BI694

Bioinformatics & Phylogenetics

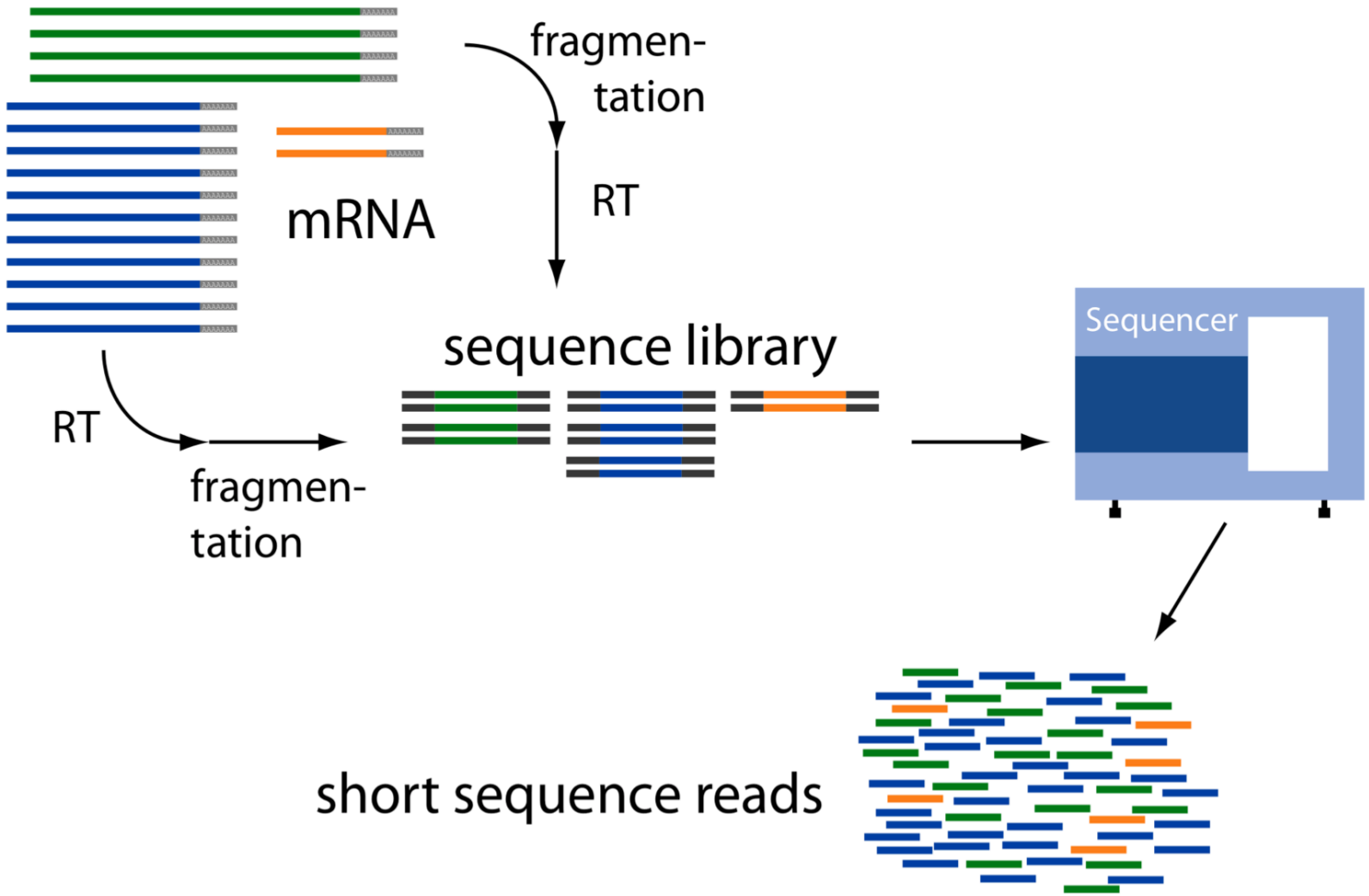
Winter Semester 2017

WEEK 9

Transcriptome Assembly, Read Mapping and DE Analysis

- This week: Transcriptome Assembly and DE analysis
- Next week: phylogenetics and revisit of MSA

Overview of RNA-Seq



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

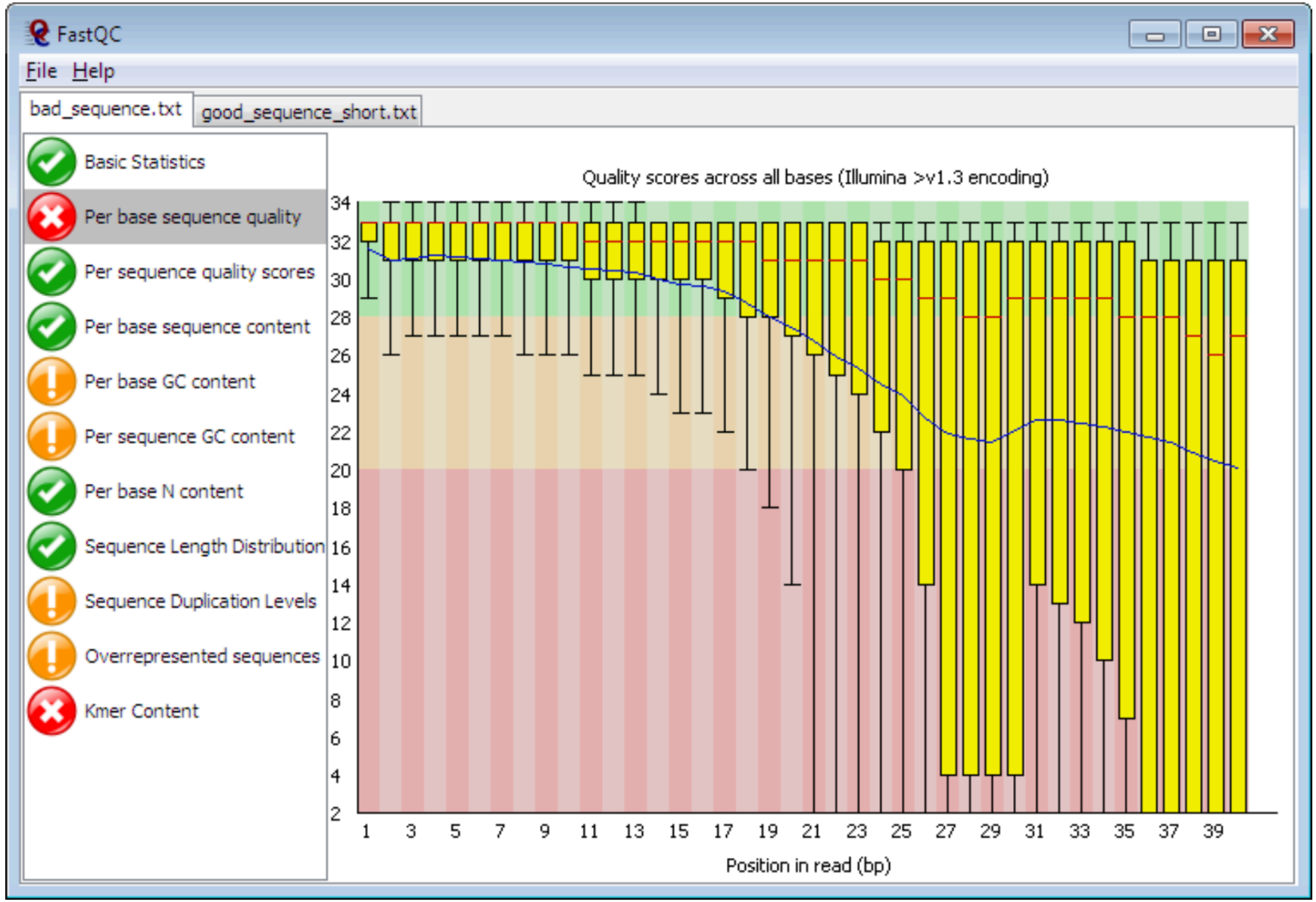
Paired-end Sequences

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

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

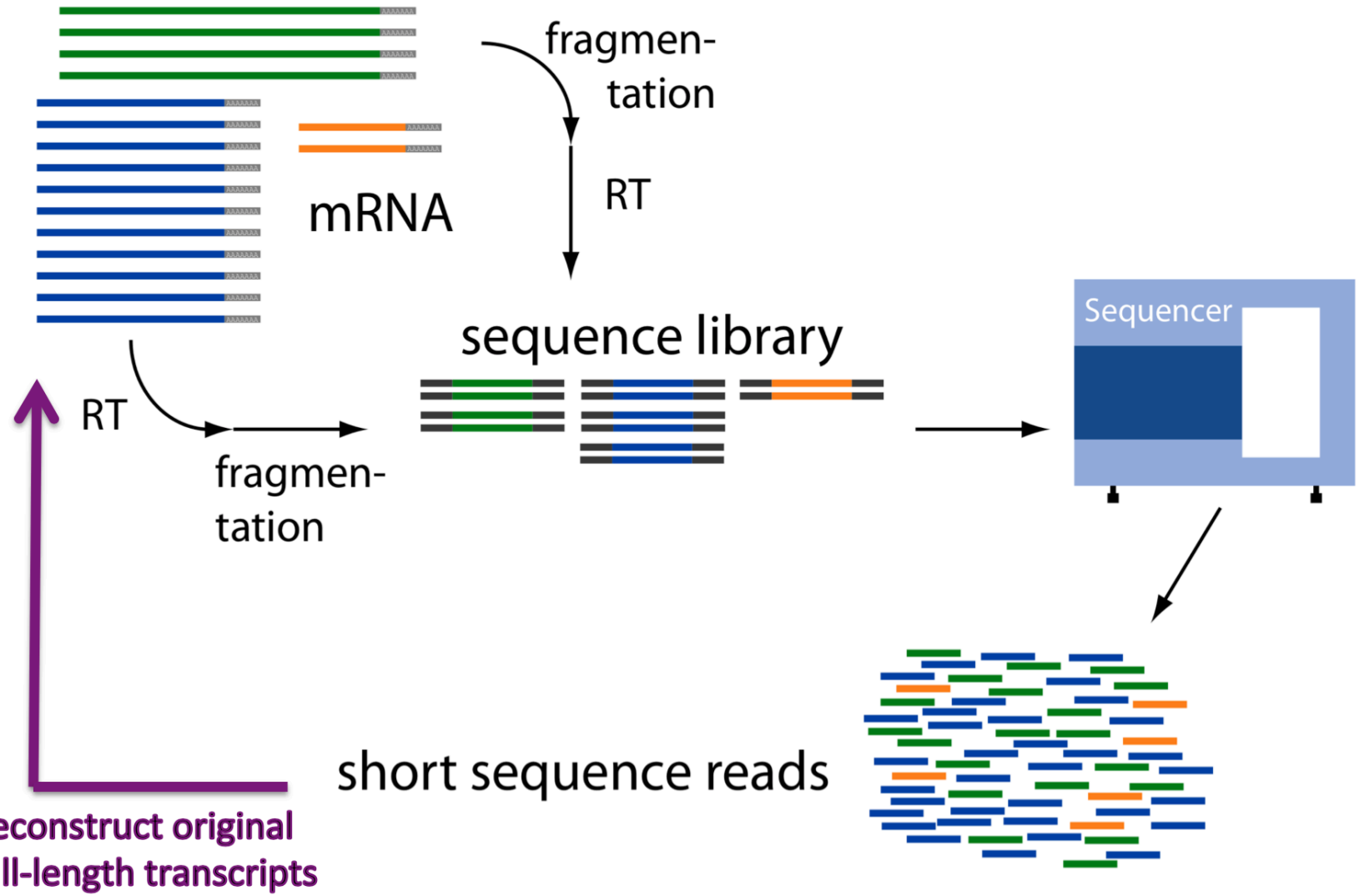
```
@61DFRAAXX100204:1:100:10494:3070/2
CTCAAATGGTTAATTCTCAGGCTGCAAATATTCGTTCAGGATGGAAGAACA
+
C<CCCCCCCACCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCBCCCC
```

Read Quality Assessment



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

Overview of RNA-Seq



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

Transcript Reconstruction from RNA-Seq Reads

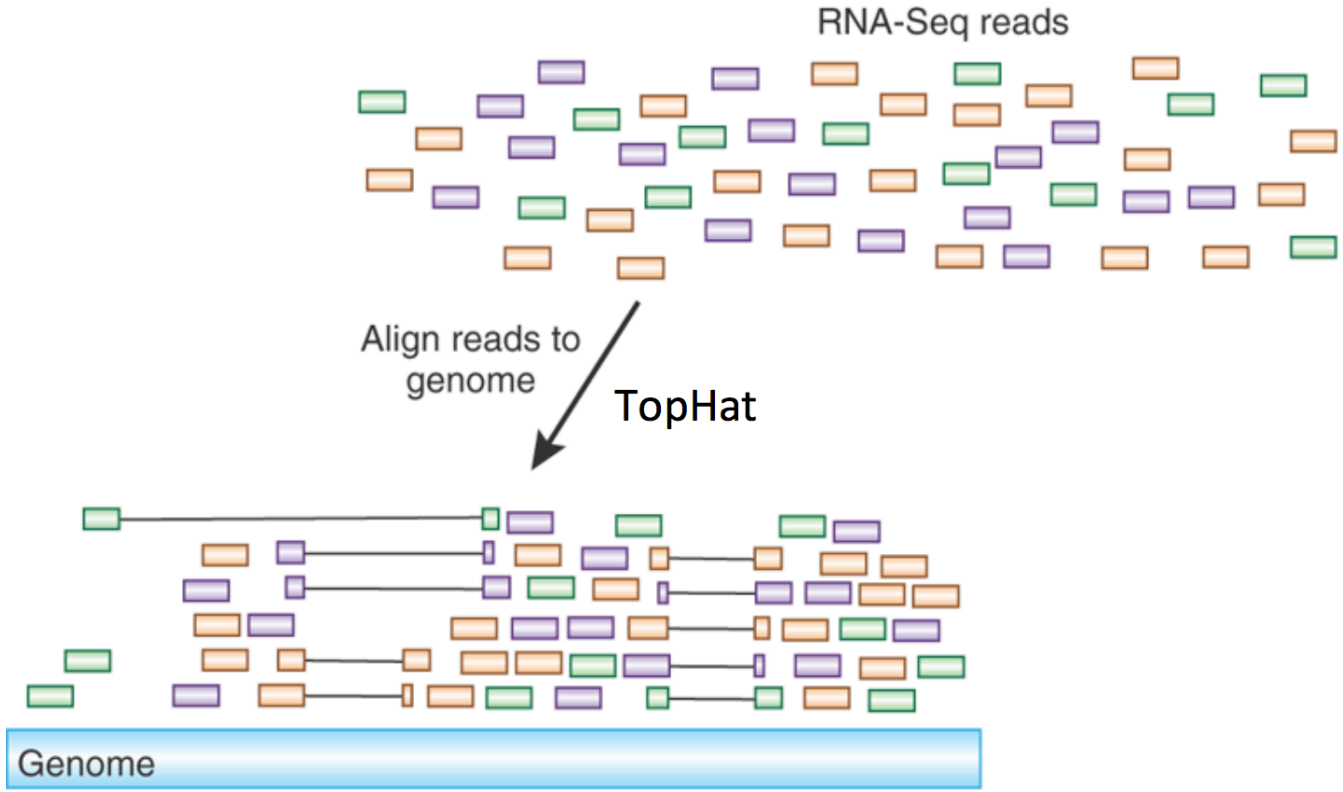
Advancing RNA-Seq analysis

Brian J Haas & Michael C Zody

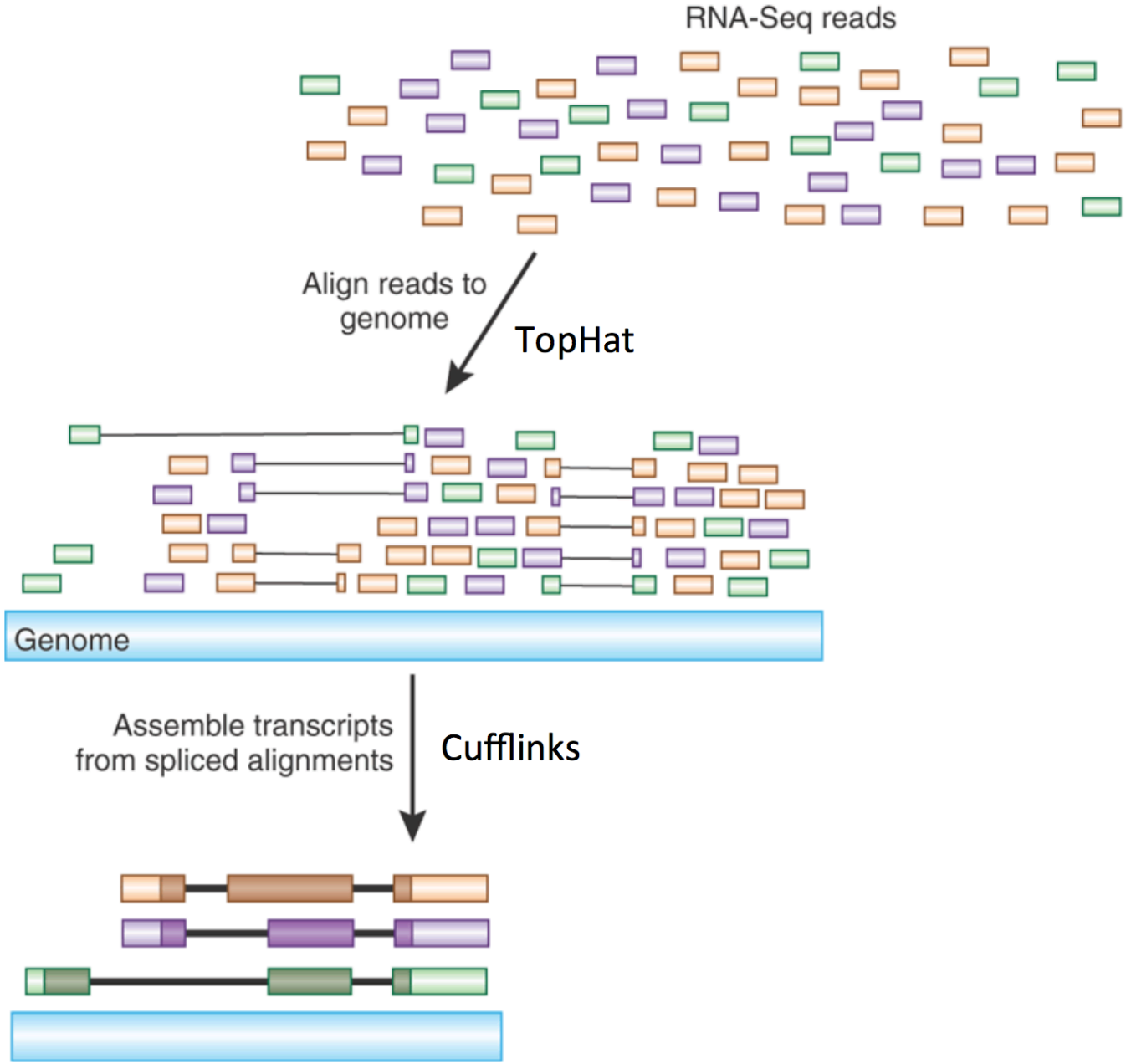
Nature Biotech, 2010

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

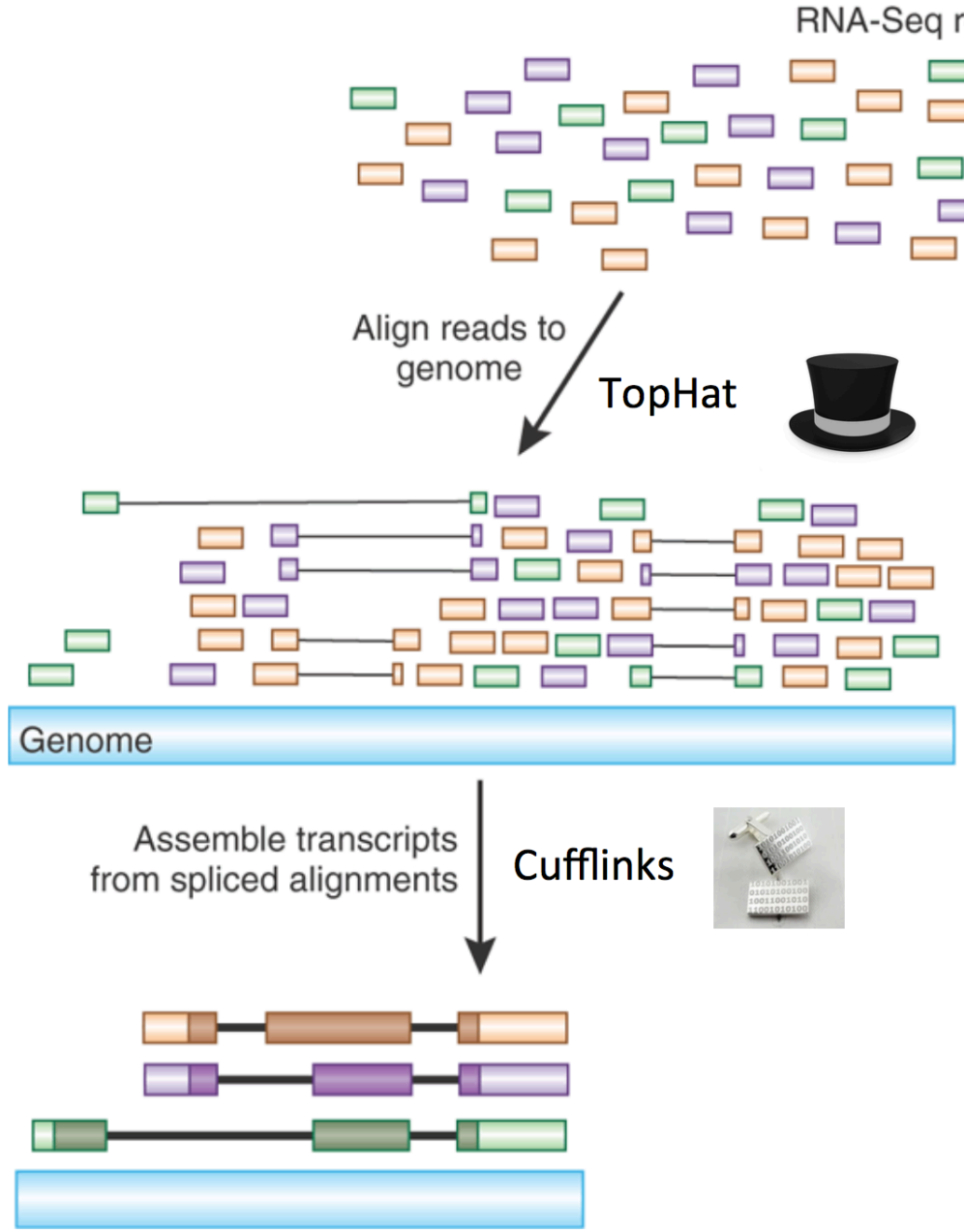
Transcript Reconstruction from RNA-Seq Reads



Transcript Reconstruction from RNA-Seq Reads



Transcript Reconstruction from RNA-Seq Reads



The Tuxedo Suite:

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

NATURE PROTOCOLS | PROTOCOL

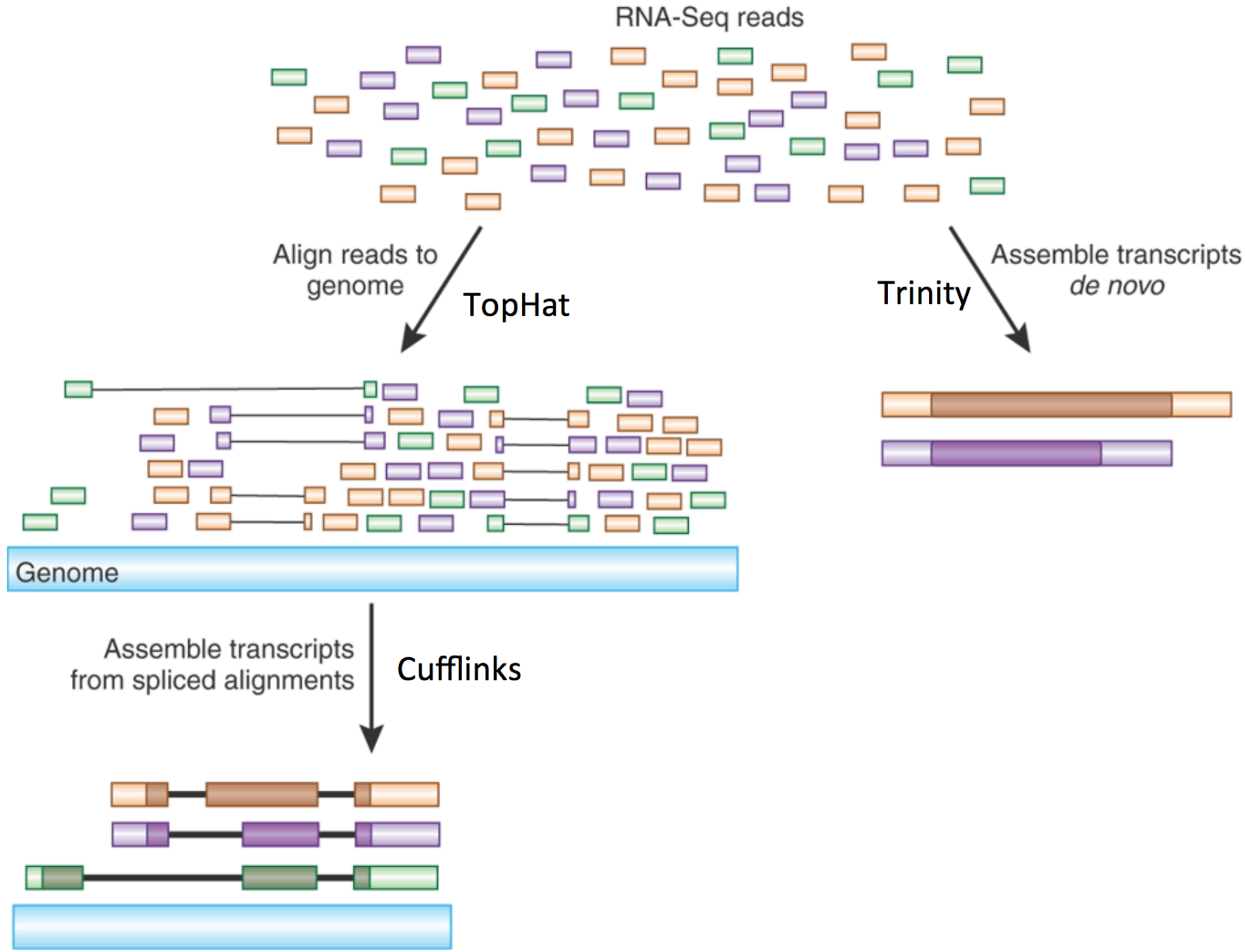
Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks

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

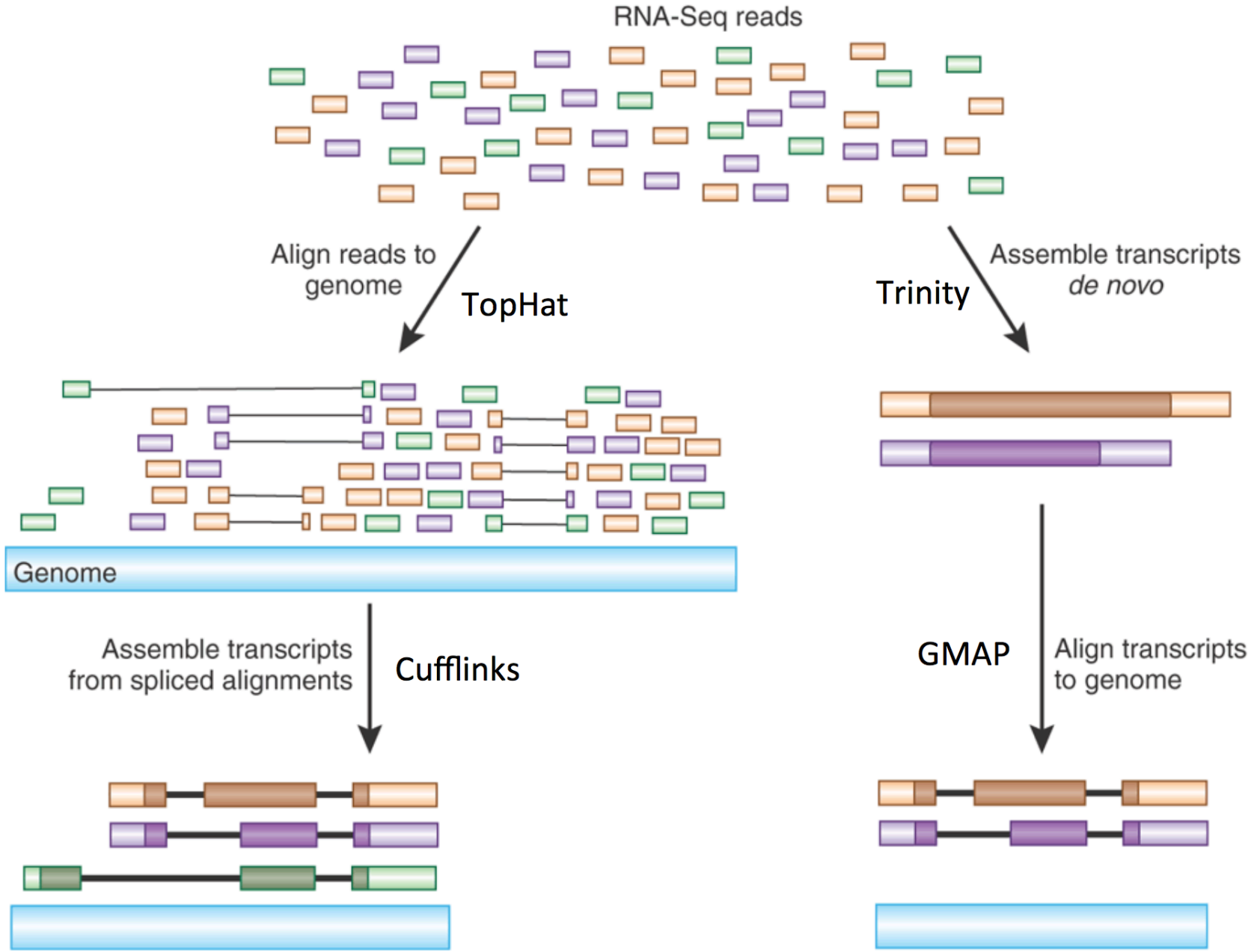
Affiliations | Contributions | Corresponding author

Nature Protocols **7**, 562–578 (2012) | doi:10.1038/nprot.2012.016
Published online 01 March 2012

Transcript Reconstruction from RNA-Seq Reads



Transcript Reconstruction from RNA-Seq Reads



Transcript Reconstruction from RNA-Seq Reads

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

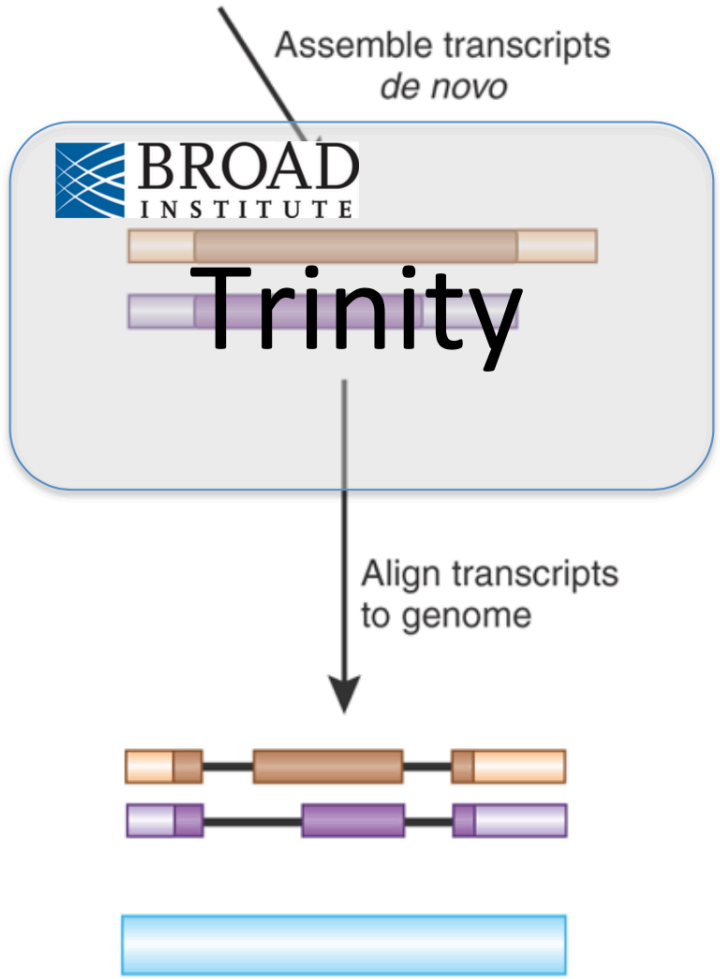
NATURE PROTOCOLS | PROTOCOL

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

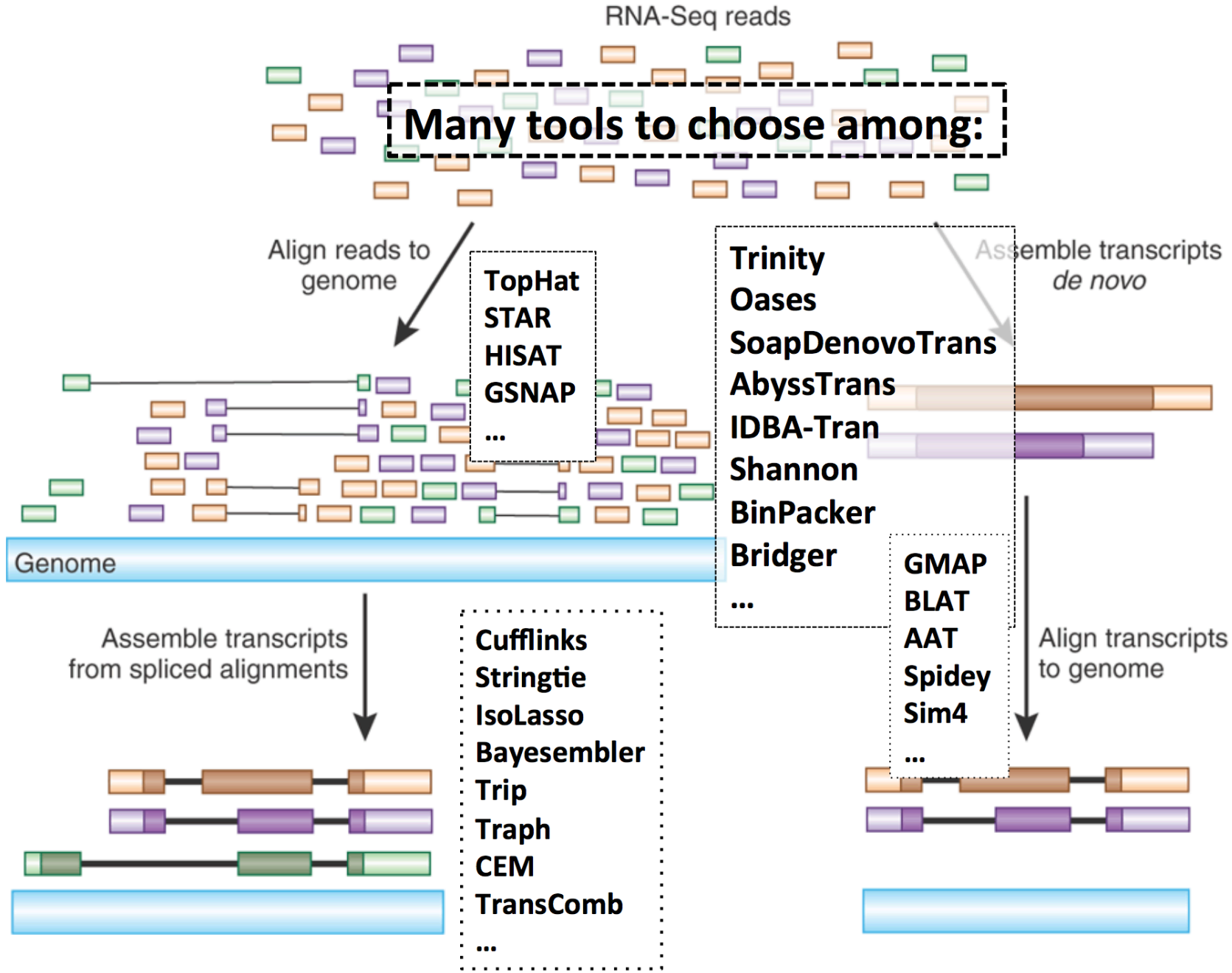
Brian J Haas, Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, Philip D Blood, Joshua Bowden, Matthew Brian Couger, David Eccles, Bo Li, Matthias Lieber, Matthew D MacManes, Michael Ott, Joshua Orvis, Nathalie Pochet, Francesco Strozzi, Nathan Weeks, Rick Westerman, Thomas William, Colin N Dewey,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



Transcript Reconstruction from RNA-Seq Reads

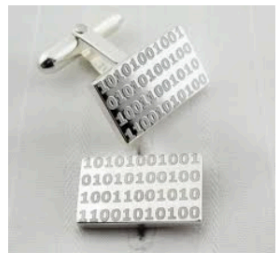


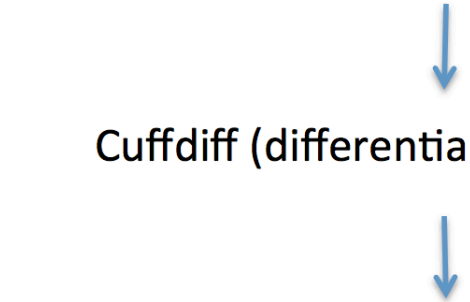
Overview of the Tuxedo Software Suite

Bowtie (fast short-read alignment)

TopHat (spliced short-read alignment)

Cufflinks (transcript reconstruction from alignments)

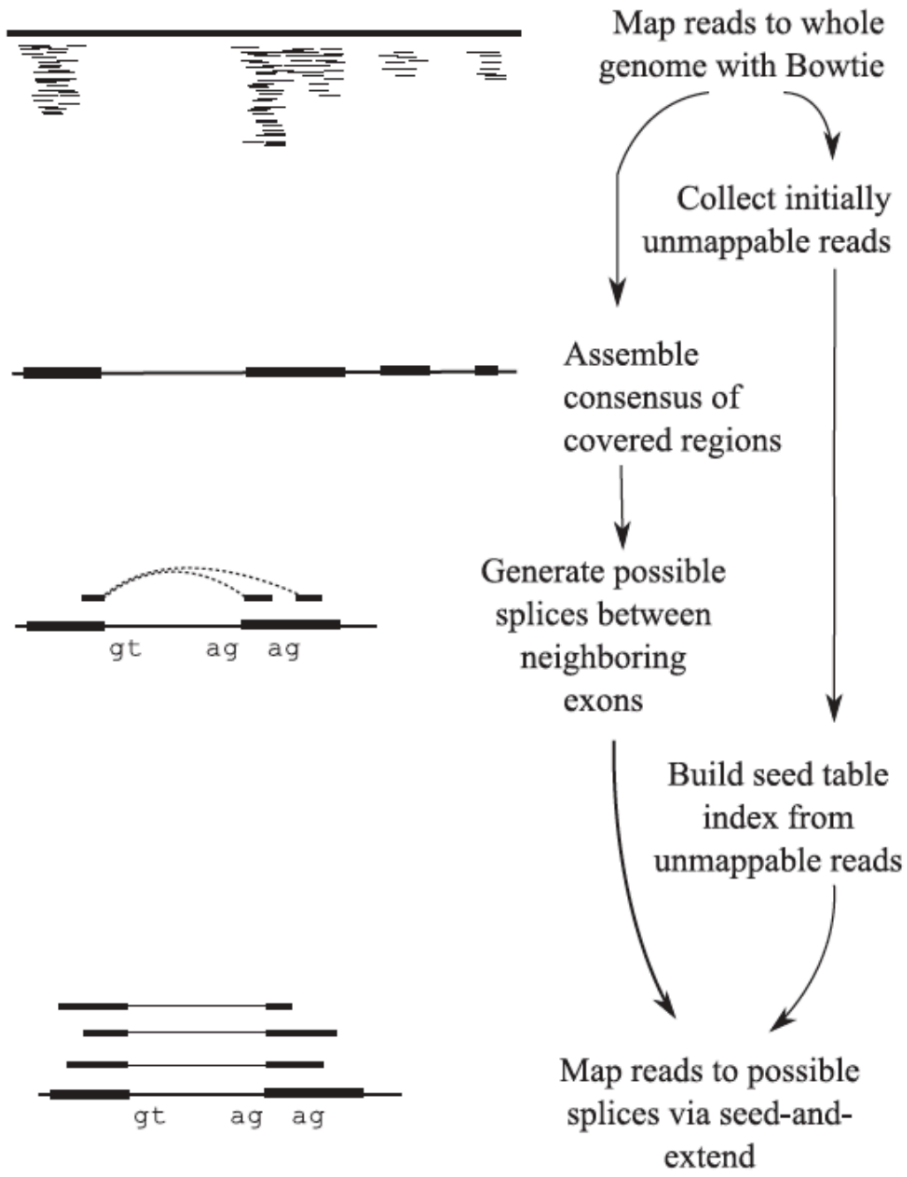




Cuffdiff (differential expression analysis)

CummeRbund (visualization & analysis)

The TopHat Pipeline



From Trapnell, Pachter, & Salzberg. Bioinformatics. 2009

De novo transcriptome assembly

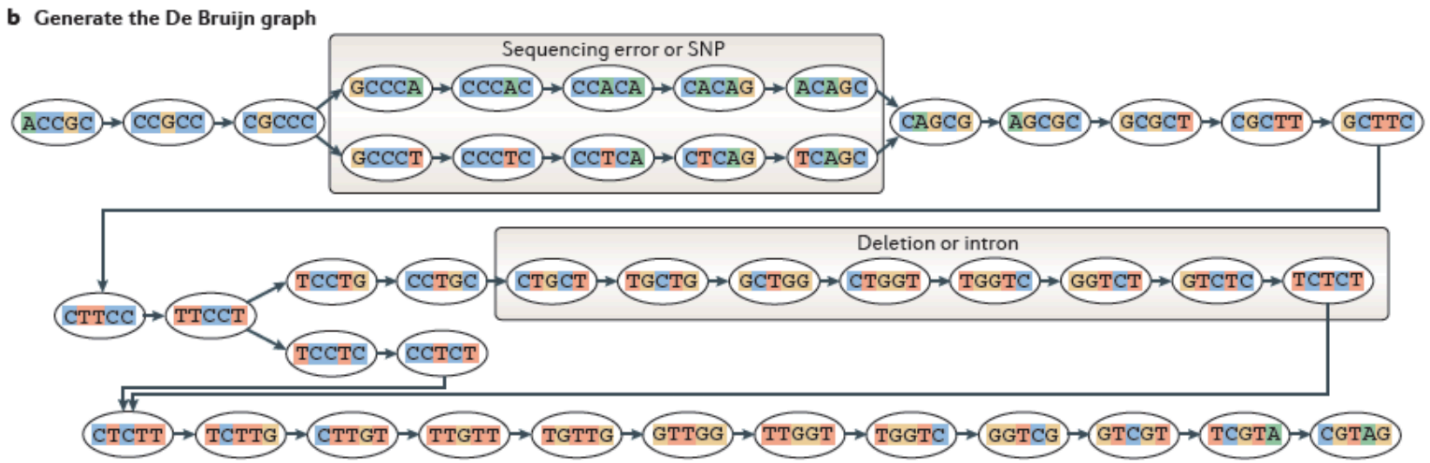
No genome required

Empower studies of non-model organisms

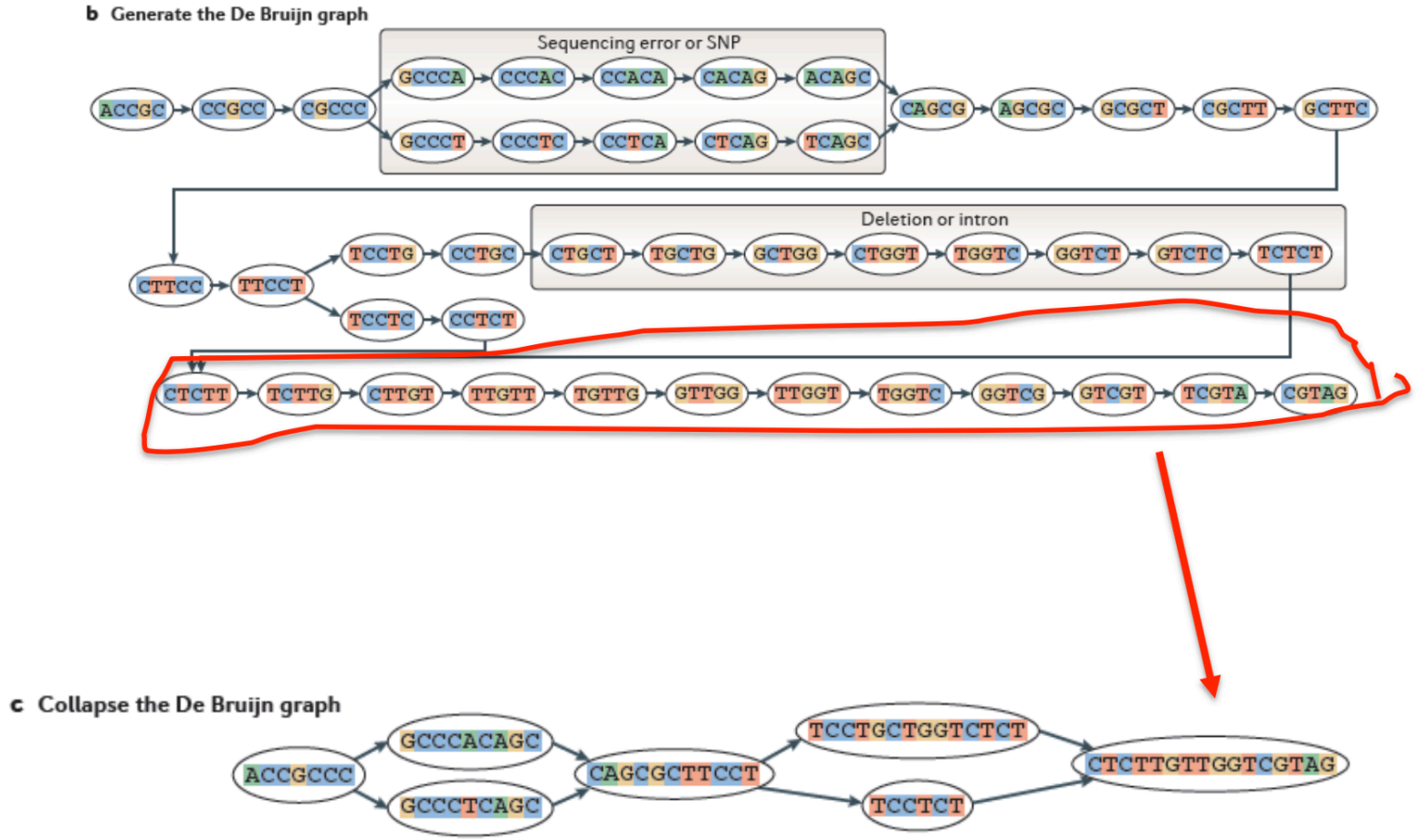
- expressed gene content
- transcript abundance
- differential expression

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

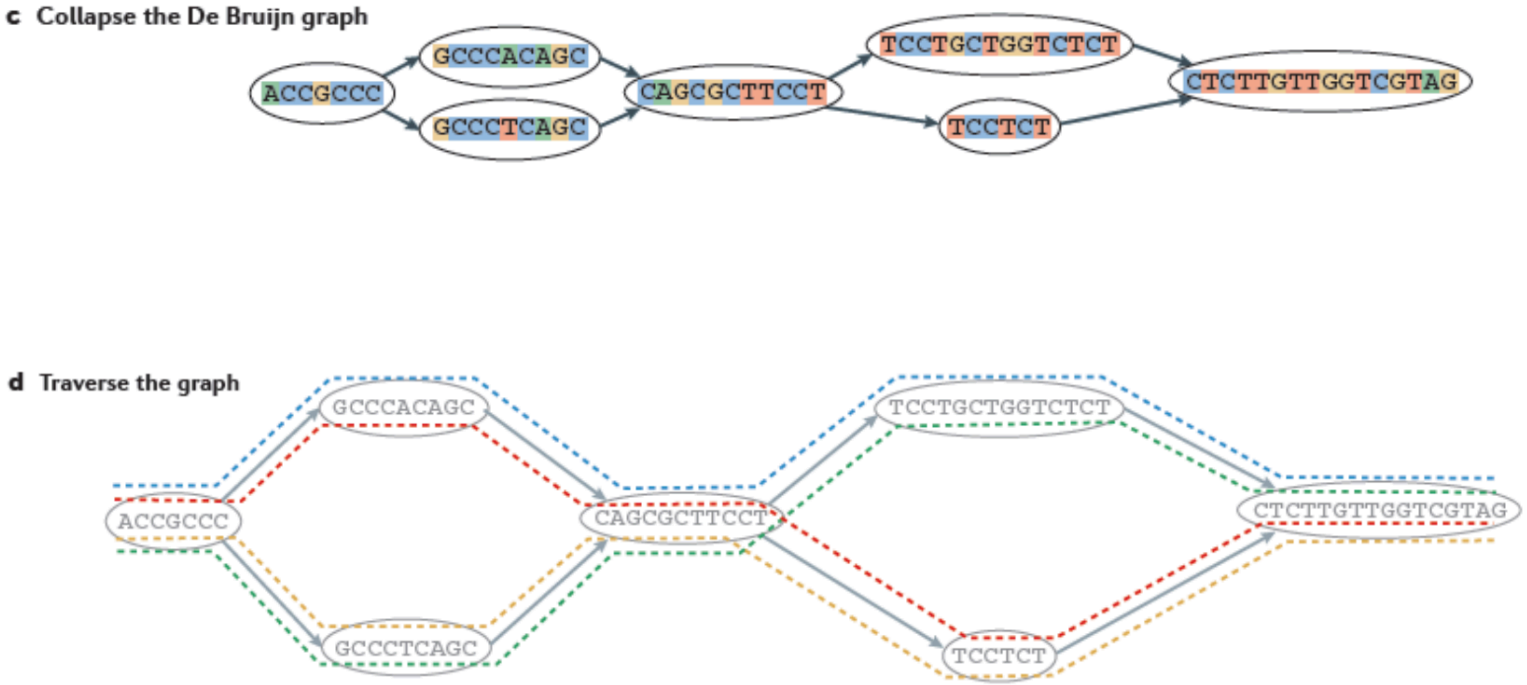
Sequence Assembly via De Bruijn Graphs



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

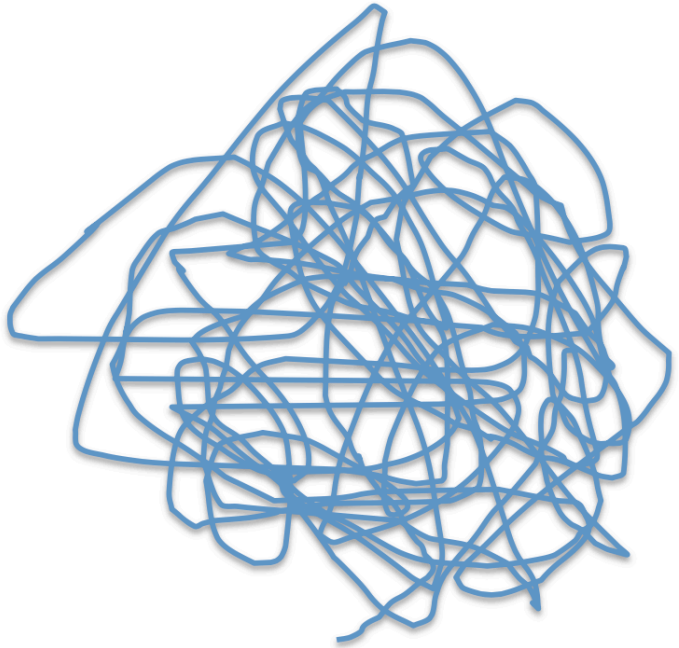
Transcriptome Assembly

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

Trinity Aggregates Isolated Transcript Graphs

Genome Assembly

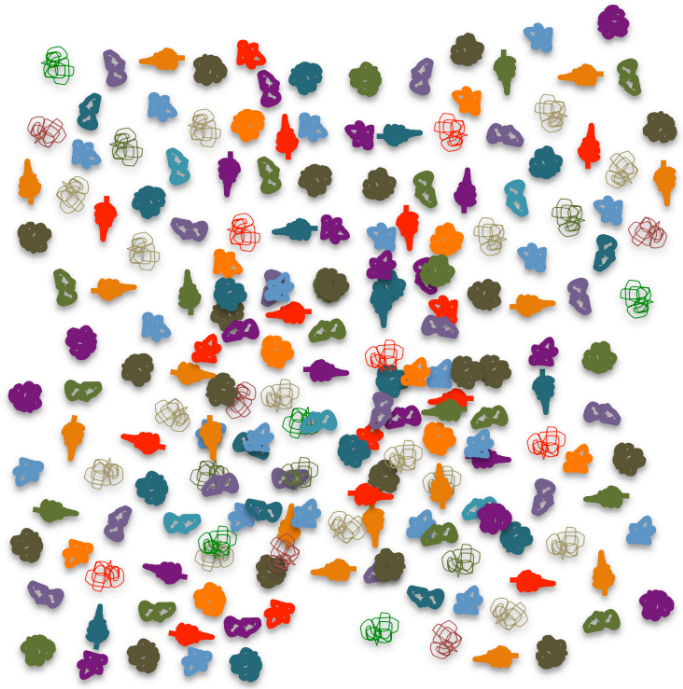
Single Massive Graph



Entire chromosomes represented.

Trinity Transcriptome Assembly

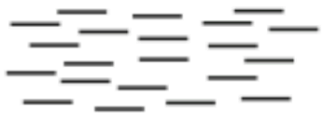
Many Thousands of Small Graphs



Ideally, one graph per expressed gene.

Trinity – How it works:

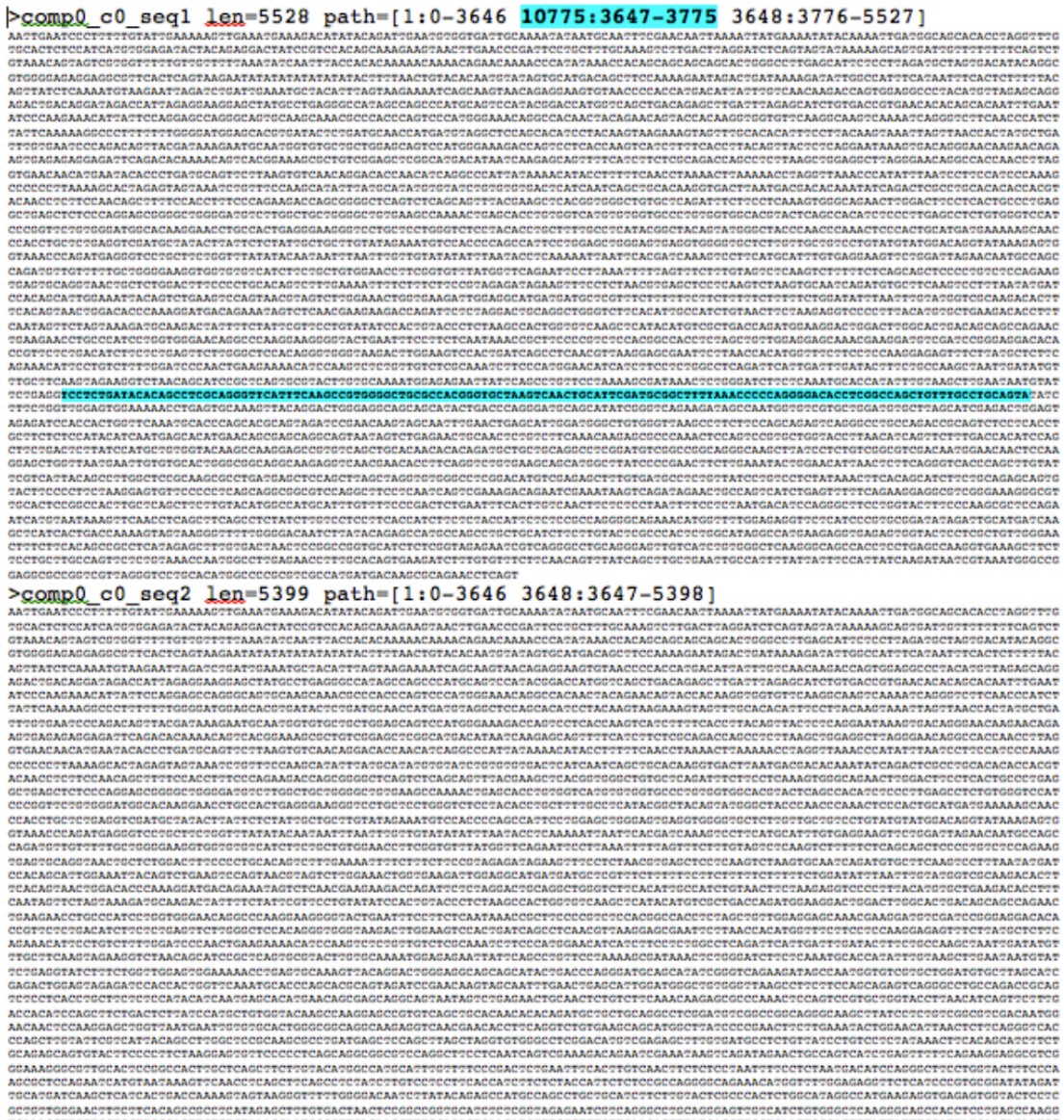
RNA-Seq reads → **Linear contigs** → **de-Bruijn graphs** → **Transcripts + Isoforms**



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

Thousands of disjoint graphs

Trinity output: A multi-fasta file



Alignments are reported in a compact representation: SAM format

0	61G9EAAXX100520:5:100:10095:16477	Read Name
1	83	
2	chr1	Alignment Target
3	51986	Position of Alignment
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)
15      SM:i:38
16      XQ:i:40
17      X2:i:0
```

Still not compact enough…
Millions to billions of reads takes up a lot of space!!

Convert SAM to binary – BAM format.

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

Samtools

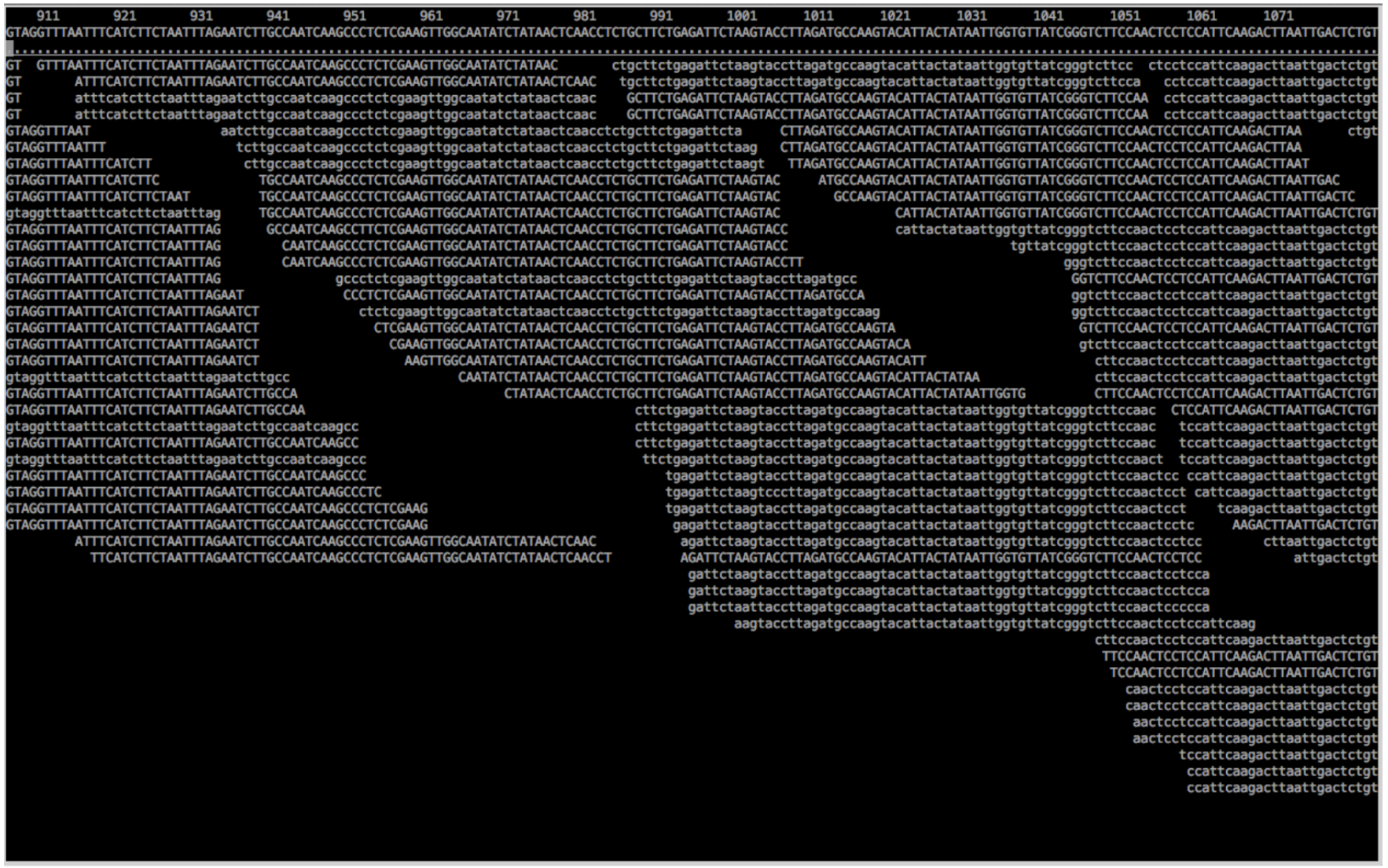
- Tools for
  - converting SAM <-> BAM
  - Viewing BAM files (eg. samtools view file.bam | less)
  - Sorting BAM files, and lots more:

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

Usage:   samtools <command> [options]

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

Visualizing Alignments of RNA-Seq reads

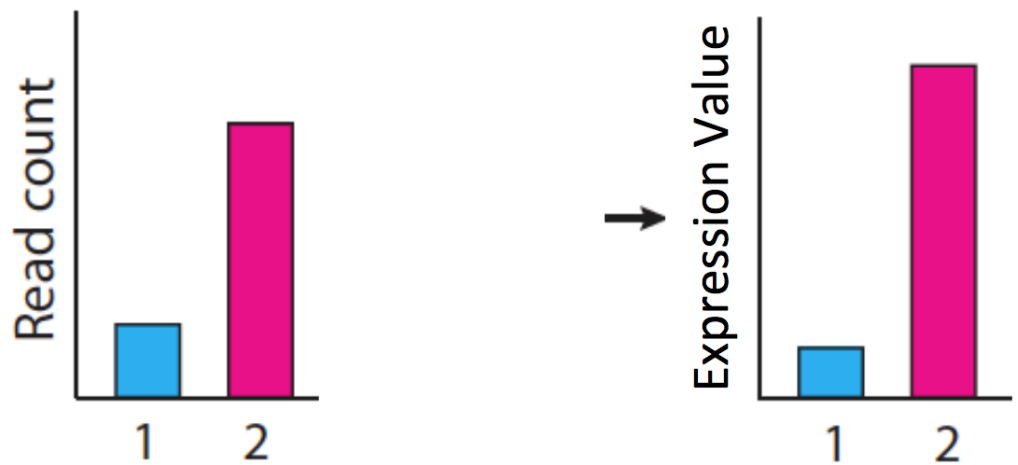
Text-based Alignment Viewer

% samtools tview alignments.bam target.fasta



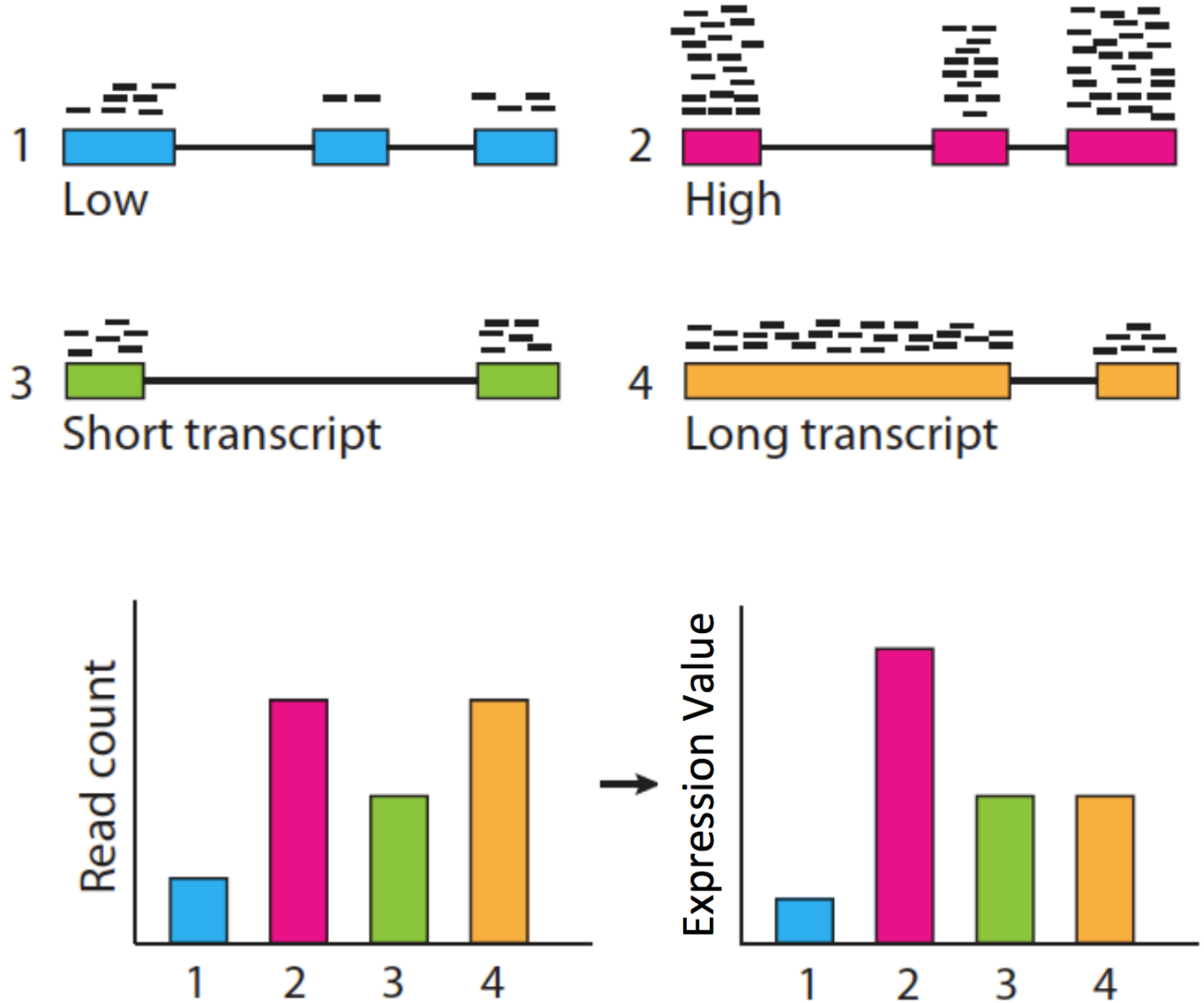
Abundance Estimation
(Aka. Computing Expression Values)

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

Transcripts per Million (TPM)

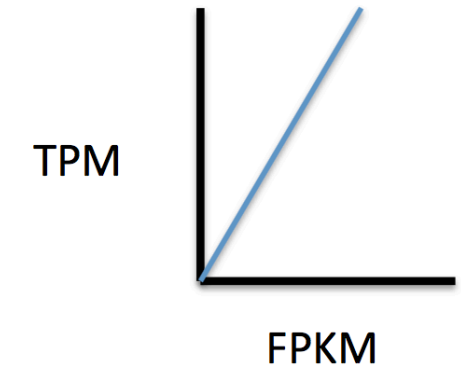
$$TPM_i = \frac{FPKM_i}{\sum_j FPKM} * 1e6$$

Preferred metric for measuring expression

- Better reflects transcript concentration in the sample.
- Nicely sums to 1 million

Linear relationship between TPM and FPKM values.

Both are valid metrics, but best to be consistent.



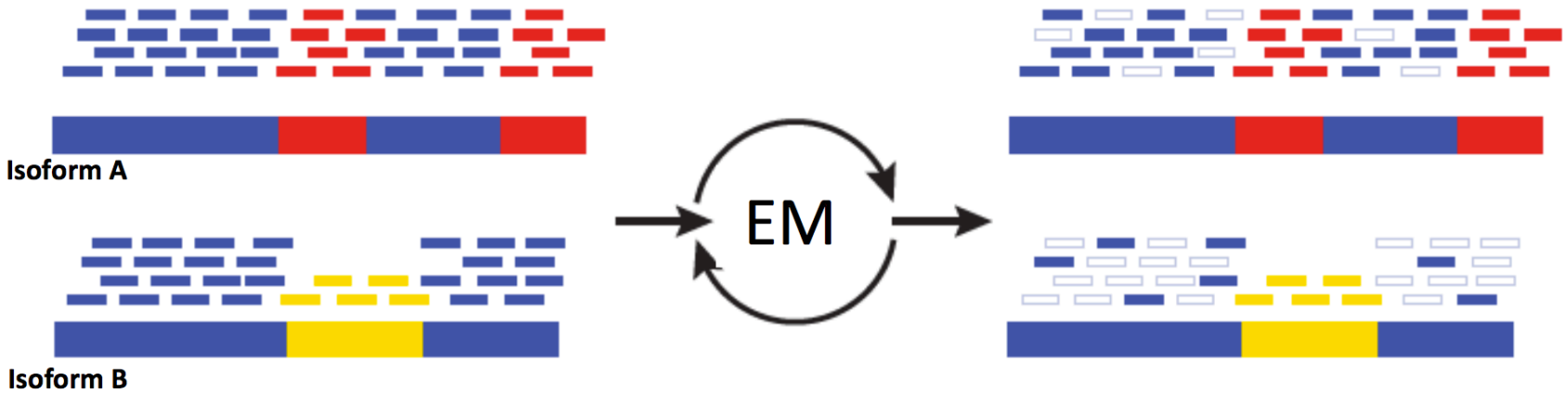
Multiply-mapped Reads Confound Abundance Estimation

Isoform A

Isoform B

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

Multiply-mapped Reads Confound Abundance Estimation



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

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

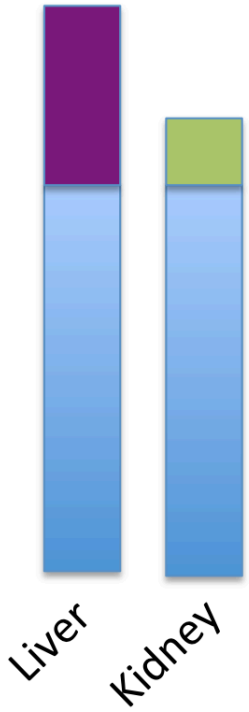
Performed by:
RSEM, eXpress, kallisto, salmon, …

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

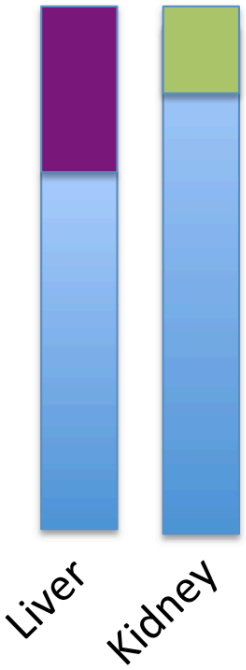
Differential Expression Analysis Using RNA-Seq

Why cross-sample normalization is important

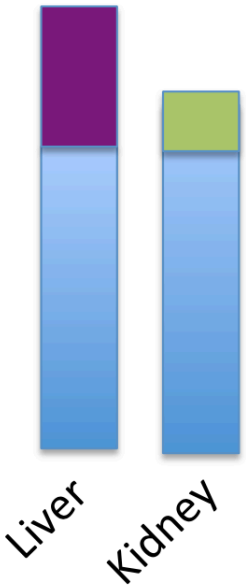
Absolute RNA quantities per cell



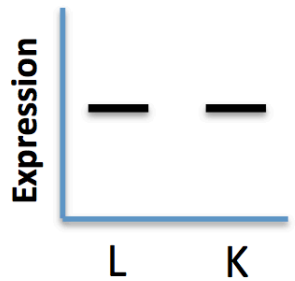
Measured relative abundance via RNA-Seq

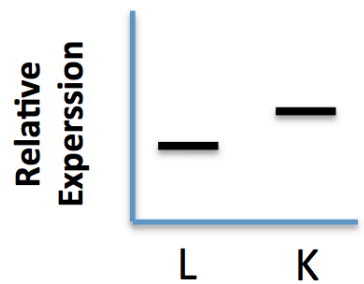


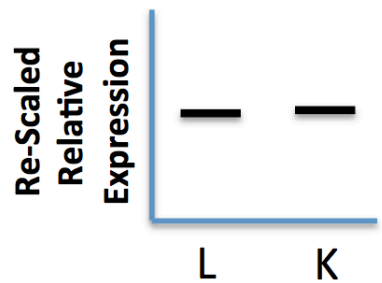
Cross-sample normalized (rescaled) relative abundance



eg. Some housekeeping gene's expression level:







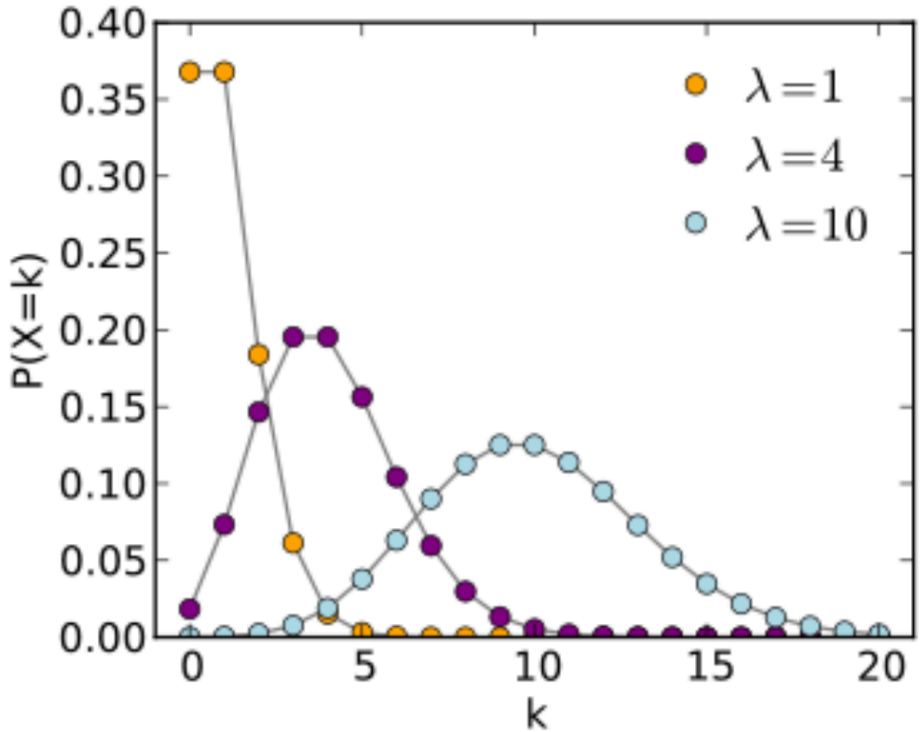
Diff. Expression Analysis Involves

- 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

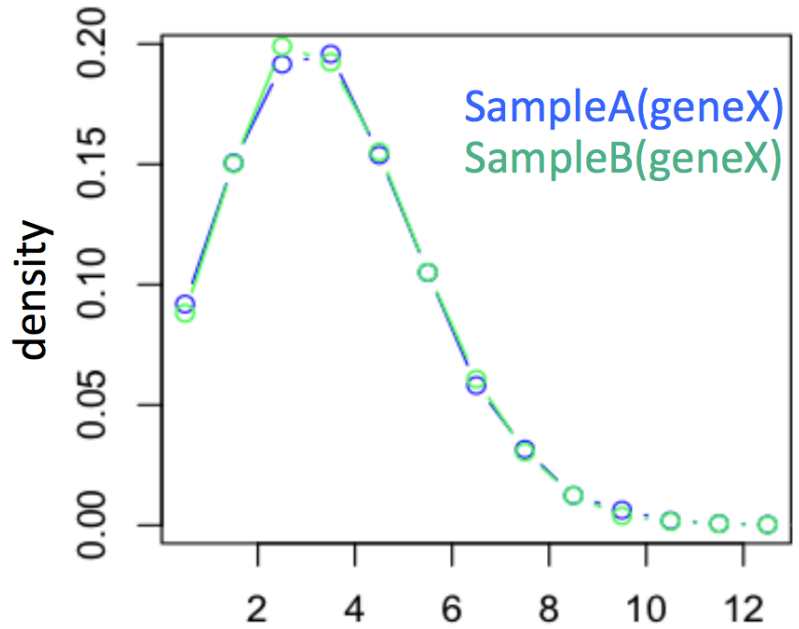
(observed read counts)

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

Example: One gene*not* differentially expressed

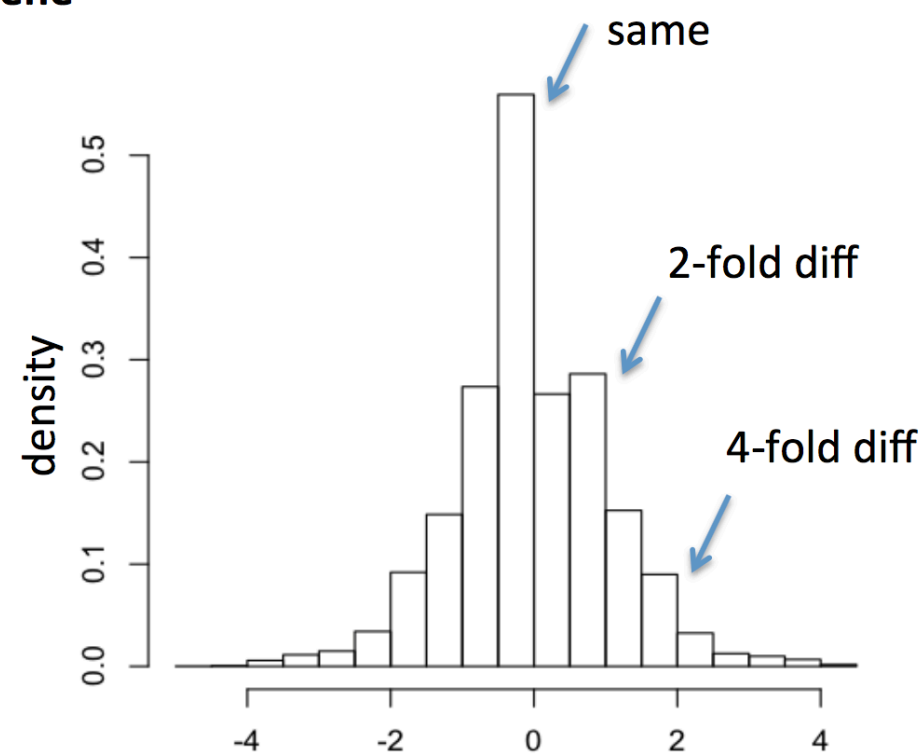
SampleA(gene) = SampleB(gene) = 4 reads

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



(k) number of reads observed

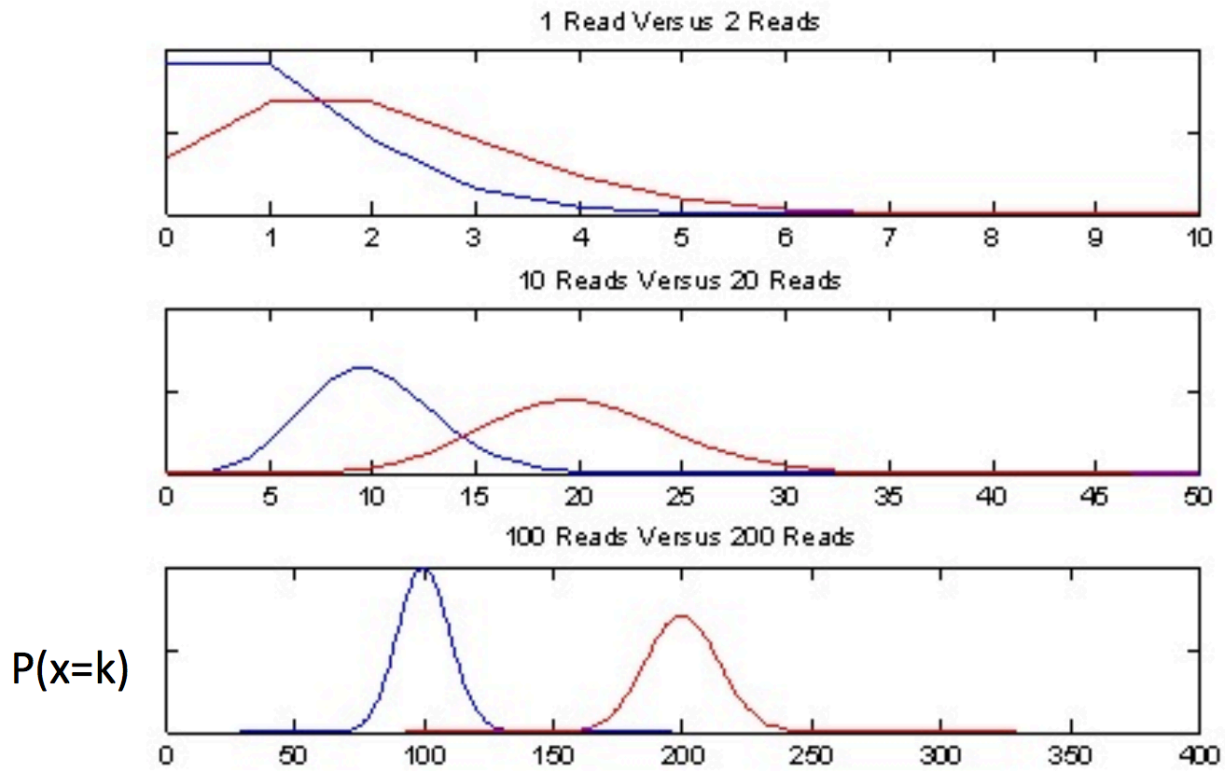
Dist. of $\log_2$(fold change) values



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

edgeR
ShrinkSeq
DESeq
baySeq
Vsf
Limma/Voom
mmdiff
cuffdiff

ROTS
TSPM
DESeq2
EBSeq
NBPSeq
SAMseq
NoiSeq

(italicized not in R/Bioconductor but stand-alone)

See: http://www.biomedcentral.com/1471-2105/14/91
A comparison of methods for differential expression analysis of RNA-seq data
Soneson & Delorenzi, 2013

Typical output from DE analysis

	logFC	logCPM	PValue	FDR
TRINITY_DN876_c0_g1_i1	-7.15049572793027	10.6197708379285	0	0
TRINITY_DN6470_c0_g1_i1	-7.26777912190146	7.03987604865422	1.687485656951e-287	6.46813252309319e-284
TRINITY_DN5186_c0_g1_i1	-7.85623682454322	9.18570464327063	1.17049180235068e-278	2.99099671894011e-275
TRINITY_DN768_c0_g1_i1	7.72884741150304	9.7514619195169	4.32504881419265e-272	8.28895605240022e-269
TRINITY_DN70_c0_g1_i1	-12.7646078189688	7.86482982471445	3.92853491279431e-253	6.02322972829624e-250
TRINITY_DN1587_c0_g1_i1	-5.89392061881667	9.07366563894607	6.32919557933429e-243	8.08660221852944e-240
TRINITY_DN3236_c0_g1_i1	-7.27029815068473	8.02209568234202	3.64955175271959e-235	3.99678053376405e-232
TRINITY_DN4631_c0_g1_i1	-7.45310693639574	6.91664918183241	4.30540921272851e-229	4.1256583780971e-226
TRINITY_DN5082_c0_g5_i1	-5.33154406167545	10.6977538760467	2.74243356676259e-225	2.33594396920022e-222
TRINITY_DN1789_c0_g3_i1	10.2032564835076	7.32607652700285	1.44273728647186e-213	1.10600240380933e-210
TRINITY_DN4204_c0_g1_i1	4.81030233739325	9.88844409410644	9.27180216086162e-205	6.46160321501501e-202
TRINITY_DN799_c0_g1_i1	-4.22044475626154	6.9937398638711	1.24746518421083e-197	7.96922341846683e-195
TRINITY_DN196_c0_g2_i1	4.60597918494257	9.86878463857276	1.9819997623131e-192	1.16877001368402e-189
TRINITY_DN5041_c0_g1_i1	-4.27126549355785	9.70894399883	1.8930437900069e-185	1.03657669244235e-182
TRINITY_DN1619_c0_g1_i1	-4.47156415953777	9.22535948721718	1.76766063029526e-181	9.03392426122899e-179
TRINITY_DN899_c0_g1_i1	-4.90914328409143	7.93768691394594	1.11054513767547e-180	5.32089939088761e-178
TRINITY_DN324_c0_g2_i1	4.87160837667488	6.84850312231775	2.20092562166991e-179	9.92487989160089e-177
TRINITY_DN3241_c0_g1_i1	-4.77760618069256	7.94111259715689	1.60585457735621e-173	6.83915621667372e-171
TRINITY_DN4379_c0_g1_i1	3.85133572453294	7.23712813663389	3.48140532848425e-164	1.4046554341137e-161
TRINITY_DN1919_c0_g1_i1	4.05998814332136	6.95937301668582	1.8588621194715e-161	7.12501850393425e-159
TRINITY_DN2504_c0_g1_i1	-6.92417817059644	6.20370039359785	2.42022459856956e-160	8.83497227268296e-158
…				

Up vs. Down regulated

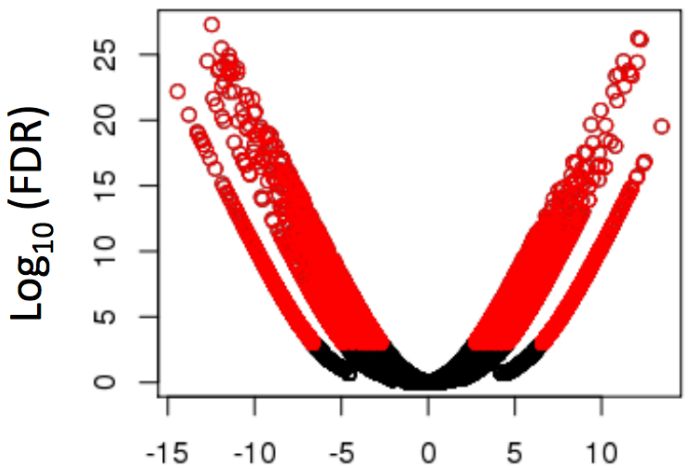
Avg. expression level

Significance

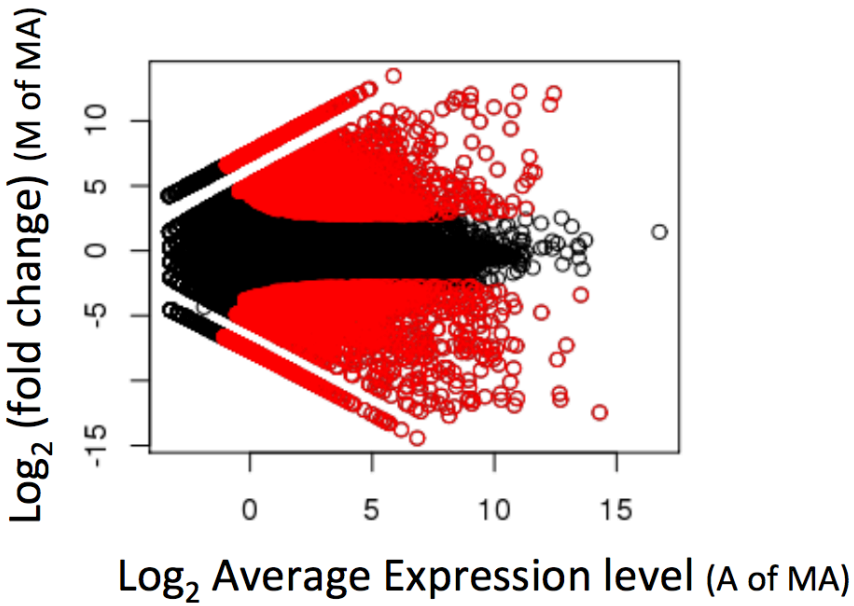
Visualization of DE results and Expression Profiling

Plotting Pairwise Differential Expression Data

Volcano plot
(fold change vs. significance)

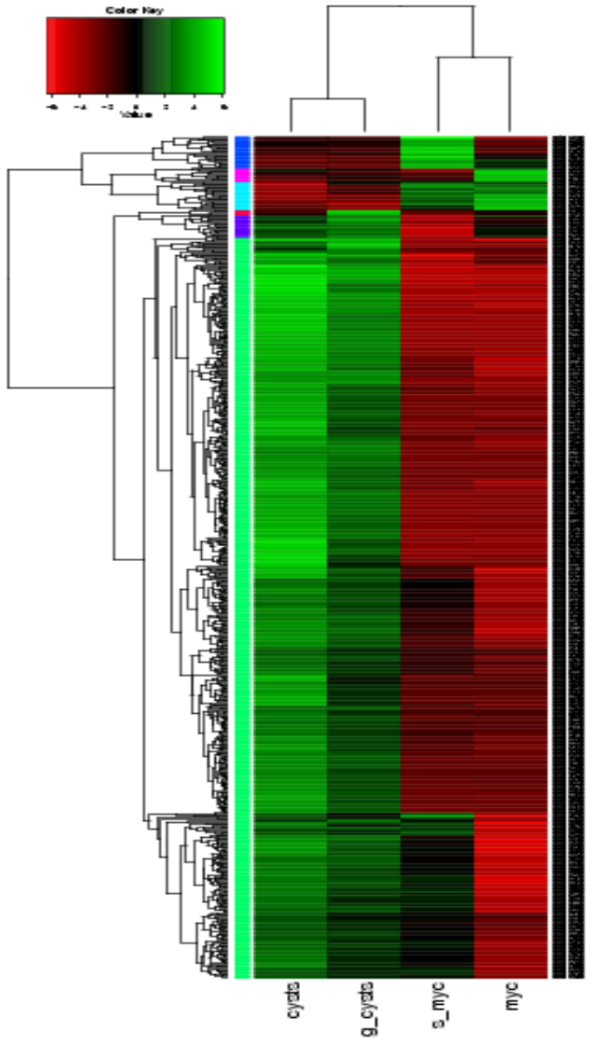


MA plot
(abundance vs. fold change)



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

Comparing Multiple Samples



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

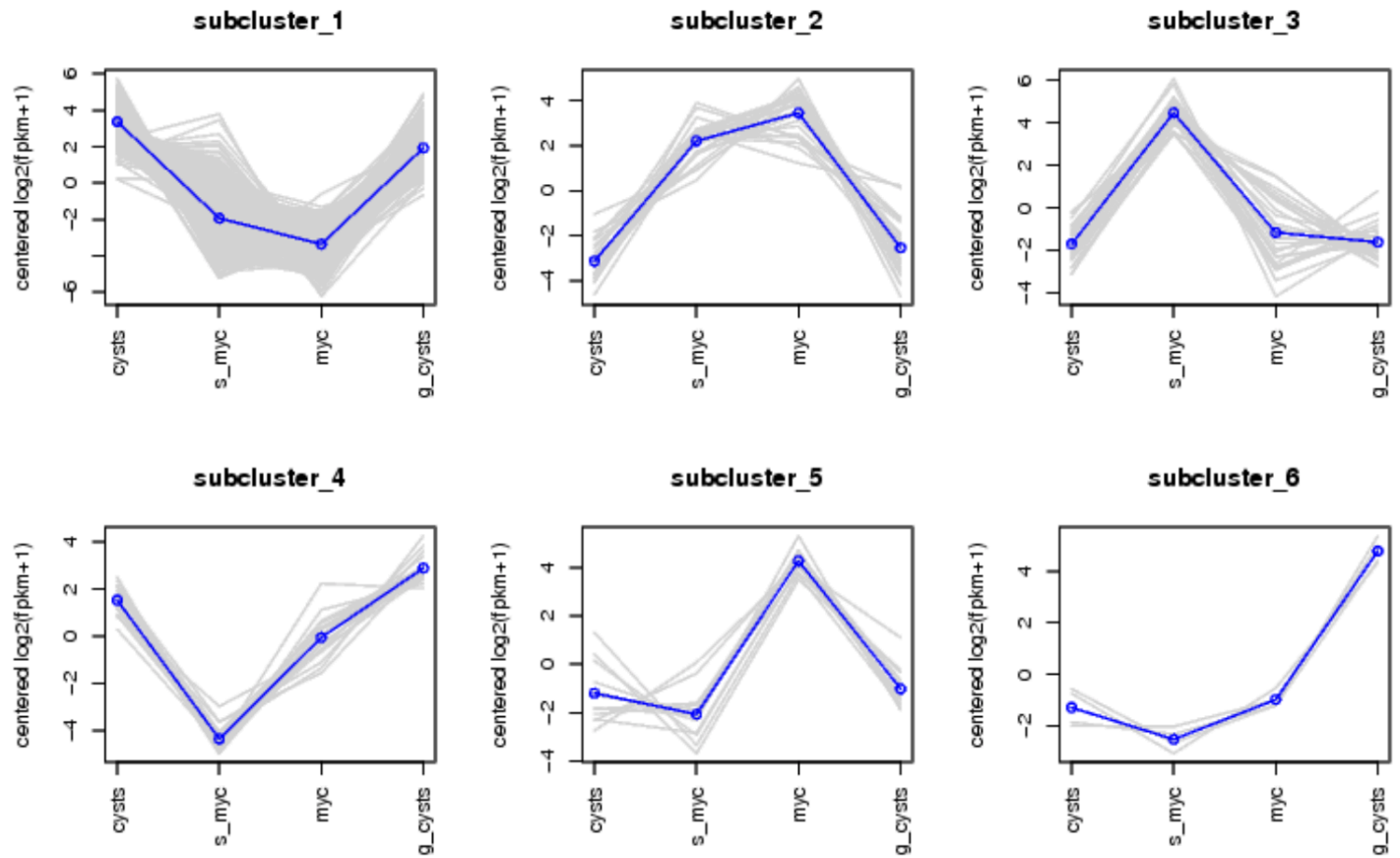
Clustering can be performed across both axes:

-cluster transcripts with similar expression patters.

-cluster samples according to similar expression values among transcripts.

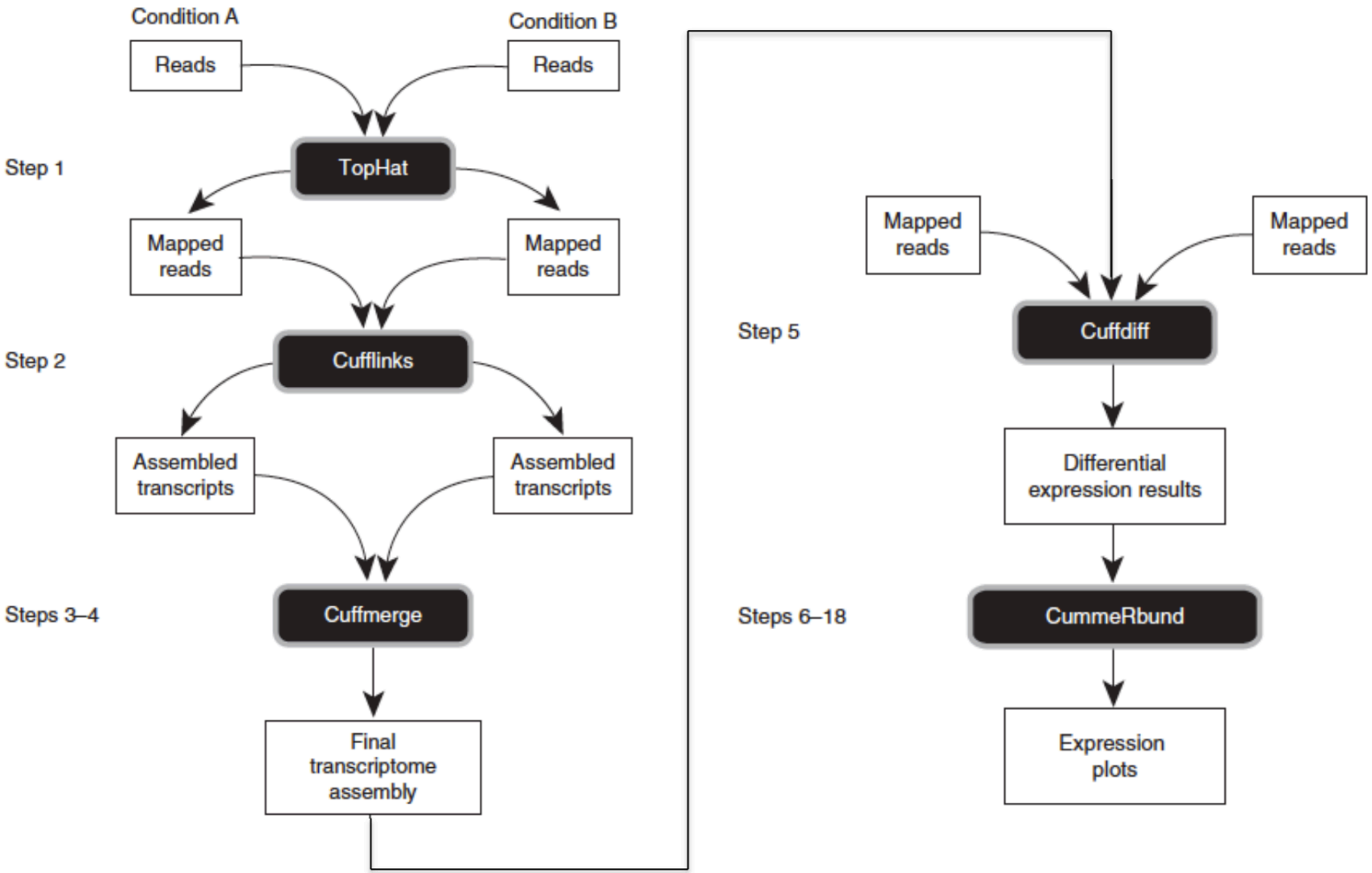
Examining Patterns of Expression Across Samples

Can extract clusters of transcripts and examine them separately.



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.

nature protocols

NATURE PROTOCOLS | PROTOCOL

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

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

Affiliations | Contributions | Corresponding authors

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