BROAD INSTITUTE

Cambridge, Massachusetts, USA.

Leveraging Trinity for *de novo* transcriptome assembly and analysis

CSHL Workshop

Brian Haas

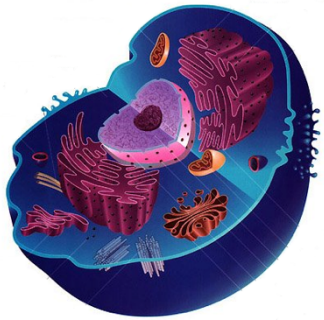
Broad Institute

Next-gen Sequencing Transforming Modern Science

Molecular Biology of the Cell

Chromatin structure
- Histone occupancy
- Transcription factor binding
- DNA 3D topology

Genes and transcripts
- gene content
- alternative splicing
- expression
- RNA-editing



Evolution

Population Genetics

Sequencing Methods

DNA-Seq
ChIP-Seq
RNA-Seq
Methyl-Seq

Algorithms and Software Tools

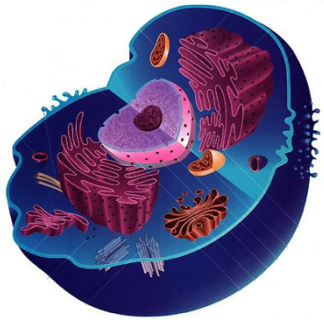
Sequence assembly
Polymorphism calling
Gene finding and
transcript reconstruction
Differential Expression analysis

RNA-Seq as a Driving Technology

Molecular Biology of the Cell

Chromatin structure
- Histone occupancy
- Transcription factor binding
- DNA 3D topology

Genes and transcripts
- gene content
- alternative splicing
- expression
- RNA-editing



Evolution

Population Genetics

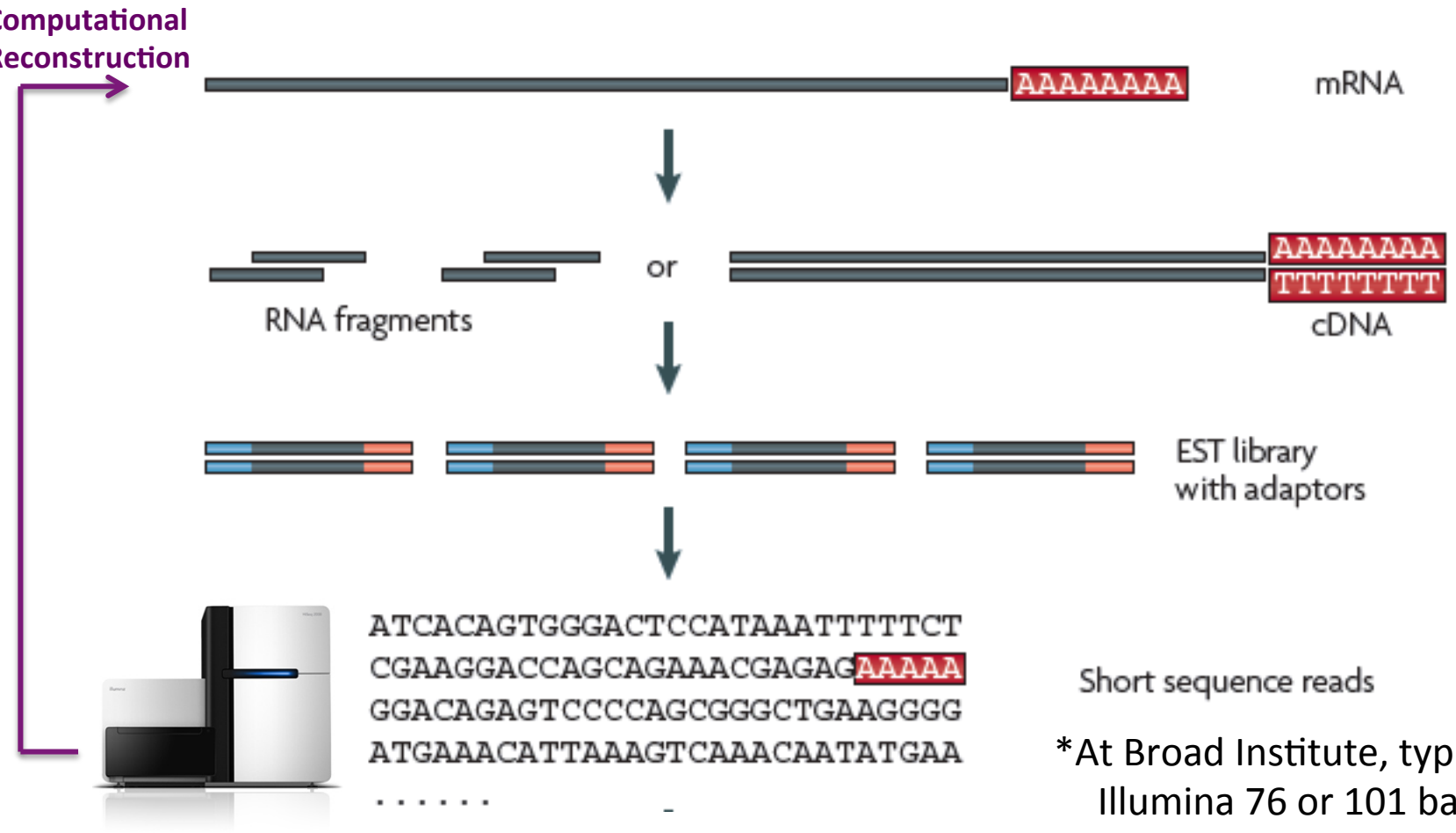
Sequencing Methods

DNA-Seq
ChIP-Seq
RNA-Seq
Methyl-Seq

Algorithms and Software Tools

Sequence assembly
Polymorphism calling
Gene finding and
transcript reconstruction
Differential Expression analysis

RNA-Sequencing Methodology



*At Broad Institute, typically Illumina 76 or 101 base reads as paired ends

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

Transcript Reconstruction from RNA-Seq Reads

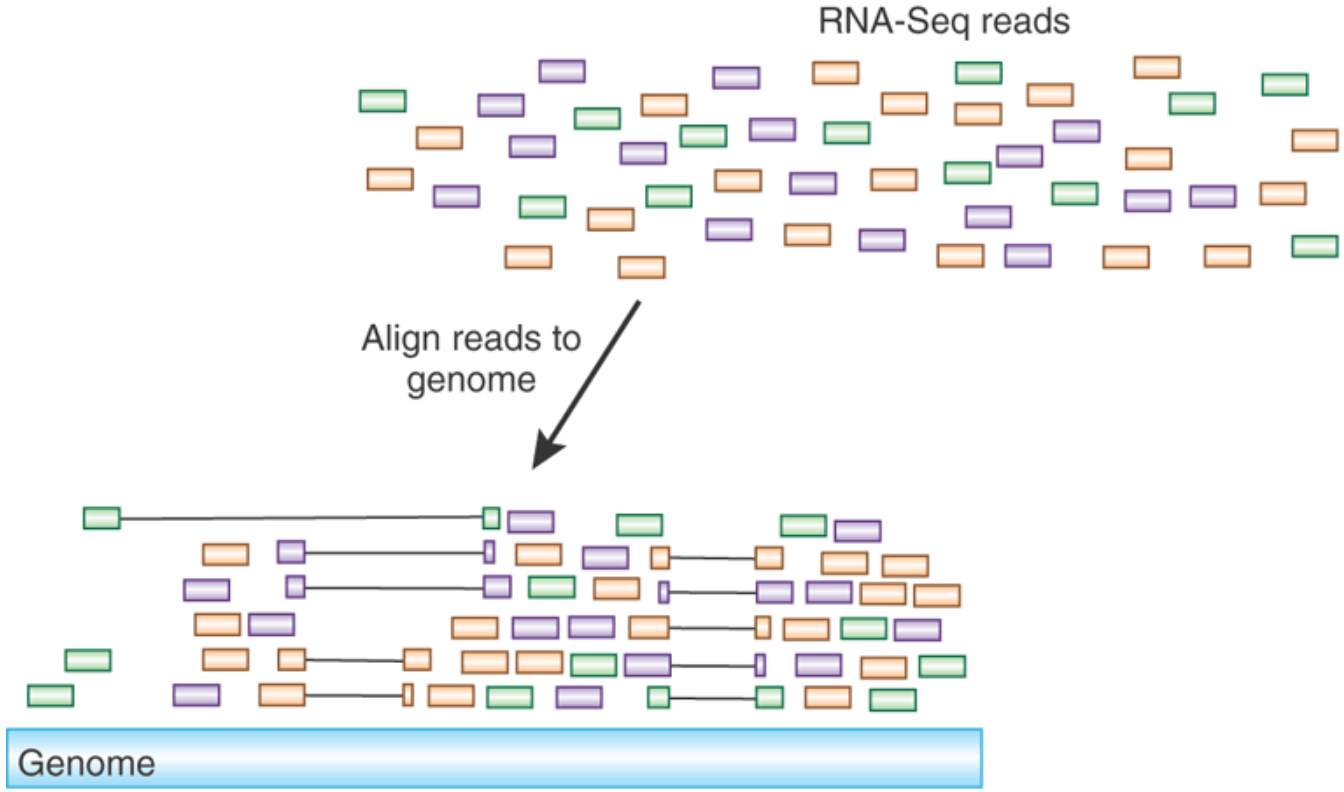
Advancing RNA-Seq analysis

Brian J Haas & Michael C Zody

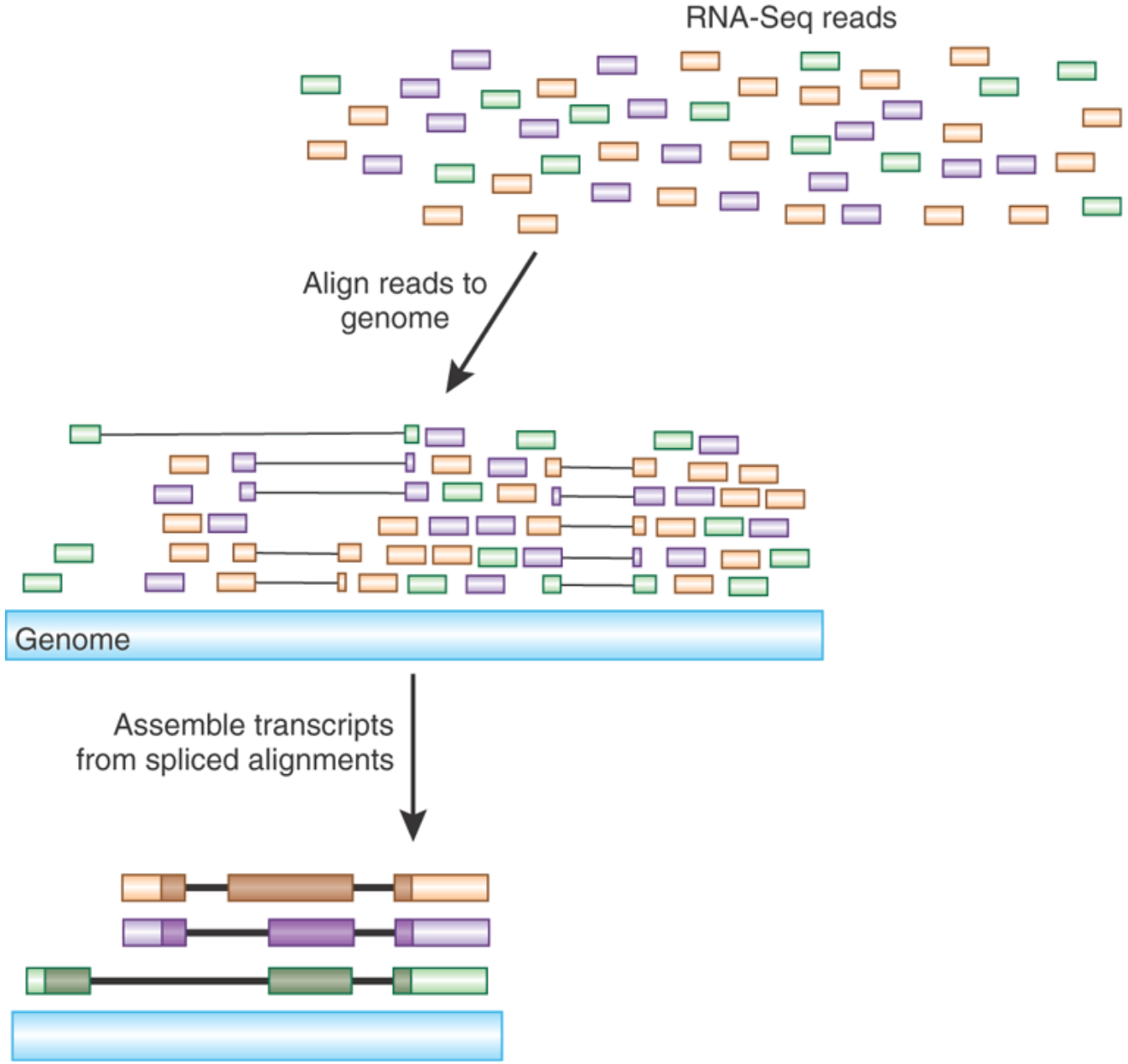
Nature Biotech, 2010

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

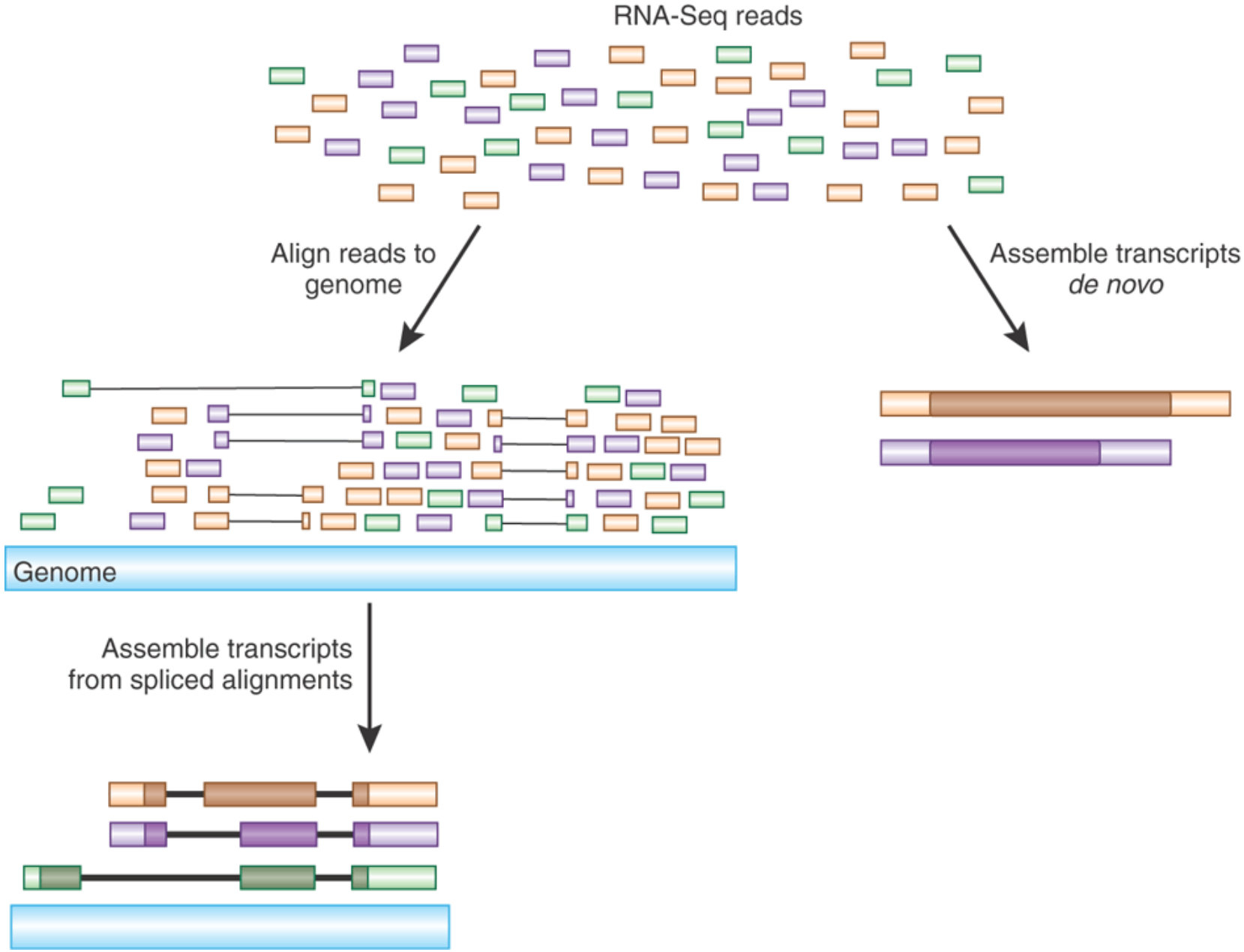
Transcript Reconstruction from RNA-Seq Reads



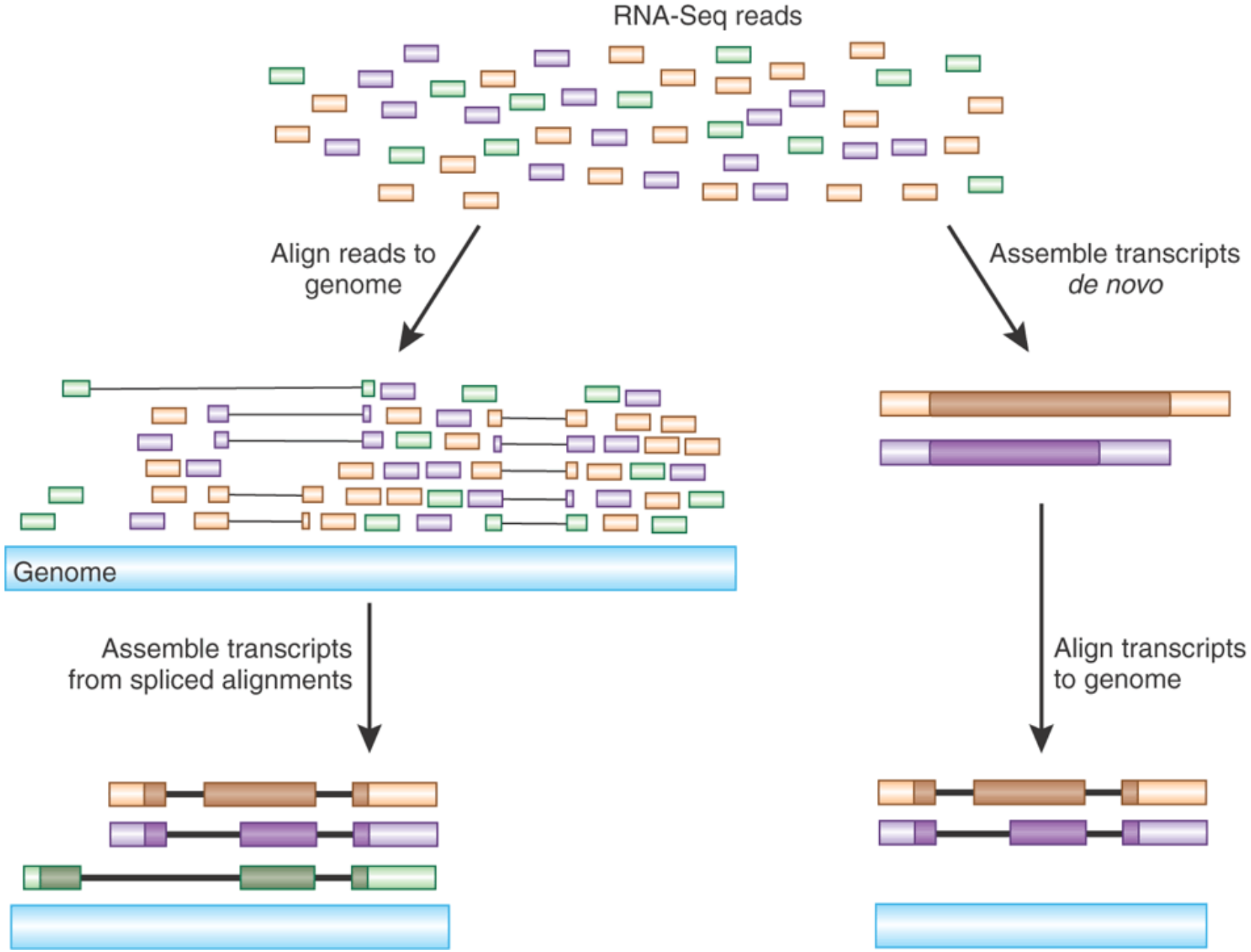
Transcript Reconstruction from RNA-Seq Reads



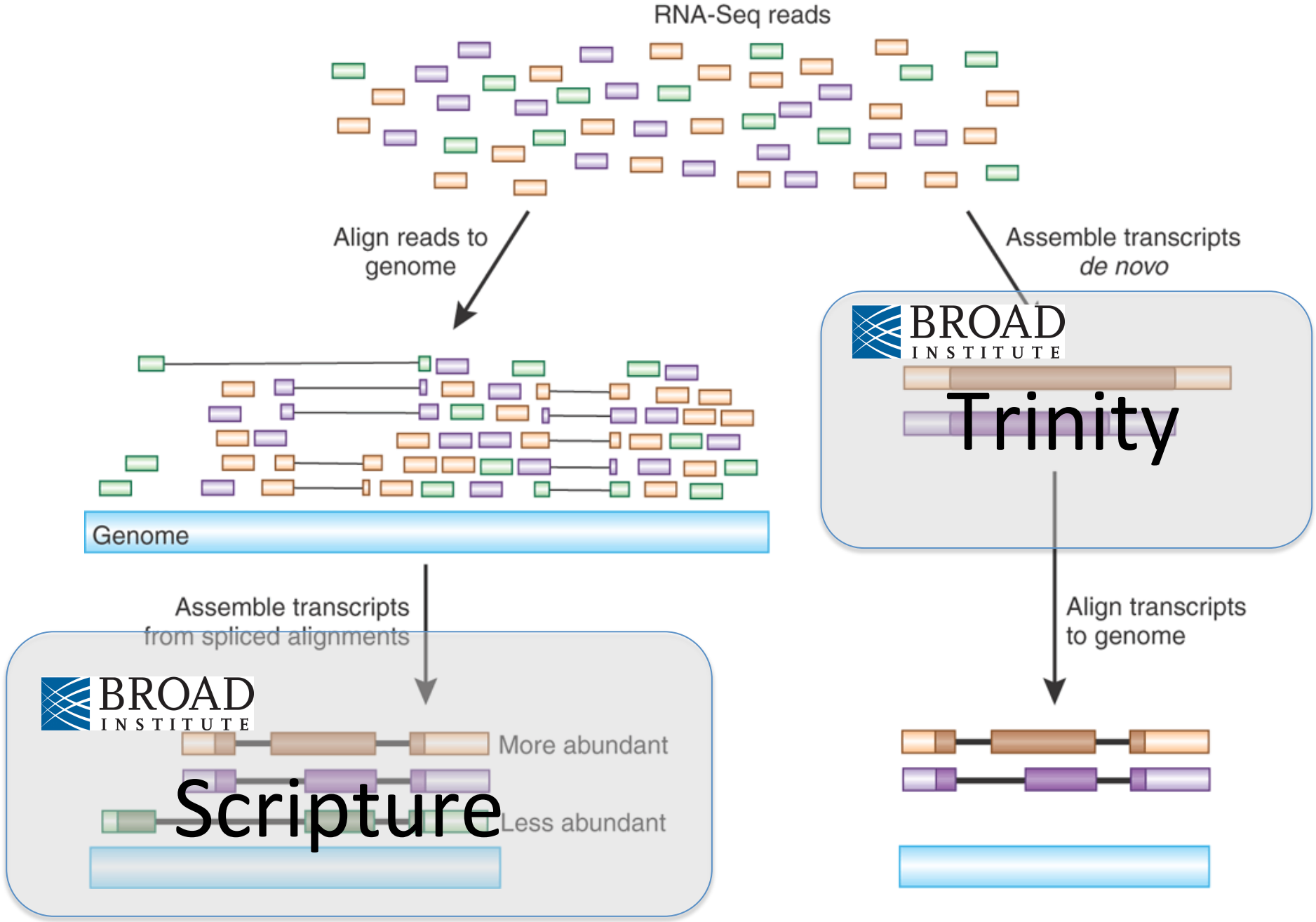
Transcript Reconstruction from RNA-Seq Reads



Transcript Reconstruction from RNA-Seq Reads



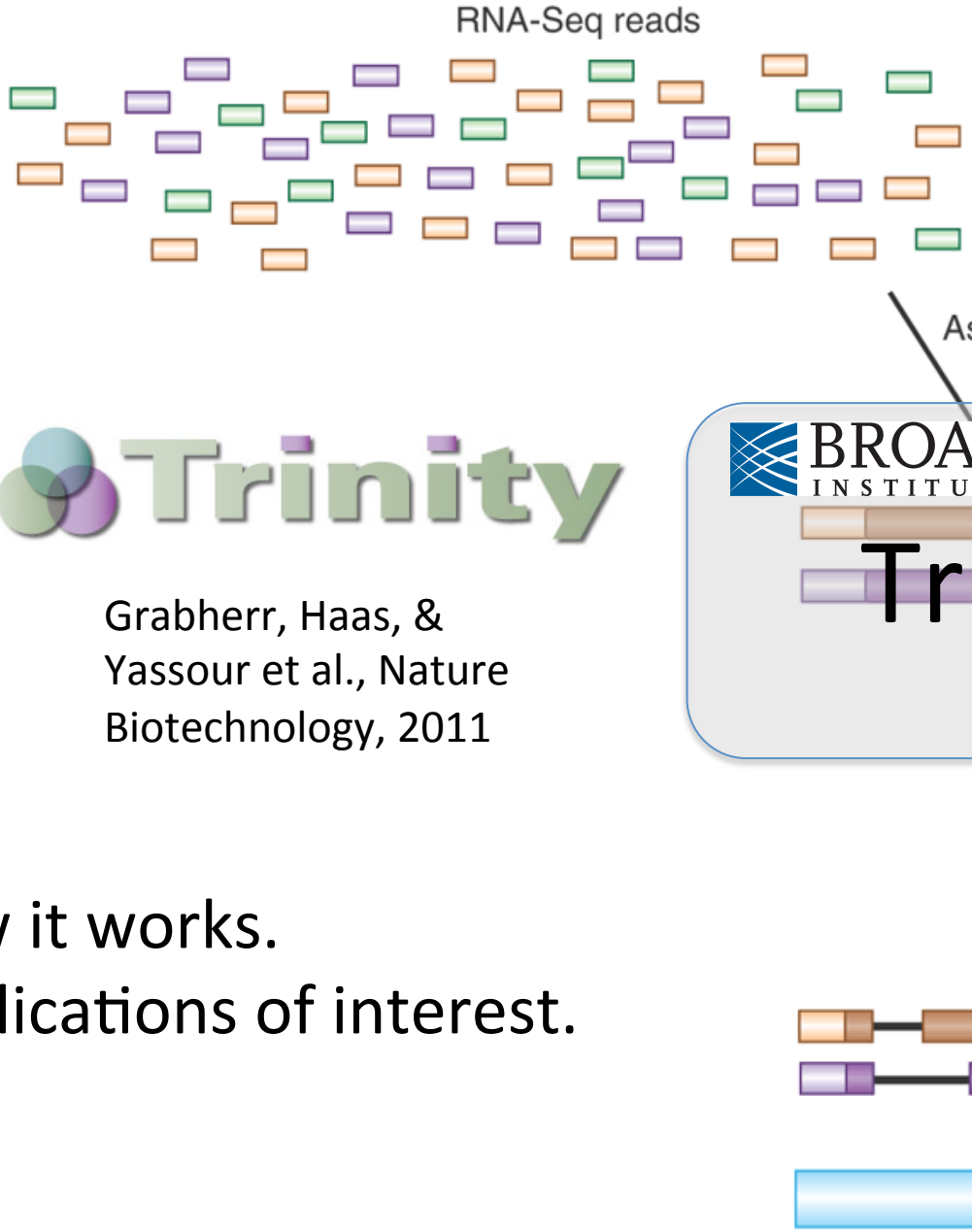
Transcript Reconstruction from RNA-Seq Reads



Transcript Reconstruction from RNA-Seq Reads

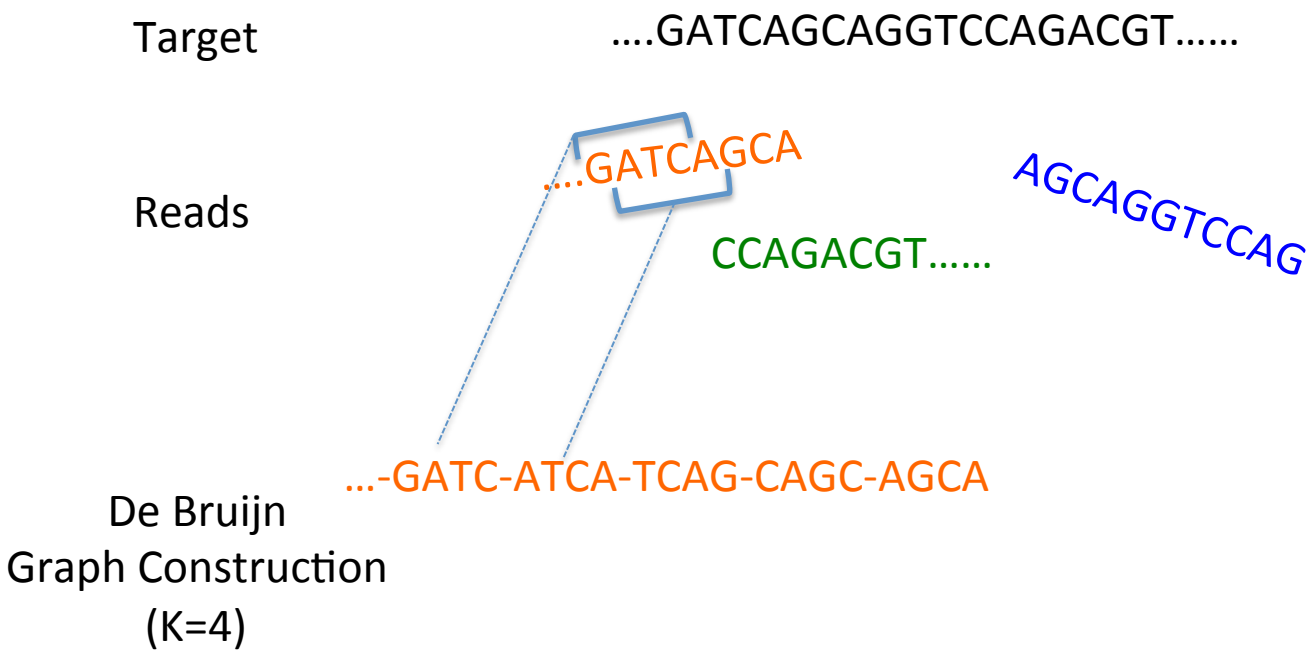
Trinity

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

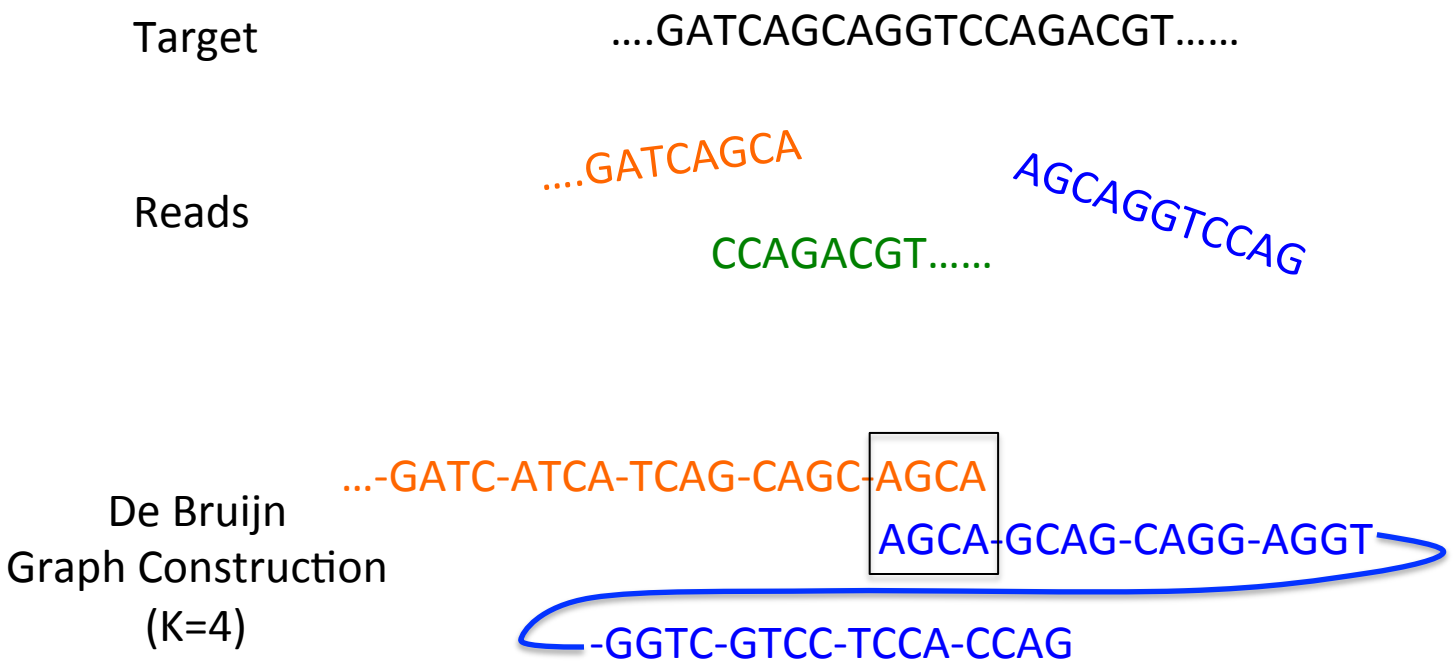


- How it works.
- Applications of interest.

Short Read Assembly Using de Bruijn Graphs



Short Read Assembly Using de Bruijn Graphs

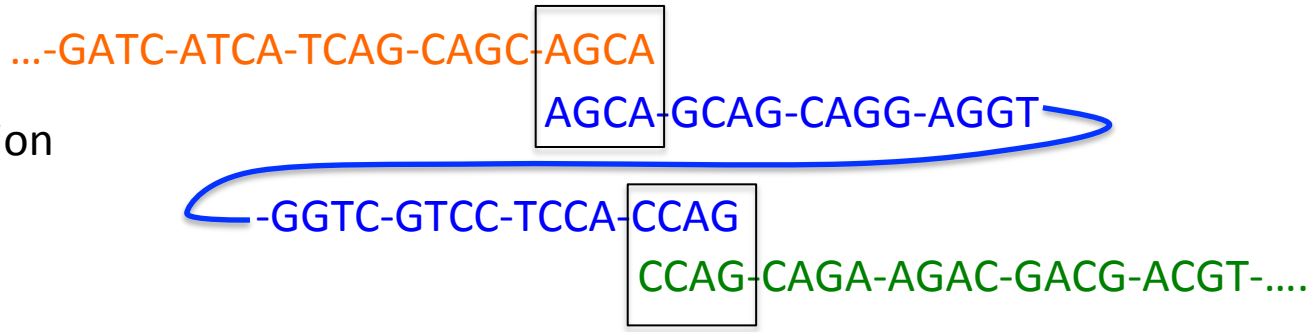


Short Read Assembly Using de Bruijn Graphs

Target ….GATCAGCAGGTCCAGACGT……

Reads ….GATCAGCA AGCAGGTCCAG CCAGACGT……

De Bruijn Graph Construction (K=4)

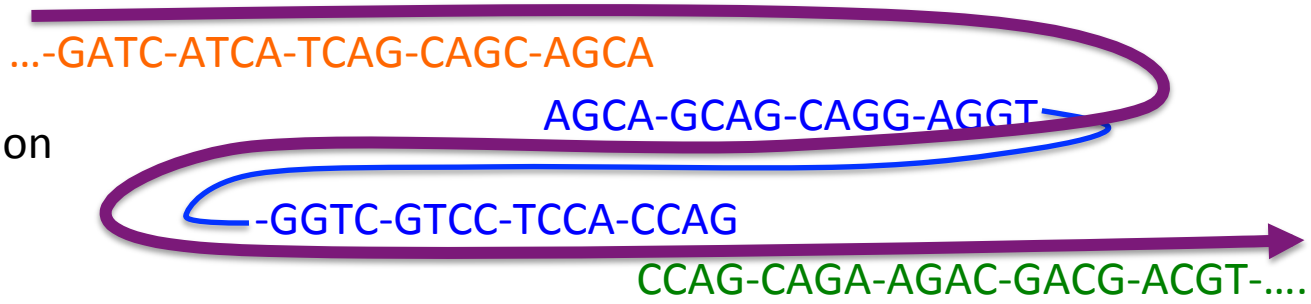


Short Read Assembly Using de Bruijn Graphs

Target ….GATCAGCAGGTCCAGACGT……

Reads ….GATCAGCA AGCAGGTCCAG CCAGACGT……

De Bruijn Graph Construction (K=4)



Sequence Reconstruction By Path Traversal **….GATCAGCAGGTCCAGACGT……**

Contrasting Genome and Transcriptome Assembly

Genome Assembly

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

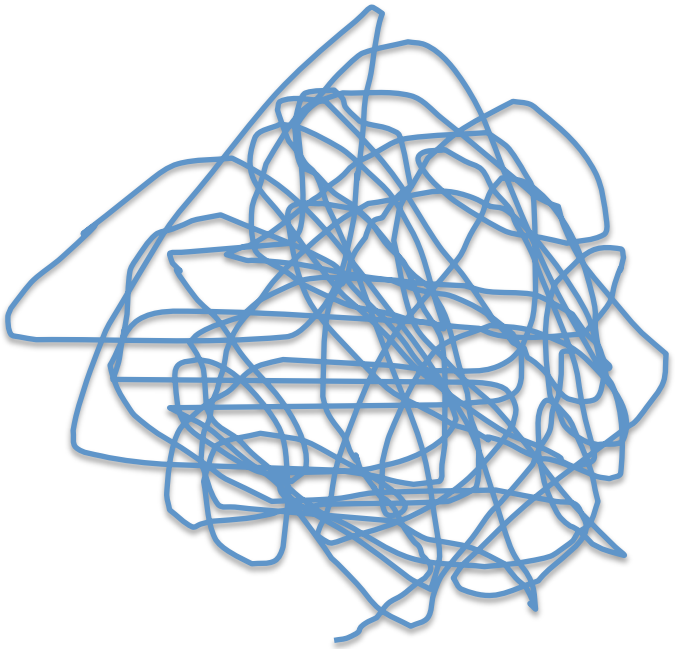
Transcriptome Assembly

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

Trinity Aggregates Isolated Transcript Graphs

Genome Assembly

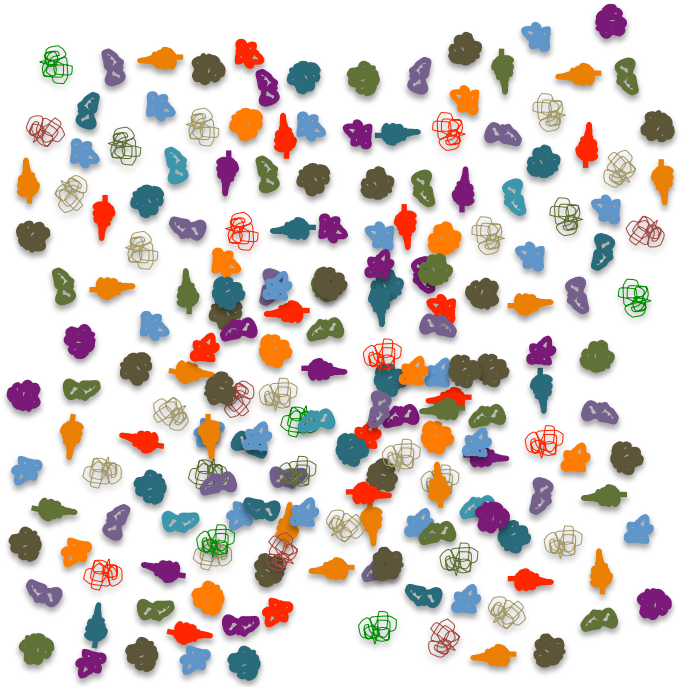
Single Massive Graph



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**

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

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

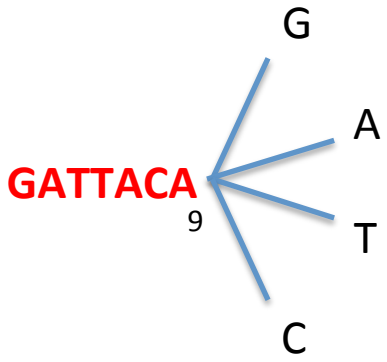
Thousands of disjoint graphs

Inchworm Algorithm

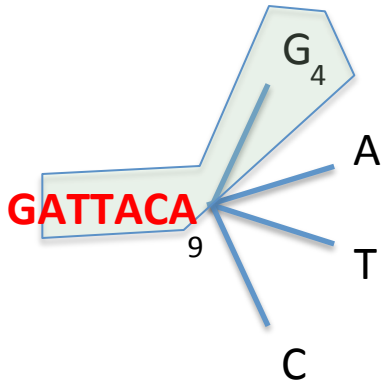
Decompose all reads into overlapping Kmers (25-mers)

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

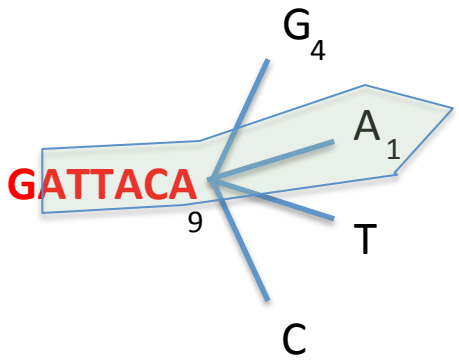
Extend kmer at 3' end, guided by coverage.



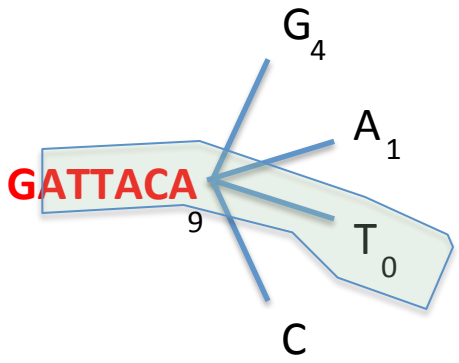
Inchworm Algorithm



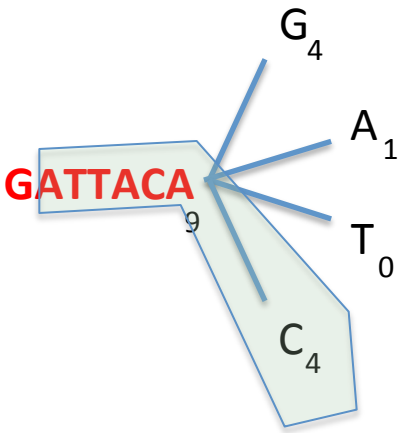
Inchworm Algorithm



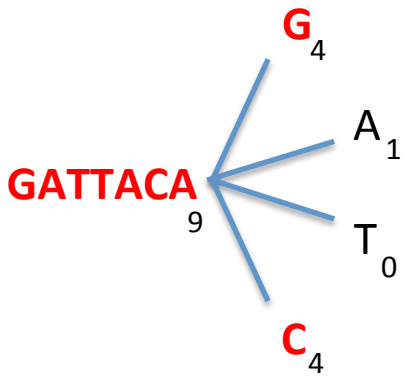
Inchworm Algorithm



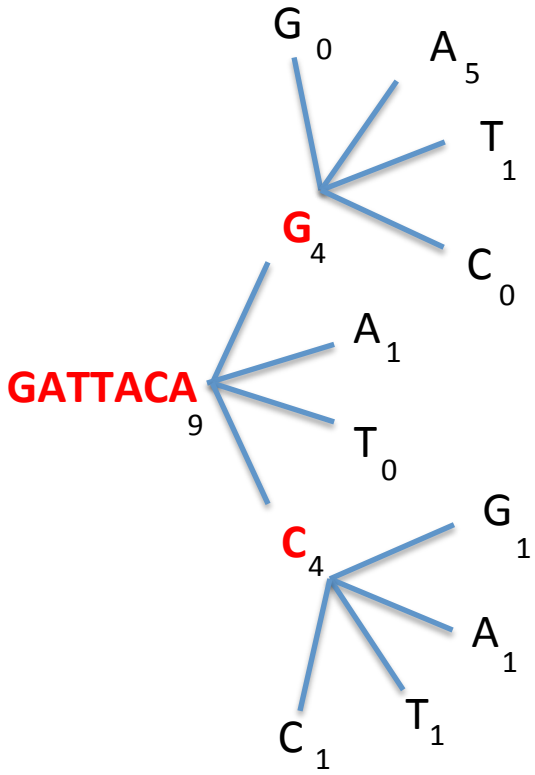
Inchworm Algorithm



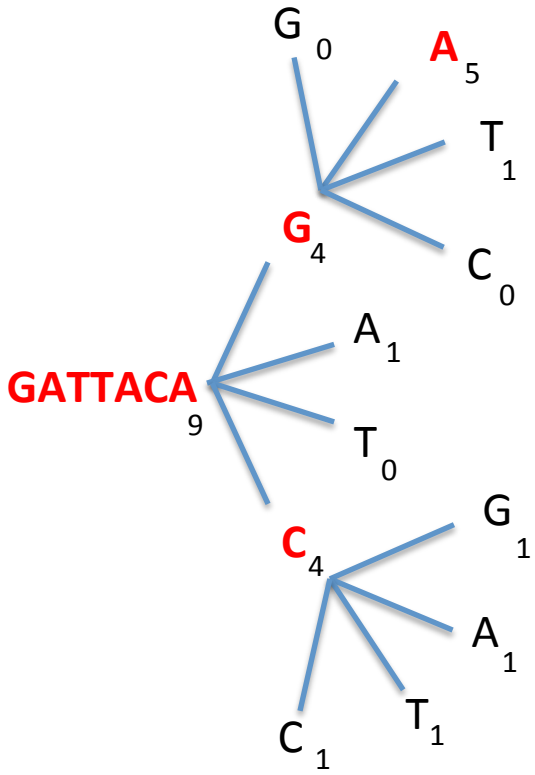
Inchworm Algorithm



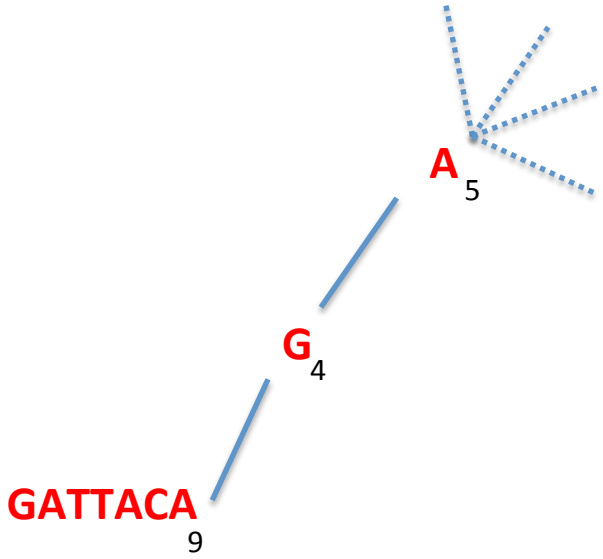
Inchworm Algorithm



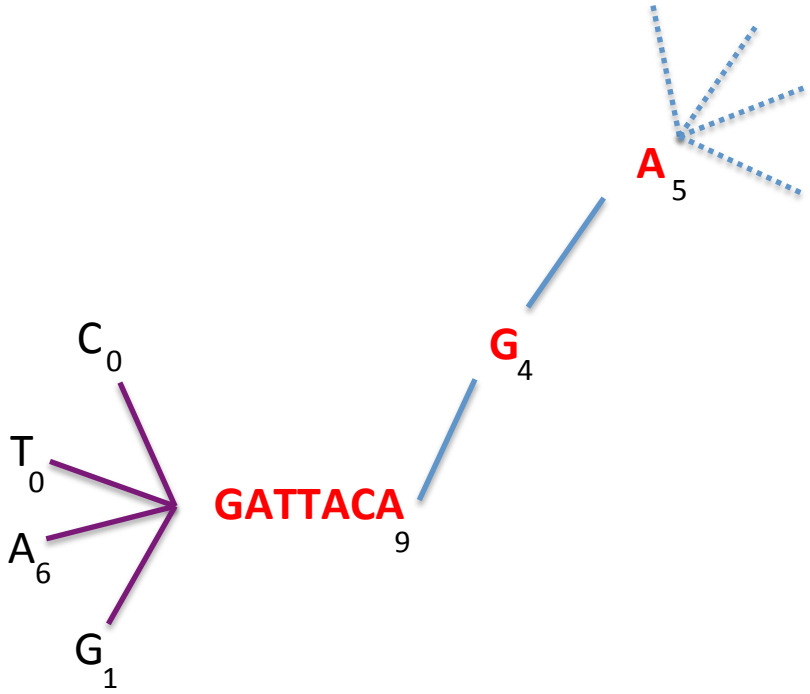
Inchworm Algorithm



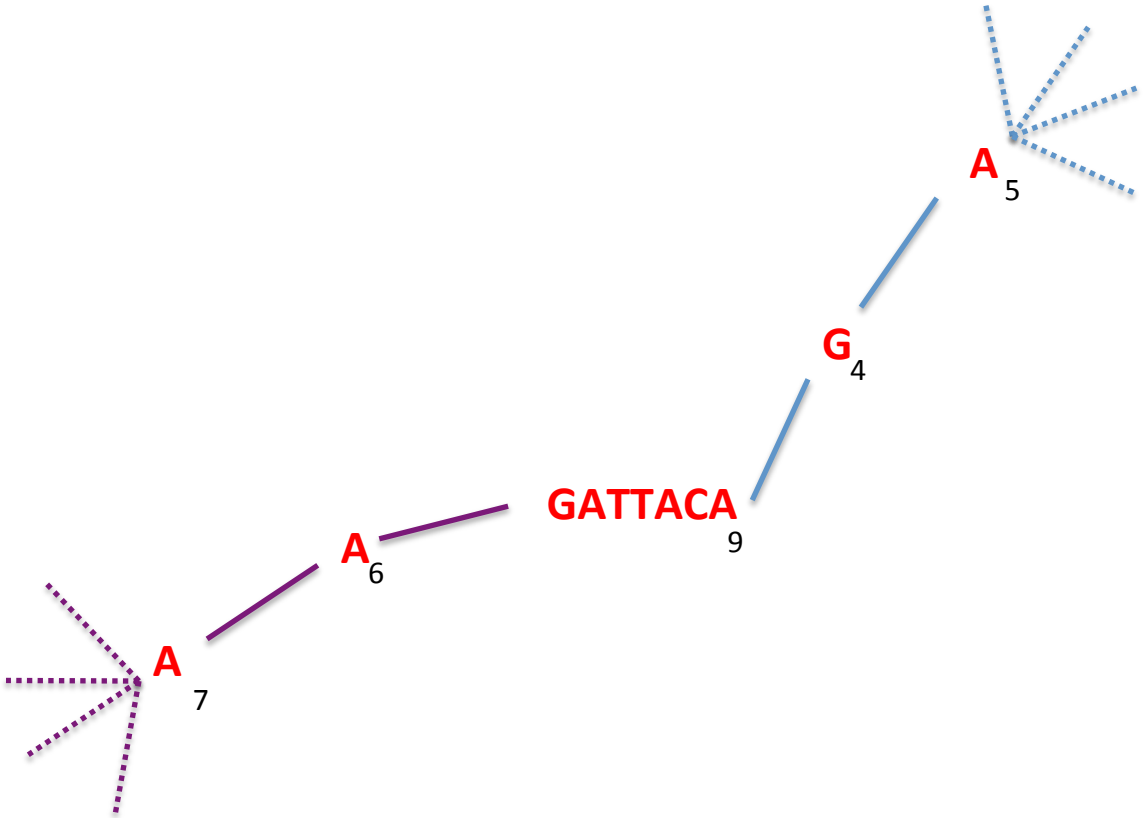
Inchworm Algorithm



Inchworm Algorithm



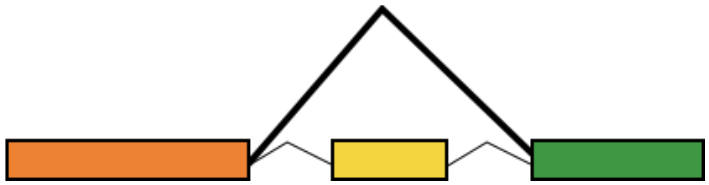
Inchworm Algorithm



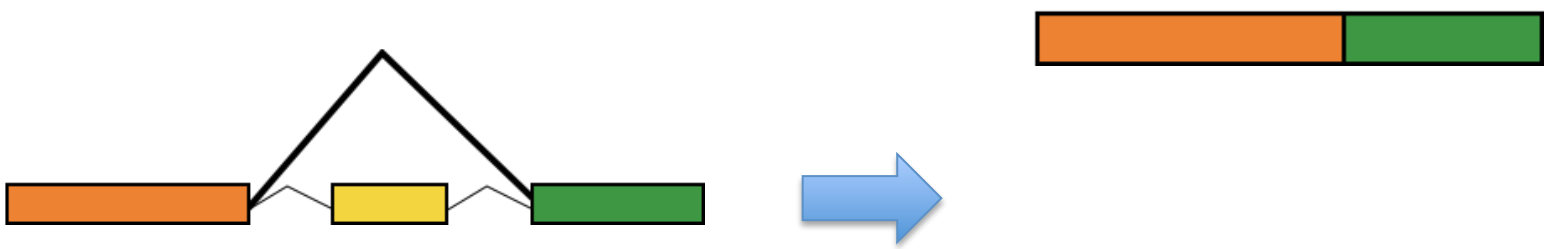
Report contig: ....AAGATTACAGA....

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

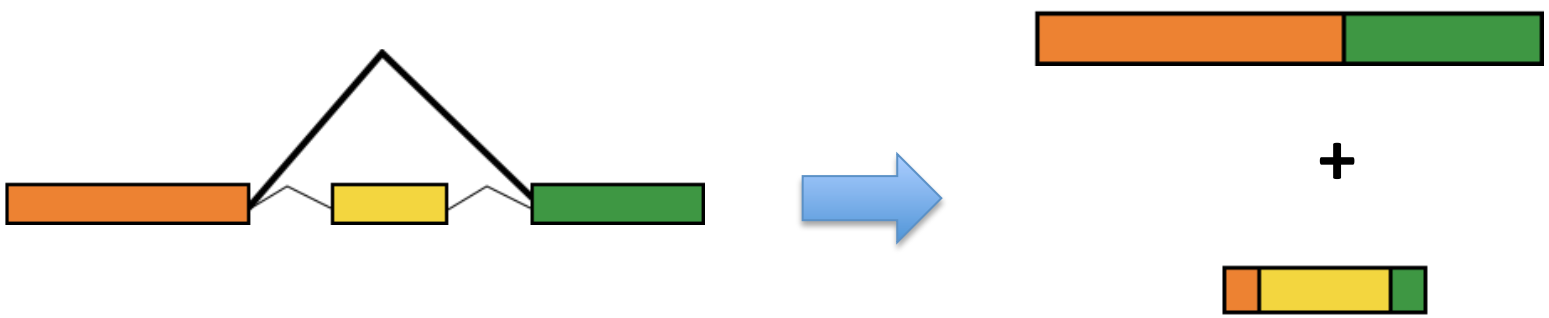
Inchworm Contigs from Alt-Spliced Transcripts



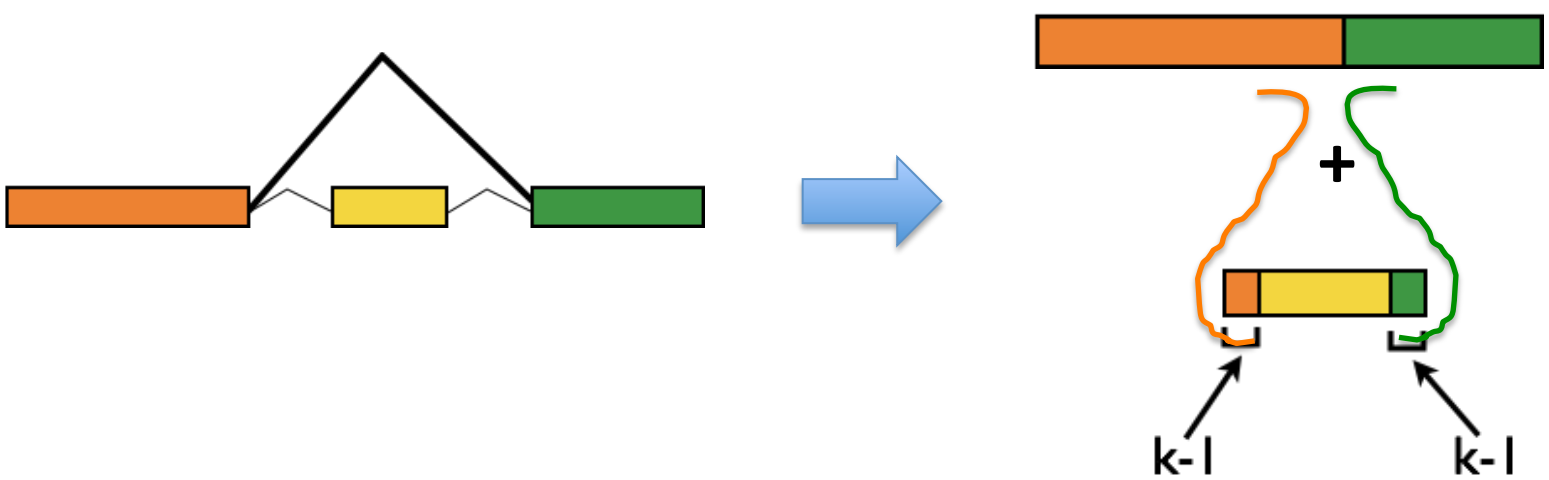
Inchworm Contigs from Alt-Spliced Transcripts



Inchworm Contigs from Alt-Spliced Transcripts



Inchworm Contigs from Alt-Spliced Transcripts

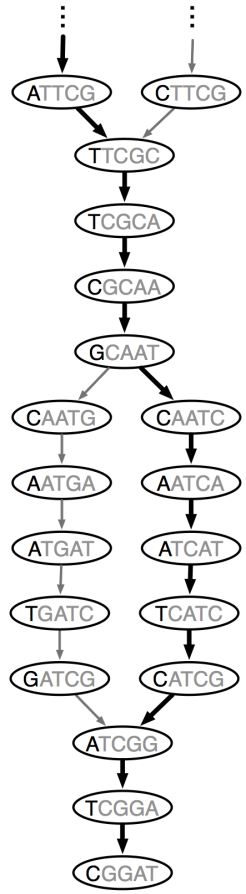


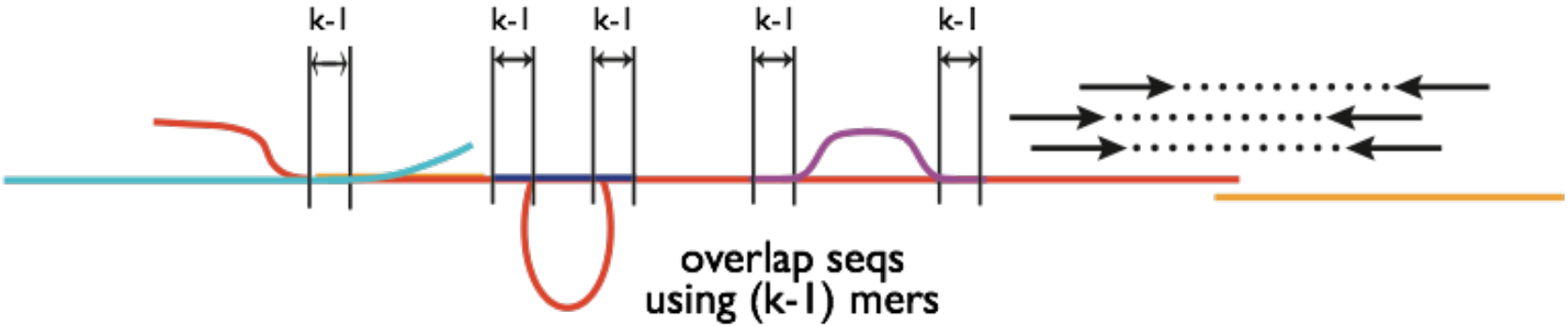
Chrysalis

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

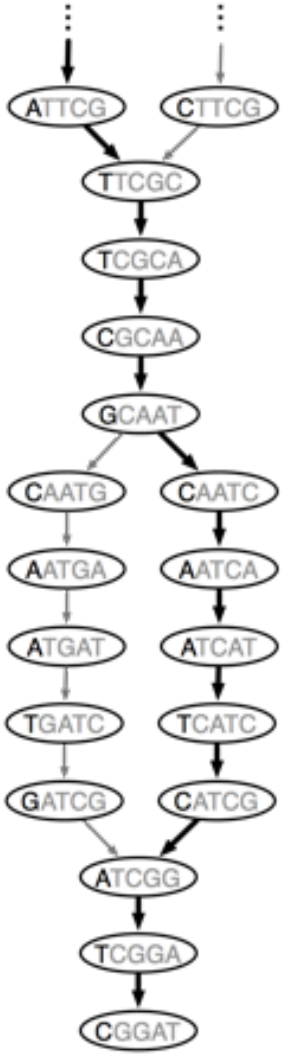
Integrate isoforms
via k-1 overlaps

Build de Bruijn Graphs
(ideally, one per gene)



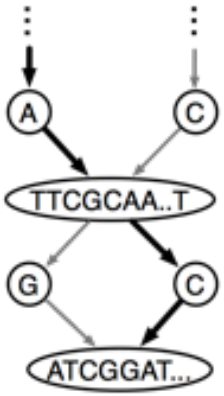


Butterfly



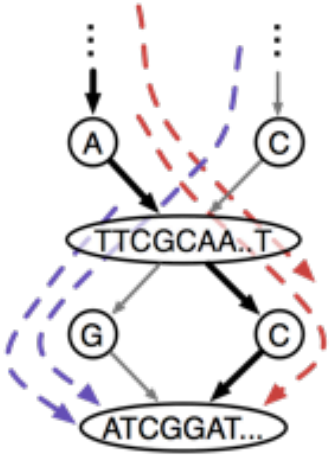
de Bruijn graph

compacting



compact graph

finding paths



compact graph with reads

extracting sequences

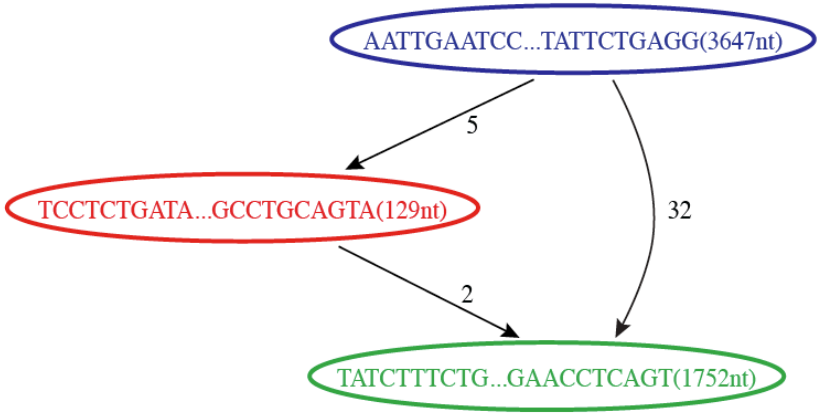
..CTTCGCAA..TGATCGGAT...

..ATTCGCAA..TCATCGGAT...

sequences
(isoforms and paralogs)

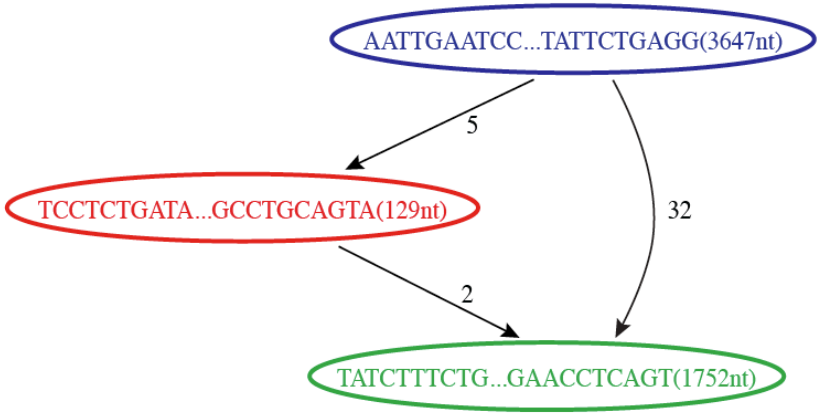
Reconstruction of Alternatively Spliced Transcripts

Butterfly's Compacted Sequence Graph



Reconstruction of Alternatively Spliced Transcripts

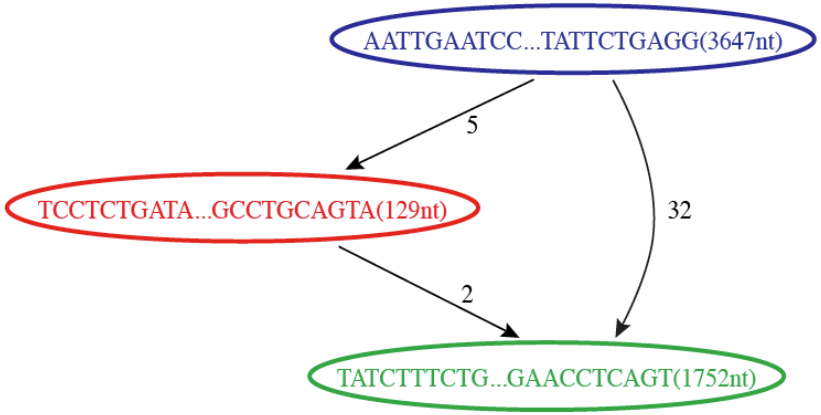
Butterfly's Compacted Sequence Graph



Reconstructed Transcripts

Reconstruction of Alternatively Spliced Transcripts

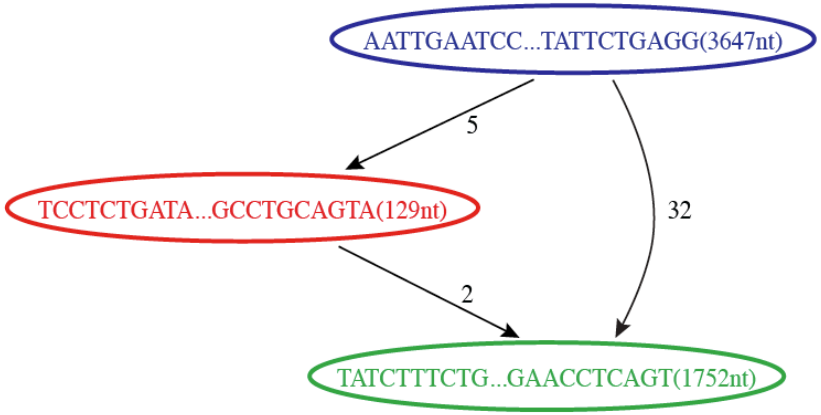
Butterfly's Compacted Sequence Graph



Reconstructed Transcripts

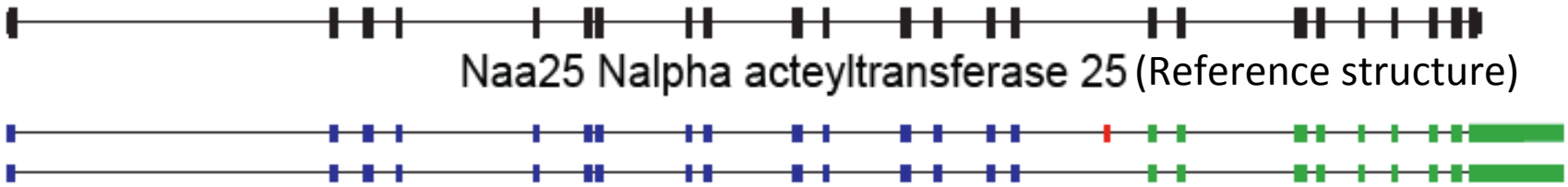
Reconstruction of Alternatively Spliced Transcripts

Butterfly's Compacted Sequence Graph



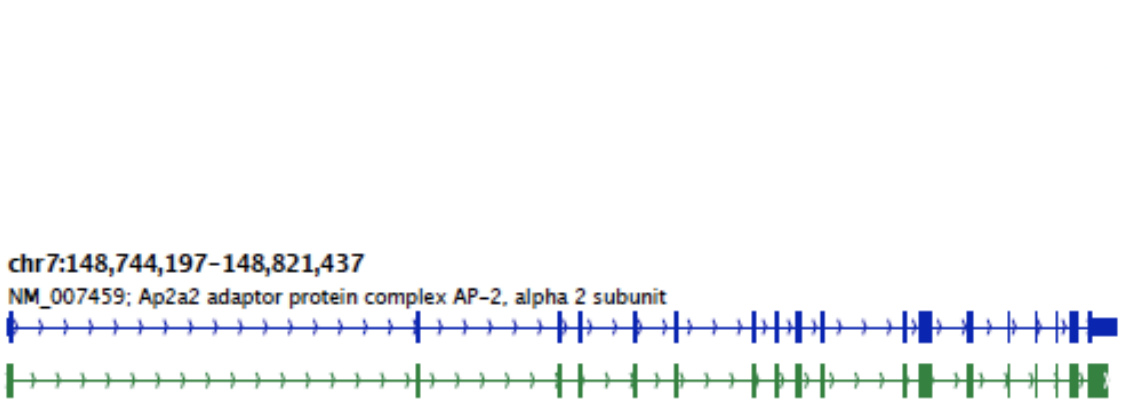
Reconstructed Transcripts

Aligned to Mouse Genome



Teasing Apart Transcripts of Paralogous Genes

Teasing Apart Transcripts of Paralogous Genes



Strand-specific RNA-Seq is Preferred

Computationally: fewer confounding graph structures:

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

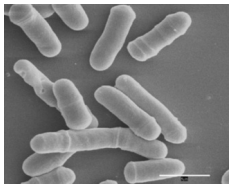
'dUTP second strand marking' identified as the leading protocol

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

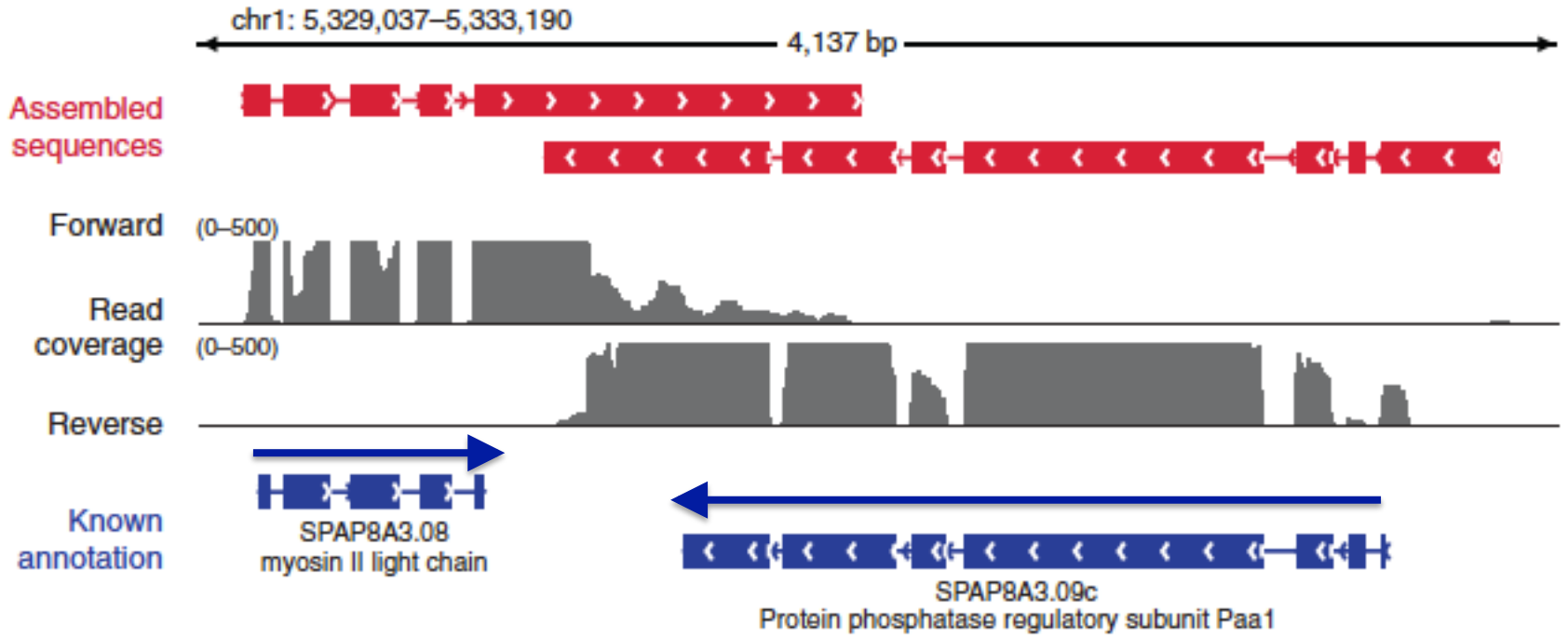
Nevertheless, direct information on the originating strand can substantially enhance the value of an RNA-seq experiment. For

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

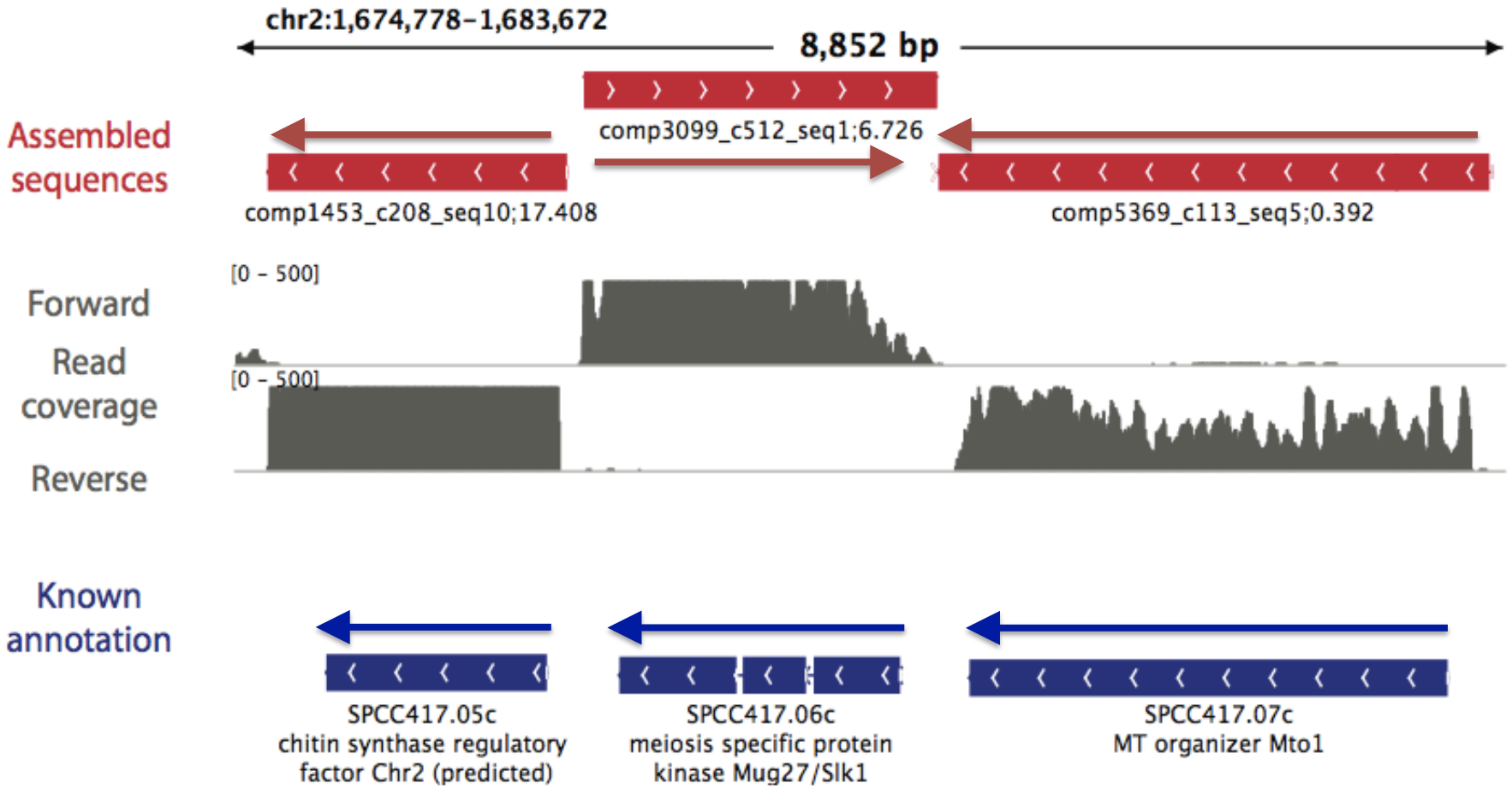
Overlapping UTRs from Opposite Strands

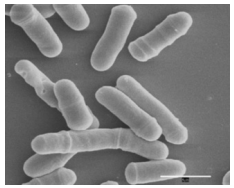


Schizosacharomyces pombe
(fission yeast)



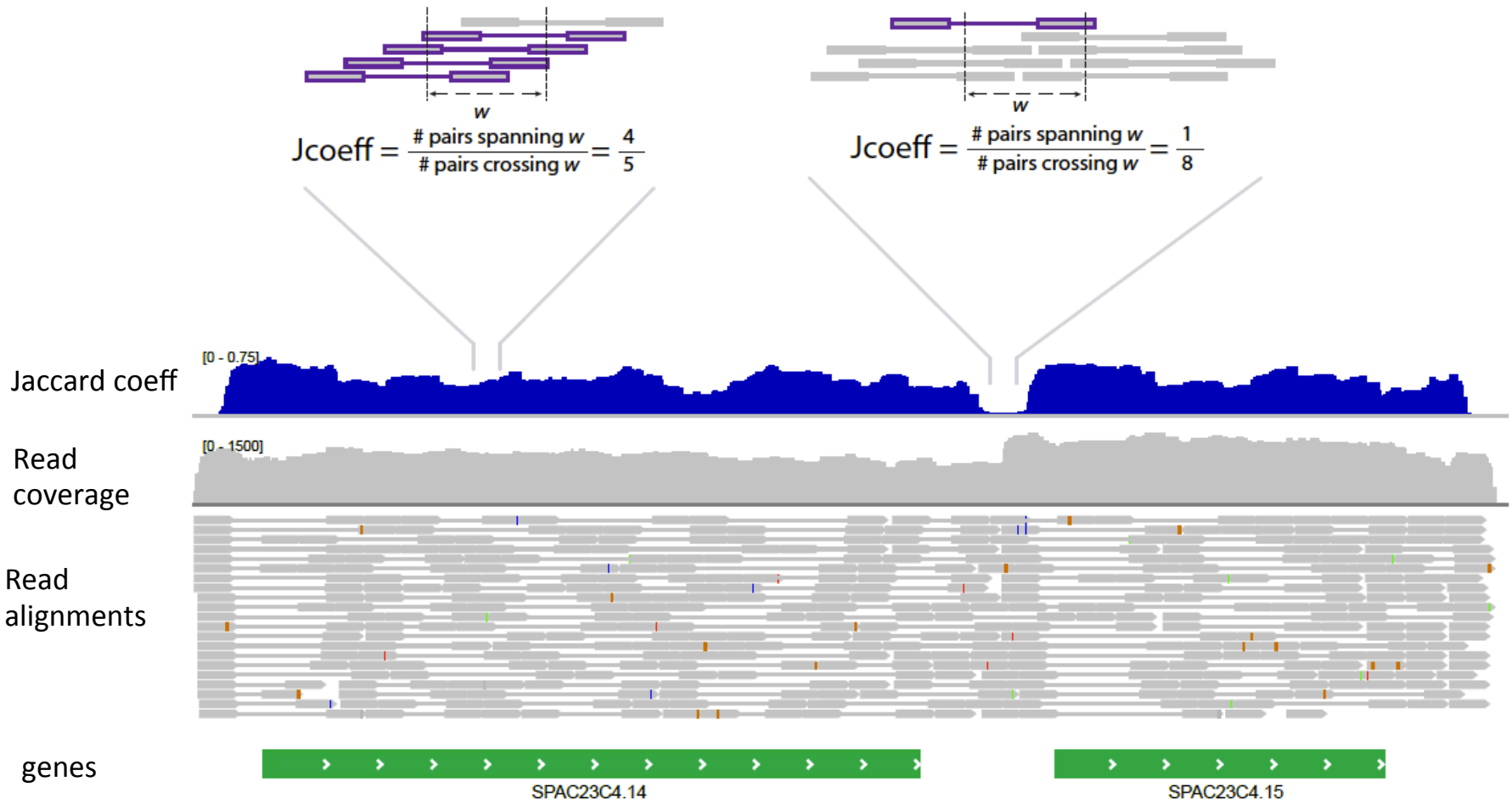
Antisense-dominated Transcription





Compact genomes and Not using strand-specific data = fusion transcripts

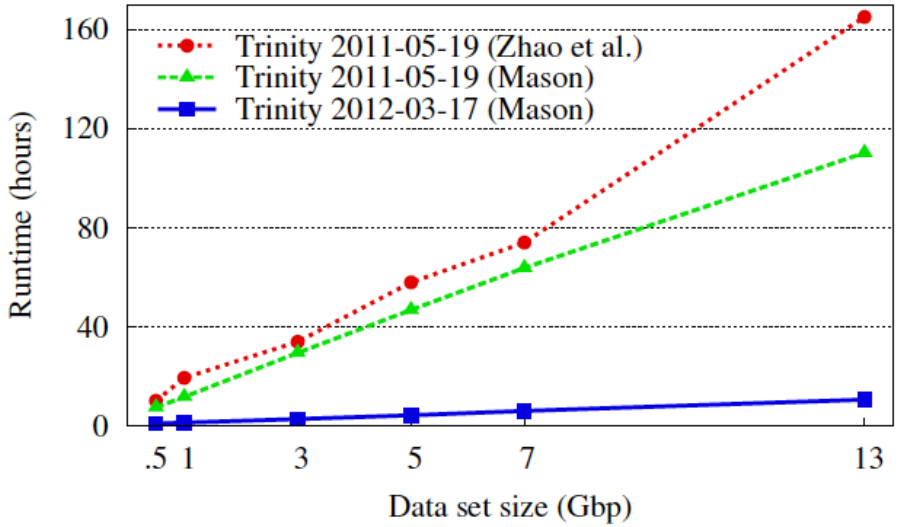
* Use the Jaccard Clip Option *



Post-publication of Trinity

- Since May, 2011:
  - Approx. 750 software downloads per month
  - More than 100 literature citations
  - Open Source software development contributions from developers world-wide.

Trinity is now 10x faster!



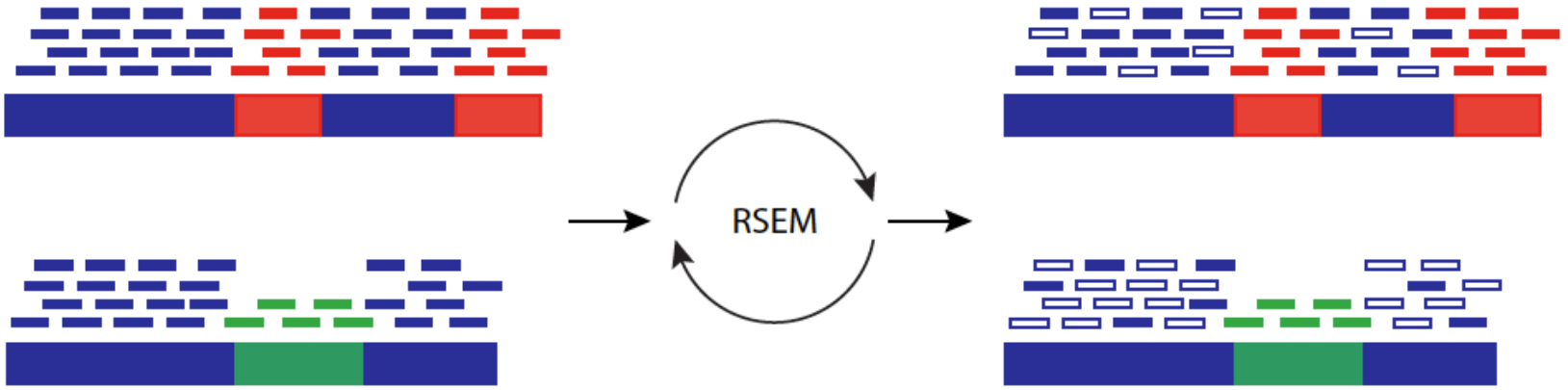
Trinity as a foundation for transcriptomics in diverse organisms

RNA-Seq →

- Differential Expression
- Alternative Splicing
- SNPs & Alleles
- Coding annotations
- Comparative Transcriptomics
- Fusion transcripts (Cancer)
- Support proteomic studies

Abundance Estimation Using Expectation Maximization

RSEM fractionally allocates multi-mapped reads according to maximum likelihood



RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.
Bo Li and Colin N Dewey.
BMC Bioinformatics. 2011; 12: 323.

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

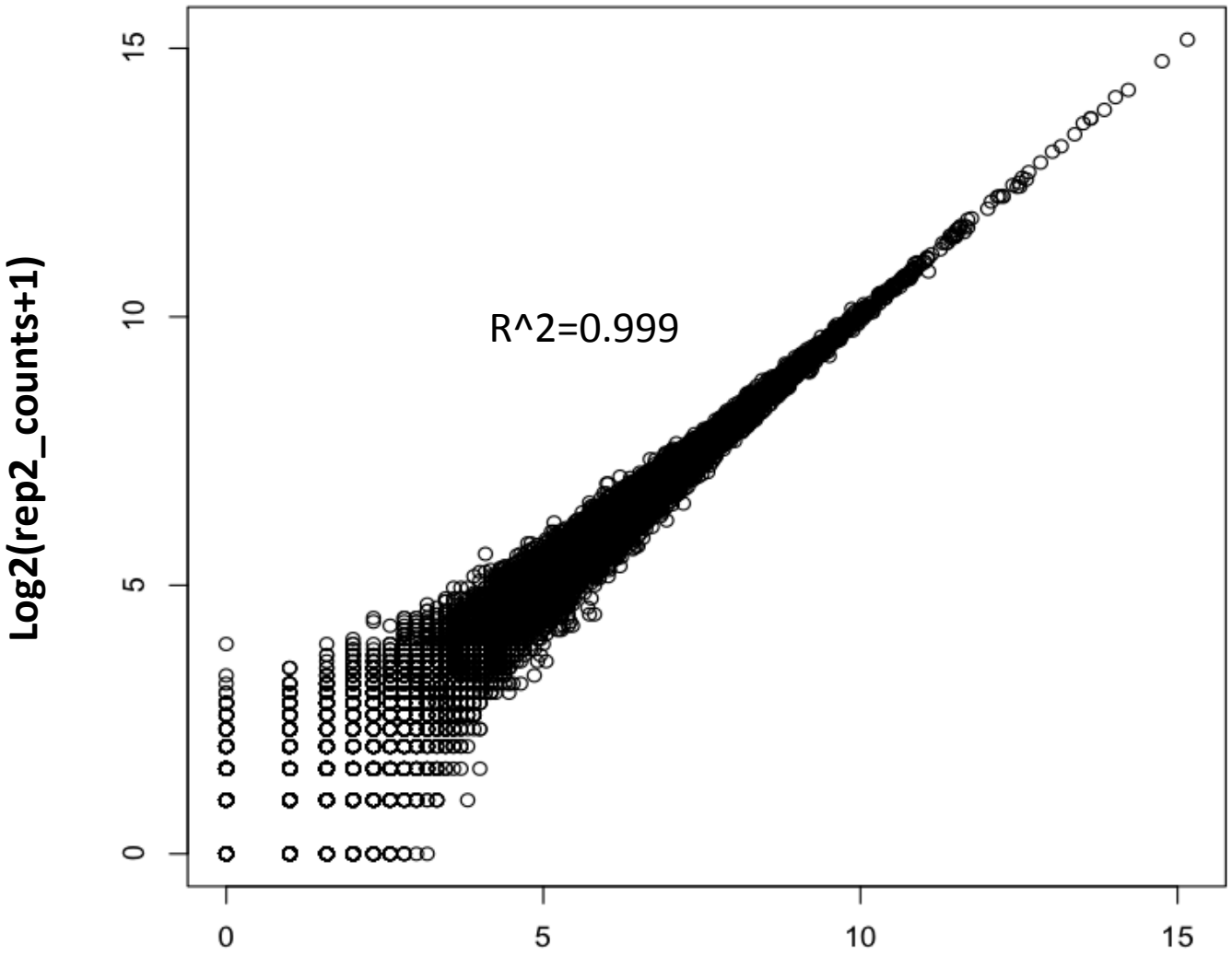
FPKM

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

RSEM Abundance Estimates

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

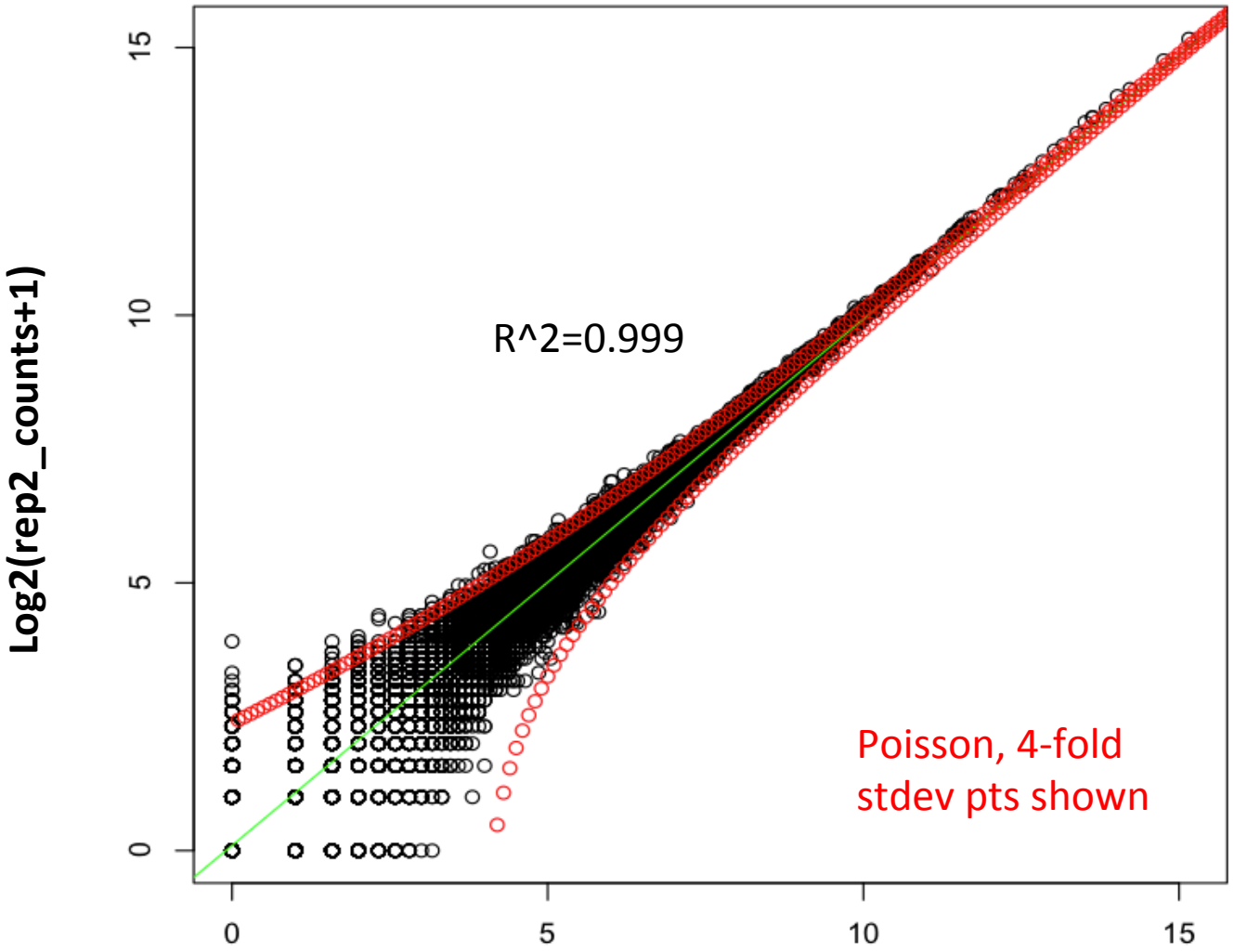


Kidney rna-seq samples

Data from Marioni et al. [2008]

Technical Replicates

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



Kidney rna-seq samples

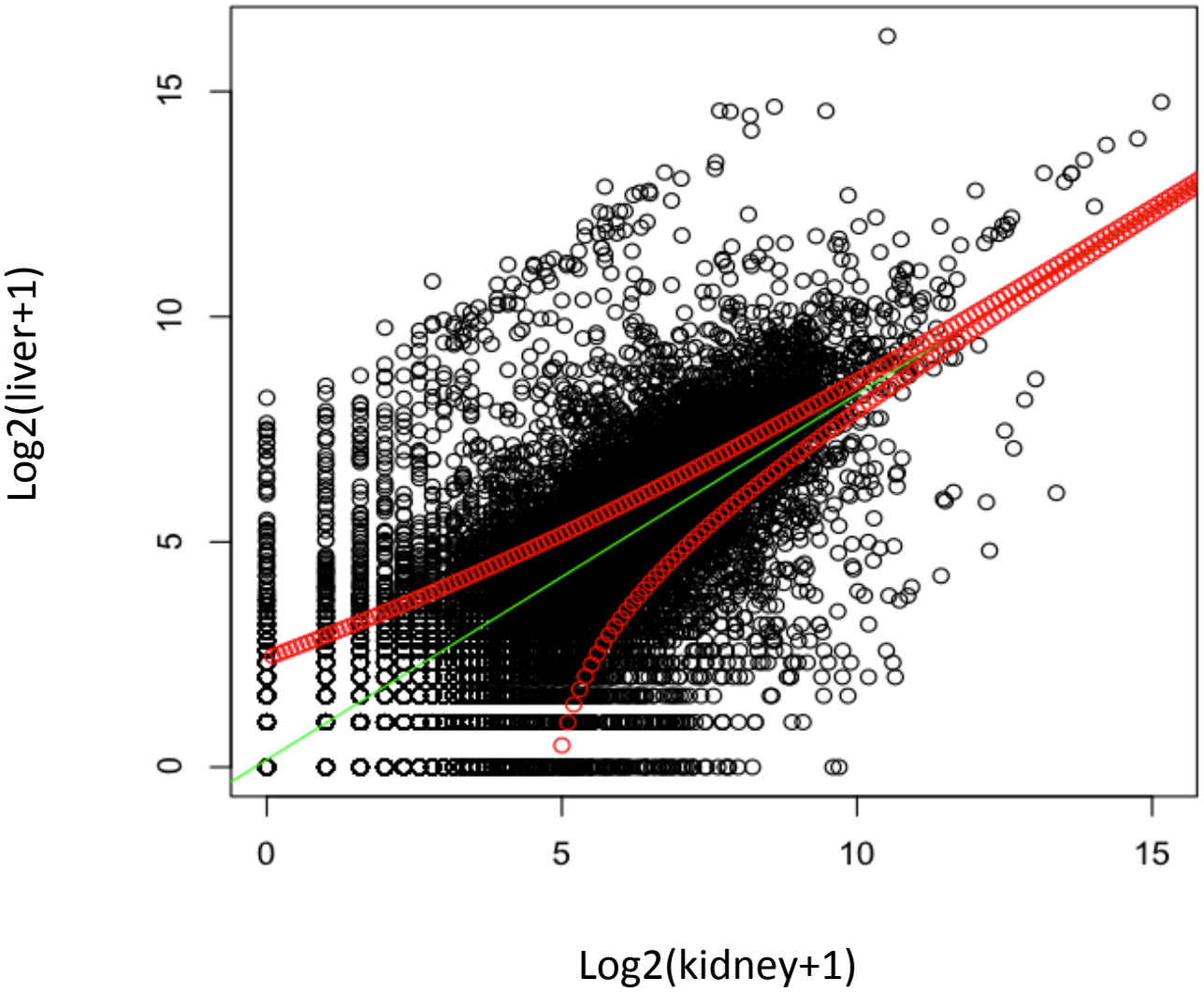
Data from Marioni et al. [2008]

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)

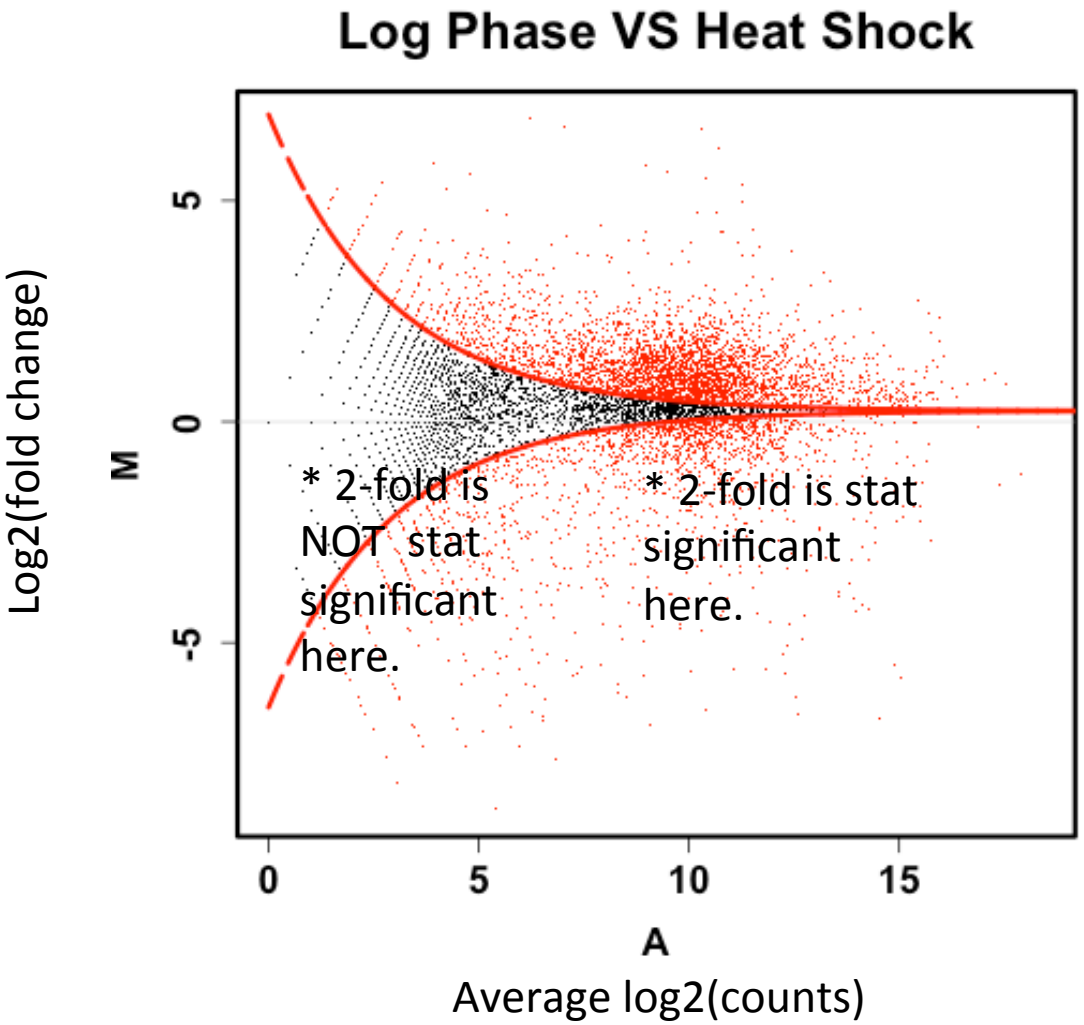
  If (P <= 0.05), significantly different
- Don't forget to adjust P-values due to false discovery rate (FDR) resulting from running many (thousands of) statistical tests. (ex. use Q-values)

Kidney vs. Liver



Increased Power for Identifying Differentially Expressed Transcripts With Deeper Sequencing

MA plot: log(Counts) vs. log(Fold change)



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

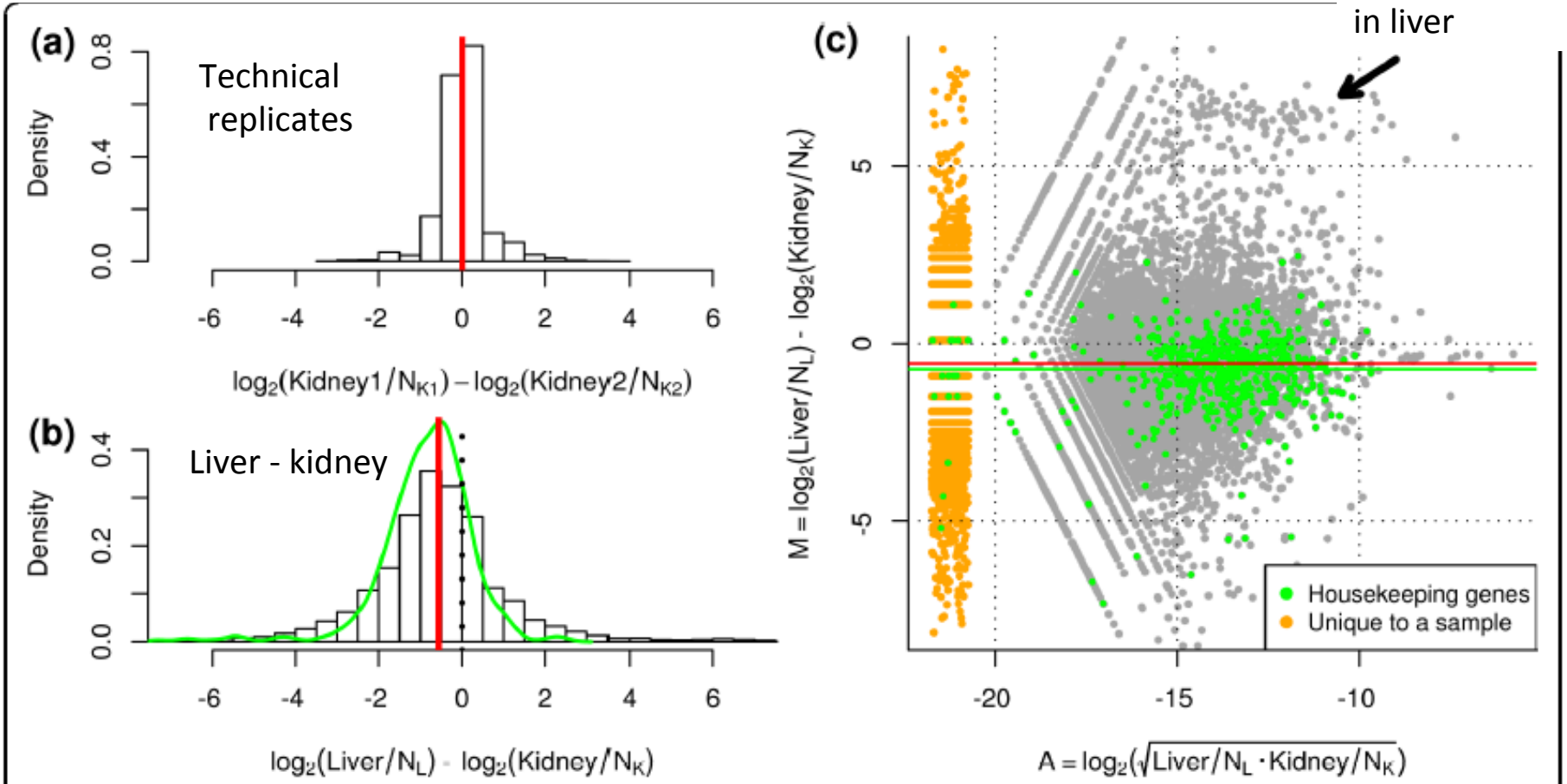
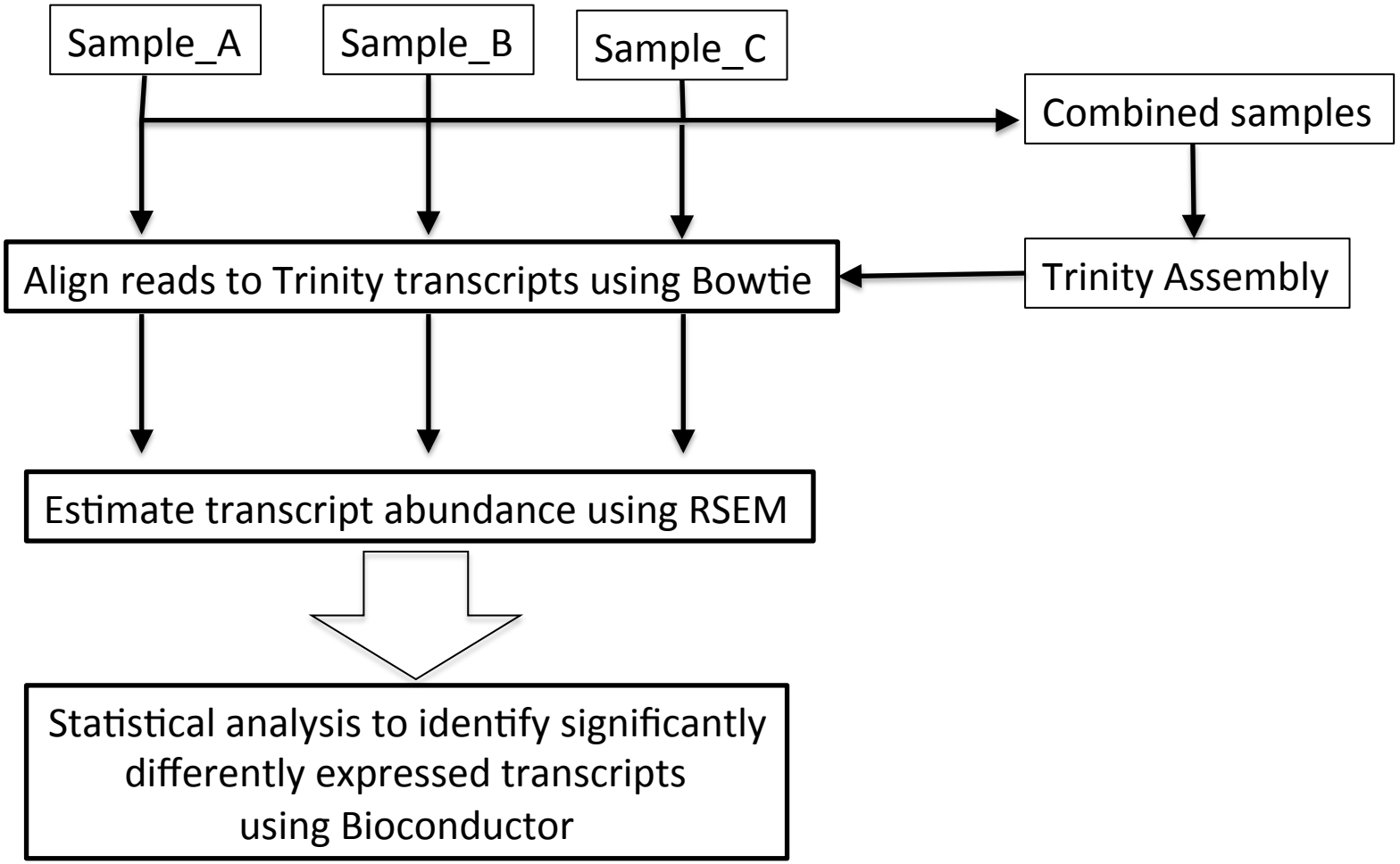


Figure 1 Normalization is required for RNA-seq data. Data from [6] comparing log ratios of **(a)** technical replicates and **(b)** liver versus kidney expression levels, after adjusting for the total number of reads in each sample. The green line shows the smoothed distribution of log-fold-changes of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

Robinson and Oshlack, Genome Biology, 2010

Differential Expression Pipeline



Statistical Analysis Software for Identifying Differentially Expressed Transcripts

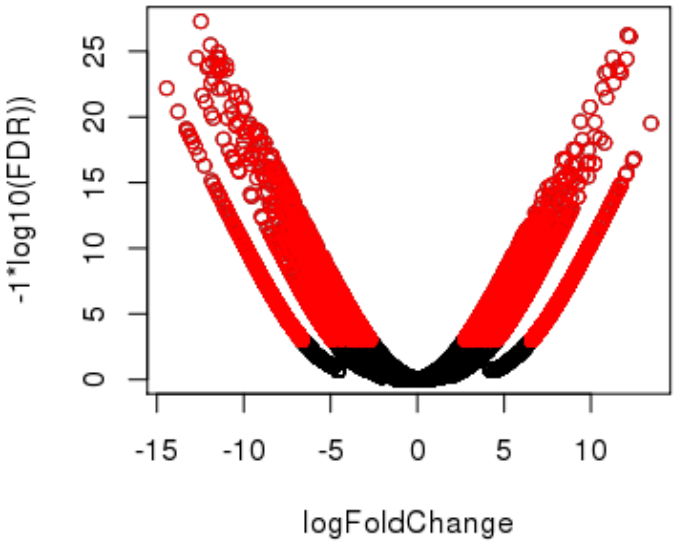
- Bioconductor
  - EdgeR
  - DEGseq
  - DESeq
  - And others…

Examples of Results (example edgeR)

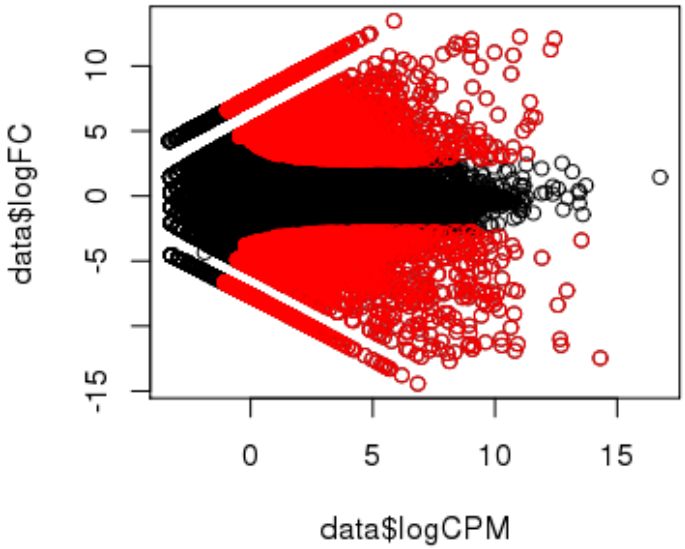
		comp217_c0_seq1	
0	logFC	6.69056684523186	
1	logCPM	16.1146897543805	(expression level)
2	PValue	2.06844466442231e-15	
3	FDR	9.01969581996253e-13	

Plotting Pairwise Differential Expression Data

Volcano plot
(fold change vs. significance)



MA plot
(abundance vs. fold change)

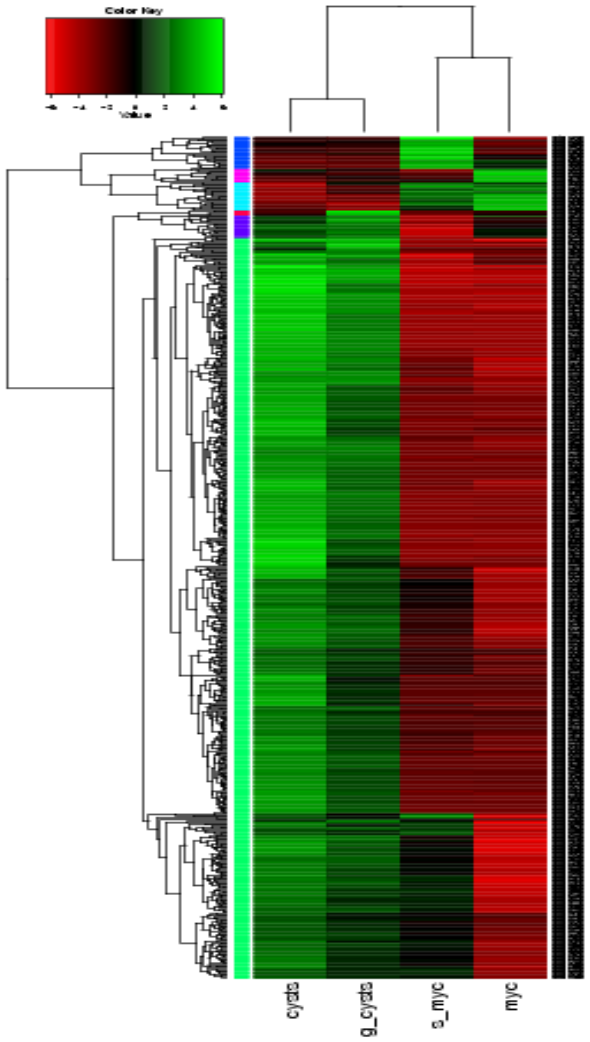


(expression level)

Significantly differently expressed transcripts have FDR <= 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.

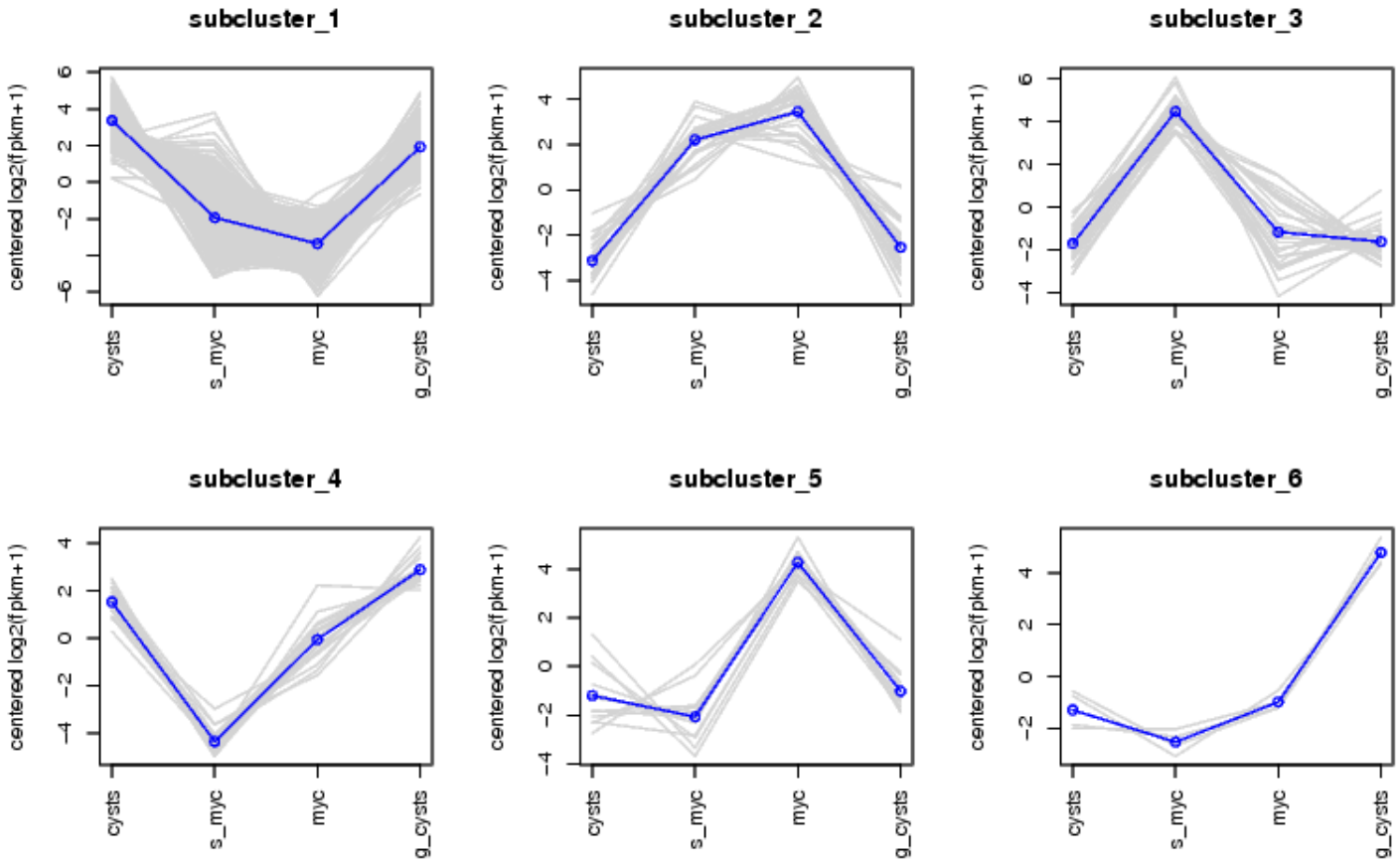
Clustering can be performed across both axes:

-cluster transcripts with similar expression 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.



Hands-on Tutorial

- Trinity
  - De novo assembly using Trinity
  - Bowtie and RSEM for abundance estimation
  - edgeR for differential expression analysis