NGS Data Analysis and Galaxy

University of Cape Town
Cape Town, South Africa
21-25 October 2013

Dave Clements, Emory University
http://galaxyproject.org/

Gerrit Botha
Computational Biology
University of Cape Town
http://cbio.uct.ac.za/

CBIO Computational Biology @ UCT

AFRICAN CENTRE FOR GENE TECHNOLOGIES

Galaxy

amazon web services™

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

South Africa Galaxy Workshop Tour

This Week

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

Tuesday Agenda

9:00 Welcome and Questions
9:15 RNA-Seq: Mapping with Tophat
10:15 Cufflinks: Transcript prediction
11:00 Break
11:30 Differential Expression Analysis with Cuffdiff
12:30 Cuffdiff alternative: DESeq, Part I
13:00 Lunch
14:00 Cuffdiff alternative: DESeq, Part II
14:30 SNP and Variant Analysis, Part I
15:30 Break
16:00 SNP and Variant Analysis, Part II
17:00 Done

Tuesday Agenda

9:00 Welcome and Questions
9:15 RNA-Seq: Mapping with Tophat
10:15 Cufflinks: Transcript prediction
11:00 Break
11:30 Differential Expression Analysis with Cuffdiff
12:30 Cuffdiff alternative: DESeq, Part I
13:00 Lunch
14:00 Cuffdiff alternative: DESeq, Part II
14:30 SNP and Variant Analysis, Part I
15:30 Break
16:00 SNP and Variant Analysis, Part II
17:00 Done

RNA-seq Exercise: Mapping with Tophat

Create a new history

Import all datasets from library:

UC Davis RNA-Seq TopHat Inputs

Get all datasets, and

UC Davis RNA-Seq Human

Get genes_chr12.gtf

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

From the 2013 UC Davis Bioinformatics Short Course

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

TopHat Manual

Mapping with Tophat: Max # of Alignments Allowed

Some reads align to more than one place equally well.

For such reads, how many should Tophat include?

If more than the specified number, Tophat will pick those with the best mapping score.

Tophat break ties randomly.

Tophat assigns equal fractional credit to all *n* mappings

Instructs TopHat to allow up to this many alignments to the reference for a given read, and choose the alignments based on their alignment scores if there are more than this number. The default is 20 for read mapping. Unless you use --report-secondary-alignments, TopHat will report the alignments with the best alignment score. If there are more alignments with the same score than this number, TopHat will randomly report only this many alignments. In case of using --report-secondary-alignments, TopHat will try to report alignments up to this option value, and TopHat may randomly output some of the alignments with the same score to meet this number.

TopHat Manual

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: 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

RNA-Seq Analysis with Galaxy
by Curtis Hendrickson, David Crossman, Jeremy Goecks

from the GCC2012 Training Day

Tophat Manual

Tuesday Agenda

9:00 Welcome and Questions
9:15 RNA-Seq: Mapping with Tophat
10:15 Cufflinks: Transcript prediction
11:00 Break
11:30 Differential Expression Analysis with Cuffdiff
12:30 Cuffdiff alternative: DESeq, Part I
13:00 Lunch
14:00 Cuffdiff alternative: DESeq, Part II
14:30 SNP and Variant Analysis, Part I
15:30 Break
16:00 SNP and Variant Analysis, Part II
17:00 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

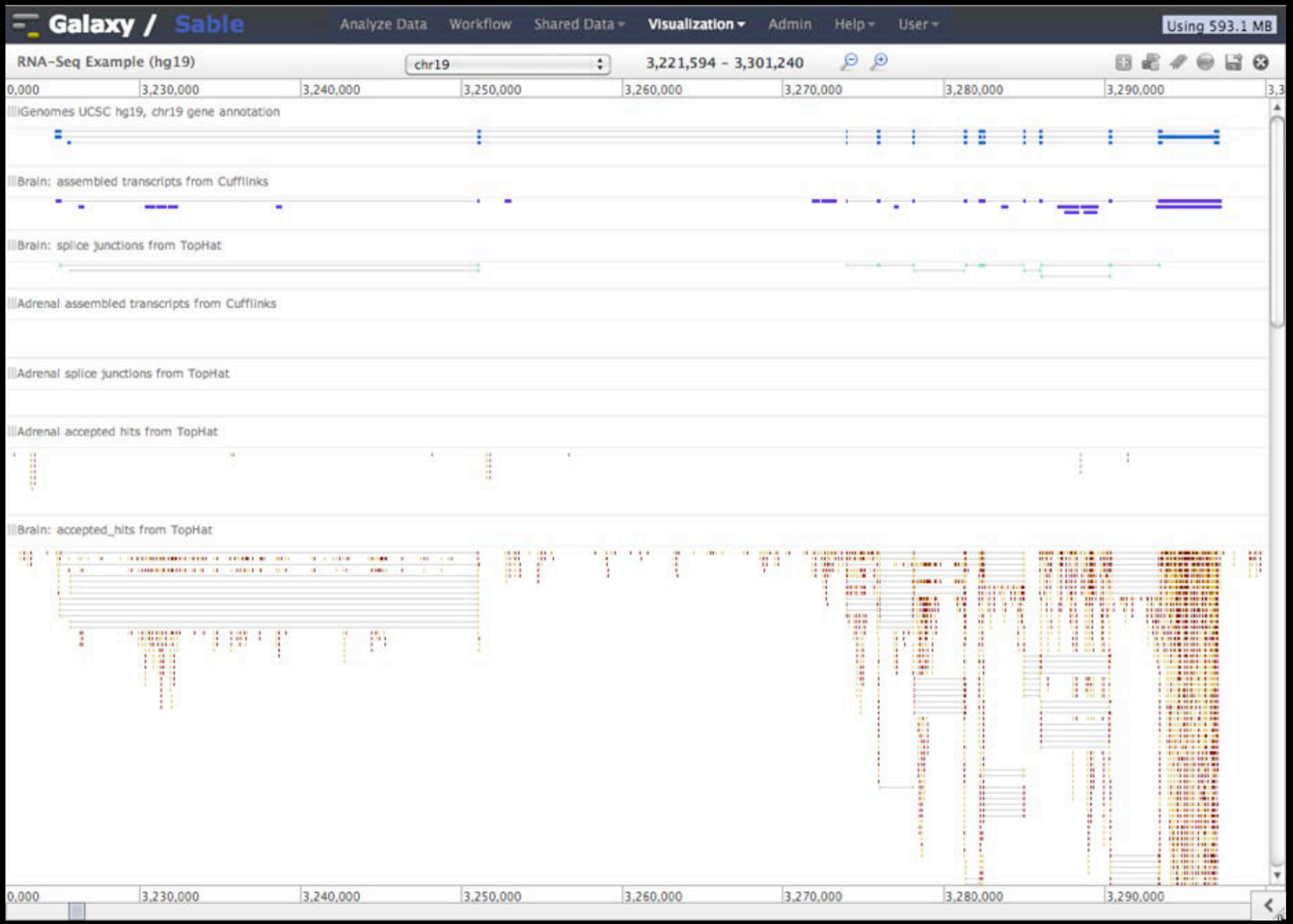
Supported external browsers

- UCSC
- Ensembl
- GBrowse
- IGB
- IGV

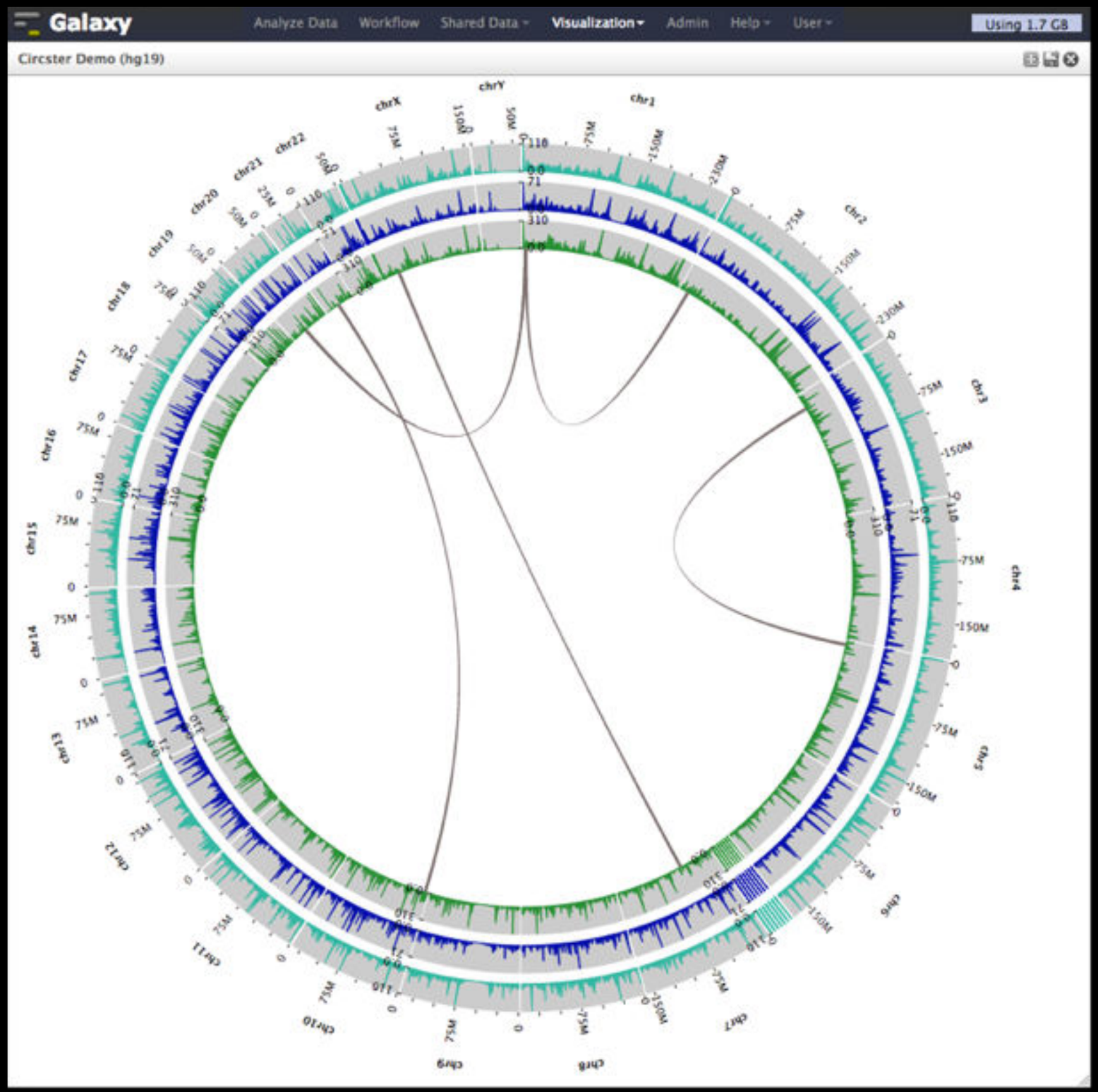
Traditional browser strengths:

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

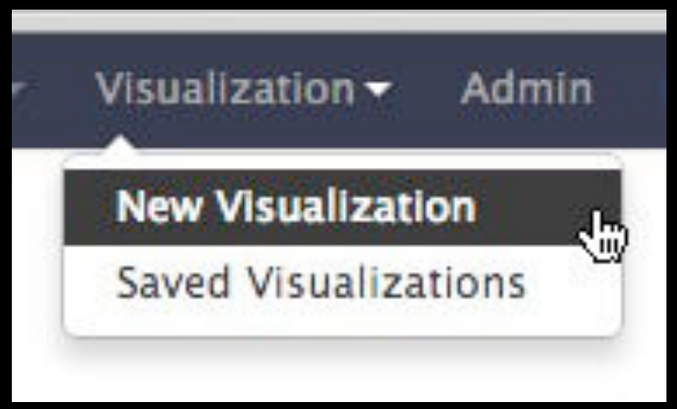
Trackster: Galaxy's embedded track browser



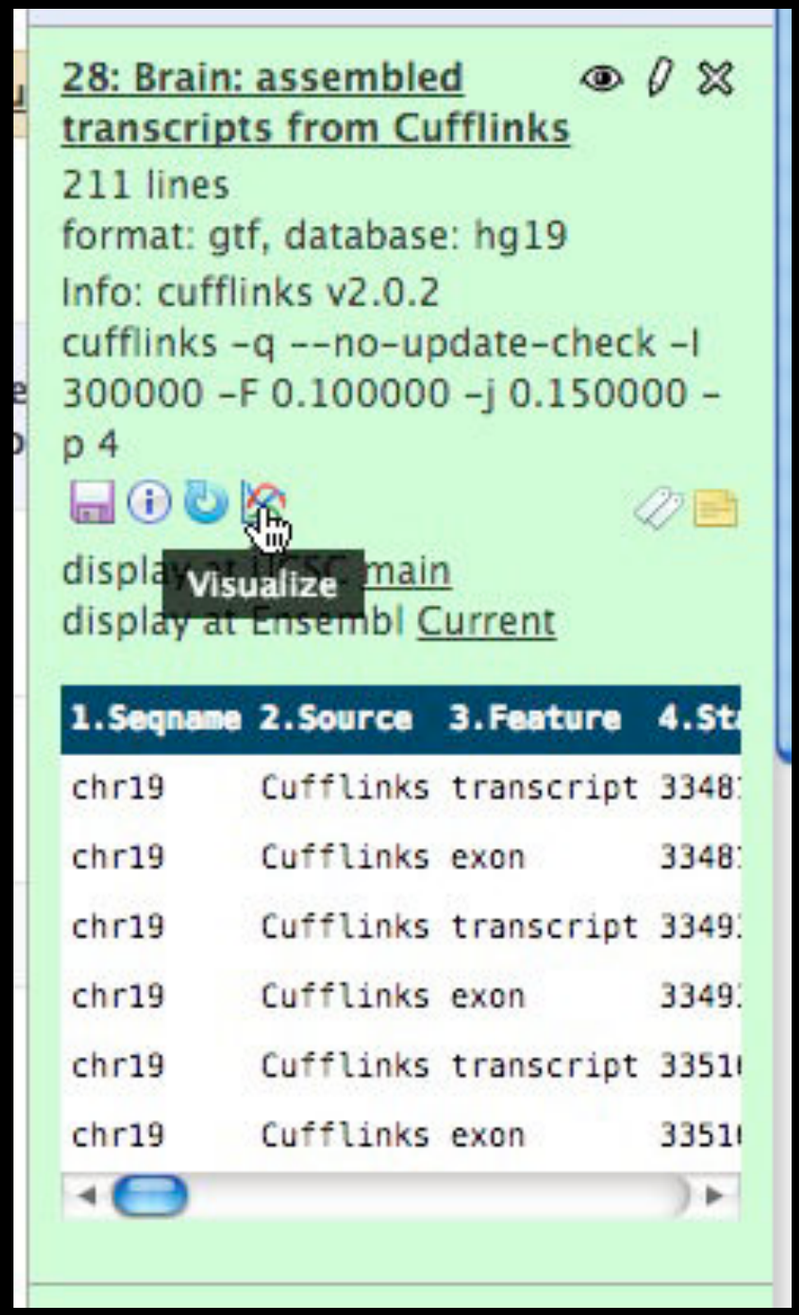
Circster


Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 1.7 GB
Circster Demo (hg19)
chrY
chrX
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22

Create a visualization in Galaxy



or



Vizualization inside Galaxy

- Levarge 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

9:00 Welcome and Questions
9:15 RNA-Seq: Mapping with Tophat
10:15 Cufflinks: Transcript prediction
11:00 Break
11:30 Differential Expression Analysis with Cuffdiff
12:30 Cuffdiff alternative: DESeq, Part I
13:00 Lunch
14:00 Cuffdiff alternative: DESeq, Part II
14:30 SNP and Variant Analysis, Part I
15:30 Break
16:00 SNP and Variant Analysis, Part II
17:00 Done

Tuesday Agenda

9:00	Welcome and Questions
9:15	RNA-Seq: Mapping with Tophat
10:15	Cufflinks: Transcript prediction
11:00	Break
11:30	Differential Expression Analysis with Cuffdiff
12:30	Cuffdiff alternative: DESeq, Part I
13:00	Lunch
14:00	Cuffdiff alternative: DESeq, Part II
14:30	SNP and Variant Analysis, Part I
15:30	Break
16:00	SNP and Variant Analysis, Part II
17:00	Done

Cuffdiff

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

Cuffdiff

- Running with 2 Groups: MeOH and R3G
- Each group has 2 replicates each

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 differential expression testing files (also care about gene FPKM files)
- Column 7 ("status") can be FAIL, NOTEST, LOWDATA or OK
- Filter and Sort → Filter
  - c7 == 'OK'
  - Column 14 ("significant") can be yes or no
- c14 == 'yes'

Tuesday Agenda

9:00 Welcome and Questions
9:15 RNA-Seq: Mapping with Tophat
10:15 Cufflinks: Transcript prediction
11:00 Break
11:30 Differential Expression Analysis with Cuffdiff
12:30 Cuffdiff alternative: DESeq, Part I
13:00 Lunch
14:00 Cuffdiff alternative: DESeq, Part II
14:30 SNP and Variant Analysis, Part I
15:30 Break
16:00 SNP and Variant Analysis, Part II
17:00 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 transciptome 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 tool 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 some alternatives in Galaxy.

Cuffdiff Alternatives: DESeq

Cuffdiff

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 Cuffdiff?

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

9:00 Welcome and Questions
9:15 RNA-Seq: Mapping with Tophat
10:15 Cufflinks: Transcript prediction
11:00 Break
11:30 Differential Expression Analysis with Cuffdiff
12:30 Cuffdiff alternative: DESeq, Part I
13:00 Lunch
14:00 Cuffdiff alternative: DESeq, Part II
14:30 SNP and Variant Analysis, Part I
15:30 Break
16:00 SNP and Variant Analysis, Part II
17:00 Done

Thanks

Dave Clements

Galaxy Project
Emory University

clements@galaxyproject.org