GCC Workshop 9
RNA-Seq with Galaxy

Curtis Hendrickson (curtish@uab.edu)
David Crossman (dkcrossm@uab.edu)
Jeremy Goecks (jeremy.goecks@emory.edu)
 Agenda

• RNA-seq flash review
  – Tophat (RNA-seq read mapping)
  – Cufflinks (Isoform assembly)
  – Cuffmerge (cross-sample consensus)
  – Cuffdiff (expression analysis)

• Galaxy flash review

• RNA-seq exercises
  – installed genome: mouse (mm9)
  – Non-standard genome: tree shrew from Ensembl
Why RNA-Seq?

- *Gene/transcript expression quantification*
- Isoform analysis (alternate splicing)
- 5’ and 3’ UTR analysis
- Sequencing of new genomes (paired genome/transcriptome sequencing)
- etc
What type of analysis?

• **Compare** gene or transcript levels between samples or conditions?
  – Sequence the whole transcriptome

• **Discovery** of alternate splice forms?
  – >= 75bp Paired End for splice site detection
  – Consider target enrichment
  – Even at 100x coverage, a rare (1:100) splice junction would only average 0.4x coverage of it’s splice junction – not enough to detect.
Reads in RNA-seq
An RNA-Seq Analysis Pipeline

TopHat

Align reads & discover transcriptome

Cufflinks

Assemble reads into isoforms

Cuffmerge

Build consensus transcriptome across samples and replicates

Cuffdiff

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

Cole Trapnell1,2, Adam Roberts3, Loyal Goff1,2,4, Geo Pertea5,6, Daehwan Kim5,7, David R Kelley1,2, Harold Pimentel3, Steven L Salzberg5,6, John L Rinn1,2 & Lior Pachter3,8,9

1Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. 2Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, Massachusetts, USA. 3Department of Computer Science, University of California, Berkeley, California, USA. 4Computer Science and Artificial Intelligence Lab, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. 5Department of Medicine, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA. 6Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland, USA. 7Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland, USA. 8Department of Mathematics, University of California, Berkeley, California, USA. 9Department 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; 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 identify new genes and new splice variants of known ones, as well as compare gene and transcript expression under two or more conditions. This protocol describes in detail how to use TopHat and Cufflinks to perform such analyses. It also covers several accessory tools and utilities that aid in managing data, including CummRbund, a tool for visualizing RNA-seq analysis results. Although the procedure assumes basic informatics skills, these tools assume little to no background with RNA-seq analysis and are meant for novices and experts alike. The protocol begins with raw sequencing reads and produces a transcriptome assembly, lists of differentially expressed and regulated genes and transcripts, and publication-quality visualizations of analysis results. The protocol's execution time depends on the volume of transcriptome sequencing data and available computing resources but takes less than 1 d of computer time for typical experiments and ~1 h of hands-on time.
Types of data

- FASTA <1G Genomes
- FASTQ 1-12G/ea Sequencing reads
- BAM 1-5G/ea Mapped Reads (binary)
- BED <1G Genomic intervals
- GTF < 1G Gene/transcript location
2-Condition Differential Expression Protocol Overview

(Trapnell et al, Nat Protocols, 2012)
2-Condition Differential Expression Protocol Overview

QC: FastQC

QC: Flatstat

Reads → Reads

TopHat

Mapped reads → Mapped reads

Cufflinks

Assembled transcripts → Assembled transcripts

Cuffmerge

Final transcriptome assembly

Cuffdiff

Differential expression results

CummeRbund

Expression plots

(Trapnell et al, Nat Protocols, 2012)
Random Galaxy icons/colors

**Colors**

- Queued
- Running
- Completed
- Failed

**Download/Save**

- 14: Control Tophat for Illumina on data 3, data 4, and data 2: accepted hits
- 14: Control Tophat for Illumina on data 3, data 4, and data 2: accepted hits

**Icons**

- Display data in browser
- Edit attributes
- Delete
- Edit dataset annotation

- View details
- Run this job again
- View in Trackster
- Edit dataset tags
In the beginning there were reads...

Get them from Shared Data

1. Click on “Shared Data” (located on top toolbar)
2. Drop down box appears; click on “Data Libraries”
3. Will see this Data Library. Click on it to expand (as shown)
So we used FASTQC to QC them...

• Tools > 🛠 > Tool Search
• Enter “fastqc”
• Select the tool
• Enter a data set & Execute
• Trick for quick entry of multiple samples:
  • Immediately hit “BACK”
  • Update parameters and re-execute
This data looks awful because this is filtered data from a much larger fastq file. Better results when using entire file!
Tophat: map reads, create transcriptome

Paired-end, so 2 FASTQ per Tophat run
- Select forward FASTQ
- Set the ref. genome
- Set PAIRed ends
- Select reverse FASTQ
- Select inner distance (get from sequencing group)
Launched 2 tophats (4 outputs each)

TIP:
> Show Structure
groups datasets produced by the same tool.

Edit “accepted_hits”
Datasets to give shorter names with sample info:
[ctrl] tophat.accepted_hits
Launch 2 flagstats: QC mapping

TIP:
You can execute a tool before it’s predecessors have finished running.

Galaxy will queue it until all the dependencies are satisfied!
Viewing Alignments

• UCSC

• IGV via “web current” (JNLP)
  – Set Visibility window & restart
  – Un-collapse gene track!

• IGV via “local”
  – Make sure it’s running on your machine first!

• Trackster
  – Build visualizations inside galaxy!
Display mapped BAM & BAI
Visually validate in IGV, or like
Cufflink: Construct Transcripts

Read data:
accepted_hits (BAM)

Use Reference Anno = yes
then pull down appears you can choose a GTF file from your history.

Reference annotation:
iGenome genes.GTF

Other Parameters
“Perform quartile normalization”&
“Perform Bias Correction”
Normally YES is best

because we're not working with a full read-set, use NO or statistics will go haywire
Cufflinks: discovery

No – for novel transcript discovery

• requires very deep sequencing – consider sample enrichment!
• Slow, memory intensive
• Discover ONLY transcripts proven by reads

Ref
• use reads *only* to quantify reference annotation
• All novel splices ignored

Ref as Guide – *most common*
• use reference annotation
• Extend reference, when there is sufficient evidence
Cufflinks outputs

TIP: hide the cmd bar!

assembled_transcripts (GTF)
  list of isoforms
transcript_expression (tab)
  isoforms w/ FPKM
gene_expression (tab)
  genes w/ FPKM

QC steps
1. Check FPKM not all 0.0!
2. Check you have gene symbols
3. Visualize assembled_transcripts in a genome browser (UCSC, IGV, IGB, etc)
assembled_transcripts Visualization
Cuffmerge: combine transcripts samples & replicates -> complete transcriptome

Run once, using both samples

Inputs
[ctrl] assembled _trans...
[drug] assembled_trans...
Reference_annotation (GTF)
**Cuffdiff**: Quantize Transcripts

Compute fold change between conditions

---

Run once, use all samples & replicates

ALWAYS create replicate groups for each condition

ALWAYS name your groups

even if you have only one replicate!
Cuffdiff: Fold Change Between Conditions

Run once, using both samples

ALWAYS create replicate groups for each condition, even if you have only one replicate!

Short names – no spaces or special characters!!!
# Cuffdiff output

(=> view structure)

## Tool: Cuffdiff

| 43: Cuffdiff on data 14, data 10, and data 32: transcript FPKM tracking |
| 42: Cuffdiff on data 14, data 10, and data 32: transcript differential expression testing |
| 41: Cuffdiff on data 14, data 10, and data 32: gene FPKM tracking |
| 40: Cuffdiff on data 14, data 10, and data 32: gene differential expression testing |
| 39: Cuffdiff on data 14, data 10, and data 32: TSS groups FPKM tracking |
| 38: Cuffdiff on data 14, data 10, and data 32: TSS groups differential expression testing |
| 37: Cuffdiff on data 14, data 10, and data 32: CDS FPKM tracking |
| 36: Cuffdiff on data 14, data 10, and data 32: CDS FPKM differential expression testing |
| 35: Cuffdiff on data 14, data 10, and data 32: CDS overloading differential expression testing |
| 34: Cuffdiff on data 14, data 10, and data 32: promoters differential expression testing |
| 33: Cuffdiff on data 14, data 10, and data 32: splicing differential expression testing |

## Tool: Cuffmerge

| 32: Cuffmerge on data 19, data 1, and data 23: merged transcripts |

## Tool: Cufflinks
Cuffdiff “gene diff”: 23,113 lines! (-> value for all genes in gene.gtf)

<table>
<thead>
<tr>
<th>test_id</th>
<th>gene_id</th>
<th>gene</th>
<th>locus</th>
<th>sample_1</th>
<th>sample_2</th>
<th>status</th>
<th>value_1</th>
<th>value_2</th>
<th>log2(fold_change)</th>
<th>test_stat</th>
<th>p_value</th>
<th>q_value</th>
<th>significant</th>
</tr>
</thead>
<tbody>
<tr>
<td>XLOC_000001</td>
<td>XLOC_000001</td>
<td>Lynpl1</td>
<td>chr1:4797973-4830810</td>
<td>ctrl</td>
<td>drug</td>
<td>NOTEST</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>XLOC_000002</td>
<td>XLOC_000002</td>
<td>Tros1</td>
<td>chr1:4847774-4887990</td>
<td>ctrl</td>
<td>drug</td>
<td>NOTEST</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>XLOC_000003</td>
<td>XLOC_000003</td>
<td>Atpv1h</td>
<td>chr1:5073259-5112630</td>
<td>ctrl</td>
<td>drug</td>
<td>NOTEST</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>XLOC_000004</td>
<td>XLOC_000004</td>
<td>Cyplk1</td>
<td>chr1:5578373-5606614</td>
<td>ctrl</td>
<td>drug</td>
<td>NOTEST</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>XLOC_000005</td>
<td>XLOC_000005</td>
<td>Rbcl1c</td>
<td>chr1:620472-6260183</td>
<td>ctrl</td>
<td>drug</td>
<td>NOTEST</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>XLOC_000006</td>
<td>XLOC_000006</td>
<td>Fam150a</td>
<td>chr1:6349411-6388412</td>
<td>ctrl</td>
<td>drug</td>
<td>NOTEST</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>XLOC_000007</td>
<td>XLOC_000007</td>
<td>St8</td>
<td>chr1:6477311-6501101</td>
<td>ctrl</td>
<td>drug</td>
<td>NOTEST</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>XLOC_000008</td>
<td>XLOC_000008</td>
<td>Porexl</td>
<td>chr1:706900-7163700</td>
<td>ctrl</td>
<td>drug</td>
<td>NOTEST</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>XLOC_000009</td>
<td>XLOC_000009</td>
<td>Psp1</td>
<td>chr1:9534888-9587356</td>
<td>ctrl</td>
<td>drug</td>
<td>NOTEST</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>XLOC_000010</td>
<td>XLOC_000010</td>
<td>Adhe1</td>
<td>chr1:9382126-9421173</td>
<td>ctrl</td>
<td>drug</td>
<td>NOTEST</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>XLOC_000011</td>
<td>XLOC_000011</td>
<td>S110003E1_A1</td>
<td>chr1:9538126-9621173</td>
<td>ctrl</td>
<td>drug</td>
<td>NOTEST</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>XLOC_000012</td>
<td>XLOC_000012</td>
<td>Atp100813</td>
<td>chr1:9713277-9763337</td>
<td>ctrl</td>
<td>drug</td>
<td>NOTEST</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>XLOC_000013</td>
<td>XLOC_000013</td>
<td>G6s3</td>
<td>chr1:9788216-9802640</td>
<td>ctrl</td>
<td>drug</td>
<td>NOTEST</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>XLOC_000014</td>
<td>XLOC_000014</td>
<td>003042MIK1</td>
<td>chr1:9898718-993033</td>
<td>ctrl</td>
<td>drug</td>
<td>NOTEST</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>XLOC_000015</td>
<td>XLOC_000015</td>
<td>Ccpp1</td>
<td>chr1:10026298-10120649</td>
<td>ctrl</td>
<td>drug</td>
<td>NOTEST</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>XLOC_000016</td>
<td>XLOC_000016</td>
<td>Pme2</td>
<td>chr1:10983545-1129763</td>
<td>ctrl</td>
<td>drug</td>
<td>NOTEST</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>XLOC_000017</td>
<td>XLOC_000017</td>
<td>493241L1</td>
<td>chr1:11464185-1196556</td>
<td>ctrl</td>
<td>drug</td>
<td>NOTEST</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>XLOC_000018</td>
<td>XLOC_000018</td>
<td>Cps1</td>
<td>chr1:11978216-12063171</td>
<td>ctrl</td>
<td>drug</td>
<td>NOTEST</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>XLOC_000019</td>
<td>XLOC_000019</td>
<td>Sufi</td>
<td>chr1:12082510-1280453</td>
<td>ctrl</td>
<td>drug</td>
<td>NOTEST</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>XLOC_000020</td>
<td>XLOC_000020</td>
<td>Tks6</td>
<td>chr1:13958682-139614</td>
<td>ctrl</td>
<td>drug</td>
<td>NOTEST</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>XLOC_000021</td>
<td>XLOC_000021</td>
<td>Xvpl2</td>
<td>chr1:15150532-15704295</td>
<td>ctrl</td>
<td>drug</td>
<td>NOTEST</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>XLOC_000022</td>
<td>XLOC_000022</td>
<td>Tef1</td>
<td>chr1:15973738-1833510</td>
<td>ctrl</td>
<td>drug</td>
<td>NOTEST</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>XLOC_000023</td>
<td>XLOC_000023</td>
<td>Rha10</td>
<td>chr1:16092259-16212681</td>
<td>ctrl</td>
<td>drug</td>
<td>NOTEST</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>XLOC_000024</td>
<td>XLOC_000024</td>
<td>D00040825Rk</td>
<td>chr1:16647562-16653259</td>
<td>ctrl</td>
<td>drug</td>
<td>NOTEST</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>XLOC_000025</td>
<td>XLOC_000025</td>
<td>Lyc1</td>
<td>chr1:16678356-1666086</td>
<td>ctrl</td>
<td>drug</td>
<td>NOTEST</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>XLOC_000026</td>
<td>XLOC_000026</td>
<td>Dap3</td>
<td>chr1:17135433-17134351</td>
<td>ctrl</td>
<td>drug</td>
<td>NOTEST</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>XLOC_000027</td>
<td>XLOC_000027</td>
<td>Pl1</td>
<td>chr1:17519938-17610120</td>
<td>ctrl</td>
<td>drug</td>
<td>NOTEST</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>XLOC_000028</td>
<td>XLOC_000028</td>
<td>Ccpp1</td>
<td>chr1:17714977-17715728</td>
<td>ctrl</td>
<td>drug</td>
<td>NOTEST</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>XLOC_000029</td>
<td>XLOC_000029</td>
<td>Ccpp1</td>
<td>chr1:19909192-19916413</td>
<td>ctrl</td>
<td>drug</td>
<td>NOTEST</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>XLOC_000030</td>
<td>XLOC_000030</td>
<td>Tgap2d</td>
<td>chr1:20047859-20068138</td>
<td>ctrl</td>
<td>drug</td>
<td>NOTEST</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>XLOC_000031</td>
<td>XLOC_000031</td>
<td>Tgap2b</td>
<td>chr1:20072840-20072606</td>
<td>ctrl</td>
<td>drug</td>
<td>NOTEST</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>XLOC_000032</td>
<td>XLOC_000032</td>
<td>493048503Rk</td>
<td>chr1:20170090-20216912</td>
<td>ctrl</td>
<td>drug</td>
<td>NOTEST</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>XLOC_000033</td>
<td>XLOC_000033</td>
<td>Dap6</td>
<td>chr1:20941766-20960522</td>
<td>ctrl</td>
<td>drug</td>
<td>NOTEST</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>yes</td>
</tr>
</tbody>
</table>

Tools > Filter and Sort > Filter: c7 <> 'NOTEST' and c14 <> 'no'
Gene diff: genes with enough reads

14 genes
Gene diff: statistically significant 9 genes
“Final” gene list

Download Excel Conditional Formatting
Visualization of Gene Expression

cummeRbund in Galaxy

Galaxy wrapper created, but not yet in tool shed

http://cvrgrid.org/node/235
http://ec2-23-20-5-163.compute-1.amazonaws.com:8080/u/boliu/h/cummerbund

(courtesy of Liu Bo @ University of Chicago)

(Trapnell et al, 2012, Nat Protocols)
References and web links

• TopHat

• Bowtie

• Cufflinks

• TopHat and Cufflinks protocol

• Illumina iGenomes: indexes and annotations for use with Cufflinks, etc.
Thanks! Questions? Contact info:

R. Curtis Hendrickson
Research Associate
Microbiology
Center for Clinical and Translational Science
University of Alabama at Birmingham

http://www.uab.edu/ccts/ResearchResources/BMI
curtish arabesque uab.edu