Canadian Bioinformatics Workshops

www.bioinformatics.ca

Afrikaans български Català Dansk Deutsch Ελληνικά English English (CA) English (GB) English (US) Esperanto
Castellano Castellano (AR) Español (CL) Castellano (CO) Español (Ecuador) Castellano (MX) Castellano (PE)
Euskara Suomeksi français français (CA) Galego עברית hrvatski Magyar Italiano 日本語 한국어 Macedonian Melayu
Nederlands Norsk Sesotho sa Leboa polski Português română slovenski jezik српски srpski (latinica) Sotho svenska
中文 華語 (台灣) isiZulu

creative commons

Attribution-Share Alike 2.5 Canada

You are free:

to Share — to copy, distribute and transmit the work

to Remix — to adapt the work

Free Cultural Works — **APPROVED FOR**

Under the following conditions:

Attribution. You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).

Share Alike. If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar licence to this one.

- For any reuse or distribution, you must make clear to others the licence terms of this work.
- Any of the above conditions can be waived if you get permission from the copyright holder.
- The author's moral rights are retained in this licence.

Disclaimer

Your fair dealing and other rights are in no way affected by the above.
This is a human-readable summary of the Legal Code (the full licence) available in the following languages:
English French

Learn how to distribute your work using this licence

Module 3
Expression and Differential Expression (tutorial)

Malachi Griffith & Obi Griffith
Informatics for RNA-seq Analysis
June 8-9, 2015
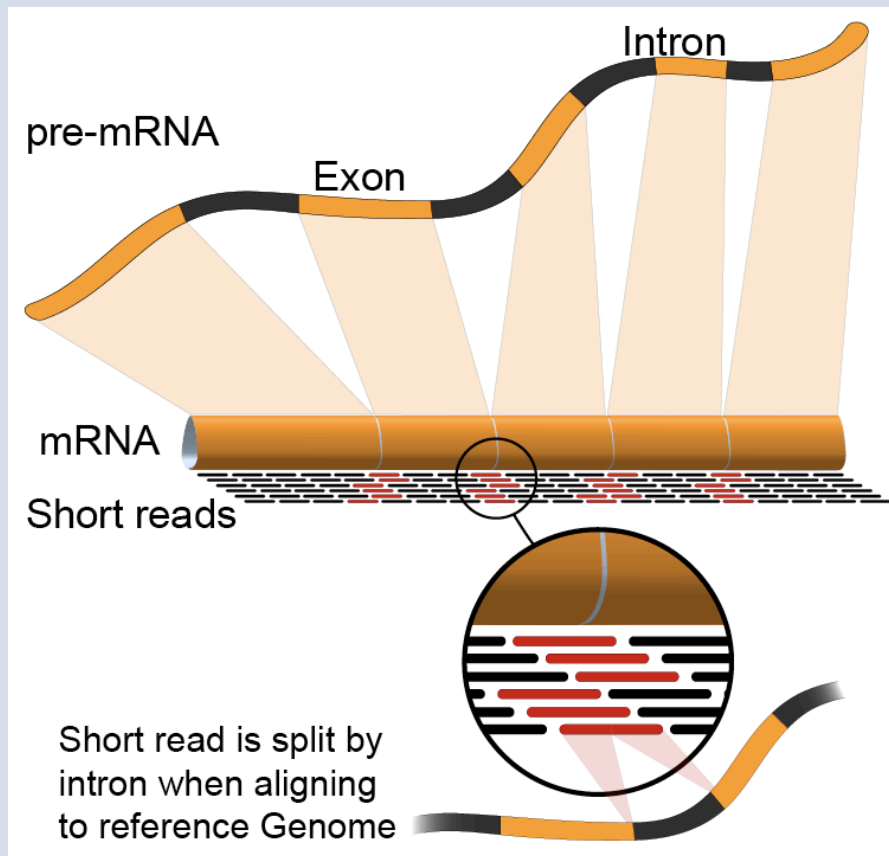
Learning Objectives of Tutorial

- Generate gene/transcript expression estimates with cufflinks

- Perform differential expression analysis with cuffmerge and cuffdiff

- Summarize and visualize results
  - cummeRbund
  - Old school R methods

4-i. 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

4-i. 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

4-ii. Perform differential expression analysis

- In this step we will use cuffmerge and cuffdiff to:
  - Combine expression estimates from our 6 libraries into more convenient files
  - Combine expression estimates across replicates
  - Compare UHR vs. HBR 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 UHR vs. HBR using all replicates, for known (reference only mode) transcripts

4-iii. 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 online tutorial for details

  - http://compbio.mit.edu/cummeRbund/

  - http://compbio.mit.edu/cummeRbund/manual_2_0.html
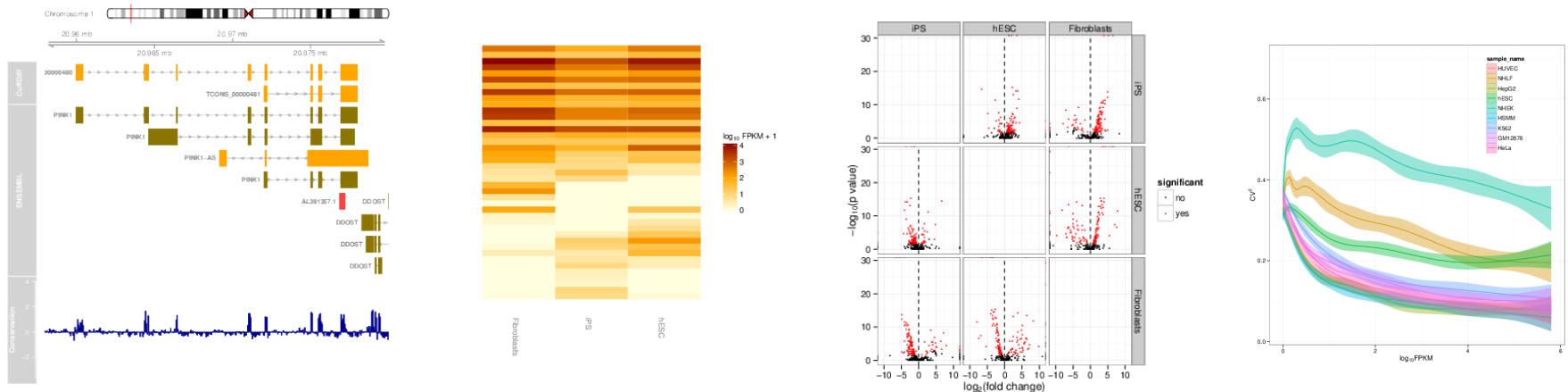
Post-process output files (optional)

- 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/

bioinformatics.ca

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 the online wiki

- 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
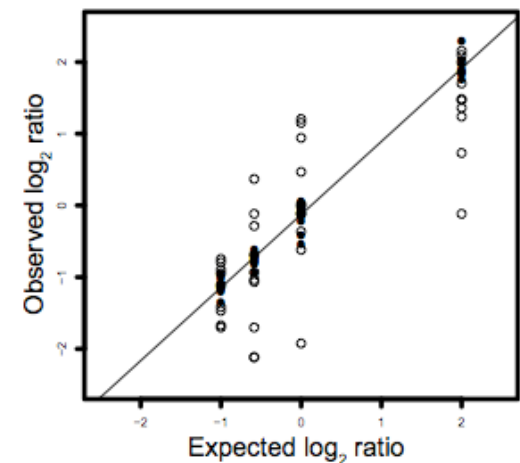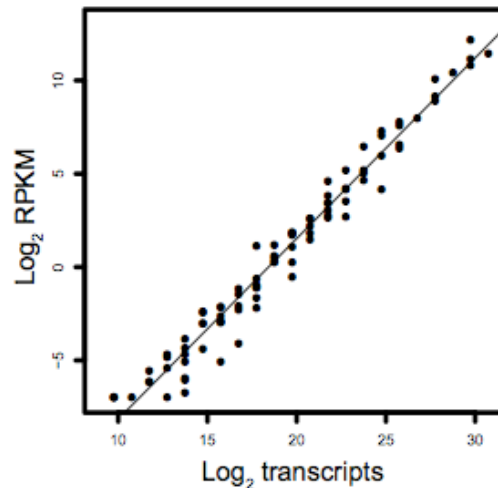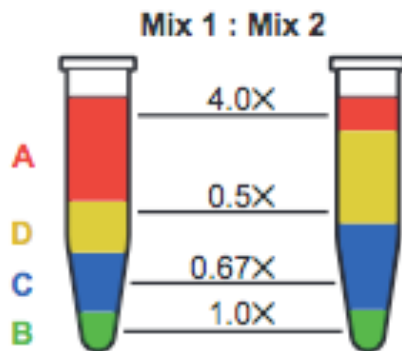
Perform differential expression analysis with edgeR using htseq output (optional)

- 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
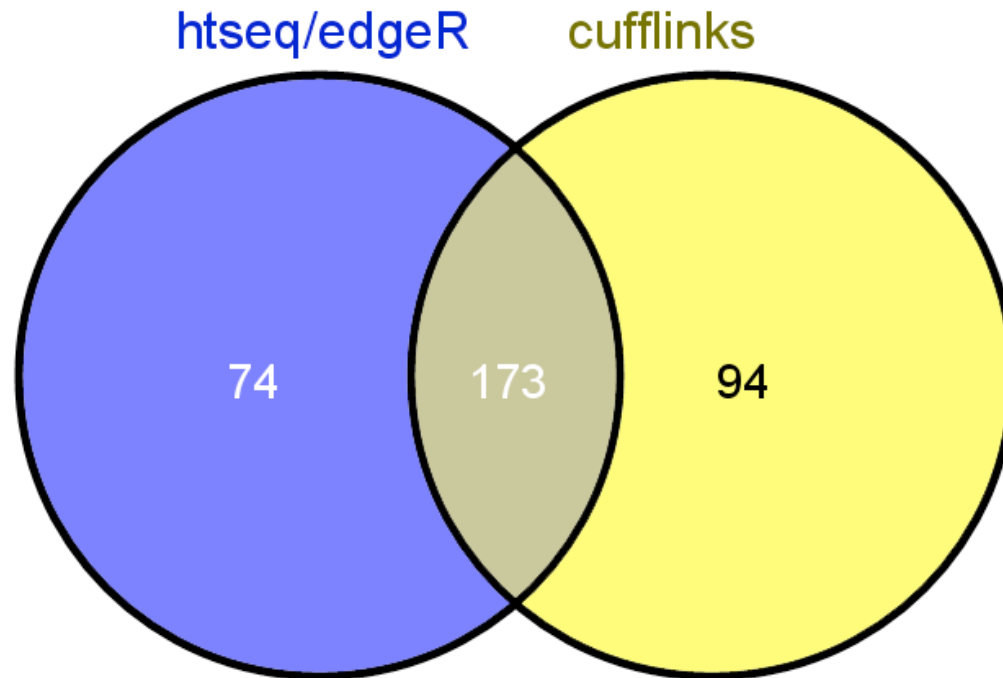
Analysis of ERCC spike-in expression and differential expression (optional)

- https://tools.lifetechnologies.com/content/sfs/manuals/cms_086340.pdf

- Lower Limit of Detection

- Dynamic Range (dose response)

- Fold-change response (DE)