

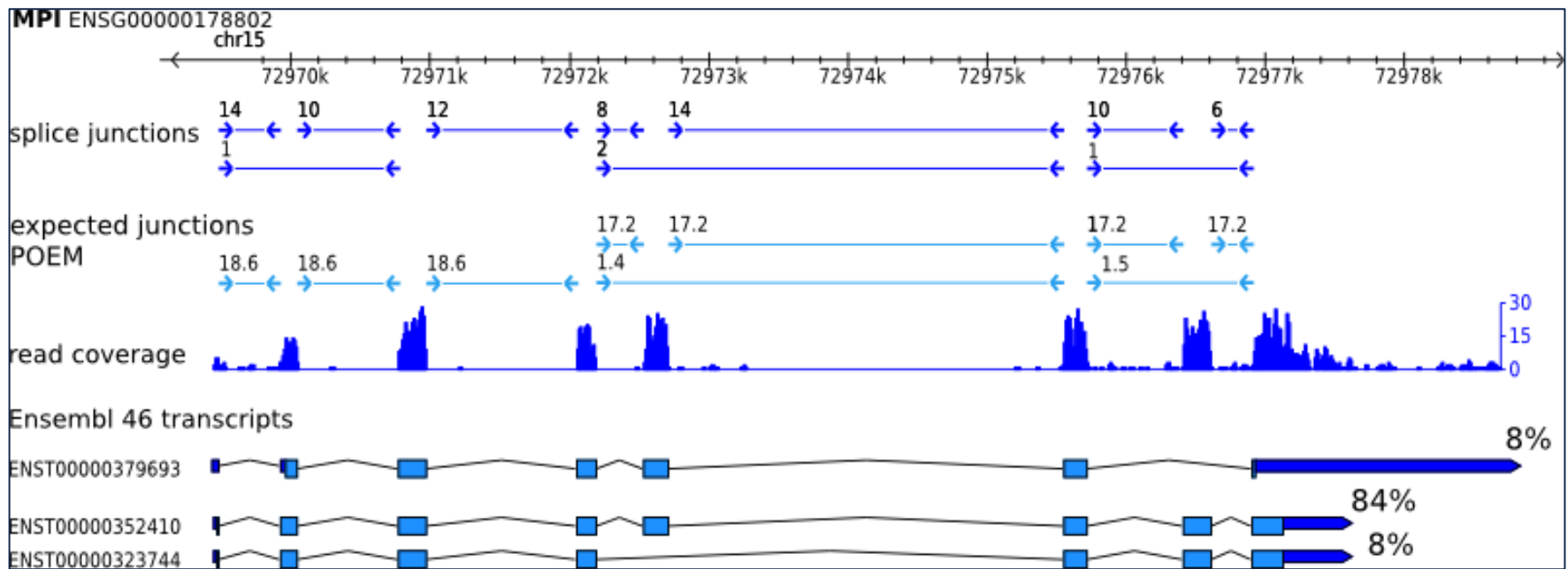
Problems with traditional gene counting software

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>



I ILLINOIS

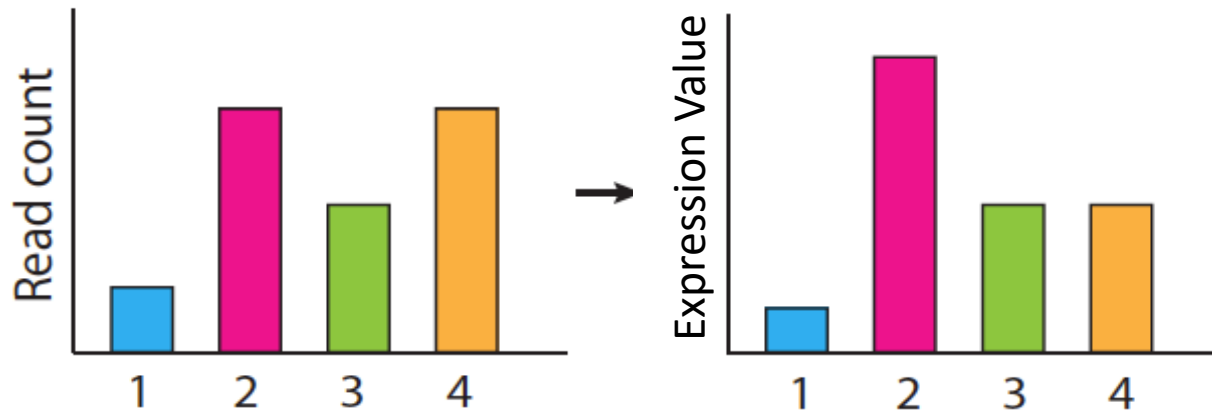
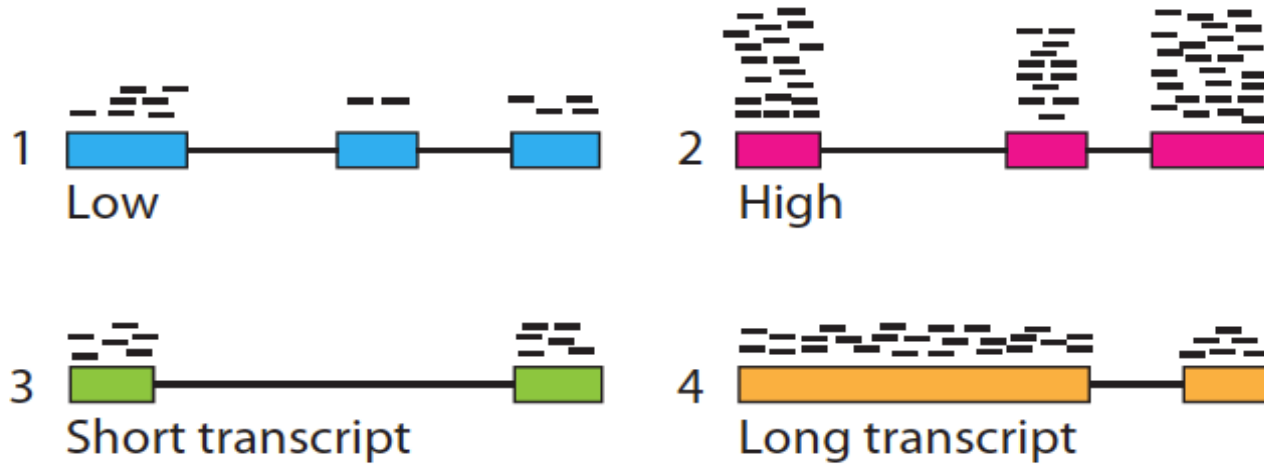
Roy J. Carver Biotechnology Center

Today's Topics

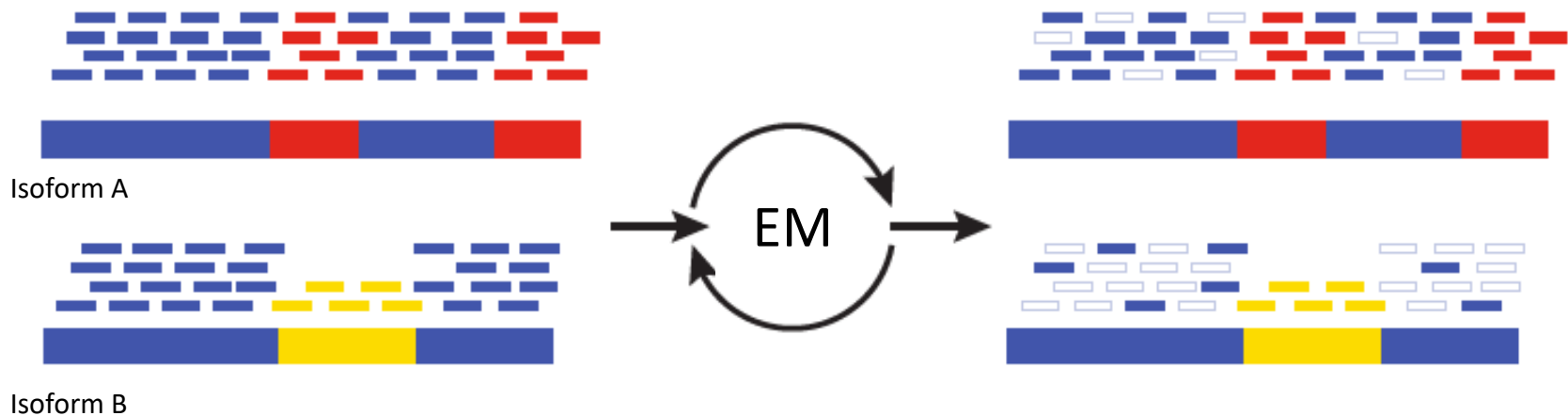
- Transcriptomics
- Traditional RNA-Seq Methods
 - Sequencing & experimental considerations
 - Traditional gene counting
 - Gene quantification
 - Statistics
- Where to find help



Calculating expression of genes and transcripts



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



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



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



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



I ILLINOIS

Roy J. Carver Biotechnology Center

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

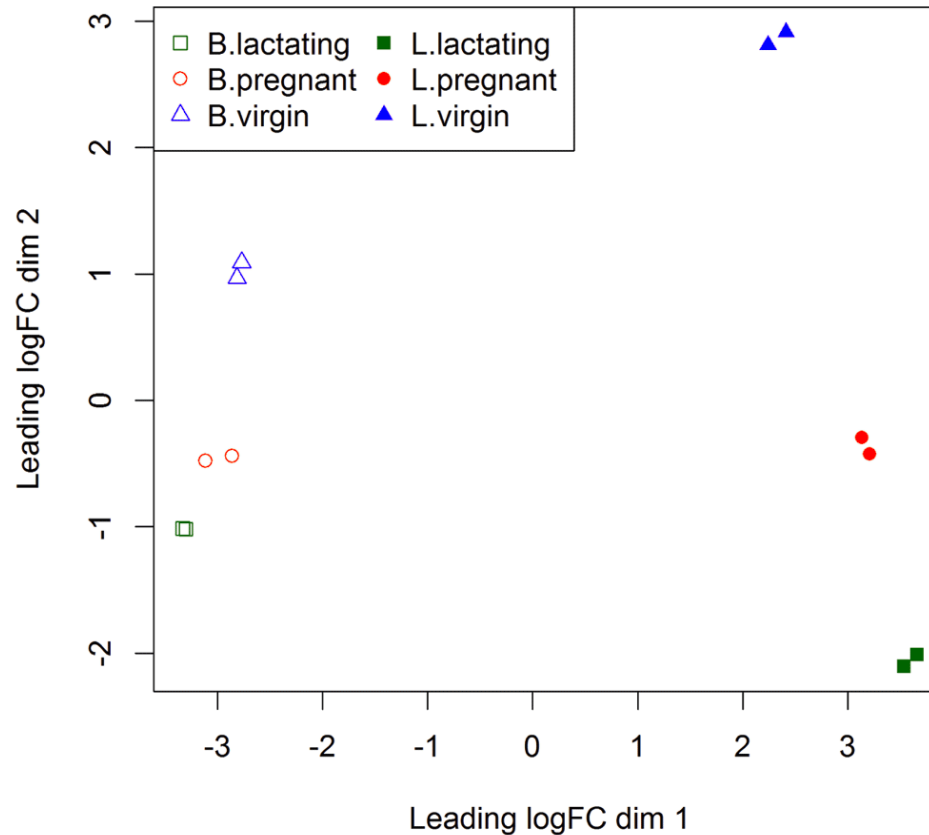


Statistical Results

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



MDS Plot

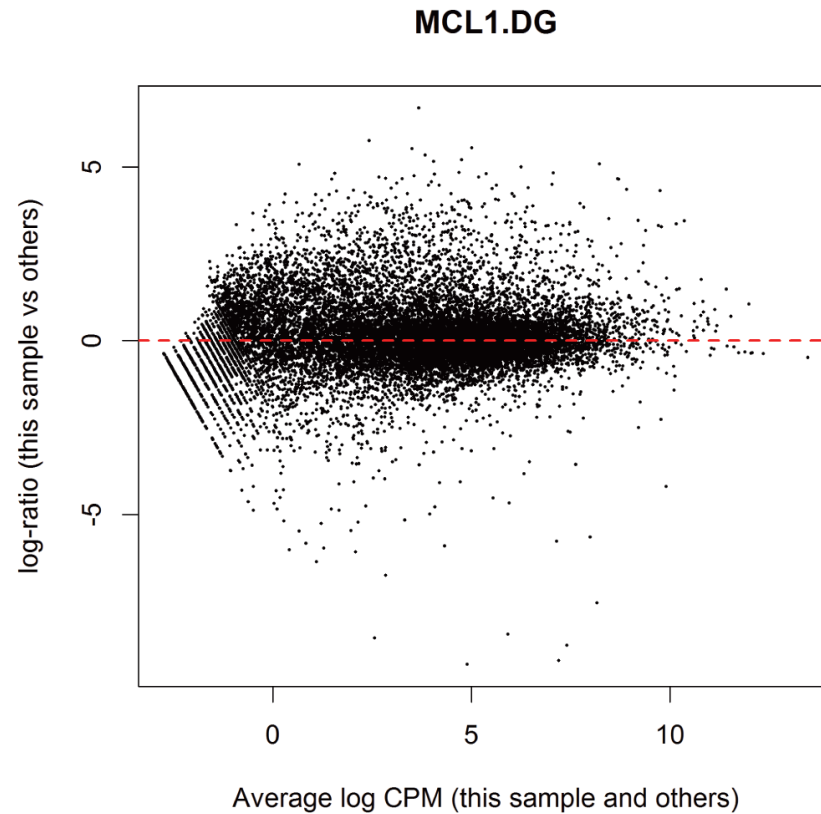


<https://f1000research.com/articles/5-1438> (doi: 10.12688/f1000research.8987.2)



Slide courtesy of Jenny Drnevich

MD Plot



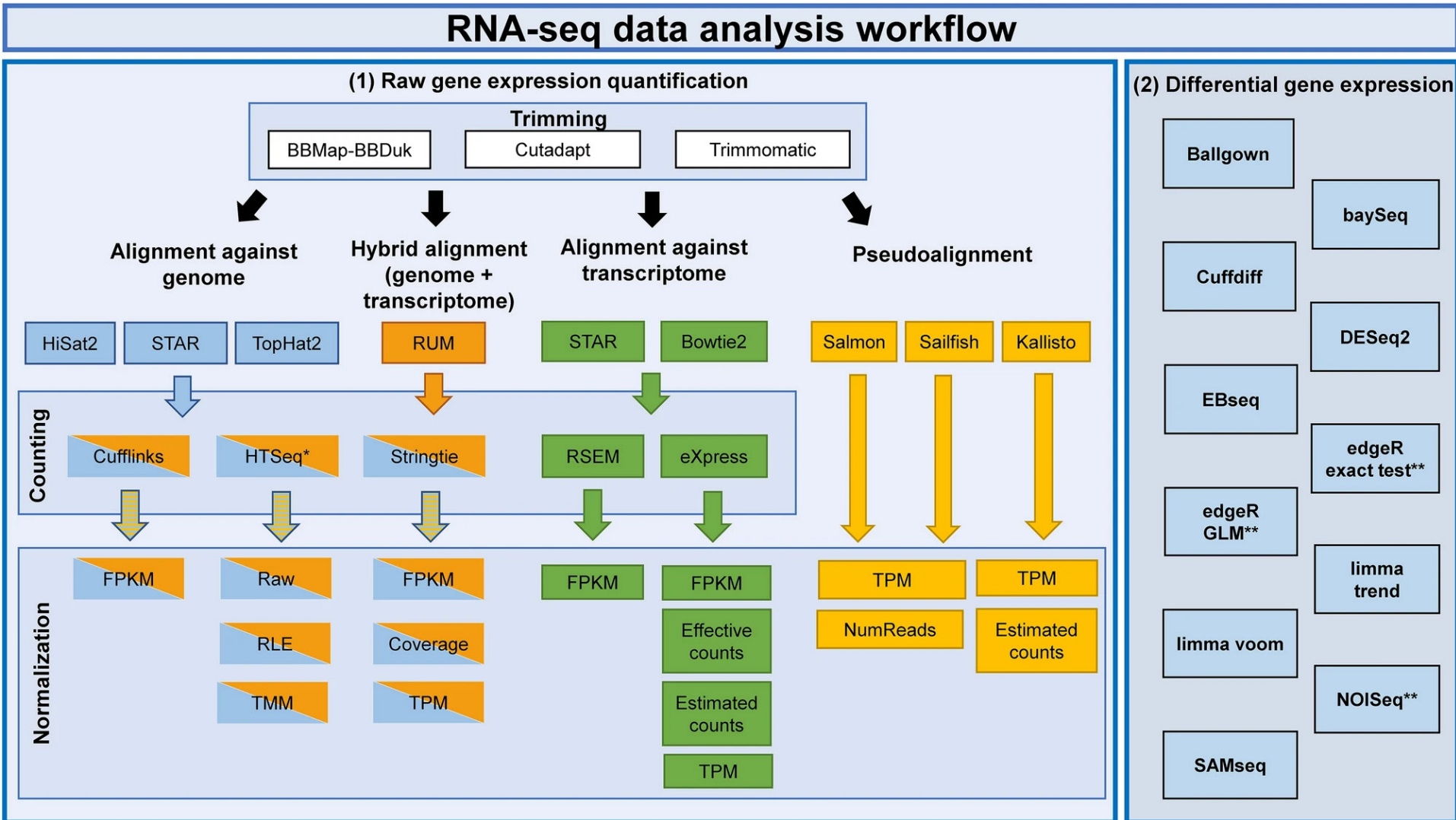
<https://f1000research.com/articles/5-1438> (doi: 10.12688/f1000research.8987.2)



Slide courtesy of Jenny Drnevich

I ILLINOIS
Roy J. Carver Biotechnology Center

So many options!



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., & Hinch, 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

Online resources for RNA-Seq analysis questions

- 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/>

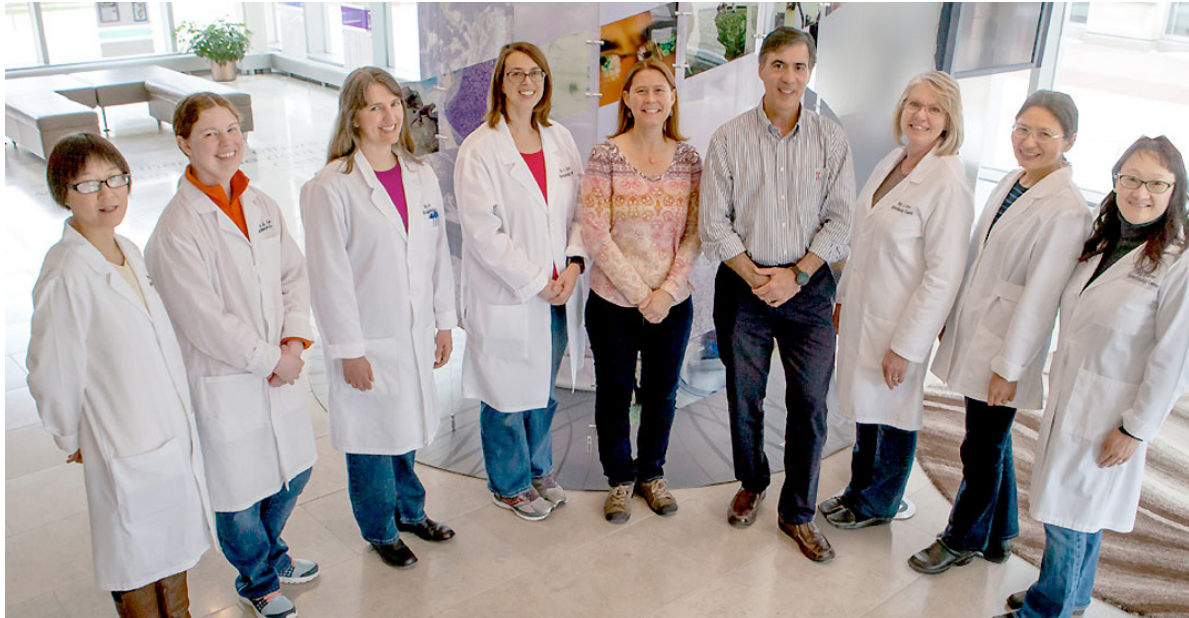
My email - jholmes5@illinois.edu



I ILLINOIS

Roy J. Carver Biotechnology Center

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!





HPCBio workshop presentations are licensed under a [Creative Commons Attribution-ShareAlike 4.0 International License](https://creativecommons.org/licenses/by-sa/4.0/).

This is a human-readable summary of (and not a substitute for) the [license](#). [Disclaimer](#).

You are free to:

- **Share** — copy and redistribute the material in any medium or format
- **Adapt** — remix, transform, and build upon the material
 - The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:

- **Attribution** — You must give [appropriate credit](#), provide a link to the license, and [indicate if changes were made](#). You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.
- **ShareAlike** — If you remix, transform, or build upon the material, you must distribute your contributions under the [same license](#) as the original.
- **No additional restrictions** — You may not apply legal terms or [technological measures](#) that legally restrict others from doing anything the license permits.

Notices:

- You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable [exception or limitation](#).
- No warranties are given. The license may not give you all of the permissions necessary for your intended use. For example, other rights such as [publicity, privacy, or moral rights](#) may limit how you use the material.

