BCB 5200 Introduction to Bioinformatics I

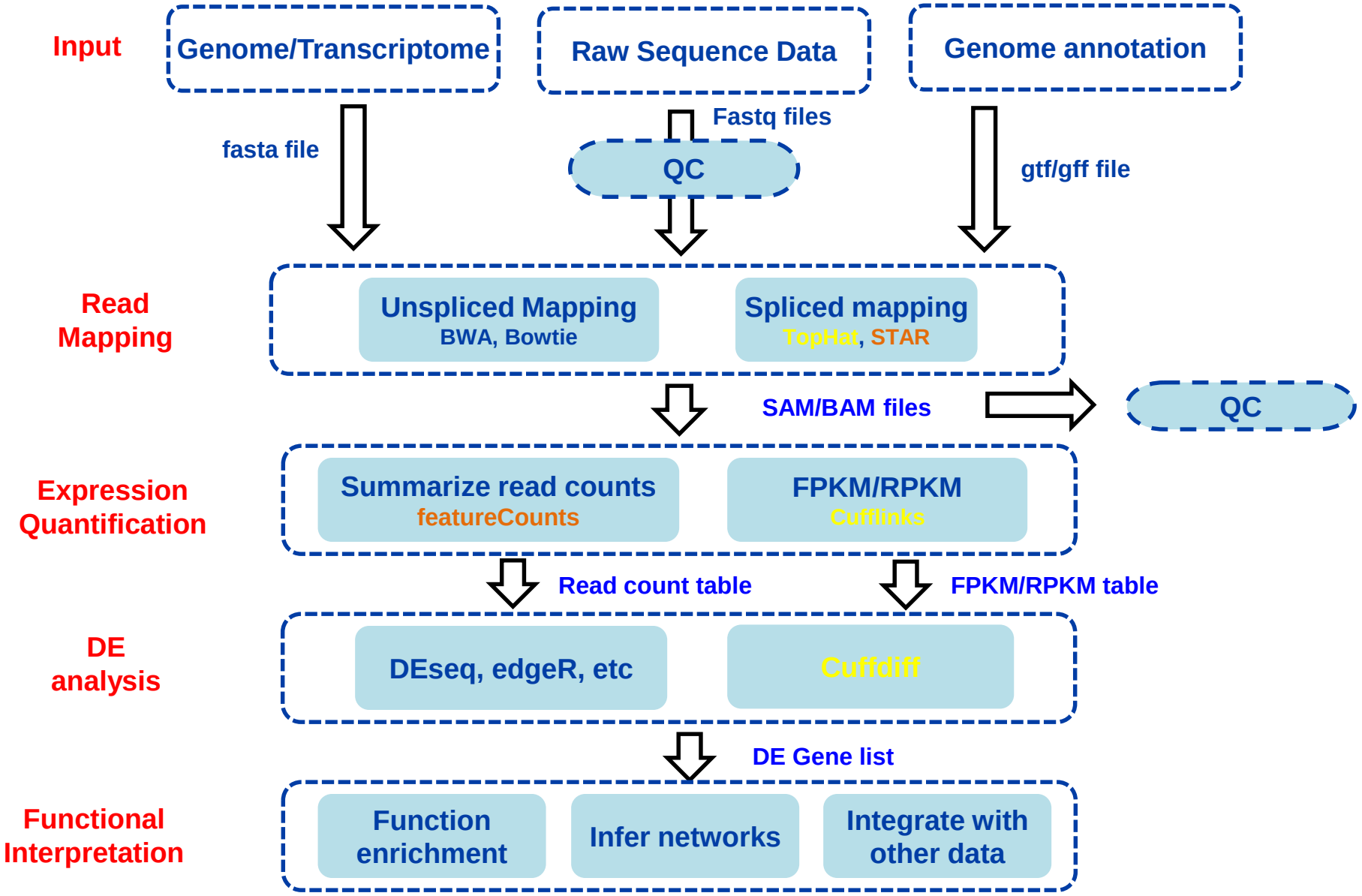
Analysis pipeline and Tuxedo tools

Zhenguo Lin, PhD
Department of Biology
Fall 2017

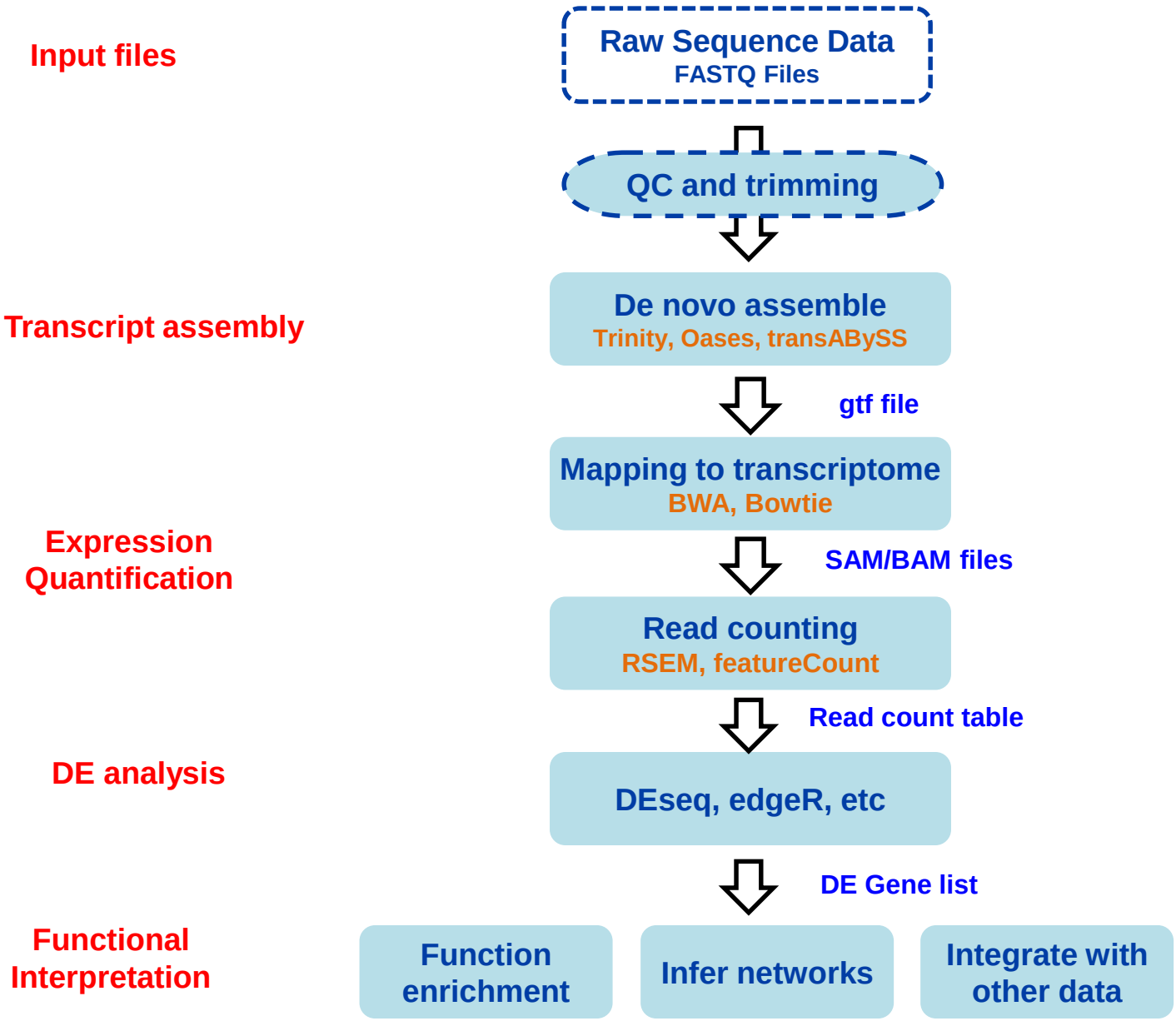
Today's Topic

- General pipeline for RNA-seq analysis
  - Referenced based: map to genome/transcriptome
  - Reference free: de novo assembly
- Tuxedo tools

From reads to differential expression: reference based



From reads to differential expression: reference free

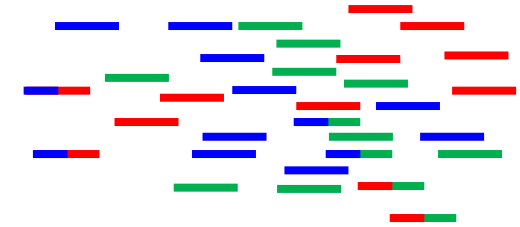


Map reads to transcriptome

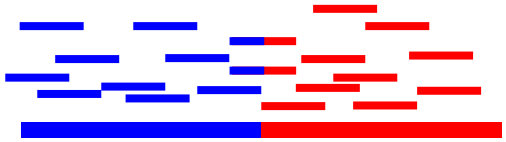
- Use unspliced aligners that do not allow large gaps may be the proper choice for accurate read mapping
- limited to the identification of known exons and junctions
- does not identify splicing events involving novel exons

Missed novel exons by mapping to known transcripts

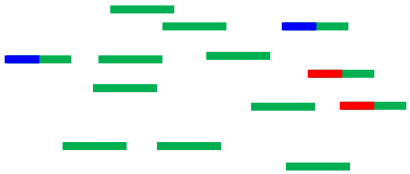
Known transcript



RNA-seq reads



Map to reads to transcript

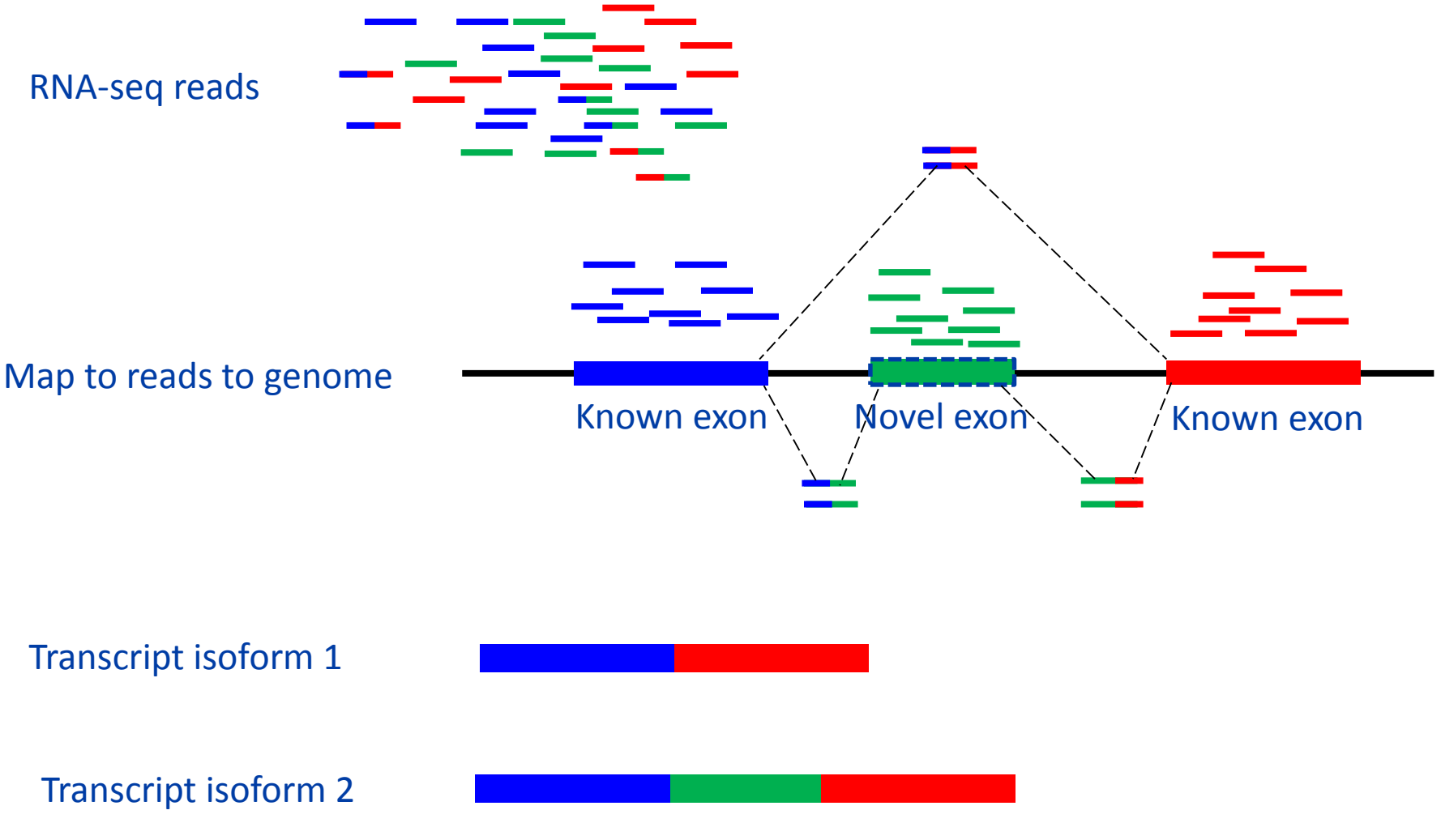


Unmapped reads

Map reads to genome

- Use spliced aligners so that reads aligned at exon-exon junctions will be split into two fragments
- Increase the chance of identifying novel transcripts generated by alternative splicing

Detected the novel transcript by genome mapping



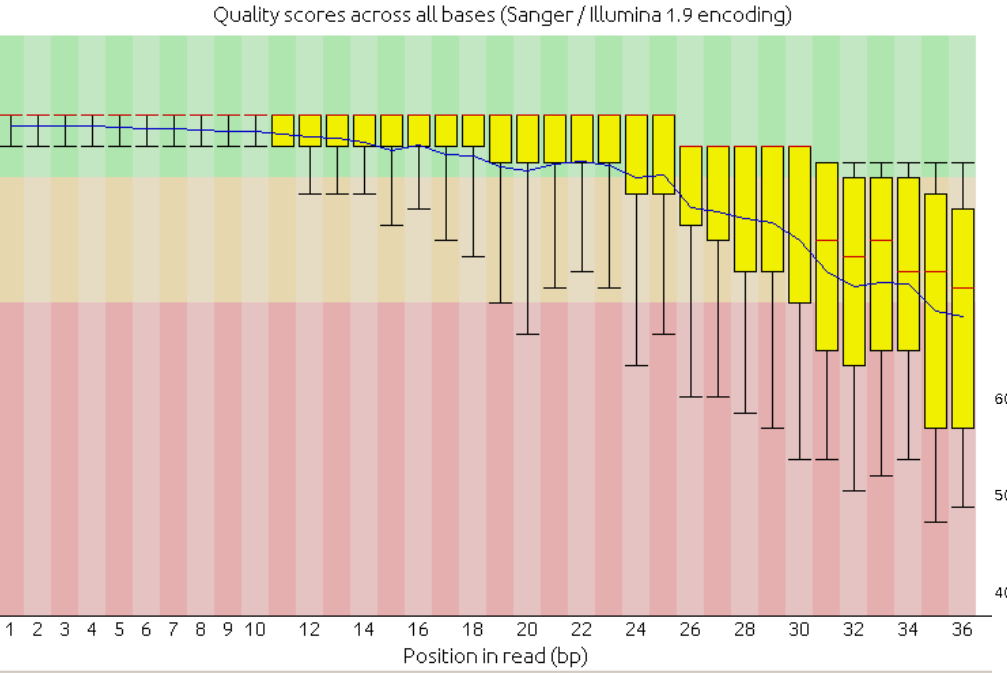
Tools: Read alignment

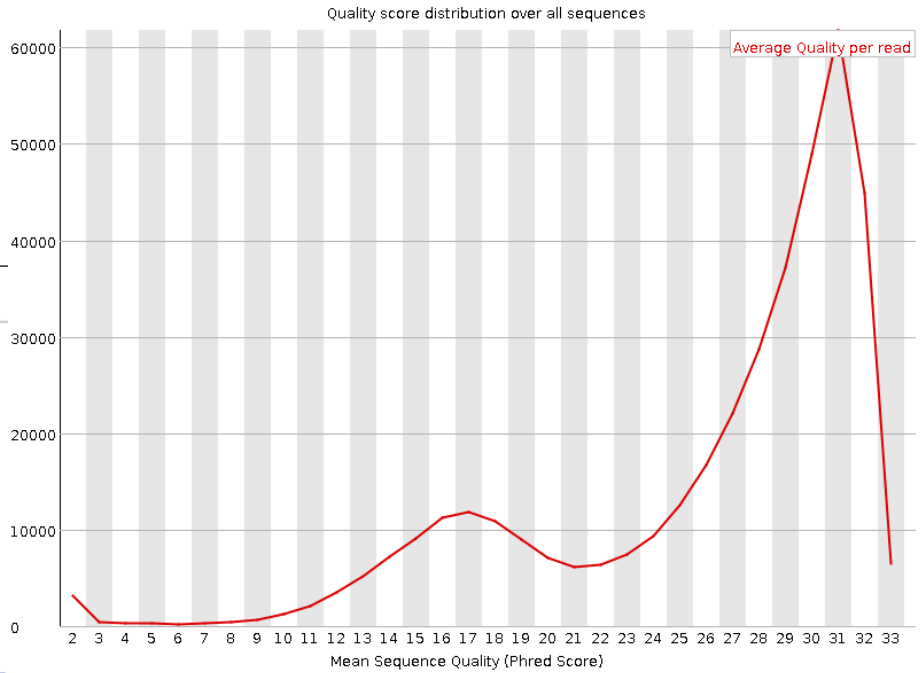
- Unspliced aligner
  - MAQ
    - Genome Res. 2008;18:1851–1858 (Cited by 2160)
  - BWA
    - Bioinformatics. 2009;25:1754–1760 (Cited by 7646)
  - Bowtie
    - Genome Biol. 2009;10:R25 (Cited by 7342)
- Spliced aligner
  - TopHat
    - Bioinformatics. 2009;25:1105–1111 (Cited by 4070)
  - STAR
    - Bioinformatics. 2013;29:15–21 (Cited by 743)
  - MapSplice
    - Nucleic Acids Res. 2010;38:e178. (Cited by 335)
  - GSNAP
    - Bioinformatics. 2010;26:873–881 (Cited by 710)

Preprocessing of Raw Data

- RNA-seq data is formatted in FASTQ
- Numerous erroneous sequence variants can be introduced during the library preparation, sequencing, and imaging steps
- QC of raw data to identify and filter out low quality reads/bases

Check quality of reads





FastQC

SAINT LOUIS UNIVERSITY.

Trim and filter reads

- Read trimming may be advisable prior to aligning the RNA-seq data
- Two common trimming strategies
  - adapter trimming: removal of the adapter sequence, typically not necessary
  - quality trimming: removes the ends of reads with low base quality scores, necessary if for SNP call

Tools: Read processing

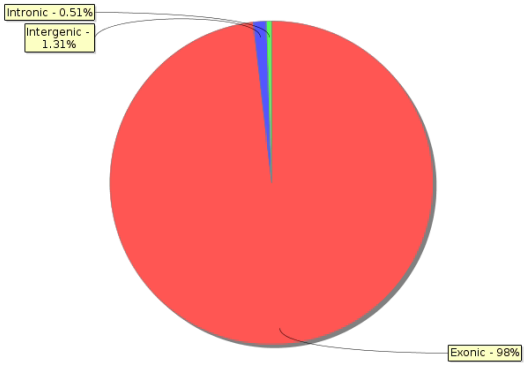
- Raw data QC
  - FastQC
    - http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
  - HTQC
    - BMC Bioinformatics. 2013;14:33 (Cited by 30)
- Read filtering and trimming
  - Trimmomatic
    - Bioinformatics. 2014 Aug 1;30(15):2114-20 (Cited by 868)
  - FASTX-Toolkit
    - http://hannonlab.cshl.edu/fastx_toolkit/
  - FLEXBAR
    - Biology (Basel) 2012;1:895–905. (Cited by 77)

Read alignment quality control

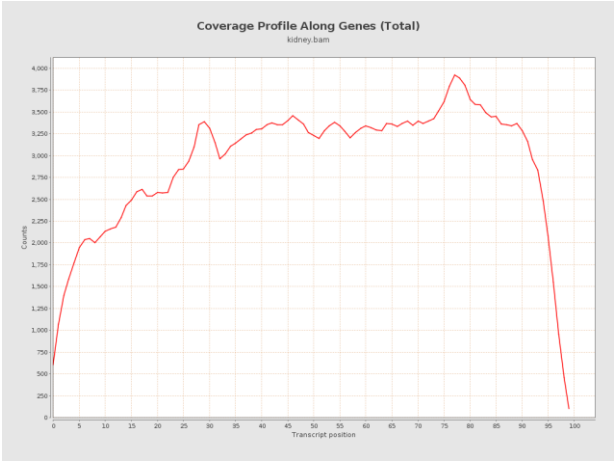
Summary	
Reads alignment	
Number of mapped reads (left/right):	12,859,872 / 12,822,712
Number of aligned pairs (without duplicates):	12,492,473
Total number of alignments:	27,150,262
Number of secondary alignments:	1,467,678
Number of non-unique alignments:	2,468,300
Aligned to genes:	20,798,807
Ambiguous alignments:	3,497,107
No feature assigned:	386,048
Not aligned:	0
Reads genomic origin	
Exonic:	20,798,807 / 98.18%
Intronic:	107,674 / 0.51%
Intergenic:	278,374 / 1.31%
Intronic/intergenic overlapping exon:	506,357 / 2.39%
Transcript coverage profile	
5' bias:	0.14
3' bias:	0.03
5'-3' bias:	2.87
Junction analysis	
Reads at junctions:	1,540,769
ACGG	4.13%
TGCT	3.88%
AGGT	3.43%
CAAC	3.25%
CGGC	3.18%
ACCA	2.9%
GCAG	2.39%
TGTG	1.97%
TAAT	1.9%
TGAA	1.89%
TGGC	1.8%

Qualimap 2 Bioinformatics. 2016 Jan 15;32(2)

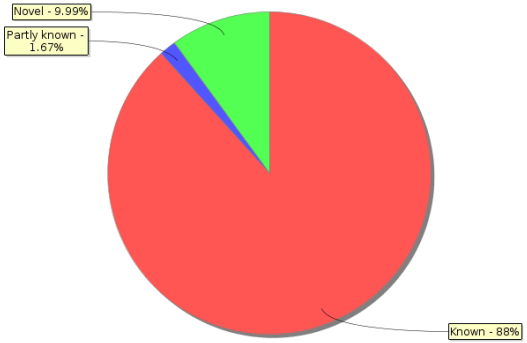
Read alignment quality control



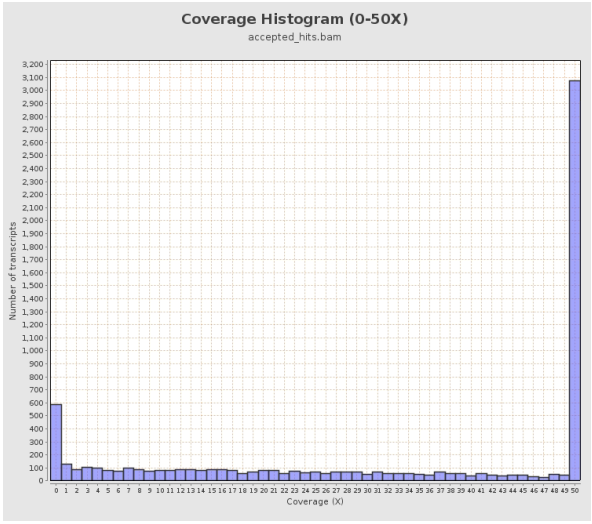
Reads Genomic Origin



Coverage Profile Along Genes (Total)



Junction Analysis



Qualimap 2 Bioinformatics. 2016 Jan 15;32(2)

Tools: Read alignment quality control

- RNA-SeQC
  - Bioinformatics. 2012;28:1530–1532. (Cited by 111)
- RSeQC
  - Bioinformatics. 2012;28:2184–2185 (Cited by 151)
- Qualimap 2
  - Bioinformatics. 2016 Jan 15;32(2):292-4 (Cited by 3)
- SAMtools
  - http://samtools.sourceforge.net/

Transcriptome Reconstruction

- Transcriptome reconstruction is the identification of all transcripts expressed in a sample
- Two strategies
  - reference-guided approach
    - alignment of raw reads to the reference
    - assembly of overlapping reads for reconstructing transcripts
  - reference-independent approach
    - uses a de novo assembly algorithm to directly build consensus transcripts

Tools: Transcriptome reconstruction

- Reference-guided
  - Cufflinks
    - Nat Biotechnol. 2010;28:511–515. (Cited by 3593)
  - Scripture
    - Nat Biotechnol. 2010;28:503–510 (Cited by 735)
  - StringTie
    - Nat Biotechnol. 2015;33:290–295 (Cited by 28)
- Reference free
  - Trinity
    - Nat Biotechnol. 2011;29:644–652 (Cited by 2696)
  - Oases
    - Bioinformatics. 2012;28:1086–1092 (Cited by 607)
  - transABySS
    - Nat Methods. 2010;7:909–912 (Cited by 427)

Expression Quantification

- Gene-level quantification
  - Requires the alignment result of reads using the transcriptome as a reference
- Isoform-level quantification
  - Requires alignment results of reads using whole genome sequences as a reference rather than the transcriptome

Tools: Expression quantification

- Gene-level quantification
  - ALEXA-seq
    - Nat Methods. 2010;7:843–847 (Cited by 195)
  - ERANGE
    - Nat Methods. 2008;5:621–628. (Cited by 5595)
  - NEUMA
    - Nucleic Acids Res. 2011;39:e9 (Cited by 84)
- Isoform-level quantification
  - Cufflinks
    - Nat Biotechnol. 2010;28:511–515. (Cited by 3593)
  - StringTie
    - Nat Biotechnol. 2015;33:290–295 (Cited by 28)
  - RSEM
    - BMC Bioinformatics. 2011;12:323 (Cited by 1118)
  - Sailfish
    - Nat Biotechnol. 2014;32:462–464 (Cited by 76)

Differential expression

- Fisher exact test :edgeR and DESeq
- Non-parametric: NOIseq and SAMseq
- T-statistic: cuffdiff
- There may be large differences between these programs and that no single method may be optimal under all experimental conditions

 Brief Bioinform. 2015 Jan;16(1):59-70

Tools: Differential expression

- Gene-level
  - NOIseq
    - Nucleic Acids Res. 2015;43:e140 (Cited by 6)
  - edgeR
    - Bioinformatics. 2010;26:139–140 (Cited by 2902)
  - DESeq
    - Genome Biol. 2010;11:R106 (Cited by 3666)
  - SAMseq
    - Stat Methods Med Res. 2013;22:519–536 (Cited by 144)
- Isoform-level
  - Cuffdiff
    - Cuffdiff 1: Nat Biotechnol. 2010;28:511–515. (Cited by 3593)
    - Cuffdiff 2: Nat Biotechnol. 2013 Jan;31(1):46-53 (Cited by 680)
  - EBSeq
    - Bioinformatics. 2013;29:1035–1043 (Cited by 171)
  - Ballgown
    - Nat Biotechnol. 2015;33:243–246 (Cited by 4)

Tuxedo tools

Bowtie
Extremely fast, general purpose short read aligner

TopHat
Aligns RNA-Seq reads to the genome using Bowtie
Discovers splice sites

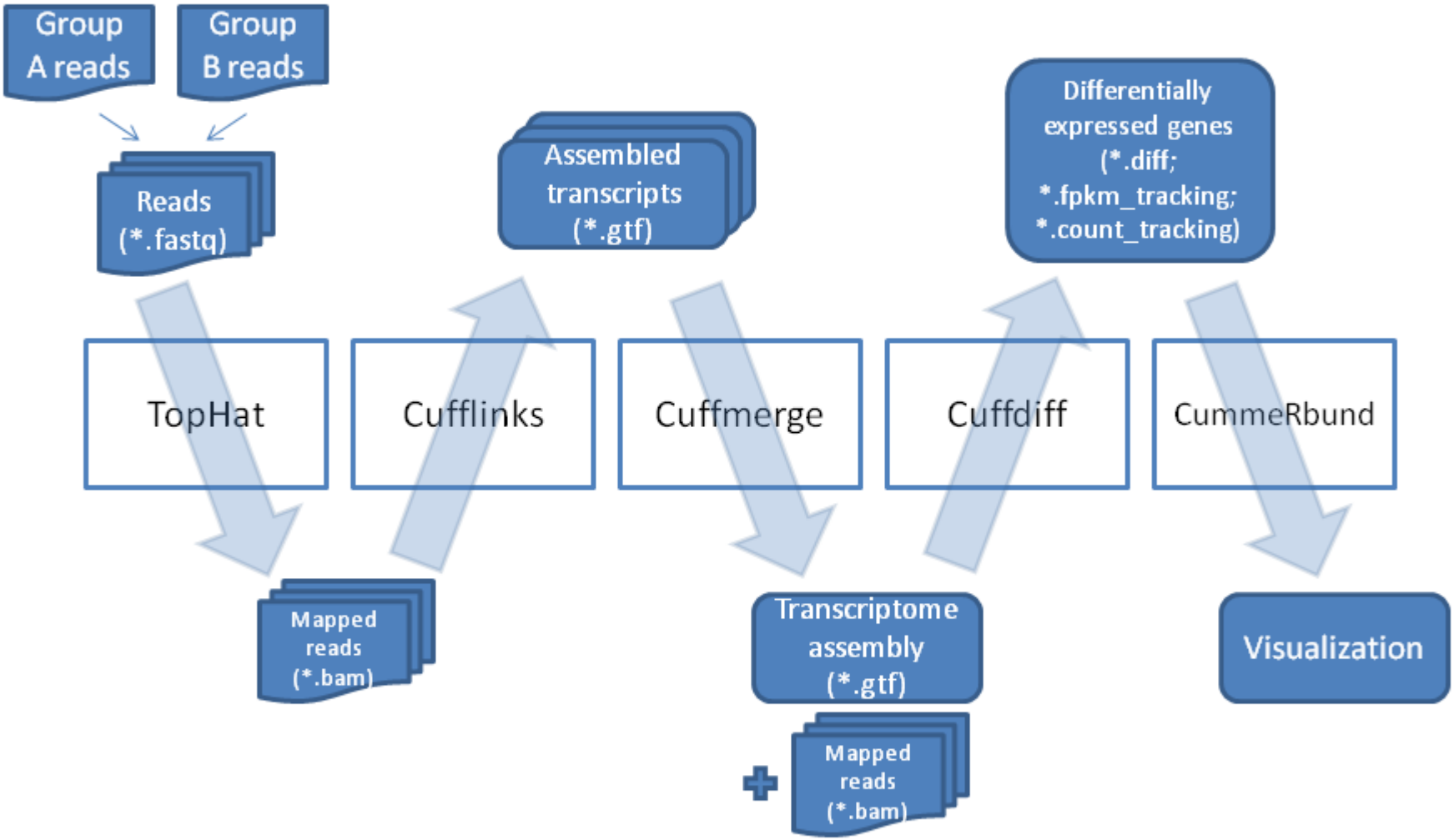
Cufflinks package

- **Cufflinks**
  Assembles transcripts
- **Cuffcompare**
  Compares transcript assemblies to annotation
- **Cuffmerge**
  Merges two or more transcript assemblies
- **Cuffdiff**
  Finds differentially expressed genes and transcripts
  Detects differential splicing and promoter use

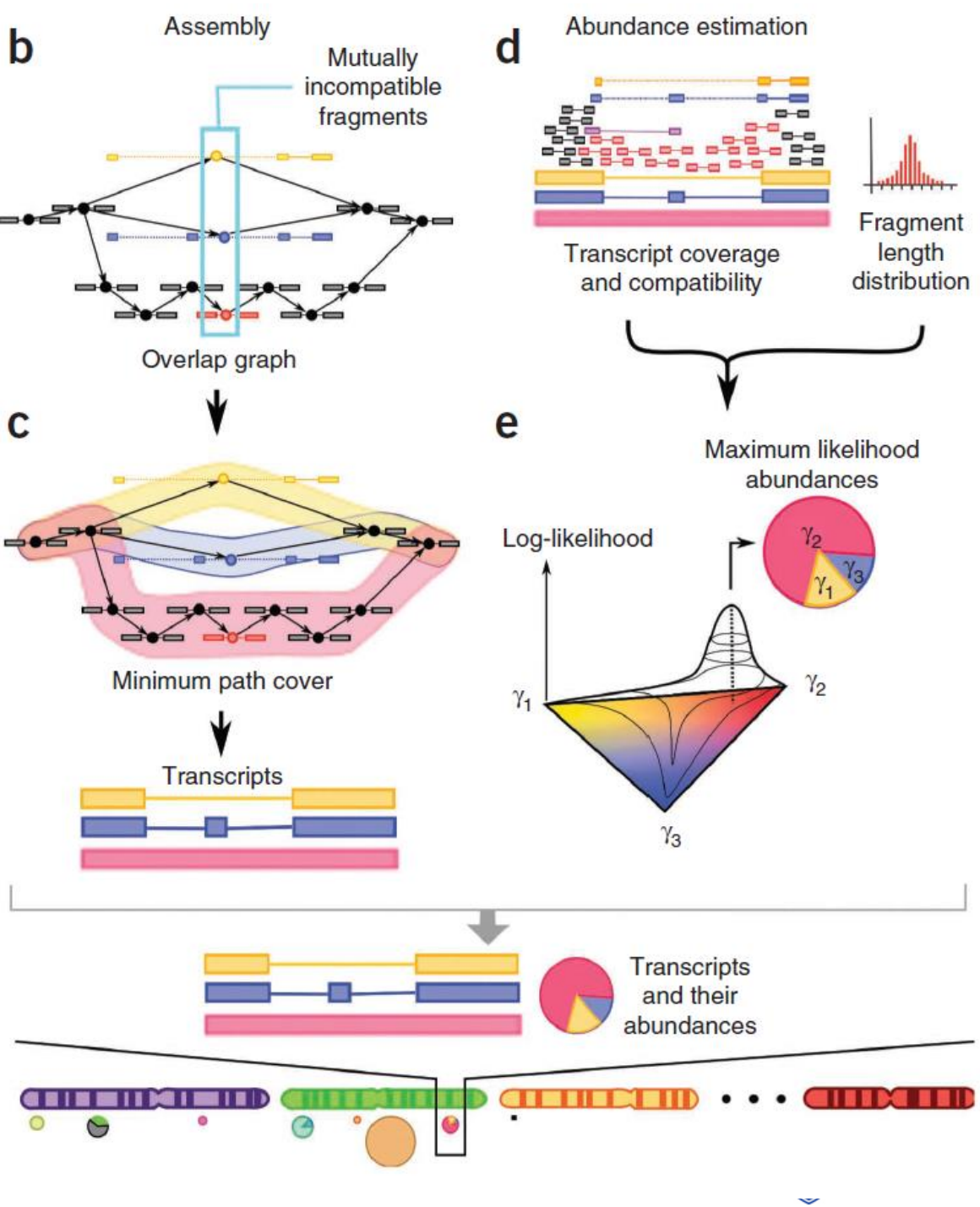
CummeRbund
Plots abundance and differential expression results from Cuffdiff

Nature Protocols 7, 562–578 (2012)

Tuxedo suite for RNA-seq differential expression analysis



http://www.labome.com/method/RNA-seq-Using-Next-Generation-Sequencing.html



Cufflinks

- Building an overlap graph from the mapped reads
- Computing minimal path cover in the overlap graph,
- Generating a minimum number of transcripts that will explain all reads in the graph
- Abundance estimation is performed by estimating the maximum likelihood abundance
- Reported isoform expression level in FPKM for paired-end and RPKM for a single-end

Nat Biotechnol. 2010;28:511-515.

Cuffmerge

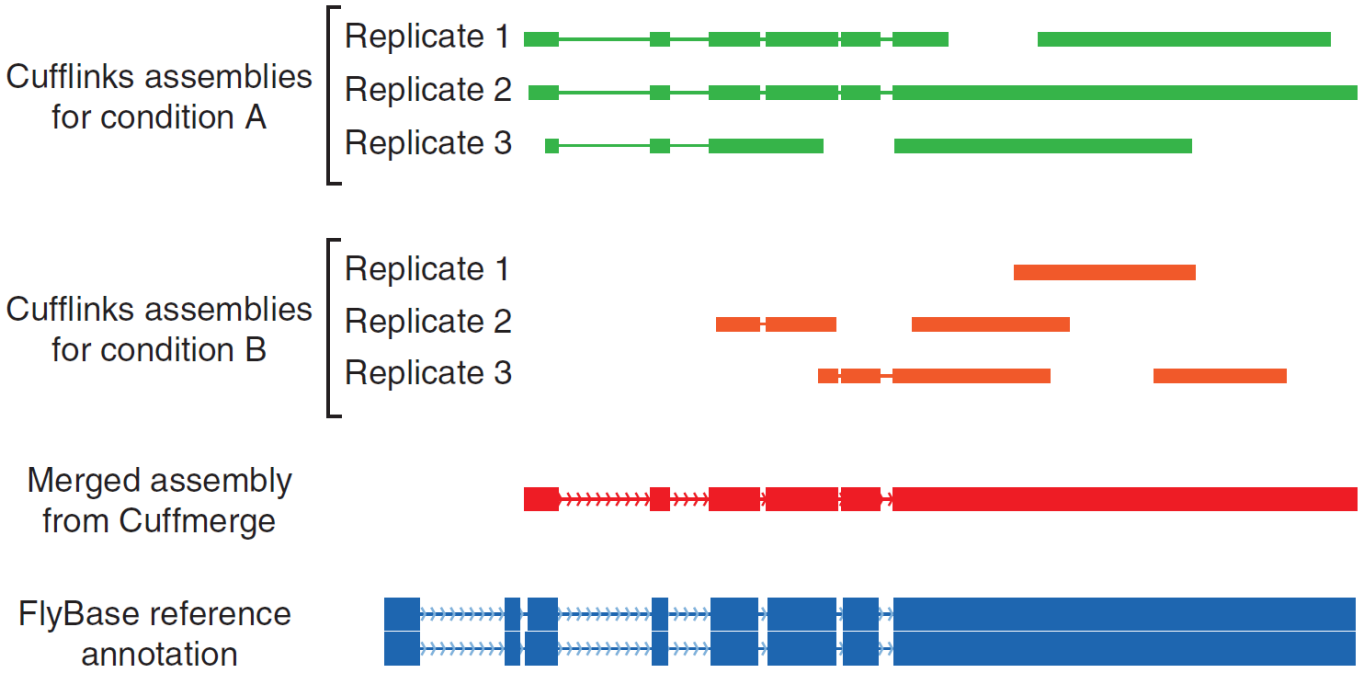
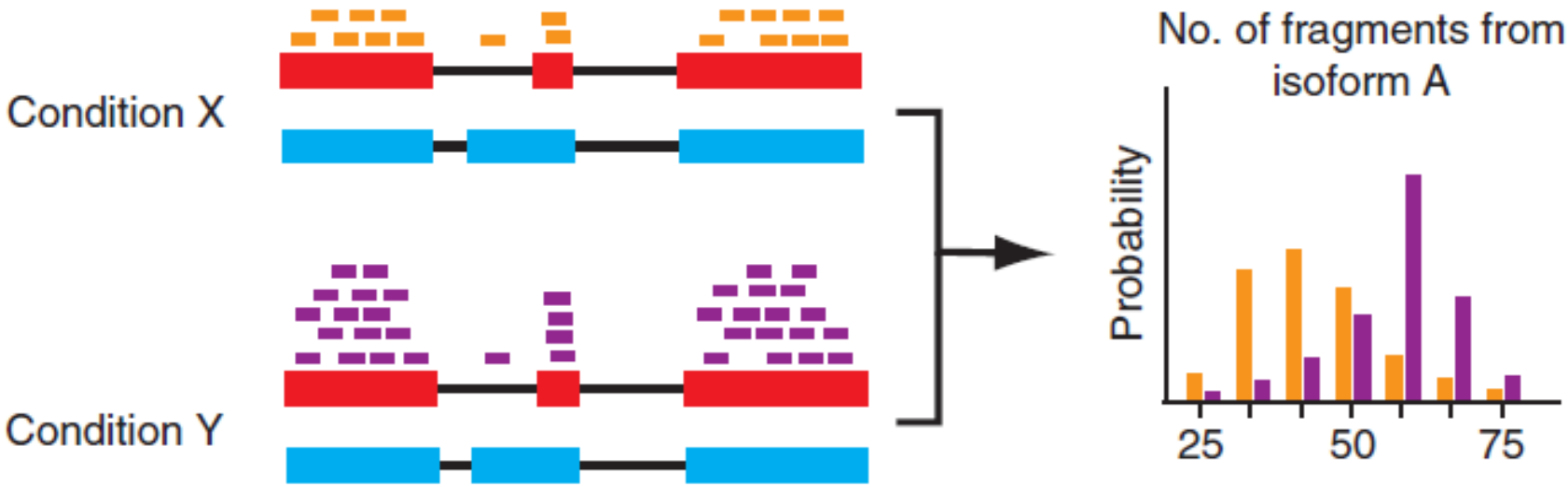


Figure 3 | Merging sample assemblies with a reference transcriptome annotation. Genes with low expression may receive insufficient sequencing depth to permit full reconstruction in each replicate. However, merging the replicate assemblies with Cuffmerge often recovers the complete gene. Newly discovered isoforms are also integrated with known ones at this stage into more complete gene models.

Trapnell et al, Nature protocols, 2012

Cuffdiff



5) Test for signficance of changes between conditions in transcript-level counts

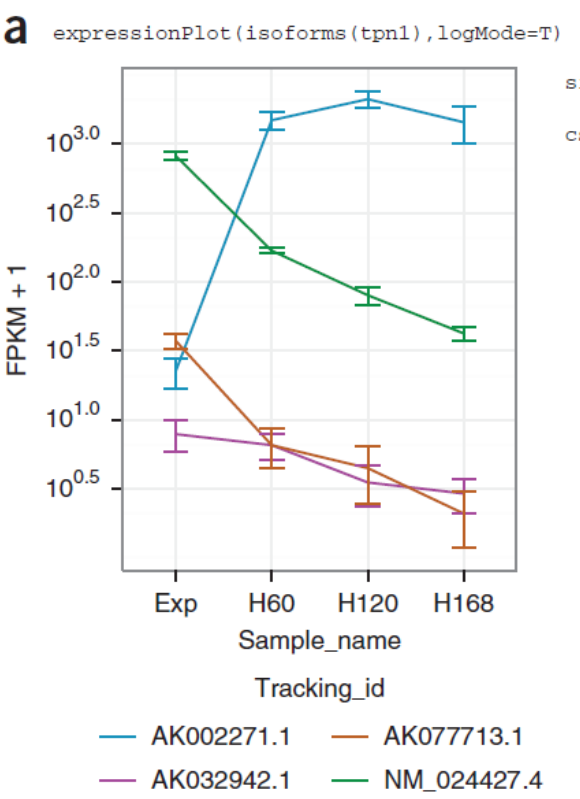
Nat Biotechnol. 2013 Jan;31(1):46-53

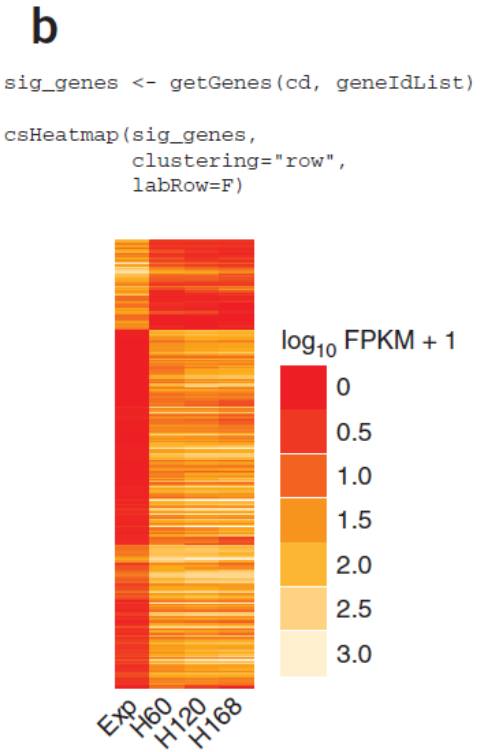
CummeRbund

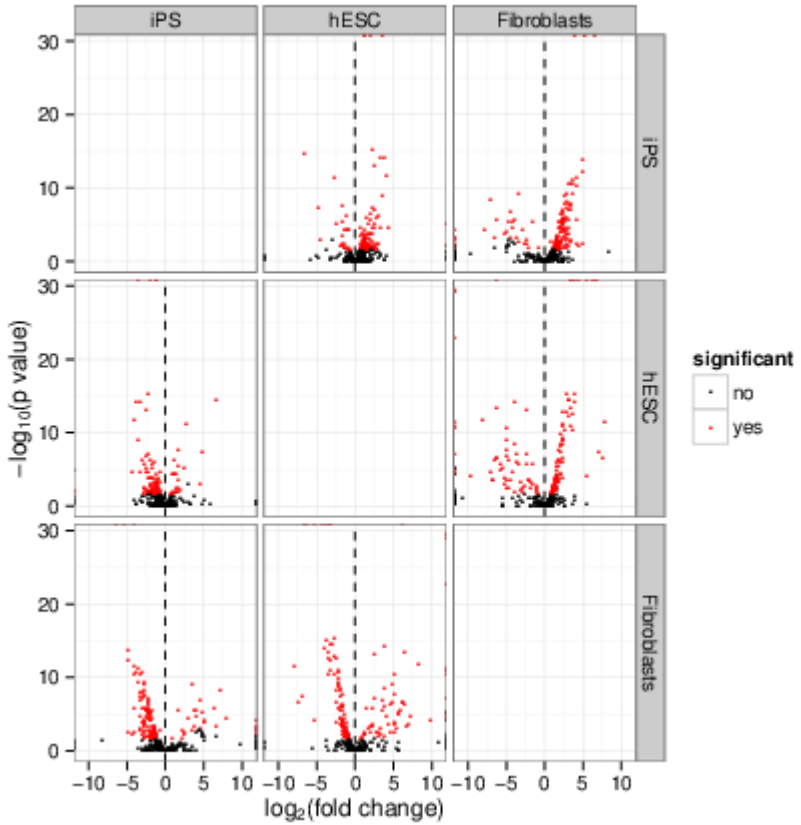
- Explore their expression data and create publication-ready plots of differentially expressed and regulated genes
- Visualize differential expression at the isoform level
- Broad patterns among large sets of genes

http://compbio.mit.edu/cummeRbund/index.html

CummeRbund

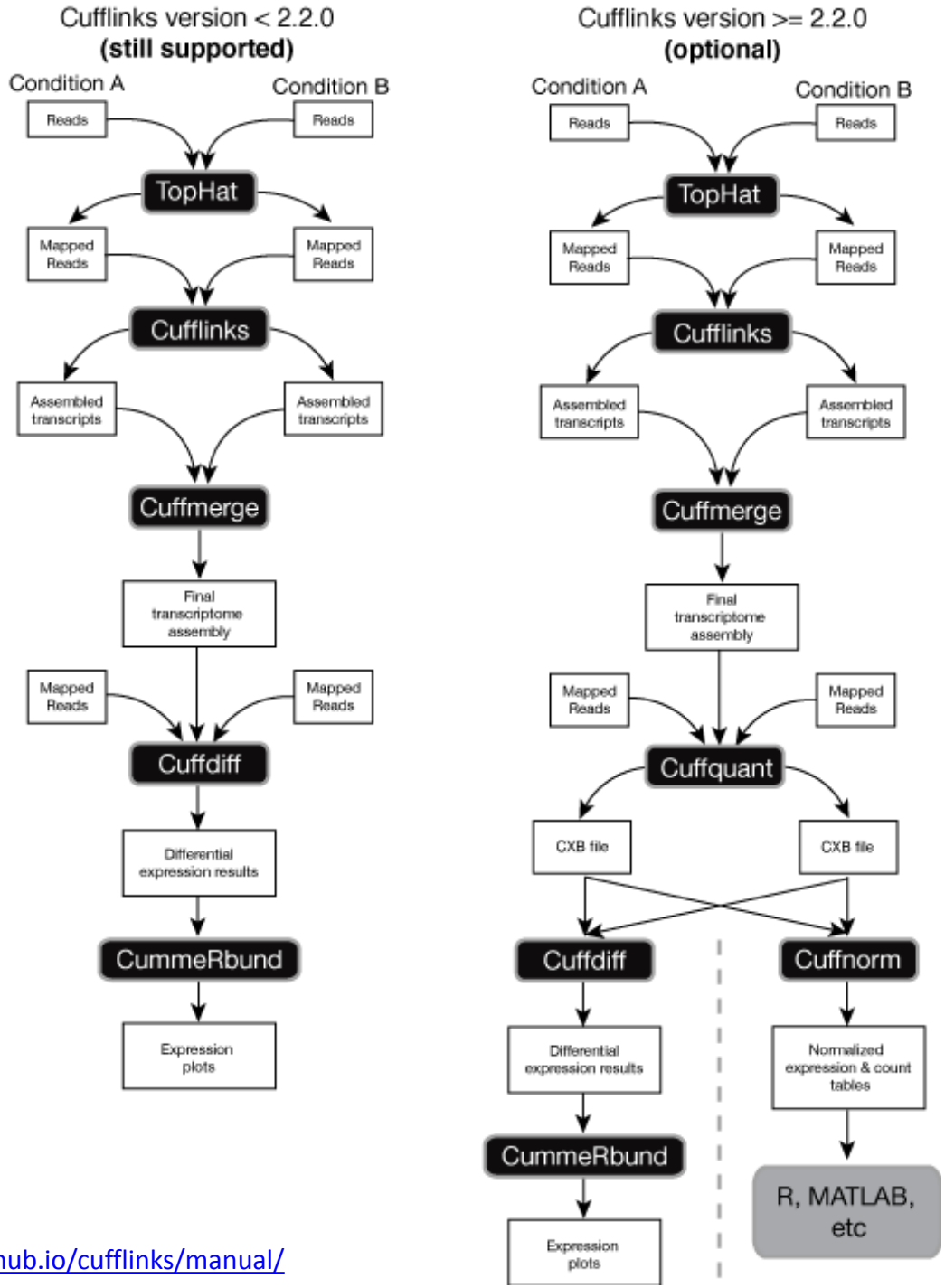






Trapnell et al, Nature protocols, 2012

SAINT LOUIS UNIVERSITY.



http://cole-trapnell-lab.github.io/cufflinks/manual/

Suggested reading: Tuxedo protocol

PROTOCOL

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

Cole Trapnell[1,2], Adam Roberts[3], Loyal Goff[1,2,4], Geo Pertea[5,6], Daehwan Kim[5,7], David R Kelley[1,2], Harold Pimentel[3], Steven L Salzberg[5,6], John L Rinn[1,2] & Lior Pachter[3,8,9]

[1]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [2]Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, Massachusetts, USA. [3]Department of Computer Science, University of California, Berkeley, California, USA. [4]Computer Science and Artificial Intelligence Lab, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [5]Department of Medicine, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA. [6]Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland, USA. [7]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland, USA. [8]Department of Mathematics, University of California, Berkeley, California, USA. [9]Department of Molecular and Cell Biology, University of California, Berkeley, California, USA. Correspondence should be addressed to C.T. (cole@broadinstitute.org).

Published online 1 March 2012; corrected after print 7 August 2014; doi:10.1038/nprot.2012.016

Recent advances in high-throughput cDNA sequencing (RNA-seq) can reveal new genes and splice variants and quantify expression genome-wide in a single assay. The volume and complexity of data from RNA-seq experiments necessitate scalable, fast and mathematically principled analysis software. TopHat and Cufflinks are free, open-source software tools for gene discovery and comprehensive expression analysis of high-throughput mRNA sequencing (RNA-seq) data. Together, they allow biologists to

ved.

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 2012;7:562-578.

Next lecture

- RNA-seq data retrieval and quality control