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

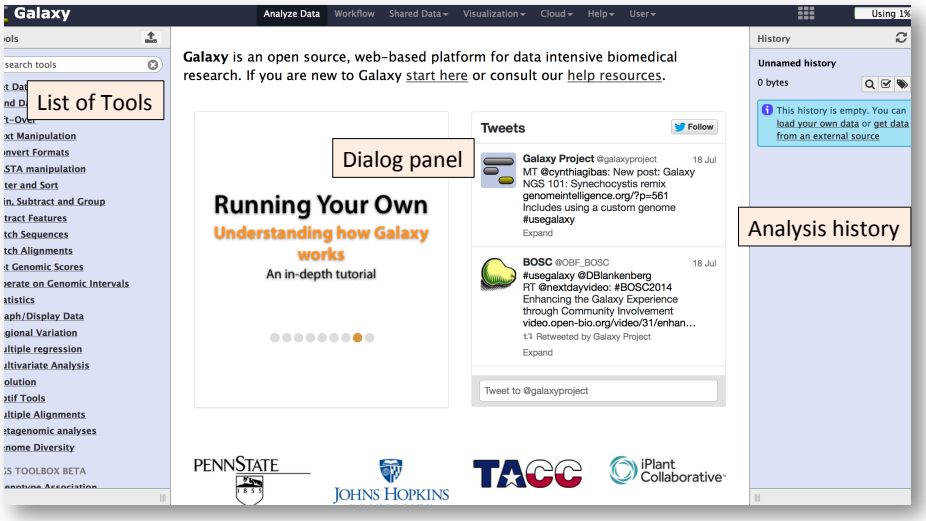
- You can do these exercises in Galaxy without an account
  - But without an account you can't save your work
- Many of the data files are LARGE and will take awhile to upload
  - Available from a public DropBox account

https://www.dropbox.com/sh/seqishl631f363x/AADoBskzYjPt_iJI21dxJqV8a

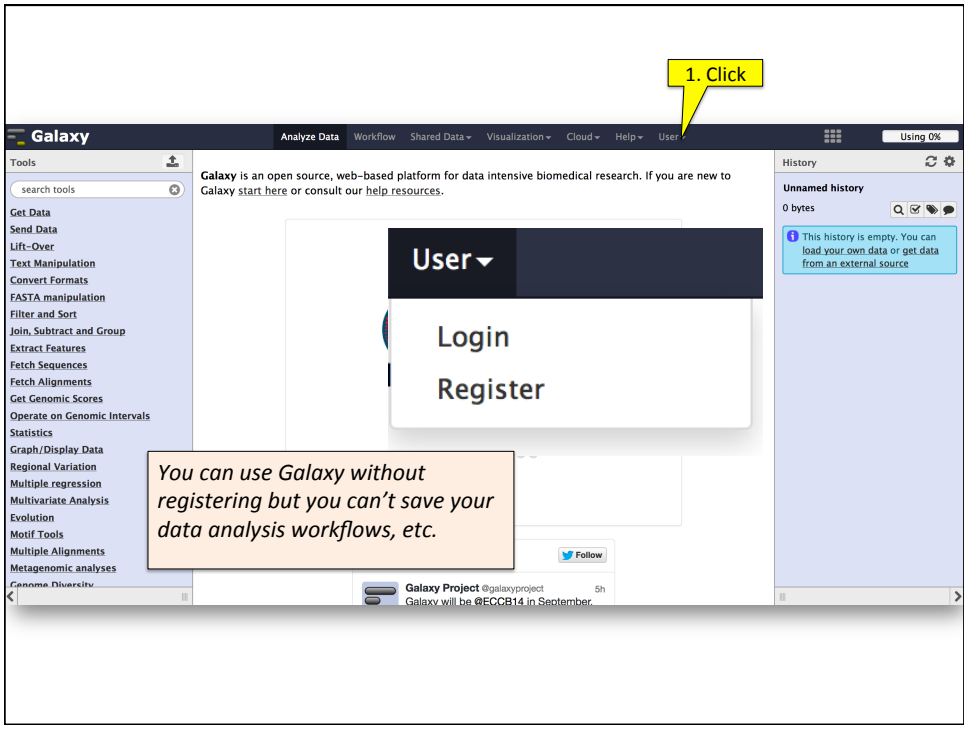
Galaxy in a Nutshell

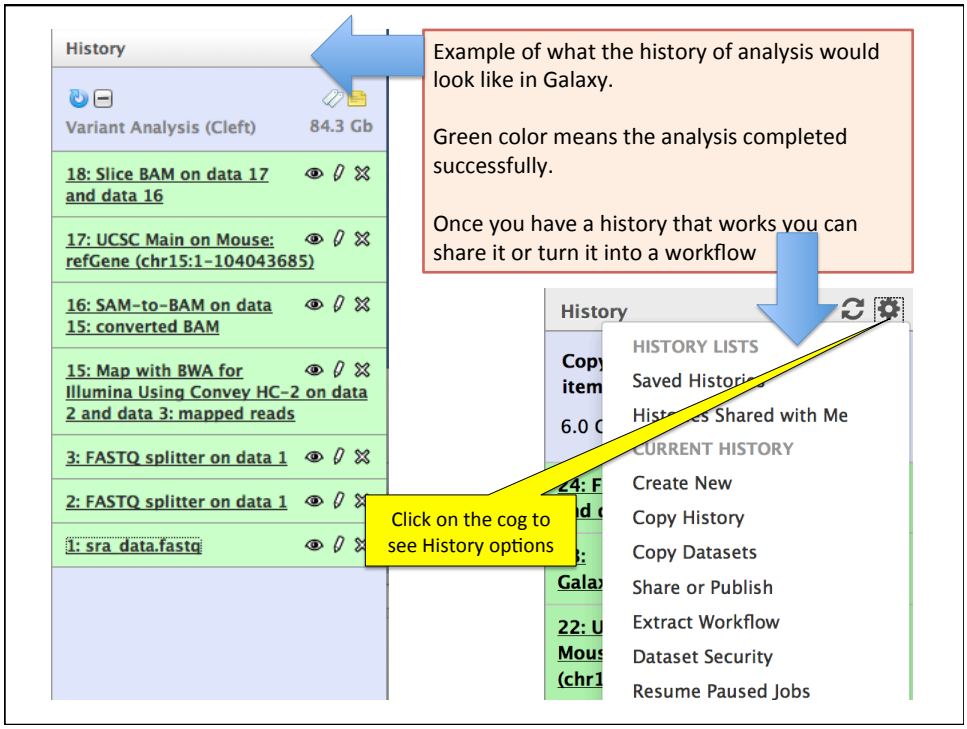
- Analyze
  - Build reusable analysis workflows for many types of data analysis needs
  - Interactive analysis
  - Reproducible analysis pipelines
  - Many analysis tools available … ready to use!
- Visualize
  - Send analysis results to standard genome browsers
- Publish and Share
  - Saved histories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

Galaxy



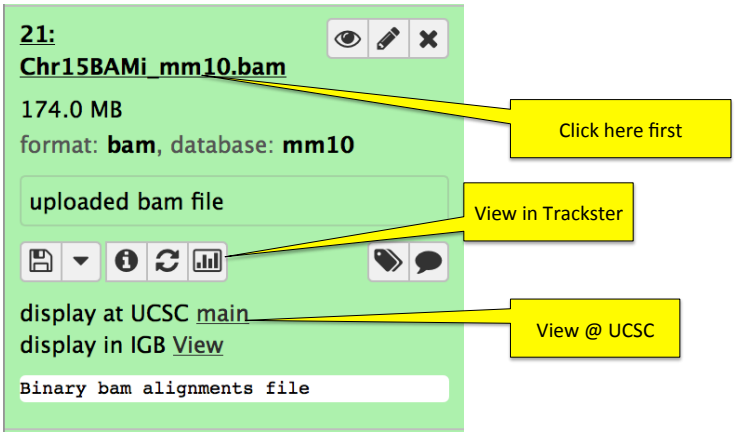
https://usegalaxy.org/


1. Click
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Cloud
Help
User
Using 0%
Tools
search tools
Get Data
Send Data
Lift-Over
Text Manipulation
Convert Formats
FASTA manipulation
Filter and Sort
Join, Subtract and Group
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Motif Tools
Multiple Alignments
Metagenomic analyses
Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy start here or consult our help resources.
User
Login
Register
You can use Galaxy without registering but you can't save your data analysis workflows, etc.
Follow
Galaxy Project @galaxyproject 5h
Galaxy will be @ECCB14 in September.
History
Unnamed history
0 bytes
This history is empty. You can load your own data or get data from an external source


History
Variant Analysis (Cleft)
84.3 Gb
18: Slice BAM on data 17 and data 16
17: UCSC Main on Mouse: refGene (chr15:1-104043685)
16: SAM-to-BAM on data 15: converted BAM
15: Map with BWA for Illumina Using Convey HC-2 on data 2 and data 3: mapped reads
3: FASTQ splitter on data 1
2: FASTQ splitter on data 1
1: sra_data.fastq
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
HISTORY LISTS
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)



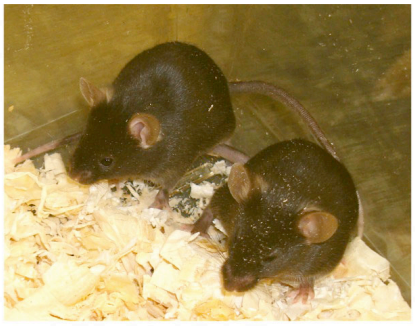
Fairfield et al. Genome Biology 2011, 12:R86
http://genomebiology.com/2011/12/9/R86

Genome Biology

METHOD Open Access

Mutation discovery in mice by whole exome sequencing

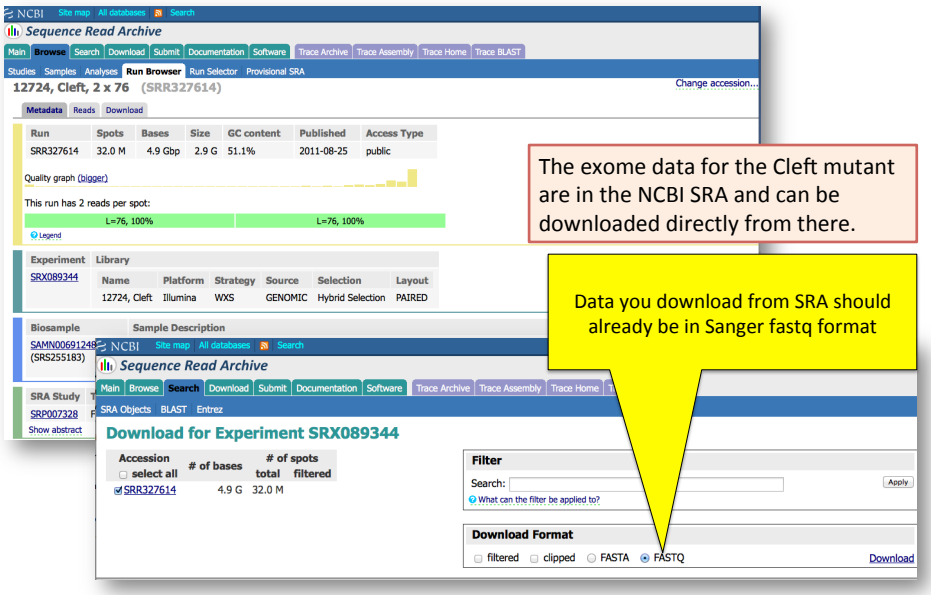
Heather Fairfield[1], Griffith J Gilbert[1], Mary Barter[1], Rebecca R Corrigan[2], Michelle Curtain[1], Yueming Ding[3], Mark D'Ascenzo[4], Daniel J Gerhardt[4], Chao He[5], Wenhui Huang[6], Todd Richmond[4], Lucy Rowe[1], Frank J Probst[2], David E Bergstrom[1], Stephen A Murray[1], Carol Bult[1], Joel Richardson[1], Benjamin T Kile[7], Ivo Gut[8], Jorg Hager[8], Snaevar Sigurdsson[9], Evan Maucell[9], Federica Di Palma[9], Kerstin Lindblad-Toh[9], Michael L Cunningham[10], Timothy C Cox[10], Monica J Justice[2], Mona S Spector[5], Scott W Lowe[5], Thomas Albert[4], Leah Rae Donahue[1], Jeffrey Jeddeloh[4], Jay Shendure[10] and Laura G Reinholdt[1*]



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

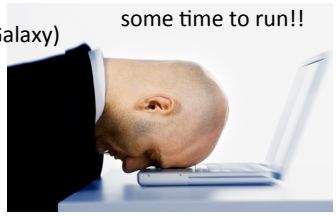


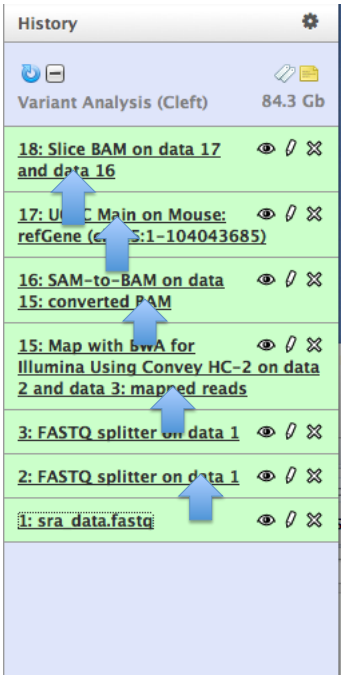
http://trace.ncbi.nlm.nih.gov/Traces/sra/?view=search_seq_name&exp=SRX089344&run=&m=search&s=seq

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!!



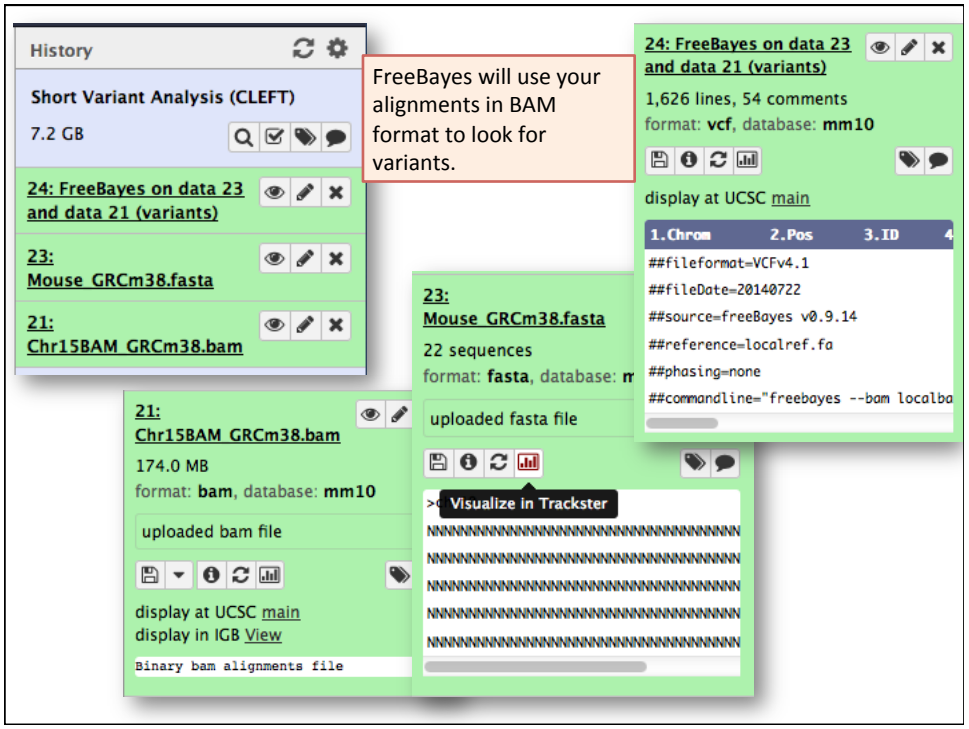


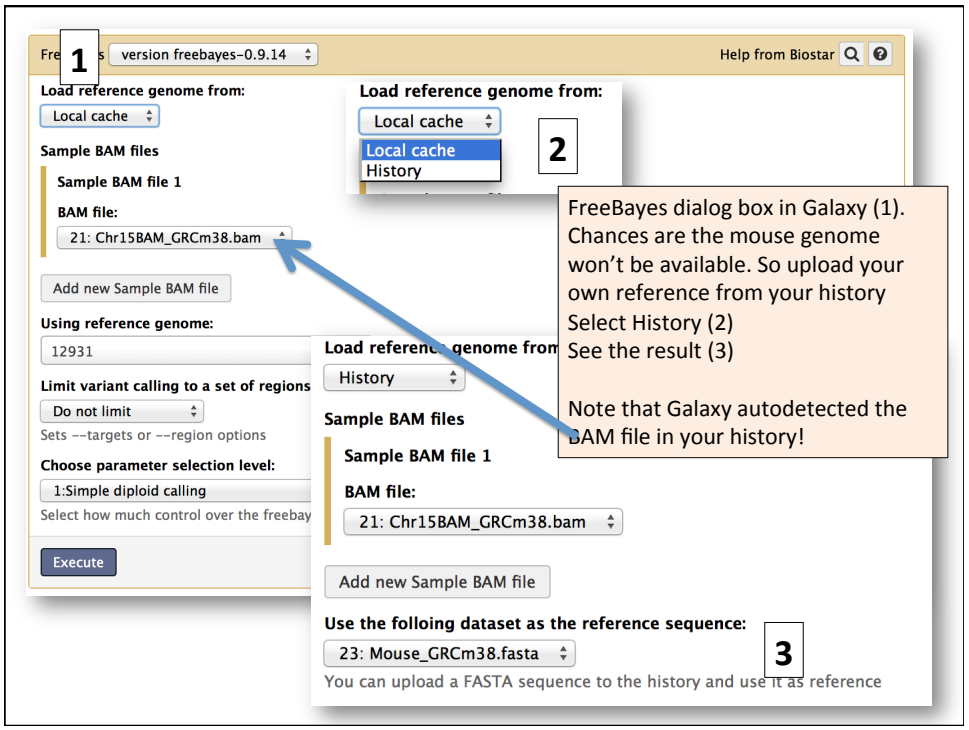
The out for one process in Galaxy becomes the input for another analysis tool.

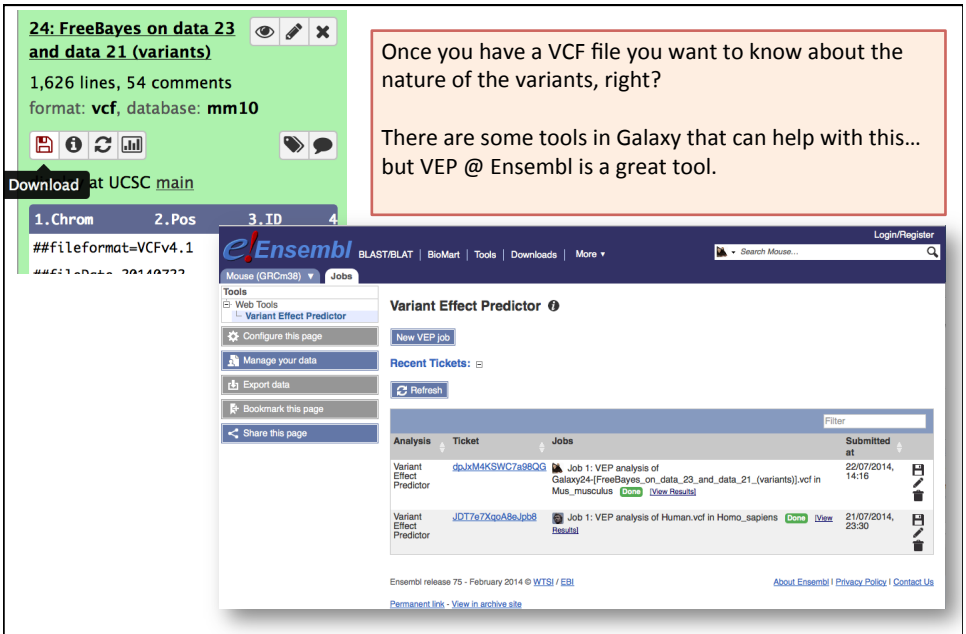
- Uploaded Cleft data from the SRA
- Split the paired end reads into separate files
- Mapped the reads to the mouse genome using BWA
- Converted the SAM file from BWA to BAM
- Extracted the subset of alignments to chromosome 15 from the whole genome BAM file

A Even Simpler Exome Analysis Workflow

- Get Data
  - **Get Data** -> (BAM file for chromosome 15)
- Get Data
  - **Get Data** -> (mouse assembly in fasta format)
- Run FreeBayes
  - **NGS: Variant Analysis** -> FreeBayes
    - For more information on FreeBayes: http://clavius.bc.edu/~erik/CSHL-advanced-sequencing/freebayes-tutorial.html
- Download FreeBayes output (VCF file) from Galaxy
- Submit VCF file to Variant Effect Predictor web site
  - http://useast.ensembl.org/info/docs/tools/vep/index.html


History
Short Variant Analysis (CLEFT)
7.2 GB
24: FreeBayes on data 23 and data 21 (variants)
23: Mouse_GRCm38.fasta
21: Chr15BAM_GRCm38.bam
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
display at UCSC main
1.Chrom 2.Pos 3.ID
##fileformat=VCFv4.1
##fileDate=20140722
##source=freeBayes v0.9.14
##reference=localref.fa
##phasing=none
##commandline="freebayes --bam localba
23: Mouse_GRCm38.fasta
22 sequences
format: fasta, database: m
uploaded fasta file
Visualize in Trackster
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
version freebayes-0.9.14
Help from Biostar
Load reference genome from:
Local cache
Sample BAM files
Sample BAM file 1
BAM file:
21: Chr15BAM_GRCm38.bam
Add new Sample BAM file
Using reference genome:
12931
Limit variant calling to a set of regions
Do not limit
Sets --targets or --region options
Choose parameter selection level:
1:Simple diploid calling
Select how much control over the freebay
Execute
Load reference genome from:
Local cache
Local cache
History
2
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!
Load reference genome from
History
Sample BAM files
Sample BAM file 1
BAM file:
21: Chr15BAM_GRCm38.bam
Add new Sample BAM file
Use the folloing dataset as the reference sequence:
23: Mouse_GRCm38.fasta
3
You can upload a FASTA sequence to the history and use it as reference



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

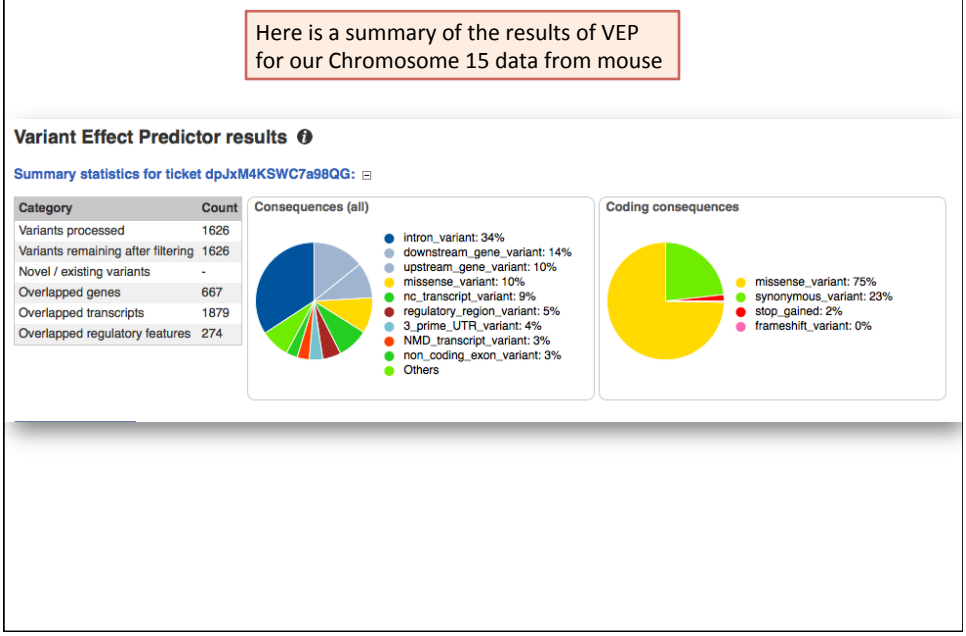
There are some tools in Galaxy that can help with this… but VEP @ Ensembl is a great tool.

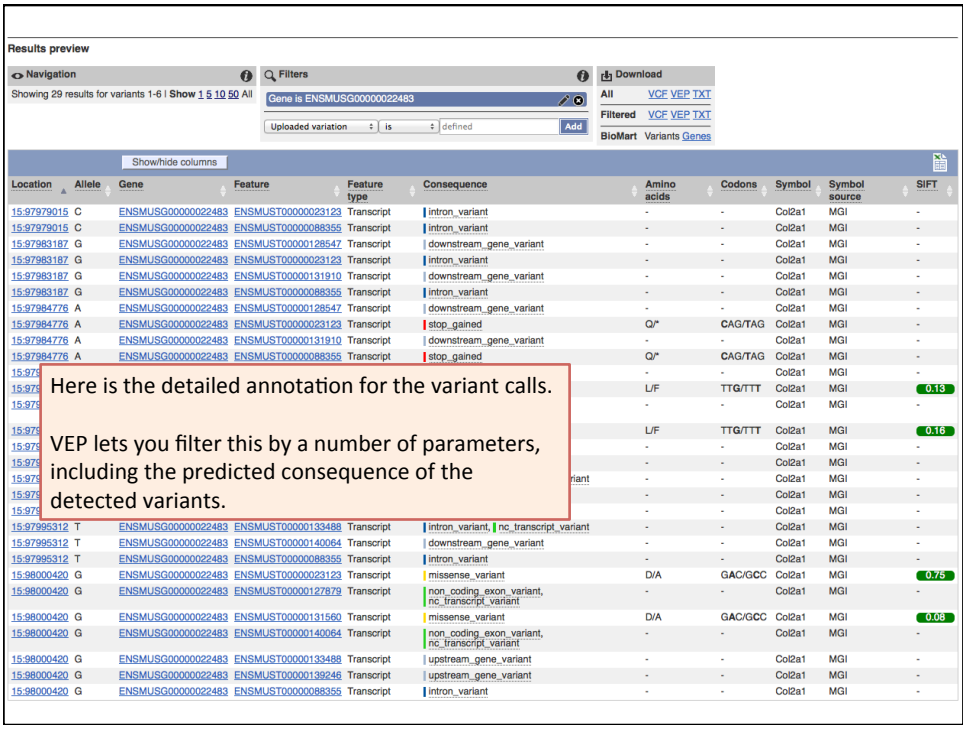
http://useast.ensembl.org/info/docs/tools/vep/index.html

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

mChr15_Cleft.vcf

Here is a summary of the results of VEP for our Chromosome 15 data from mouse





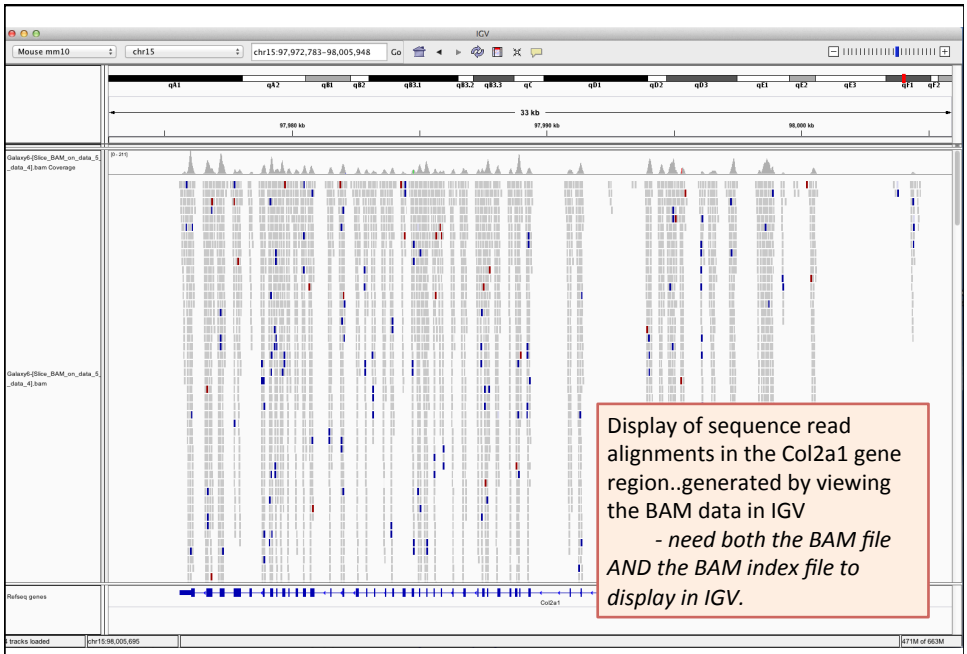
Here is the detailed annotation for the variant calls.

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 mis-sense 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.

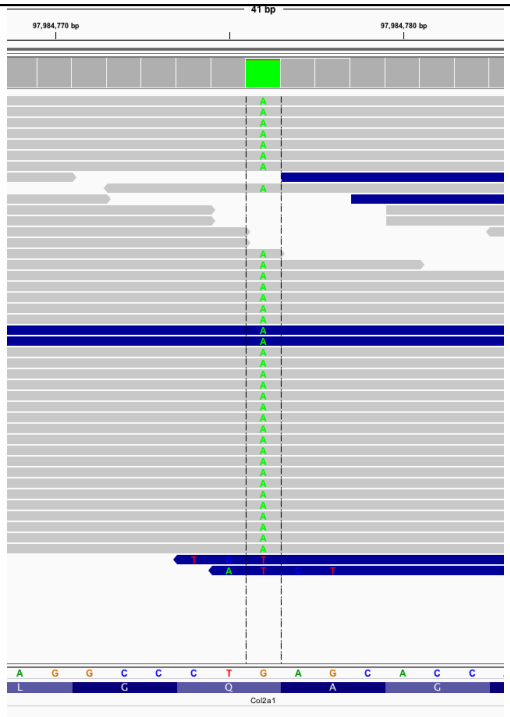
The location of the variants in this paper refer to NCBIm37…but I mapped the reads to GRCm38.p1. How do you map Build 37 coordinates to Build 38?



Display of sequence read alignments in the Col2a1 gene region..generated by viewing the BAM data in IGV

- need both the BAM file AND the BAM index file to display in IGV.

http://www.broadinstitute.org/igv/home



Zoomed in view using IGV for chr15:97984776

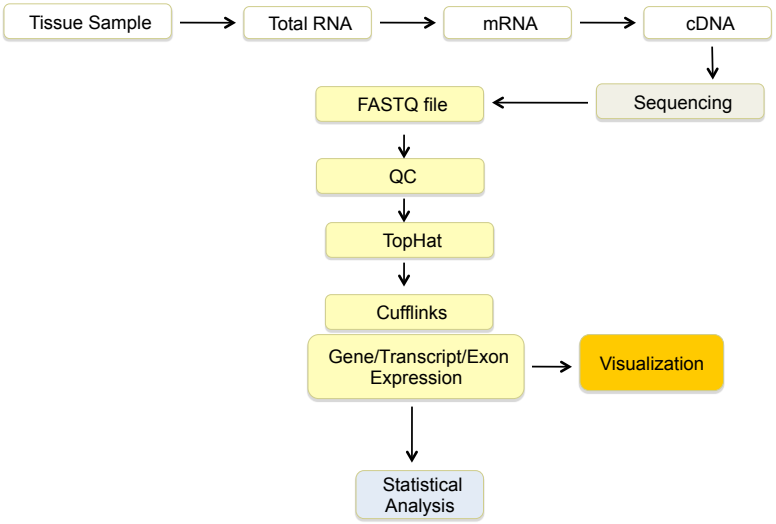
Your Turn

- 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

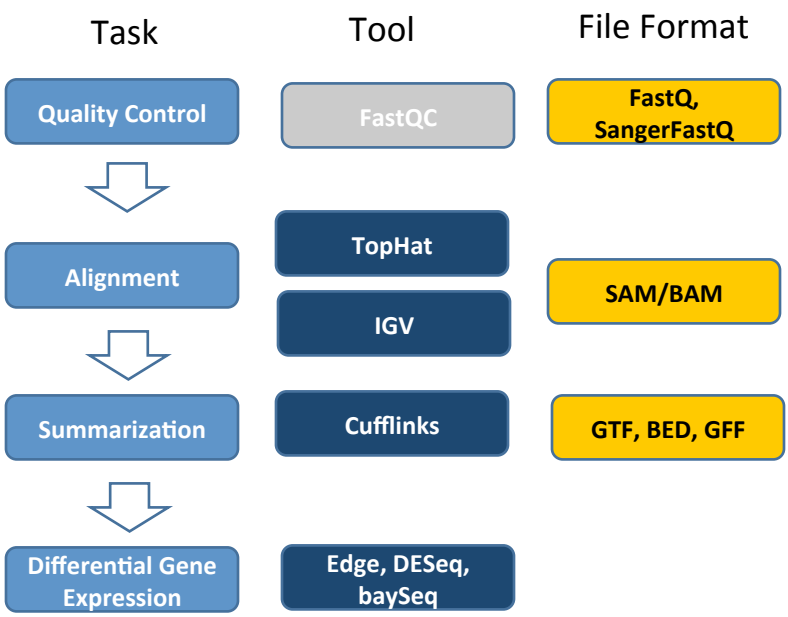
A Simple RNA Seq Workflow

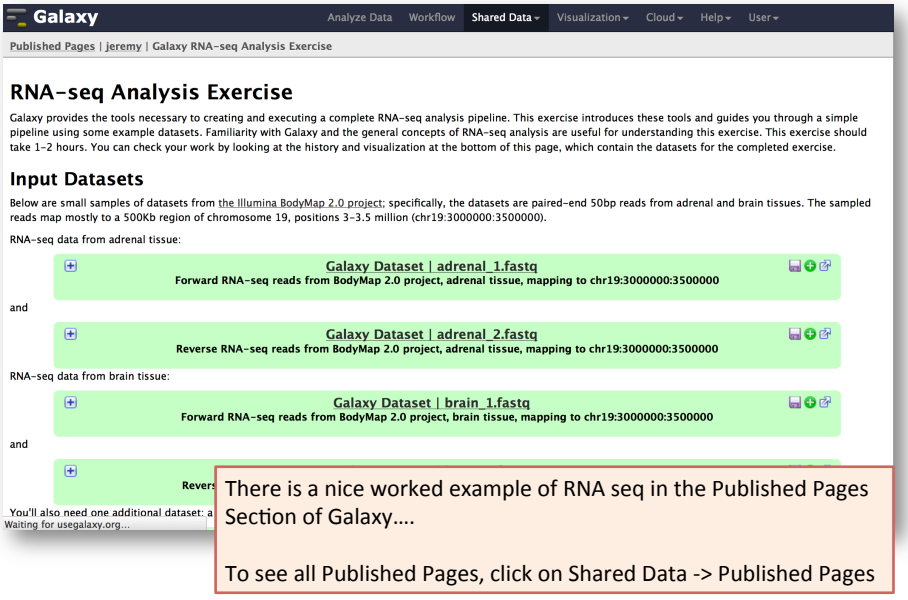
- Get sequence data
- Align sequence data to a reference
  - TopHat is commonly used because it is tuned to aligning transcripts to a genome (splice site aware)
  - http://ccb.jhu.edu/software/tophat/index.shtml
- Associate aligned reads to transcript models/ annotations
  - Cufflinks
- Quantitation 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

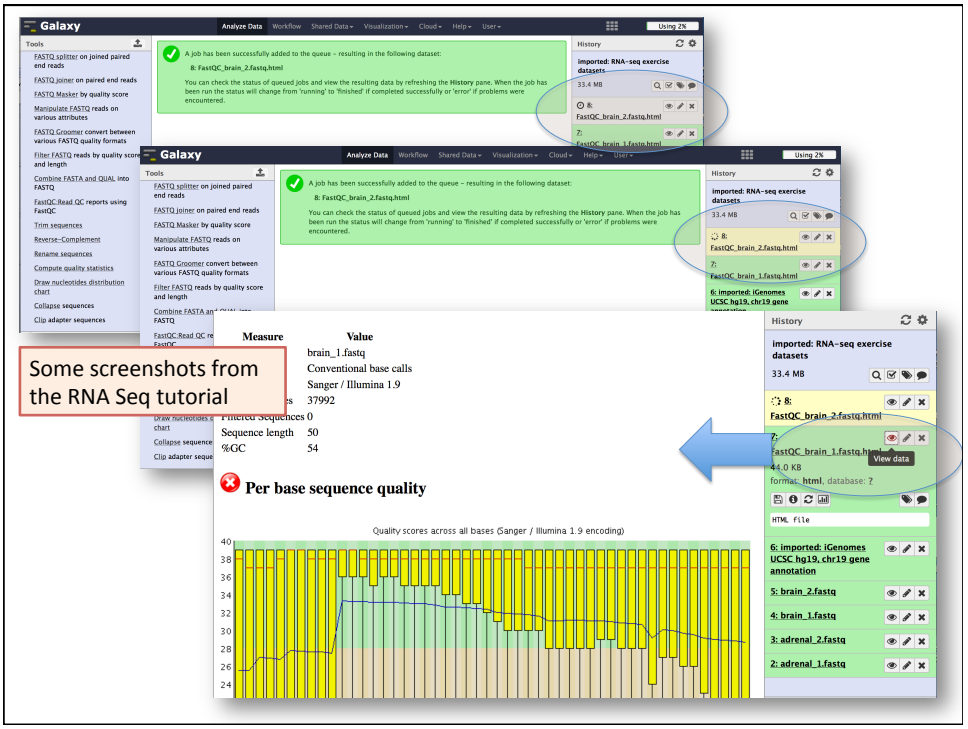


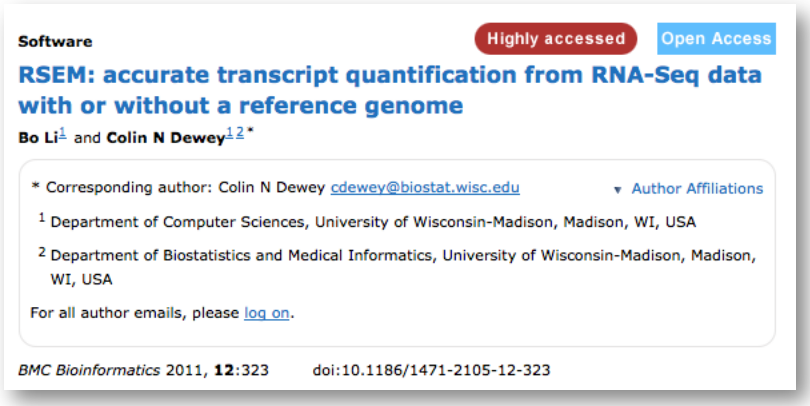


There is a nice worked example of RNA seq in the Published Pages Section of Galaxy….

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

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





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.

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