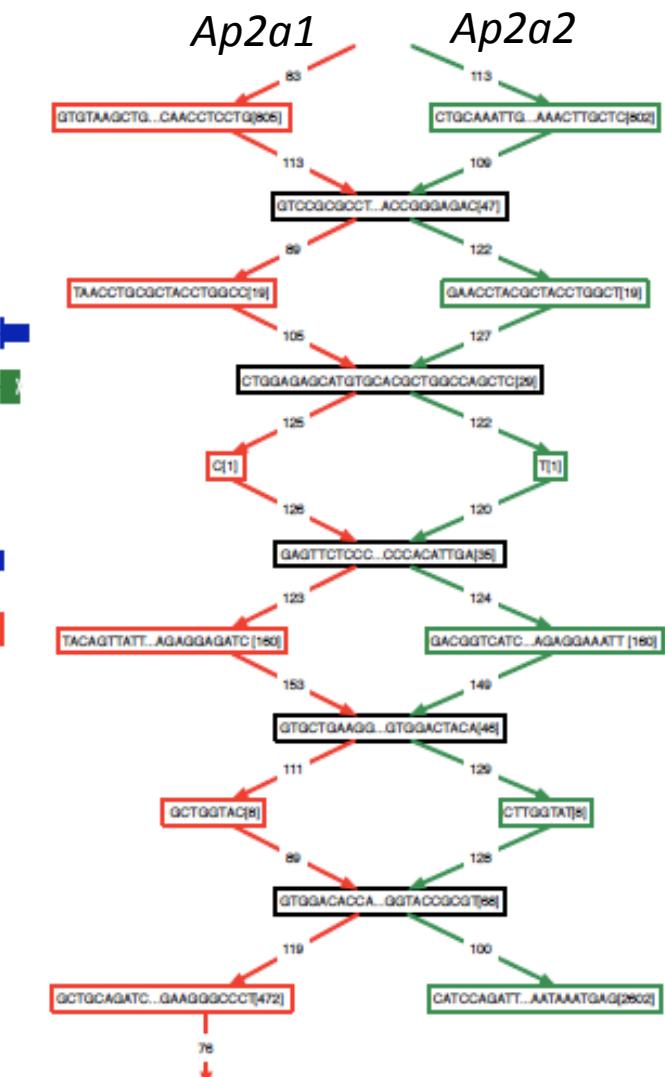
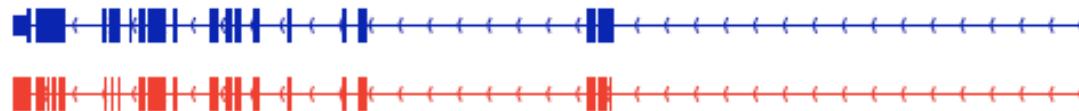
chr7:148,744,197–148,821,437

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

- 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

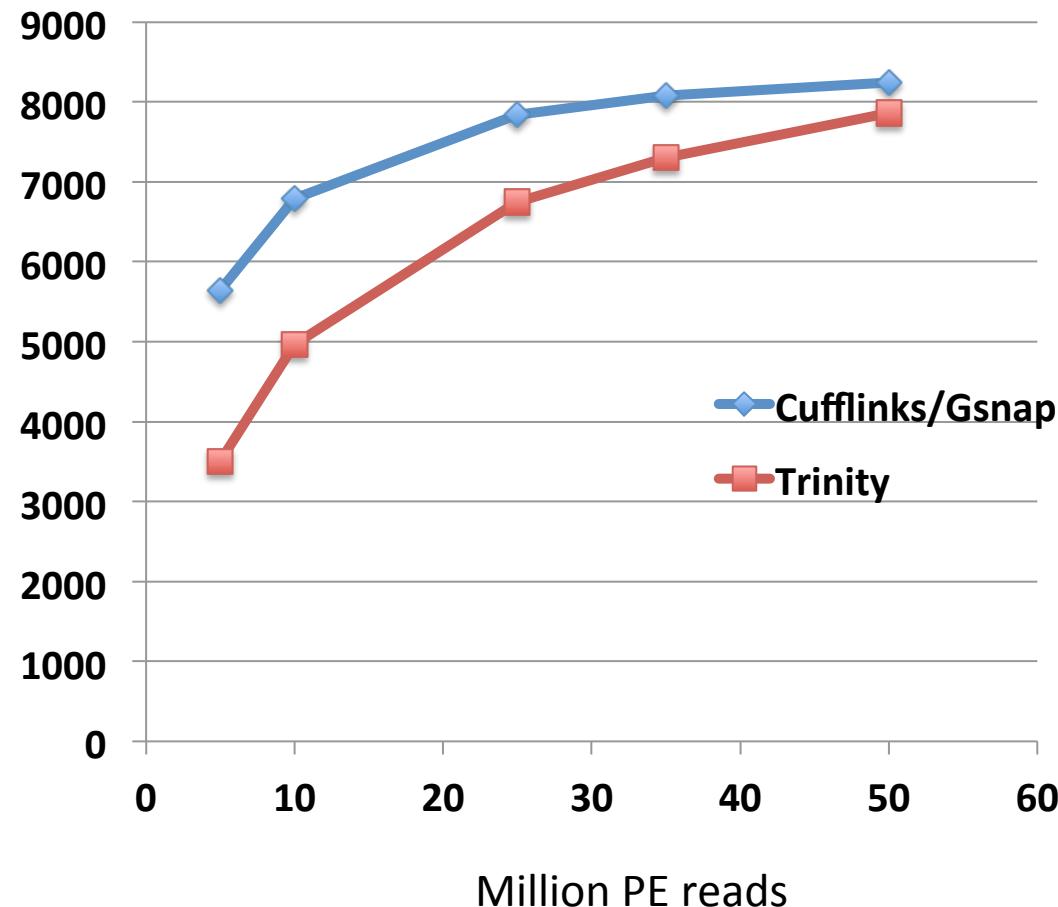
Improved reconstruction with deeper sequencing depth and

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

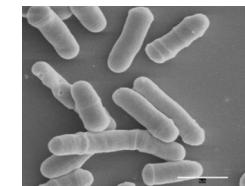
Nevertheless, direct information on the originating strand can substantially enhance the value of an RNA-seq experiment. For

'dUTP second strand marking' identified as the leading protocol

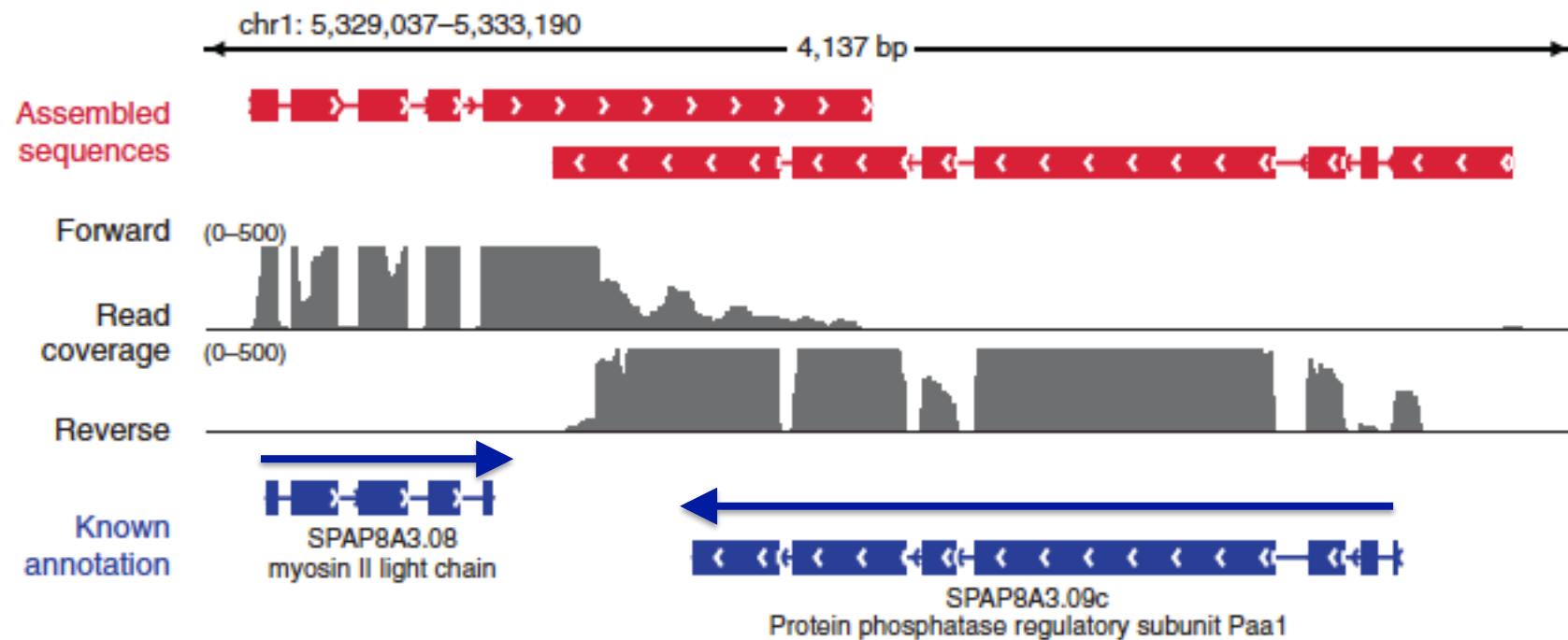
to choose between them. Here we developed a comprehensive computational pipeline to compare library quality metrics from any RNA-seq method. Using the well-annotated *Saccharomyces cerevisiae* transcriptome as a benchmark, we compared seven library-construction protocols, including both published and

transcribed strand or other noncoding RNAs; delineate 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

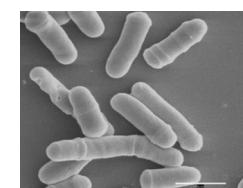
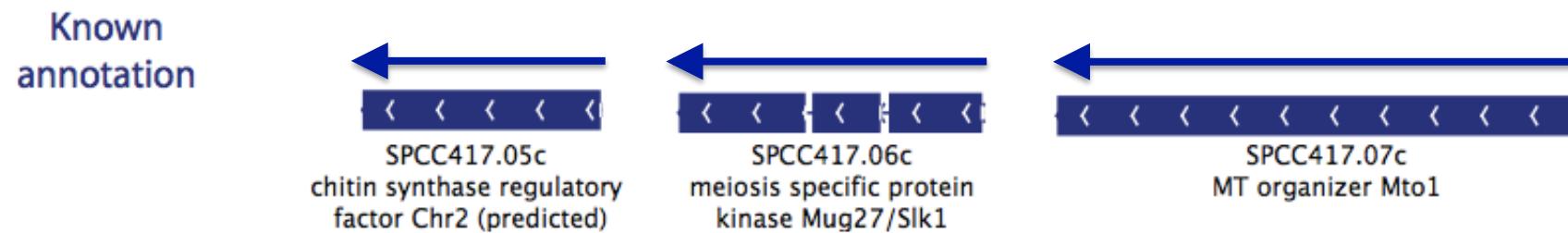
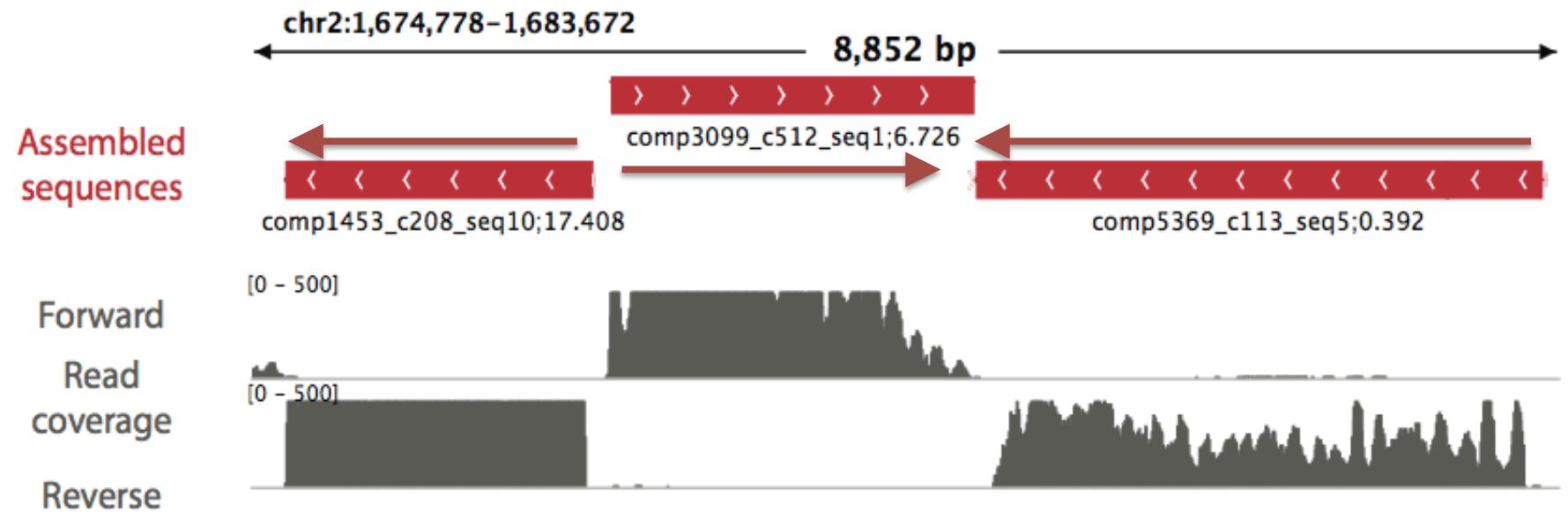
Overlapping UTRs from Opposite Strands



Schizosaccharomyces pombe
(fission yeast)



Antisense-dominated Transcription



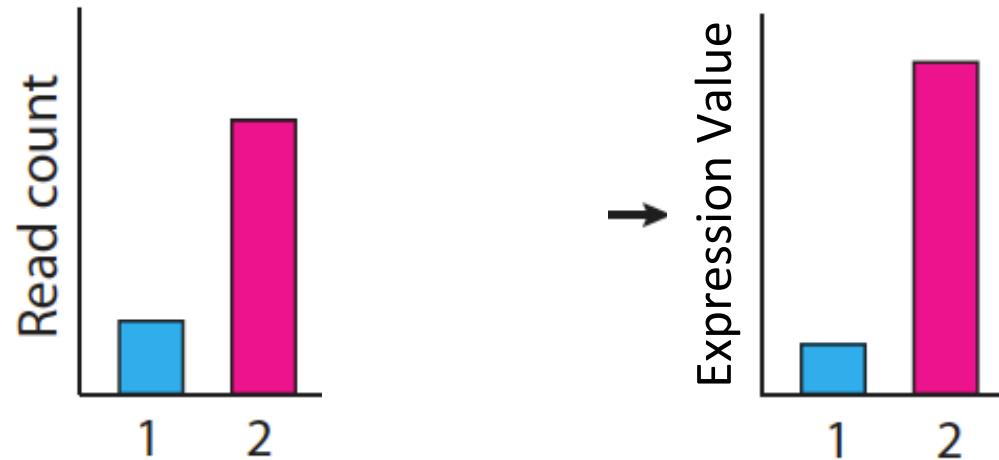
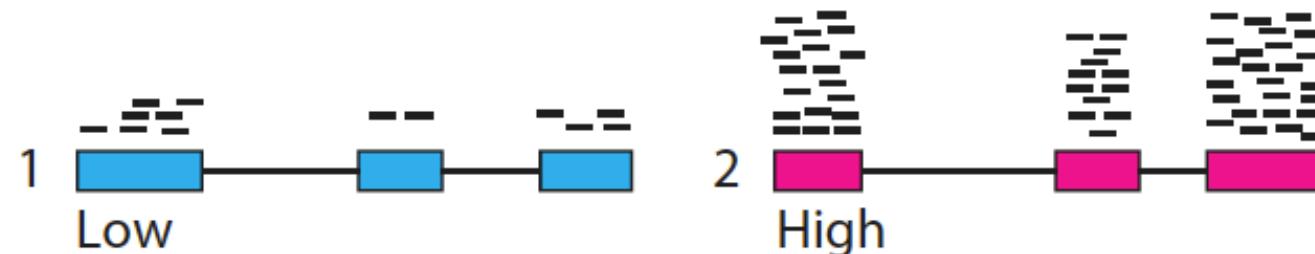
Summary

- 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

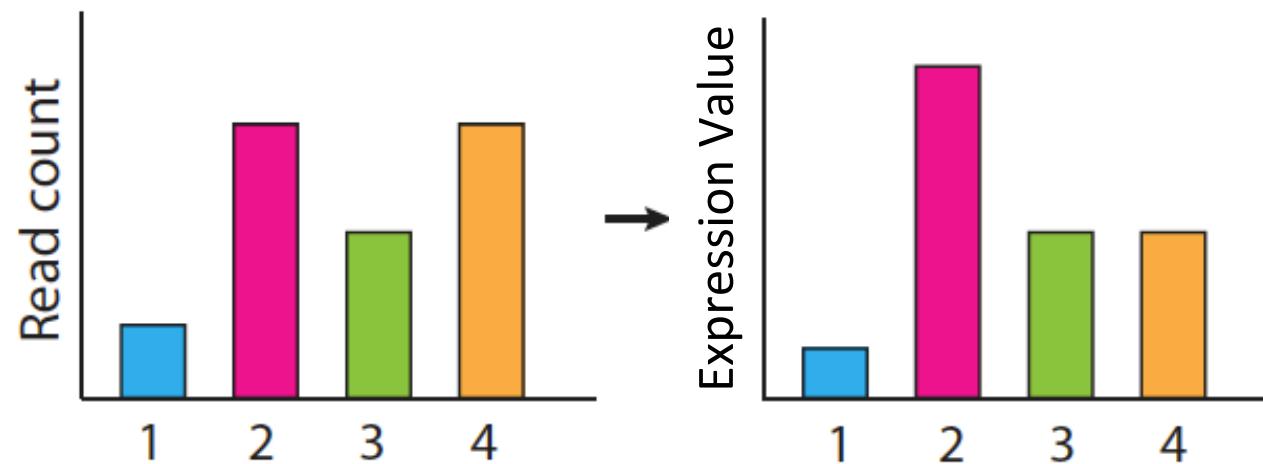
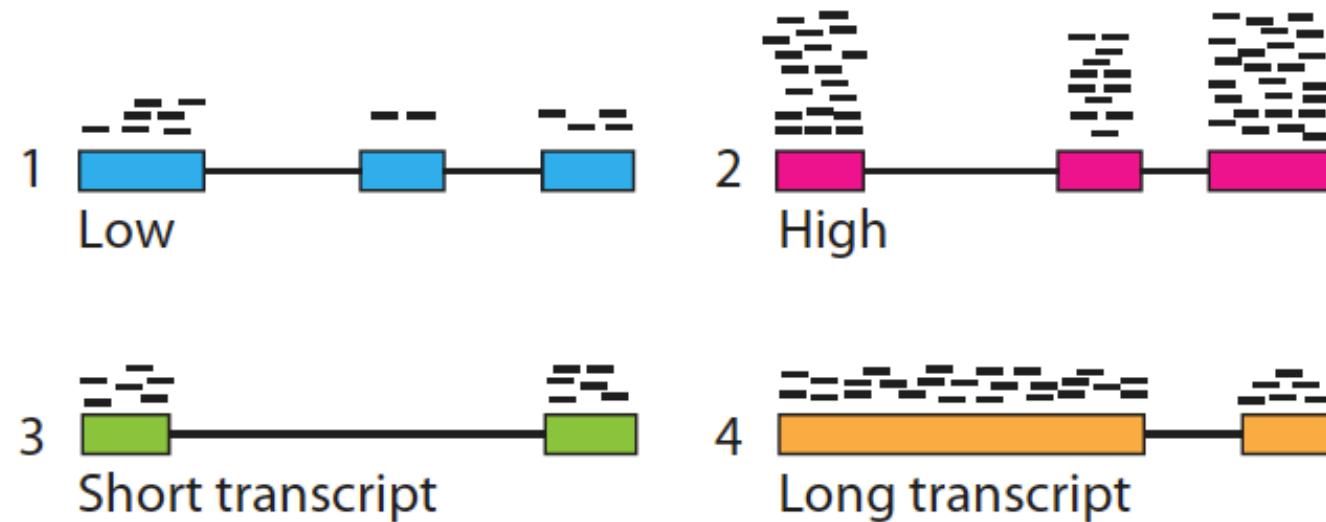
(Aka. Computing Expression Values)

Calculating expression of genes and transcripts



Slide courtesy of Cole Trapnell

Calculating expression of genes and transcripts



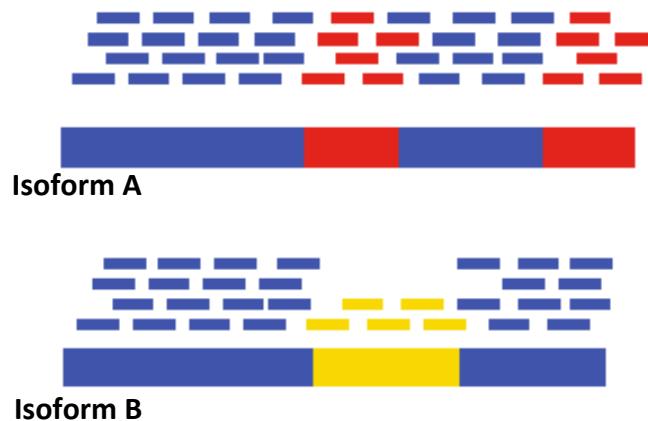
Slide courtesy of Cole Trapnell

Normalized Expression Values

- Normalized for both length of the transcript and total depth of sequencing.
- Number of RNA-Seq Fragments
Per Kilobase of transcript
per total Million fragments mapped

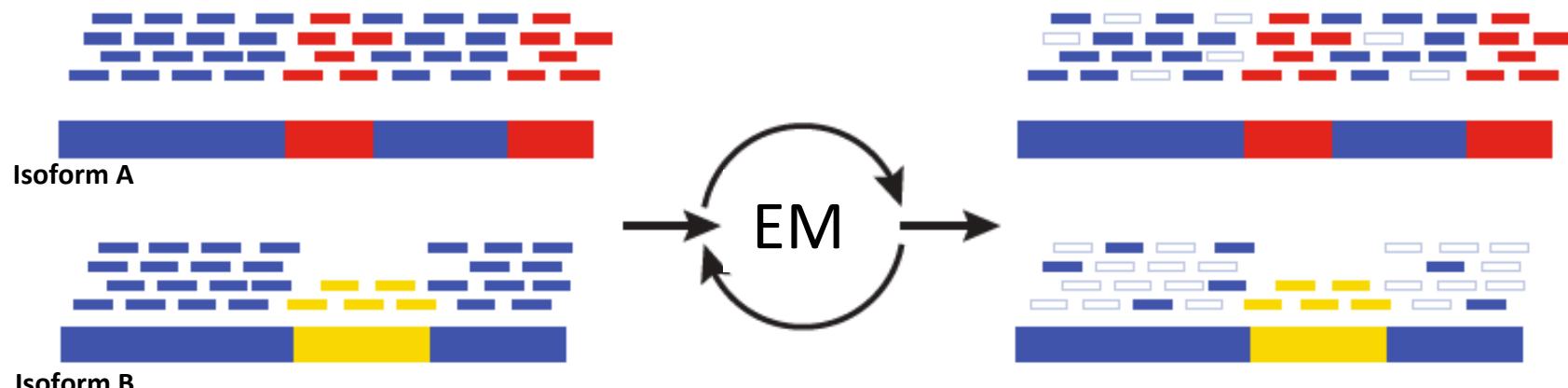
FPKM

Multiply-mapped Reads Confound Abundance Estimation



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

Multiply-mapped Reads Confound Abundance Estimation



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

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

Performed by:

- Cufflinks and Cuffdiff (Tuxedo)
- RSEM
- eXpress

Differential Expression Analysis Using RNA-Seq

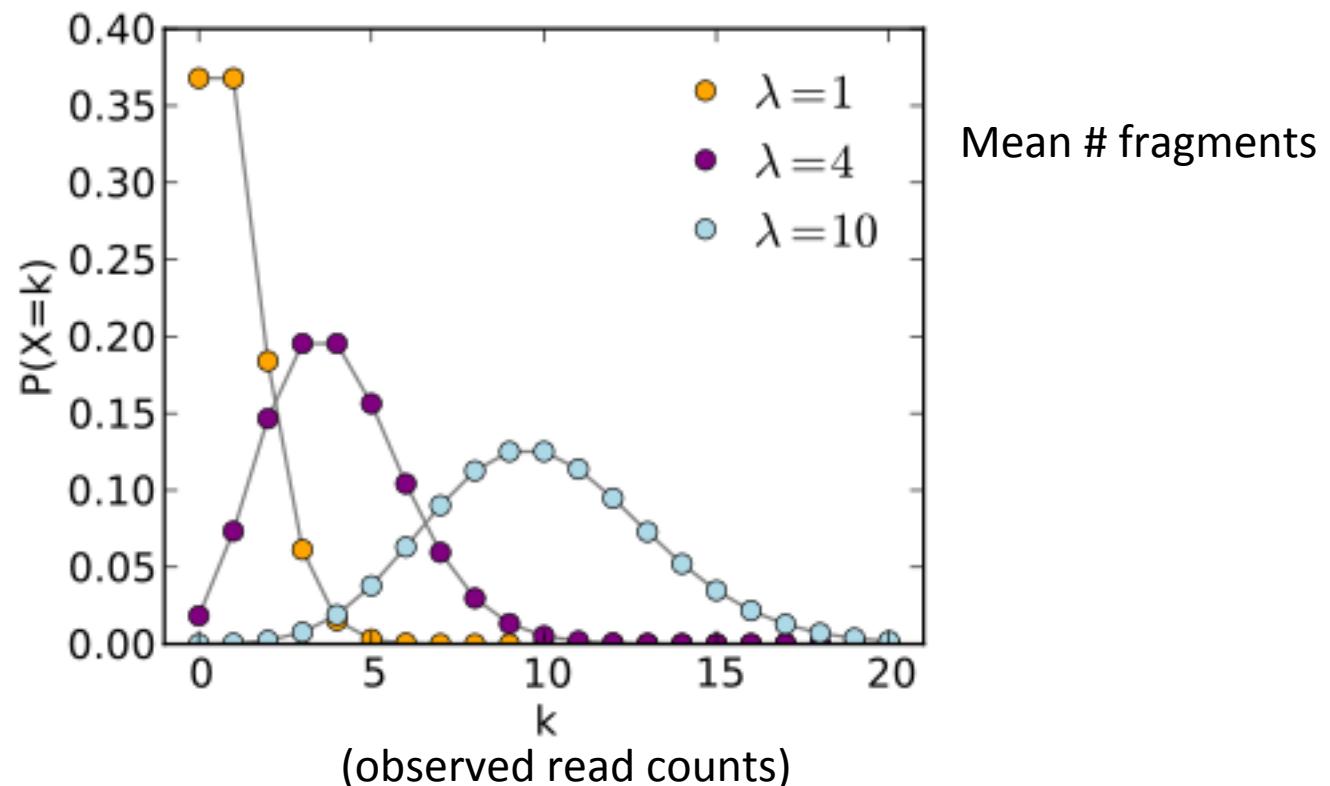
Diff. Expression Analysis Involves

- 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

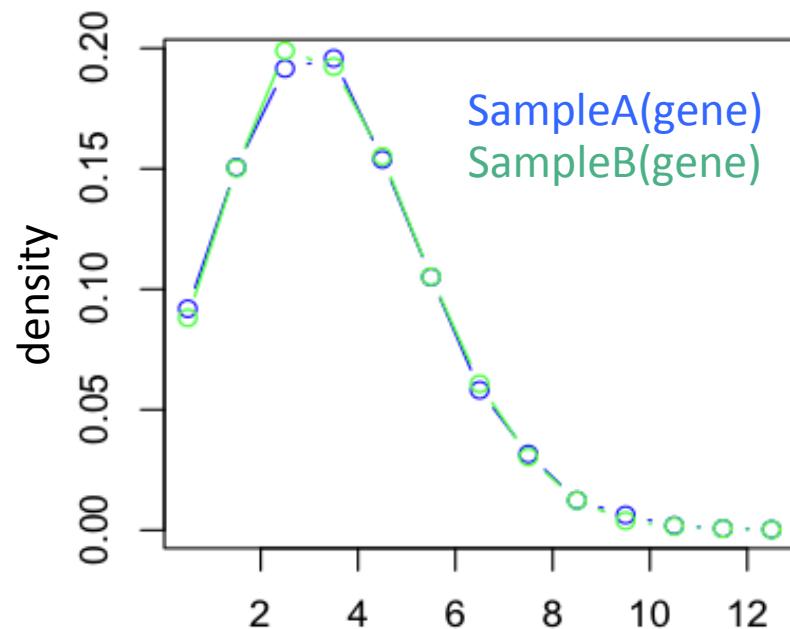


See: http://en.wikipedia.org/wiki/Poisson_distribution

Example: One gene*not* differentially expressed

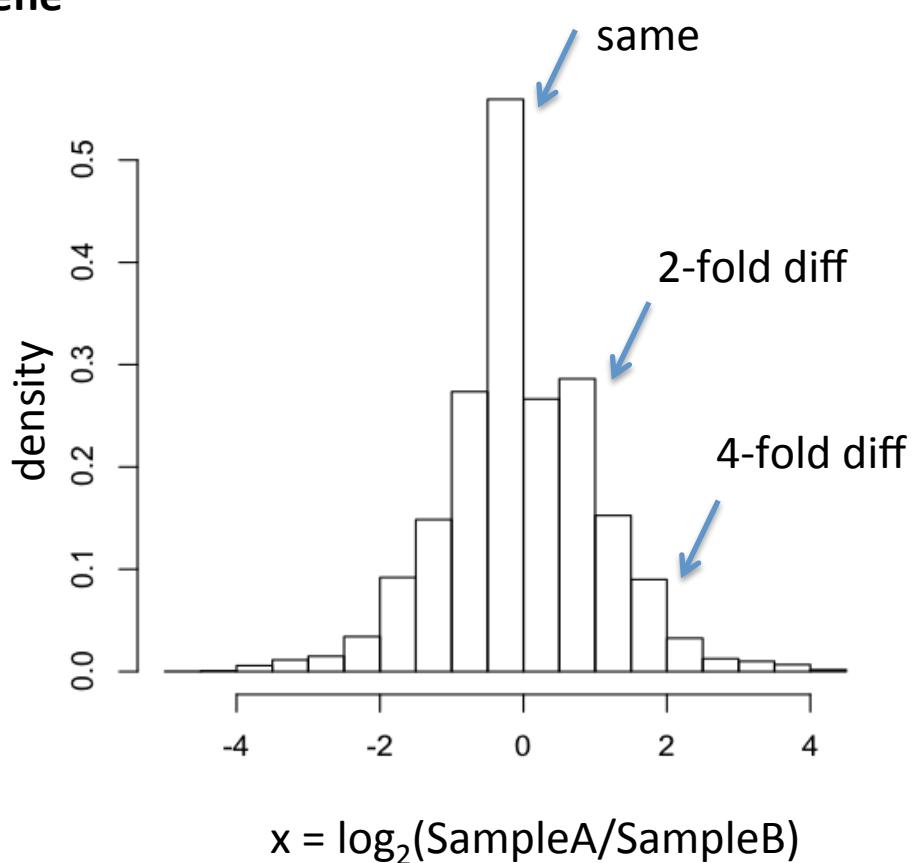
SampleA(gene) = SampleB(gene) = 4 reads

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



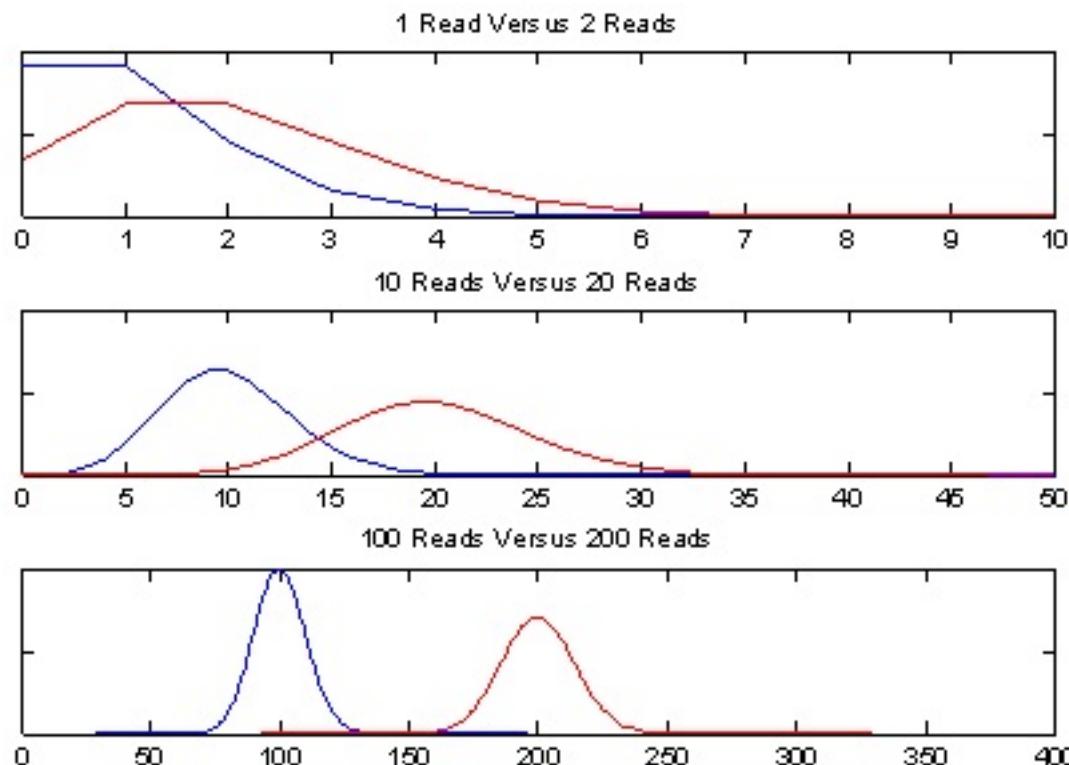
(k) number of reads observed

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.

	Sample A	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



ShrinkSeq
NoiSeq
baySeq
Vsf
Voom
SAMseq
TSPM
DESeq
EBSeq
NBPSeq
edgeR

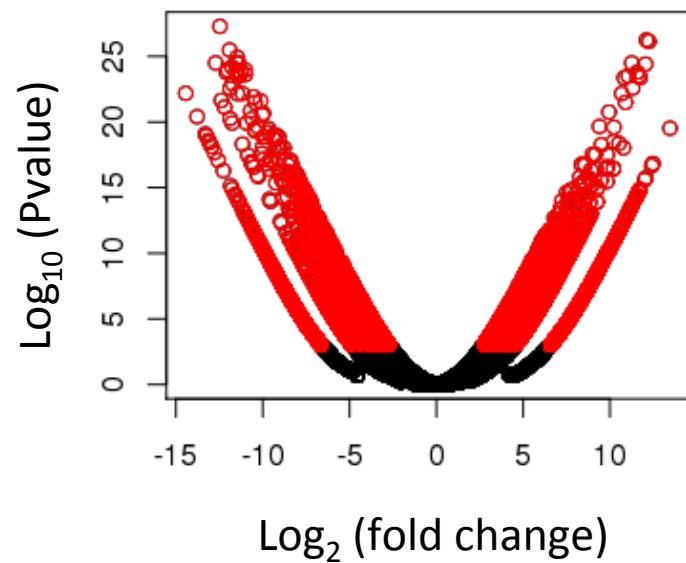
+ other (not-R)
including CuffDiff

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

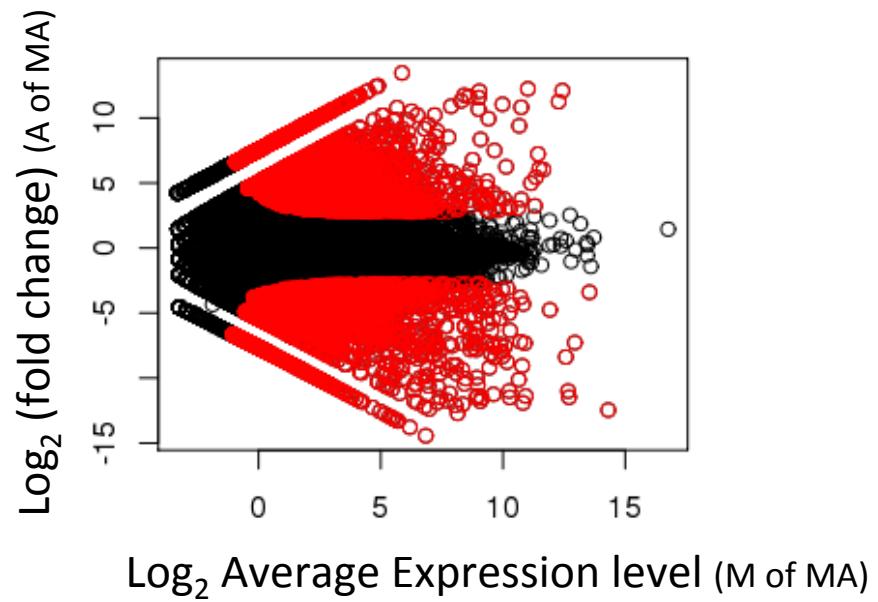
Visualization of DE results and Expression Profiling

Plotting Pairwise Differential Expression Data

Volcano plot
(fold change vs. significance)

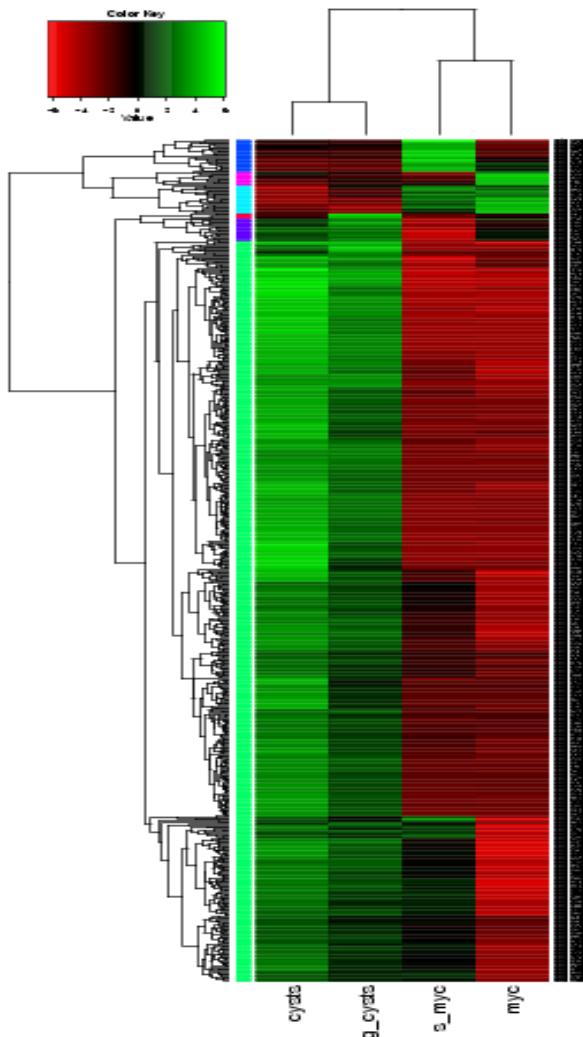


MA plot
(abundance vs. fold change)



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

Comparing Multiple Samples

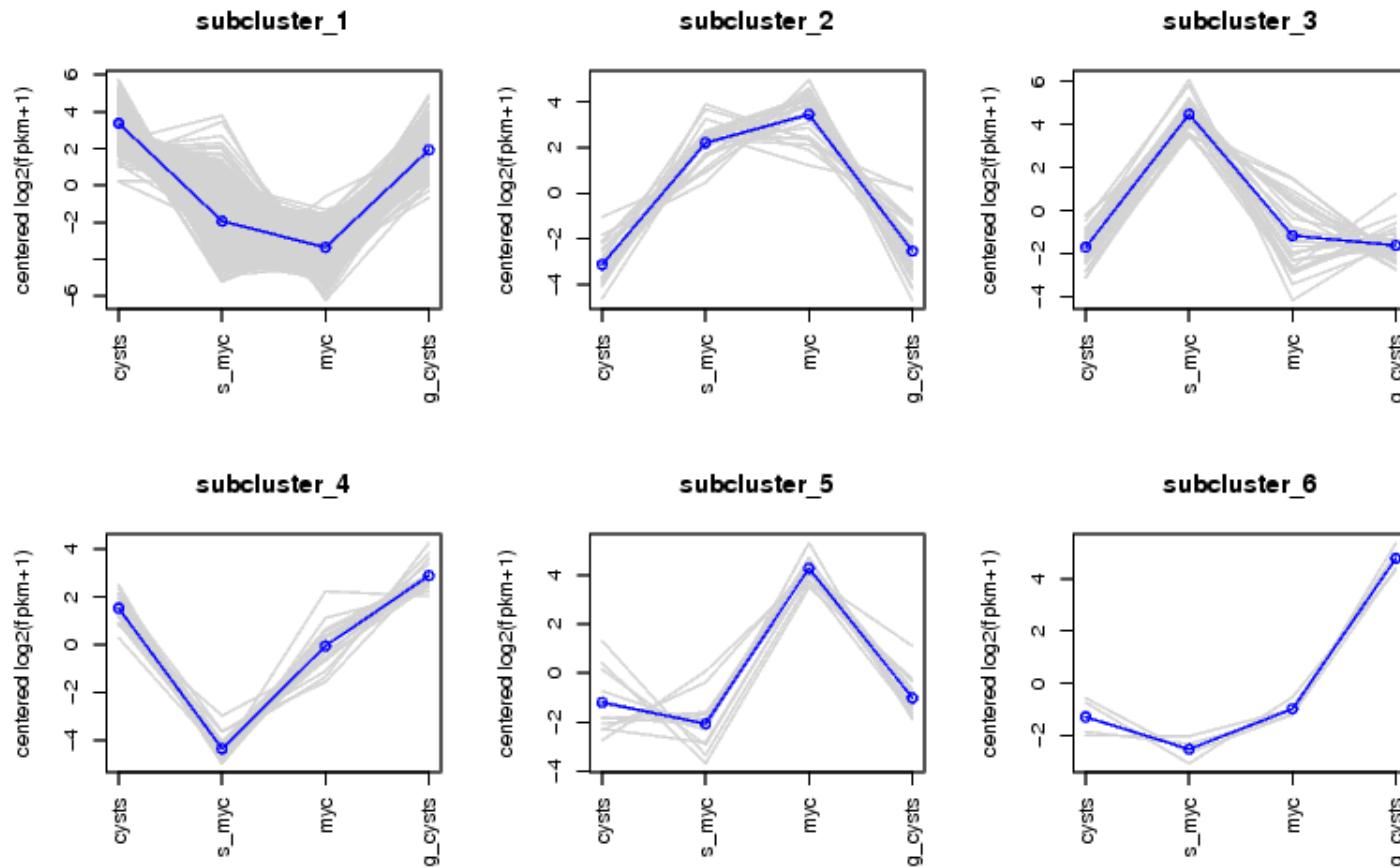


Heatmaps provide an effective tool for navigating differential expression across multiple samples.

Clustering can be performed across both axes:
-cluster transcripts with similar expression patterns.
-cluster samples according to similar expression values among transcripts.

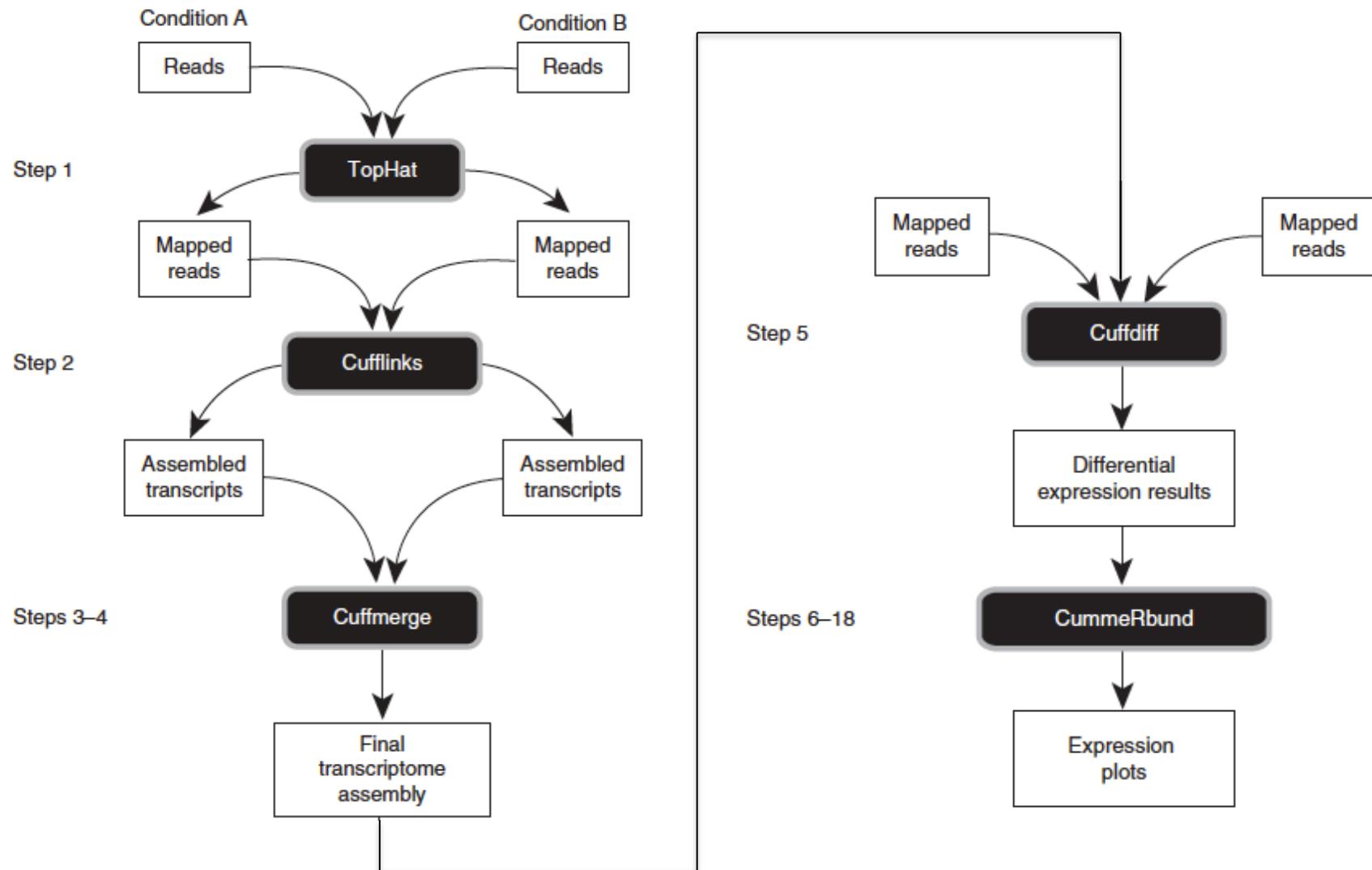
Examining Patterns of Expression Across Samples

Can extract clusters of transcripts and examine them separately.



RNA-Seq Analysis Frameworks

Tuxedo Framework for Transcriptome Analysis

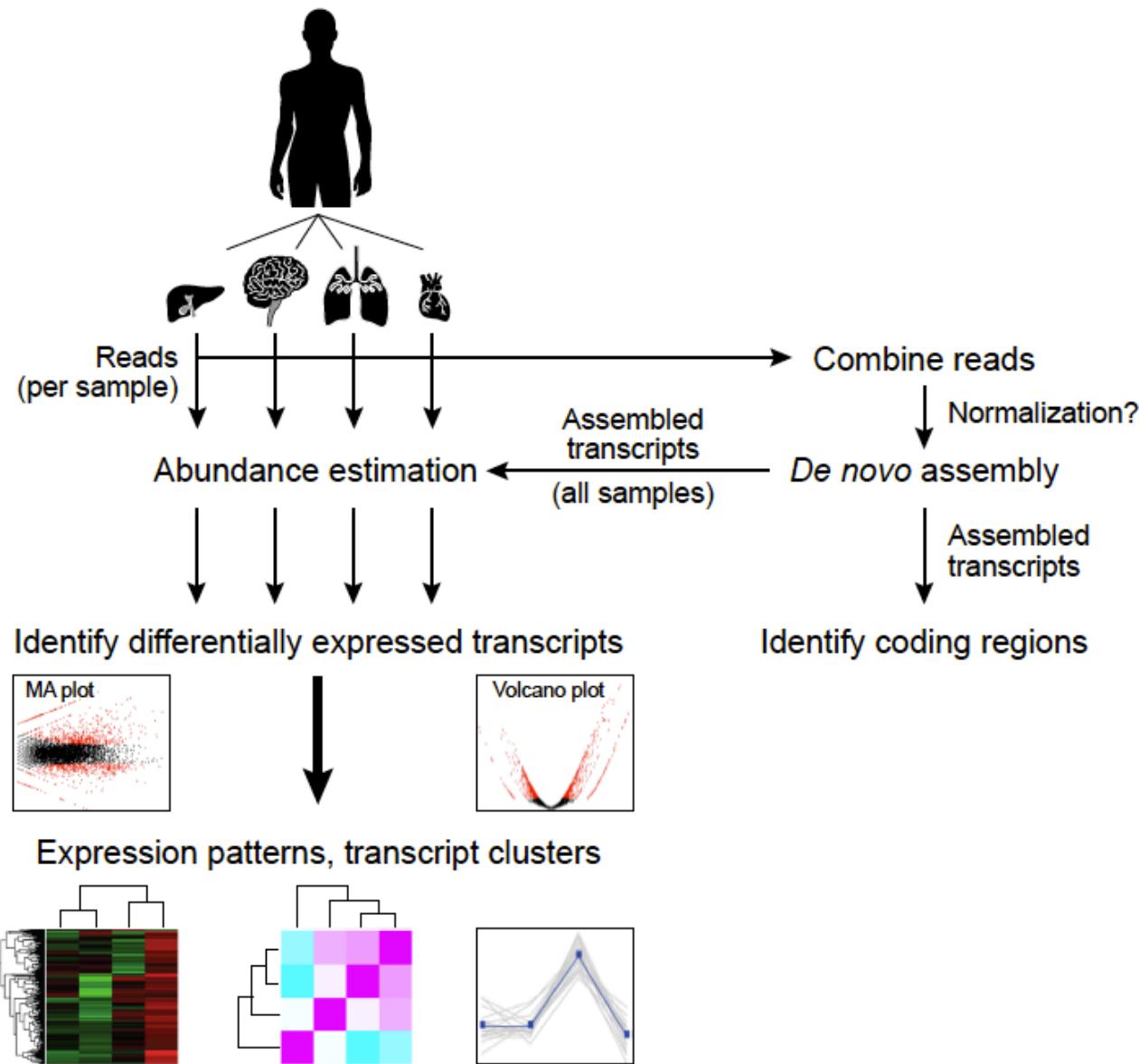


Derived from: Nat Protoc. 2012 Mar 1;7(3):562-78. doi: 10.1038/nprot.2012.016.

Full Tuxedo Framework Demo

- See: Tuxedo_workshop_activities.pdf

Trinity Framework for Transcriptome Analysis



Full Trinity Framework Demo

- See Trinity_workshop_activities.pdf

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.

Software Links

- Tuxedo
 - Tophat: <http://tophat.cbcb.umd.edu/>
 - Cufflinks: <http://cufflinks.cbcb.umd.edu/>
- Trinity
 - <http://trinityrnaseq.sourceforge.net/>
- Visualization
 - IGV: <http://www.broadinstitute.org/igv/>
 - Genomeview: <http://genomeview.org>
- GMAP
 - <http://research-pub.gene.com/gmap/>
- Samtools
 - <http://samtools.sourceforge.net/>

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/>