Cold Spring Harbor Laboratory

Advanced Sequencing Technologies & Applications

http://meetings.cshl.edu/courses.html

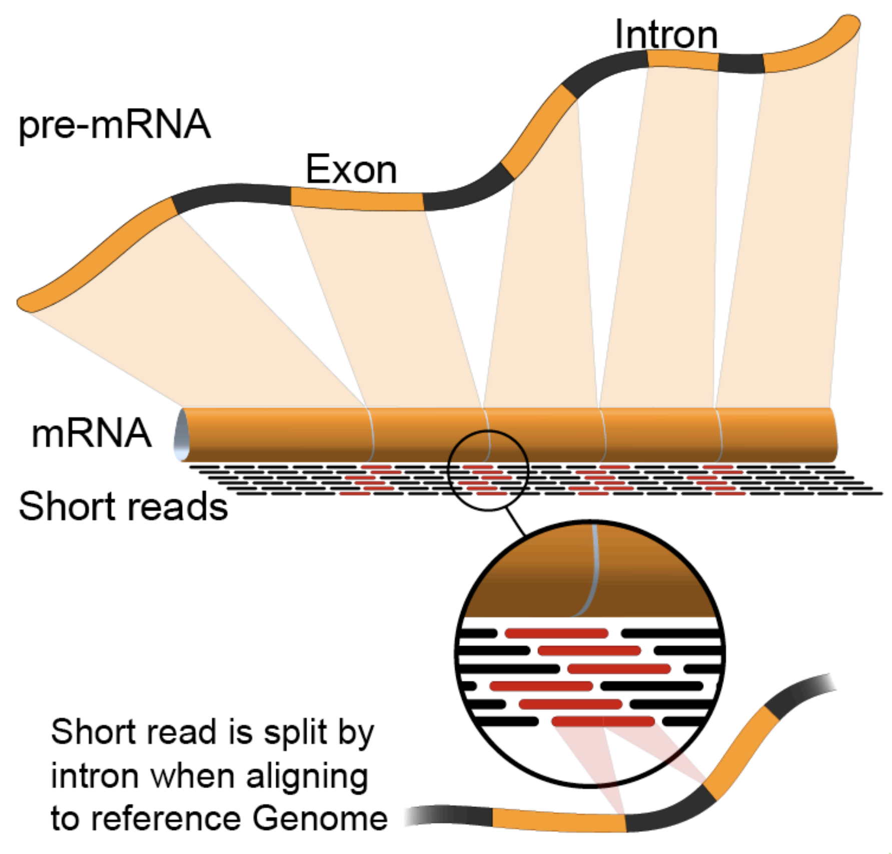
Cold Spring Harbor Laboratory

Module 3
Expression and Differential Expression (tutorial)

Malachi Griffith, Obi Griffith, Jason Walker, Ben Ainscough

Advanced Sequencing Technologies & Applications

November 11-23, 2014



The Genome Institute at Washington University

Learning Objectives of Tutorial

- Generate gene/transcript expression estimates with cufflinks
- Perform differential expression analysis with cuffmerge and cuffdiff
- Summarize and visualize results
  - cummeRbund
  - Old school R methods

Tutorial files

- Three parts
  - **Tutorial_Module3_Part1_Linux.txt**
    - Use Cufflinks to generate expression estimates
    - Use Cuffmerge and Cuffdiff to compare tumor and normal conditions
  - **Tutorial_Module3_Part2_CummeRbund.txt**
    - Summarize and visualize expression and differential expression data
  - **Tutorial_Module3_Part3_Supplementary_R.txt**
    - Optional alternative to cummeRbund for interpreting expression and differential expression data
  - **Tutorial_Module3_Part4_edgeR.txt**
    - Optional alternative to process raw counts with edgeR instead of using cufflinks/cuffdiff

9. Generate expression estimates

- The alignment SAM/BAM files generated in the previous step will now be used by cufflinks to calculate expression estimates
  - For all transcripts on the target chromosome
- For this step an option, confusingly also called '-G' is used
  - Forces cufflinks to calculate expression values for known transcripts
  - To discover novel transcripts with Cufflinks you should:
    - **Not use the '-G' option. De novo transcript assembly and estimation will be performed. (we will try this in Module 4) OR ...**
    - Use the '-G' option along with the '-g' option. Known transcripts will be used as a 'guide', but novel transcripts will also be predicted
- This step will generate one isoform and one gene expression file for each library
  - Expression values are reported as 'FPKM', or '**F**ragments **P**er **K**ilobase of exon per million fragments **M**apped'
  - Where each 'fragment' corresponds to a read-pair mapped to the genome

9. Generate expression estimates (Optional Alternatives)

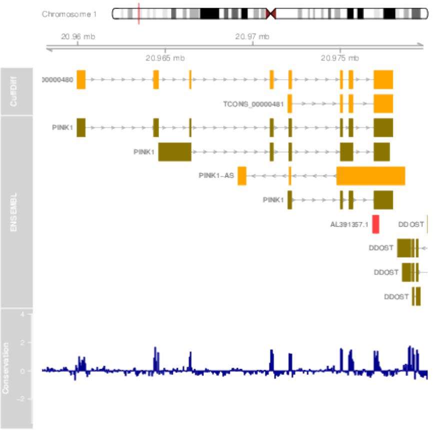
- The Alignment SAM/BAM files generated from STAR can also be used in cufflinks to generate expression estimates – exactly as above
- Another alternative we will explore is a count-based method
  - We will use a program called htseq-count
    - Requires name-sorted SAM file
    - We will count at the gene level (transcript-level is also possible)
- In the end we will have three expression estimates for each sample
  - Tophat/cufflinks
  - STAR/cufflinks
  - Tophat/Htseq-count

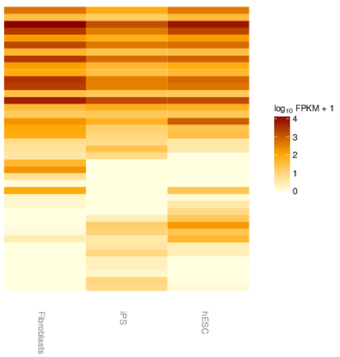
10. Perform differential expression analysis

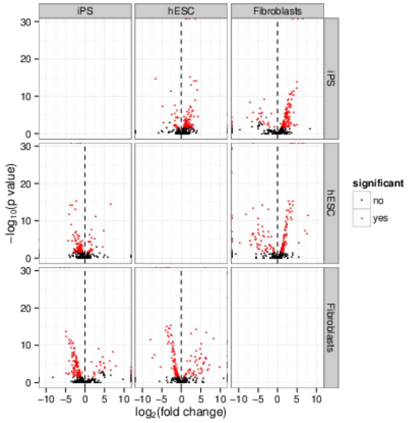
- In this step we will use cuffmerge and cuffdiff to:
  - Combine expression estimates from our 4 libraries into more convenient files
  - Combine expression estimates across replicates
  - Compare tumor vs. normal and identify significantly differentially expressed genes and isoforms (transcripts)
- Note that these commands can get quite complicated when you have replicates
  - The positioning of spaces and commas, and grouping of libraries matters!
- Comparisons
  - Compare Tumor vs. Normal using all replicates, for known (reference only mode) transcripts

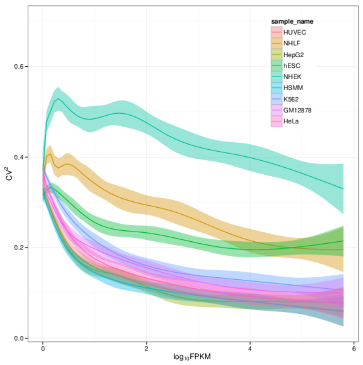
11. Summarize and visualize results

- In this step we will run the R package cummeRbund to visualize our expression and differential expression results from Cuffdiff.
  - See **Tutorial_Module3_Part2_CummeRbund.R** for details
  - http://compbio.mit.edu/cummeRbund/
  - http://compbio.mit.edu/cummeRbund/manual_2_0.html









12a. Post-process output files (optional)

- Cufflinks and Cuffdiff output various file formats
  - .fpkm_tracking, transcrips.gtf, and .diff files
- In this step, we will explore the content of these files at the linux command line before importing them into R for more advanced summarization, plotting, etc.
  - If you are unfamiliar with R, this is an interactive statistical programming interface that can also be used for graphing and file data manipulation (i.e. an alternative to ‘excel’)
  - http://cran.r-project.org/

12b. Summarize and visualize results (optional)

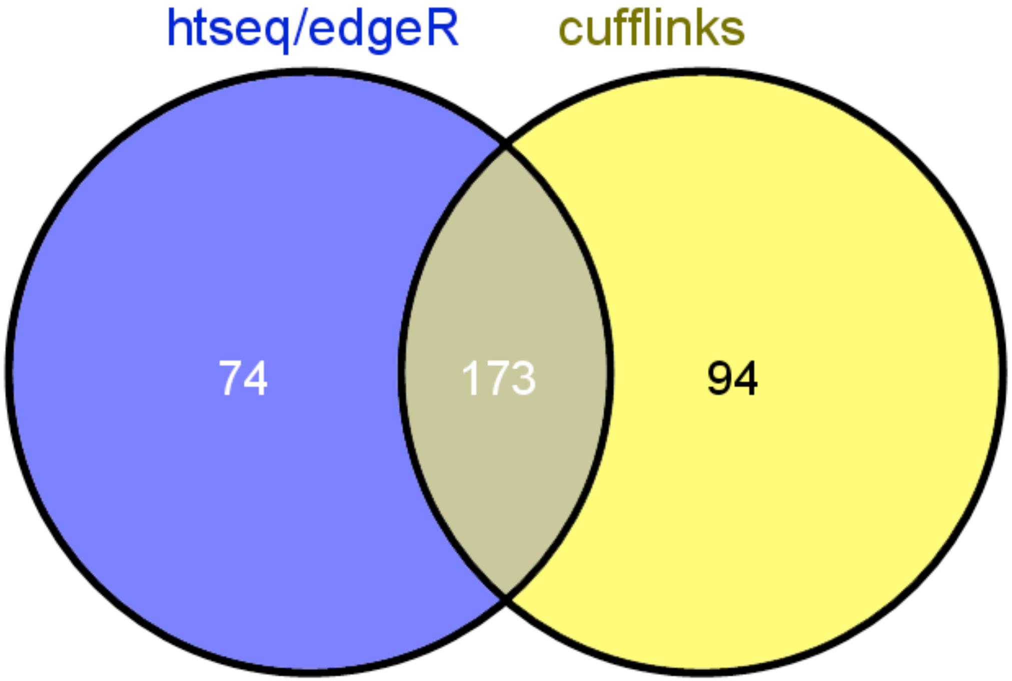
- In this step we will use R to summarize and visualize the results of the previous steps
- Explanation of the R commands is provided in a separate tutorial file '**Tutorial_Module3_Part3_Supplementary_R.R**'
- Examples of the tasks performed:

– Examine the expression estimates
  - How reproducible are the technical replicates?
  - How well do the different library construction methods correlate?
  - Visualize the differences between/among replicates, library prep methods and tumor versus normal

– Examine the differential expression estimates
  - Visualize the expression estimates and highlight those genes that appear to be differentially expressed according to cuffdiff
  - Generate a list of the top differentially expressed genes

12c/d. Perform differential expression analysis with edgeR using htseq output (optional)

- Make use of raw counts generated by htseq-count
- Load into R and process with edgeR package
- Compare significantly differentially expressed genes from two methods



Break