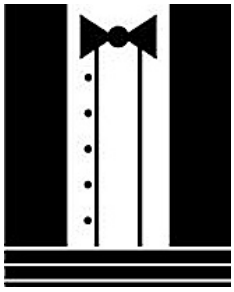
Genome-Based and Genome-Free Transcript Reconstruction and Analysis Using RNA-Seq Data

Brian Haas

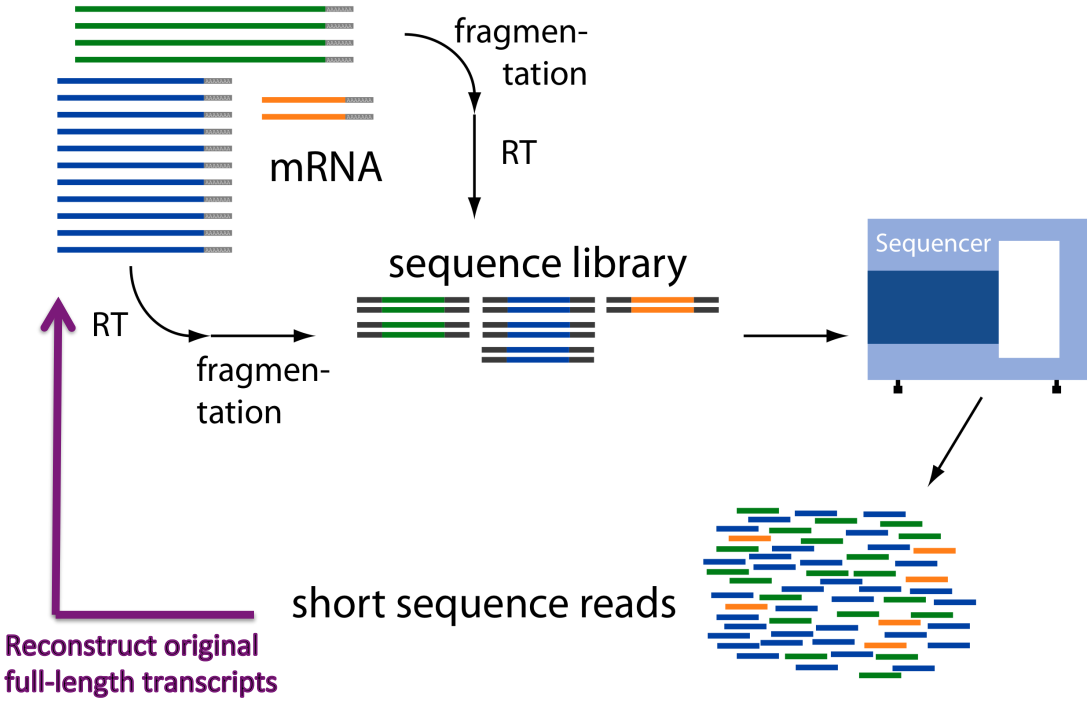
Broad Institute



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
AAACAACAGGGCACATTGTCACTCTTGTATTTGAAAAACACTTTCCGGCCAT
```

FASTQ format:

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

Read

Quality values

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

AsciiEncodedQual ('C') = **64**

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

Paired-end Sequences

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

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

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

Transcript Reconstruction from RNA-Seq Reads

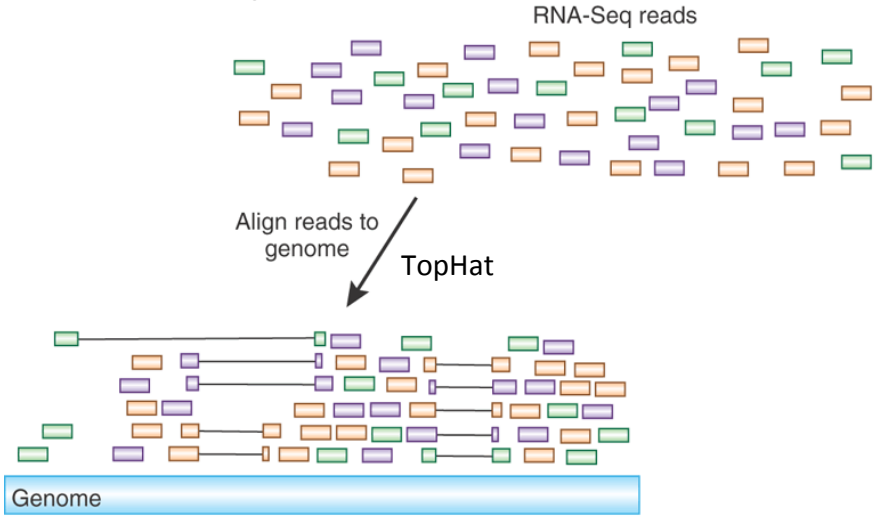
Advancing RNA-Seq analysis

Brian J Haas & Michael C Zody

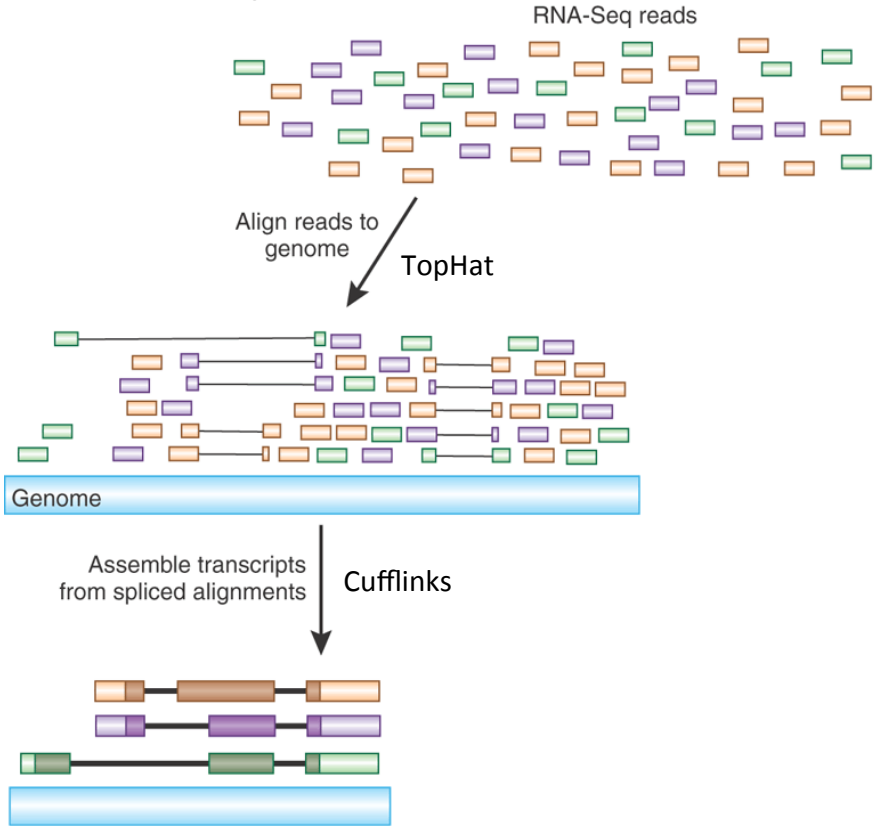
Nature Biotech, 2010

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

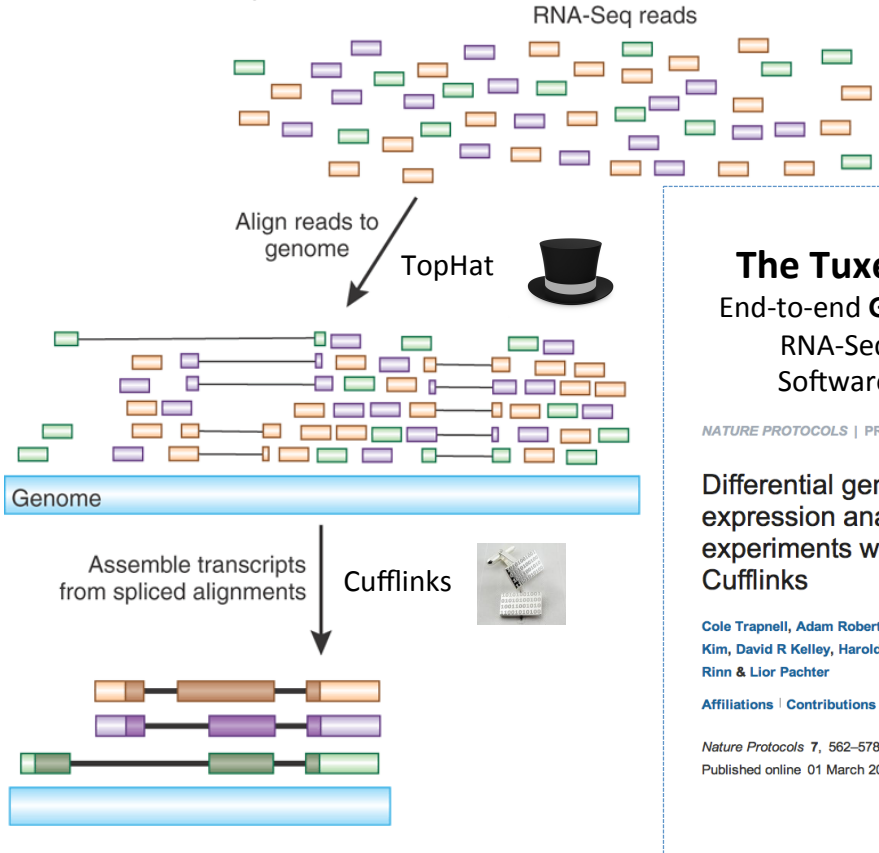
Transcript Reconstruction from RNA-Seq Reads



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

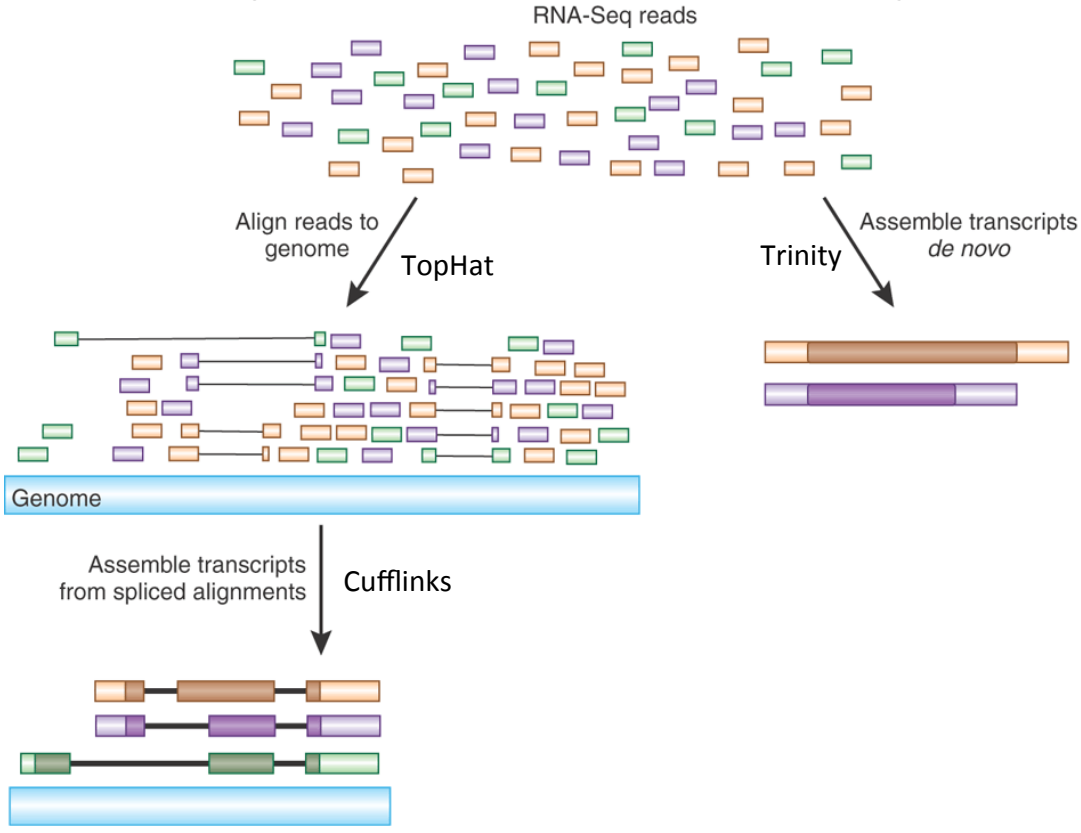
Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks

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

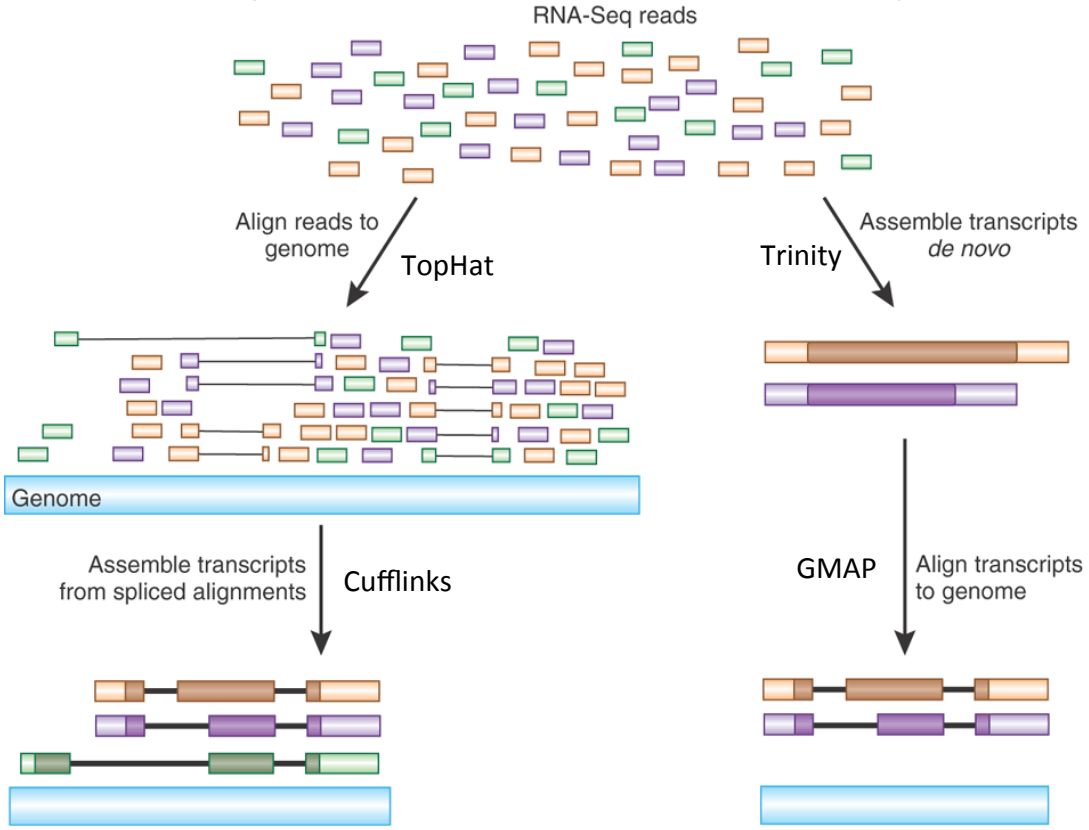
Affiliations | Contributions | Corresponding author

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

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

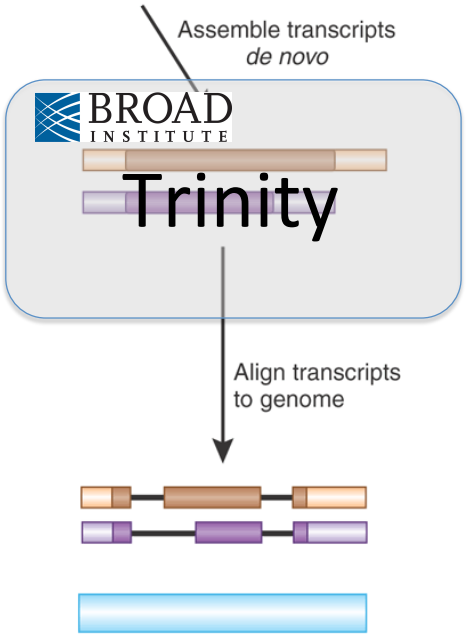
NATURE PROTOCOLS | PROTOCOL

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

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

Affiliations | Contributions | Corresponding authors

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



Overview of the Tuxedo Software Suite

Bowtie (fast short-read alignment)

TopHat (spliced short-read alignment)

↓

Cufflinks (transcript reconstruction from alignments)

↓

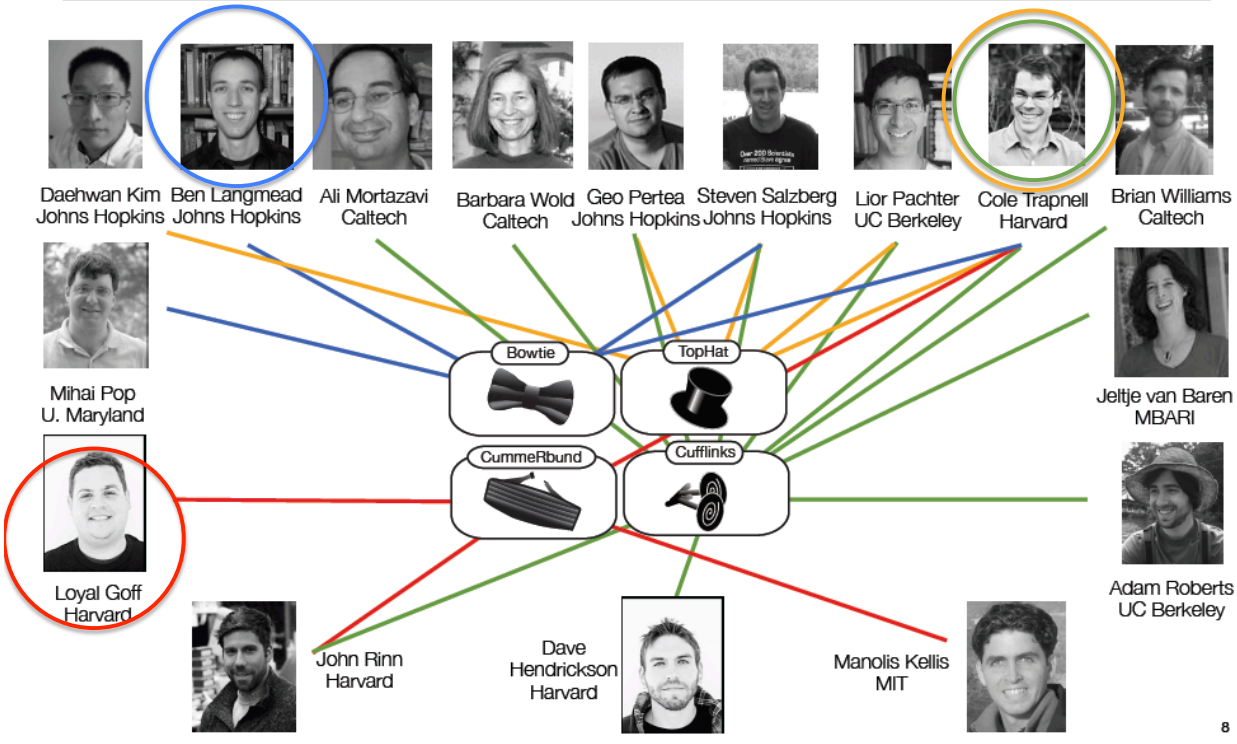
Cuffdiff (differential expression analysis)

↓

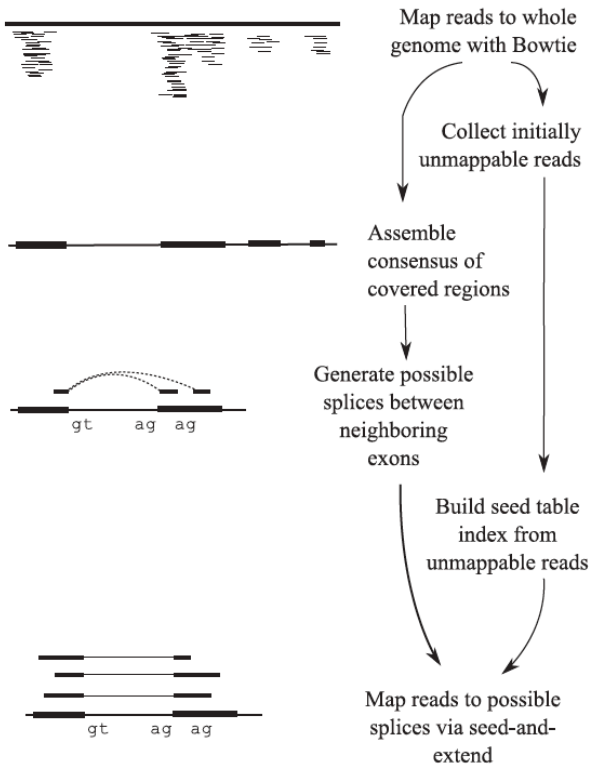
CummeRbund (visualization & analysis)

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       CCCAAACAAGCCGAACTAGCTGATTTGGCTCGTAAAGACCCGGAAA
10      ###CB?=ADDBCBCDEEFFDEFFFDEFFGDBEFGEDGCFGFGGGGG
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
9       CCCAAACAAGCCGAACTAGCTGATTTGGCTCGTAAAGACCCGGAAA   (read sequence, oriented according to the forward alignment)
10      ###CB?=ADDBCBCDEEFFDEFFFDEFFGDBEFGEDGCFGFGGGGG   (base quality values)
11      MD:Z:67
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
16 XQ:i:40
17 X2:i:0

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

Samtools

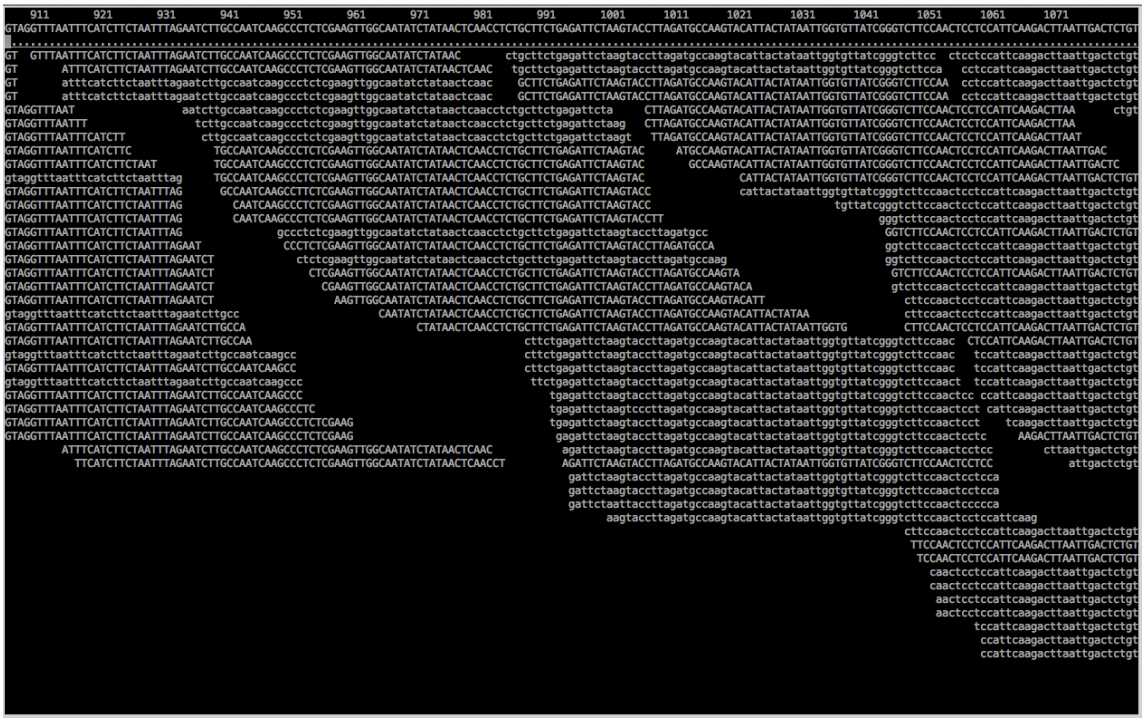
- Tools for
  - converting SAM <-> BAM
  - Viewing BAM files (eg. samtools view file.bam | less)
  - Sorting BAM files, and lots more:

```
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
```

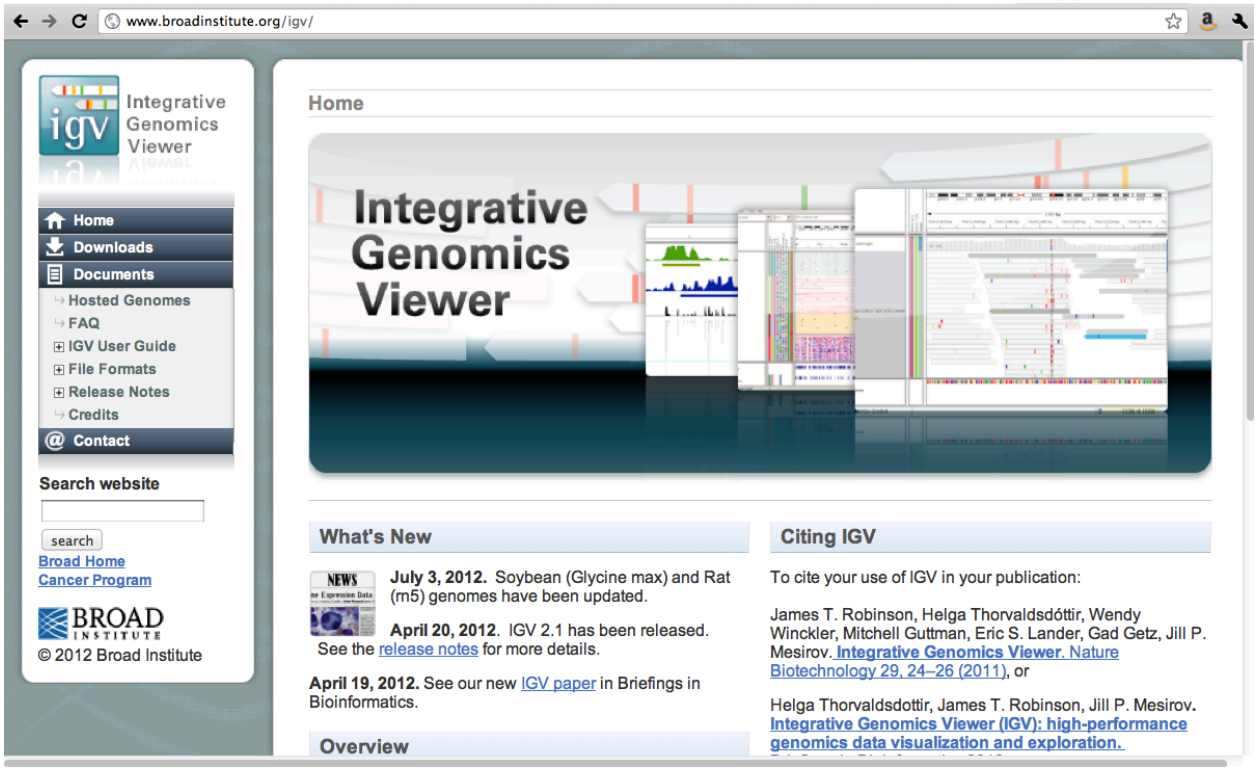
Visualizing Alignments of RNA-Seq reads

Text-based Alignment Viewer

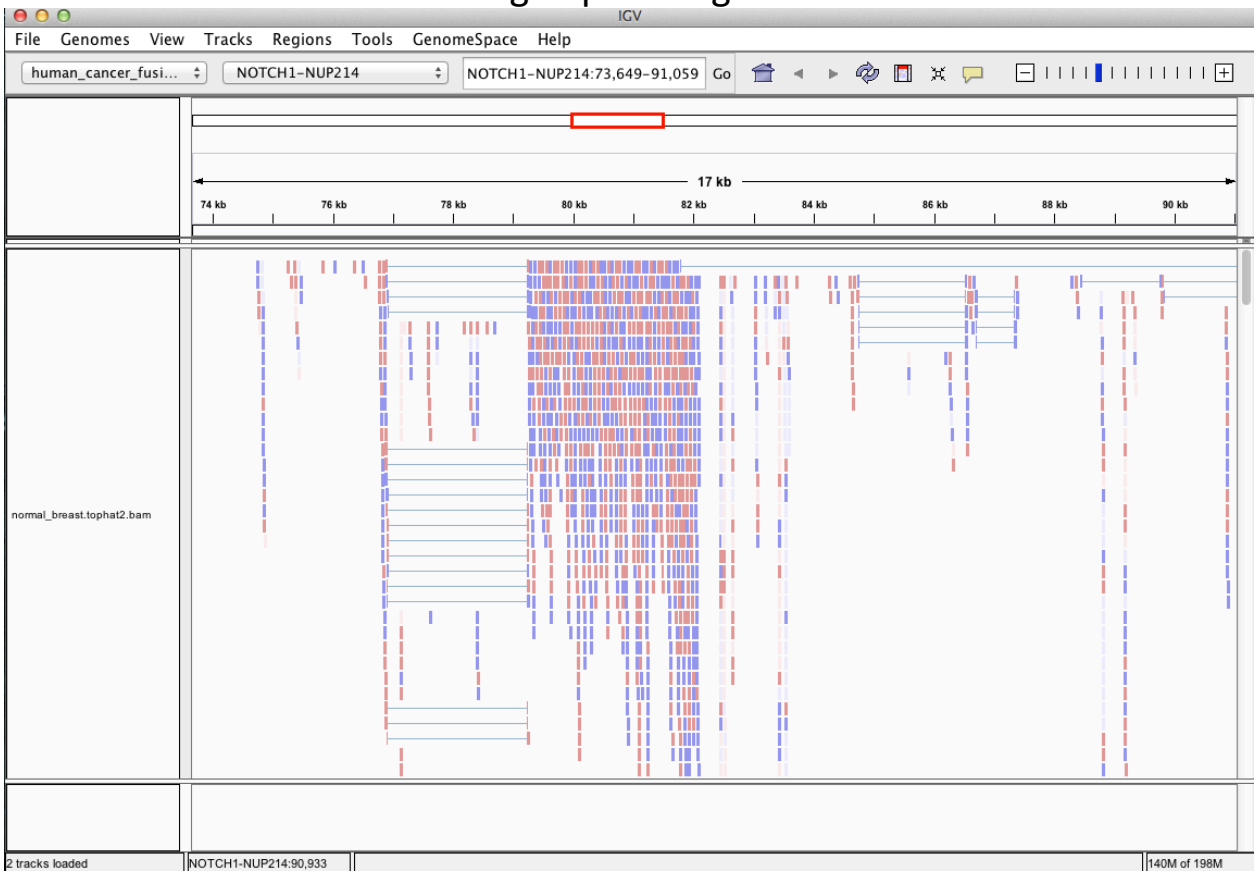
% samtools tview alignments.bam target.fasta



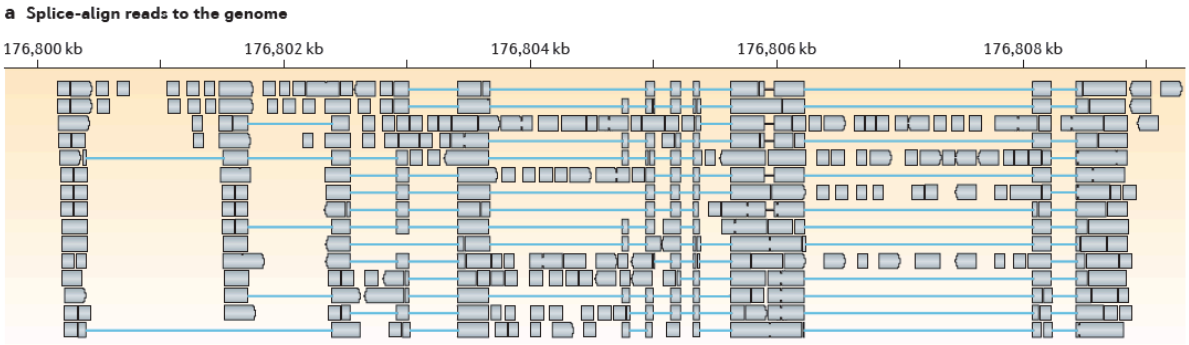
IGV



IGV: Viewing Tophat Alignments

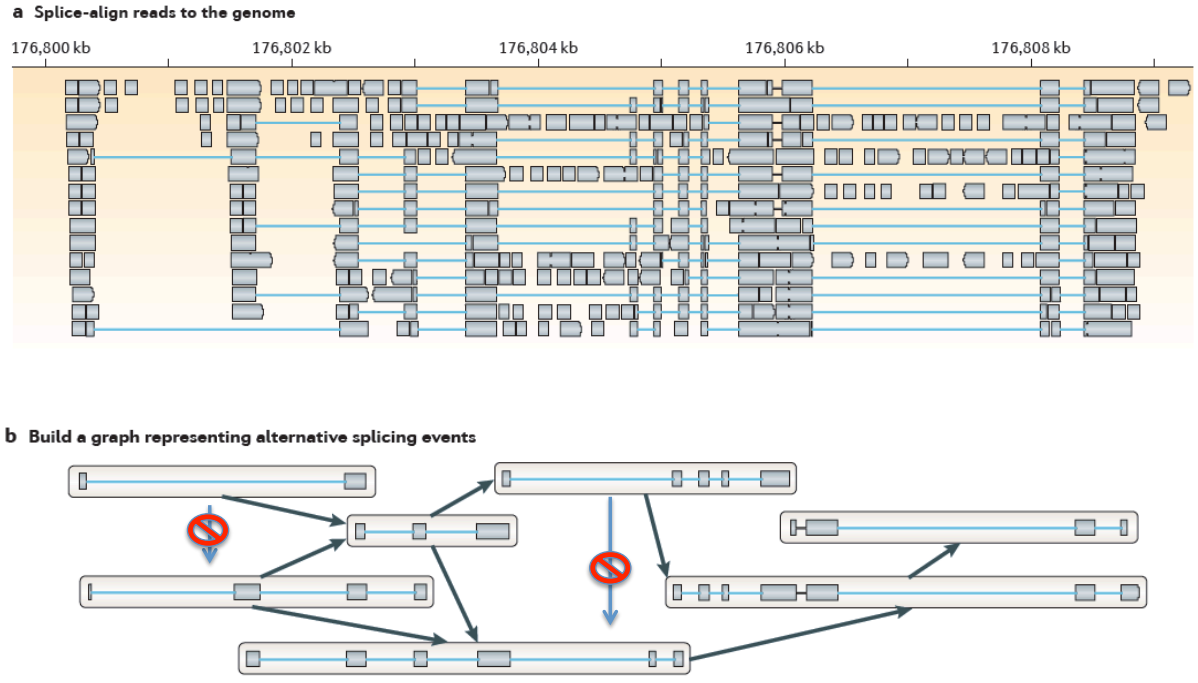


Transcript Reconstruction Using Cufflinks



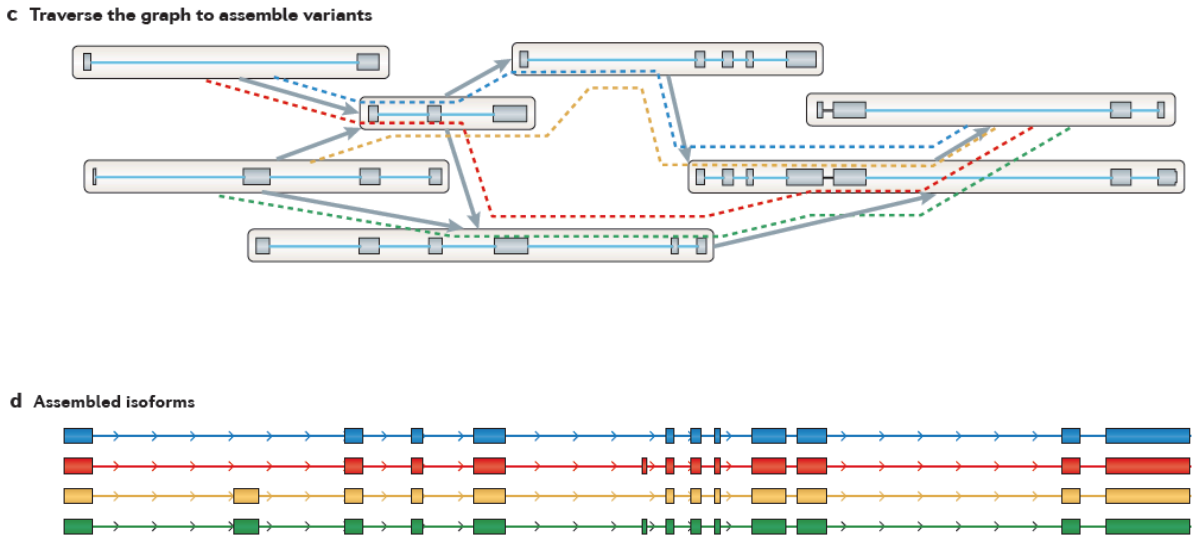
From Martin & Wang. Nature Reviews in Genetics. 2011

Transcript Reconstruction Using Cufflinks



From Martin & Wang. Nature Reviews in Genetics. 2011

Transcript Reconstruction Using Cufflinks



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   (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

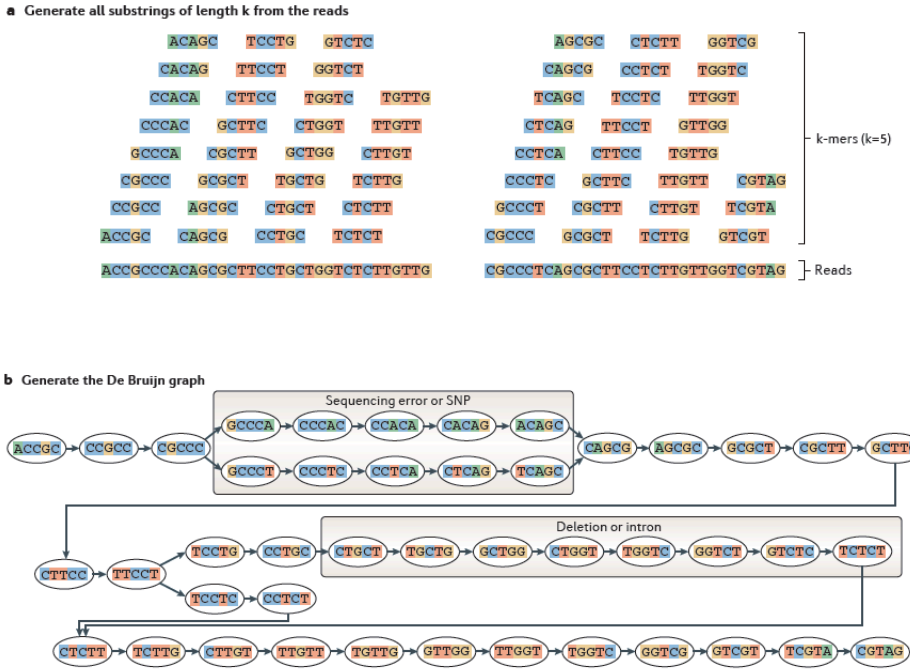
No genome required

Empower studies of non-model organisms

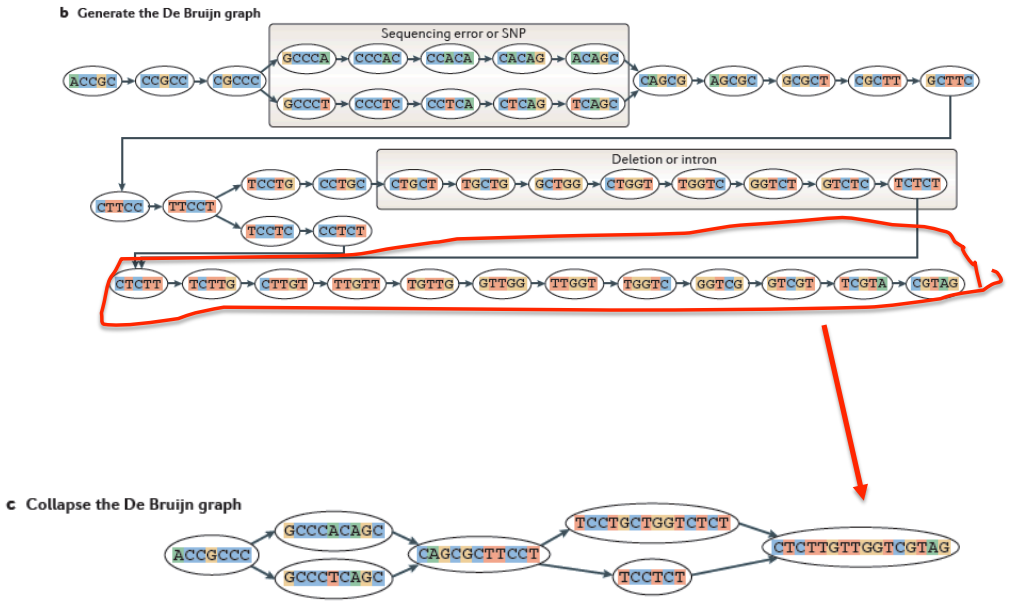
- expressed gene content
- transcript abundance
- differential expression

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

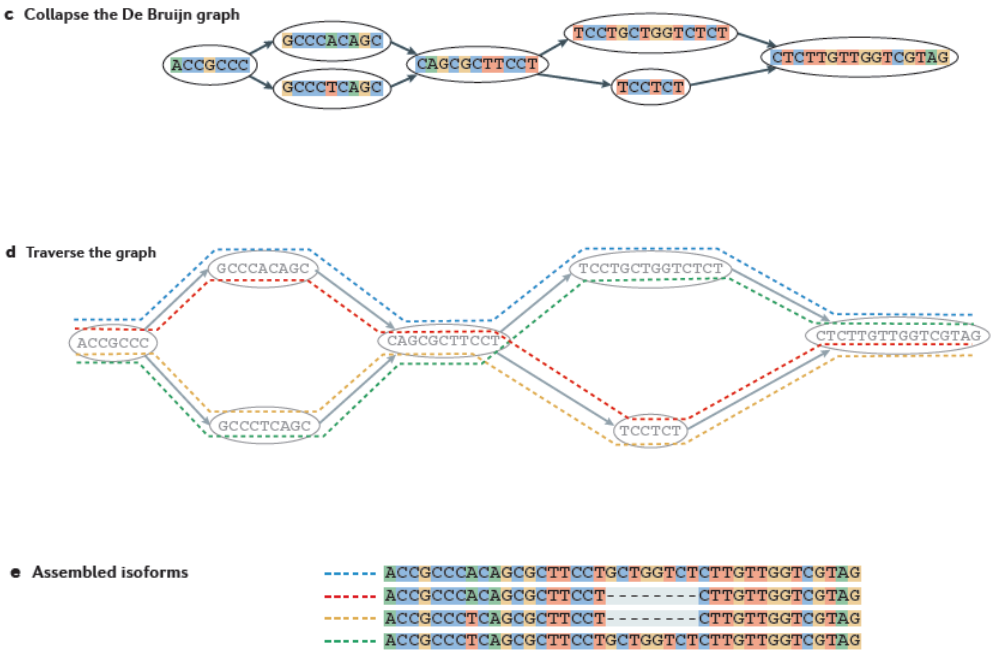
Sequence Assembly via De Bruijn Graphs



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



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



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

Contrasting Genome and Transcriptome Assembly

Genome Assembly

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

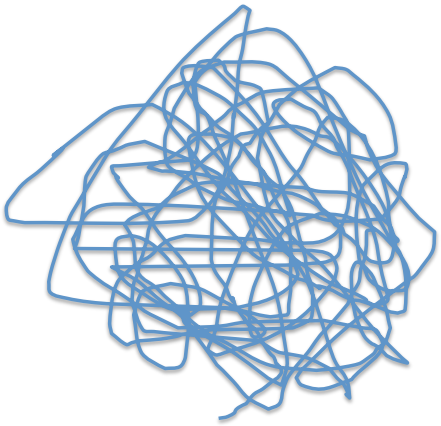
Transcriptome Assembly

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

Trinity

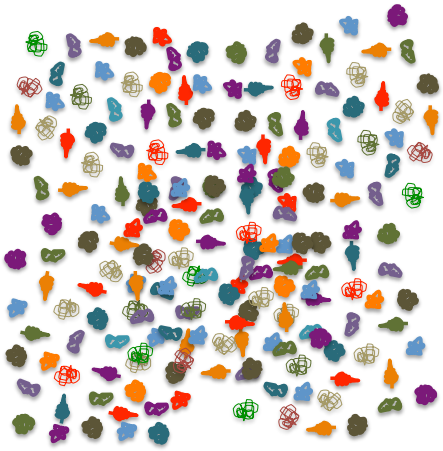
Trinity Aggregates Isolated Transcript Graphs

Genome Assembly
Single Massive Graph



Entire chromosomes represented.

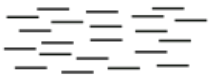
Trinity Transcriptome Assembly
Many Thousands of Small Graphs

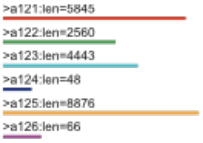


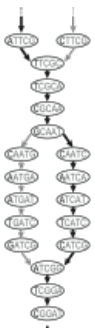
Ideally, one graph per expressed gene.

Trinity – How it works:

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







Thousands of disjoint graphs

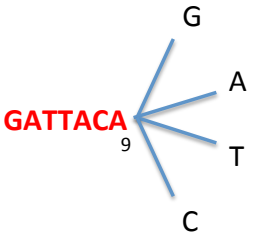
...CTTCGCAA...TGATCGGAT...
...ATTCGCAA...TCATCGGAT...

Inchworm Algorithm

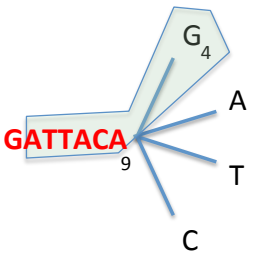
Decompose all reads into overlapping Kmers (25-mers)

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

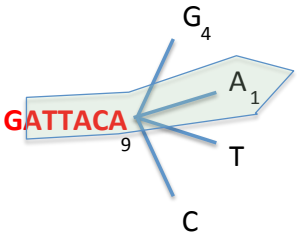
Extend kmer at 3’ end, guided by coverage.



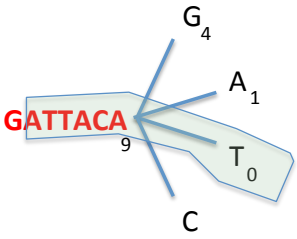
Inchworm Algorithm



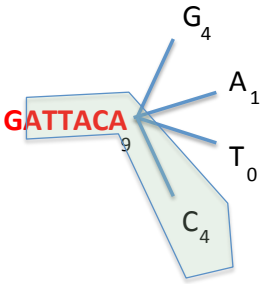
Inchworm Algorithm



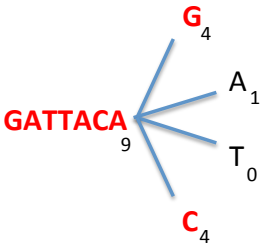
Inchworm Algorithm



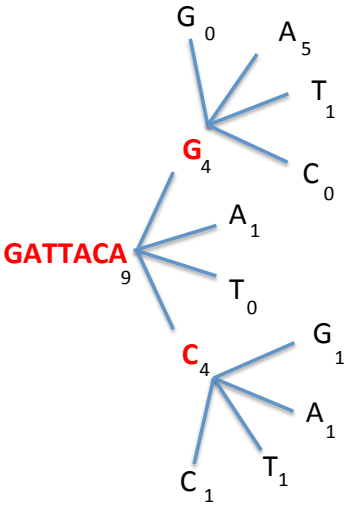
Inchworm Algorithm



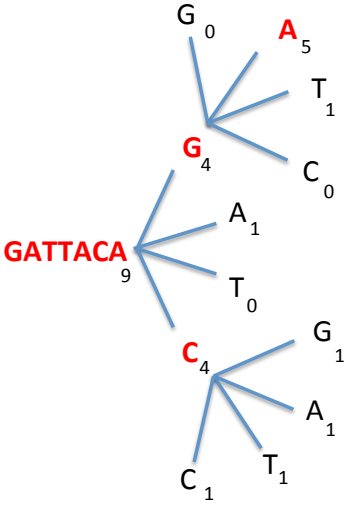
Inchworm Algorithm



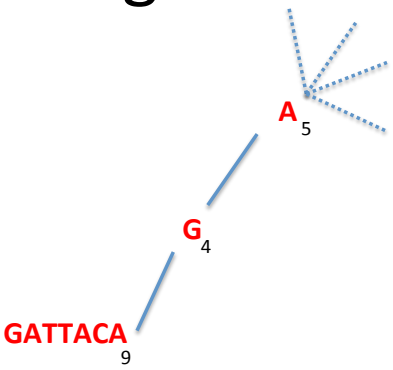
Inchworm Algorithm



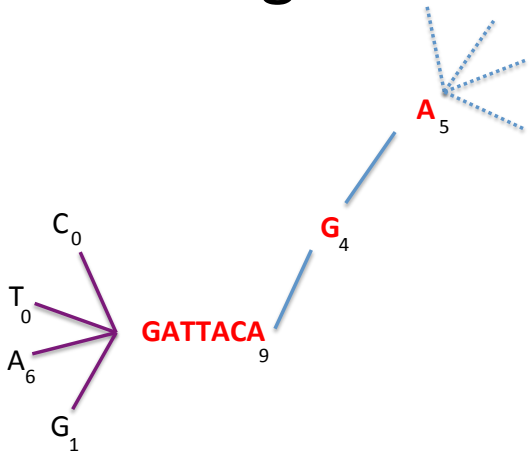
Inchworm Algorithm



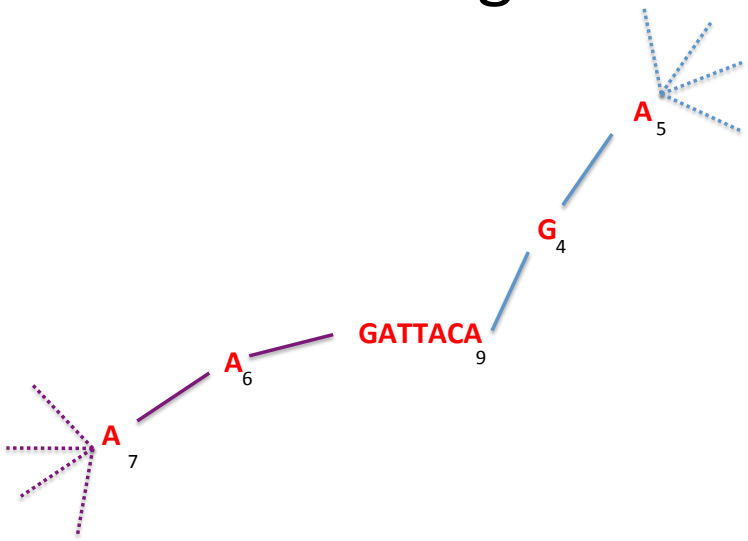
Inchworm Algorithm



Inchworm Algorithm



Inchworm Algorithm



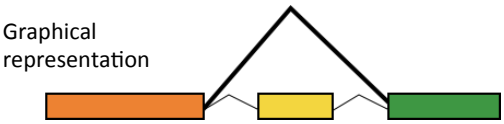
Report contig: ....AAGATTACAGA....

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

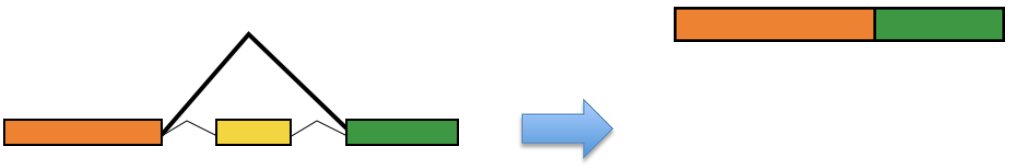
Inchworm Contigs from Alt-Spliced Transcripts

Expressed isoforms

Inchworm Contigs from Alt-Spliced Transcripts

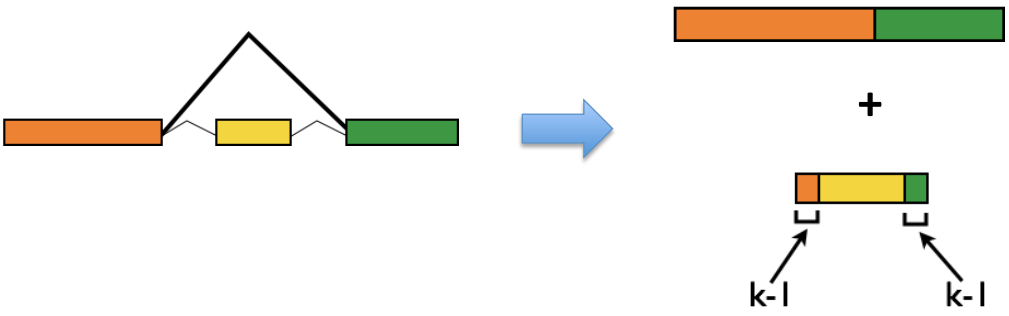


Inchworm Contigs from Alt-Spliced Transcripts

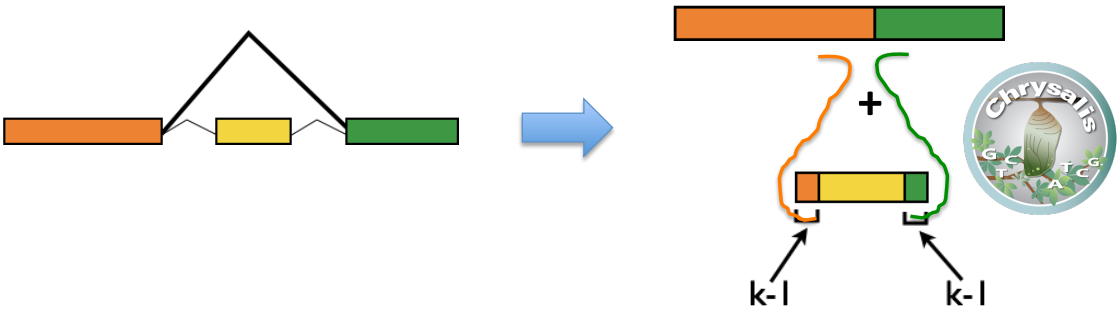


Inchworm Contigs from Alt-Spliced Transcripts

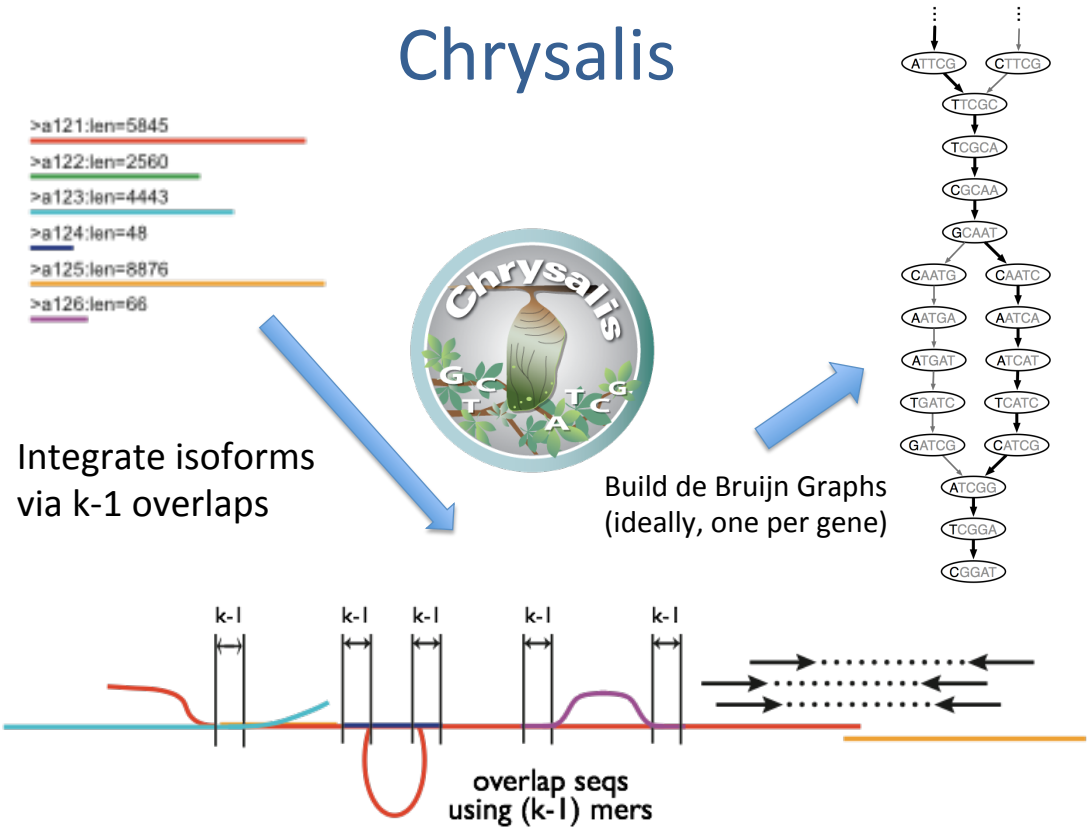
Inchworm Contigs from Alt-Spliced Transcripts

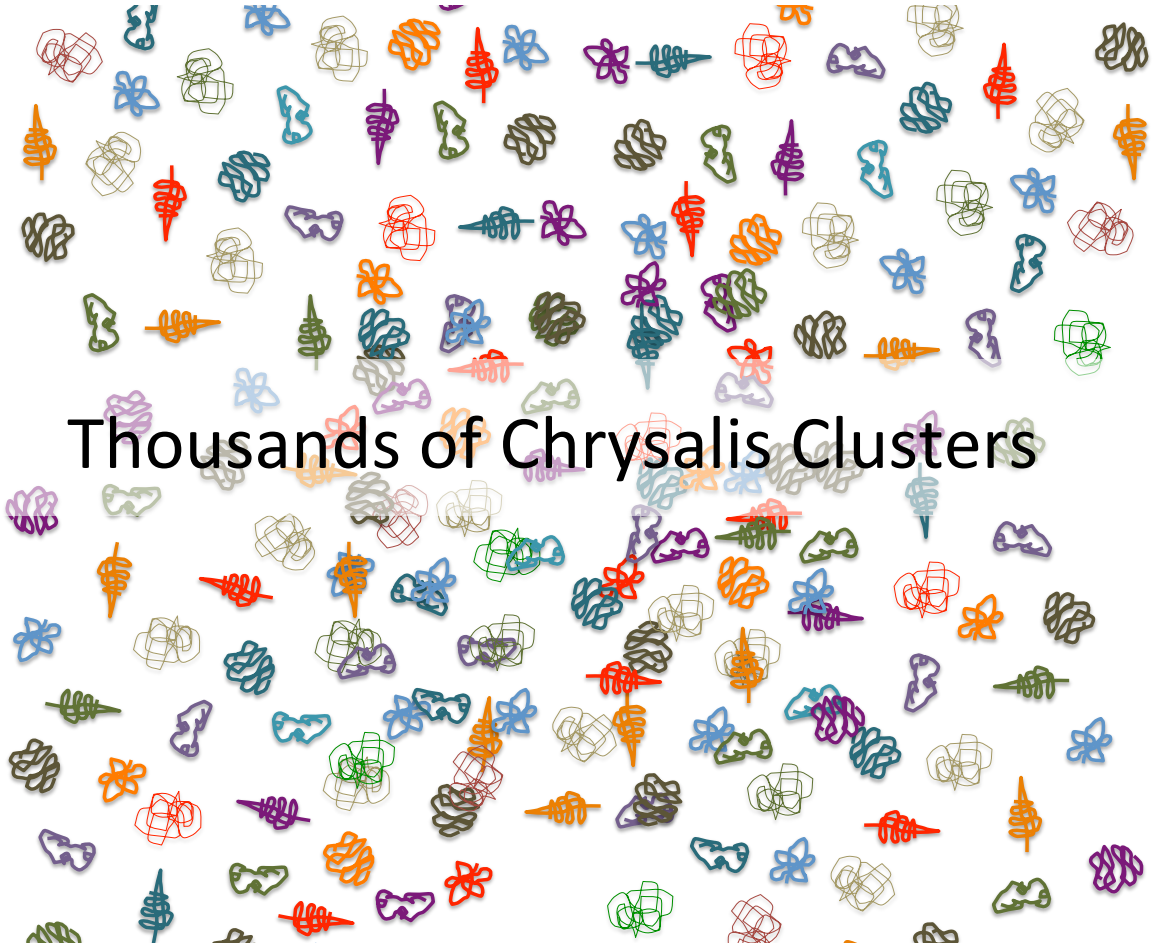


Chrysalis Re-groups Related Inchworm Contigs

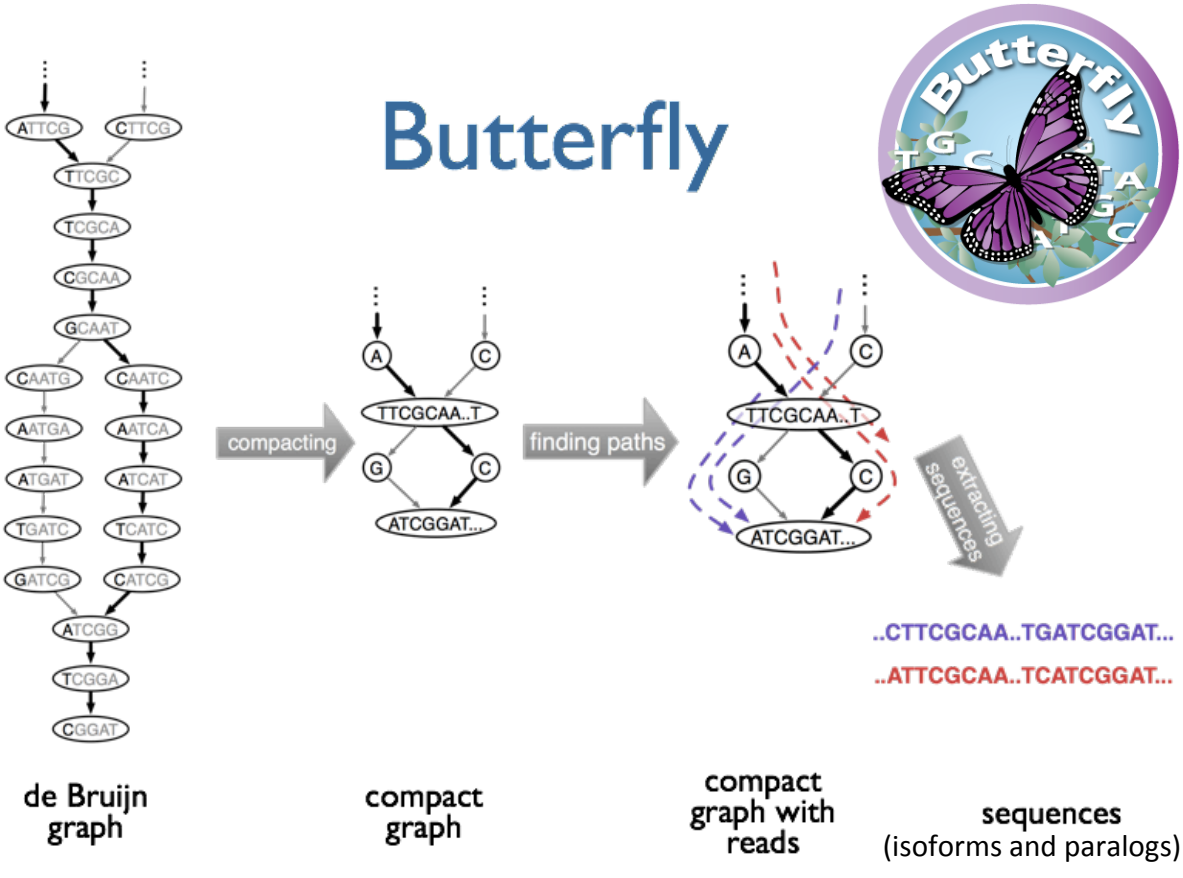


Chrysalis uses (k-1) overlaps and read support to link related Inchworm contigs



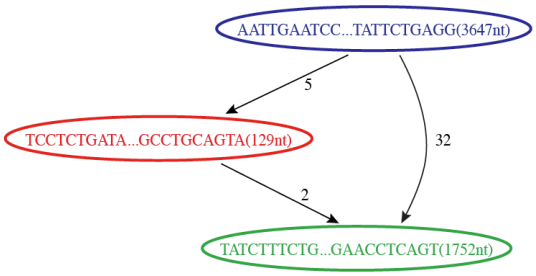


Thousands of Chrysalis Clusters



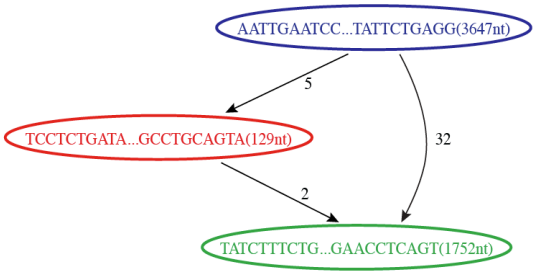
Butterfly Example 1:
Reconstruction of Alternatively Spliced Transcripts

Butterfly's Compacted Sequence Graph



Reconstruction of Alternatively Spliced Transcripts

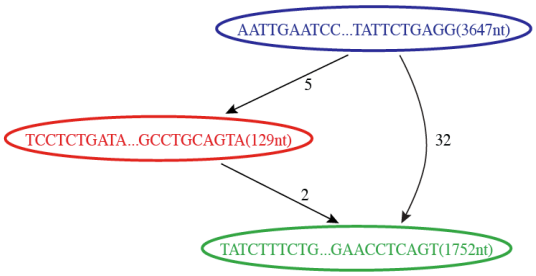
Butterfly's Compacted Sequence Graph



Reconstructed Transcripts

Reconstruction of Alternatively Spliced Transcripts

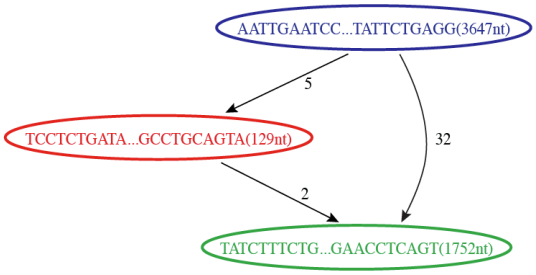
Butterfly's Compacted Sequence Graph



Reconstructed Transcripts

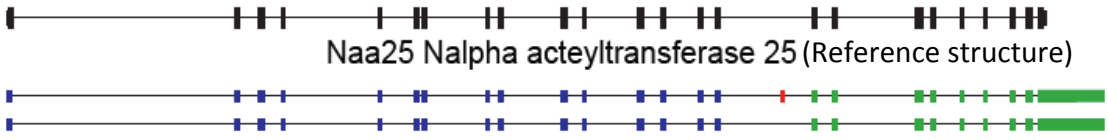
Reconstruction of Alternatively Spliced Transcripts

Butterfly's Compacted Sequence Graph



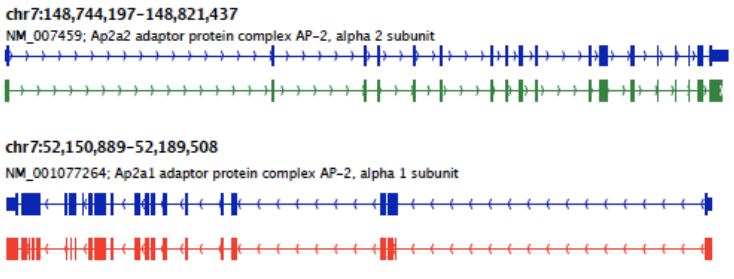
Reconstructed Transcripts

Aligned to Mouse Genome

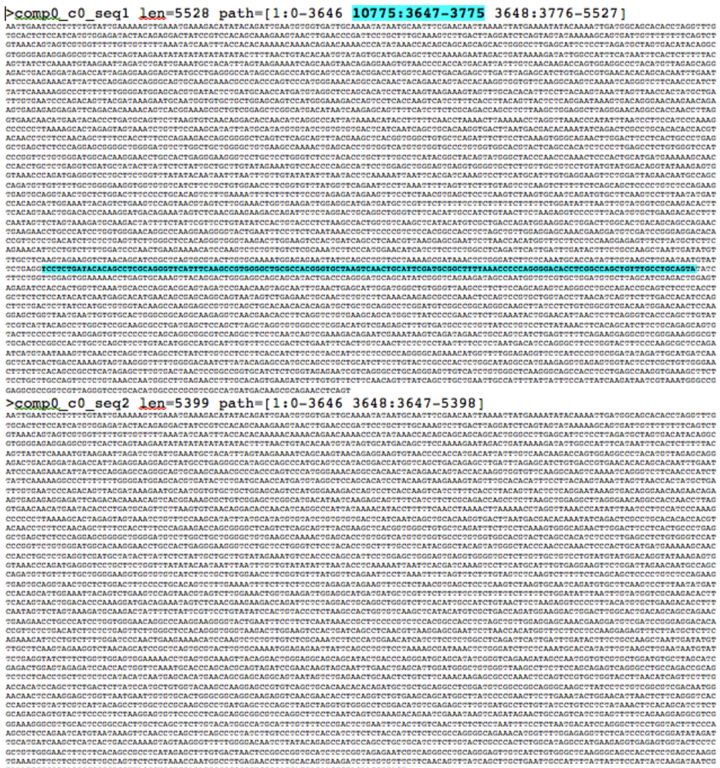


Butterfly Example 2:
Teasing Apart Transcripts of Paralogous Genes

Teasing Apart Transcripts of Paralogous Genes

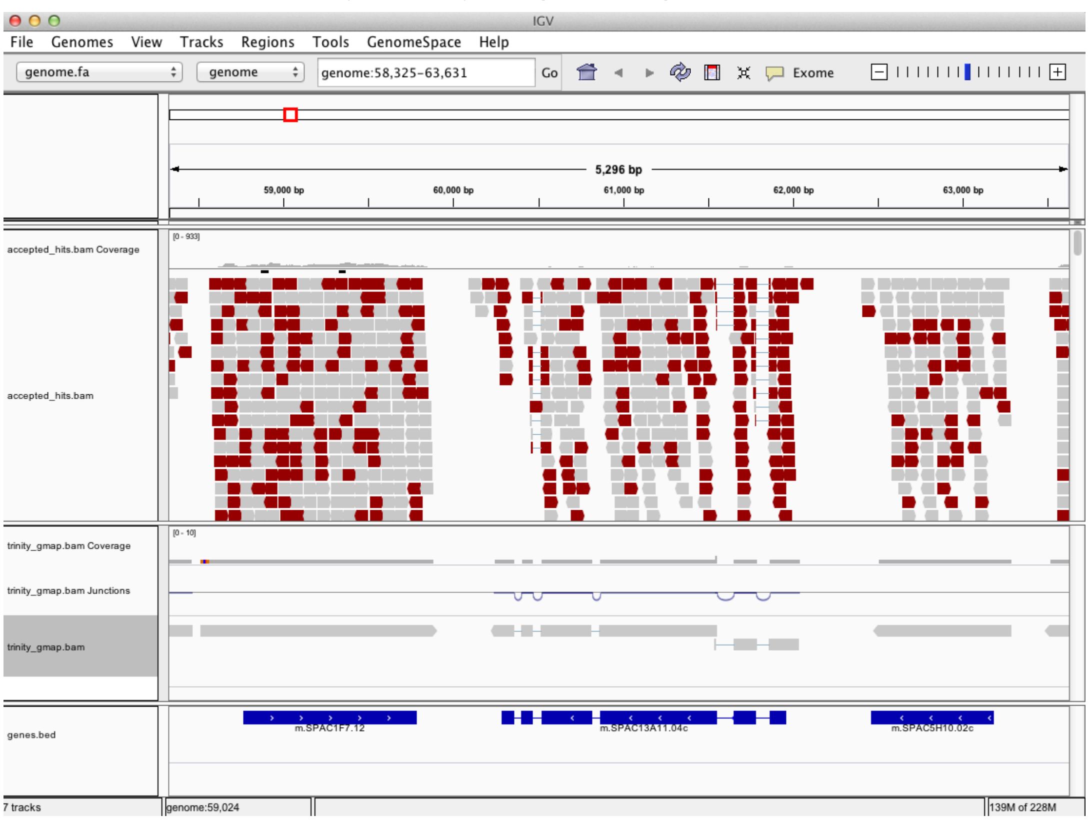


Trinity output: A multi-fasta file



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

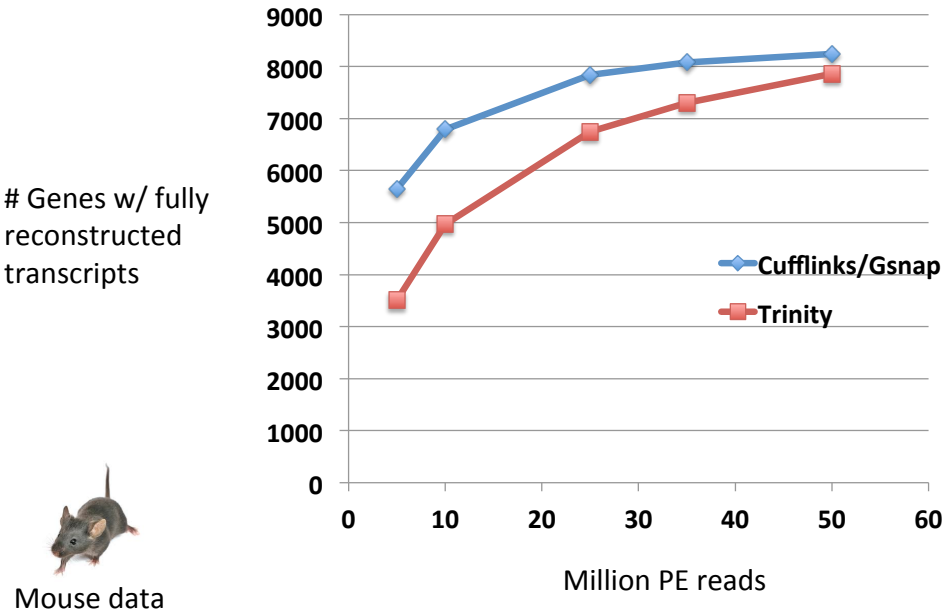
(Trinity transcripts aligned using GMAP)



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



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 |

BROAD INSTITUTE

Comprehensive comparative analysis of strand-specific RNA sequencing methods

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

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

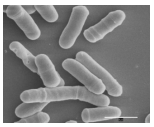
'dUTP second strand marking' identified as the leading protocol

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

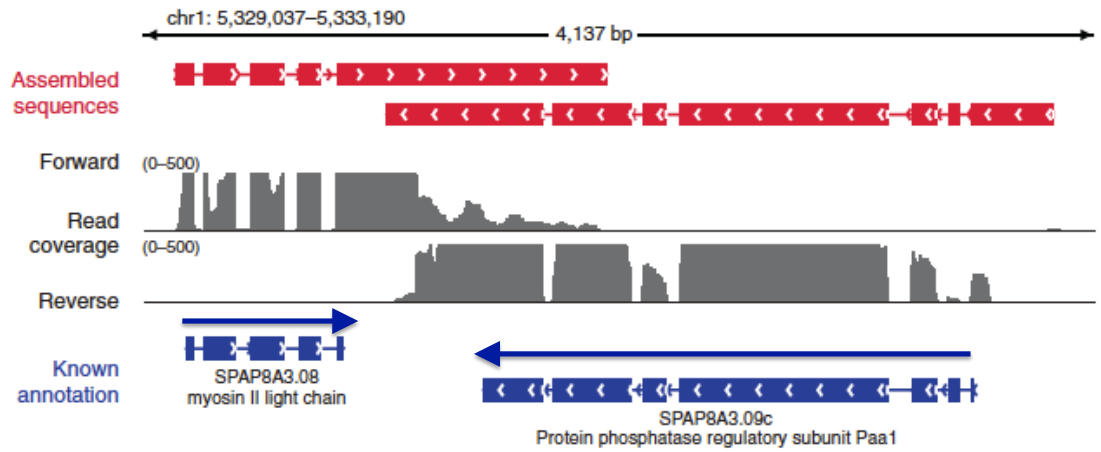
Nevertheless, direct information on the originating strand can substantially enhance the value of an RNA-seq experiment. For

boundaries of adjacent genes transcribed on opposite strands and resolve the correct expression levels of coding or noncoding overlapping transcripts. These tasks are particularly challenging in small microbial genomes, prokaryotic and eukaryotic, in which

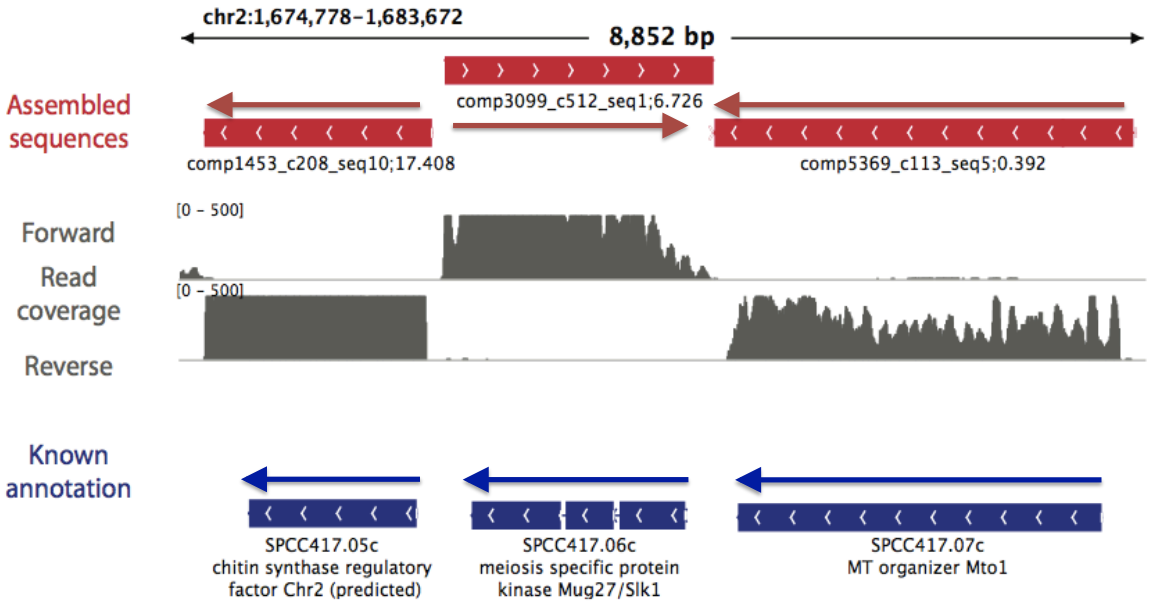
Overlapping UTRs from Opposite Strands

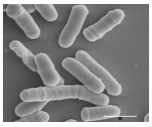


Schizosacharomyces pombe
(fission yeast)



Antisense-dominated Transcription





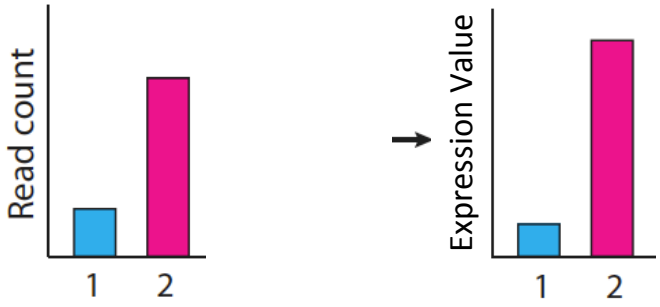
Summary

- 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.

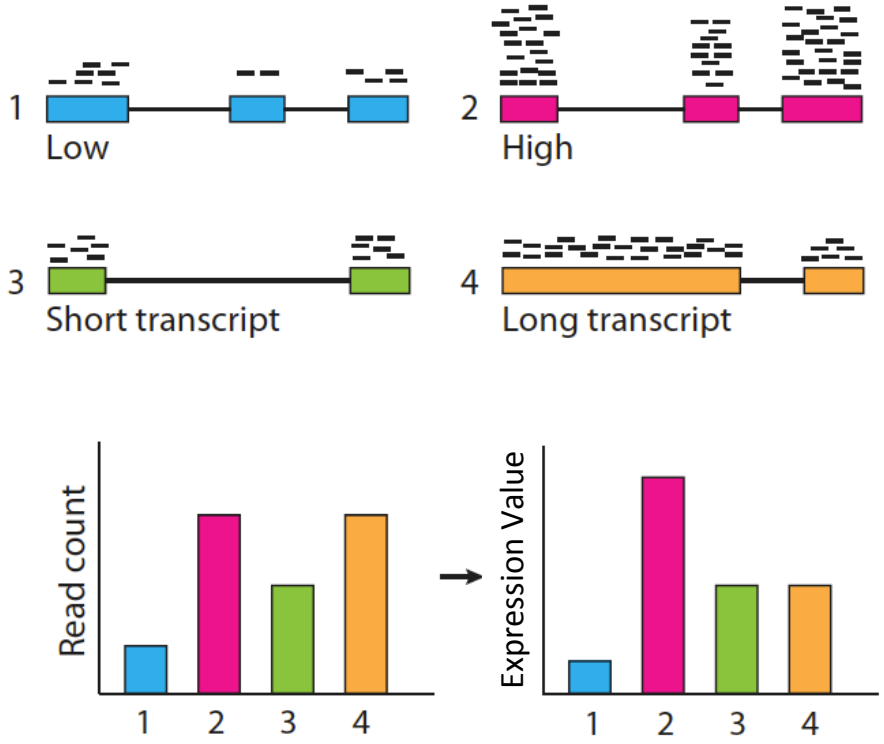
Abundance Estimation
(Aka. Computing Expression Values)

Calculating expression of genes and transcripts



Slide courtesy of Cole Trapnell

Calculating expression of genes and transcripts



Slide courtesy of Cole Trapnell

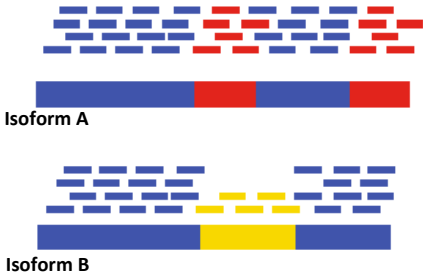
Normalized Expression Values

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

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

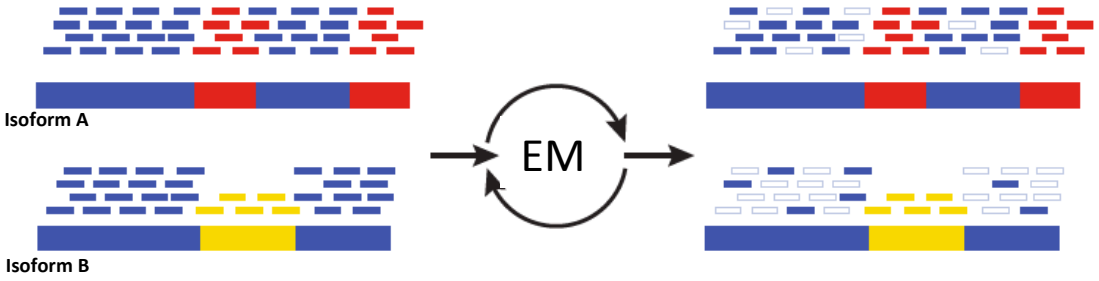
FPKM

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

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

Subset of genes highly expressed in liver

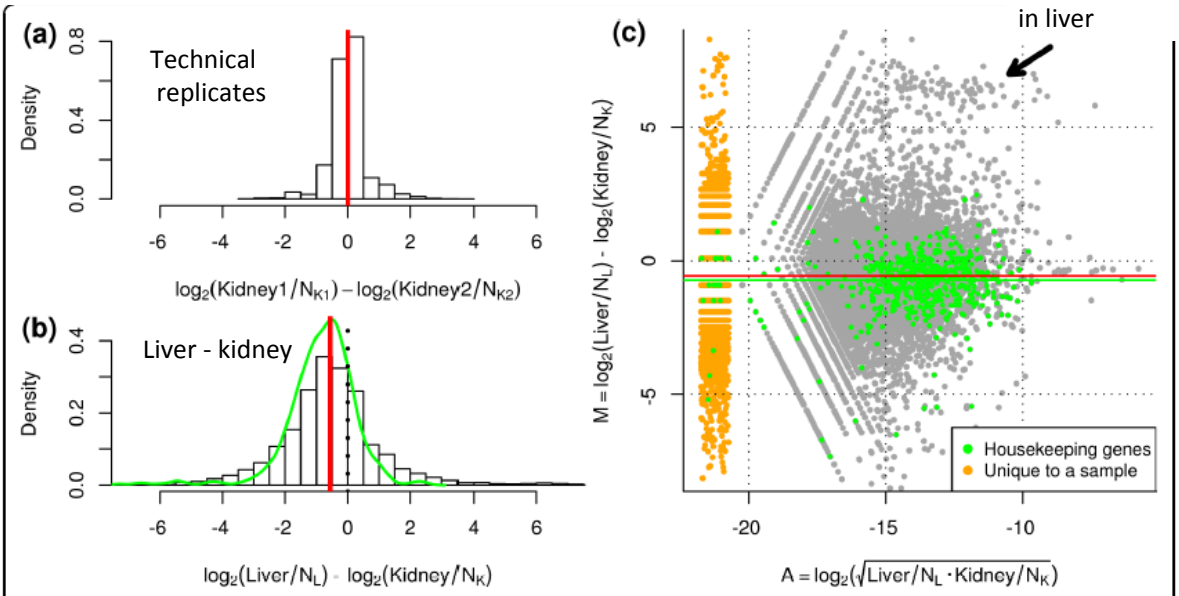


Figure 1 Normalization is required for RNA-seq data. Data from [6] comparing log ratios of **(a)** technical replicates and **(b)** liver versus kidney expression levels, after adjusting for the total number of reads in each sample. The green line shows the smoothed distribution of log-fold-changes of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

Robinson and Oshlack, Genome Biology, 2010

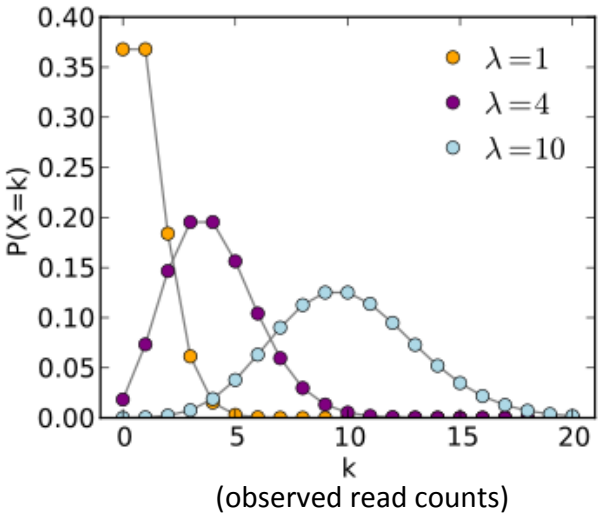
Diff. Expression Analysis Involves

- 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



Mean # fragments

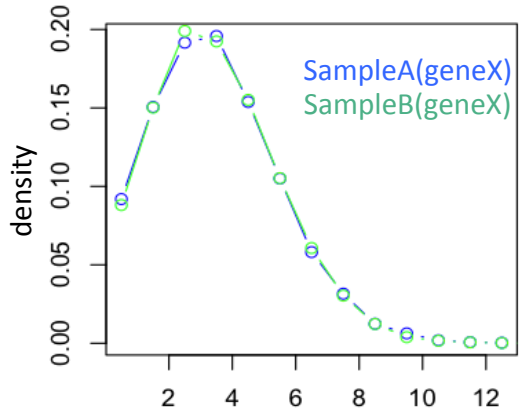
(observed read counts)

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

Example: One gene*not* differentially expressed

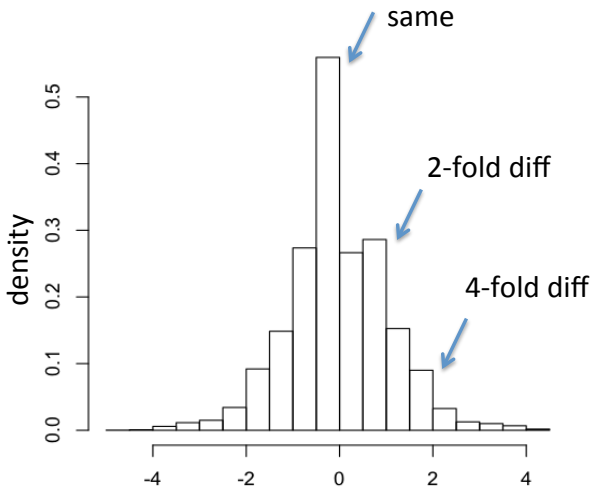
SampleA(gene) = SampleB(gene) = 4 reads

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



(k) number of reads observed

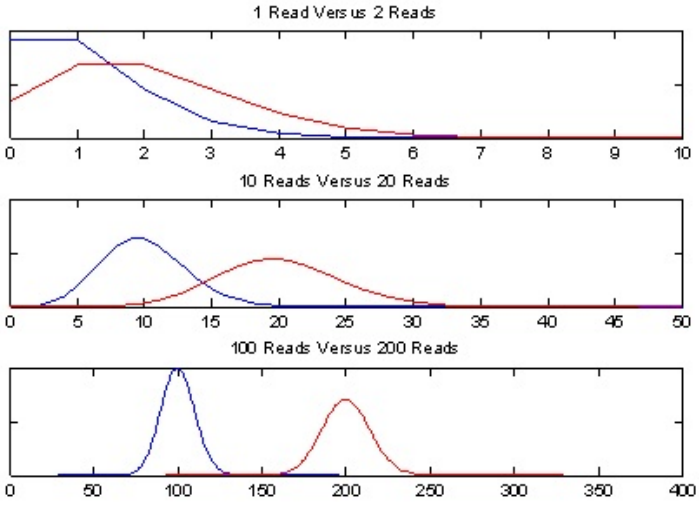
Dist. of $\log_2$(fold change) values



x = $\log_2$(SampleA/SampleB)

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.

From: http://gkno2.tumblr.com/post/24629975632/thinking-about-rna-seq-experimental-design-for

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

ShrinkSeq
NoiSeq
baySeq
Vsf
Voom
SAMseq
TSPM
DESeq
EBSeq
NBPSeq
edgeR

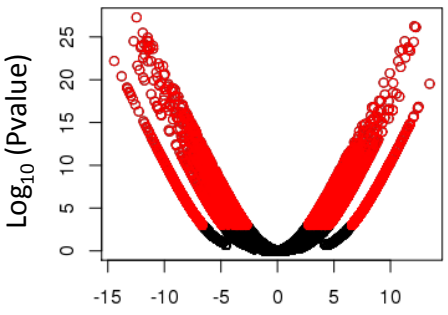
+ other (not-R) including CuffDiff

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

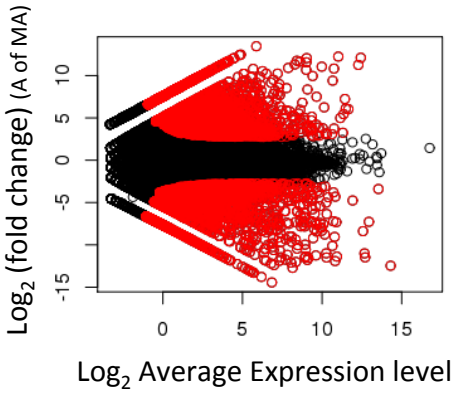
Visualization of DE results and Expression Profiling

Plotting Pairwise Differential Expression Data

Volcano plot
(fold change vs. significance)

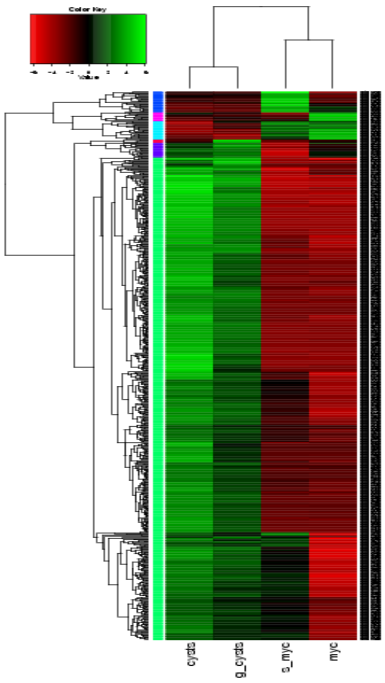


MA plot
(abundance vs. fold change)



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.

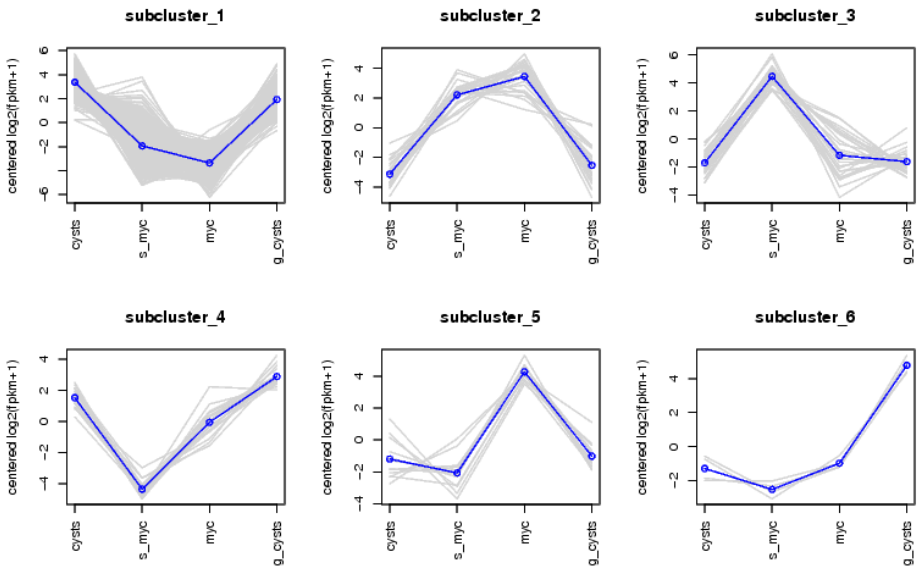
Clustering can be performed across both axes:

-cluster transcripts with similar expression patters.

-cluster samples according to similar expression values among transcripts.

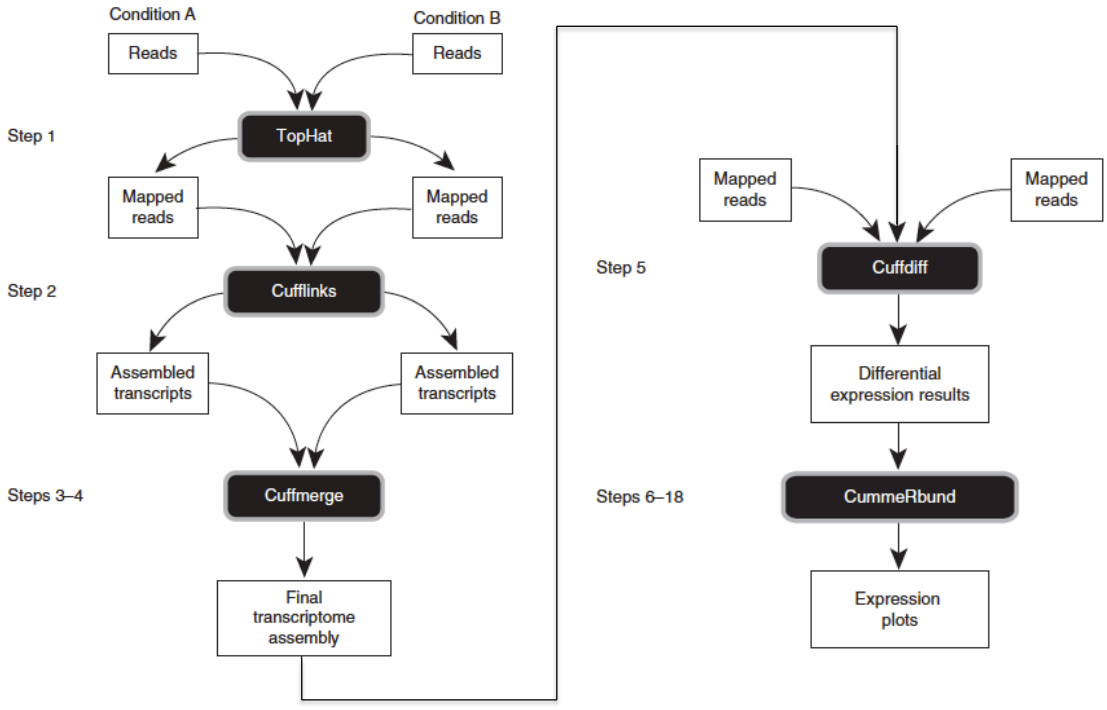
Examining Patterns of Expression Across Samples

Can extract clusters of transcripts and examine them separately.



RNA-Seq Analysis Frameworks

Tuxedo Framework for Transcriptome Analysis

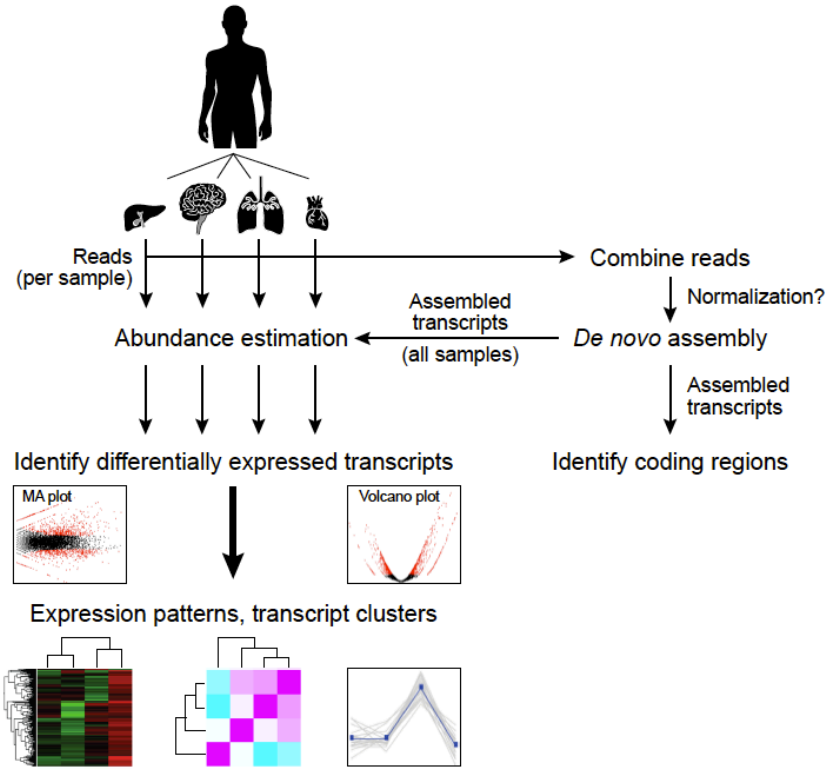


Derived from: Nat Protoc. 2012 Mar 1;7(3):562-78. doi: 10.1038/nprot.2012.016.

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
  - Bowtie: http://bowtie-bio.sourceforge.net/index.shtml
  - Tophat: http://tophat.cbcb.umd.edu/
  - Cufflinks: http://cufflinks.cbcb.umd.edu/
- Trinity
  http://trinityrnaseq.sourceforge.net/
- IGV for Visualization
  http://www.broadinstitute.org/igv/
- 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/
  - http://www.nature.com/nprot/journal/v8/n8/full/nprot.2013.084.html