Galaxy Basics: DataSet Manipulation inside of Galaxy

IIHG Bioinformatics Course 2013
Ann Black-Ziegelbein
annblack@eng.uiowa.edu

Lab 1: **Using Galaxy to manipulate large data sets & creating a BED file for experimental design**

http://galaxy.hpc.uiowa.edu
logon: hawkid
password: hawkid password

Lab Goals: **By the end of Lab 1 you should:**

- Be familiar with the overall Galaxy web interface
- Understand how to access Shared Data Libraries
- Have basic understanding of the BED file format and how to manipulate it with Galaxy Tools
- Have an introduction to the UCSC Browser and how to download data from it into Galaxy
- Familiarity with manipulating data in Galaxy

Lab Steps:

1. Import/Load Data into Galaxy
2. Explore Existing Target and Bait Interval Files
3. Reverse Engineer Target Regions to Gene Names
4. Extract a target bed file from a list of gene names to use in analysis
5. Compare targeted capture design (bed) file against a whole exome bait interval file

Steps 1-2 : Performed Live in Lab

Summary of Steps 1-2:

You have now loaded two bed files into your Galaxy history from the IIHG Bioinformatics course Galaxy shared data library. You have inspected the file content and structure as well as the file attributes.

File	Description
otoscope_v4.bed	Target regions bed interval file for targeted capture of Deafness Specific Genes
SureSelect_50MB_exome.bed	Agilent Whole Exome Bait Target Intervals (From capture kit)

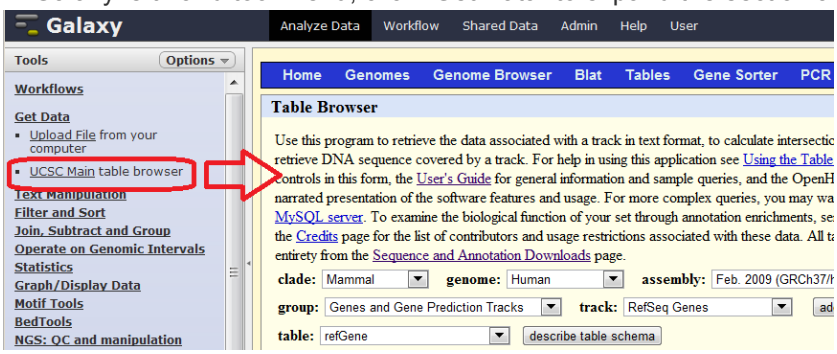
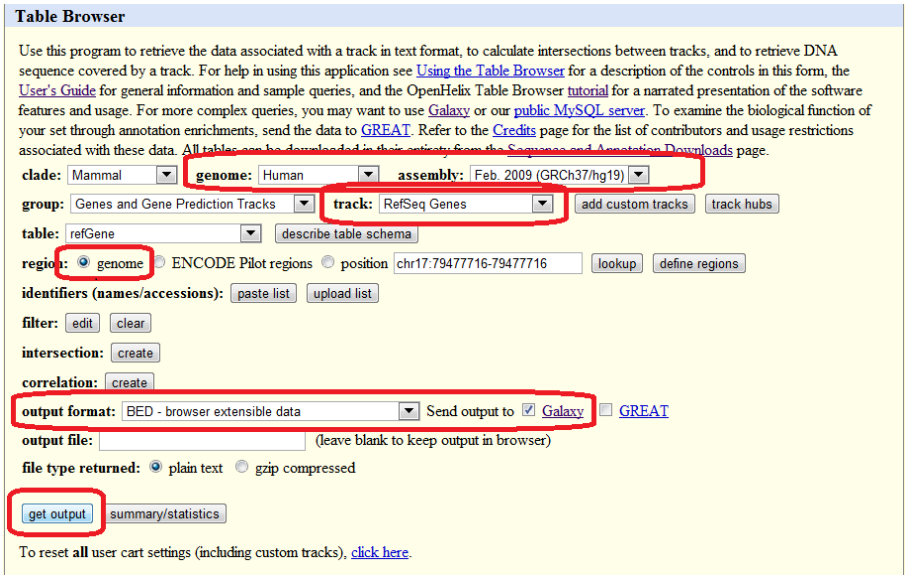
Please raise your hand for assistance at this time if you do not have the files in the table above in your Galaxy history.

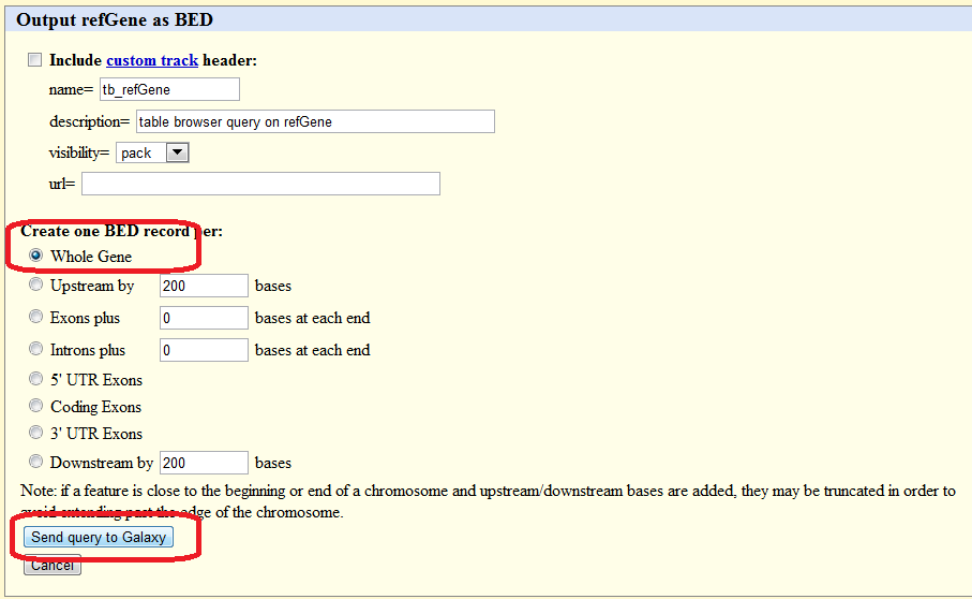
Step 3: Reverse Engineer Target Regions to Gene Names

For this portion of the lab, we will use the otoscope_v4.bed target region bed file to find out the list of gene names that are being targeted.

Step 3.A: Locate all the Gene IDs for the Human Genome

We are going to use the UCSC Table Browser to pull down a list of all HG19 gene ids and their corresponding chromosome locations. We will use this to annotate and compare the otoscope_v4 regions with.

3.A.1	In Galaxy left hand tool menu, click "Get Data" to expand the section and click "UCSC Main table browser". 
3.A.2	Next we are going to configure the data retrieval wizard to get the data we need for our genome of interest & from the right location. We want to get Gene and Gene Prediction Tracks for Human HG19 data. In the table browser make sure: • Genome = Human • Assembly = Feb 2009 (GRCh37/HG19) • Table = refGene • Group/Track = Gene and Gene Prediction Tracks/RefSeq Genes • Output Format = BED & Send Output = Galaxy Click "get output" once you have the UCSC wizard configured to retrieve the data we need. 
3.A.3	After clicking "get output" UCSC table browser allows you to further refine the data regions you want to download. We want whole genome information to get a list of all gene ids. • Make sure "Whole Genome" is selected • Click on "Send Query to Galaxy" button to begin downloading the data in BED format to your Galaxy History

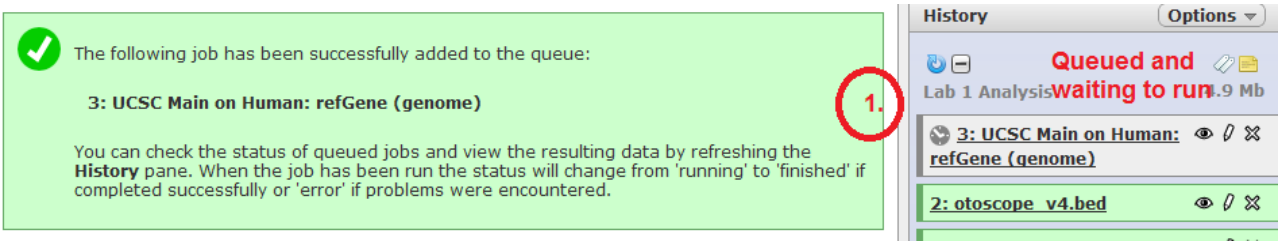


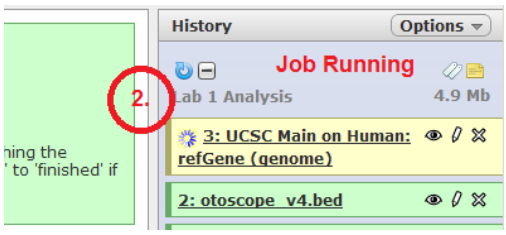
3.A.4

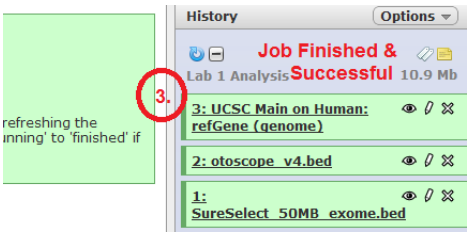
Galaxy has built in integration with UCSC Browser such that data can be automatically downloaded and accessible from your Galaxy current history without additional steps. After clicking "Send Query to Galaxy" you will see a new dataset/task created in your current history.

Tasks in galaxy are dispatched as jobs to the compute cluster. You will be able to check status of the job as it progresses through the following states:

- Queued = grey
- In Progress = yellow
- Green = successful
- Red = error occurred





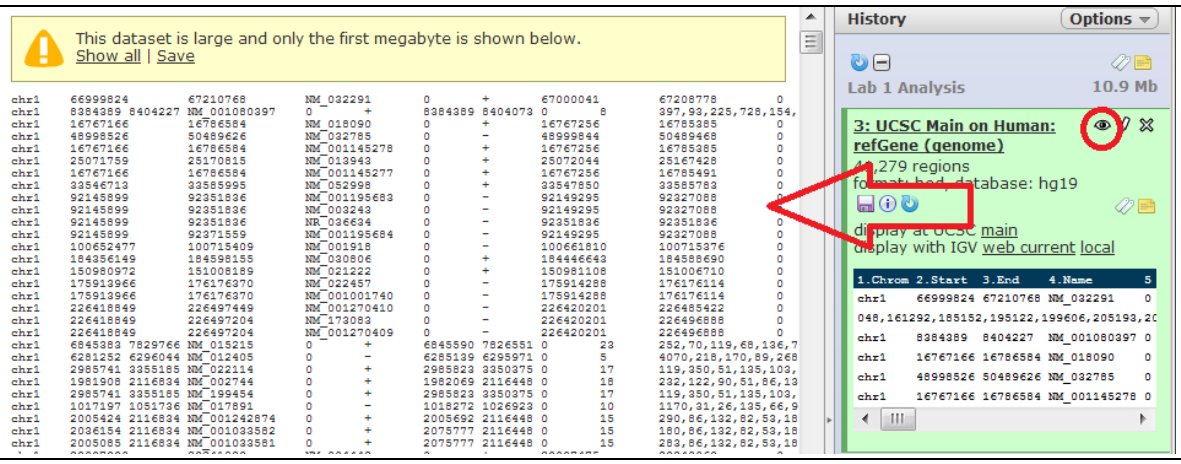


3.A.5

You can click on the name of the new UCSC Main dataset in your history to expand it for a quick snapshot of information about the dataset as well as column headers. "Poke the eye" to see the file's contents in the Galaxy work area portion of the browser. This file contains all the gene ids from the HG 19 Genome and their corresponding location (range) in the genome.

This follows the tab delimiter BED file format with optional columns:

{CHR} {START} {END} {GENE_ID} {SCORE} {STRAND} {CODE_START} {CODE_END} {EXON_FRAME} {EXON_COUNT} {EXON_STARTS} {EXON_ENDS}

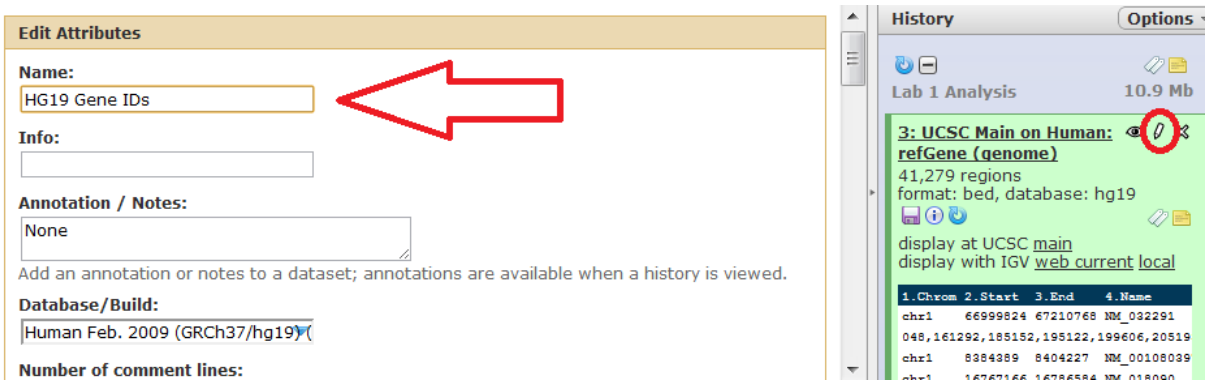


3.A.6

Galaxy names new datasets based on the action/tool used to generate the dataset. Let's rename this dataset to a more useful name.

BEST PRACTICE: When working with a large number of datasets in a history, this helps you remember what the dataset represents better than Galaxy default names.

- Click on the Pencil Icon to see the dataset attributes.
- Give the dataset a more meaningful name: "HG19 Gene IDs"
- Scroll down to the bottom of Attributes and click "Save"



3.A.7

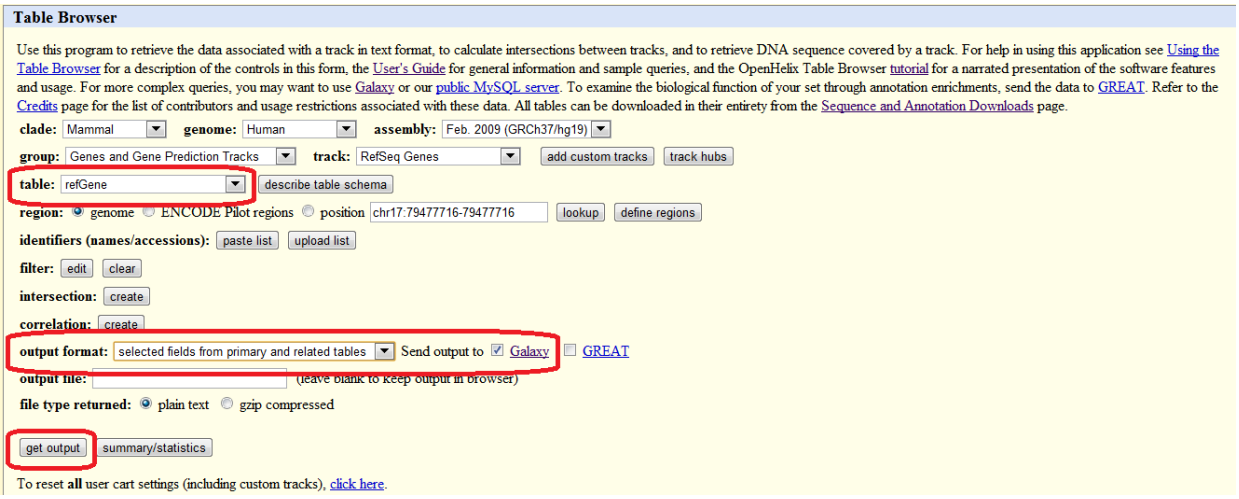
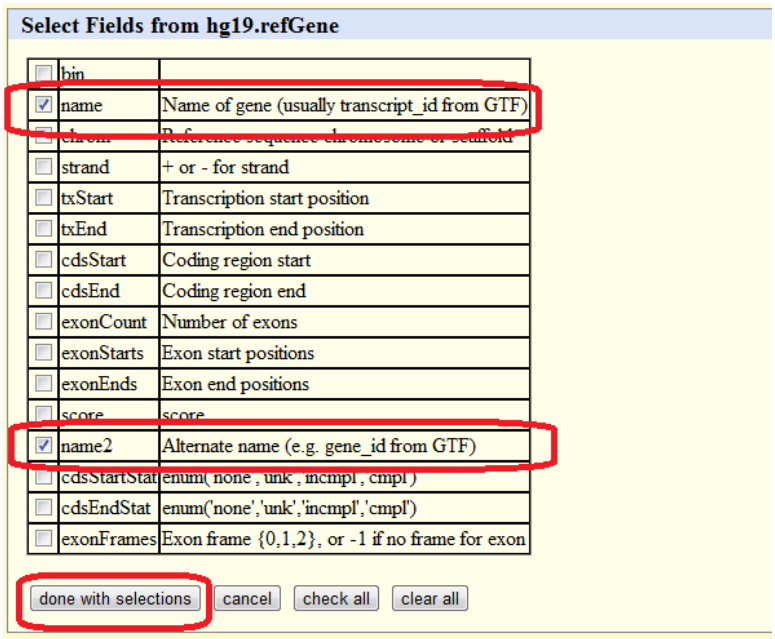
If you want to see a description of the data in this file, we can head back to the UCSC Browser and take a look at the schema of the table this information was pulled down from:

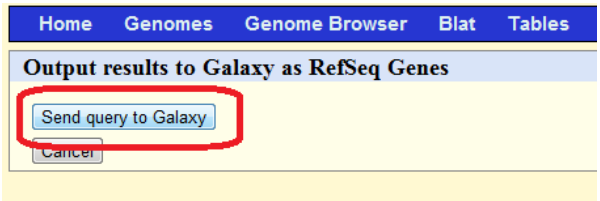
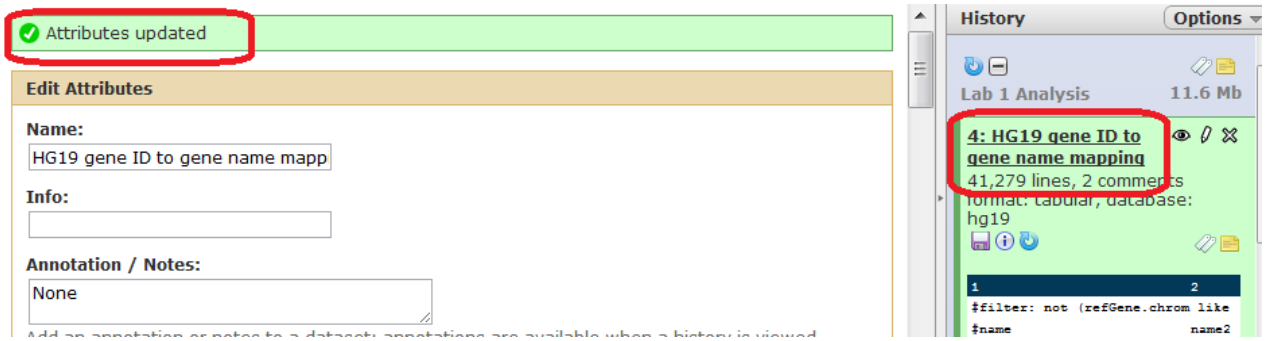
- From the left hand "Tools" Menu, expand "Get Data" -> "UCSC Browser Main"
- Make sure the wizard is filled out as in step 3.A.2
- Select the "describe table schema" button next to the table = refGene
- You can review the table content and column descriptions.
- Note, the schema does not show the data in the same order as you see it in the BED file format.

Step 3.B: Locate Gene Names to Gene ID Mappings

During step 3.A.5, you may have noticed that the genes are identified with the GTF identifier. However, we would like to see the corresponding gene name (more human readable/understandable). If you inspected

what data is available from the refGene table in UCSC browser in step 3.A.7, there is a “Name2” that is stored in addition to id. We are going to use the USCS Table Browser to pull down a list of all HG19 gene ids and their corresponding names to annotate the OtoSCOPE bed file with.

3.B.1	In Galaxy left hand tool menu, click “Get Data” to expand the section and click “UCSC Main table browser”.
3.B.2	We want to get Gene information about HG19 genome again, but this time we need specific fields. In the table browser make sure: • Genome = Human • Assembly = Feb 2009 (GRCh37/HG19) • Table = refGene • Group/Track = Gene and Gene Prediction Tracks/RefSeq Genes • **Output Format = “Selected fields from primary and related tables”** • Send Output = Galaxy Click “get output” once you have the UCSC wizard configured to retrieve the data we need. 
3.B.3	After “get output” is clicked, you will have the opportunity to select which data fields you want to download. We need the following fields: • Name • Name2 • Click “done with selections” 

3.B.4	Now submit the job such that data can be downloaded to your Galaxy history: • Click “Send query to Galaxy” 
3.B.5	After the job successfully runs (grey = Queued, yellow= In Progress, green = Successful), we can once again rename the dataset (See step 3.A.6 for photos). • Click the “pencil” icon to show dataset attributes. • Rename the dataset to something more remember-able like “HG19 Gene IDs to Gene Name mapping” • Click Save button at bottom 
3.B.6	You can inspect the dataset in the Galaxy browser by clicking on the “eye” icon (“Poke it in the eye”). You will see it is a simple two column list of gene ids mapped to gene names.

Step 3.C: Join/Annotate OtoSCOPE regions with gene IDs.

You should now have all the input data you need in your local history:

- **The otoscope_v4.bed file which depicts the target regions of interest for deafness related genes (chr, start, end)**
- **A hg19 interval bed file of all gene regions with corresponding gene id**
- **A tab delimited file listing each gene id and its corresponding gene name**

Now we will work to annotate the otoscope_v4.bed file with corresponding gene ids. We will use Galaxy to find from the interval locations in the OtoScope Bed file, the gene id that it belongs to and then from the gene *ID* we will find the gene *Name*.

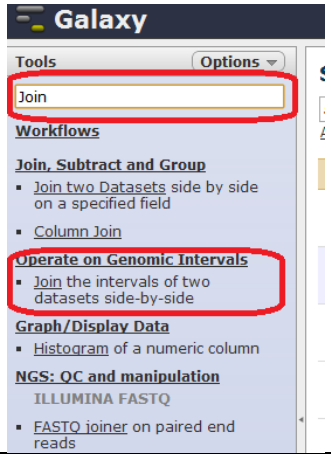
3.C.1

Galaxy has a large variety of tools that can help you in manipulating datasets. We are looking for a tool that will allow us to join two files together based on the chr, start, and end regions.

TIP: Use the Tool search to help find tools (Options -> Show Tool Search)

The Galaxy tool search field is an easy way to sift through the volume of tools exposed in Galaxy. You can show/hide the tool search field through the Tools -> Options menu.

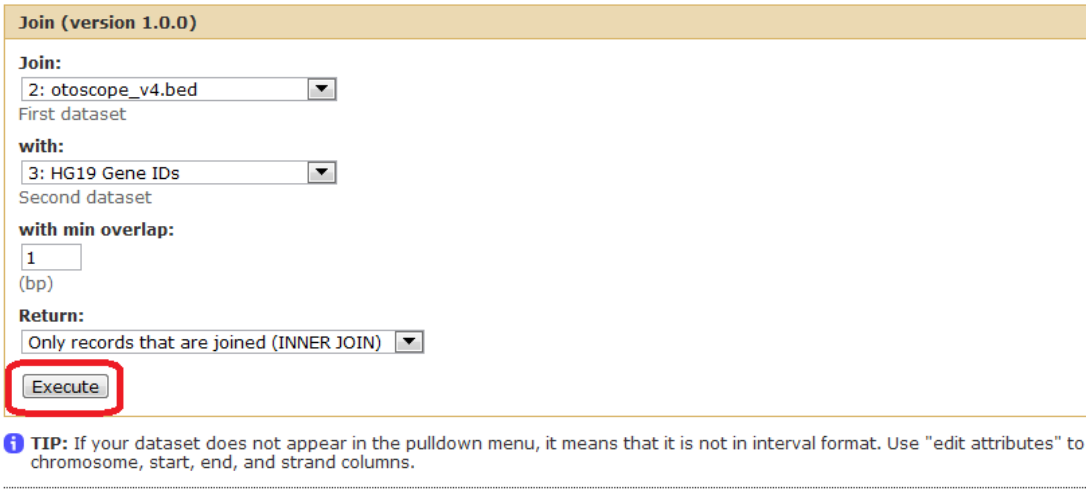
With the tool search menu shown, search for tools that can **join** datasets:



Underneath "Operate on Genomic Intervals" tool category, we find a tool named "Join" and the description is exactly what we need. Click "Join" under "Operate on Genomic Intervals" to take this action.

3.C.2

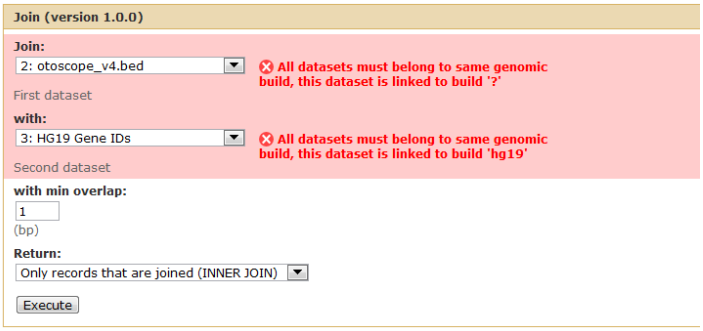
We want to find regions in our otoscope_v4.bed file that overlap with regions from the hg19 gene id regions and join them together in a new interval file:



1. Configure the tool with the appropriate datasets
2. Return "Only records that are joined"
3. click "Execute"

3.C.3

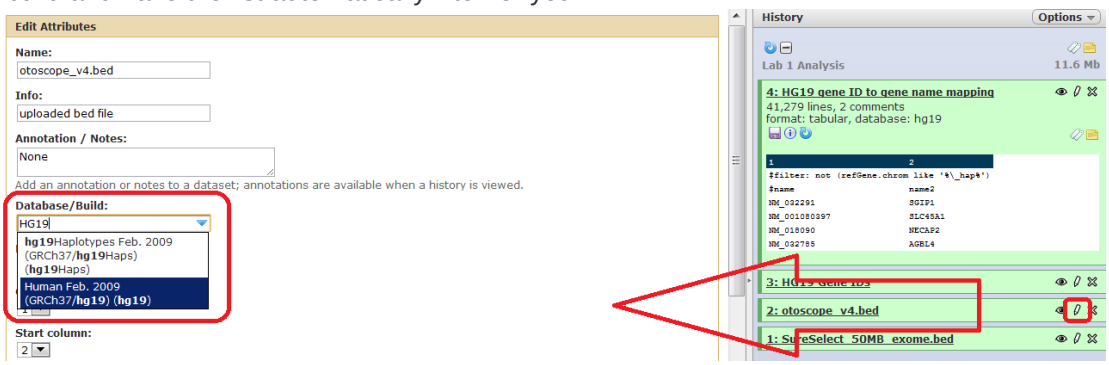
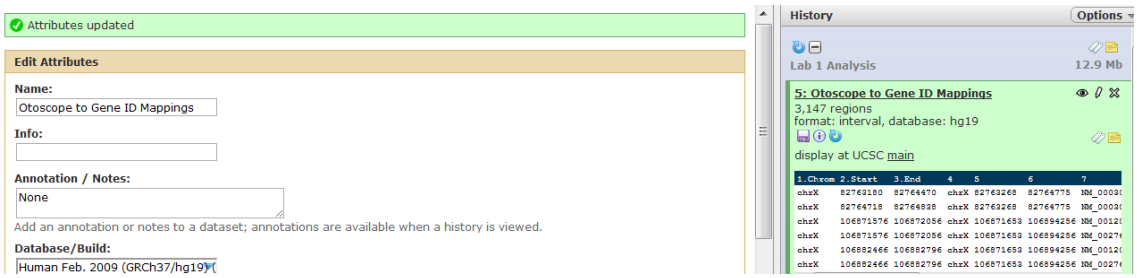
Uh – Oh! Error!



The error messages are telling us that the two files do not appear to be of the same genomic build. The tool has intelligence to make sure you are joining interval files that make sense.
The genomic build information is stored as dataset attributes

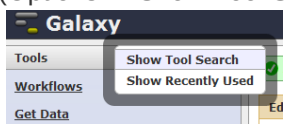
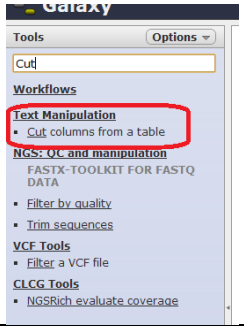
3.C.4

Lets correct this problem and run again. Look at the error messages, the otoscope_v4.bed file has a build of '?' but we know it is human HG19 intervals. To correct it:

	1. Select the “Pencil” icon next of the otoscope_v4.bed dataset. 2. Change the genome build drop down to HG19 (NOT HG19 hap) 3. Click “Save” **TIP:** When setting the genomic build via the drop down selector, you can start to type the reference genome build and have the list automatically filter for you. 
3.C.5	Now re-do step 3.C.4. You should see a new dataset appear and change status from queued (grey) -> in progress (yellow) -> completed (green).
3.C.6	Inspect the results by clicking the “eye” icon on the new dataset to “Poke it in the eye” and review the results. You should see that for each row in the original OtoScope bed file, it has been joined with a row from the HG19 Gene IDs file that has a matching overlapping region. This gives us Otoscope region -> gene ID mapping.
3.C.7	Rename the datset to something more useful like “ Otoscope to Gene ID Mappings” 

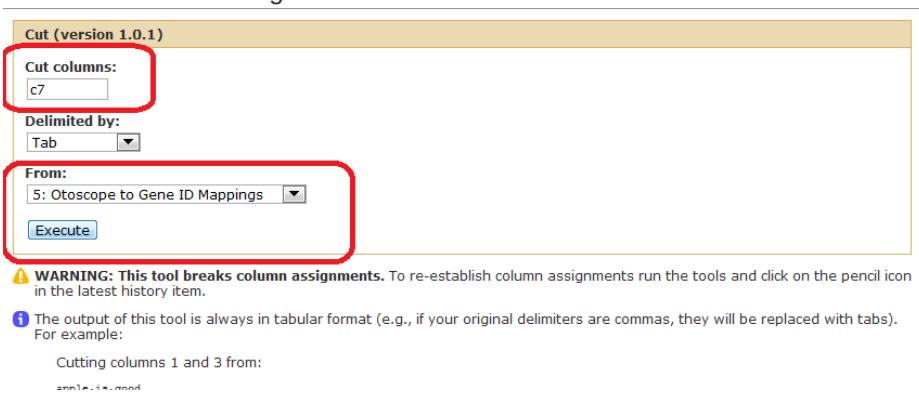
Step 3.D: Make a List of All Unique Gene IDs associated with OtoSCOPE regions

Now we have a file that shows all regions from the OtoSCOPE design file and their corresponding gene id(s). Next we will just pull out the list of unique gene ids from this file.

3.D.1	**TIP:** Use the Tool search to help find tools (Options -> Show Tool Search)  Using the Galaxy tool search field, let’s find a field that allows us to cut out a specific column from a delimited file.  For this lab, we just want column 7 which holds the gene id that is associated with the OtoSCOPE region. The search for “Cut” shows a tool under “Text Manipulation” that has a short description that matches what we need it to do. Select “Text Manipulation” -> “Cut”.

3.D.2

Expand the dataset in your current history by clicking on “Otoscope to Gene ID Mappings” dataset name. Part of the summary information shown is the first few lines of data headered with the column name/number. Find the column number that holds the gene id data. This will be the column to cut from the file. (This should be column 7).



For the “Cut” tool execution parameters:

1. Enter in the column number that holds the gene id (found from the dataset summary)
2. Select the “Otoscope to Gene ID Mappings” as the dataset to cut from
3. Click “Execute”

3.D.3

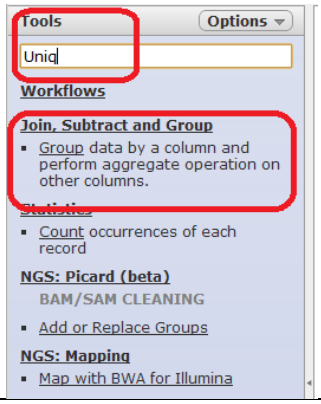
The “cut” job and new dataset should transition from queued -> in progress -> completed/successful. View the dataset’s content in the Galaxy browser by “Poking the eye” (click on the “eye” icon). Notice that there are duplicate entries in the list.

3.D.4

Rename the dataset to something more meaningful: “All Otoscope Gene IDs”

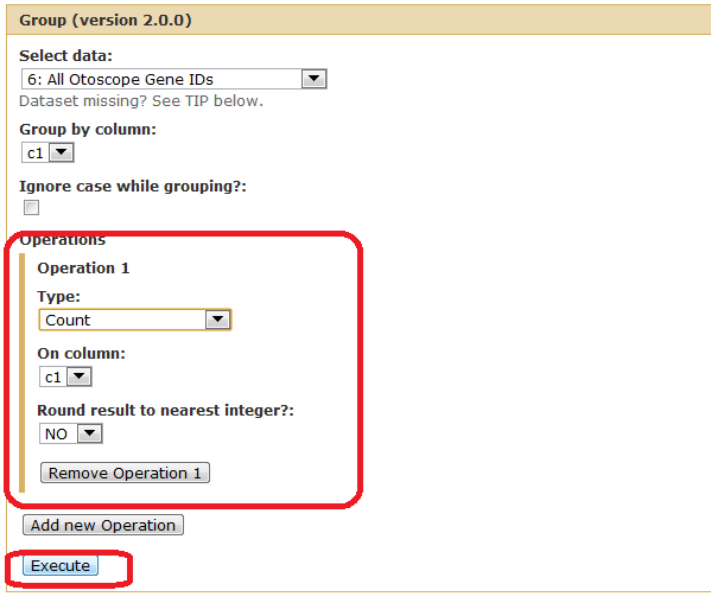
3.D.5

Now lets consolidate the list of Gene IDs to a unique list and get rid of all the duplicates. On a unix command prompt this would be done with commands like ‘uniq’. Use the tool search field to find tools that might help us make a unique list. By looking through the tool names, “Group” (under Join, Subtract and Group) looks like it will do what we need. We can group by gene id to build a consolidated list.



3.D.6

Click on the “Group” tool under “Join, Subtract and Group” tool category.



The data you want to group on is from the “All Otoscope Gene IDs” list that was created in step 3.C.2. You can “Add new Operation” to generate statistics/aggregations when grouping.

1. Set Select Data to “All Otoscope Gene IDs”
2. Group by c1 (the column with the gene ids)
3. “Add new Operation”
4. Type = “Count” On Column = “c1”
5. Click “Execute” to launch the job.

This will consolidate the list of gene ids to only the unique subset while at the same time count how many regions in the OtoSCOPE Bed file overlapped with the gene.

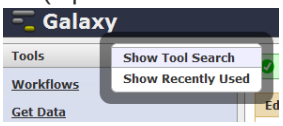
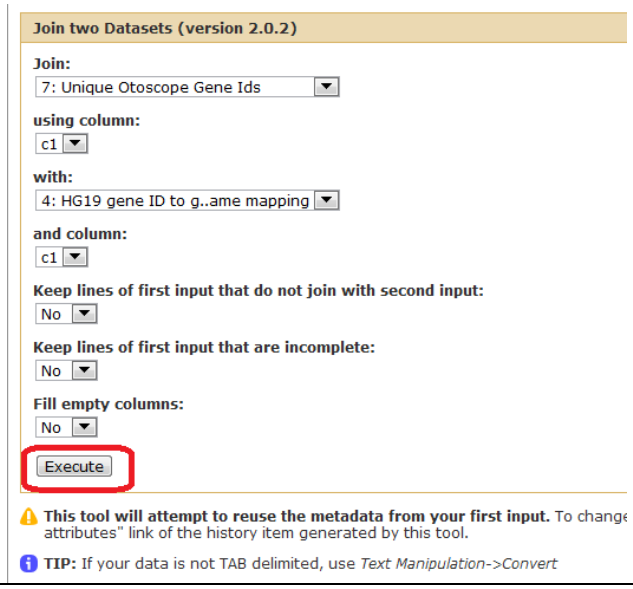
3.D.7

After the job completes, rename the dataset with a more meaningful name : “Unique Otoscope Gene IDs” and view

	the dataset's contents in Galaxy ("Poke it in the eye")

Step 3.E: Make a List of All Unique Gene NAMEs associated with OtoSCOPE regions

Now you have a file that contains the list of unique gene ids targeted by OtoSCOPE for genetic hearing loss. Next lets map them to their common names.

3.E.1	**TIP:** Use the Tool search to help find tools (Options -> Show Tool Search)  In Step 3.B you downloaded from UCSC Browser a tabular file that mapped the gene id to its common name. We need to associate these gene names to the OtoSCOPE gene ids we listed in step 3.D. Use the tool search to list all JOIN tools in Galaxy once again:  This time, we are not working with interval files, but with tabular data. From the list of "Join" related tools found in our search, "Join two Datasets" under "Join, Subtract, and Group" provides a good match.
3.E.2	**Click on "Join" tool underneith the "Join, Subtract, and Group" tool category.**  **Join the list of "Unique Otoscope Gene Ids" with the "HG19 gene ids to gene name mapping" dataset downloaded from the UCSC Browser in step 3.B. When joining, you select the column from each dataset that must match for the join to occur.** **Click "Execute" once you have configured the tool parameters like the screen catpure.**
3.E.3	**Once the job/task completes, rename the dataset to something more meaningful: "Otoscope Unique Gene IDs to Gene Name Mapping". View the dataset contents in the Galaxy browser by clicking the "eye" icon of the dataset.**
3.E.4	**When reviewing the list of Otoscope gene ids to gene names, notice that there are duplicate gene names. Build a unique gene name list by performing steps 3.D.3 to 3.D.6 but on the "Otoscope Unique Gene IDs to Gene Name Mapping" dataset. This time do not add a count grouping operation such that you end up with a file with only one column of unique gene names.**
3.E.5	**Once you have the dataset of unique gene names completed, rename it: "Unique Otoscope Gene Names"**

Step 4: Building a target region interval file from a set of Gene Names

Reverse engineering otoscope_v4.bed file to a set of unique gene names gave you an opportunity to explore various tools in Galaxy and manipulate data files. However, for experimental design, you may want to create a target regions file (similar to otoscope_v4.bed) but for specific biological areas of interest to your specific research. In Step 4 of Lab 1, we will now learn how to create a target bed file from the set of gene names we generated in Step 3.

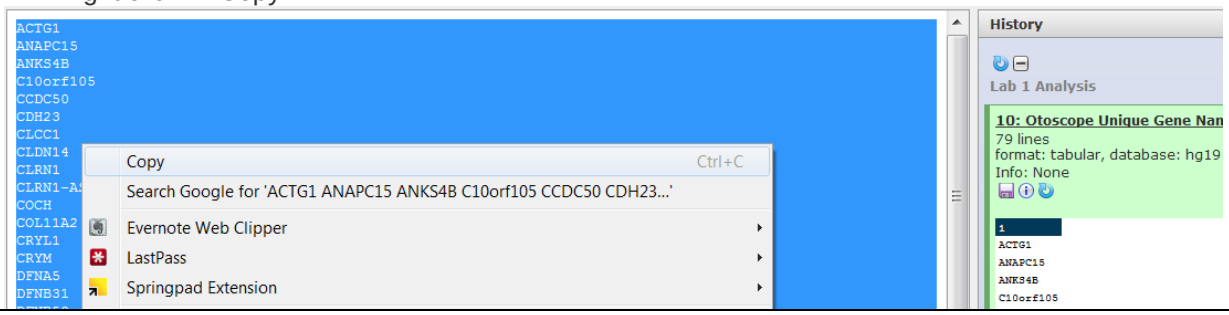
Step 4.A: Back to the UCSC Browser

We are going to use the UCSC Table Browser to pull down target regions that match specific criteria that we have.

4.A.1

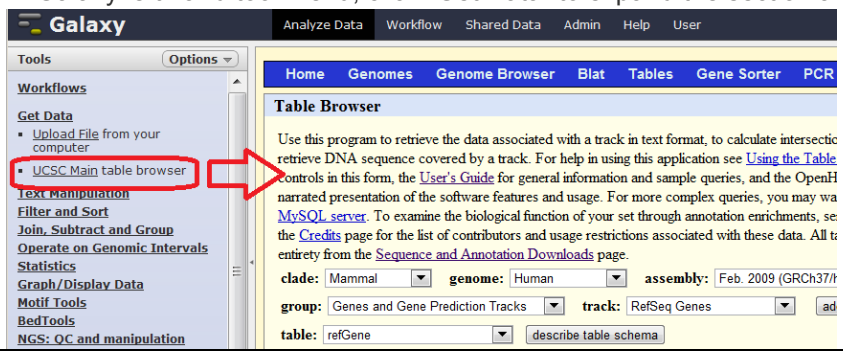
First lets copy into our computers memory the list of gene names we generated from Lab 1 Step 3:

1. Find "Otoscope unique gene names" dataset in your current history
2. Click the "eye" icon to view file contents in the Galaxy browser
3. Select all the gene names in the browser window
4. Right click -> Copy



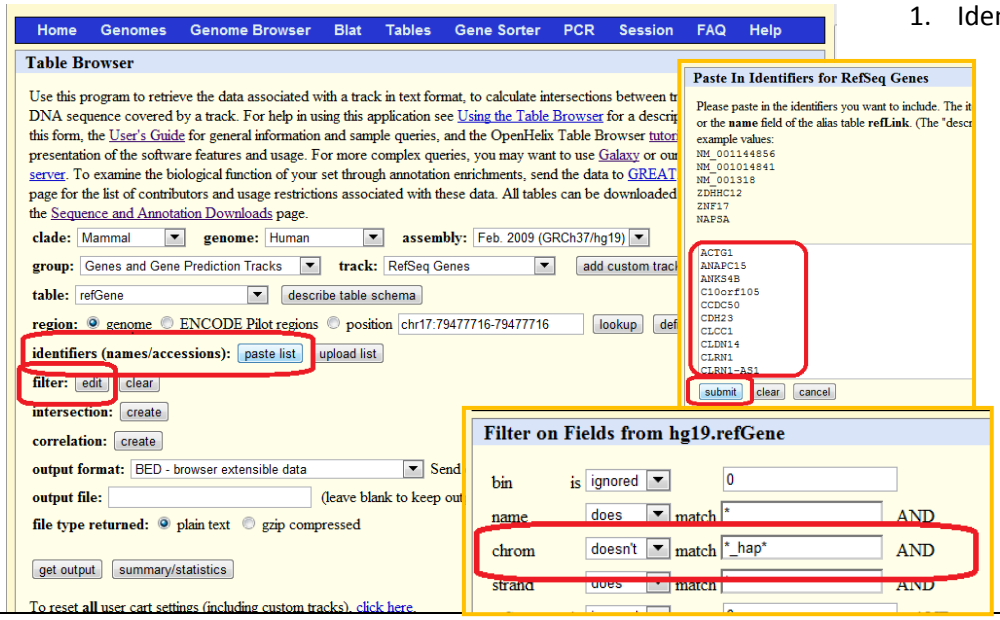
4.A.2

In Galaxy left hand tool menu, click "Get Data" to expand the section and click "UCSC Main table browser".



4.A.3

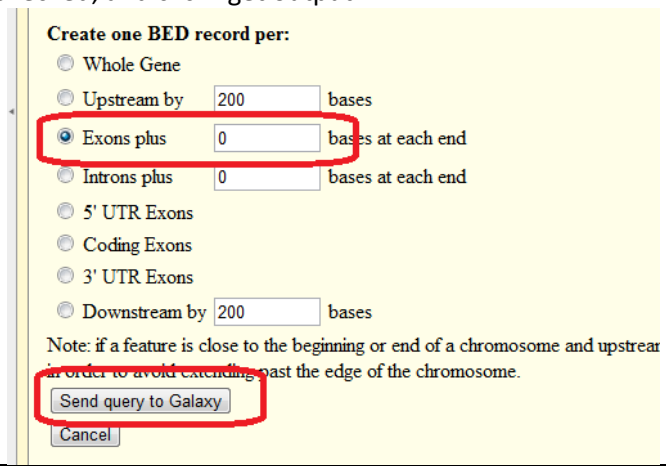
This time, we don't want to get regions all HG19 genes, only a subset.



1. Identify the areas of interest by name. Click on "identifiers" -> "paste list"
2. Paste in the list of gene names from step 4.A.1
3. Click "Submit"
4. Click "filter"
5. Let's filter out all hap chroms.

Change the chrom field to "doesn't" match "*_hap*".

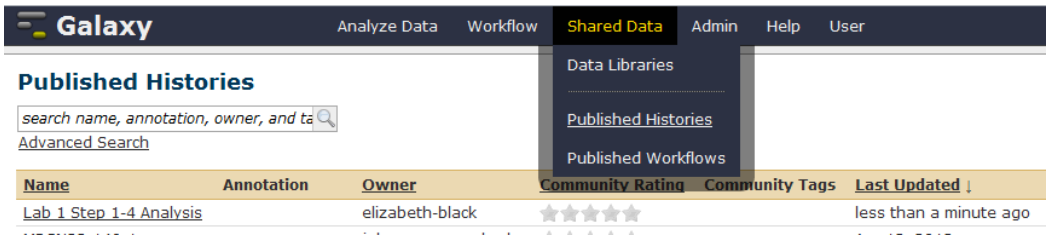
6. Click "Submit"

4.A.4	Now that you have entered in filter and match criteria, make sure output format is “BED”, Send to Galaxy is checked, and click “get output”  Lets further limit the data being retrieved to only “Exons” by selecting “Exons Plus” radio button. Click “Send query to Galaxy”.
4.A.5	Once the job is completed, you can inspect its contents by clicking the “eye” icon of the dataset. You have now created a target region file from specific criteria. UCSC Table Browser supports a variety of filtering and critera matching configuration to pull data, more than what we had time to use in this simple lab. Please ask questions about your specific needs.

Lab 1 Steps 1-4 are available in a published History for you to import and view:

https://galaxy.hpc.uiowa.edu/u/elizabeth-black/h/lab-1-step-1-4-analsis

or: Shared Data -> Published Histories -> Lab 1 Step 1-4 Analysis



Step 5: Comparing Intervals from the Otoscope v4 bed file, and the Whole Exome Bait Interval File

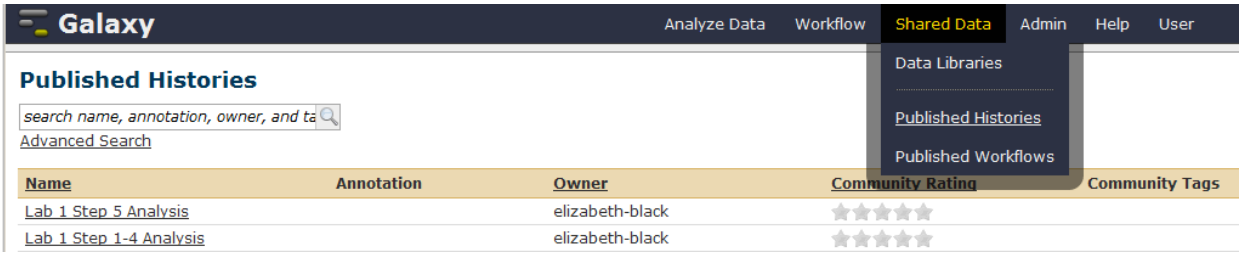
Using similar techniques to Step 3-4, compare the two interval files to see the differences between the target capture regions and a whole exome bait file. The steps to compare the files will not be documented in detail, but here are some high level steps for you to try:

1. Copy the two bed file datasets into a new Galaxy History (see Options -> Copy Datasets)
2. Load the new history as your “current history” in Galaxy
3. Subtract the whole exome bait file from the OtoSCOPE target file to find the non-overlapping interval pieces of the OtoSCOPE targets
4. Intersect the OtoSCOPE targets with the whole exome bait file to find the overlapping interval pieces

Got Stuck? Check out the published history of the competed steps:

https://galaxy.hpc.uiowa.edu/u/elizabeth-black/h/lab-1-step-5-analysis

or:


Galaxy
Analyze Data
Workflow
Shared Data
Admin
Help
User
Data Libraries
Published Histories
Published Workflows
Published Histories
search name, annotation, owner, and ta
Advanced Search
Name
Annotation
Owner
Community Rating
Community Tags
Lab 1 Step 5 Analysis
elizabeth-black
Lab 1 Step 1-4 Analysis
elizabeth-black