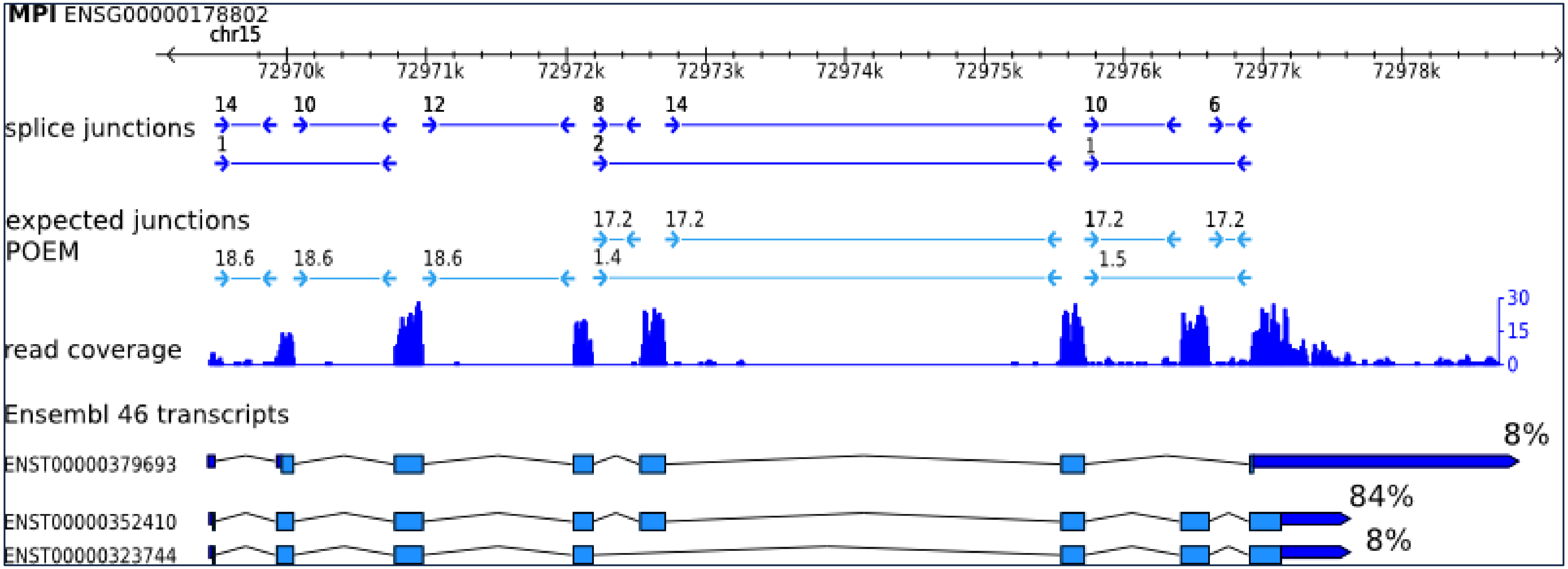
Problems with traditional gene counting software

1. Multi-mapping reads not used, leading to underestimation of gene abundances, particularly for genes with more shared sequence
2. Some genes are completely unquantifiable with featureCounts due to duplication (e.g. paralogous genes)
3. Genes that change relative isoform usage can have erroneous results due to changes in isoform length
4. Uses a lot of hard disk space

Don’t use STAR/featureCounts at transcript level

If you want to count at transcript level, many more reads will now be ambiguous due to shared sequence, and will be discarded



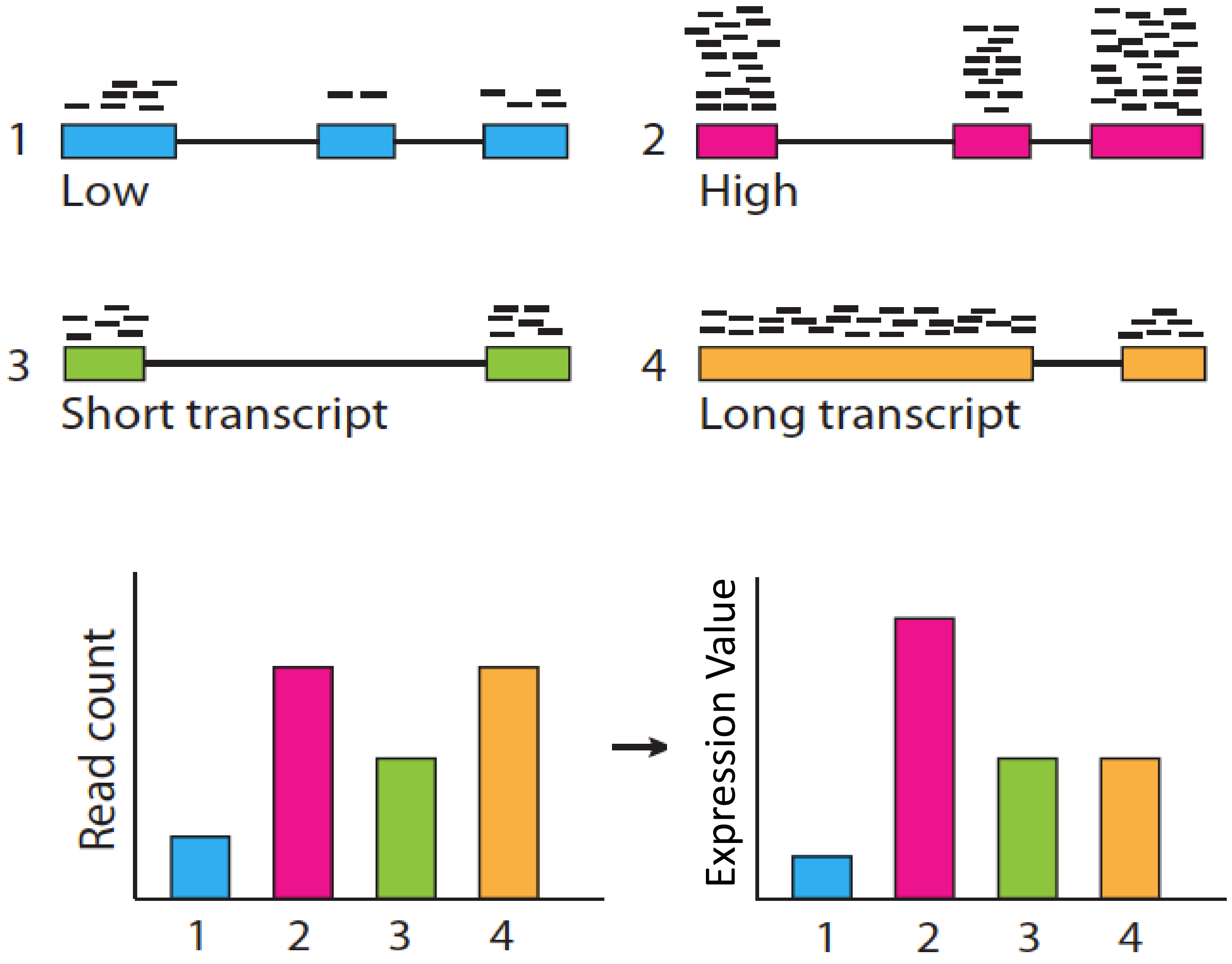
http://www.cs.cmu.edu/~maschulz/projects.html

Today's Topics

- Transcriptomics
- Traditional RNA-Seq Methods
  - Sequencing & experimental considerations
  - Traditional gene counting
  - Gene quantification
  - Statistics
- Where to find help

ILLINOIS
Roy J. Carver Biotechnology Center

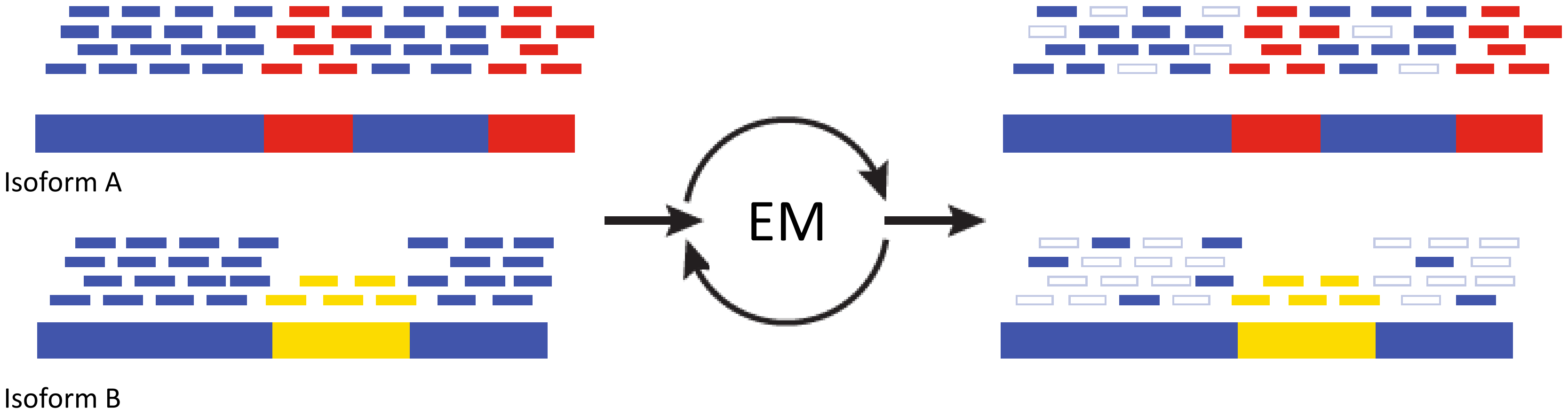
Calculating expression of genes and transcripts



Slide courtesy of Cole Trapnell

ILLINOIS
Roy J. Carver Biotechnology Center

Solution: Expectation Maximization algorithms



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:

- Cuffdiff
- RSEM
- eXpress
- Salmon
- Kallisto

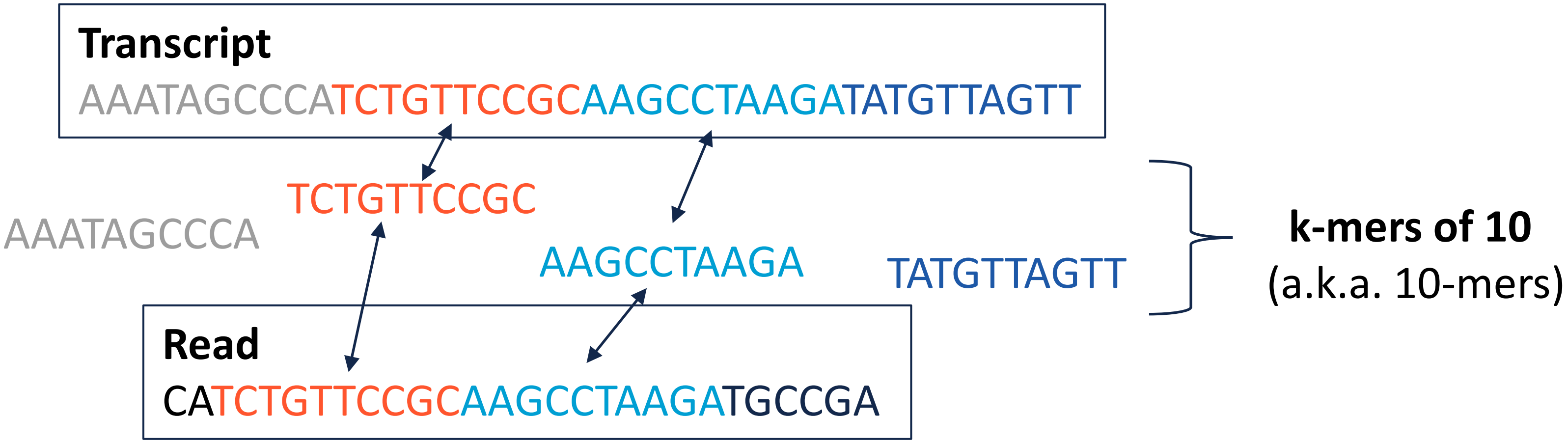
slide from Brian Haas; title modified

Quasi-mapping / pseudoalignment methods

- Use transcriptome sequences, rather than genomic sequences and annotations
  - No need for splice-aware alignment
  - Indexed, and focused on a smaller amount of sequence
- Very fast
- Better for distinguishing alternative isoforms (differently spliced transcripts from same gene)

Quasi-mapping / pseudoalignment methods

Instead of comparing whole reads to the transcriptome, finds k-mers in common between reads and transcripts



Transcript

AAATAGCCCATCTGTTCCGCAAGCCTAAGATATGTTAGTT

AAATAGCCCA TCTGTTCCGC AAGCCTAAGA TATGTTAGTT

k-mers of 10

(a.k.a. 10-mers)

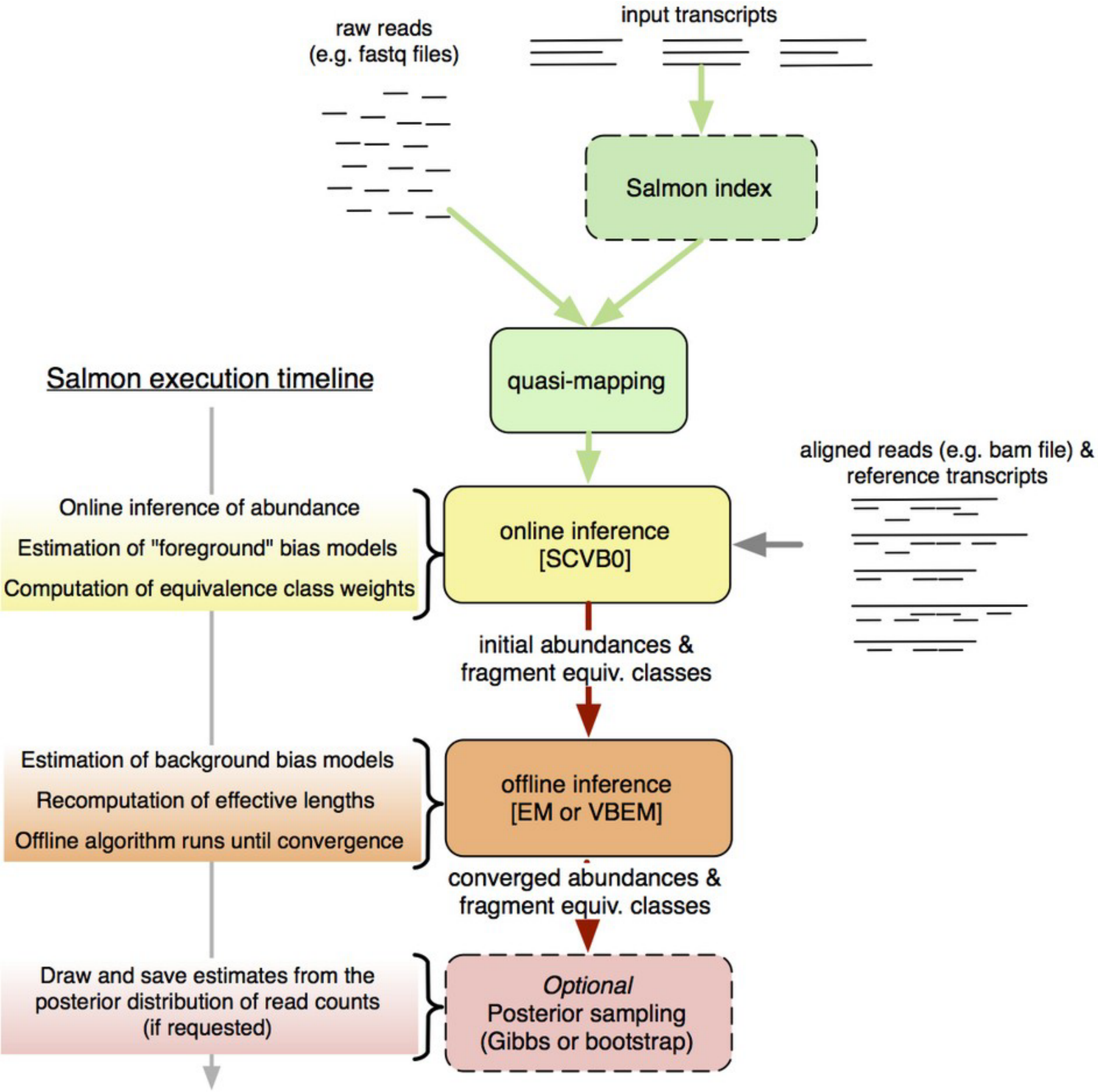
Read

CATCTGTTCCGCAAGCCTAAGATGCCGA

Salmon

- Patro, Duggal, Love, Irizarry, and Kingsford (2017)
- Adjusts for GC content, position within transcript, and other sources of bias
- Can start with raw reads or BAM files
- Can estimate uncertainty in transcript abundances

Salmon



raw reads
(e.g. fastq files)
input transcripts
Salmon index
quasi-mapping
Salmon execution timeline
aligned reads (e.g. bam file) &
reference transcripts
Online inference of abundance
Estimation of "foreground" bias models
Computation of equivalence class weights
online inference
[SCVB0]
initial abundances &
fragment equiv. classes
Estimation of background bias models
Recomputation of effective lengths
Offline algorithm runs until convergence
offline inference
[EM or VBEM]
converged abundances &
fragment equiv. classes
Draw and save estimates from the
posterior distribution of read counts
(if requested)
Optional
Posterior sampling
(Gibbs or bootstrap)

Fig. from Patro, R., Duggal, G., Love, M. *et al.* https://doi.org/10.1038/nmeth.4197

Salmon counts

- Quantifies at the **transcript level** rather than the gene level *(multiple transcripts per gene)*
- Transcript counts, as well as abundances **adjusted by transcript length**
- Transcript counts are **non-integers** because multi-mapping reads are partially assigned to multiple transcripts
- These counts **can be grouped to the gene-level**, which improves accuracy (even more than traditional methods like STAR + featureCounts)

When to use either method

Gene Quantification (Salmon)	Traditional Gene Counting (STAR + featureCounts)
By default, since counts grouped by gene are the most accurate	Reads with retained introns (e.g. cancer and rapidly developing tissues like embryos) that you'd like to count (consider that they may be low quality)
Genome duplications present	Need to find novel transcripts/splice junctions
Lots of gene families present	Want to visualize alignments on genome
When ever you have a large percentage (>15%) of multi-mapped reads	

Today’s Topics

- Transcriptomics
- Traditional RNA-Seq Methods
  - Sequencing & experimental considerations
  - Traditional gene counting
  - Gene quantification
  - Statistics
- Where to find help

DGE Statistical Analyses

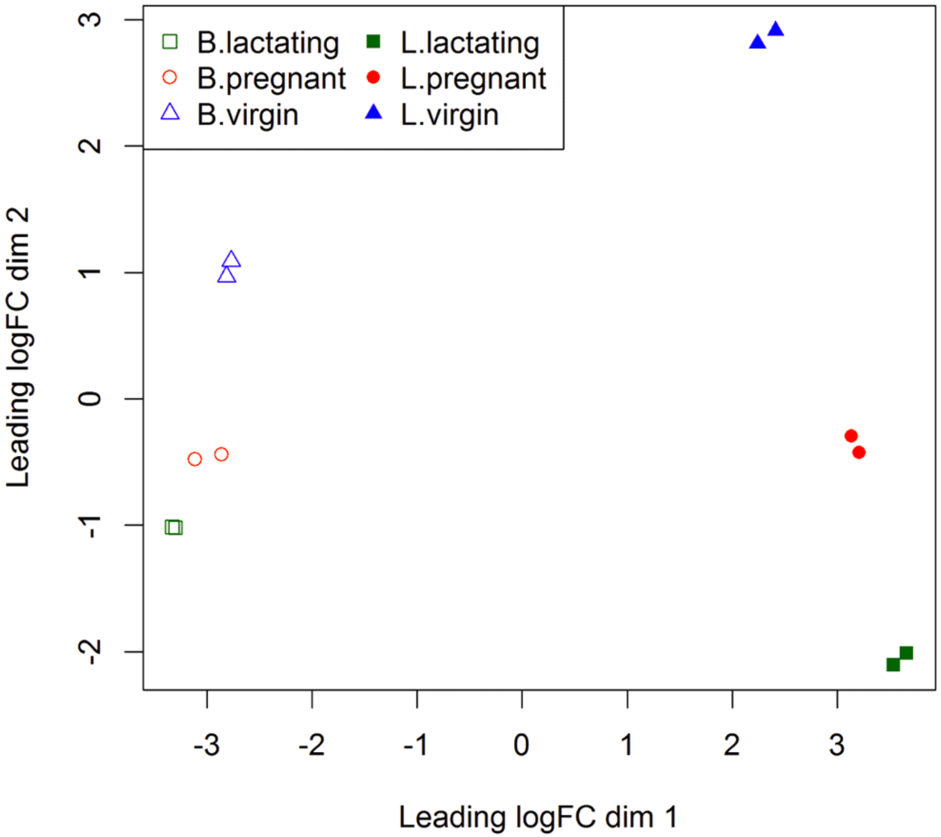
1. The first step is proper normalization of the data
   - ✧ Often the statistical package you use will have a normalization method that it prefers and uses exclusively (e.g. Voom, FPKM, TMM (used by EdgeR))
2. Is your experiment a pairwise comparison?
   - ✧ Ballgown, EdgeR, DESeq
3. Is it a more complex design?
   - ✧ limma, EdgeR, DESeq, and other R/Bioconductor packages

Slide courtesy of Rhadika Khetani, text edited

Statistical Results

- A list of significantly differentially expressed genes
- Venn Diagrams
- Heatmaps
- WGCNA
- Advanced annotation
- ... and more!

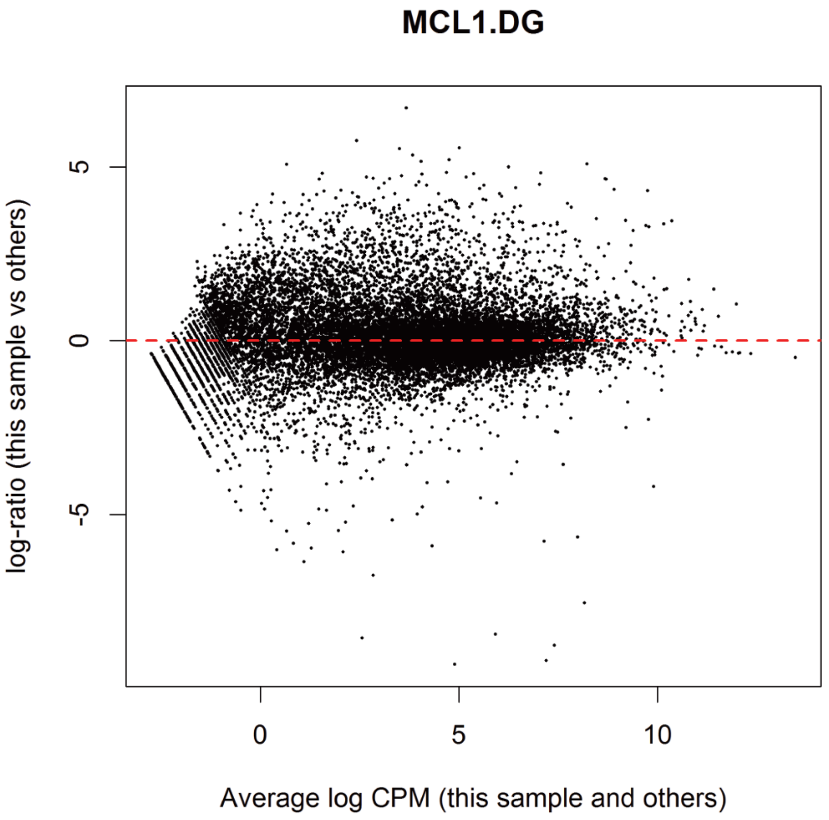
MDS Plot



https://f1000research.com/articles/5-1438 (doi: 10.12688/f1000research.8987.2)

Slide courtesy of Jenny Drnevich

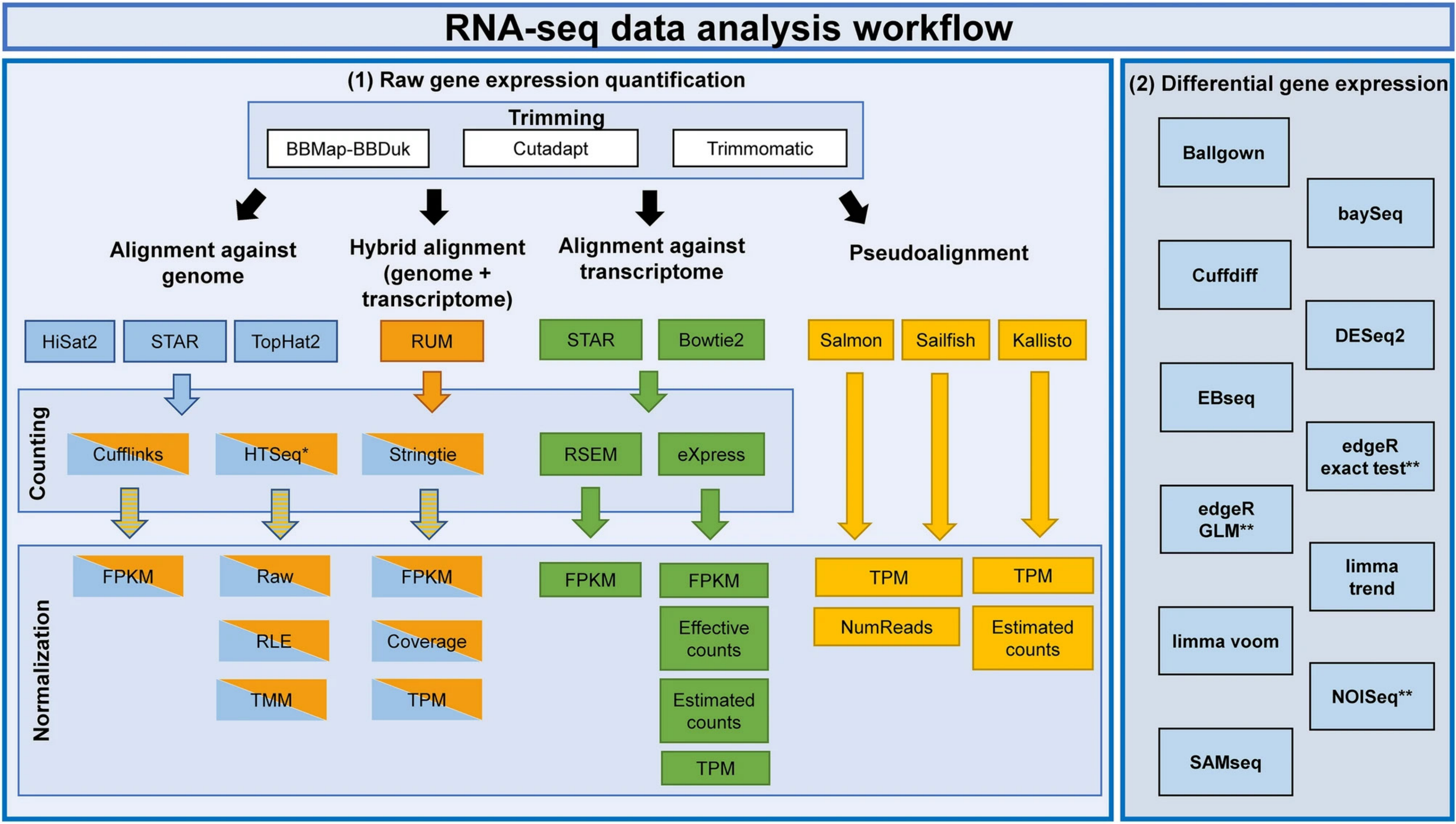
MD Plot



https://f1000research.com/articles/5-1438 (doi: 10.12688/f1000research.8987.2)

Slide courtesy of Jenny Drnevich

So many options!



Corchete, L.A., Rojas, E.A., Alonso-López, D. *et al.* Systematic comparison and assessment of RNA-seq procedures for gene expression quantitative analysis. *Sci Rep* **10,** 19737 (2020). https://doi.org/10.1038/s41598-020-76881-x

What does HPCBio use?

- Quality Check - **FASTQC** or **fastp**
- Trimming - **Trimmomatic** or **fastp**
- Splice-aware alignment - **STAR**
- Bacterial alignment - **BWA** or **Novoalign**
- Counting reads per gene - **featureCounts**
- Counting reads per isoform - **Salmon*** *Can also group these counts by gene for even more accuracy*
- DGE Analysis - **limma** and/or **edgeR**
- De novo transcriptome assembly – **Trinity**
- Reference-based transcriptome assembly – **StringTie**

Still not sure?

Recent RNA-Seq software comparison articles:

Corchete, L.A., Rojas, E.A., Alonso-López, D. *et al.* Systematic comparison and assessment of RNA-seq procedures for gene expression quantitative analysis. *Sci Rep* **10,** 19737 (2020). https://doi.org/10.1038/s41598-020-76881-x

Schaarschmidt, S., Fischer, A., Zuther, E., & Hincha, D. K. (2020). Evaluation of Seven Different RNA-Seq Alignment Tools Based on Experimental Data from the Model Plant *Arabidopsis thaliana*. *International journal of molecular sciences*, *21*(5), 1720. https://doi.org/10.3390/ijms21051720

Zhang, C., Zhang, B., Lin, LL. *et al.* Evaluation and comparison of computational tools for RNA-seq isoform quantification. *BMC Genomics* **18,** 583 (2017). https://doi.org/10.1186/s12864-017-4002-1

Today's Topics

- Transcriptomics
- Traditional RNA-Seq Methods
  - Sequencing & experimental considerations
  - Traditional gene counting
  - Gene quantification
  - Statistics
- Where to find help

How do I learn more about these steps?

- Your lab will briefly go through some R code that we use to perform DGE analysis on Salmon transcript counts
- HPCBio do offer a longer detailed workshop on the entire RNA-Seq pipeline
    - Bulk RNA-Seq Analysis workshop
- Check https://biotech.illinois.edu/hpcbio-core/hpcbio-workshop/ for updates

Documentation and Support

<u>Online resources for RNA-Seq analysis questions</u>

- Software manuals
    - Most tools also have a dedicated lists/forums and/or github pages
- Biostar (Bioinformatics explained) - http://www.biostars.org/
- SEQanswers (the next generation sequencing community) - http://seqanswers.com/

Reproducible Notebook

- You should have a (virtual) computational notebook like you have a lab notebook
- Every detail of the data analysis needs to be recorded so that **you** or anyone else could reproduce the end results
  - Software, versions, options specified
  - All data manipulations/normalizations/transformations
  - Exact statistical methods used
  - How each figure/table was calculated and made
  - Anything else?

HPCBio Bioinformatics Consulting

Contact us at:

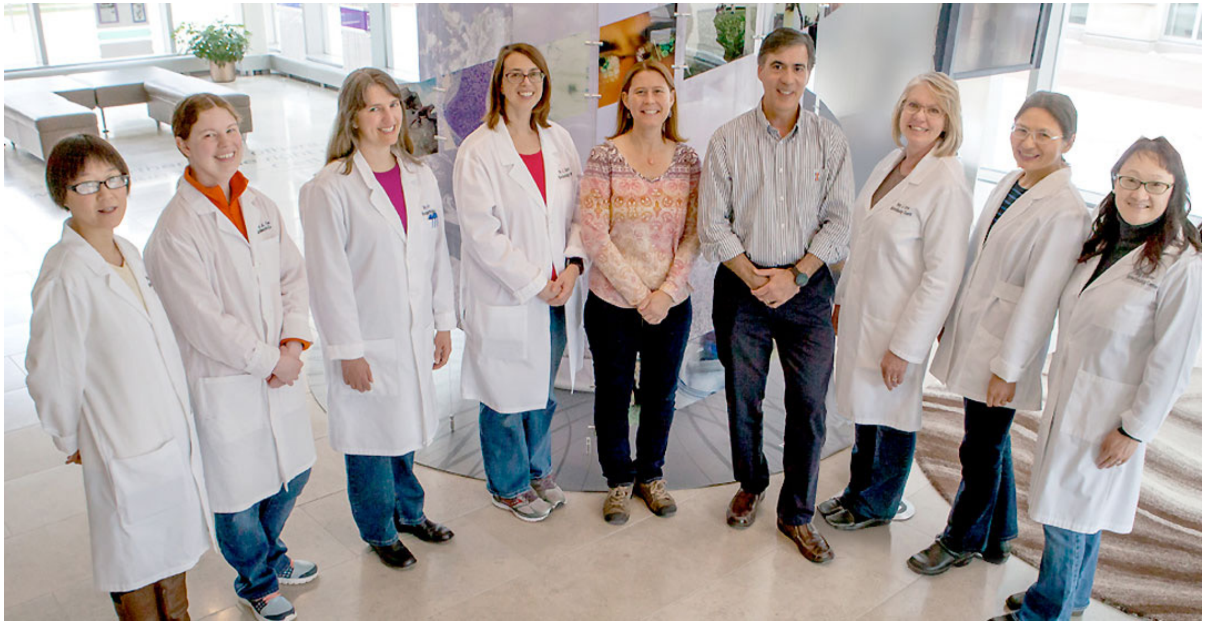
Help desk - hpcbio@biotech.illinois.edu

Training questions - hpcbio-training@biotech.illinois.edu

HPCBio website - https://biotech.illinois.edu/hpcbio-core/

My email - jholmes5@illinois.edu

DNA Services Laboratory



Director: Alvaro Hernandez

aghernan@Illinois.edu

329 ERML

217-244-3480

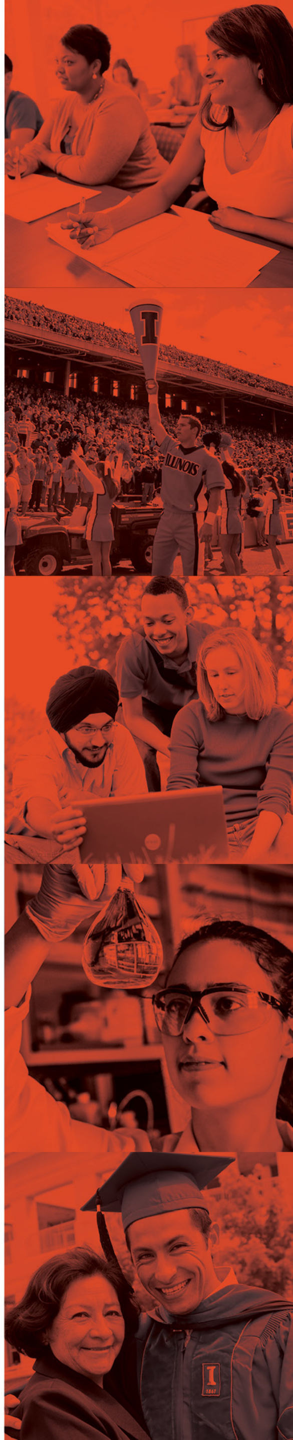
Associate Director: Chris Wright

clwright@Illinois.edu

334 ERML

217-333-4372

https://biotech.illinois.edu/dna-services-core/



Thank you!