Use Case 18: Mapping ontogenetic pathways of cellular differentiation using Human Epigenome Atlas data and the epigenome toolset within the Genboree Workbench

American Society of Human Genetics
Boston, MA 2013

Presented by the
Bioinformatics Research Laboratory

Baylor College of Medicine

Summary of Use Case 18

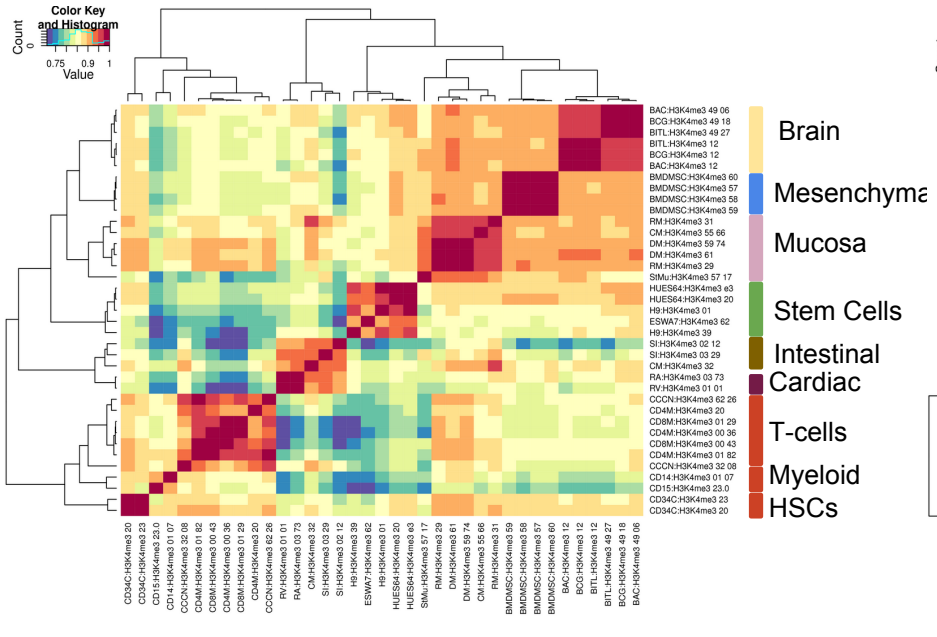
Background: The epigenome plays a key role in establishing and maintaining cellular phenotype during cellular differentiation. The wealth of data from large-scale sequencing projects provides a resource for biological discovery and analysis. The Human Epigenome Atlas, developed as part of the NIH Epigenome Roadmap Project, contains Chip-Seq data from over 100 different cell-types and tissues. This data repository provides a rich resource for ongoing comparative analysis on data generated outside of the NIH Epigenome Roadmap project. We demonstrate here the use of Atlas data to further our understanding of cellular differentiation by examining cell lineage relationships based on clustering over functional genetic elements, such as promoters, enhancers, and lincRNAs.

Results: Analysis validates current knowledge regarding H3K4me3 signals at protein coding gene promoters by clustering cell types of similar lineage[1]. Cluster analysis also reveals H3K4me1 signal at lincRNA promoters able to discriminate cell types/lineages, suggesting an important role for lincRNA in maintaining cellular identity.

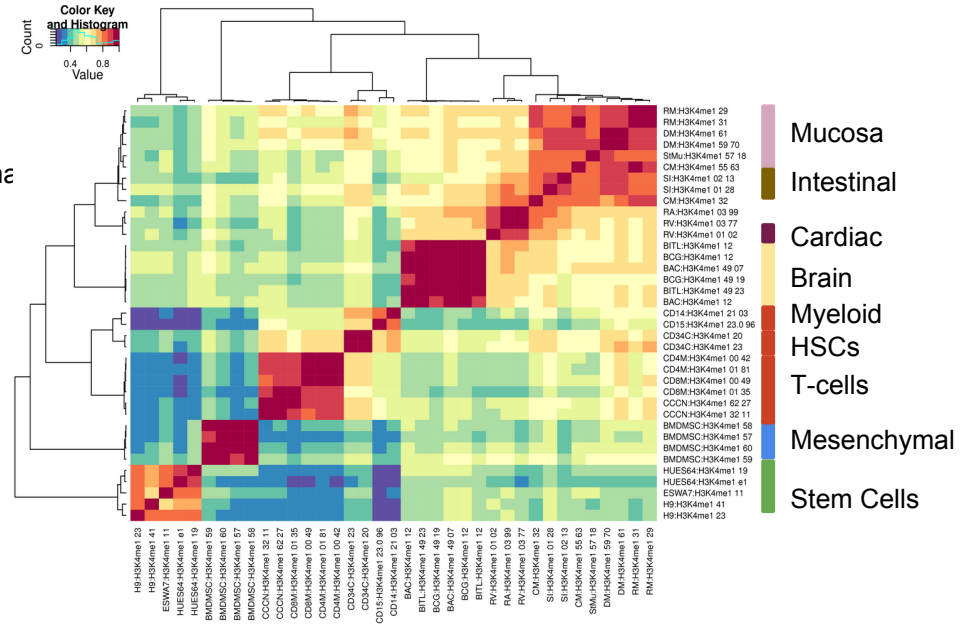
1. Guenther, M. G., Levine, S. S., Boyer, L. A., Jaenisch, R. & Young, R. A. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell* **130,** 77–88 (2007).

Summary of Results

H3K4me3 signal over protein coding gene promoters on the NIH Roadmap Epigenome data



H3K4me1 signal over lincRNA gene promoters on the NIH Roadmap Epigenome data



Cluster analysis of the epigenomes profiled in the NIH Roadmap Epigenome project suggests that lincRNAs play an important role in maintaining cellular identity.

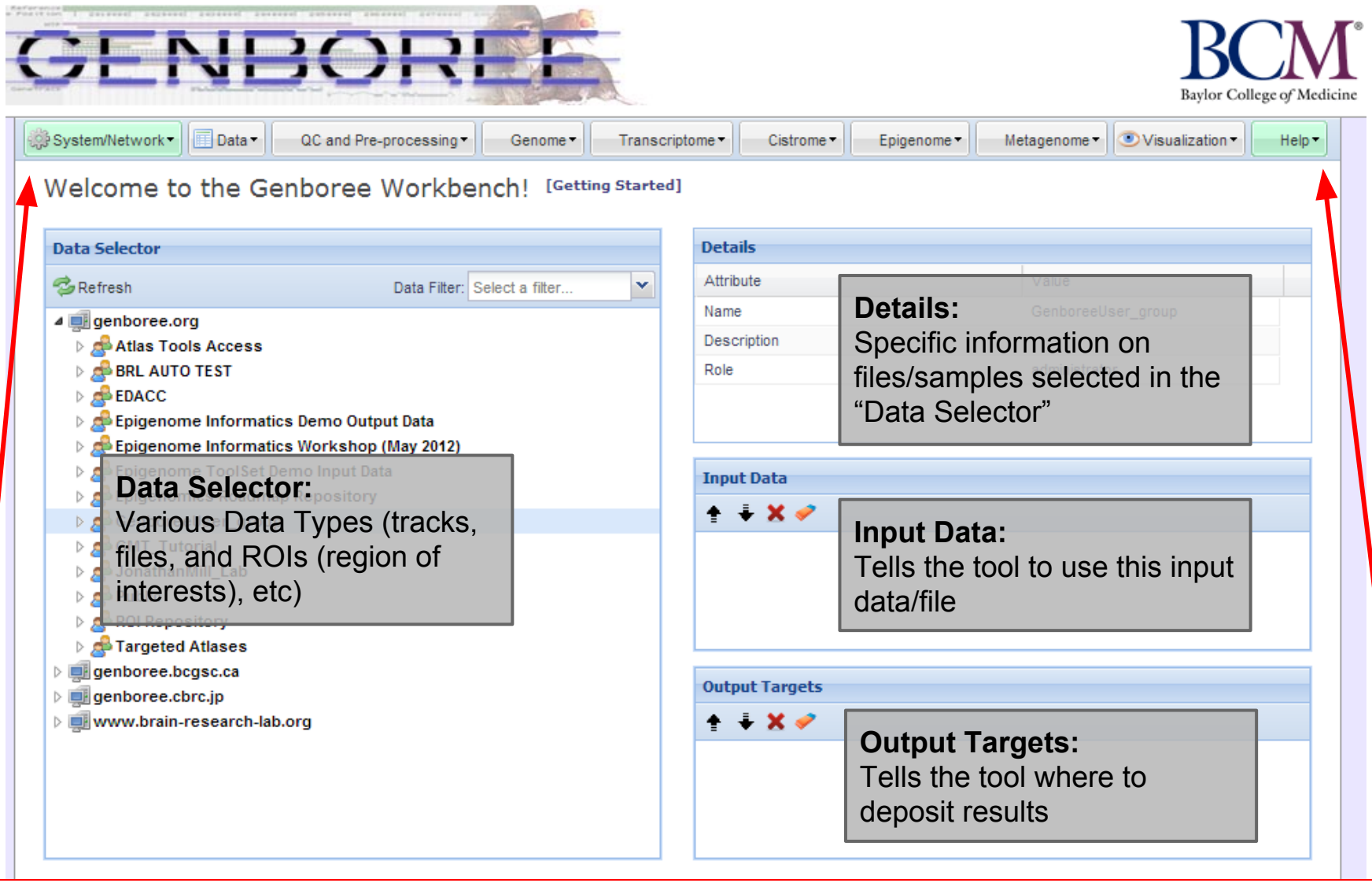
Use Case Overview

New Genboree Users - Slides 5-13 provide steps for new Genboree users on how to create a database, a project page, and view track grid of data generated in the NIH Roadmap Epigenome Project.

Existing Genboree Users - If you have attended past Genboree Workshops or are familiar with the Genboree Workbench then you may briefly review these slides and start on slide 14 for the actual use case

- Methodology
- Steps for reproducing the results

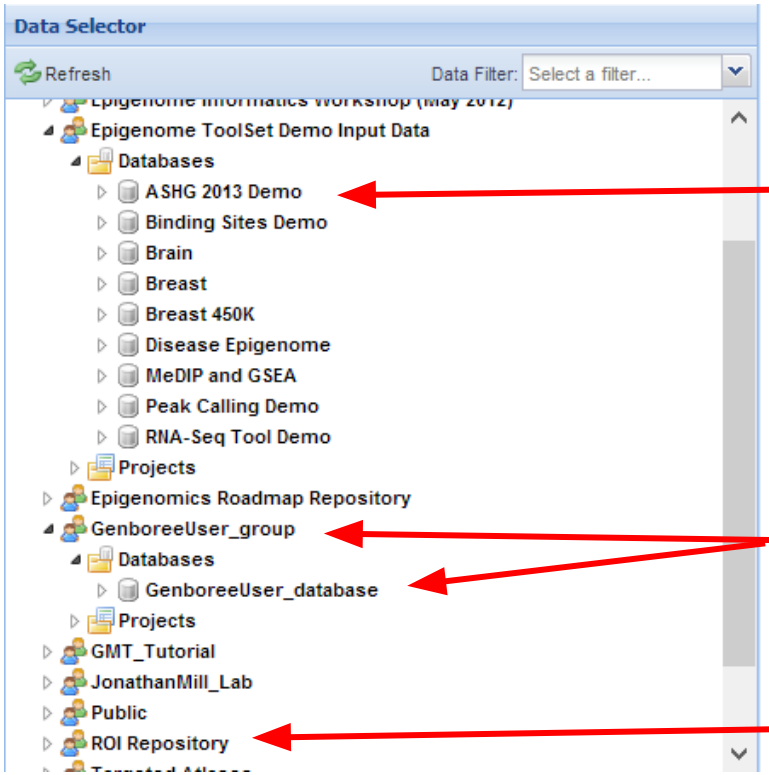
The Genboree Workbench: Web-based Data Management & Analysis



Important: **Toolset Menu turns GREEN when "Input Data" and "Output Targets" are properly populated for a tool to run. Please note that "System/Network" and "Help" options are always green since "User Profile", "Jobs", and "Request Feature" are always available for use and do not need "Input Data" and "Output Targets" to be populated.**

Preparation Prior to Starting the Use Case

- "***GenboreeUser*_group**" is a name template for an automatically created Genboree user group **for you** where "***GenboreeUser***" is **your user name**.
- Similarly, "***GenboreeUser*_database**' is a name template for your database.
- Of course, you may create many more databases and may create and be member of many other groups.

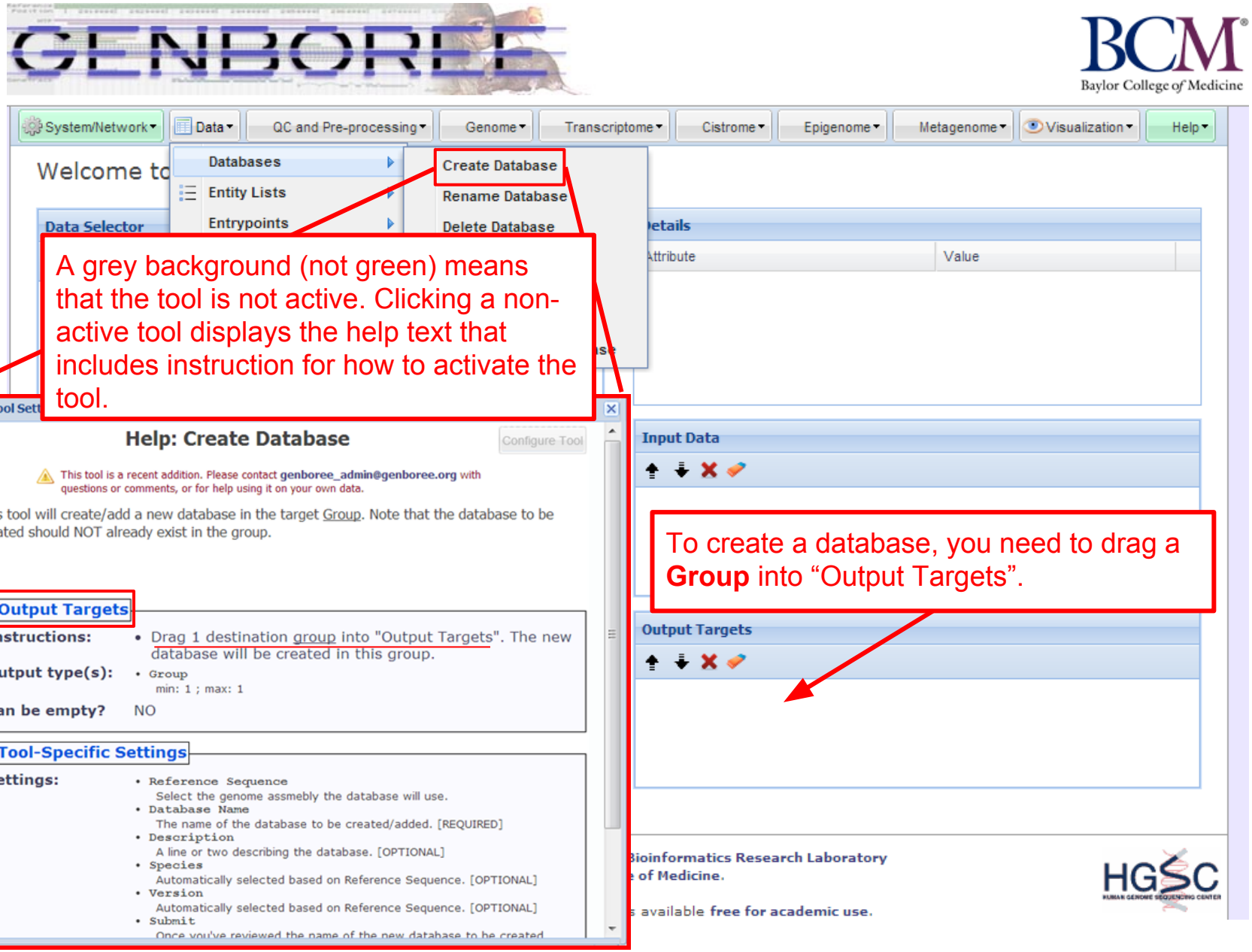


Under "Epigenome Toolset Demo Input Data" you will find "ASHG 2013 Demo" database, where we have provided you with sample data to try out the use cases

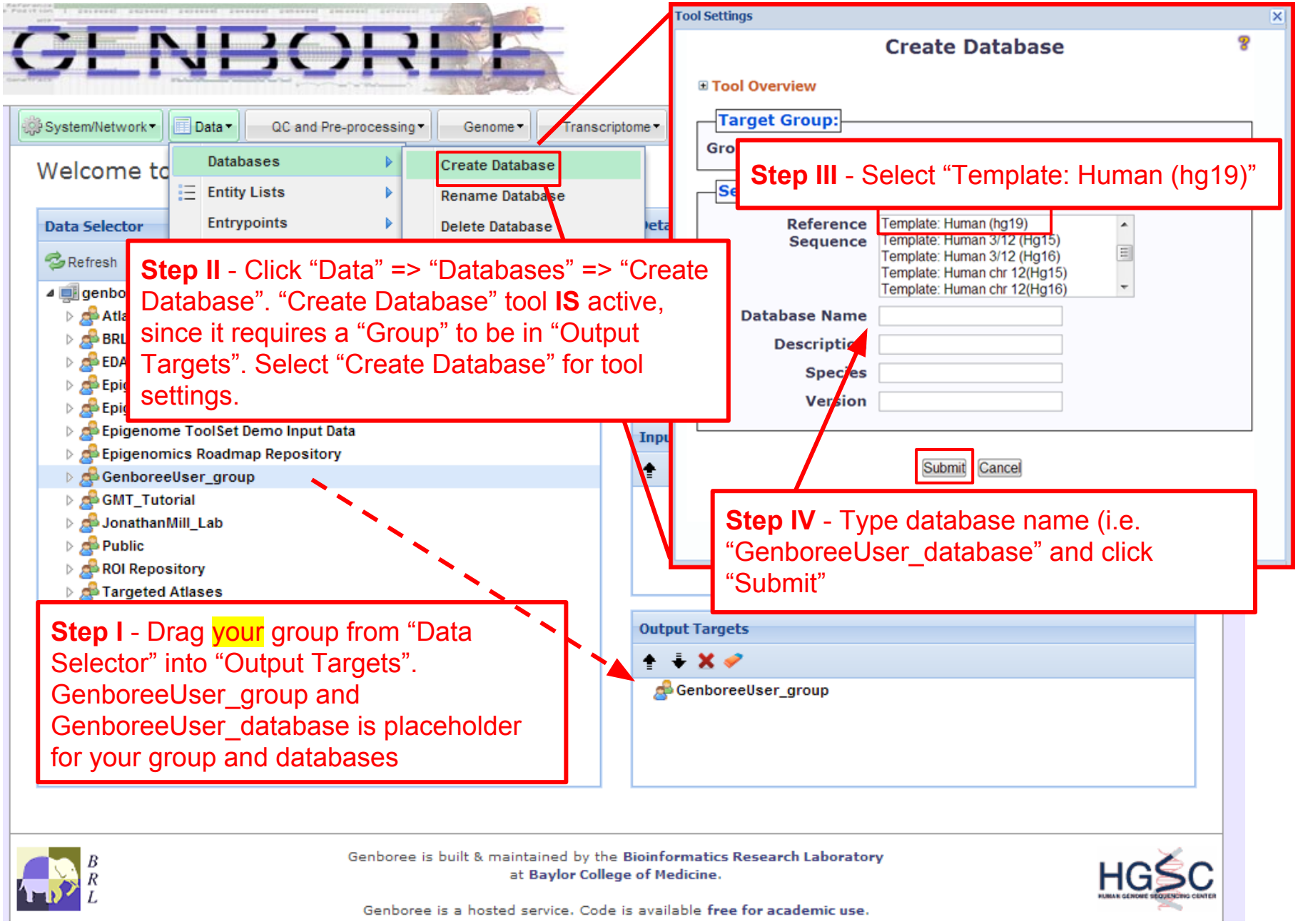
When making screenshots for providing instructions we have used "GenboreeUser_group" and "GenboreeUser_database" as output targets, however, you will use your own group and database. Following slides will show you how to create database and project

ROI Repository contains database of annotated regions (eg. Gencode and Refseq annotations)

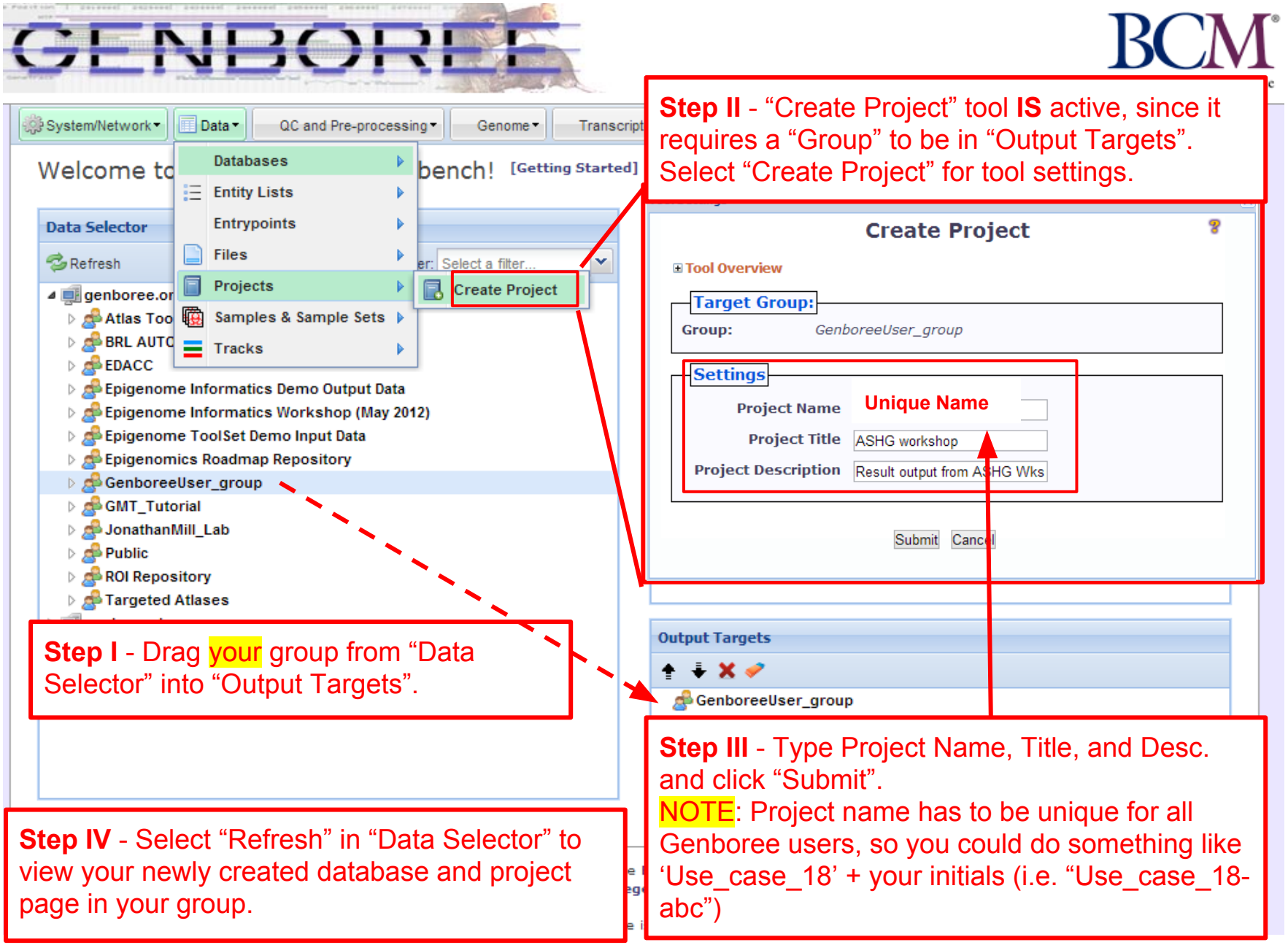
Display Tool Setting “Help” dialogue box in the Workbench



Steps for Creating a Database



Steps for Creating a Project page



Step I - Drag your group from "Data Selector" into "Output Targets".

Step II - "Create Project" tool **IS** active, since it requires a "Group" to be in "Output Targets". Select "Create Project" for tool settings.

Step III - Type Project Name, Title, and Desc. and click "Submit".
NOTE: Project name has to be unique for all Genboree users, so you could do something like 'Use_case_18' + your initials (i.e. "Use_case_18-abc")

Step IV - Select "Refresh" in "Data Selector" to view your newly created database and project page in your group.

Description of Regions of Interest (ROIs) Tracks

Source of ROIs that will be used in this analysis:

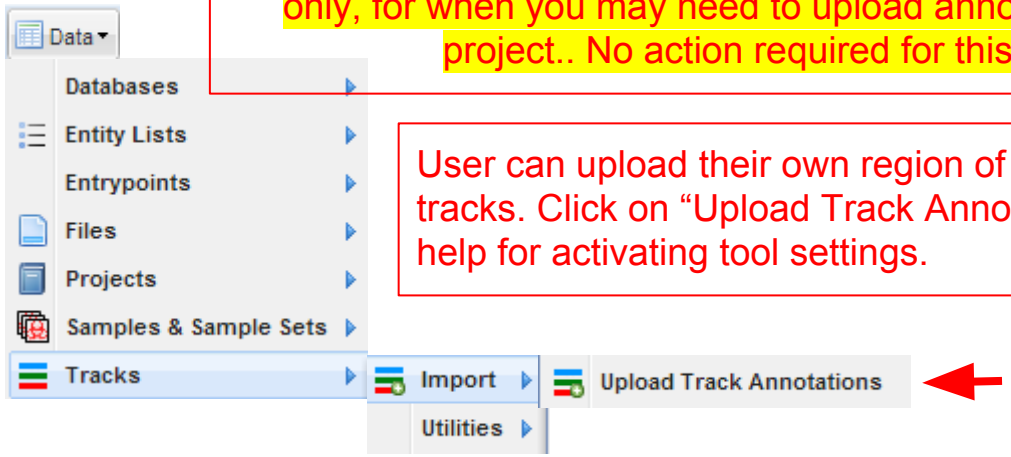
- **Track "GeneProteinCoding_promoter":** This track contains promoters of protein coding genes that were defined using Gencode V10 annotations (www.gencodegenes.org/), with transcription start sites (TSS) +/- 1500bp. The track contains 20,007 promoters from human genome build Hg19.

- **Track "GeneLincRNA_promoter":** This track contains promoters of lincRNAs that were defined using Gencode V10 annotations, with transcription start site (TSS) +/- 1500bp. the track contains 5,484 promoters from human genome build Hg19.

- **Track "ChromHMM:Enhancers":** This track contains enhancers obtained from Manolis Kellis. Enhancers here are defined by ChromHMM using the NIH Roadmap Consortium data (www.epigenomeatlas.org).[1]

Please note: Upload of annotations is provided for your information only, for when you may need to upload annotations for your own project.. No action required for this use case.

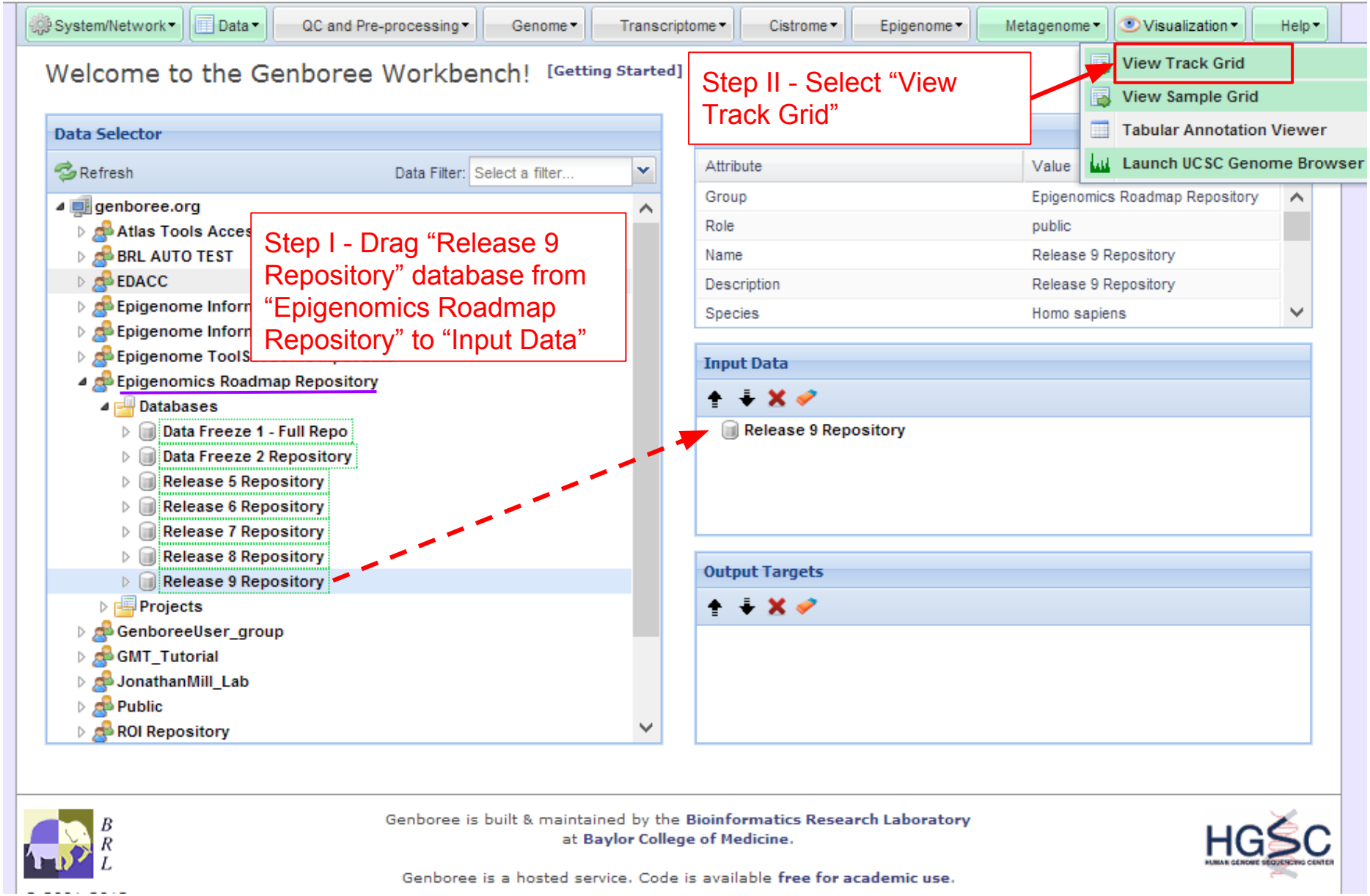
Upload your own Track Annotations:



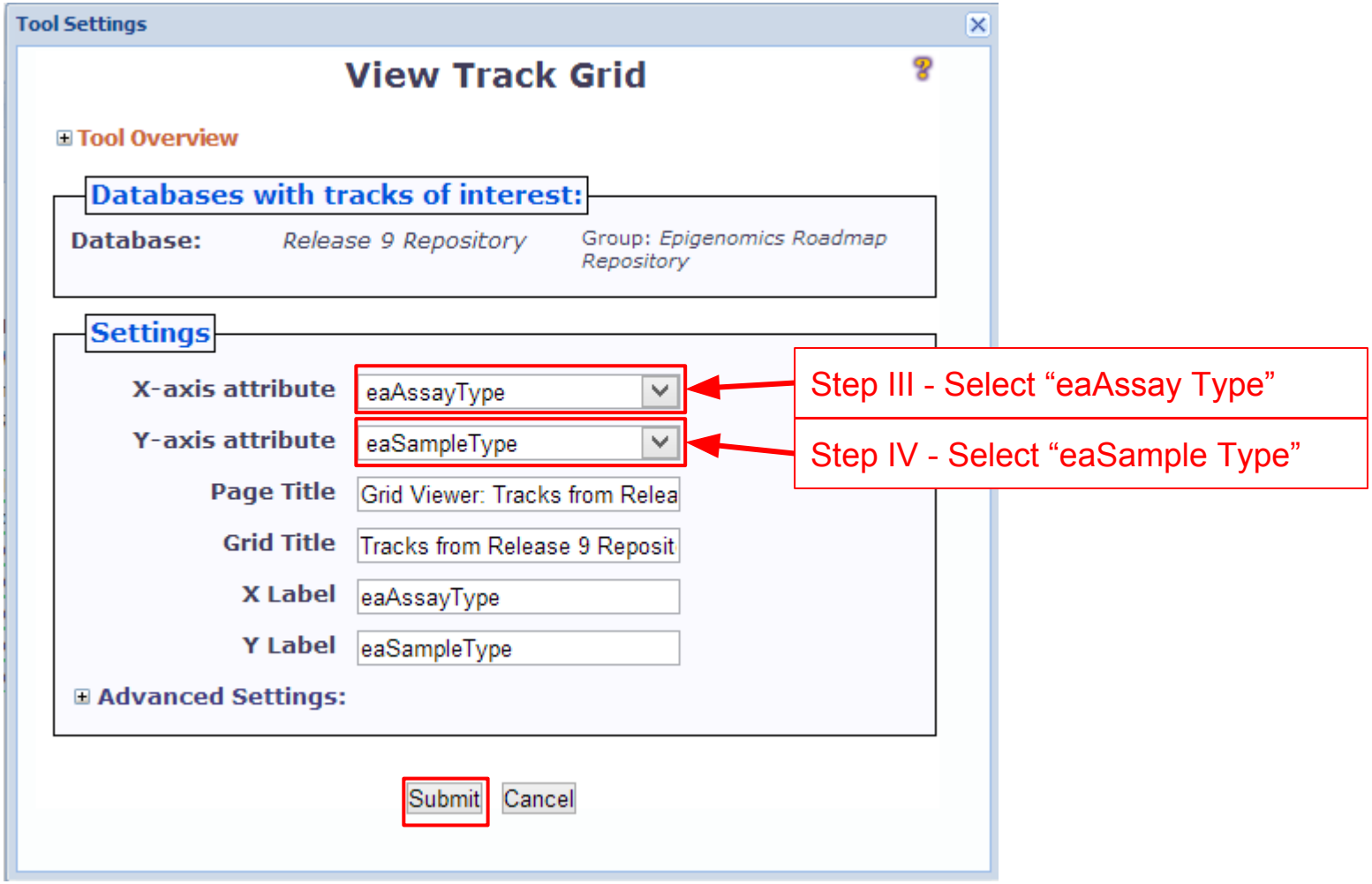
User can upload their own region of interests tracks. Click on "Upload Track Annotations" to see help for activating tool settings.

1. Ernst, J. & Kellis, M. "ChromHMM: automating chromatin-state discovery and characterization". *Nat. Methods* **9,** 215–216 (2012).

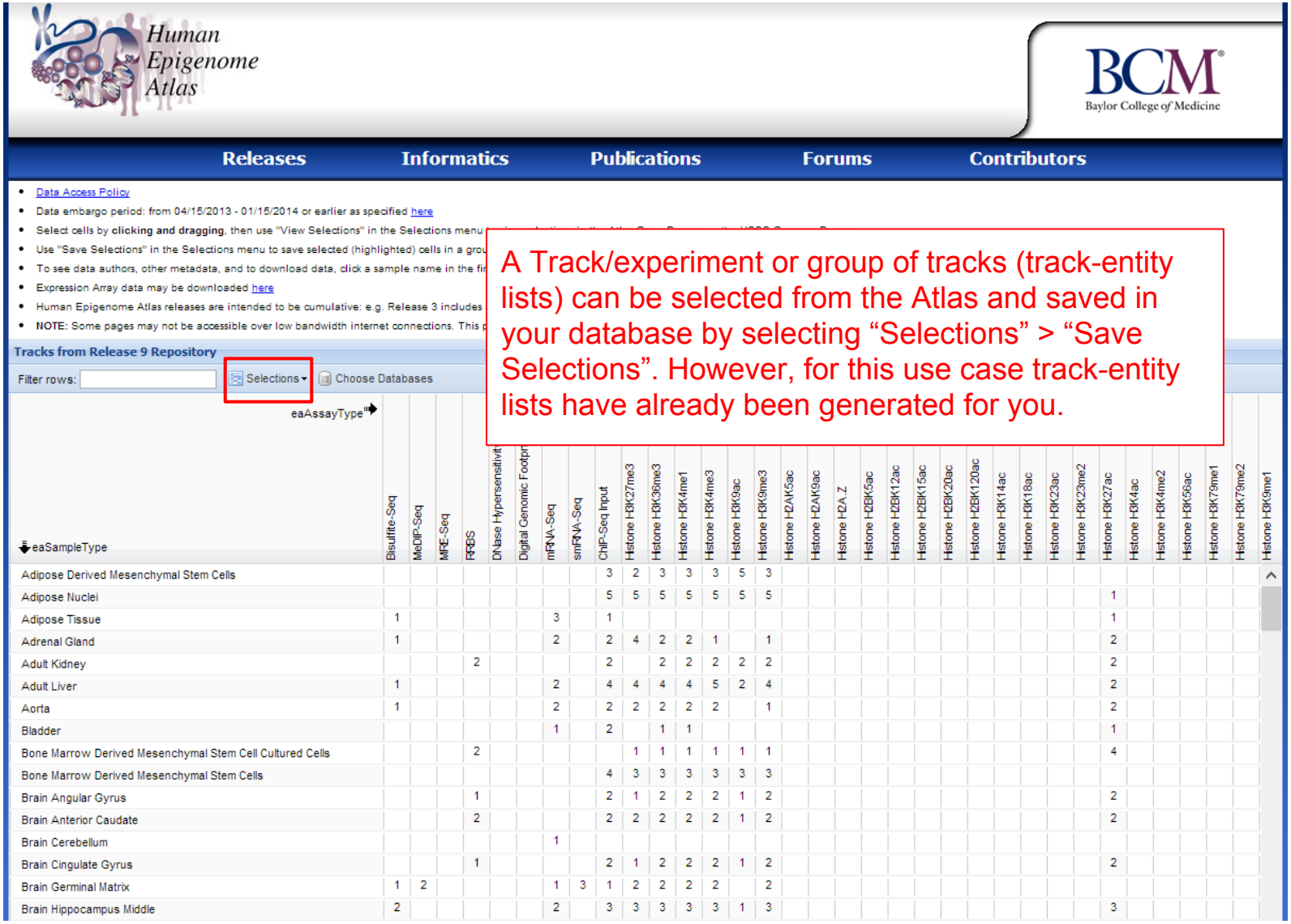
Follow these steps to view Track Grid of data from the Roadmap Epigenome Project



Select how you want the tracks displayed in the “View Track Grid” tool.



Track Grid view of the data from Roadmap Epigenome Project



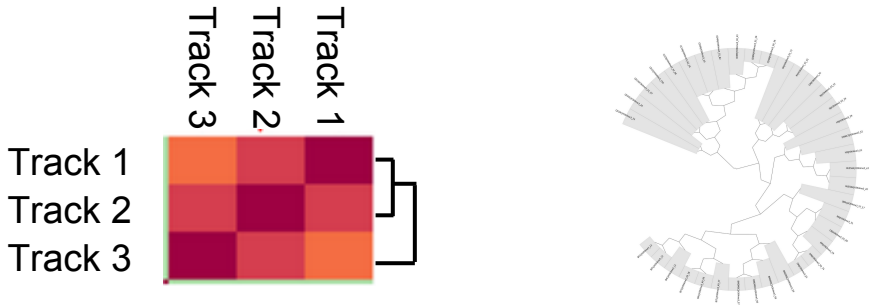
A Track/experiment or group of tracks (track-entity lists) can be selected from the Atlas and saved in your database by selecting “Selections” > “Save Selections”. However, for this use case track-entity lists have already been generated for you.

Methodology Overview

H3K4me3 Atlas signal projected over 20,007 protein coding gene promoters (Gencode defined)

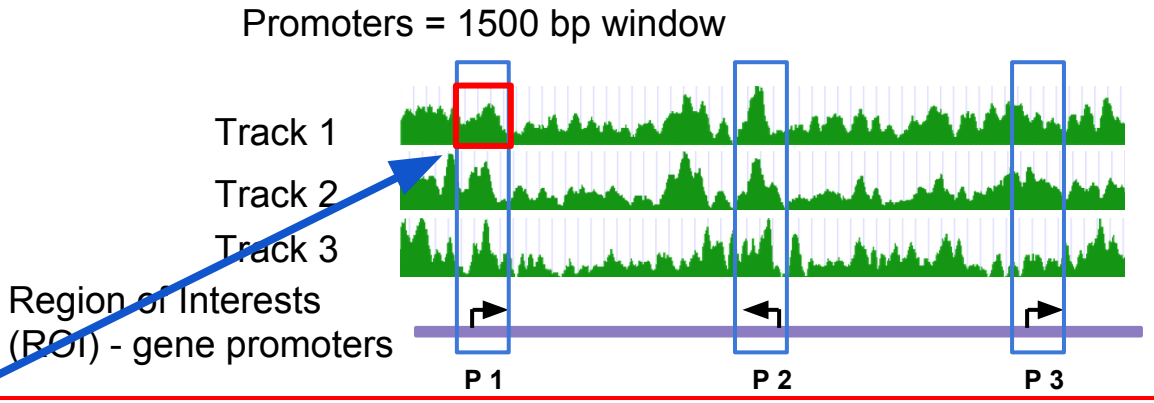
H3K4me1 Atlas signal projected over 5,484 lincRNA gene promoters (Gencode defined)

Cluster analysis (Pearson Correlation) and display as Heatmap and Newick file



Methodology: Clustering/Heatmap

I. Data Selection



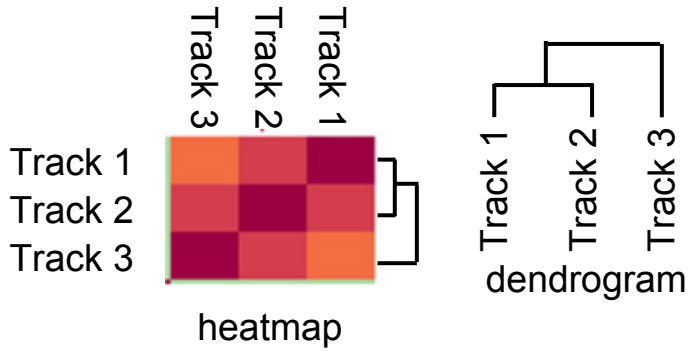
II. Signal processing- Heatmap Tool calculates average signal for each track and ROI (promoter) combination

	Track 1	Track 2	Track 3
P 1	0.8	0.7	0.3
P 2	0.7	0.6	0.7
P 3	0.6	0.8	0.2

Data matrix (3 x 3 shown here) with average signal is normalized and a correlation matrix generated. Correlation matrix is used to calculate distance measure and hierarchical clustering to group samples based on similarity to plot dendrogram.

III. Visualization of results

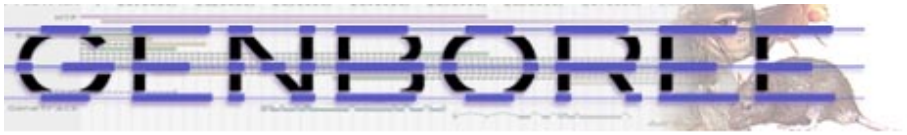
Correlation values are dynamically scaled and represented in heatmap.



Note: Heatmap and dendrogram are shown as result. Genboree generates dendrogram separately, since if the output contains many tracks, row/column labels may be hard to visualize in heatmap.

Rationale for Selection of H3K4me3 Histone Mark

In this use case, we will use H3K4me3 signal tracks of various cell types and tissues profiled in the NIH Roadmap Epigenomics project. The H3K4me3 mark is associated with active promoters. We sought to ask how the various cell types/tissues in the Human Epigenome Atlas cluster using a H3K4me3-promoter combination. Here, the promoters are defined as the transcription start site (TSS) +/- 1500 base pairs using Gencode annotations.



BCM
Baylor College of Medicine

System/Network | Data | QC and Pre-processing | Genome | Transcriptome | Cistrome | Epigenome | Metagenome | Visualization | Help

Welcome to the Genboree Workbench! [Getting Started]

Data Selector

Refresh Data Filter: Select a filter...

- Epigenomics Roadmap Repository
- GenboreeUser_group
- GMT_Tutorial
- JonathanMill_Lab
- Public
- ROI Repository
  - Databases
    - ROI Repository - hg18
    - ROI Repository - hg19
      - All Annotations in Database
      - Tracks
        - Class: Affymetrix
        - Class: Agilent
        - Class: ENCODE
        - Class: ENCODE - T.f. Binding Sites
        - Class: Enhancer
        - Class: GC
        - Class: Gencode
          - Gene:LincRNA_promoter
          - Gene:ProteinCoding_promoter
        - Class: Gene
        - Class: Gene Model

Step1 - Drag Promoter Track to Input Data

Details

Attribute	Value
Group	ROI Repository
Database	ROI Repository - hg19
Name	Gene:ProteinCoding_promoter
Description	
BigBed	none

Input Data

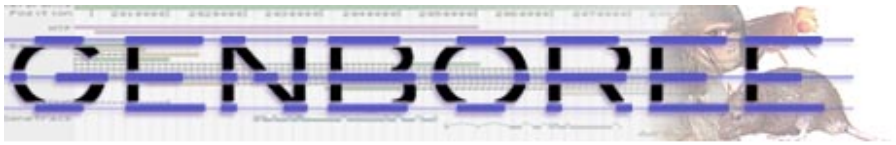
Gene:ProteinCoding_promoter

Output Targets

Genboree is built & maintained by the **Bioinformatics Research Laboratory** at **Baylor College of Medicine.**

Genboree is a hosted service. Code is available **free for academic use.**

HGSC
HUMAN GENOME SEQUENCING CENTER



BCM
Baylor College of Medicine

System/Network | Data | QC and Pre-processing | Genome | Transcriptome | Cistrome | Epigenome | Metagenome | Visualization | Help

Welcome to the [Started]

Step 2 - Drag "release9_H3K4me3_subset" tracks to Input Data

We are providing the heatmap tool with two types of tracks: data tracks and annotation tracks. The signal intensity from the "Release9_H3K4me3_subset", 37 data tracks, are projected over the "Gene: ProteinCoding_promoter" annotation tracks. The Pearson correlation metric is calculated and the similarity matrix is presented visually as a heatmap.

Data Selector

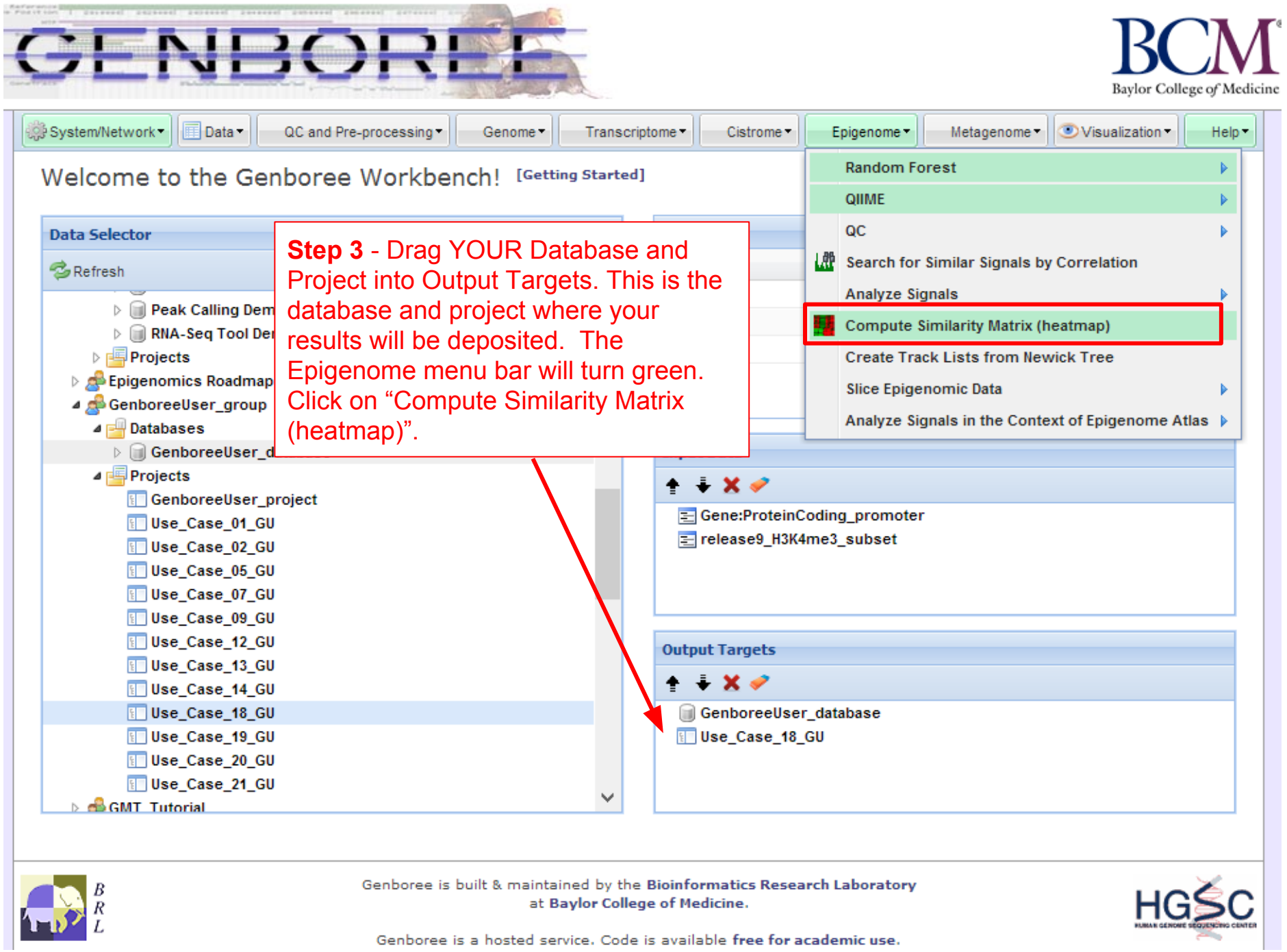
Refresh

- Epigenome ToolSet Demo Input Data
  - Databases
    - ASHG 2013 Demo
      - All Annotations in Database
      - Tracks
      - Lists & Selections
        - Lists of Tracks
          - HL60_Tracks
          - HSC_Tracks
          - Immune_HL60_H3K4me1
          - Myeloid_Tracks
          - release9_H3K27me3
          - release9_H3K36me3
          - release9_H3K4me1
          - release9_H3K4me1_subset
          - release9_H3K4me3
          - release9_H3K4me3_subset
          - release9_H3K9me3
      - SampleSets
      - Samples
      - Files
      - Queries

Input Data

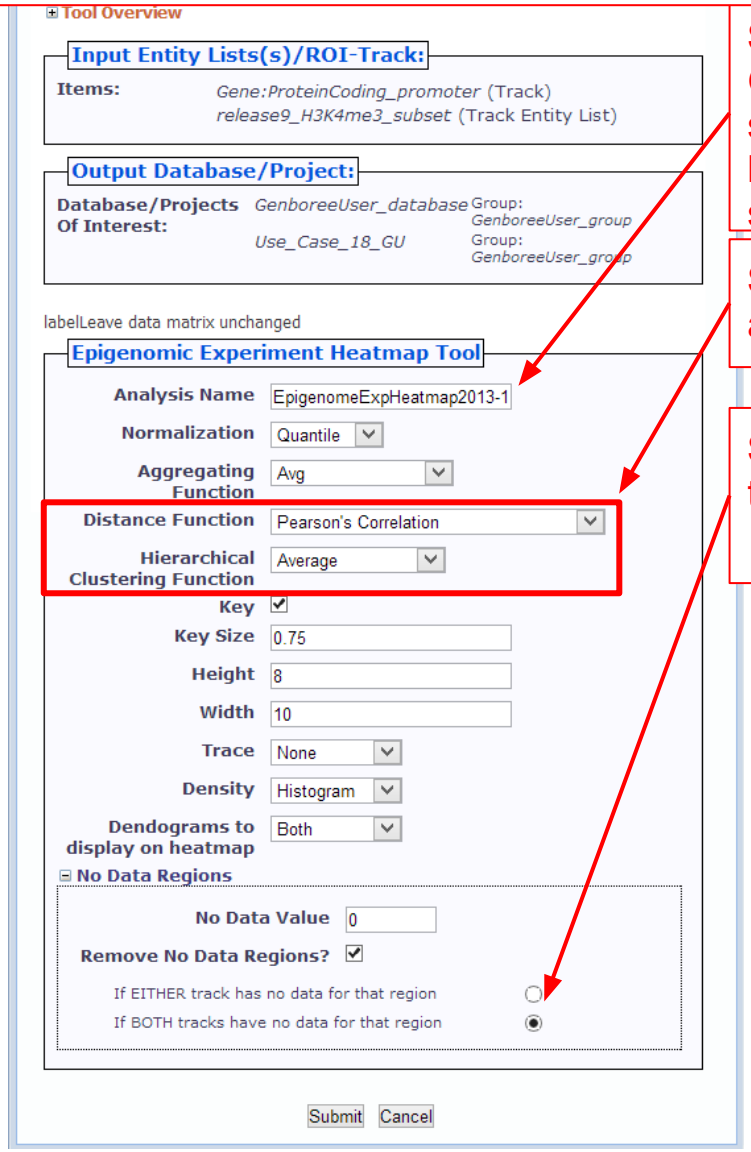
- Gene:ProteinCoding_promoter
- release9_H3K4me3_subset

Output Targets


GENBOREE
BCM
Baylor College of Medicine
System/Network
Data
QC and Pre-processing
Genome
Transcriptome
Cistrome
Epigenome
Metagenome
Visualization
Help
Welcome to the Genboree Workbench! [Getting Started]
Random Forest
QIIME
QC
Search for Similar Signals by Correlation
Analyze Signals
Compute Similarity Matrix (heatmap)
Create Track Lists from Newick Tree
Slice Epigenomic Data
Analyze Signals in the Context of Epigenome Atlas
Step 3 - Drag YOUR Database and Project into Output Targets. This is the database and project where your results will be deposited. The Epigenome menu bar will turn green. Click on "Compute Similarity Matrix (heatmap)".
Data Selector
Refresh
Peak Calling Dem
RNA-Seq Tool Der
Projects
Epigenomics Roadmap
GenboreeUser_group
Databases
GenboreeUser_d
Projects
GenboreeUser_project
Use_Case_01_GU
Use_Case_02_GU
Use_Case_05_GU
Use_Case_07_GU
Use_Case_09_GU
Use_Case_12_GU
Use_Case_13_GU
Use_Case_14_GU
Use_Case_18_GU
Use_Case_19_GU
Use_Case_20_GU
Use_Case_21_GU
GMT_Tutorial
Gene:ProteinCoding_promoter
release9_H3K4me3_subset
Output Targets
GenboreeUser_database
Use_Case_18_GU
Genboree is built & maintained by the Bioinformatics Research Laboratory at Baylor College of Medicine.
Genboree is a hosted service. Code is available free for academic use.
BRL
HGSC

Check that the "Input Files Directory and "Output Database" and "Project" are correct (based on what you named them). The default parameters will be used unless noted otherwise.

A default "Analysis Name" is generated by Genboree. It is recommended that all text and the time stamp be kept, and that you append some unique text to the beginning to help you distinguish different jobs run from the same tool.

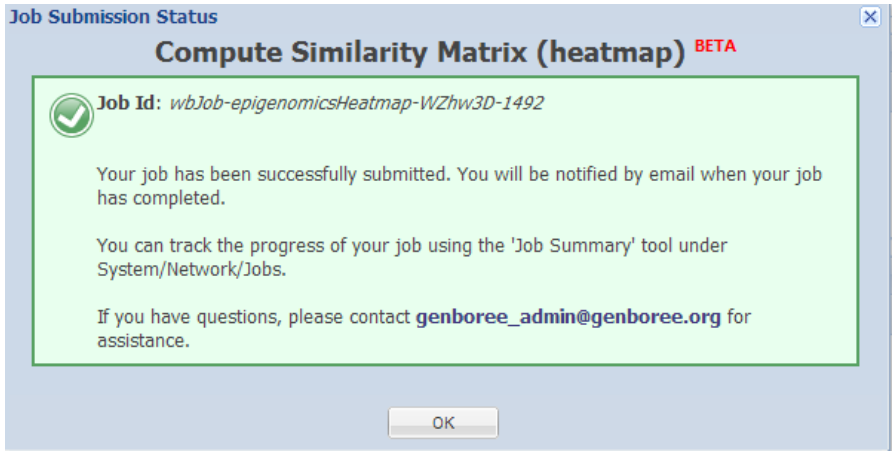


Step 4 - A default "Analysis Name" is generated by Genboree. It is recommended that all text and the time stamp be kept, and that you append some unique text to the beginning to help you distinguish different jobs run from the same tool.

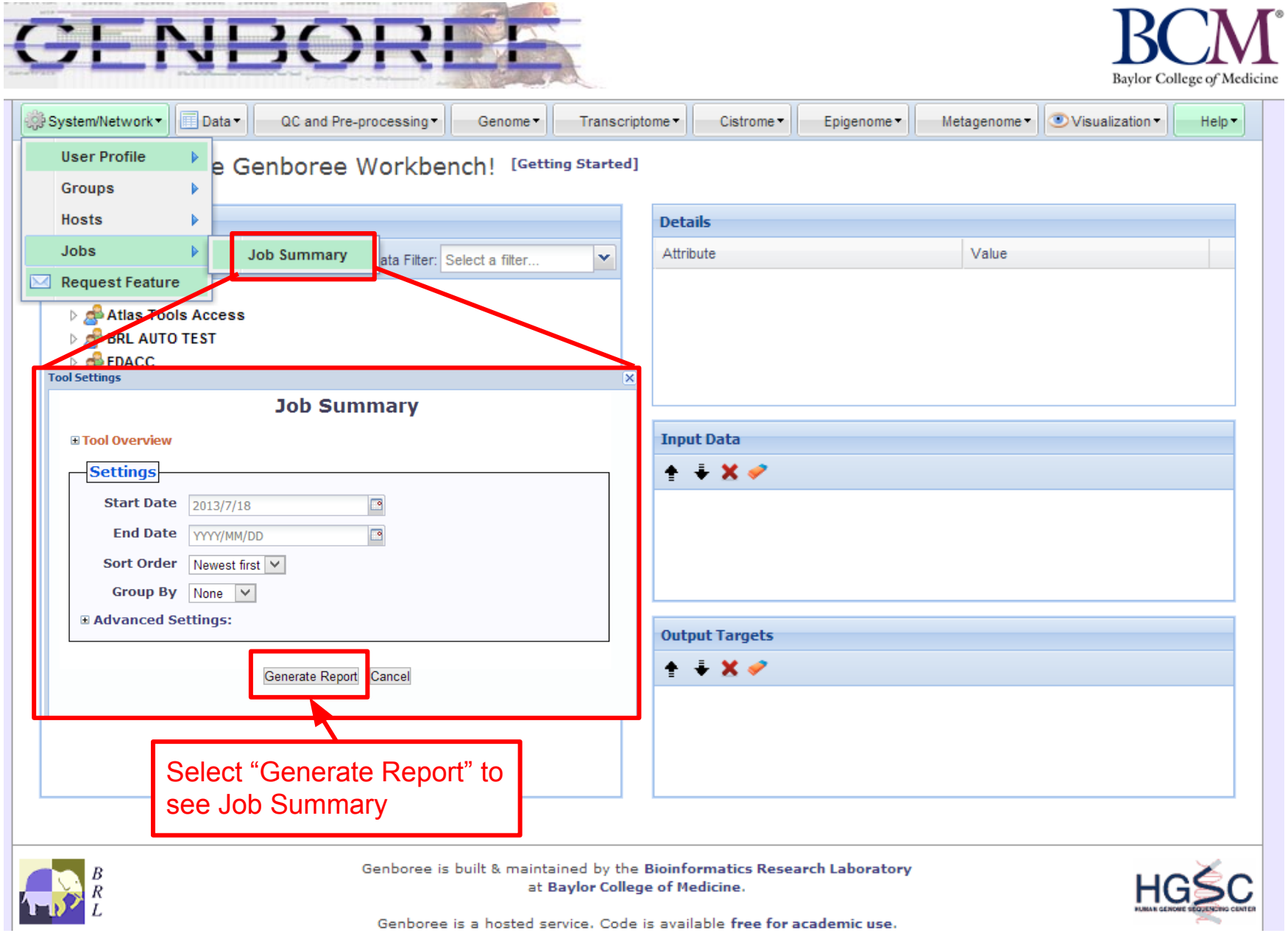
Step 5 - Select "Pearson's Correlation " as distance function and Average as hierarchical clustering function

Step 6 - Expand "No Data Regions" and select "If BOTH tracks have no data for that region". Click "Submit".

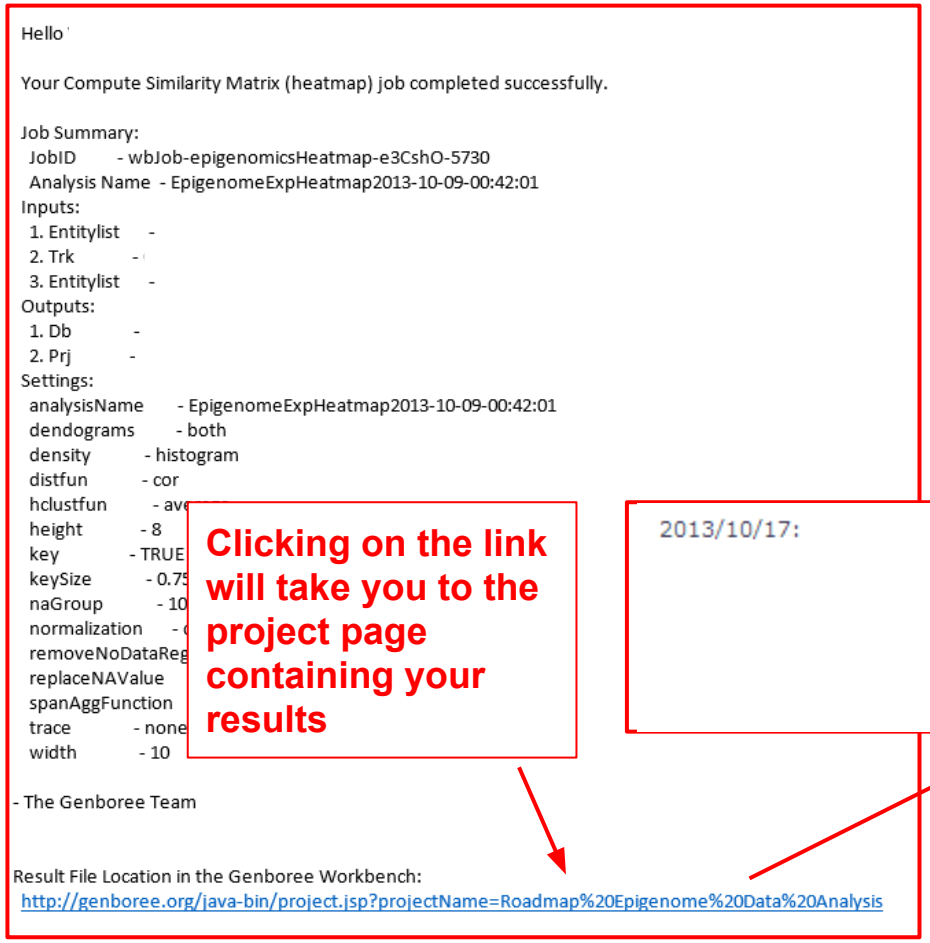
You will see the message below upon successful submission of your heatmap job:



Running time of the job will vary based on the data tracks you choose for analysis. Job you submitted will take around 10 mins to complete once it starts running. You can status of the job through Job Summary



You will get the following e-mail message when your job is completed



Hello

Your Compute Similarity Matrix (heatmap) job completed successfully.

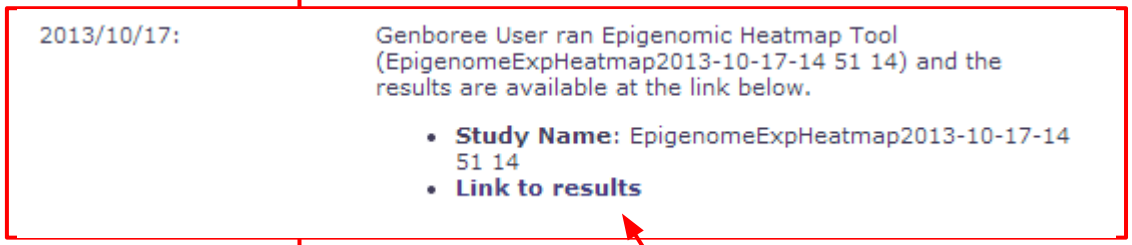
Job Summary:
JobID - wbJob-epigenomicsHeatmap-e3CshO-5730
Analysis Name - EpigenomeExpHeatmap2013-10-09-00:42:01
Inputs:
1. Entitylist -
2. Trk -
3. Entitylist -
Outputs:
1. Db -
2. Prj -
Settings:
analysisName - EpigenomeExpHeatmap2013-10-09-00:42:01
dendograms - both
density - histogram
distfun - cor
hclustfun - ave
height - 8
key - TRUE
keySize - 0.75
naGroup - 10
normalization -
removeNoDataReg
replaceNAValue
spanAggFunction
trace - none
width - 10

- The Genboree Team

Result File Location in the Genboree Workbench:
http://genboree.org/java-bin/project.jsp?projectName=Roadmap%20Epigenome%20Data%20Analysis

Clicking on the link will take you to the project page containing your results

Project Page



2013/10/17: Genboree User ran Epigenomic Heatmap Tool (EpigenomeExpHeatmap2013-10-17-14 51 14) and the results are available at the link below.

- **Study Name**: EpigenomeExpHeatmap2013-10-17-14 51 14
- **Link to results**

Click on the “Link to results” in your Project page

Link to Results on Your Project Page

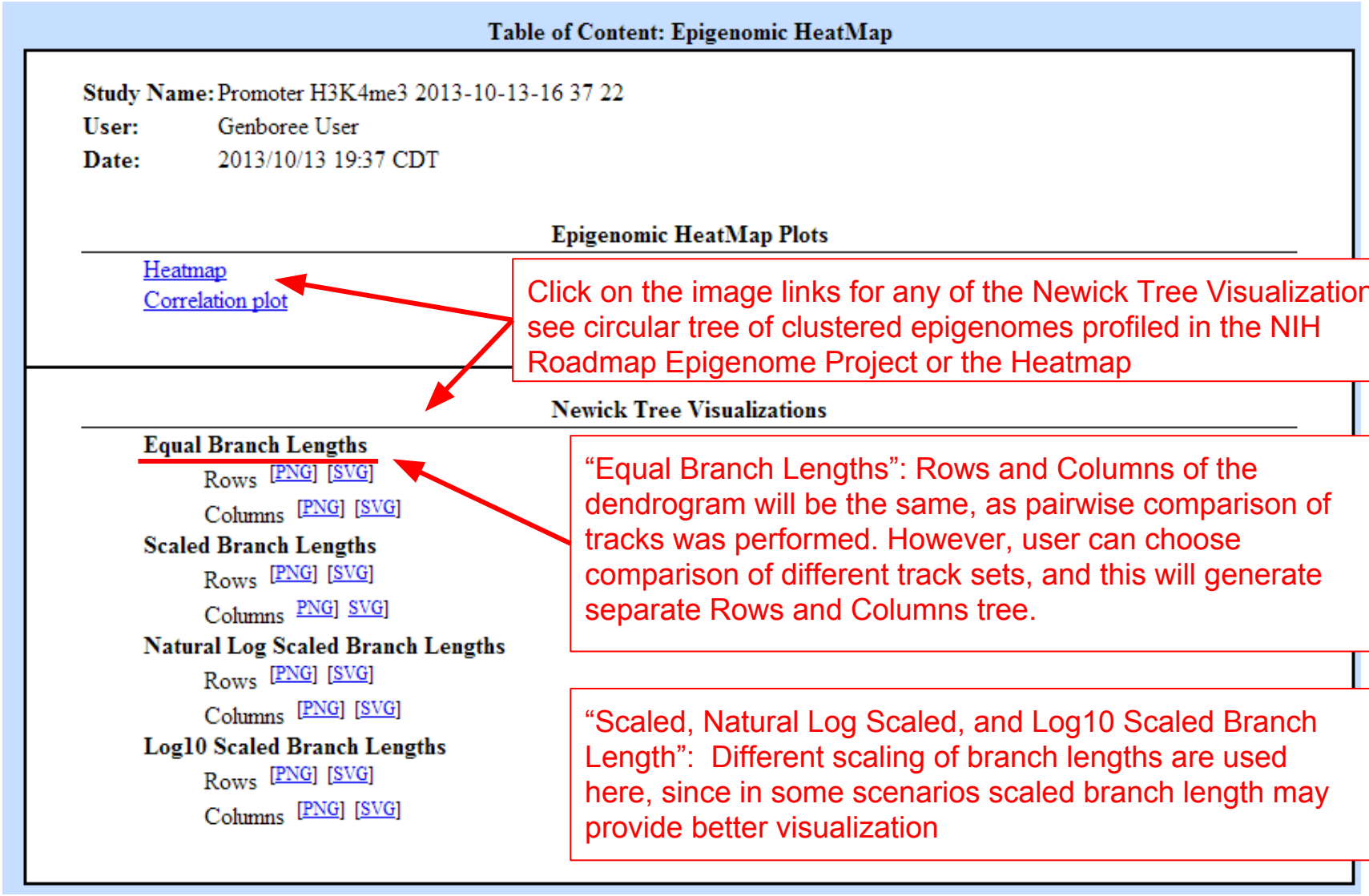


Table of Content: Epigenomic HeatMap

Study Name: Promoter H3K4me3 2013-10-13-16 37 22
User: Genboree User
Date: 2013/10/13 19:37 CDT

Epigenomic HeatMap Plots

Heatmap
Correlation plot

Newick Tree Visualizations

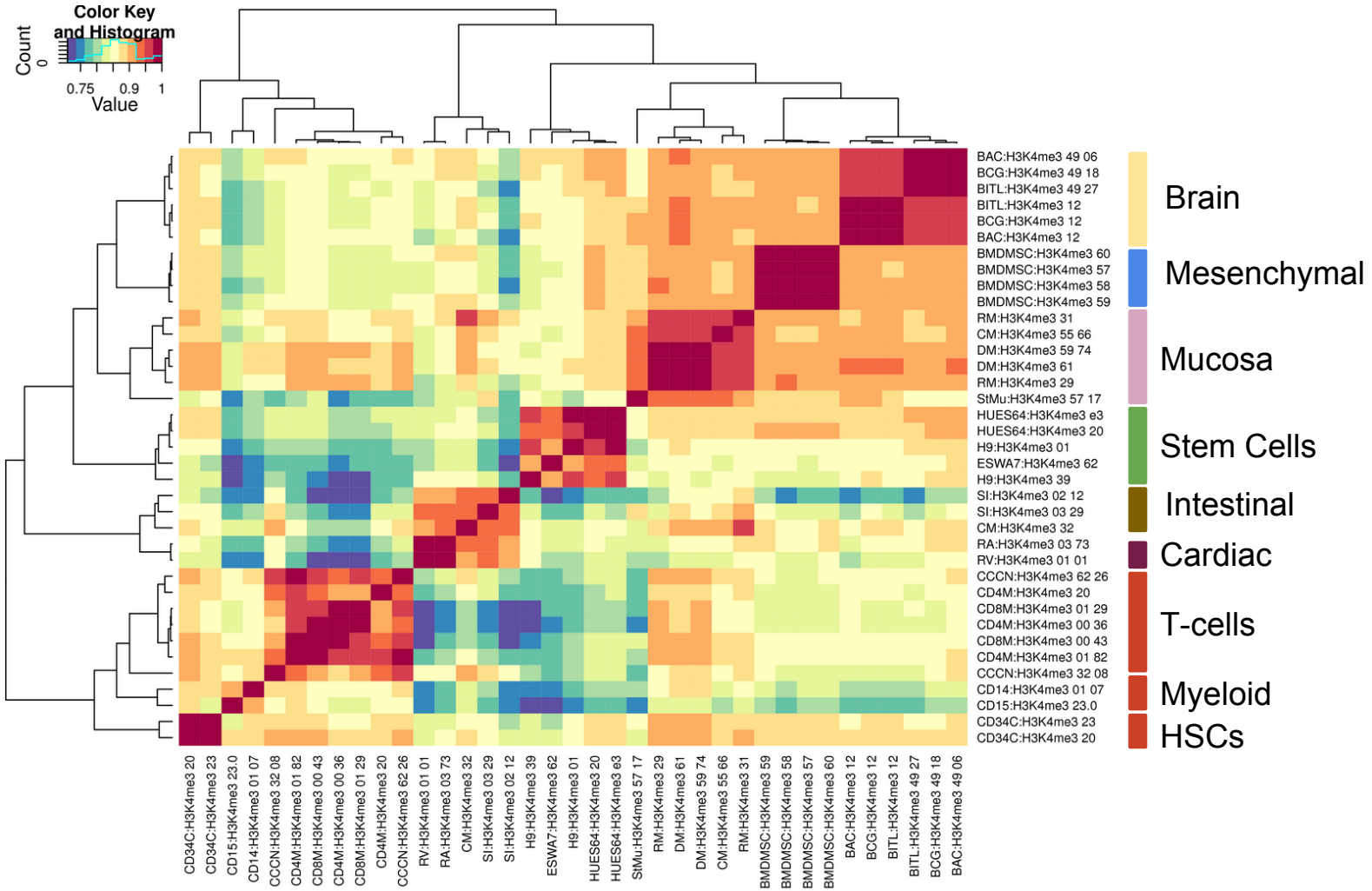
Equal Branch Lengths
Rows [PNG] [SVG]
Columns [PNG] [SVG]
Scaled Branch Lengths
Rows [PNG] [SVG]
Columns PNG] SVG]
Natural Log Scaled Branch Lengths
Rows [PNG] [SVG]
Columns [PNG] [SVG]
Log10 Scaled Branch Lengths
Rows [PNG] [SVG]
Columns [PNG] [SVG]

Click on the image links for any of the Newick Tree Visualizations to see circular tree of clustered epigenomes profiled in the NIH Roadmap Epigenome Project or the Heatmap

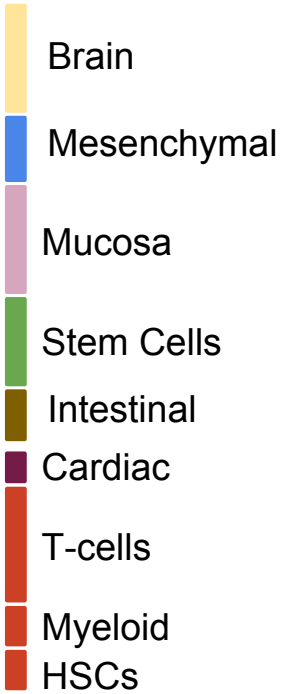
"Equal Branch Lengths": Rows and Columns of the dendrogram will be the same, as pairwise comparison of tracks was performed. However, user can choose comparison of different track sets, and this will generate separate Rows and Columns tree.

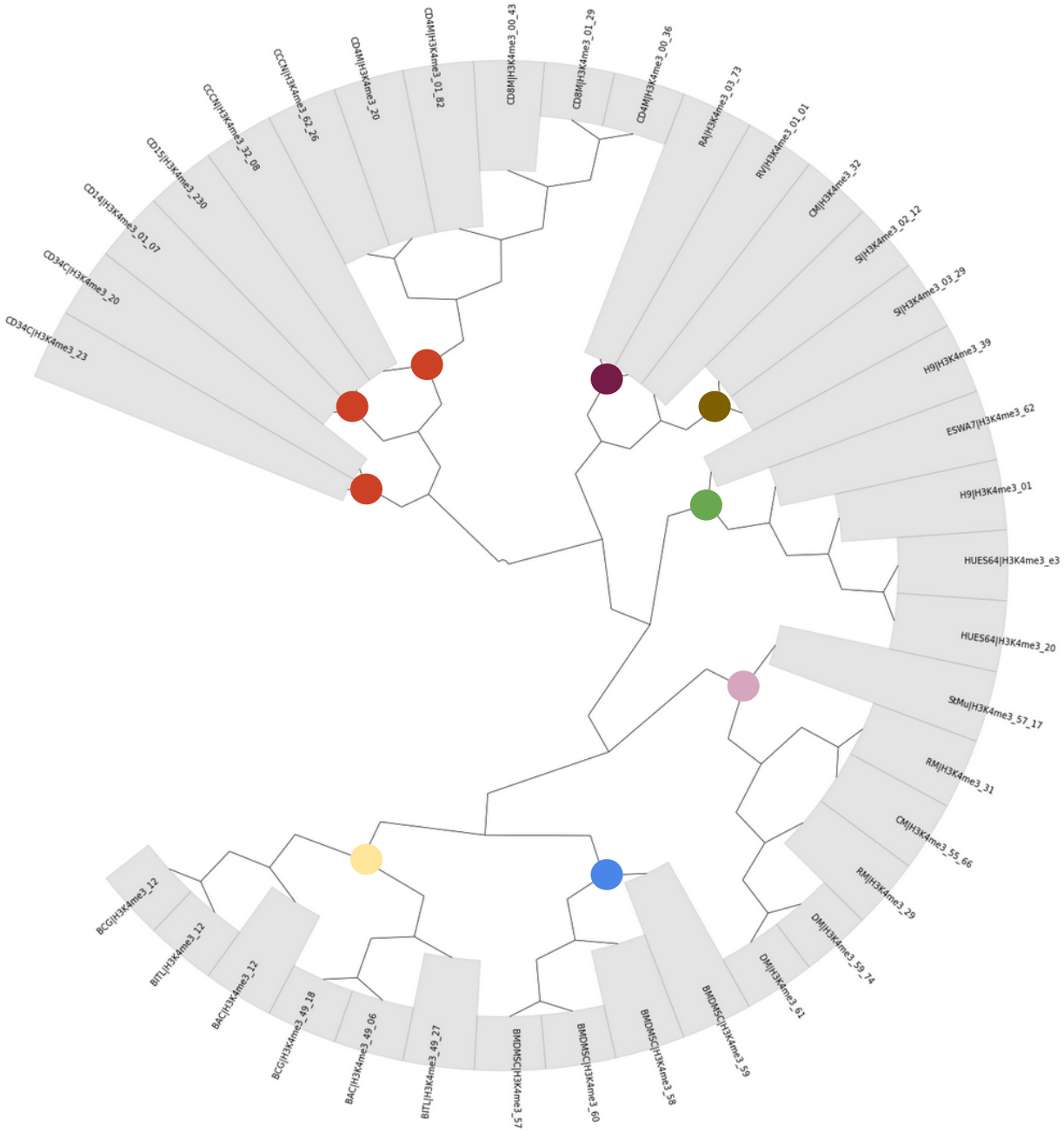
"Scaled, Natural Log Scaled, and Log10 Scaled Branch Length": Different scaling of branch lengths are used here, since in some scenarios scaled branch length may provide better visualization

H3K4me3 signal at protein-coding promoter region can distinguish different cell types/tissues of origin



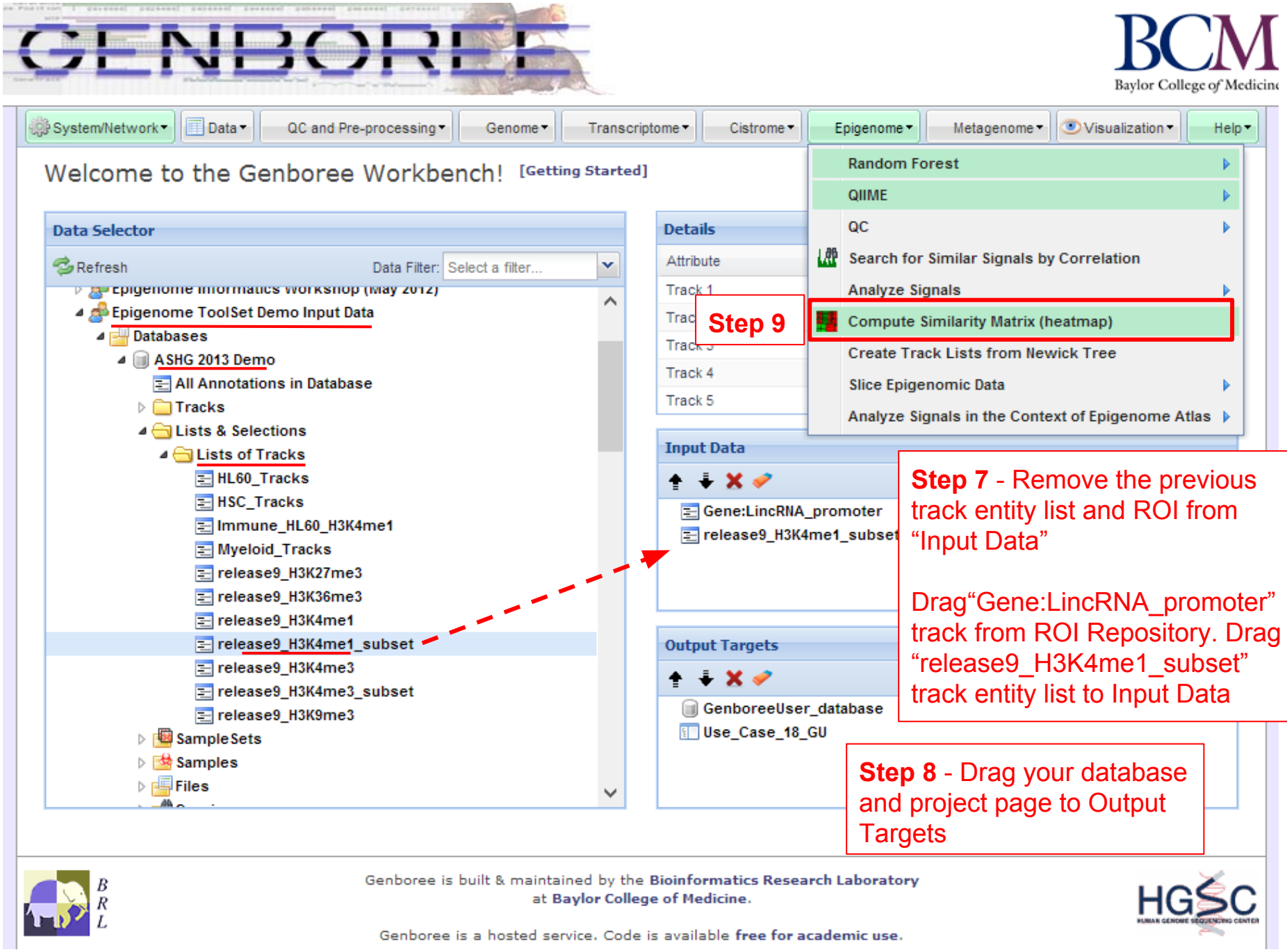
Clustering Results Presented in a Circular Dendrogram

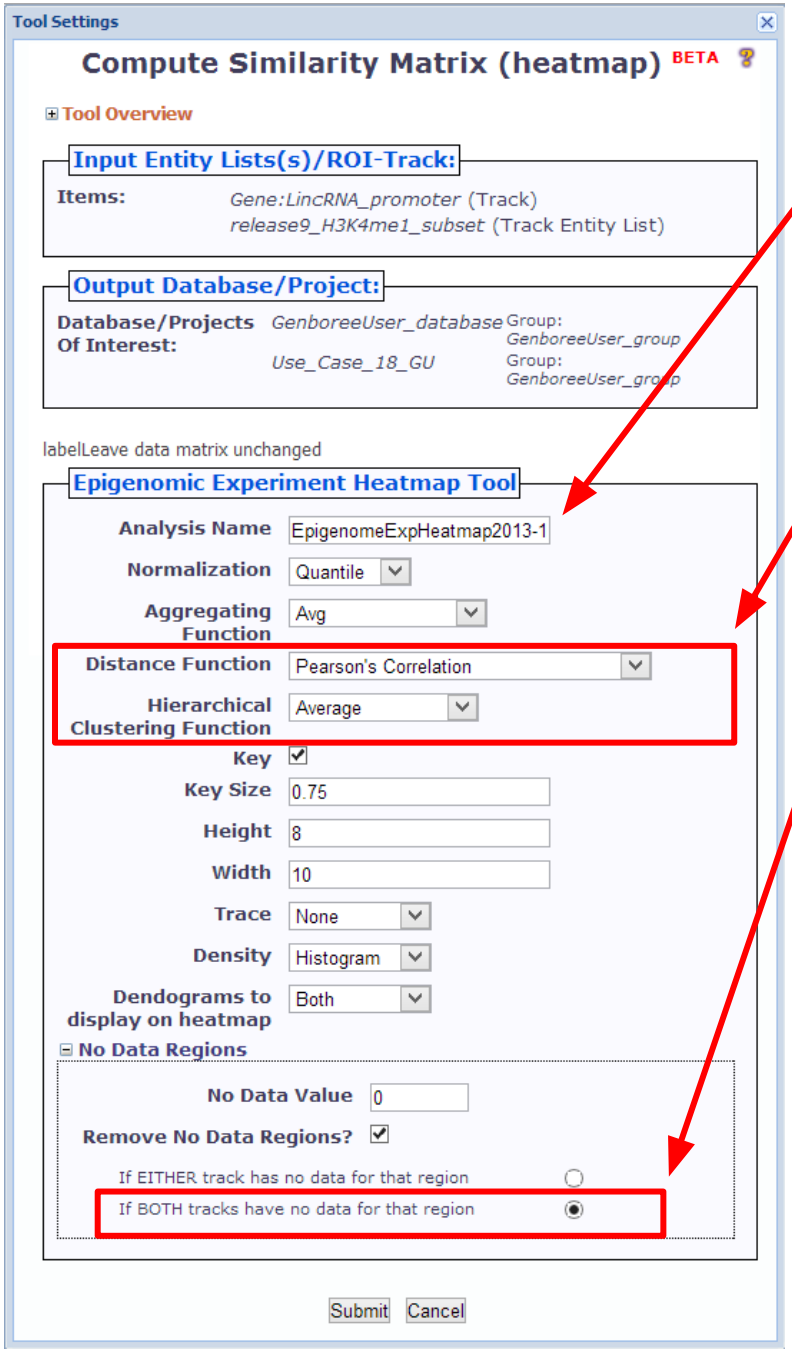




Rationale for Selection of H3K4me1 Histone Mark

Next, we will like to see how these different cell types/tissues cluster over different histone modification signal and region of interests. Here we use H3K4me1 signal tracks profiled in the NIH Roadmap Epigenome project and lincRNA promoter. H3K4me1 mark is associated with active enhancers. The lincRNA promoters were defined using Gencode annotations.


GENBOREE
BCM
Baylor College of Medicine
System/Network
Data
QC and Pre-processing
Genome
Transcriptome
Cistrome
Epigenome
Metagenome
Visualization
Help
Welcome to the Genboree Workbench!
[Getting Started]
Random Forest
QIIME
QC
Search for Similar Signals by Correlation
Analyze Signals
Compute Similarity Matrix (heatmap)
Create Track Lists from Newick Tree
Slice Epigenomic Data
Analyze Signals in the Context of Epigenome Atlas
Step 9
Data Selector
Refresh
Data Filter:
Select a filter...
Epigenome Informatics Workshop (May 2012)
Epigenome ToolSet Demo Input Data
Databases
ASHG 2013 Demo
All Annotations in Database
Tracks
Lists & Selections
Lists of Tracks
HL60_Tracks
HSC_Tracks
Immune_HL60_H3K4me1
Myeloid_Tracks
release9_H3K27me3
release9_H3K36me3
release9_H3K4me1
release9_H3K4me1_subset
release9_H3K4me3
release9_H3K4me3_subset
release9_H3K9me3
SampleSets
Samples
Files
Details
Attribute
Track 1
Track 4
Track 5
Input Data
Gene:LincRNA_promoter
release9_H3K4me1_subset
Output Targets
GenboreeUser_database
Use_Case_18_GU
Step 7 - Remove the previous track entity list and ROI from "Input Data"
Drag"Gene:LincRNA_promoter" track from ROI Repository. Drag "release9_H3K4me1_subset" track entity list to Input Data
Step 8 - Drag your database and project page to Output Targets
Genboree is built & maintained by the Bioinformatics Research Laboratory at Baylor College of Medicine.
Genboree is a hosted service. Code is available free for academic use.
BRL
HGSC

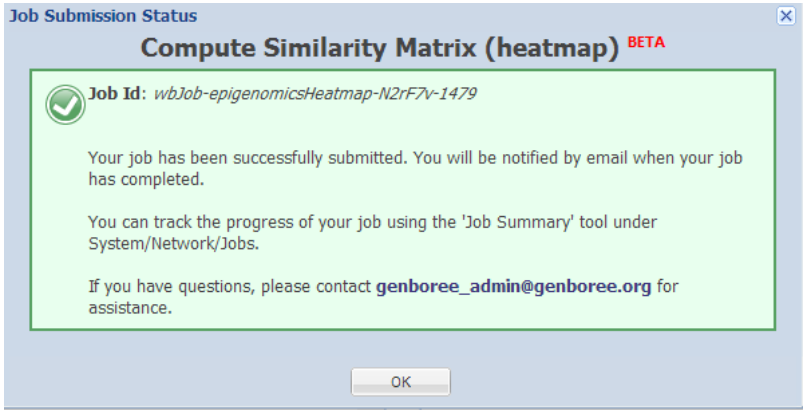


Step 10 - A default "Analysis Name" is generated by Genboree. It is recommended that all text and the time stamp be kept, and that you append some unique text to the beginning to help you distinguish different jobs run from the same tool.

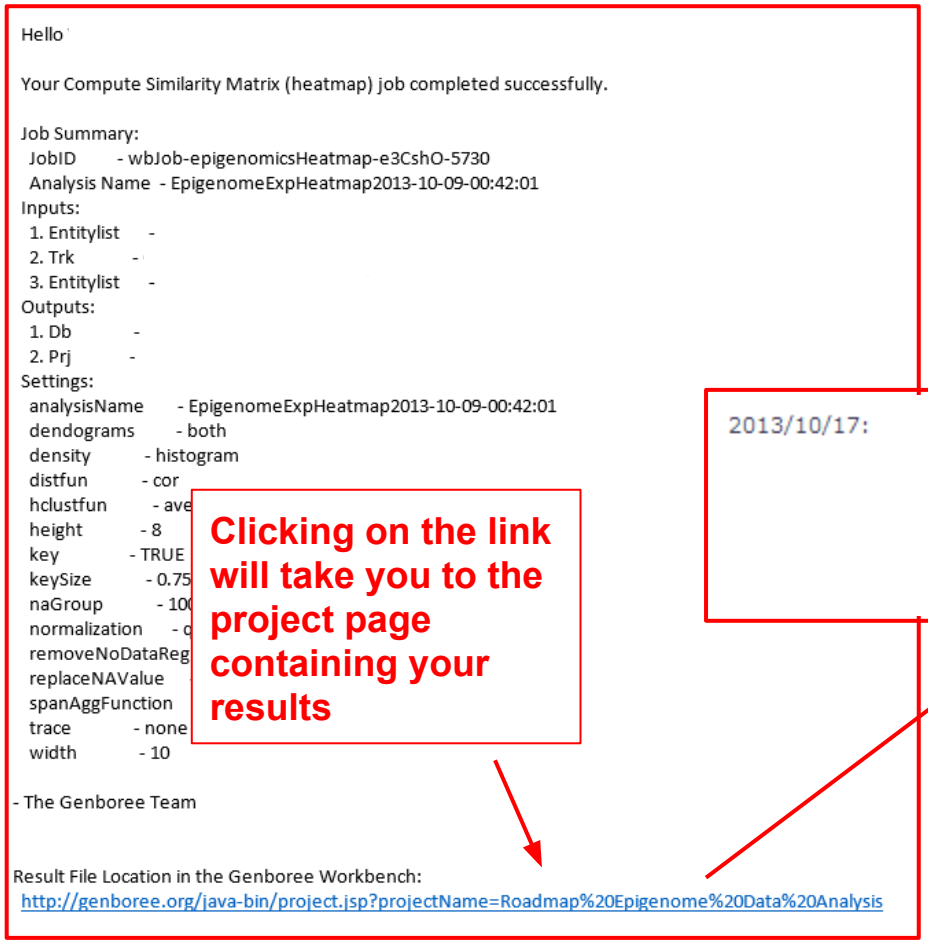
Step 11 - Select Pearson's Correlation (Absolute Value) as distance function and Average as hierarchical clustering function

Step 12 - Select to remove data if both tracks have no data for that region

You will see the message below upon successful submission of your heatmap job:



You will get following e-mail message when job is completed



Hello

Your Compute Similarity Matrix (heatmap) job completed successfully.

Job Summary:
JobID - wbJob-epigenomicsHeatmap-e3CshO-5730
Analysis Name - EpigenomeExpHeatmap2013-10-09-00:42:01
Inputs:
1. Entitylist -
2. Trk -
3. Entitylist -
Outputs:
1. Db -
2. Prj -
Settings:
analysisName - EpigenomeExpHeatmap2013-10-09-00:42:01
dendograms - both
density - histogram
distfun - cor
hclustfun - ave
height - 8
key - TRUE
keySize - 0.75
naGroup - 10
normalization - c
removeNoDataReg
replaceNAValue
spanAggFunction
trace - none
width - 10

- The Genboree Team

Result File Location in the Genboree Workbench:
http://genboree.org/java-bin/project.jsp?projectName=Roadmap%20Epigenome%20Data%20Analysis

Clicking on the link will take you to the project page containing your results

Project Page

2013/10/17: Genboree User ran Epigenomic Heatmap Tool (EpigenomeExpHeatmap2013-10-17-14 51 14) and the results are available at the link below.

- **Study Name**: EpigenomeExpHeatmap2013-10-17-14 51 14
- **Link to results**

Click on the “Link to results” in your Project page

Table of Content: Epigenomic HeatMap

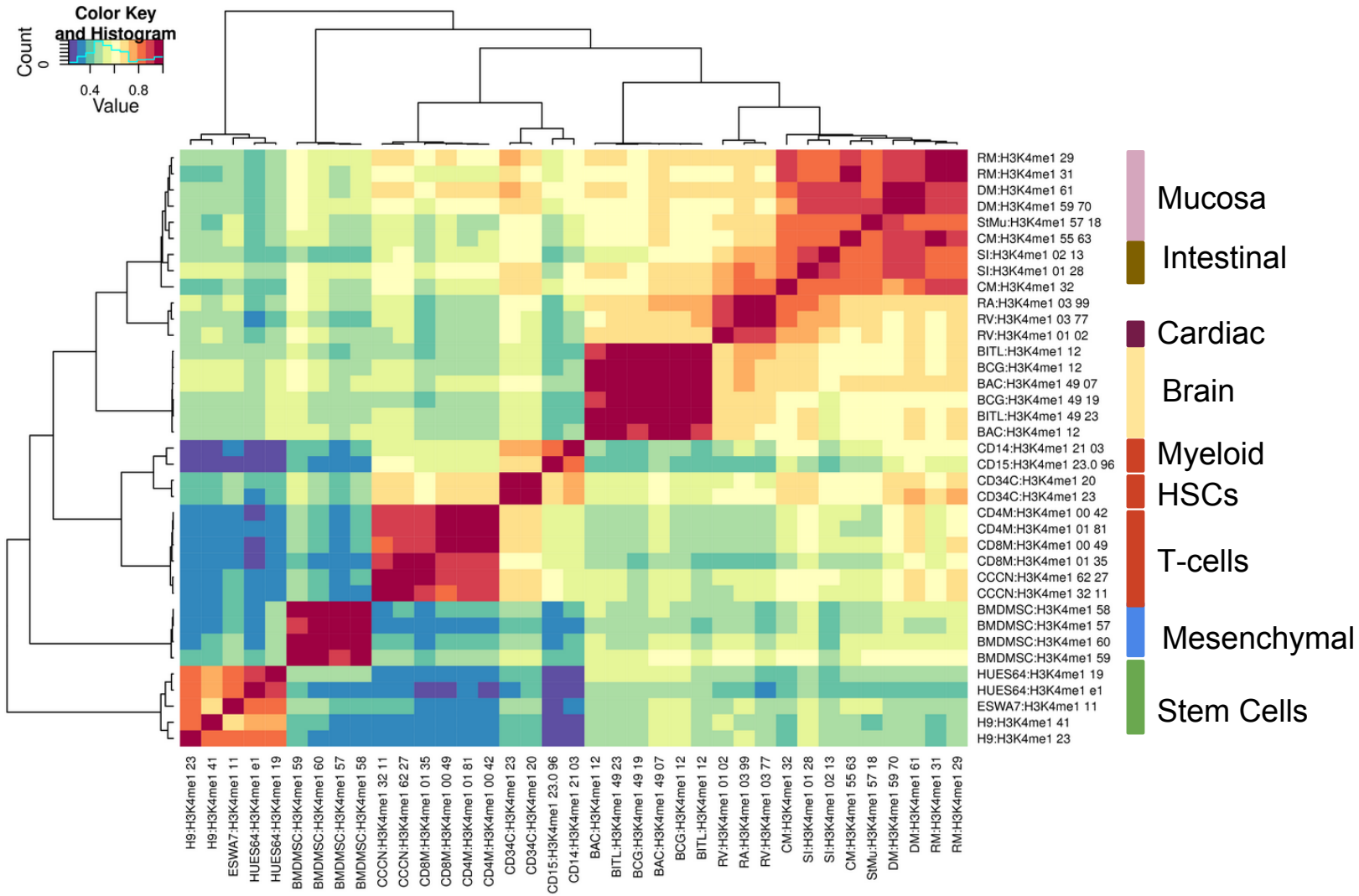
Study Name: lincRNA H3K4me1 2013-10-13-17 00 03
User: Genboree User
Date: 2013/10/13 18:10 CDT

Epigenomic HeatMap Plots

Newick Tree Visualizations

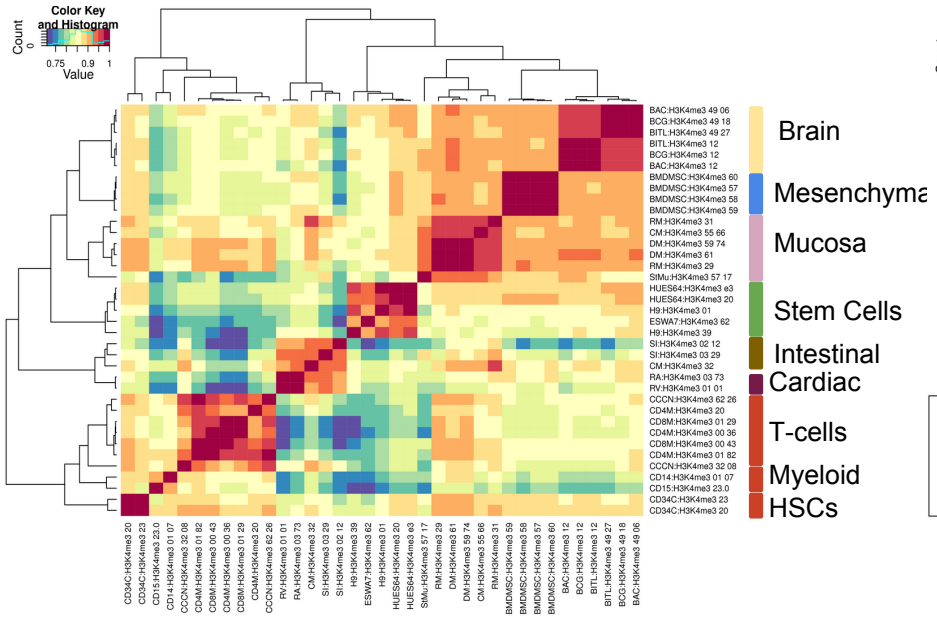
Click here to see circular tree of clustered epigenomes profiled in the NIH Roadmap Epigenome Project

Similar to H3K4me3 signal at protein-coding promoter region, H3K4me1 signal at lincRNA promoter region can also distinguish different cell types/tissues of origin and thus indicating role of lincRNA in cellular identity

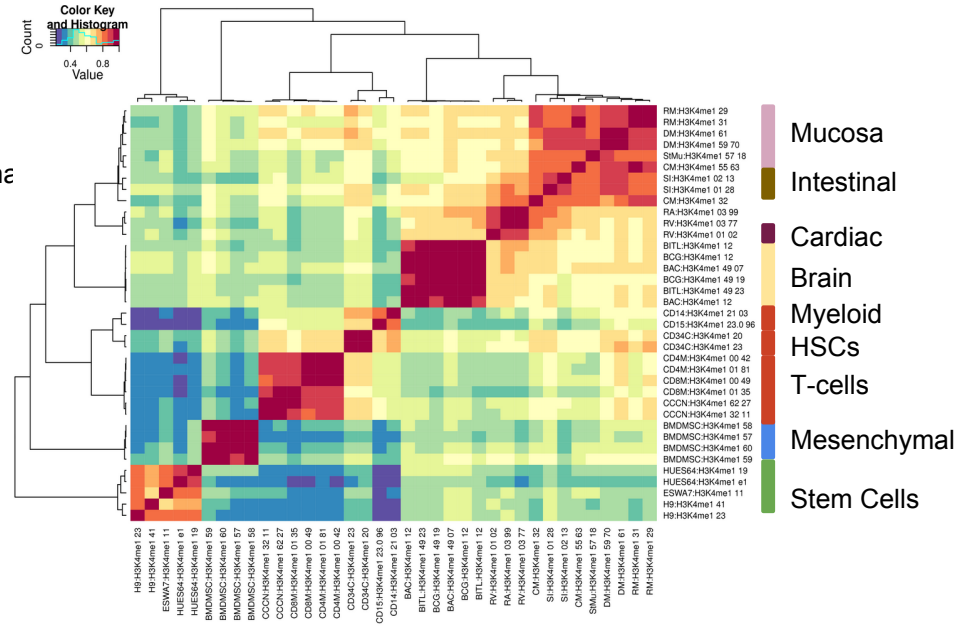


Summary of Results

H3K4me3 signal over protein coding gene promoters on the NIH Roadmap Epigenome data



H3K4me1 signal over lincRNA gene promoters on the NIH Roadmap Epigenome data



Cluster analysis of the epigenomes profiled in the NIH Roadmap Epigenome project suggests that lincRNAs play an important role in maintaining cellular identity.

If you are interested in learning about how Genboree tools can help identify regions that are specific for given lineage (Use case 19 supplementary slides) and how you can study regions that are undergoing epigenomic changes during cell differentiation, go on to use case 19.

Help us improve Genboree. Please provide a comment or request feature.