

RNA-Seq Analysis Workshop

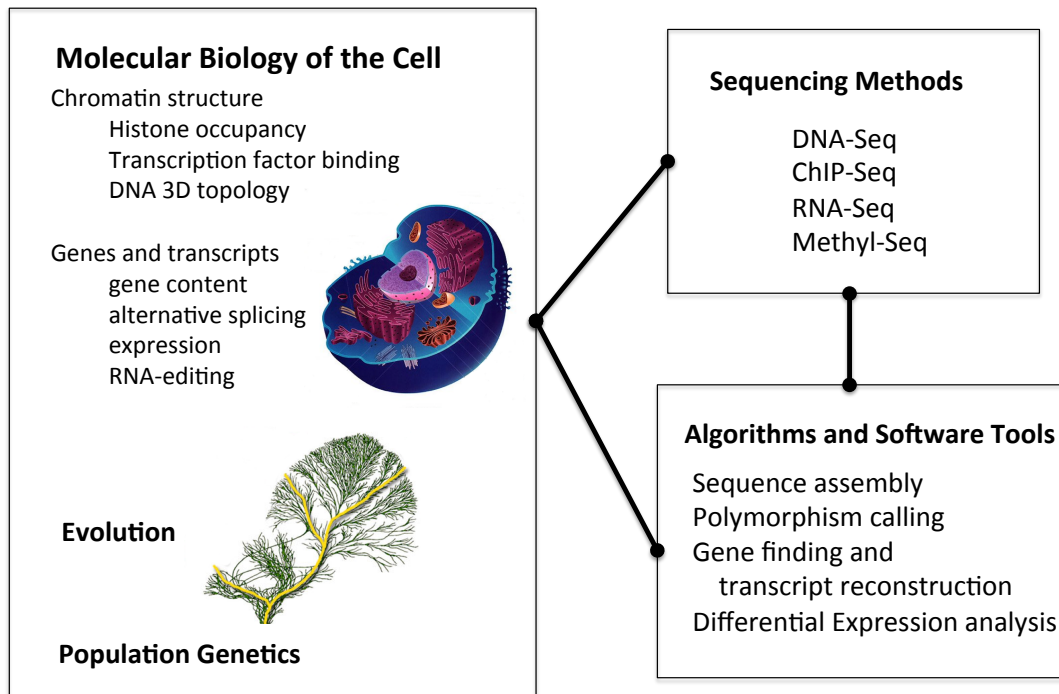
Tuxedo and Trinity for Next-Gen Transcriptome Studies

CSHL 2012-10

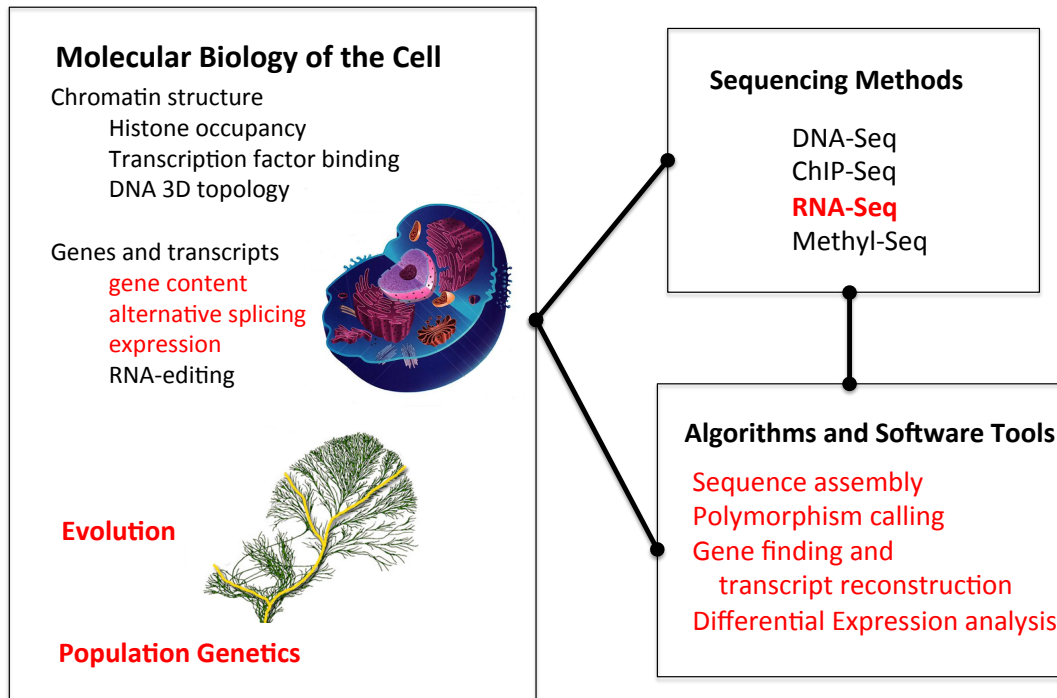
Brian Haas

Broad Institute

Next-gen Sequencing Transforming Modern Science



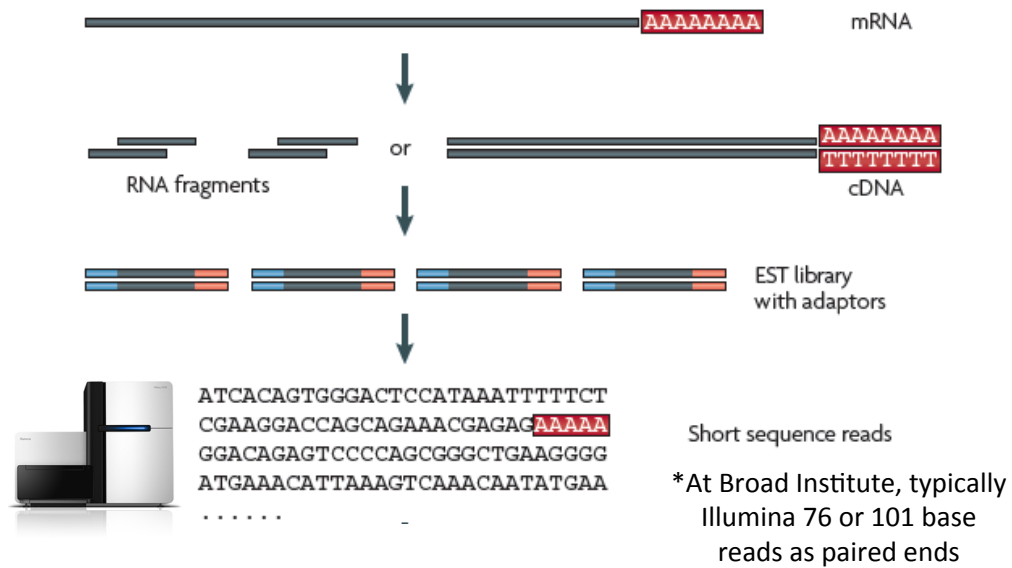
RNA-Seq as a Driving Technology



Outline

- RNA-Seq basics
- Analysis paradigms
- Genome-based rna-seq studies
- Data formats and visualization
- De novo transcriptome-based rna-seq studies
- Transcript Quantification
- Differential Expression Analysis

RNA-Sequencing Methodology



*Adapted from Wang, Gerstein, and Snyder, Nature Reviews Genetics, 2009

Common Data Formats for RNA-Seq

FastA format:

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

FastQ format:

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

AsciiEncodedQual = $-10 * \log_{10}(P_{\text{wrong}}) + 30$

↑
Ascii ('C') = 64

So, $P_{\text{wrong}}('C') = 10^{((64-30)/(-10))} = 10^{-3.4} = \mathbf{0.0004}$

Transcript Reconstruction from RNA-Seq Reads



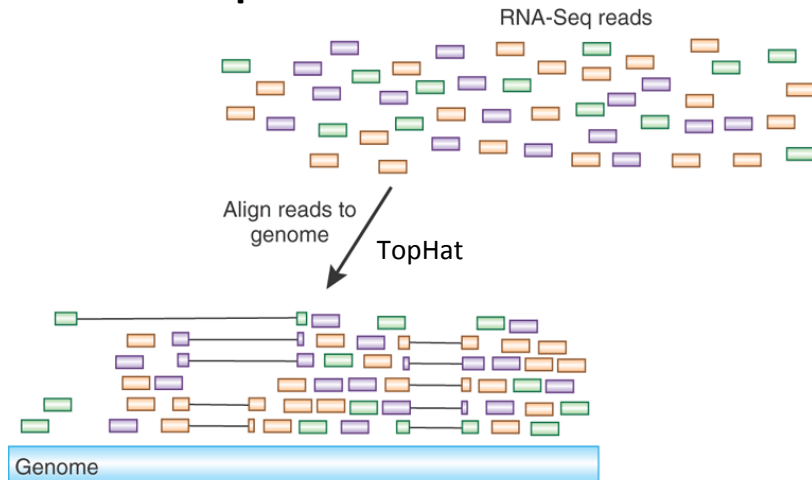
Advancing RNA-Seq analysis

Brian J Haas & Michael C Zody

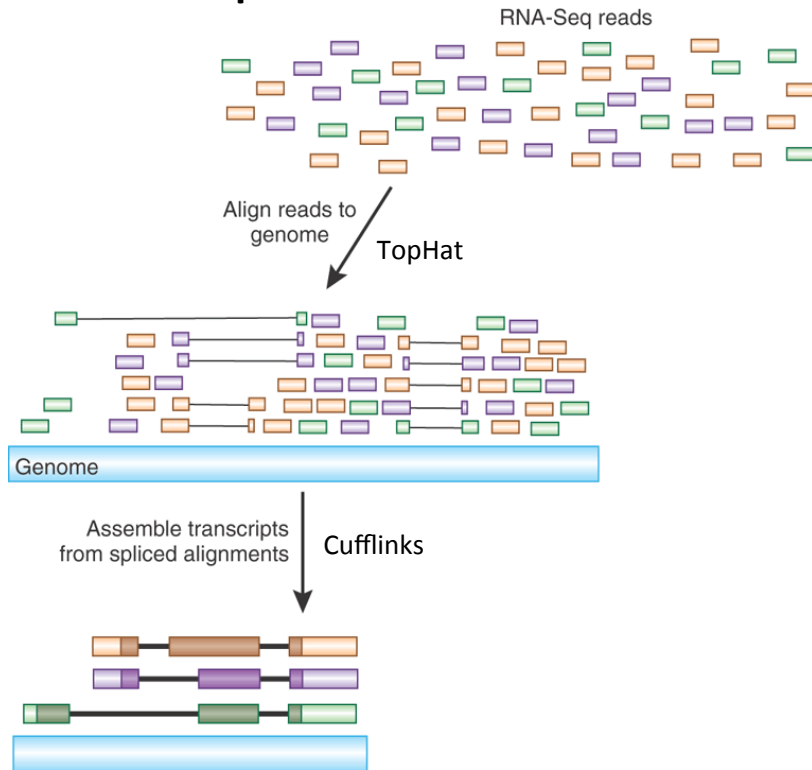
Nature Biotech, 2010

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

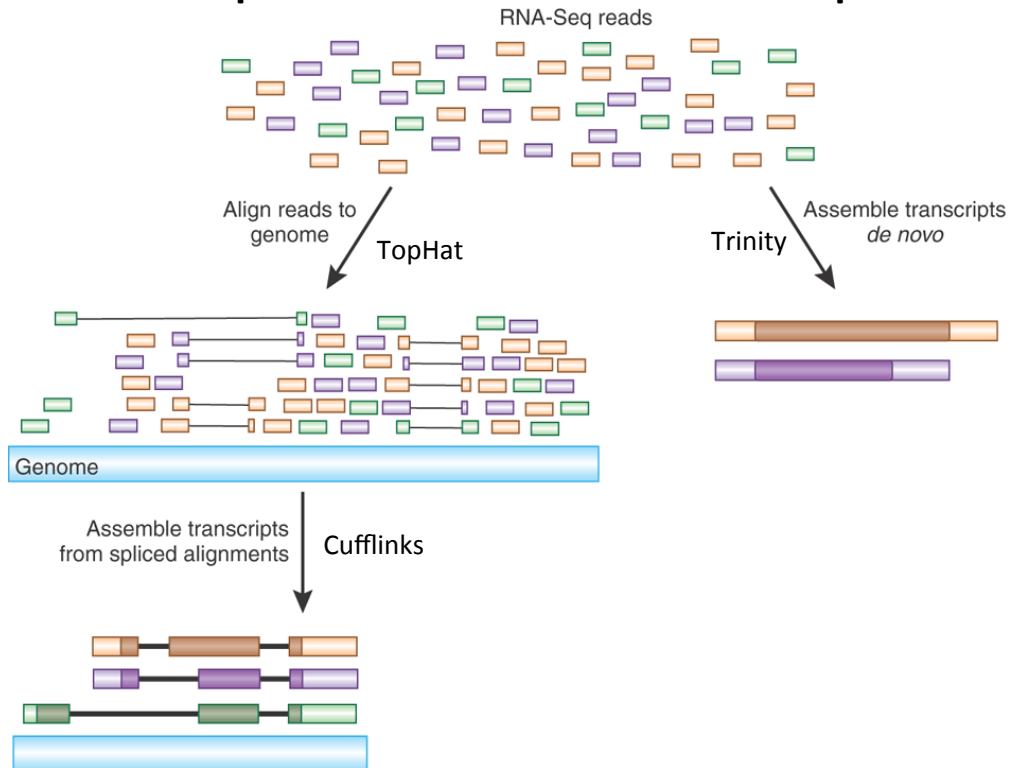
Transcript Reconstruction from RNA-Seq Reads



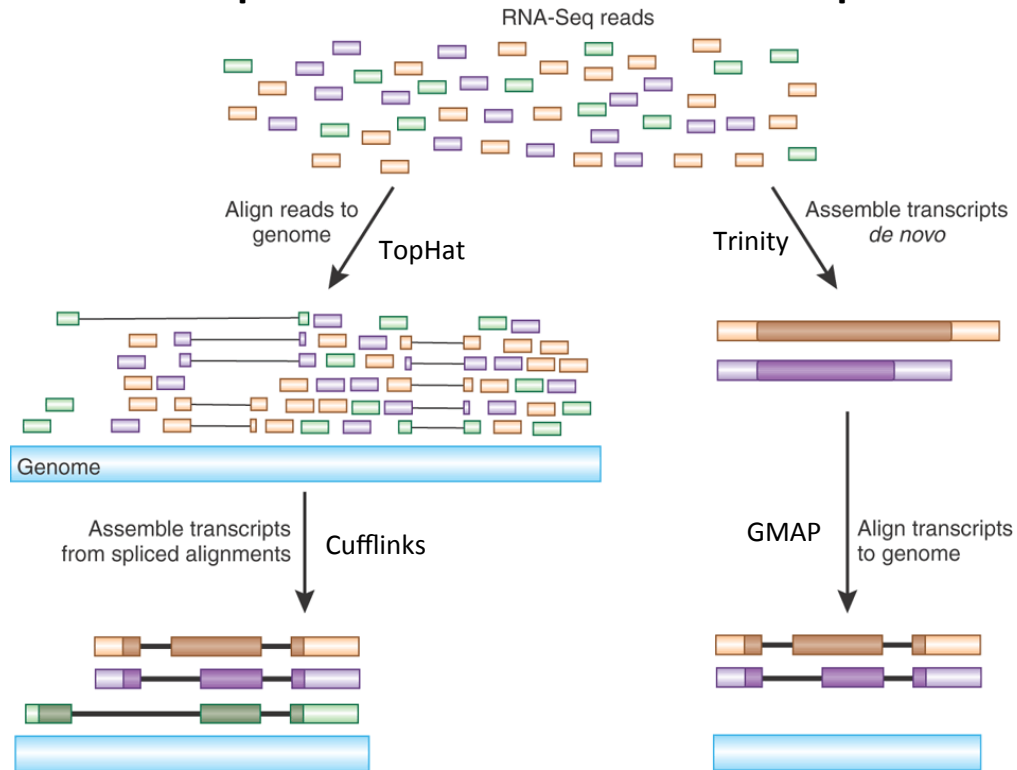
Transcript Reconstruction from RNA-Seq Reads



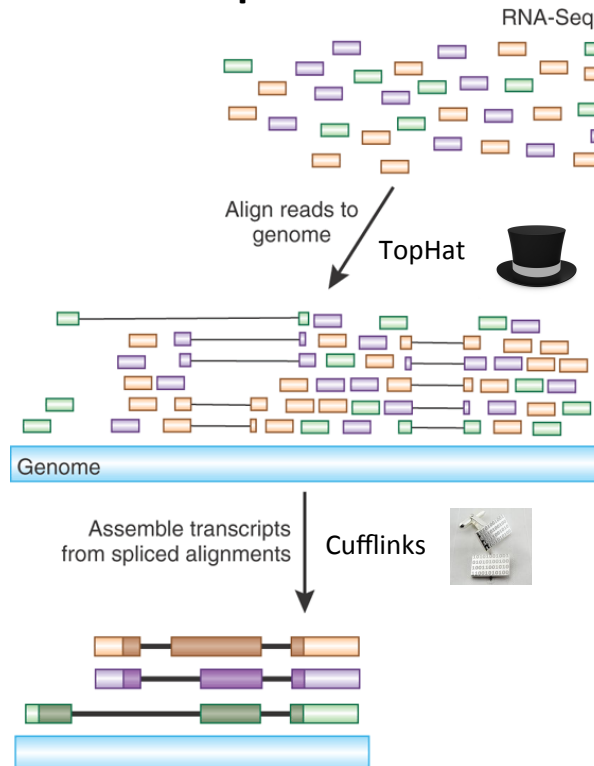
Transcript Reconstruction from RNA-Seq Reads



Transcript Reconstruction from RNA-Seq Reads



Transcript Reconstruction from RNA-Seq Reads



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

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)

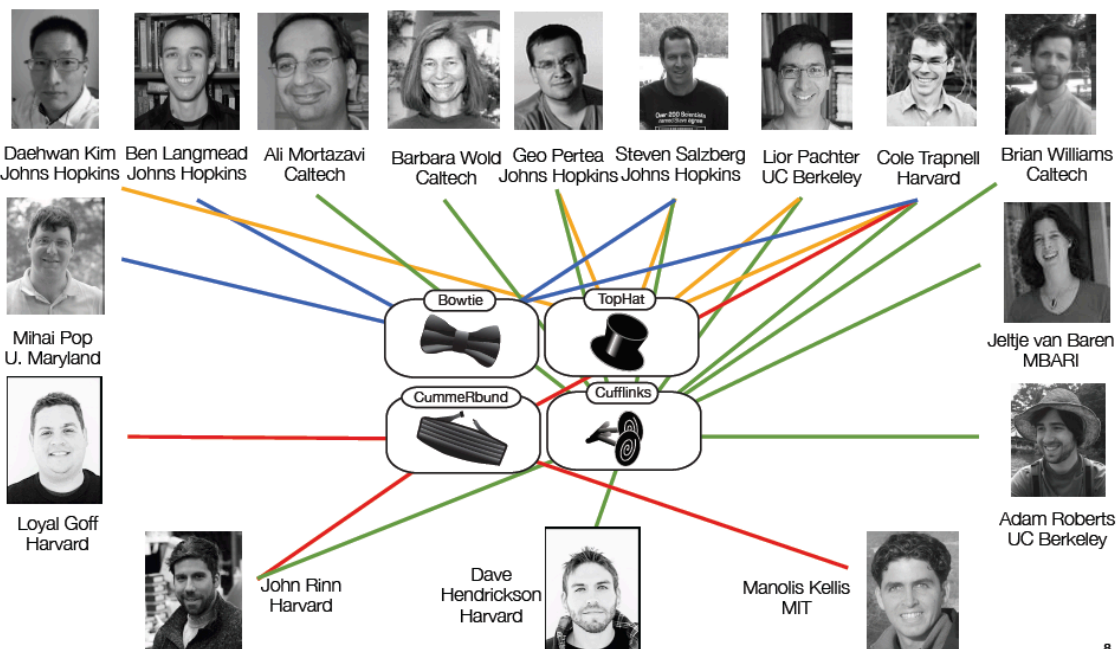


CummeRbund (visualization & analysis)

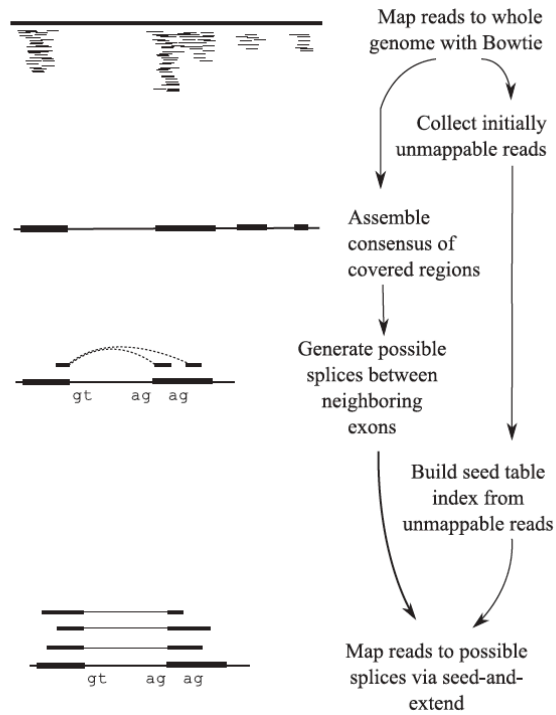


Slide courtesy of Cole Trapnell

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      CCCAAACAAGCCGAAGCTGATTTGGCTCGTAAAGACCCGAAA
10     ###CB?=ADDBCBCDEEFFDEFFDEFFGDBEFGEDGCFGFGGGG
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
```

[Link to SAM format description](#)

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      CCCAAACAAGCCGAAGTAGCTGATTTGGCTCGTAAAGACCCGAAA
10     ###CB?=ADDBCBCDEEFFDEFFDEFFGDBEFGEFGCFGFGGGGG
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
```

[Link to SAM format description](#)

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
```

[Link to SAM format description](#)

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

```
911 921 931 941 951 961 971 981 991 1001 1011 1021 1031 1041 1051 1061 1071
GTAGTTTAATTGATCTCTAATTTAGAATCTGGCAATCAAGCCCTCTGGAAGTTGGCAATCTATAACTCAACCTCTGCTCTGAGATTCTAGTACCTTAGATGCAAGTACATTACTATAATGGTGTATCGGGTCTTCCAACTCTCATTCAAGACTTAATGACTCTGT
GT GTTTAATTCATCTCTAATTTAGAATCTGGCAATCAAGCCCTCTGGAAGTTGGCAATCTATAACTCAAC ctgctctgagattctaaagcttagatgccaagtaacattactataatgggtgtatcgggtctcc ctctccatcaagacttaattgactctgt
GT ATTGATCTCTAATTTAGAATCTGGCAATCAAGCCCTCTGGAAGTTGGCAATCTATAACTCAAC GCTCTGAGATTCTAAAGTACCTTAGATGCAAGTACATTACTATAATGGTGTATCGGGTCTTCCAA ctctccatcaagacttaattgactctgt
GT atttcaattctcaattagaatctggcaatcaagccctctggaagttggcaatctataactcaac GCTCTGAGATTCTAAAGTACCTTAGATGCAAGTACATTACTATAATGGTGTATCGGGTCTTCCAA ctctccatcaagacttaattgactctgt
GTAGTTTAAT aactctgcaatcaagccctctggaagttggcaatctataactcaacctctgctctgagattctta CTTAGATGCAAGTACATTACTATAATGGTGTATCGGGTCTTCCAACTCTCATTCAAGACTTAA ctgt
GTAGTTTAATTT ctctgcaatcaagccctctggaagttggcaatctataactcaacctctgctctgagattctaa CTTAGATGCAAGTACATTACTATAATGGTGTATCGGGTCTTCCAACTCTCATTCAAGACTTAA
GTAGTTTAATTTGATCTCT ctctgcaatcaagccctctggaagttggcaatctataactcaacctctgctctgagattctaa TTAGATGCAAGTACATTACTATAATGGTGTATCGGGTCTTCCAACTCTCATTCAAGACTTAA
GTAGTTTAATTTGATCTCT TGGCAATCAAGCCCTCTGGAAGTTGGCAATCTATAACTCAACCTCTGCTCTGAGATTCTAAAGTAC ATGCCAAGTACATTACTATAATGGTGTATCGGGTCTTCCAACTCTCATTCAAGACTTAAATGAC
GTAGTTTAATTTGATCTCTAAT TGGCAATCAAGCCCTCTGGAAGTTGGCAATCTATAACTCAACCTCTGCTCTGAGATTCTAAAGTAC GCAAGTACATTACTATAATGGTGTATCGGGTCTTCCAACTCTCATTCAAGACTTAAATGACTC
gtaggttaattcaactctcaattag TGGCAATCAAGCCCTCTGGAAGTTGGCAATCTATAACTCAACCTCTGCTCTGAGATTCTAAAGTAC CATTACTATAATGGTGTATCGGGTCTTCCAACTCTCATTCAAGACTTAAATGACTCTGT
GTAGTTTAATTTGATCTCTAATTTAG GCAATCAAGCCCTCTGGAAGTTGGCAATCTATAACTCAACCTCTGCTCTGAGATTCTAAAGTAC cacttataatgggtgtatcgggtctcccaactctccatcaagacttaattgactctgt
GTAGTTTAATTTGATCTCTAATTTAG CAATCAAGCCCTCTGGAAGTTGGCAATCTATAACTCAACCTCTGCTCTGAGATTCTAAAGTAC tggatcgggtctcccaactctccatcaagacttaattgactctgt
GTAGTTTAATTTGATCTCTAATTTAG CAATCAAGCCCTCTGGAAGTTGGCAATCTATAACTCAACCTCTGCTCTGAGATTCTAAAGTACCTT gggctctcccaactctccatcaagacttaattgactctgt
GTAGTTTAATTTGATCTCTAATTTAG gcccctcgaagttggcaatctataactcaacctctgctctgagattctaaagcttagatgccc GGCTTCCAACTCTCATTCAAGACTTAAATGACTCTGT
GTAGTTTAATTTGATCTCTAATTTAG CACTCTGGAAGTTGGCAATCTATAACTCAACCTCTGCTCTGAGATTCTAAAGTACCTTAGATGCA ggtcttcccaactctccatcaagacttaattgactctgt
GTAGTTTAATTTGATCTCTAATTTAG ctctcgaagttggcaatctataactcaacctctgctctgagattctaaagcttagatgccc ggtcttcccaactctccatcaagacttaattgactctgt
GTAGTTTAATTTGATCTCTAATTTAG ctctcgaagttggcaatctataactcaacctctgctctgagattctaaagcttagatgccc ggtcttcccaactctccatcaagacttaattgactctgt
GTAGTTTAATTTGATCTCTAATTTAG GAAATGGCAATCTATAACTCAACCTCTGCTCTGAGATTCTAAAGTACCTTAGATGCAAGTACA gttctcccaactctccatcaagacttaattgactctgt
GTAGTTTAATTTGATCTCTAATTTAG AAGTTGGCAATCTATAACTCAACCTCTGCTCTGAGATTCTAAAGTACCTTAGATGCAAGTACTT cttcccaactctccatcaagacttaattgactctgt
gtaggttaattcaactctcaattagaactctggc CAATATCTATAACTCAACCTCTGCTCTGAGATTCTAAAGTACCTTAGATGCAAGTACTATAAA ctcccaactctccatcaagacttaattgactctgt
GTAGTTTAATTTGATCTCTAATTTAGACTTCTAATTTAGAATCT CTATAACTCAACCTCTGCTCTGAGATTCTAAAGTACCTTAGATGCAAGTACTATAATGGTGT CTTCCAACTCTCATTCAAGACTTAAATGACTCTGT
GTAGTTTAATTTGATCTCTAATTTAGACTTCTGCA CAATGGCAATCTATAACTCAACCTCTGCTCTGAGATTCTAAAGTACCTTAGATGCAAGTACA gttctcccaactctccatcaagacttaattgactctgt
gtaggttaattcaactctcaattagaactctggc ctctcgaagttggcaatctataactcaacctctgctctgagattctaaagcttagatgccc tctcccaactctccatcaagacttaattgactctgt
GTAGTTTAATTTGATCTCTAATTTAGACTTCTGCAAGCC ctctcgaagttggcaatctataactcaacctctgctctgagattctaaagcttagatgccc tctcccaactctccatcaagacttaattgactctgt
GTAGTTTAATTTGATCTCTAATTTAGACTTCTGCAAG GCAAGTGGCAATCTATAACTCAACCTCTGCTCTGAGATTCTAAAGTACCTTAGATGCAAGTACA gttctcccaactctccatcaagacttaattgactctgt
GTAGTTTAATTTGATCTCTAATTTAGACTTCTGCAAG GCAAGTGGCAATCTATAACTCAACCTCTGCTCTGAGATTCTAAAGTACCTTAGATGCAAGTACA gttctcccaactctccatcaagacttaattgactctgt
ATTGATCTCTAATTTAGAATCTGGCAATCAAGCCCTCTGGAAGTTGGCAATCTATAACTCAAC TTGATCTCTAATTTAGAATCTGGCAATCAAGCCCTCTGGAAGTTGGCAATCTATAACTCAACCT
aagtaaccttagatgccaagtaacattactataatgggtgtatcgggtcttcccaactctccatcaag actcccaactctccatcaagacttaattgactctgt
TCCAACTCTCATTCAAGACTTAAATGACTCTGT TCCAACTCTCATTCAAGACTTAAATGACTCTGT
TCCAACTCTCATTCAAGACTTAAATGACTCTGT caactctccatcaagacttaattgactctgt
caactctccatcaagacttaattgactctgt
aactctccatcaagacttaattgactctgt
aactctccatcaagacttaattgactctgt
tccattcaagacttaattgactctgt
ccattcaagacttaattgactctgt
cattcaagacttaattgactctgt
```

IGV

www.broadinstitute.org/igv/

Integrative Genomics Viewer

Home

Integrative Genomics Viewer

What's New

- July 3, 2012. Soybean (Glycine max) and Rat (m5) genomes have been updated.
- April 20, 2012. IGV 2.1 has been released. See the [release notes](#) for more details.
- April 19, 2012. See our new [IGV paper](#) in Briefings in Bioinformatics.

Citing IGV

To cite your use of IGV in your publication:

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

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

© 2012 Broad Institute

GenomeView

The screenshot shows the GenomeView website homepage. At the top, there is a navigation bar with links for Demos, Plug-ins, JAnnot API, Join mailing list, Support - Frequently asked questions, and Cite us. The main content area is divided into several sections:

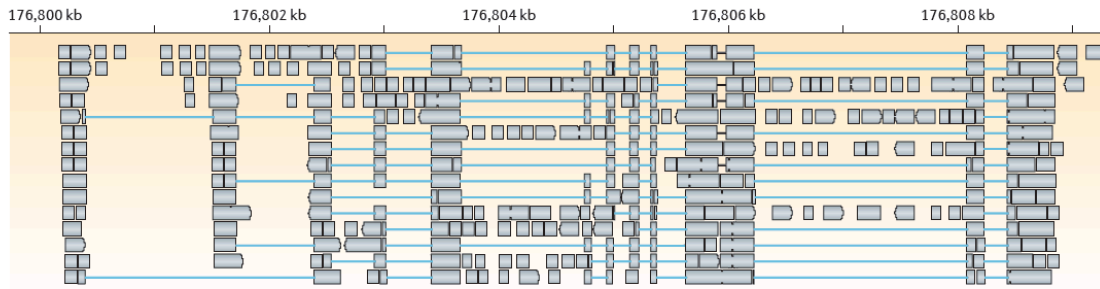
- Start Now!**: Includes a "Webstart:" section with a "Launch" button and a "High-mem webstart" section with another "Launch" button. Below this is an "Applet:" section with a "Launch" button.
- Documentation**: A list of links including "Quick start guide", "Manual", "Advanced manual", and "Tutorials".
- Navigation**: A list of links including "Download", "Demos", and "Plug-ins".
- Getting started**: A section with a clock icon and text: "Get started with a five minute quick-start guide that will get up and running in no time".
- Web start**: A section with a "Launch" button and text: "Click the launch button to start GenomeView".
- Download**: A section with a folder icon and text: "Download the current release. You can also start GenomeView".
- Support**: A section with text: "If you experience any issues, head over to the support section, we like to help you."
- Recent questions**: A list of questions such as "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?", and "How do I integrate GenomeView in my".
- Awards**: A circular badge for "Most Creative Visualization iDEA Challenge 2011 Academic" and a banner for "Most creative visualization award @ Illumina iDEA challenge 2011" with a "KILLER APP AWARD" graphic.

GenomeView: viewing TopHat alignments

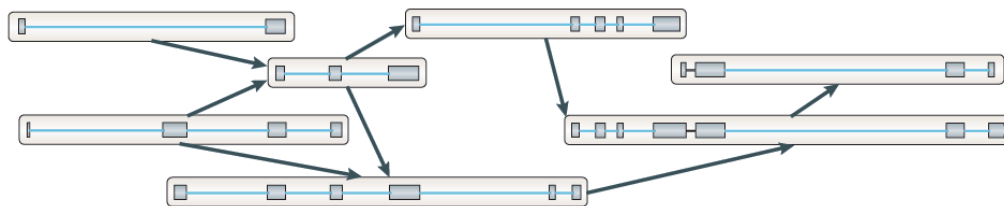
The screenshot shows the GenomeView application interface. The main window displays a genomic track for chromosome 7 (7000000090838467) with a 2.1 Kb region selected. The track shows a gene structure (Gene structure) and TopHat alignments (accepted_hits.bam). The alignments are represented by colored horizontal bars (red, green, blue, purple) indicating read positions. A red warning message "Max stacking depth reached!" is visible at the bottom of the alignment track. The right sidebar contains a "Track list" with "Ruler", "Gene structure", and "accepted_hits.bam". Below the track list, there are "Features" (set to "CDS") and "Details on selected items:" sections. The bottom status bar shows the coordinates "7000000090838467:159289:161368" and a progress indicator "21 / 123 (Mb)".

Transcript Reconstruction Using Cufflinks

a Splice-align reads to the genome



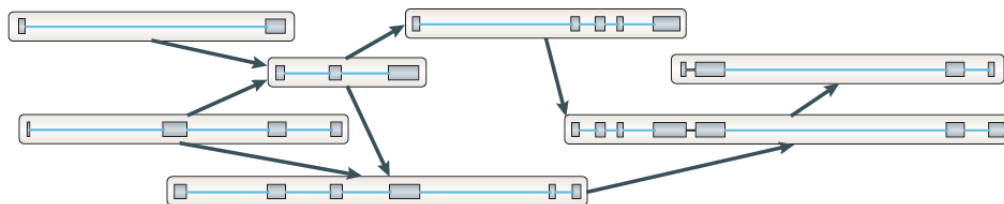
b Build a graph representing alternative splicing events



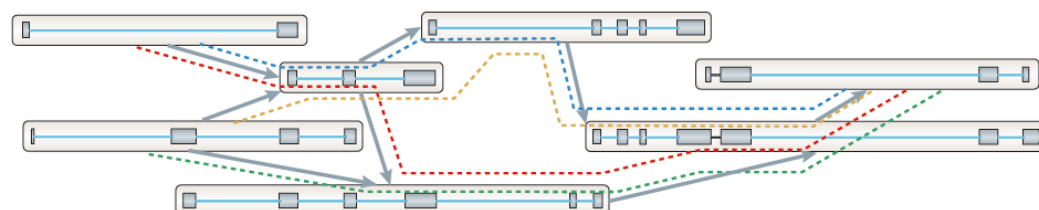
From Martin & Wang. Nature Reviews in Genetics. 2011

Transcript Reconstruction Using Cufflinks

b Build a graph representing alternative splicing events



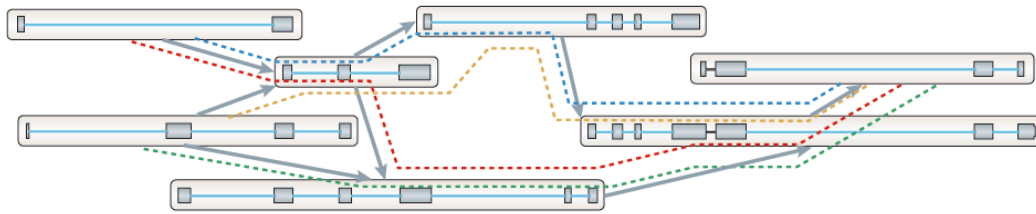
c Traverse the graph to assemble variants



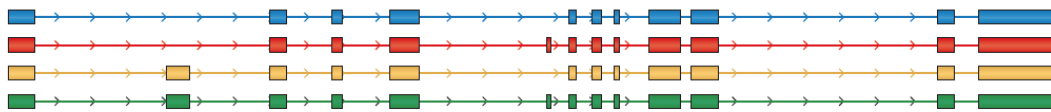
From Martin & Wang. Nature Reviews in Genetics. 2011

Transcript Reconstruction Using Cufflinks

c Traverse the graph to assemble variants



d Assembled isoforms



From Martin & Wang. Nature Reviews in Genetics. 2011

Transcript Structures in **GTF** Format

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

```
0 7000000090838467
1 Cufflinks
2 transcript
3 101
4 5716
5 1000
6 .
7 .
8 gene_id "CUFF.1"; transcript_id "CUFF.1.1"; FPKM "378.0239937260"

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"
```

De novo transcriptome assembly

No genome required

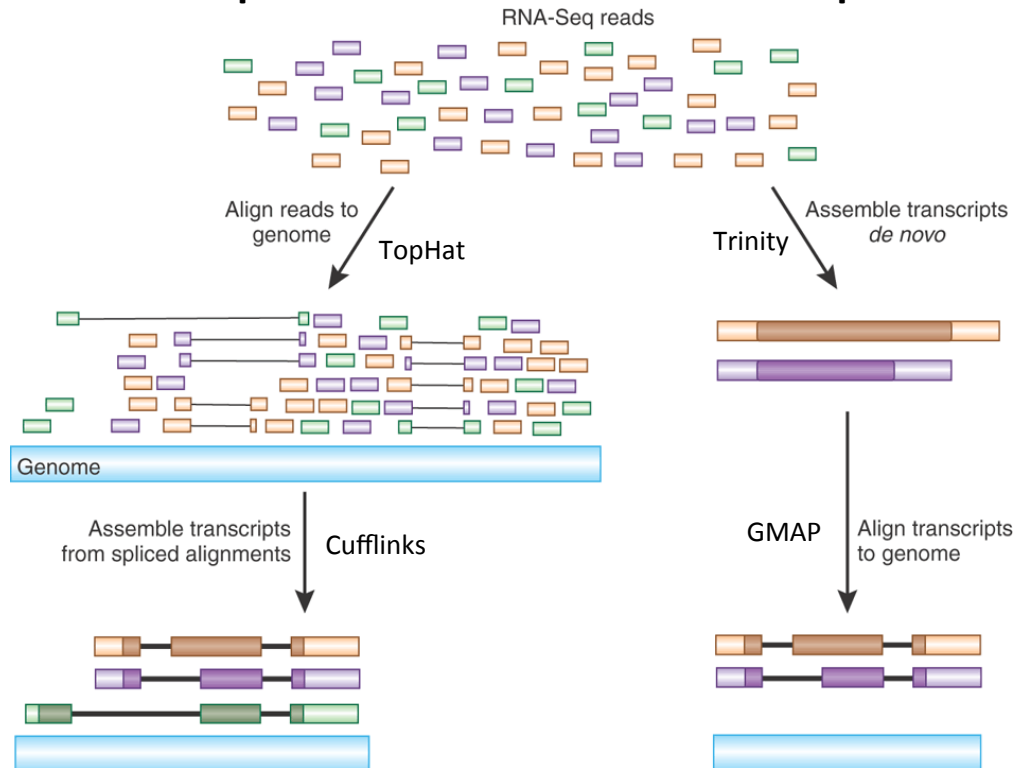
Empower studies of non-model organisms

expressed gene content

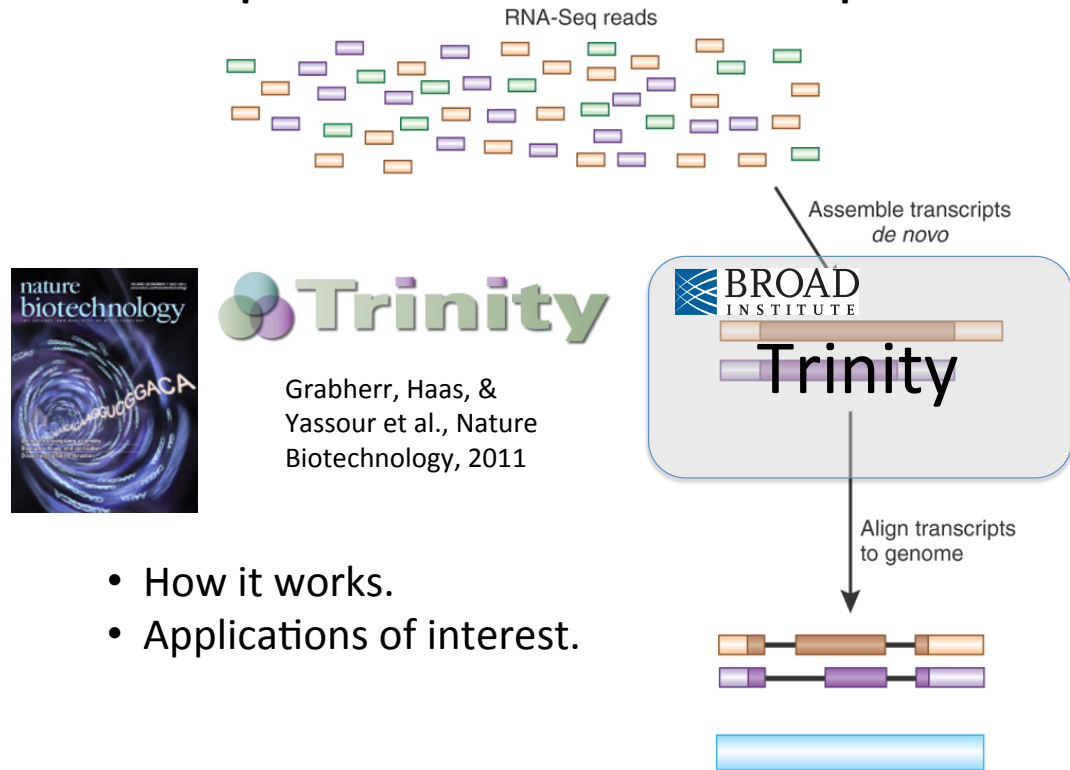
transcript abundance

differential expression

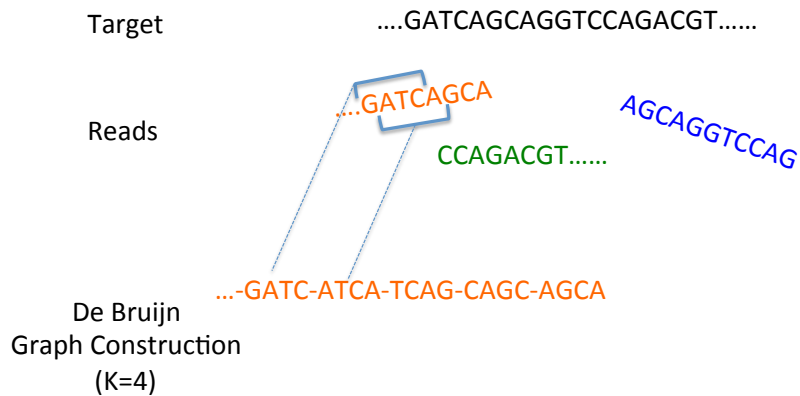
Transcript Reconstruction from RNA-Seq Reads



Transcript Reconstruction from RNA-Seq Reads

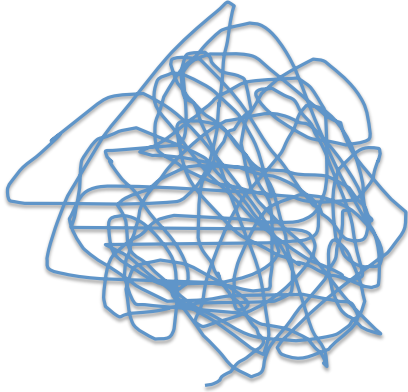


Short Read Assembly Using de Bruijn Graphs



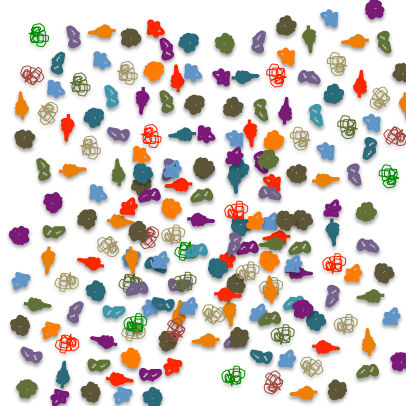
Trinity Aggregates Isolated Transcript Graphs

Genome Assembly Single Massive Graph



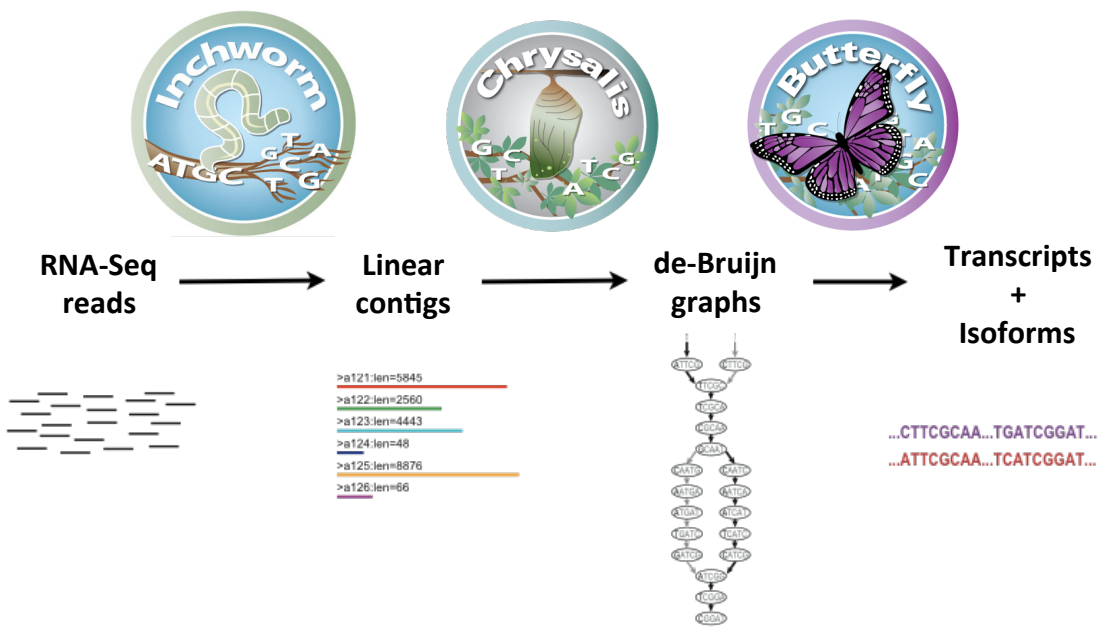
Entire chromosomes represented.

Trinity Transcriptome Assembly Many Thousands of Small Graphs



Ideally, one graph per expressed gene.

Trinity



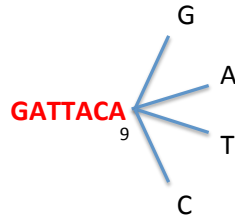


Inchworm Algorithm

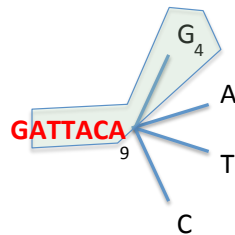
Decompose all reads into overlapping Kmers (25-mers)

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

Extend kmer at 3' end, guided by coverage.

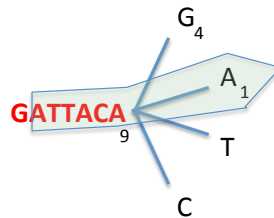


Inchworm Algorithm

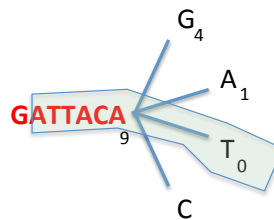




Inchworm Algorithm

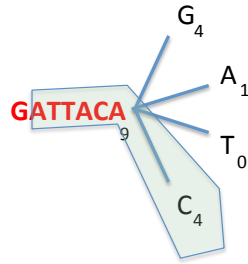


Inchworm Algorithm

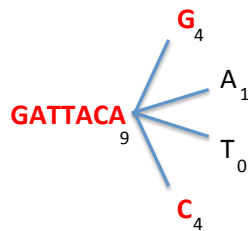




Inchworm Algorithm

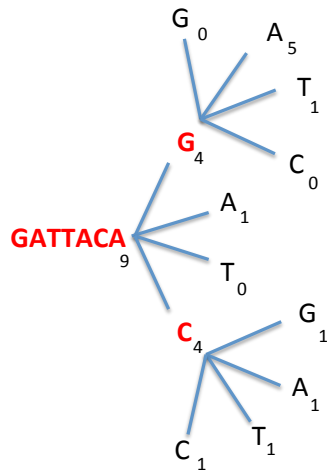


Inchworm Algorithm

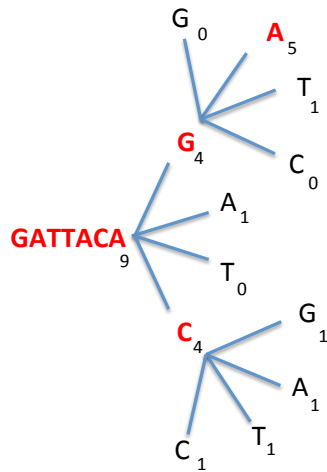




Inchworm Algorithm

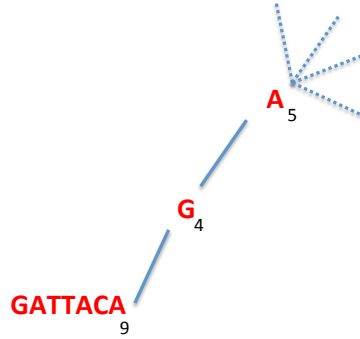


Inchworm Algorithm

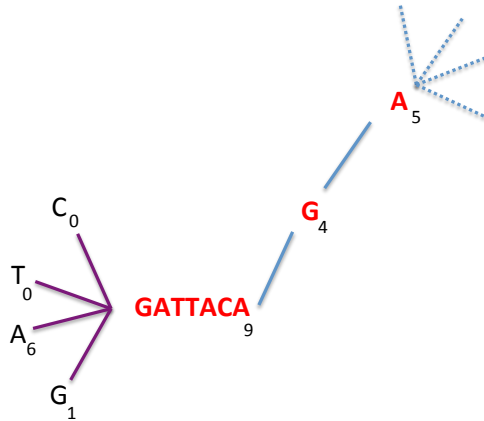




Inchworm Algorithm

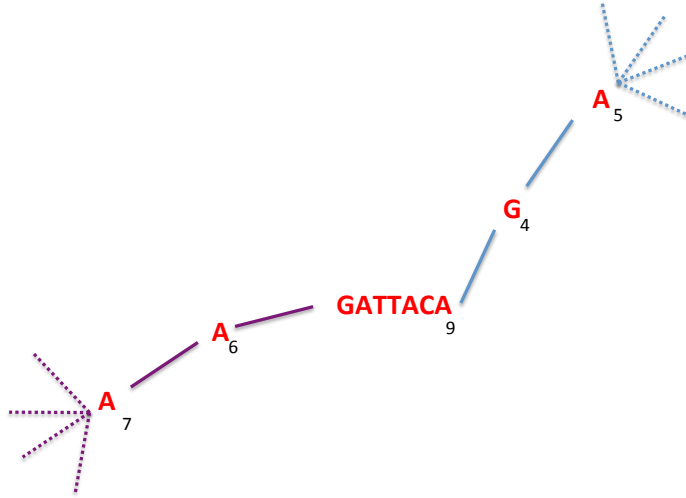


Inchworm Algorithm





Inchworm Algorithm



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

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



Inchworm Contigs from Alt-Spliced Transcripts

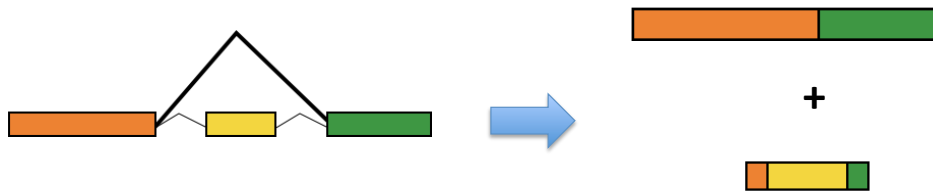




Inchworm Contigs from Alt-Spliced Transcripts

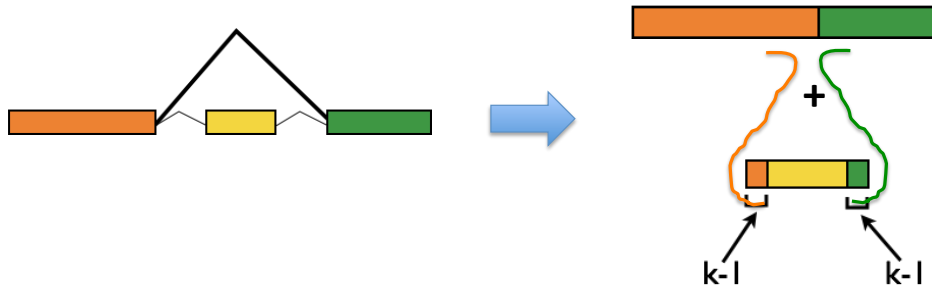


Inchworm Contigs from Alt-Spliced Transcripts





Inchworm Contigs from Alt-Spliced Transcripts



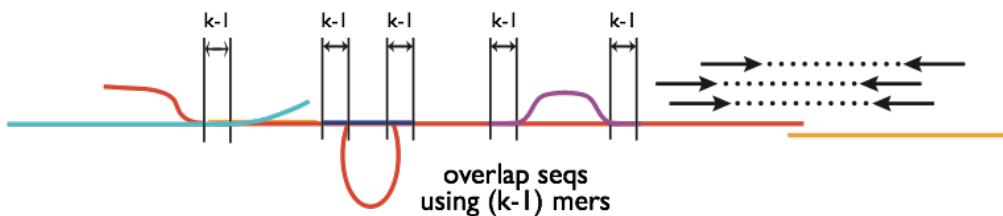
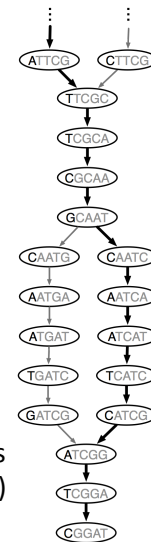
Chrysalis

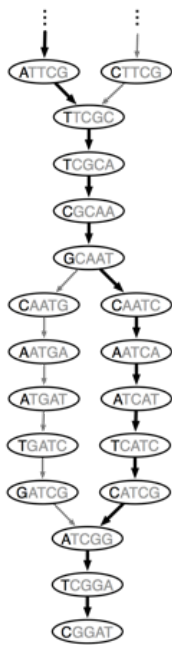
```
>a121:len=5845  
>a122:len=2560  
>a123:len=4443  
>a124:len=48  
>a125:len=8876  
>a126:len=68
```

Integrate isoforms
via $k-1$ overlaps



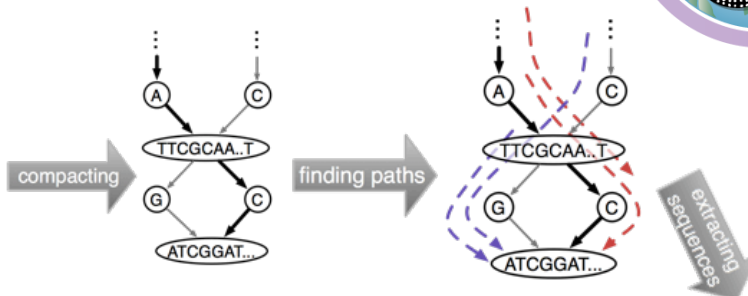
Build de Bruijn Graphs
(ideally, one per gene)



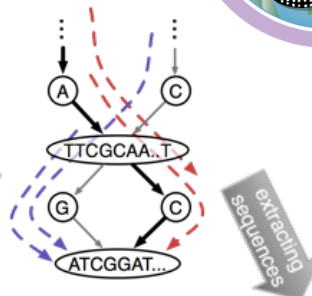


de Bruijn graph

Butterfly



compact graph



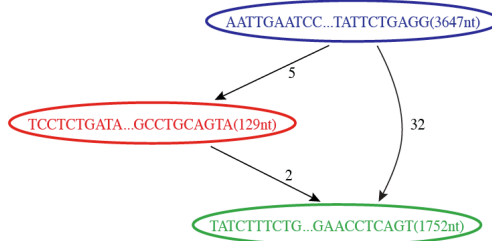
compact graph with reads

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

sequences (isoforms and paralogs)

Reconstruction of Alternatively Spliced Transcripts

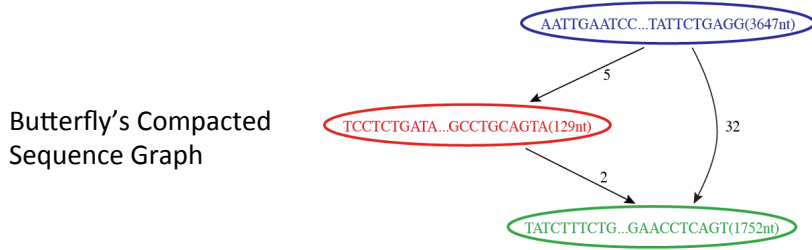
Butterfly's Compacted Sequence Graph



Reconstructed Transcripts



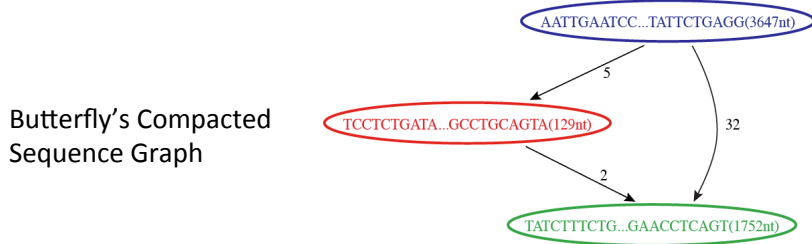
Reconstruction of Alternatively Spliced Transcripts



Reconstructed Transcripts



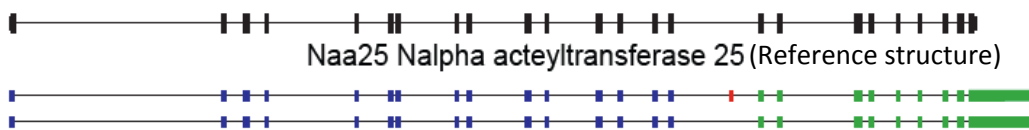
Reconstruction of Alternatively Spliced Transcripts



Reconstructed Transcripts



Aligned to Mouse Genome



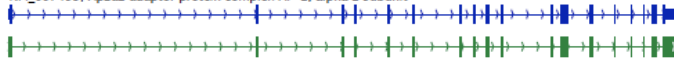
Teasing Apart Transcripts of Paralogous Genes



Teasing Apart Transcripts of Paralogous Genes

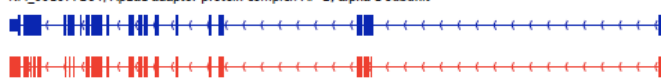
chr7:148,744,197-148,821,437

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

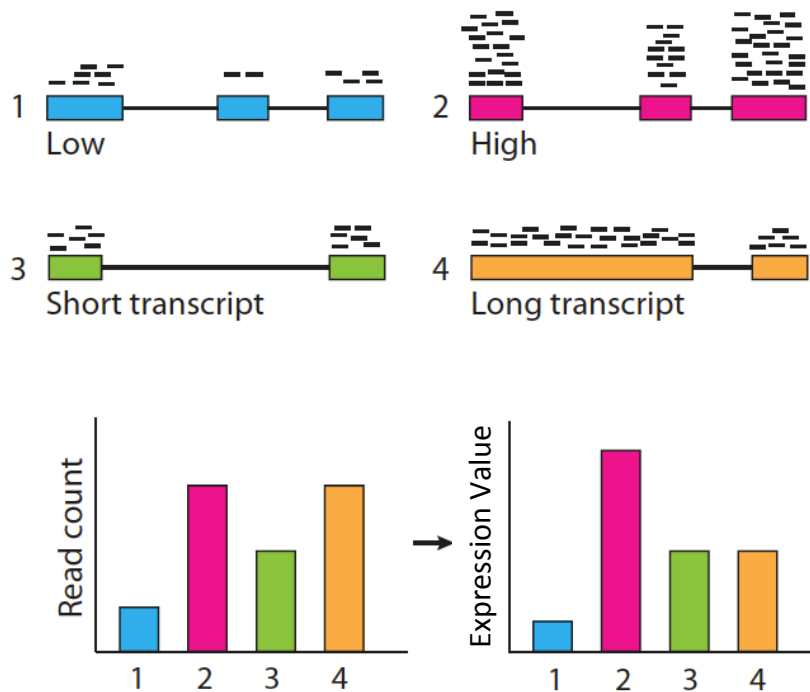


chr7:52,150,889-52,189,508

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



Calculating expression of genes and transcripts



Slide courtesy of Cole Trapnell

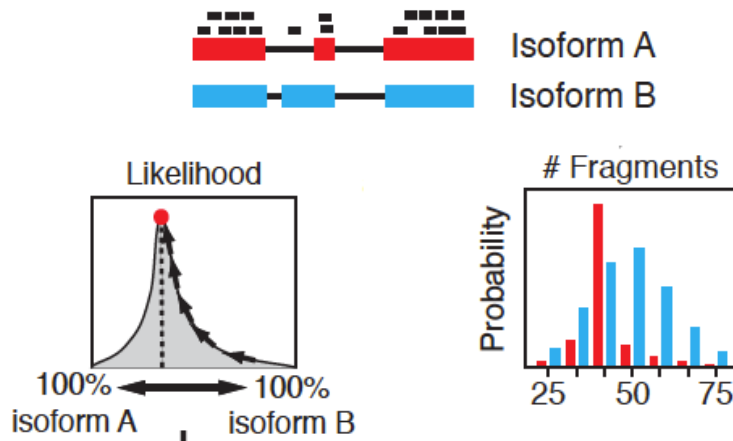
Normalized Expression Values

- Normalized for both length of the transcript and total depth of sequencing.
- Number of RNA-Seq **F**ragments
Per **K**ilobase of transcript
per total **M**illion fragments mapped

FPKM

Note, **RPKM** : **R**eads per ... instead of Fragments is often used with single-end reads.

Sophisticated computations are required to estimate isoform expression where there is read mapping ambiguity.



Model considers unique and ambiguously mapping reads and the length of transcripts.

Illustrations courtesy of Cole Trapnell

Tools that perform abundance estimation

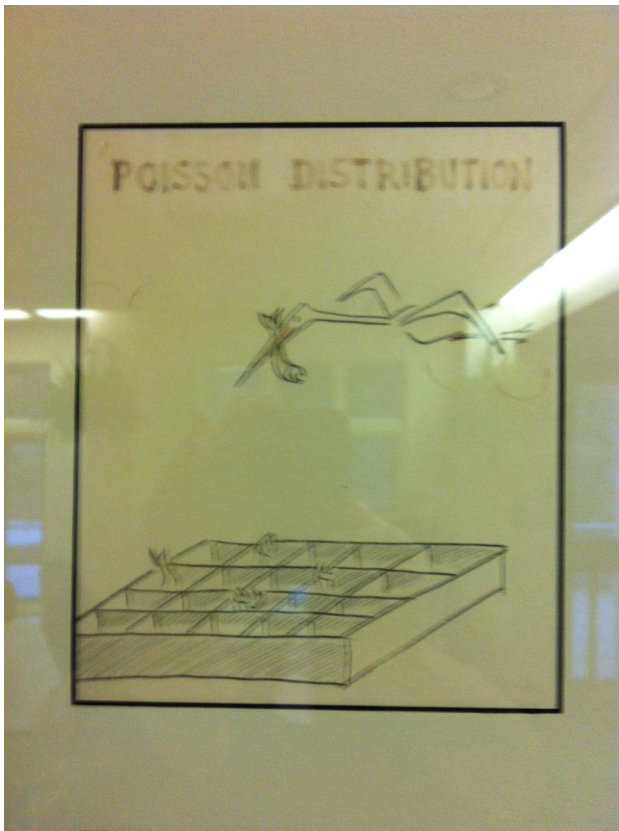
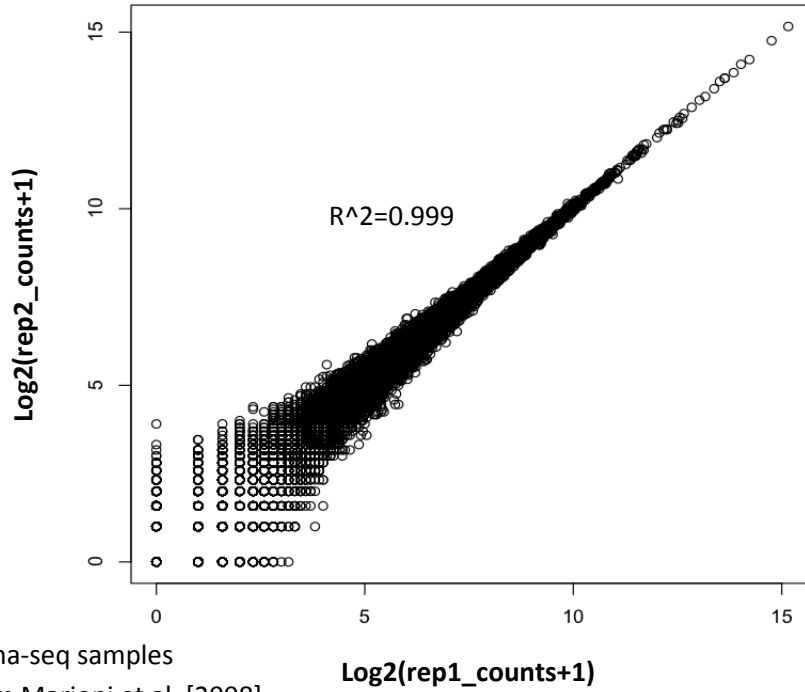
Cuffdiff

0	tracking_id	XLOC_000001
1	class_code	-
2	nearest_ref_id	-
3	gene_id	XLOC_000001
4	gene_short_name	-
5	tss_id	TSS1
6	locus	Chr1:180422-180902
7	length	-
8	coverage	-
9	condA_FPKM	10042.1
10	condA_conf_lo	0
11	condA_conf_hi	20319.6
12	condA_status	OK

RSEM

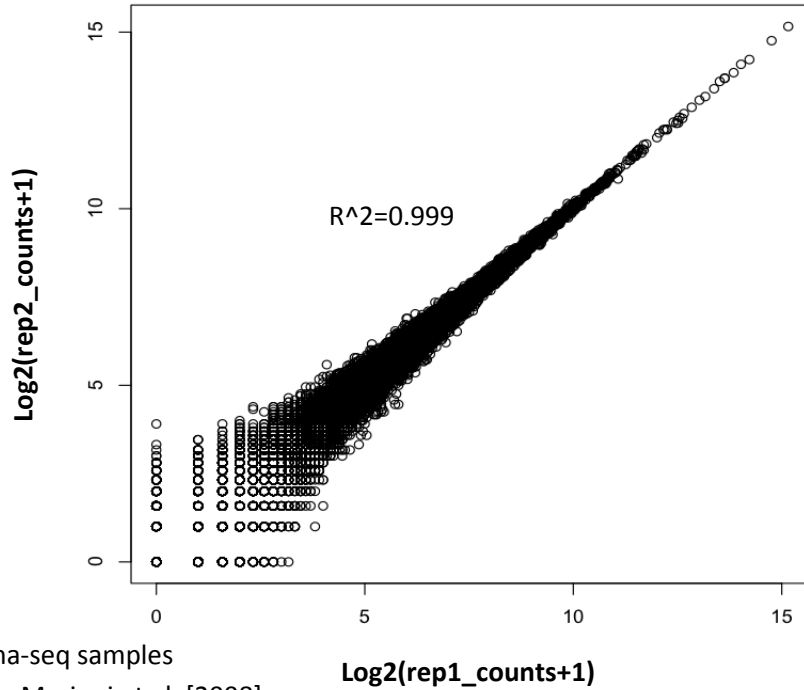
0	transcript_id	comp100_c0_seq1
1	gene_id	comp100_c0
2	length	727
3	effective_length	534.74
4	expected_count	14.00
5	TPM	328.11
6	FPKM	532.77
7	IsoPct	100.00

Technical Replicates



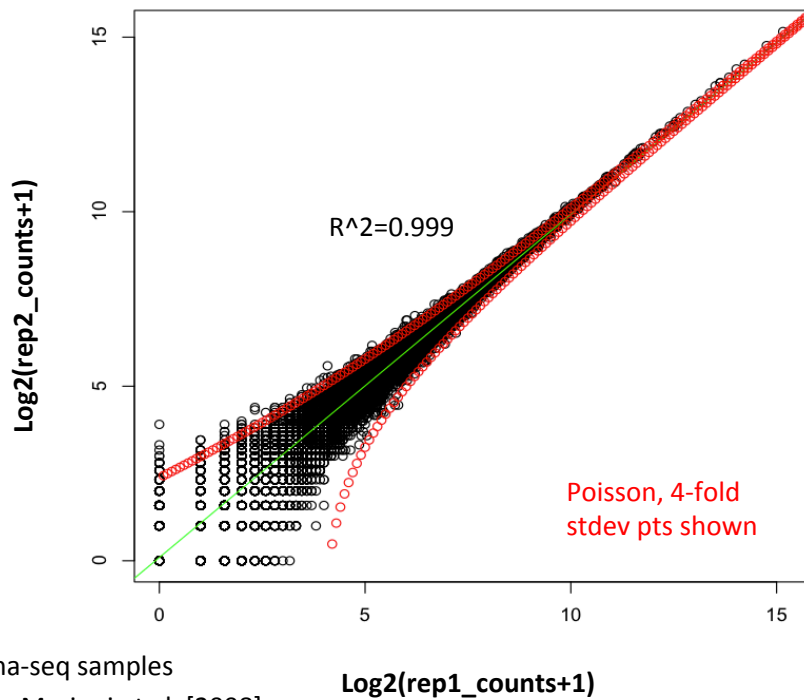
See above the toasters in
Blackford Hall, CSHL

Technical Replicates



Technical Replicates

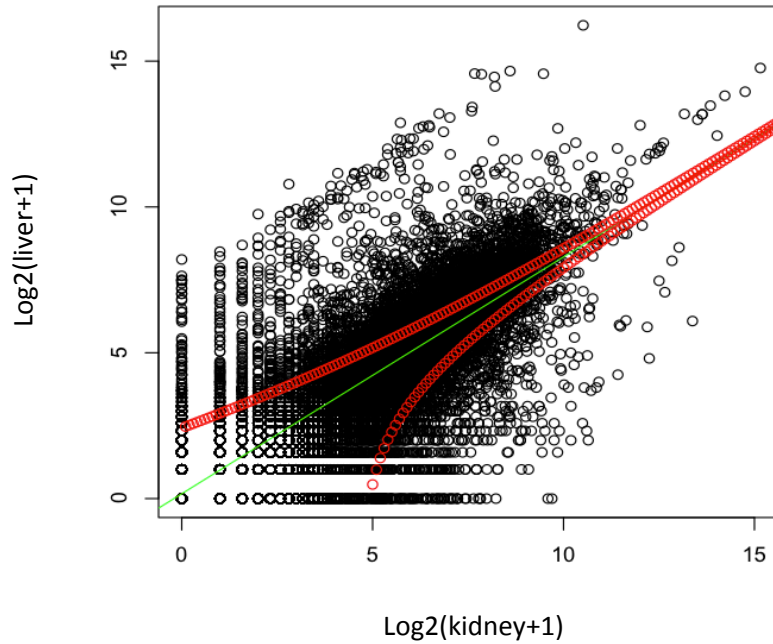
Variation observed matches expectations due to random sampling (Poisson distribution)



- Poisson well-describes variation observed in technical replicates.
- Negative binomial (overly dispersed poisson) better models biological replicates.

Comparing Samples and Identifying Differentially Expressed Transcripts

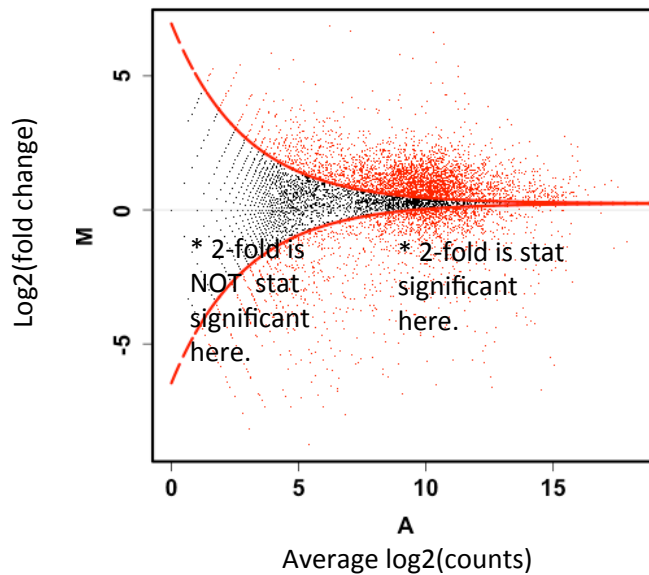
Kidney vs. Liver



Increased Power for Identifying Differentially Expressed Transcripts With Deeper Sequencing

MA plot: $\log(\text{Counts})$ vs. $\log(\text{Fold change})$

Log Phase VS Heat Shock



Normalization Required

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

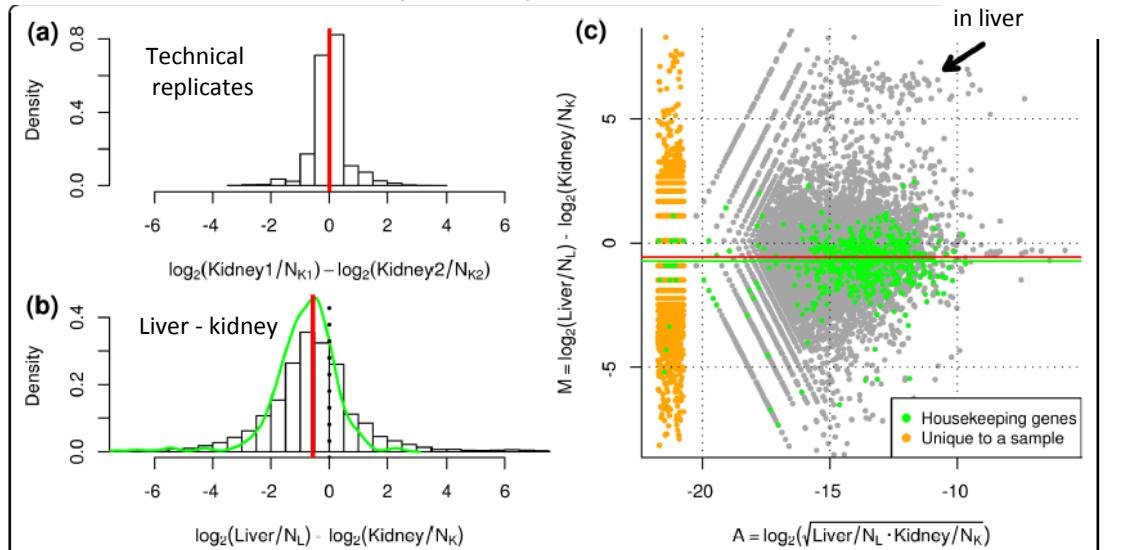


Figure 1 Normalization is required for RNA-seq data. Data from [6] comparing log ratios of (a) technical replicates and (b) liver versus kidney expression levels, after adjusting for the total number of reads in each sample. The green line shows the smoothed distribution of log-fold-changes of the housekeeping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

Robinson and Oshlack, Genome Biology, 2010

Identifying Differentially Expressed Transcripts

- Statistical tests performed on fragment counts (not FPKM values).
- Given observed read counts for a transcript in each of two samples, what's the probability they were derived from the same distribution (null hypothesis)? (ex. Fishers exact test)
If ($P \leq 0.05$), significantly different
- Don't forget to adjust P-values due to false discovery rate (FDR) resulting from running many (thousands of) statistical tests. (ex. use Q-values)

Experimental Design

- Forego technical replicates
- Ideally, have at least 3 biological replicates
- Without biological replicates, can still model variation based on parametric distributions (eg. Negative binomial), but expect lower accuracy.

Statistical Analysis Software for Identifying Differentially Expressed Transcripts

- Bioconductor
 - EdgeR
 - DEGseq
 - DESeq
 - And others...
- Tuxedo suite
 - Cuffdiff
 - (analysis enabled with CummeRbund/Bioconductor)

Examples of Results

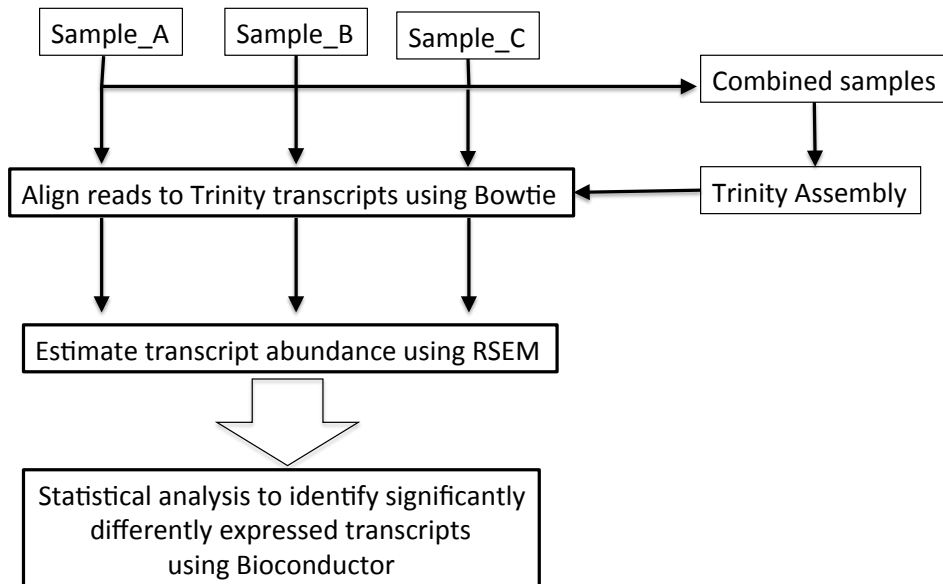
(Cuffdiff)

0	test_id	XLOC_000024
1	gene_id	XLOC_000024
2	gene	-
3	locus	7000000090838467:1335927-1338056
4	sample_1	condA
5	sample_2	condB
6	status	OK
7	value_1	680.167
8	value_2	68932
9	log2(fold_change)	6.66314
10	test_stat	-2.91993
11	p_value	0.00350111
12	q_value	0.0424377
13	significant	yes

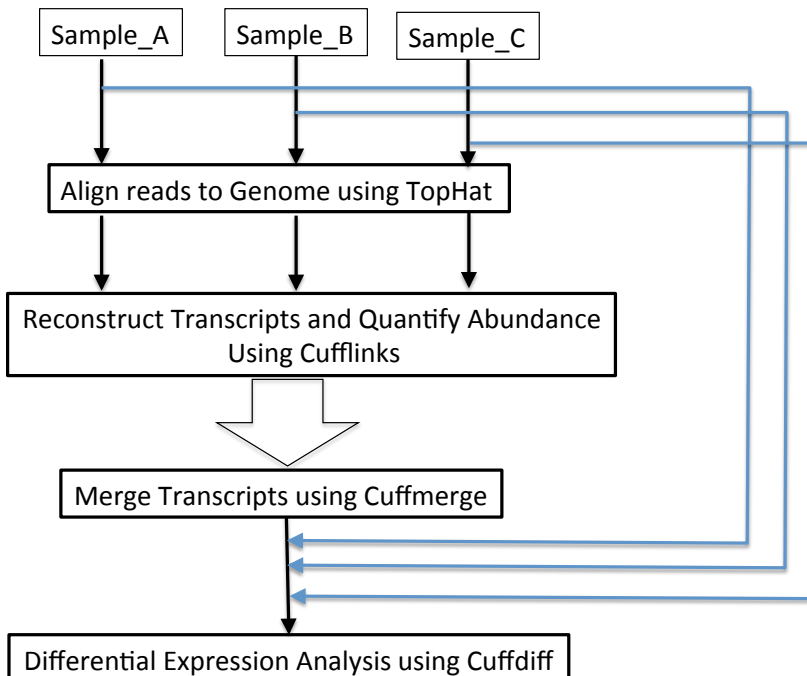
Examples of Results (example edgeR)

		comp217_c0_seq1
0	logFC	6.69056684523186
1	logCPM	16.1146897543805
2	PValue	2.06844466442231e-15
3	FDR	9.01969581996253e-13

Trinity Differential Expression Workflow

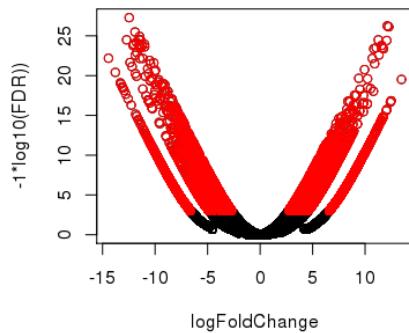


Tuxedo Differential Expression Workflow

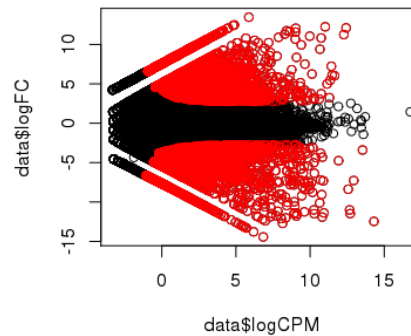


Plotting Pairwise Differential Expression Data

Volcano plot
(fold change vs. significance)



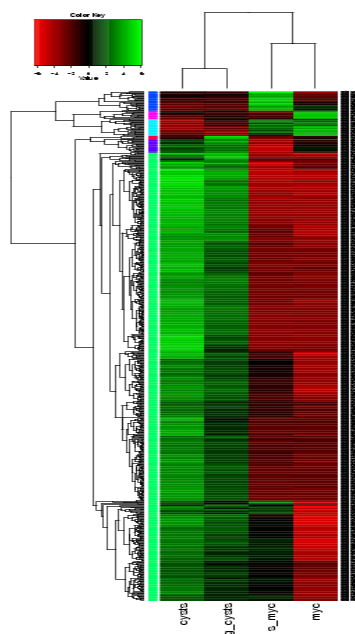
MA plot
(abundance vs. fold change)



Significantly differentially expressed transcripts have $FDR \leq 0.001$
(shown in red)

No replicates available, so modeled by edgeR using the
Negative Binomial with dispersion manually set to 0.1

Comparing Multiple Samples

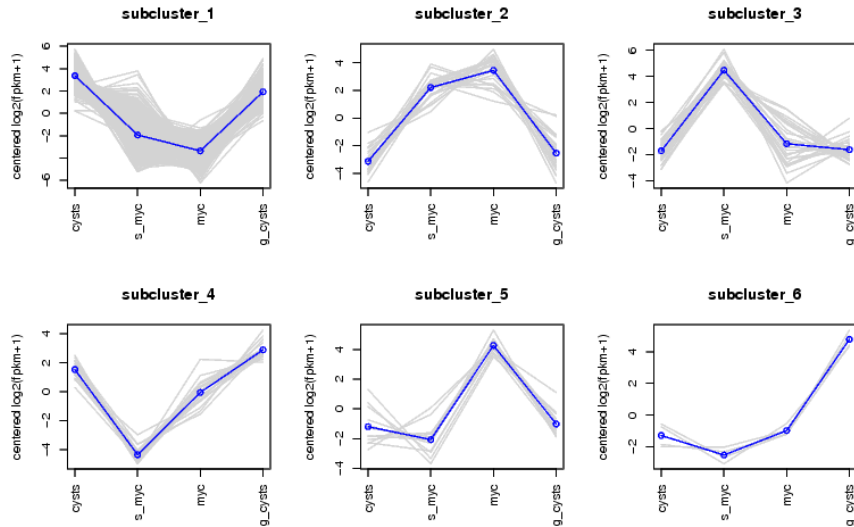


Heatmaps provide an effective tool
for navigating differential expression across
multiple samples.

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.



Hands-on Tutorials

- Tuxedo
 - Tophat alignment
 - Cufflinks transcript reconstructions
 - GenomeView for navigating the alignments
 - Cuffdiff for differential expression analysis
 - cummeRbund for exploring diff. express. results.
- Trinity
 - De novo assembly using Trinity
 - Bowtie and RSEM for abundance estimation
 - edgeR for differential expression analysis