

Introduction to Exome Analysis in Galaxy

Carol Bult, Ph.D.
Professor
Deputy Director, JAX Cancer Center

Short Course Bioinformatics Workshops
2014

Disclaimer...I am on the Galaxy Advisory Board

- You can do these exercises in Galaxy without an account
 - But without an account you can't save your work
- Many of the data files are LARGE and will take awhile to upload
 - Available from a public DropBox account

https://www.dropbox.com/sh/seqishl631f363x/AADoBskzYjPt_ijI21dxJqV8a

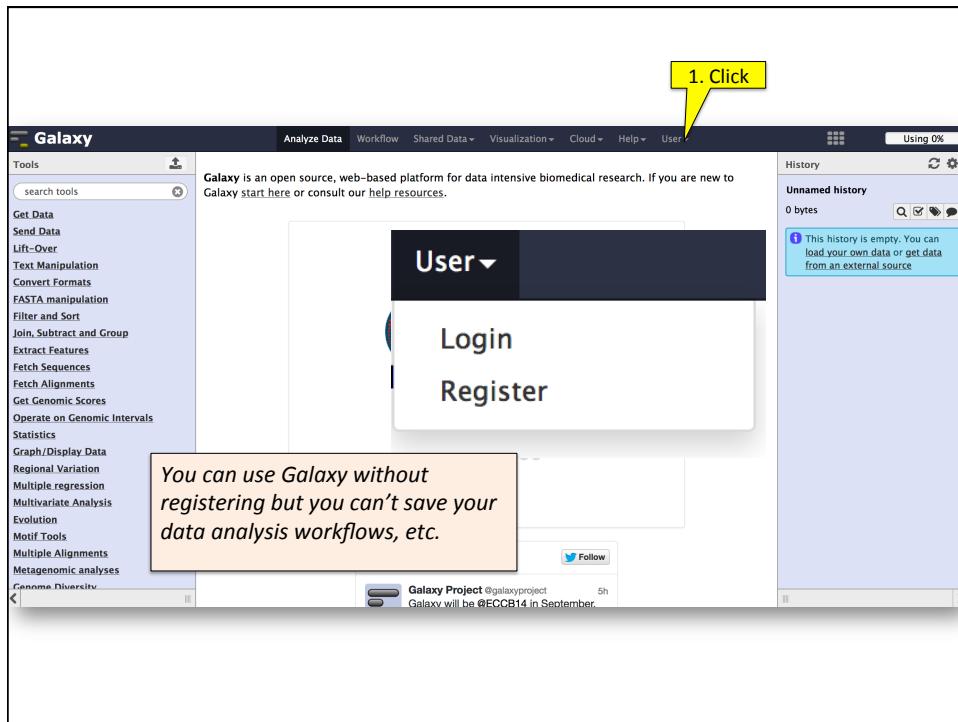
Galaxy

The screenshot shows the Galaxy web interface. On the left, there's a sidebar titled "List of Tools" containing various analysis tools like "Sequence Manipulation", "Genomic Scores", and "Multiple Alignments". The main content area features a "Dialog panel" with the title "Running Your Own Understanding how Galaxy works" and a "Tweets" sidebar showing tweets from the Galaxy Project and BOSC. On the right, there's an "Analysis history" panel which is currently empty. Logos for Penn State, Johns Hopkins, TACC, and iPlant Collaborative are at the bottom.

<https://usegalaxy.org/>

Galaxy in a Nutshell

- Analyze
 - Build reusable analysis workflows for many types of data analysis needs
 - Interactive analysis
 - Reproducible analysis pipelines
 - Many analysis tools available ... ready to use!
- Visualize
 - Send analysis results to standard genome browsers
- Publish and Share
 - Saved histories record the details of your analysis steps
 - Open access to data and analysis results to colleagues
 - Package analysis results for publication



Example of what the history of analysis would look like in Galaxy.

Green color means the analysis completed successfully.

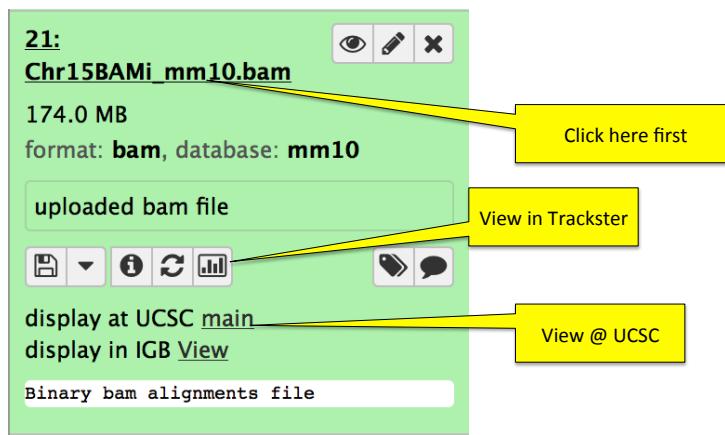
Once you have a history that works you can share it or turn it into a workflow

Click on the cog to see History options

- HISTORY LISTS
- Copy item
- Saved Histories
- Histories Shared with Me
- CURRENT HISTORY
- Create New
- Copy History
- Copy Datasets
- Share or Publish
- Extract Workflow
- Dataset Security
- Resume Paused Jobs

Visualization

- Galaxy results include links to visualization tools
 - UCSC Genome Browser, IGV (external)
 - Trackster (Galaxy's internal viz tool)



Fairfield et al. *Genome Biology* 2011, **12**:R86
<http://genomebiology.com/2011/12/9/R86>

Genome Biology

METHOD Open Access

Mutation discovery in mice by whole exome sequencing

Heather Fairfield¹, Griffith J Gilbert¹, Mary Barter¹, Rebecca R Corrigan², Michelle Curtain¹, Yueming Ding³, Mark D'Ascenzo⁴, Daniel J Gerhardt⁴, Chao He⁵, Wenhui Huang⁶, Todd Richmond⁴, Lucy Rowe¹, Frank J Probst², David E Bergstrom¹, Stephen A Murray¹, Carol Bult¹, Joel Richardson¹, Benjamin T Kile⁷, Ivo Gut⁸, Jorg Hager⁸, Snaevar Sigurdsson⁹, Evan Mauceli⁹, Federica Di Palma⁹, Kentin Lindblad-Toh⁹, Michael L Cunningham¹⁰, Timothy C Cox¹⁰, Monica J Justice², Mona S Spector⁵, Scott W Lowe⁵, Thomas Albert⁴, Leah Rae Donahue¹, Jeffrey Jeddeloh⁴, Jay Shendre¹⁰ and Laura G Reinholdt¹



A heterozygous mutant +/M2J on the left and littermate control on the right.

Example data for exome analysis from Fairfield et al., 2011

Focus on the Cleft mutant

The exome data for the Cleft mutant are in the NCBI SRA and can be downloaded directly from there.

Data you download from SRA should already be in Sanger fastq format

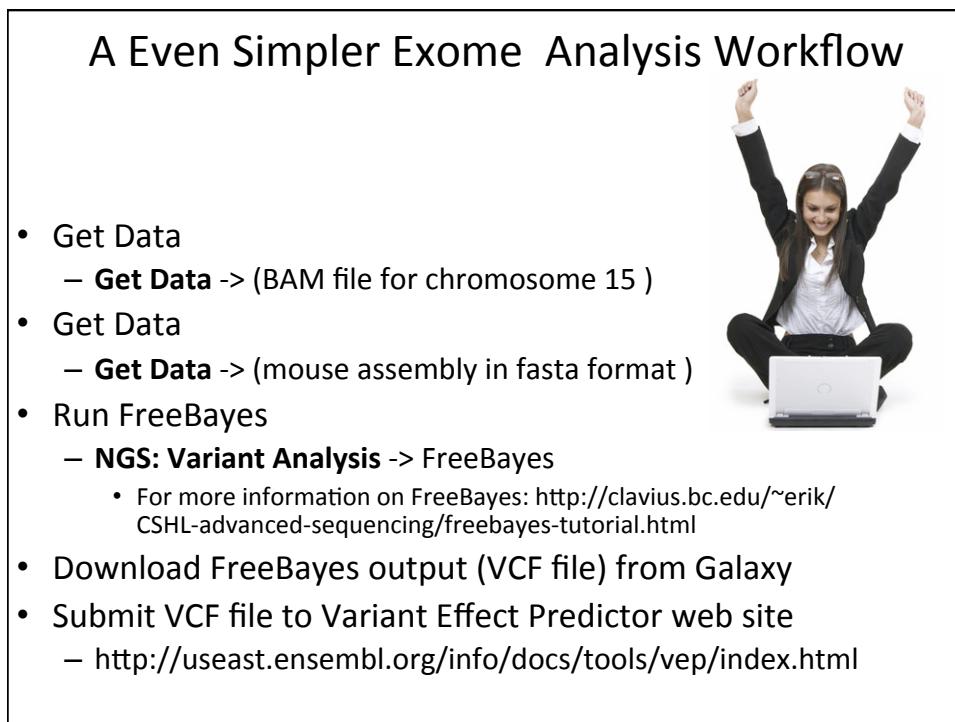
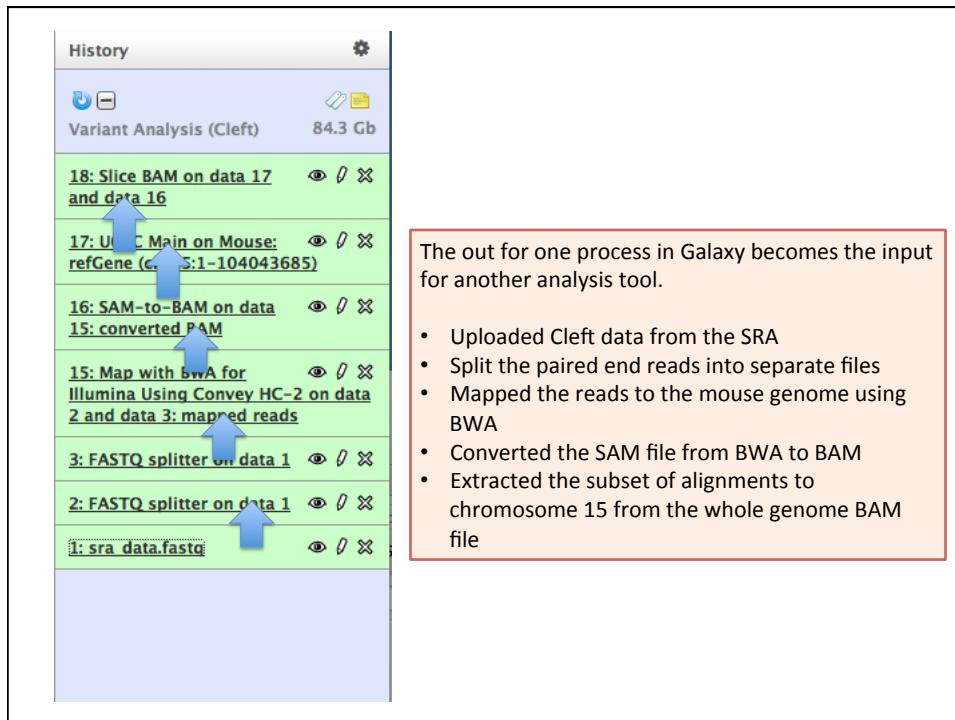
[http://trace.ncbi.nlm.nih.gov/Traces/sra/?
view=search_seq_name&exp=SRX089344&run=&m=search&s=seq](http://trace.ncbi.nlm.nih.gov/Traces/sra/?view=search_seq_name&exp=SRX089344&run=&m=search&s=seq)

A Simple Exome Analysis Workflow in Galaxy

- Get Data into Galaxy
 - **Get Data** -> (we'll get this from the short read archive @ NCBI)
- Split merged paired end sequence data file into forward and reverse (if necessary)
 - **NGS: QC and manipulation** -> Fastq splitter
- Run QC analysis
 - **NGS: QC and manipulation** -> Fastqc: Fastq QC
- Map sequence reads to reference genome
 - **NGS: Mapping** -> Map with BWA for Illumina
 - Select appropriate parameters
- Convert SAM to BAM
 - **NGS: SAM Tools** -> SAM-to-BAM
- Visualize Alignments
 - Select UCSC genome browser or Trackster in Galaxy OR...
 - Download BAM and BAM index files AND
 - Download and Install IGV from Broad
- Upload a Reference Genome (if one isn't already in Galaxy)
 - **Get Data** -> (Mouse_GRCm38p1.fasta)
- Call Variants
 - **NGS: Variant Detection** -> FreeBaye
- Annotate Variants
 - Download VCF file
 - Upload to Variant Effect Predictor (VEP) @ ENSEMBL

This workflow will take some time to run!!





FreeBayes will use your alignments in BAM format to look for variants.

24: FreeBayes on data 23 and data 21 (variants)

1,626 lines, 54 comments
format: vcf, database: mm10

23: Mouse_GRCm38.fasta

22 sequences
format: fasta, database: mm10

21: Chr15BAM_GRCm38.bam

174.0 MB
format: bam, database: mm10

uploaded bam file

display at UCSC main
display in IGB View
Binary bam alignments file

24: FreeBayes on data 23 and data 21 (variants)

1,626 lines, 54 comments
format: vcf, database: mm10

23: Mouse_GRCm38.fasta

22 sequences
format: fasta, database: mm10

uploaded fasta file

21: Chr15BAM_GRCm38.bam

174.0 MB
format: bam, database: mm10

uploaded bam file

display at UCSC main
display in IGB View
Binary bam alignments file

1 FreeBayes version freebayes-0.9.14

Load reference genome from:
Local cache

Sample BAM files
Sample BAM file 1
BAM file:
21: Chr15BAM_GRCm38.bam

Using reference genome:
12931

Limit variant calling to a set of regions
Do not limit

Choose parameter selection level:
1:Simple diploid calling
Select how much control over the freebayes

Execute

2 Load reference genome from:
Local cache
History

3 Use the following dataset as the reference sequence:
23: Mouse_GRCm38.fasta

You can upload a FASTA sequence to the history and use it as reference

FreeBayes dialog box in Galaxy (1). Chances are the mouse genome won't be available. So upload your own reference from your history Select History (2) See the result (3)
Note that Galaxy autodetected the BAM file in your history!

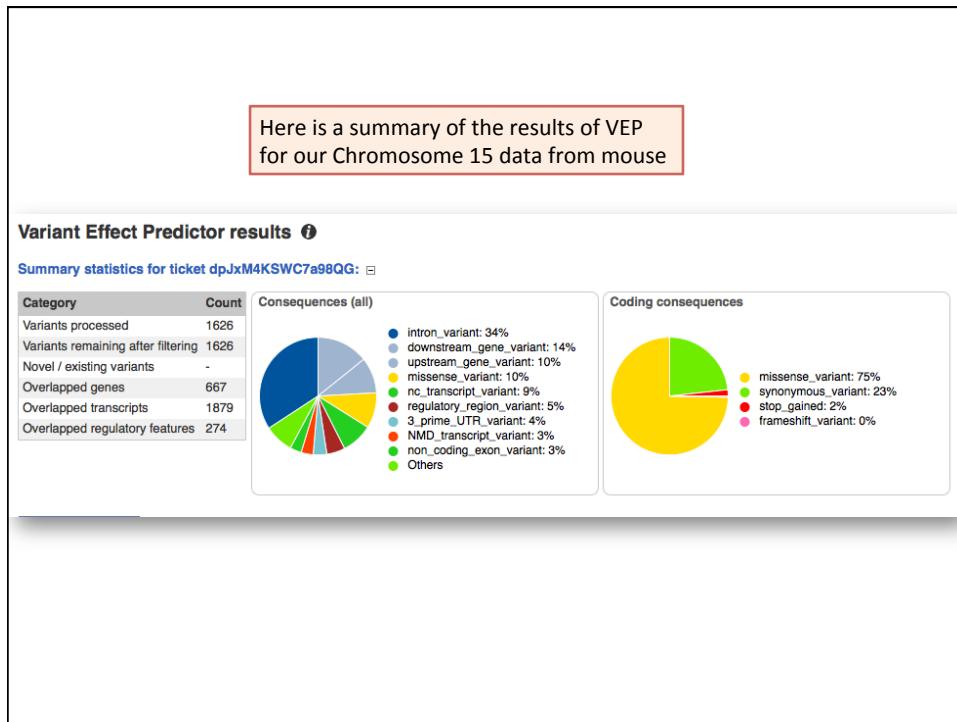
Once you have a VCF file you want to know about the nature of the variants, right?

There are some tools in Galaxy that can help with this... but VEP @ Ensembl is a great tool.

<http://useast.ensembl.org/info/docs/tools/vep/index.html>

There is a VCF file already “done” on the DropBox site that you can try with VEP.

mChr15_Cleft.vcf



Results preview

Navigation ⌂ **Filters** ⌂ **Download**

Showng 29 results for variants 1-6 | Show 1 5 10 50 All Gene is ENSMUSG00000022483

Uploaded variation is defined Add BioMart Variants Genes

Show/hide columns

Location	Allele	Gene	Feature	Feature type	Consequence	Amino acids	Codons	Symbol	Symbol source	SIFT
15:97979015	C	ENSMUSG00000022483	ENSMUST00000023123	Transcript	intron_variant	-	-	Col2a1	MGI	-
15:97979015	C	ENSMUSG00000022483	ENSMUST00000088355	Transcript	intron_variant	-	-	Col2a1	MGI	-
15:97983167	G	ENSMUSG00000022483	ENSMUST00000128547	Transcript	downstream_gene_variant	-	-	Col2a1	MGI	-
15:97983167	G	ENSMUSG00000022483	ENSMUST00000023123	Transcript	intron_variant	-	-	Col2a1	MGI	-
15:97983167	G	ENSMUSG00000022483	ENSMUST00000131910	Transcript	downstream_gene_variant	-	-	Col2a1	MGI	-
15:97983167	G	ENSMUSG00000022483	ENSMUST00000088355	Transcript	intron_variant	-	-	Col2a1	MGI	-
15:97984776	A	ENSMUSG00000022483	ENSMUST00000128547	Transcript	downstream_gene_variant	-	-	Col2a1	MGI	-
15:97984776	A	ENSMUSG00000022483	ENSMUST00000023123	Transcript	stop_gained	Q/*	CAG/TAG	Col2a1	MGI	-
15:97984776	A	ENSMUSG00000022483	ENSMUST00000131910	Transcript	downstream_gene_variant	-	-	Col2a1	MGI	-
15:97984776	A	ENSMUSG00000022483	ENSMUST00000088355	Transcript	stop_gained	Q/*	CAG/TAG	Col2a1	MGI	-
15:97984776	A	ENSMUSG00000022483	ENSMUST00000088355	Transcript	stop_gained	-	-	Col2a1	MGI	-
15:97984776	A	ENSMUSG00000022483	ENSMUST00000088355	Transcript	stop_gained	L/F	TTG/TTT	Col2a1	MGI	0.13
15:97984776	A	ENSMUSG00000022483	ENSMUST00000088355	Transcript	stop_gained	-	-	Col2a1	MGI	-
15:97984776	A	ENSMUSG00000022483	ENSMUST00000088355	Transcript	stop_gained	L/F	TTG/TTT	Col2a1	MGI	0.16
15:97984776	A	ENSMUSG00000022483	ENSMUST00000088355	Transcript	stop_gained	-	-	Col2a1	MGI	-
15:97984776	A	ENSMUSG00000022483	ENSMUST00000088355	Transcript	stop_gained	-	-	Col2a1	MGI	-
15:97984776	A	ENSMUSG00000022483	ENSMUST00000088355	Transcript	stop_gained	-	-	Col2a1	MGI	-
15:97984776	A	ENSMUSG00000022483	ENSMUST00000088355	Transcript	stop_gained	D/A	GAC/GCC	Col2a1	MGI	0.75
15:97984776	A	ENSMUSG00000022483	ENSMUST00000127879	Transcript	non_coding_exon_variant, nc_transcript_variant	-	-	Col2a1	MGI	-
15:98000420	G	ENSMUSG00000022483	ENSMUST00000023123	Transcript	missense_variant	-	-	Col2a1	MGI	-
15:98000420	G	ENSMUSG00000022483	ENSMUST00000140064	Transcript	downstream_gene_variant	-	-	Col2a1	MGI	-
15:97995312	T	ENSMUSG00000022483	ENSMUST00000140064	Transcript	intron_variant	-	-	Col2a1	MGI	-
15:97995312	T	ENSMUSG00000022483	ENSMUST00000088355	Transcript	intron_variant	-	-	Col2a1	MGI	-
15:98000420	G	ENSMUSG00000022483	ENSMUST00000023123	Transcript	missense_variant	D/A	GAC/GCC	Col2a1	MGI	0.08
15:98000420	G	ENSMUSG00000022483	ENSMUST00000127879	Transcript	non_coding_exon_variant, nc_transcript_variant	-	-	Col2a1	MGI	-
15:98000420	G	ENSMUSG00000022483	ENSMUST000000131560	Transcript	missense_variant	D/A	GAC/GCC	Col2a1	MGI	0.08
15:98000420	G	ENSMUSG00000022483	ENSMUST00000140064	Transcript	non_coding_exon_variant, nc_transcript_variant	-	-	Col2a1	MGI	-
15:98000420	G	ENSMUSG00000022483	ENSMUST00000133488	Transcript	upstream_gene_variant	-	-	Col2a1	MGI	-
15:98000420	G	ENSMUSG00000022483	ENSMUST00000139246	Transcript	upstream_gene_variant	-	-	Col2a1	MGI	-
15:98000420	G	ENSMUSG00000022483	ENSMUST00000088355	Transcript	intron_variant	-	-	Col2a1	MGI	-

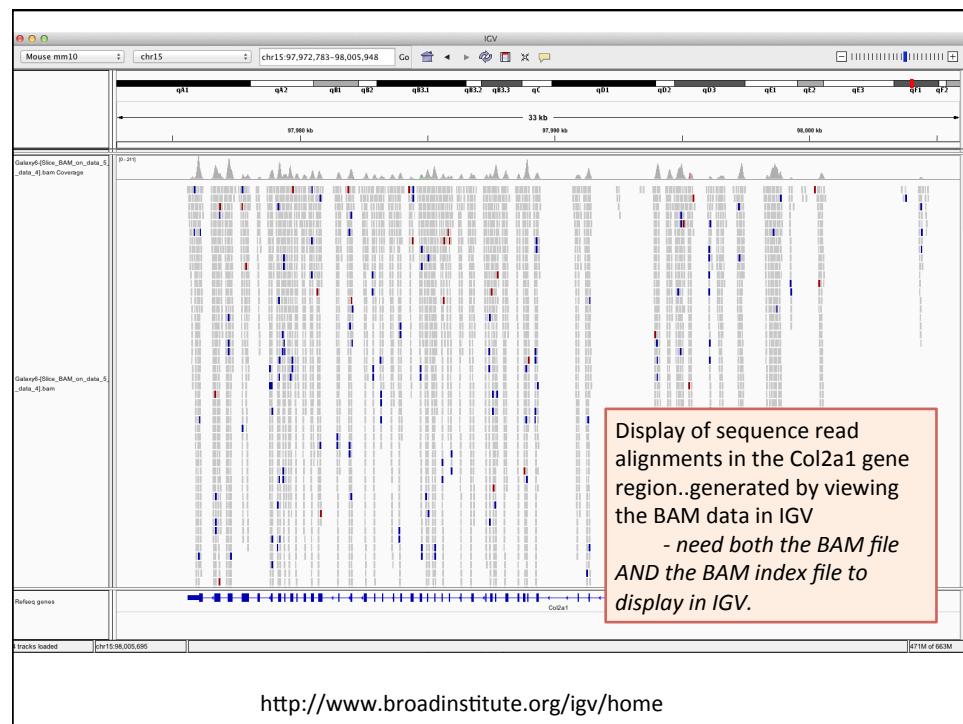
Here is the detailed annotation for the variant calls.

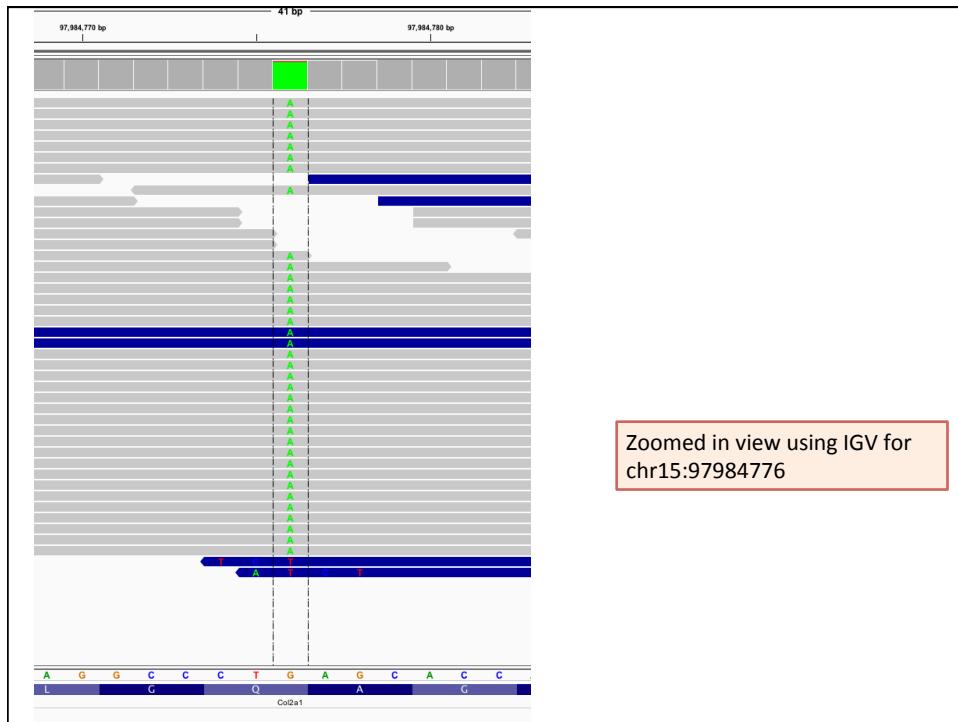
VEP lets you filter this by a number of parameters, including the predicted consequence of the detected variants.

So...which might be the causative mutation?
Not a push button answer....

Cleft is a dominant craniofacial ENU mutation that causes cleft palate. Of the two variants that were nominated for validation, both were SNVs residing in Col2a1, a gene coding for type II procollagen. Both SNVs reside within 10 kb of each other (Chr15:97815207 and Chr15:97825743) in Col2a1, a gene coding for type II procollagen, and not surprisingly were found to be concordant with the phenotype when multiple animals from the pedigree were genotyped. The most likely causative lesion (G to A at Chr15:97815207) is a nonsense mutation that introduces a premature stop codon at amino acid 645. The second closely linked variant is an A to T transversion in intron 12 that could potentially act as a cryptic splice site. However, since RTPCR did not reveal splicing abnormalities, it is more likely that the nonsense mutation is the causative lesion (Figure 2b). Mice homozygous for targeted deletions in Col2a1 and mice homozygous for a previously characterized, spontaneous missense mutation, Col2a1sedc, share similar defects in cartilage development to Cleft mutants, including recessive peri-natal lethality and orofacial clefting [19,20], providing further support that the Cleft phenotype is the result of a mutation in Col2a1.

*The location of the variants in this paper refer to NCBI³⁷...but I mapped the reads to GRCm38.p1.
How do you map Build 37 coordinates to Build 38?*





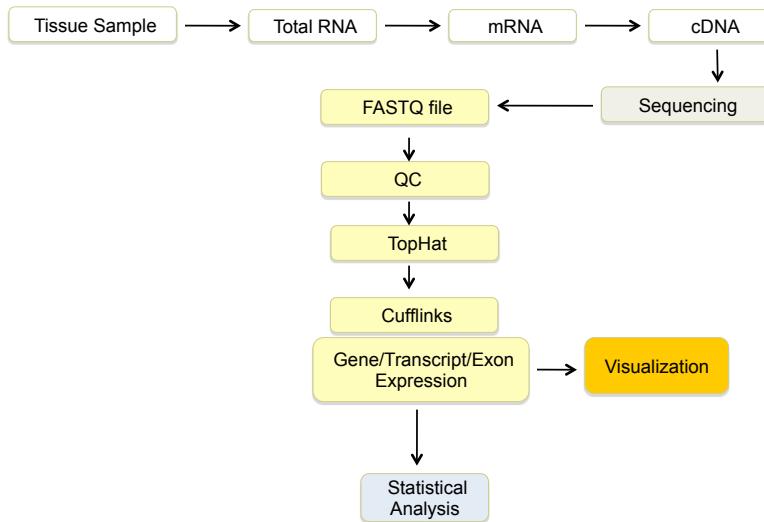
Your Turn

- We've supplied data files to let you do the longer or shorter exome workflow.
- **Start with the simple workflow (Chr15BAM + GRCm38.p1 -> FreeBayes).**
 - Note...the BAM file for the entire exome data set (not just chr 15) for the Cleft mutant is also available
 - MMR_12724_GES_JAX_Lmerged_aln.bam

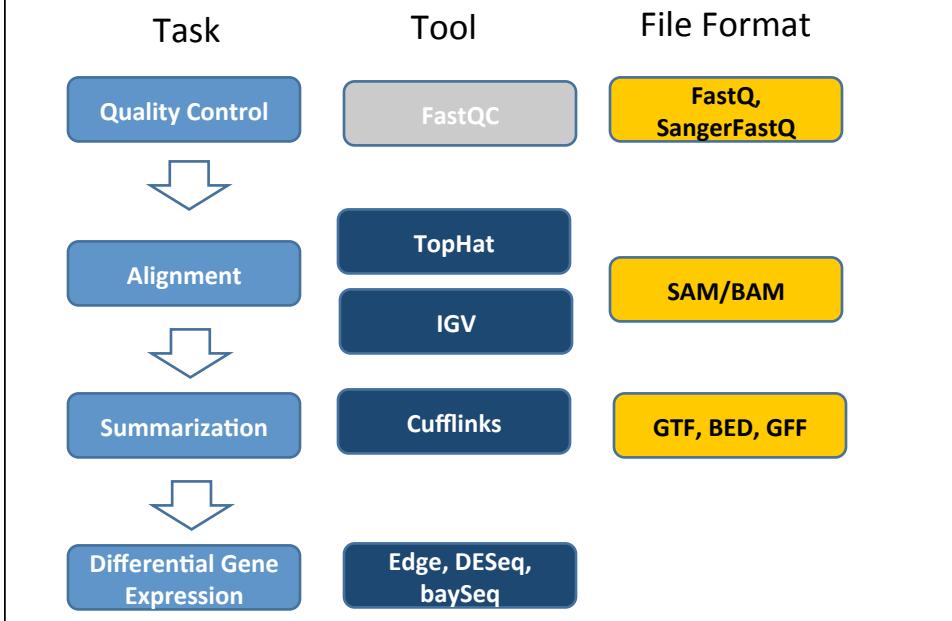
A Simple RNA Seq Workflow

- Get sequence data
- Align sequence data to a reference
 - TopHat is commonly used because it is tuned to aligning transcripts to a genome (splice site aware)
 - <http://ccb.jhu.edu/software/tophat/index.shtml>
- Associate aligned reads to transcript models/ annotations
 - Cufflinks
- Quantitation of expression/Differential gene expression
 - Cuffmerge/Cuffdiff
 - <http://cufflinks.cbcb.umd.edu/tutorial.html>

Typical RNA_Seq Project Work Flow



RNASeq Tasks, Tools and File Formats



Galaxy Analyze Data Workflow Shared Data Visualization Cloud Help User

Published Pages | jeremy | Galaxy RNA-seq Analysis Exercise

RNA-seq Analysis Exercise

Galaxy provides the tools necessary to creating and executing a complete RNA-seq analysis pipeline. This exercise introduces these tools and guides you through a simple pipeline using some example datasets. Familiarity with Galaxy and the general concepts of RNA-seq analysis are useful for understanding this exercise. This exercise should take 1–2 hours. You can check your work by looking at the history and visualization at the bottom of this page, which contain the datasets for the completed exercise.

Input Datasets

Below are small samples of datasets from the Illumina BodyMap 2.0 project; specifically, the datasets are paired-end 50bp reads from adrenal and brain tissues. The sampled reads map mostly to a 500kb region of chromosome 19, positions 3–3.5 million (chr19:3000000:3500000).

RNA-seq data from adrenal tissue:

- Galaxy Dataset | adrenal_1.fastq
Forward RNA-seq reads from BodyMap 2.0 project, adrenal tissue, mapping to chr19:3000000:3500000
- and
- Galaxy Dataset | adrenal_2.fastq
Reverse RNA-seq reads from BodyMap 2.0 project, adrenal tissue, mapping to chr19:3000000:3500000

RNA-seq data from brain tissue:

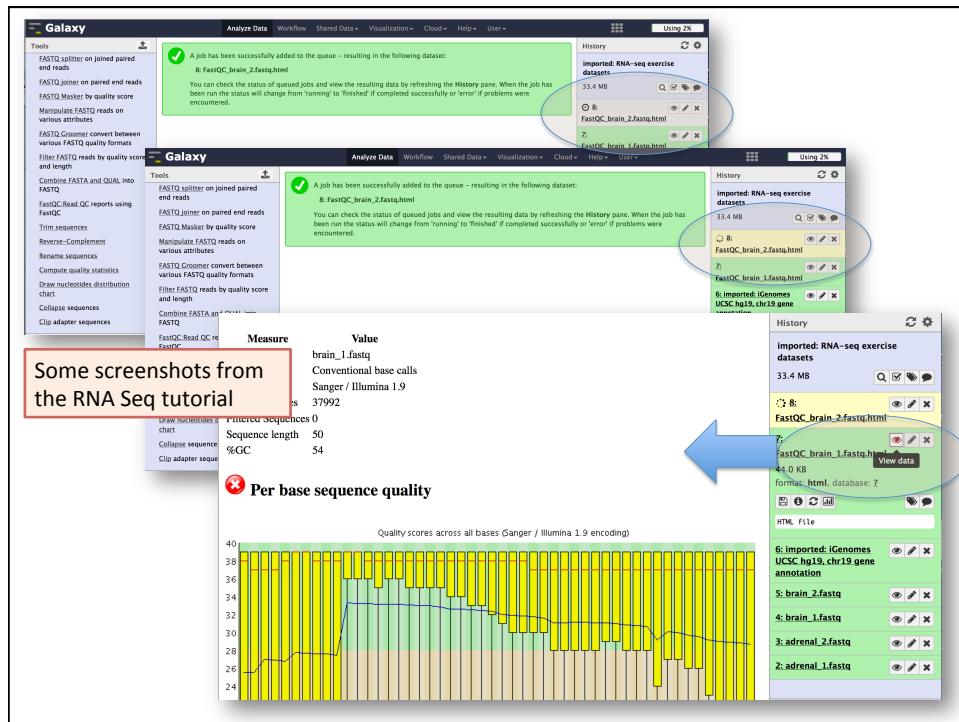
- Galaxy Dataset | brain_1.fastq
Forward RNA-seq reads from BodyMap 2.0 project, brain tissue, mapping to chr19:3000000:3500000
- and
- Revers...

You'll also need one additional dataset: a Waiting for usegalaxy.org...

There is a nice worked example of RNA seq in the Published Pages Section of Galaxy....

To see all Published Pages, click on Shared Data -> Published Pages

<https://usegalaxy.org/u/jeremy/p/galaxy-rna-seq-analysis-exercise>



Software

RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome

Bo Li¹ and Colin N Dewey^{1,2*}

* Corresponding author: Colin N Dewey cdewey@biostat.wisc.edu

1 Department of Computer Sciences, University of Wisconsin-Madison, Madison, WI, USA

2 Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI, USA

For all author emails, please [log on](#).

BMC Bioinformatics 2011, **12**:323 doi:10.1186/1471-2105-12-323

!

The TopHat/Cufflinks RNA Seq tools are commonly used..but they aren't the only ones out there.

It is possible to add new tools to Galaxy via the Galaxy ToolShed but this requires some programming experience.

14

Additional Notes on Galaxy

- You can try different parameters for alignment or variant calling and visualize the differences in the results
- Your history helps you “remember” the parameter settings when you publish your data

Many Galaxy Tutorials Available

- User support
 - <https://biostar.usegalaxy.org/>
- Tutorials
 - <https://usegalaxy.org/u/aun1/p/galaxy101>
 - <https://wiki.galaxyproject.org/Learn>