

# NGS Data Analysis and Galaxy

---

University of Pretoria  
Pretoria, South Africa  
14-18 October 2013

Dave Clements, Emory University

<http://galaxyproject.org/>

Fourie Joubert, Burger van Jaarsveld  
Bioinformatics & Computational Biology Unit  
University of Pretoria

<http://science.up.ac.za/>



UNIVERSITEIT VAN PRETORIA  
UNIVERSITY OF PRETORIA  
YUNIBESITHI YA PRETORIA



# This Week

|           |  |
|-----------|--|
| Monday    | Welcome, Project Intro, Basic Galaxy Usage<br>NGS QualityControl   |
| Tuesday   | RNA-Seq - Mapping and Transcript Prediction<br>RNA-Seq: Differential expression and<br>Alternative Pipelines; SNP & Variant Analysis |
| Wednesday | SNP & Variant Analysis<br>Chip-Seq Analysis  |
| Thursday  | Genome Assembly<br>Install your own Galaxy on Amazon Cloud   |
| Friday    | Customizing Galaxy, Galaxy Tool Shed, and<br>Wrapping Tools for Galaxy   |

# Tuesday Agenda

- 8:30 **Welcome and Questions**
- 8:45 RNA-Seq: Mapping with Tophat
- 9:45 Cufflinks: Transcript prediction
- 10:30 Break
- 11:00 Differential Expression Analysis with Cuffdiff
- 12:00 Cuffdiff alternative: DESeq, Part I
- 12:30 Lunch
- 13:30 Cuffdiff alternative: DESeq, Part II
- 14:00 SNP and Variant Analysis, Part I
- 15:00 Break
- 15:30 SNP and Variant Analysis, Part II
- 16:30 Done

# Tuesday Agenda

- 8:30 **Welcome and Questions**
- 8:45 **RNA-Seq: Mapping with Tophat**
- 9:45 Cufflinks: Transcript prediction
- 10:30 Break
- 11:00 Differential Expression Analysis with Cuffdiff
- 12:00 Cuffdiff alternative: DESeq, Part I
- 12:30 Lunch
- 13:30 Cuffdiff alternative: DESeq, Part II
- 14:00 SNP and Variant Analysis, Part I
- 15:00 Break
- 15:30 SNP and Variant Analysis, Part II
- 16:30 Done

# RNA-seq Exercise: Mapping with Tophat

Create a new history

Import all datasets from library:

[UC Davis RNA-Seq TopHat Inputs](#)

**NGS: RNA Analysis → TopHat for Illumina**

<http://cloud1.galaxyproject.org/>

<http://cloud2.galaxyproject.org/>

<http://cloud3.galaxyproject.org/>

# RNA-seq Exercise: Mapping with Tophat

- Tophat looks for best place(s) to map reads, and best places to insert introns
- *Imagine pages and pages of discussion on the intricacies and pitfalls of RNA-seq mapping here.*

# Mapping with Tophat: **mean inner distance**

Expected distance between paired ends

- Has to be provided to you by sequencing core!
- We'll use **90\*** for **mean inner distance**
- We'll use **50** for **standard deviation**

\* The library was constructed with the typical Illumina TruSeq protocol, which is supposed to have an average insert size of 200 bases. Our reads are 55 bases (R1) plus 55 bases (R2). So, the Inner Distance is estimated to be  $200 - 55 - 55 = 90$

# Mapping with Tophat: Use Existing Annotations?

You can bias Tophat towards known annotations

- Use Own Junctions → Yes
  - Use Gene Annotation → Yes
  - Gene Model Annotation → genes\_chr12.gtf
- Use Raw Junctions → Yes (tab delimited file)
- Only look for supplied junctions → Yes



# Mapping with Tophat: **Make it quicker?**

Warning: Here be dragons!

- **Allow indel search** → **No**
- **Use Coverage Search** → **No** (wee dragons)

TopHat generates its database of possible splice junctions from two sources of evidence. The first and strongest source of evidence for a splice junction is when two segments from the same read (for reads of at least 45bp) are mapped at a certain distance on the same genomic sequence or when an internal segment fails to map - again suggesting that such reads are spanning multiple exons. With this approach, "GT-AG", "GC-AG" and "AT-AC" introns will be found ab initio. The second source is pairings of "coverage islands", which are distinct regions of piled up reads in the initial mapping. Neighboring islands are often spliced together in the transcriptome, so TopHat looks for ways to join these with an intron. **We only suggest users use this second option (--coverage-search) for short reads (< 45bp) and with a small number of reads (<= 10 million).** This latter option will only report alignments across "GT-AG" introns

Mapping with Tophat: **How did we do?**

NGS: SAM Tools → flagstat

Mapping with Tophat: **Lets do it some more!**

NGS: RNA Analysis → TopHat

for the remaining 3 replicates

# RNA-Seq Mapping With Tophat

## Further reading & Resources

[RNA-Seq Concepts, Terminology, and Work Flows](#)

by Monica Britton

[Aligning PE RNA-Seq Reads to a Genome](#)

by Monica Britton

both from the [UC Davis 2013 Bioinformatics Short Course](#)

[RNA-Seq Analysis with Galaxy](#)

by [Jeroen F.J. Laros](#), [Wibowo Arindrarto](#), [Leon Mei](#)

from the [GCC2013 Training Day](#)

[Tophat Manual](#)

# Tuesday Agenda

- 8:30 **Welcome and Questions**
- 8:45 **RNA-Seq: Mapping with Tophat**
- 9:45 **Cufflinks: Transcript prediction**
- 10:30 Break
- 11:00 Differential Expression Analysis with Cuffdiff
- 12:00 Cuffdiff alternative: DESeq, Part I
- 12:30 Lunch
- 13:30 Cuffdiff alternative: DESeq, Part II
- 14:00 SNP and Variant Analysis, Part I
- 15:00 Break
- 15:30 SNP and Variant Analysis, Part II
- 16:30 Done

# Transcript Prediction: Cufflinks

- Run Cufflinks on Tophat output to assemble reads into transcripts
  - Tophat does not make any predictions about how the reads it mapped assemble together into transcripts.
  - *Imagine pages and pages of discussion on the intricacies and pitfalls of RNA-seq transcript prediction here*

**NGS: RNA Analysis → Cufflinks**

# Cufflinks: **Min Isoform Fraction**

Cufflinks can predict many different transcripts for a gene.

One transcript is likely to dominate.

**Min Isoform Fraction** tells Cufflinks to ignore any isoforms that fall below this level of expression, *relative to the dominant isoform*.

Higher values: less noise; less likely to report/discover low-expression transcripts.

# Cufflinks: Pre mRNA Fraction

## From the Cufflinks Manual

“Some RNA-Seq protocols produce a significant amount of reads that originate from incompletely spliced transcripts, and these reads can confound the assembly of fully spliced mRNAs. Cufflinks uses this parameter to filter out alignments that lie within the intronic intervals implied by the spliced alignments. The minimum depth of coverage in the intronic region covered by the alignment is divided by the number of spliced reads, and if the result is lower than this parameter value, the intronic alignments are ignored. The default is 15%.”

Basically, sets your tolerance for noise/novel constructs in intronic regions.

# Cufflinks: Normalization and Correction

How hard should Cufflinks work to do the right thing?

**Quartile Optimization:** Attempt to compensate for skew caused by highly expressed genes

**Bias Correction:** Attempt to compensate for known issues with use of random hexamers in library preparation.\*

**Multi-Read Correct:** Try to make reads that mapped to multiple locations more useful\*\*

\* see Kasper D. Hansen, Steven E. Brenner, Sandrine Dudoit, Biases in Illumina transcriptome sequencing caused by random hexamer priming Nucleic Acids Research, Volume 38, Issue 12 (2010)

\*\* see <http://cufflinks.cbcb.umd.edu/howitworks.html#hmul>



# Cufflinks: Reference Annotation

How biased should we be, based on what we already know?

**Reference Annotation:** Use the reference annotation as dogma.  
Only doing quantification of known transcripts

**Reference Annotation as Guide:** Take advantage of what we already know, but be open to novel transcripts, if there is sufficient evidence

**No:** Transcript prediction will be based entirely on mapped reads in this dataset.

# Transcript Prediction: Cuffmerge

- Each Cufflinks run creates a set of transcript predictions.
- **Cuffmerge** unifies all those predictions into a single set.
- Makes this incredibly tedious task easy.

# Transcript Prediction: Cufflinks

- Run Cufflinks on Tophat output to assemble reads into transcripts
  - *Imagine pages and pages of discussion on the intricacies and pitfalls of RNA-seq transcript prediction here.*
  - **Visualize and refine our analysis**

# Visualizing Genomics

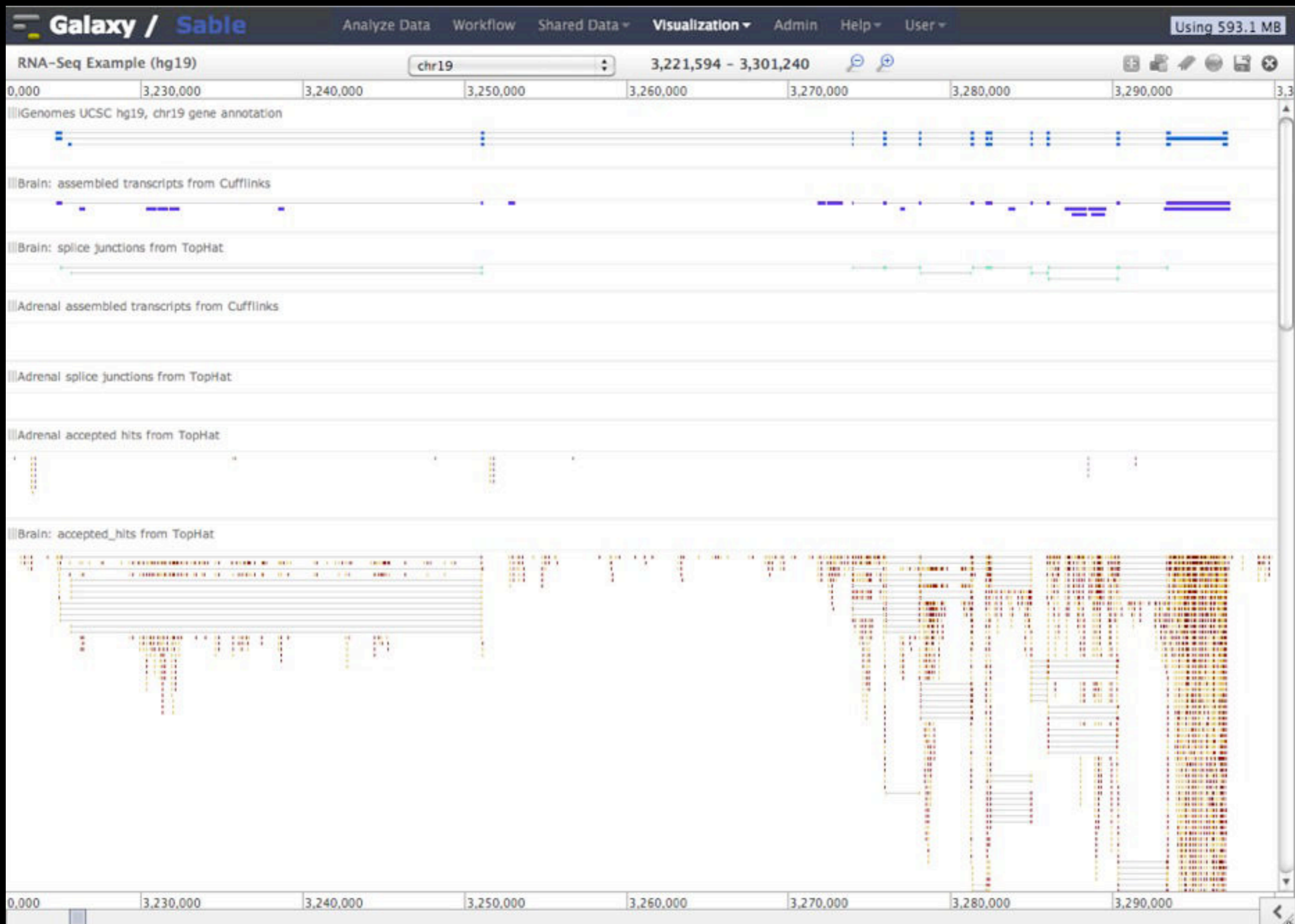
## Supported external browsers

- UCSC
- Ensembl
- GBrowse
- IGB
- IGV

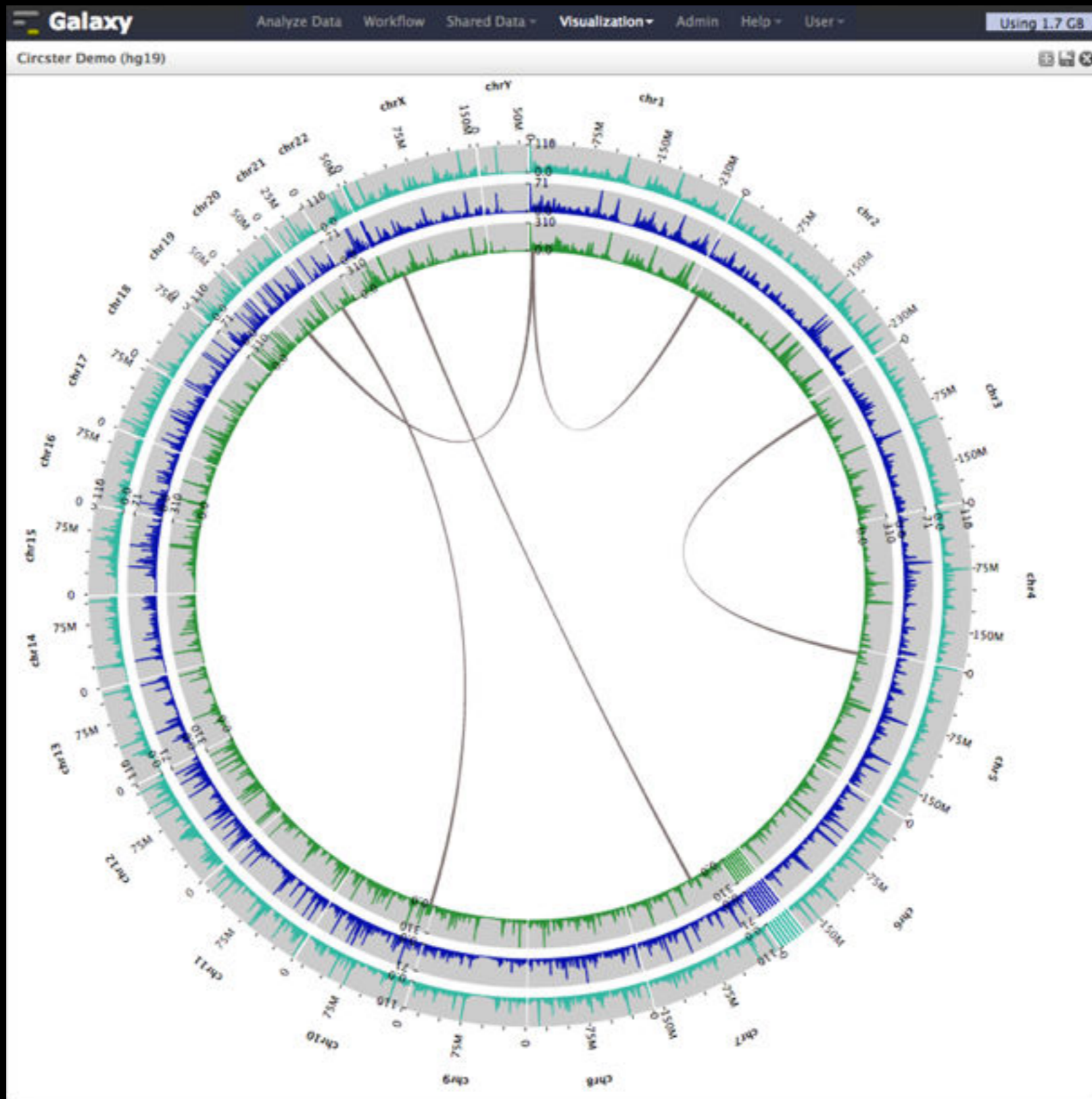
## Traditional browser strengths:

- Showing what is nearby
- what else is happening here
- highlighting correlations
- integrating many datasets

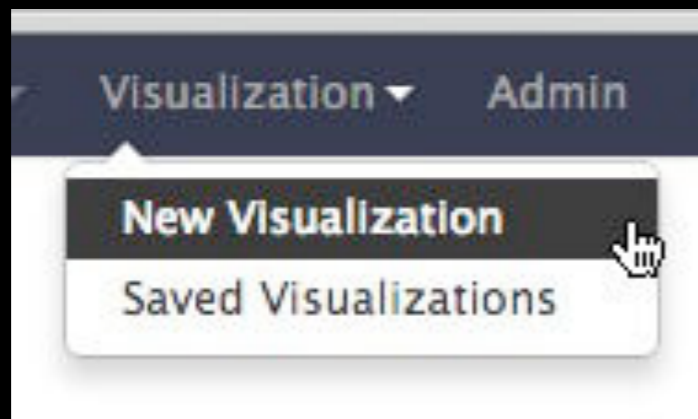
# Trackster: Galaxy's embedded track browser



# Circster



# Create a visualization in Galaxy



or

A screenshot of a Galaxy visualization panel. The panel title is '28: Brain: assembled transcripts from Cufflinks'. It shows 211 lines of data in gtf format for the hg19 database, generated by cufflinks v2.0.2. The command used is 'cufflinks -q --no-update-check -l 300000 -F 0.100000 -j 0.150000 -p 4'. The panel includes a 'Visualize' button and a 'main' link. Below the command, there is a table with columns: 1. Seqname, 2. Source, 3. Feature, and 4. Start. The table lists several transcripts and exons on chromosome 19.

| 1. Seqname | 2. Source | 3. Feature | 4. Start |
|------------|-----------|------------|----------|
| chr19      | Cufflinks | transcript | 33480    |
| chr19      | Cufflinks | exon       | 33480    |
| chr19      | Cufflinks | transcript | 33490    |
| chr19      | Cufflinks | exon       | 33490    |
| chr19      | Cufflinks | transcript | 33510    |
| chr19      | Cufflinks | exon       | 33510    |

# Vizualization inside Galaxy

- Leverage visualization to **evaluate and refine analyses**
- Make the *analyze-visualize-refine* loop seamless and **fast**
- Enable **experimenting with tools and their parameter space**
- Support **custom genome browsers**



# Transcript Prediction: Further Reading & Resources

[Princeton HTSEQ Users RNA-Seq Tutorial](#)

by Lance Parsons

[Gene Construction](#)

By Monica Britton

[Web-based visual analysis for high-throughput genomics](#)

by Goecks, *et al.*

[Cufflinks Manual](#)

# Tuesday Agenda



- 8:30 **Welcome and Questions**
- 8:45 **RNA-Seq: Mapping with Tophat**
- 9:45 **Cufflinks: Transcript prediction**
- 10:30 **Break**
- 11:00 **Differential Expression Analysis with Cuffdiff**
- 12:00 **Cuffdiff alternative: DESeq, Part I**
- 12:30 **Lunch**
- 13:30 **Cuffdiff alternative: DESeq, Part II**
- 14:00 **SNP and Variant Analysis, Part I**
- 15:00 **Break**
- 15:30 **SNP and Variant Analysis, Part II**
- 16:30 **Done**

# Tuesday Agenda

- 8:30 **Welcome and Questions**
- 8:45 **RNA-Seq: Mapping with Tophat**
- 9:45 **Cufflinks: Transcript prediction**
- 10:30 **Break**
- 11:00 **Differential Expression Analysis with Cuffdiff**
- 12:00 Cuffdiff alternative: DESeq, Part I
- 12:30 Lunch
- 13:30 Cuffdiff alternative: DESeq, Part II
- 14:00 SNP and Variant Analysis, Part I
- 15:00 Break
- 15:30 SNP and Variant Analysis, Part II
- 16:30 Done

# Cuffdiff

- Identifies differential expression between multiple datasets
- Uses RPKM/FPKM as its guiding statistic
- RPKM/FPKM attempts to track expression levels of each feature relative to total expression in the dataset

# Cuffdiff

- Which Transcript definitions to use?
  - Official
  - MeOH or R3G Cufflinks transcripts
  - Results of **Cuffmerge** on MeOH & R3G Cufflinks transcripts
- Depends on what you care about

**NGS: RNA Analysis → Cuffdiff**

# Cuffdiff

- Produces 11 output files, all explained in doc
- We'll focus on gene/transcript differential expression testing files (also care about gene/transcript FPKM files)
- Column 7 ("status") can be FAIL, NOTEST, LOWDATA or OK
  - Filter and Sort → **Filter**
    - **c7 == 'OK' or C7 == 'LOWDATA'**
- Column 14 ("significant") can be yes or no
  - **c14 == 'yes'**

# Tuesday Agenda

- 8:30 Welcome and Questions
- 8:45 RNA-Seq: Mapping with Tophat
- 9:45 Cufflinks: Transcript prediction
- 10:30 Break
- 11:00 Differential Expression Analysis with Cuffdiff
- 12:00 Cuffdiff alternative: DESeq, Part I
- 12:30 Lunch
- 13:30 Cuffdiff alternative: DESeq, Part II
- 14:00 SNP and Variant Analysis, Part I
- 15:00 Break
- 15:30 SNP and Variant Analysis, Part II
- 16:30 Done

# Alternatives

- Yesterday we used **Tophat** (calling **Bowtie**) to **map RNA-Seq reads to the genome**
- Today we used **Cuffdiff** to **identify differentially expressed genes** across two experimental conditions
- Tophat, Bowtie and Cuffdiff are widely installed on many Galaxy instances, including CloudMan based instances
- but ...



# Alternatives

Lindner R, Friedel CC (2012) "A Comprehensive **Evaluation of Alignment Algorithms** in the Context of RNA-Seq." *PLoS ONE* 7(12): e52403. doi:10.1371/journal.pone.0052403

reviews **14 packages** (for slightly different problem of transcriptome alignment)

Rapaport, *et al.*, "Comprehensive **evaluation of differential gene expression analysis** methods for RNA-seq data." *Genome Biology* 2013, 14:R95 doi:10.1186/gb-2013-14-9-r95

reviews **7 packages**

Each has it's own strengths and weaknesses.

**What's a biologist to do?**

# Alternatives: **What's a biologist to do?**

Learn the strengths and weaknesses of the tools you have ready access to. Are they a good match for the questions you are asking?

If not, then research alternatives, identify good options and then work with your bioinformatics/systems people to get access to those tools.

Note: on Friday we will learn how to install alternatives in Galaxy.

# Cuffdiff Alternatives: DESeq

## Cufflinks

Uses **FPKM/RPKM** as a central statistic.

Total # mapped reads heavily influences FPKM/RPKM.

Can lead to challenges when you have very highly expressed genes in the mix.

## DESeq (and edgeR)

DESeq is an R based differential expression analysis package where expression analysis is much more effectively isolated between features.

# Cuffdiff Alternatives: DESeq

Takes a simple, tab delimited list of features and read counts across different samples.

First, have to create that list.

NGS: SAM Tools → htseq-count  
once for each BAM file

Join the 4 HTSeq datasets together on gene name

Cut out the duplicate gene name columns

NGS: RNA Analysis → DE Seq

# Cuffdiff Alternatives: DESeq

DESeq output is a list of genes,  
sorted by adjusted P value,  
with lowest P values listed first

## Guided Exercise:

How many genes have an adjusted P value  $< 0.05$  ?

How does this gene list compare to the list  
produced by Cufflinks?

# Cuffdiff Alternatives: Further Reading & Resources

Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data

by Rapaport, *et al.*

DESeq Reference Manual

DESeq Galaxy Wrapper

by Nikhil Joshi

htseq-count Galaxy Wrapper

by Lance Parsons

# Tuesday Agenda

- 8:30 Welcome and Questions
- 8:45 RNA-Seq: Mapping with Tophat
- 9:45 Cufflinks: Transcript prediction
- 10:30 Break
- 11:00 Differential Expression Analysis with Cuffdiff
- 12:00 Cuffdiff alternative: DESeq, Part I
- 12:30 Lunch
- 13:30 Cuffdiff alternative: DESeq, Part II
- 14:00 SNP and Variant Analysis, Part I
- 15:00 Break
- 15:30 SNP and Variant Analysis, Part II
- 16:30 Done

**Thanks**



**Dave Clements**

**Galaxy Project  
Emory University**

**[clements@galaxyproject.org](mailto:clements@galaxyproject.org)**