Use Case 2: Clustering of breast cell types (using the Epigenome Atlas)

Epigenome Informatics Workshop
Bioinformatics Research Laboratory

BCM Baylor College of Medicine

Reference and Credits

Reference:

Human Epigenome Atlas and the Genboree Epigenomic Toolset for Comparative Epigenome Analysis

Coarfa C[1], Harris RA[1], Jackson AR[1], Pichot CS[2], Raghuraman S[1], Paithankar S[1], Lee AV[3], McGuire SE[2], Milosavljevic A[1]

[1]NIH Roadmap Epigenomics Data Analysis and Coordination Center (EDACC), Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas. [2]Department of Molecular and Cellular Biology, Baylor College of Medicine, Houston, Texas. [3]Department of Pharmacology and Chemical Biology, University of Pittsburgh, Pittsburgh, Pennsylvania..

The NIH Roadmap Epigenomics Program Investigators' Meeting, May 14-15 2012, Bethesda, Maryland, USA.

Funding:

NIH Roadmap Epigenomics (NIH Common Fund)

Use Case 2: Clustering of breast cell types

Objective: To illustrate how one may use public datasets for comparative epigenomics.

Use Case #2 is similar to Use Case #1 in that the one objective is to evaluate the classification of samples based on differential methylation. Another key objective is to illustrate how one may execute integrative analysis using large public data repositories (the Epigenome Atlas in this instance).

The samples of interest here are breast luminal, breast myoepithelial, and breast stem cells.

Public repositories provide important data to which researchers can assess their own data by comparing methylation status and biological pathways of interest. In this use case, differential methylation based on MeDIP signals will be used to differentiate different breast samples from the Epigenome Roadmap Initiative.

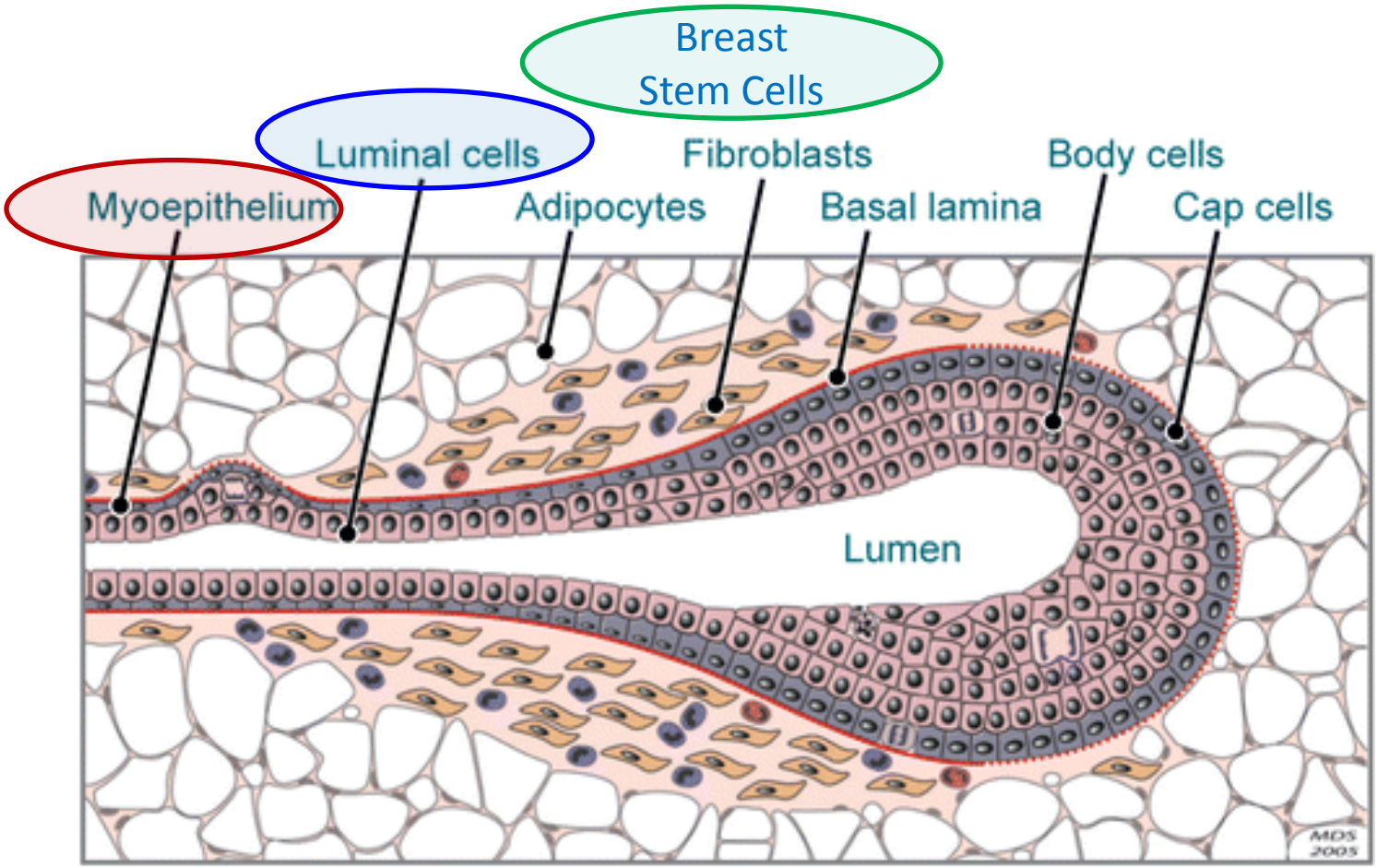
Promoter DNA Methylation in the Human Genome

Source of ROIs found in "Class: Regulation" in the Data Selector

- Enriched methylated DNA from human primary fibroblasts using methylated DNA immunoprecipitation (MeDIP) + microarray detection
- 15,609 promoters evaluated in primary somatic and germline cells
- **HCPs** (high-CpG promoters) – contain 500 bp region with CpG ratio above 0.75 and GC content >55%
- **LCP** (low-CpG promoters) – do not contain a 500 bp region with a CpG ratio above 0.48
- **ICP** (intermediate CpG promoters) – are neither HCPs or LCPs. ICP class contains many "subthreshold" CpG islands, meaning small CpG islands (<500 bp), moderate CpG richness and/or GC content <55%

Weber et al, "Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome" *Nature Genetics*, 39 (4), April 2007

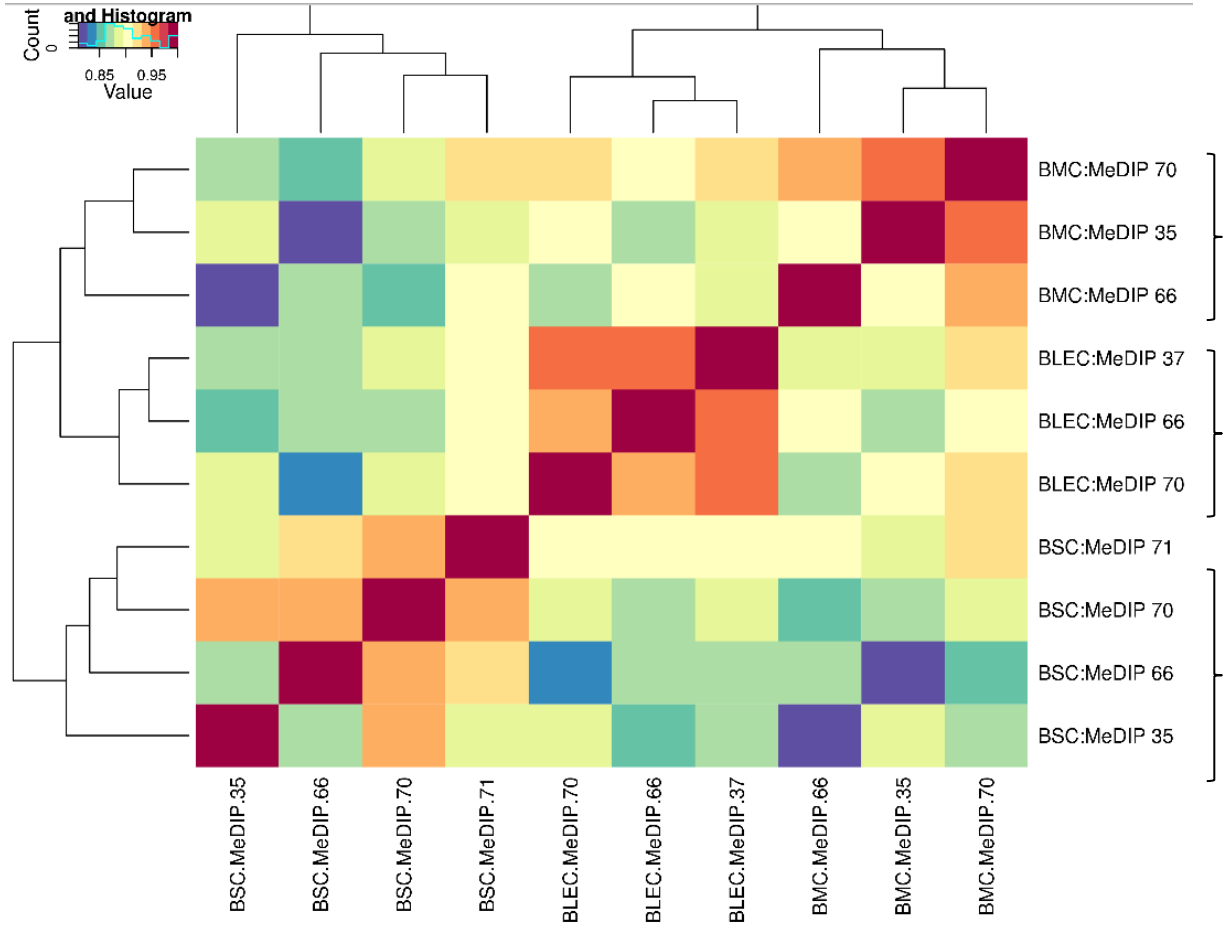
Breast Cell Type Differentiation



Hebner C, et al. 2008.
Annu. Rev. Pathol. Mech. Dis. 3:313–39

Use Case 2 Results: Breast Cell Types Cluster Based on MeDIP Profile

(Epigenome Atlas and UCSF REMC data)



Breast myoepithelial cells

Breast luminal epithelial cells

Breast stem cells

Data from: Epigenome Atlas, Release 5

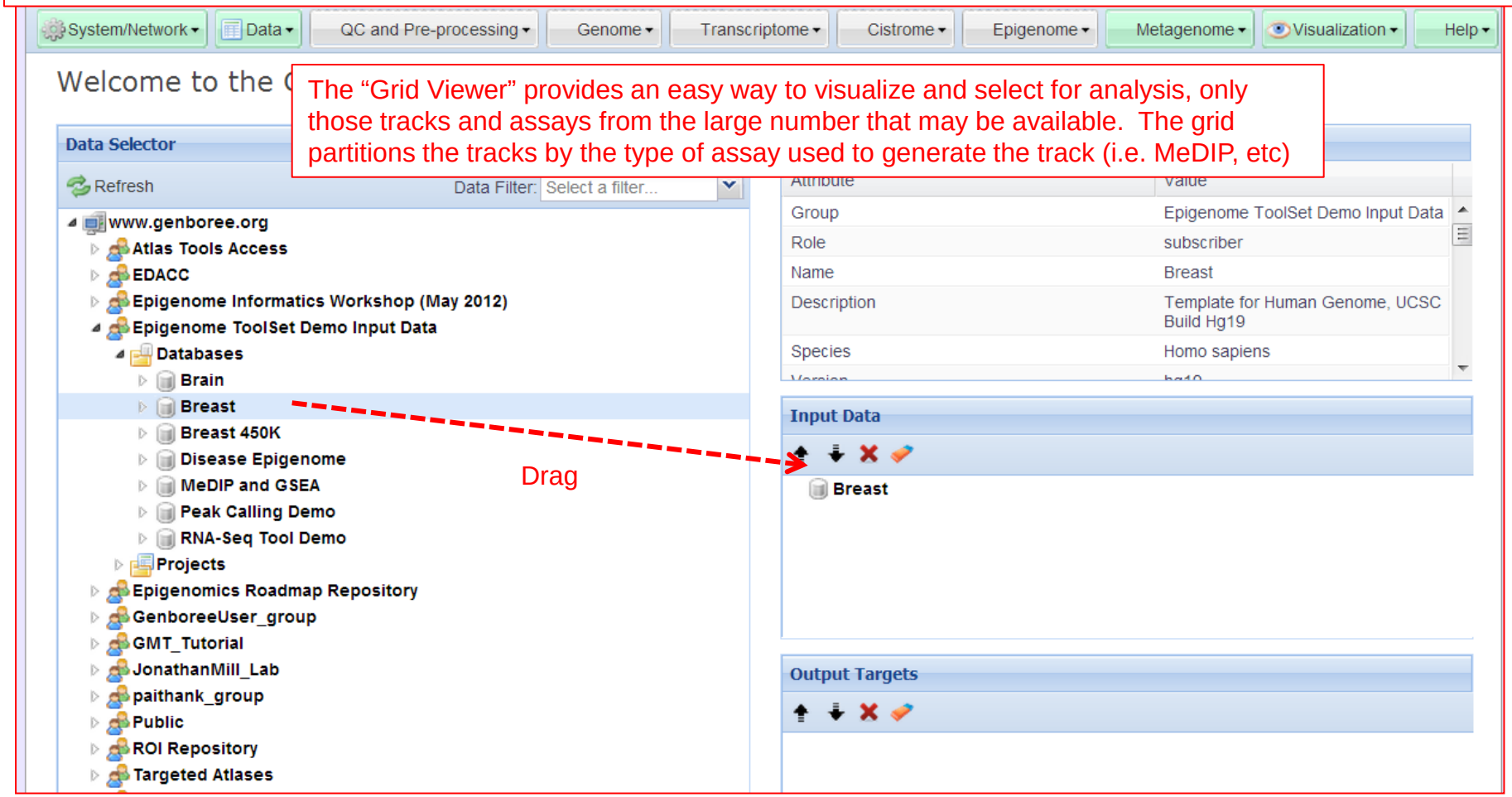
The following slides walk you through the process of generating the output displayed in the previous slide.

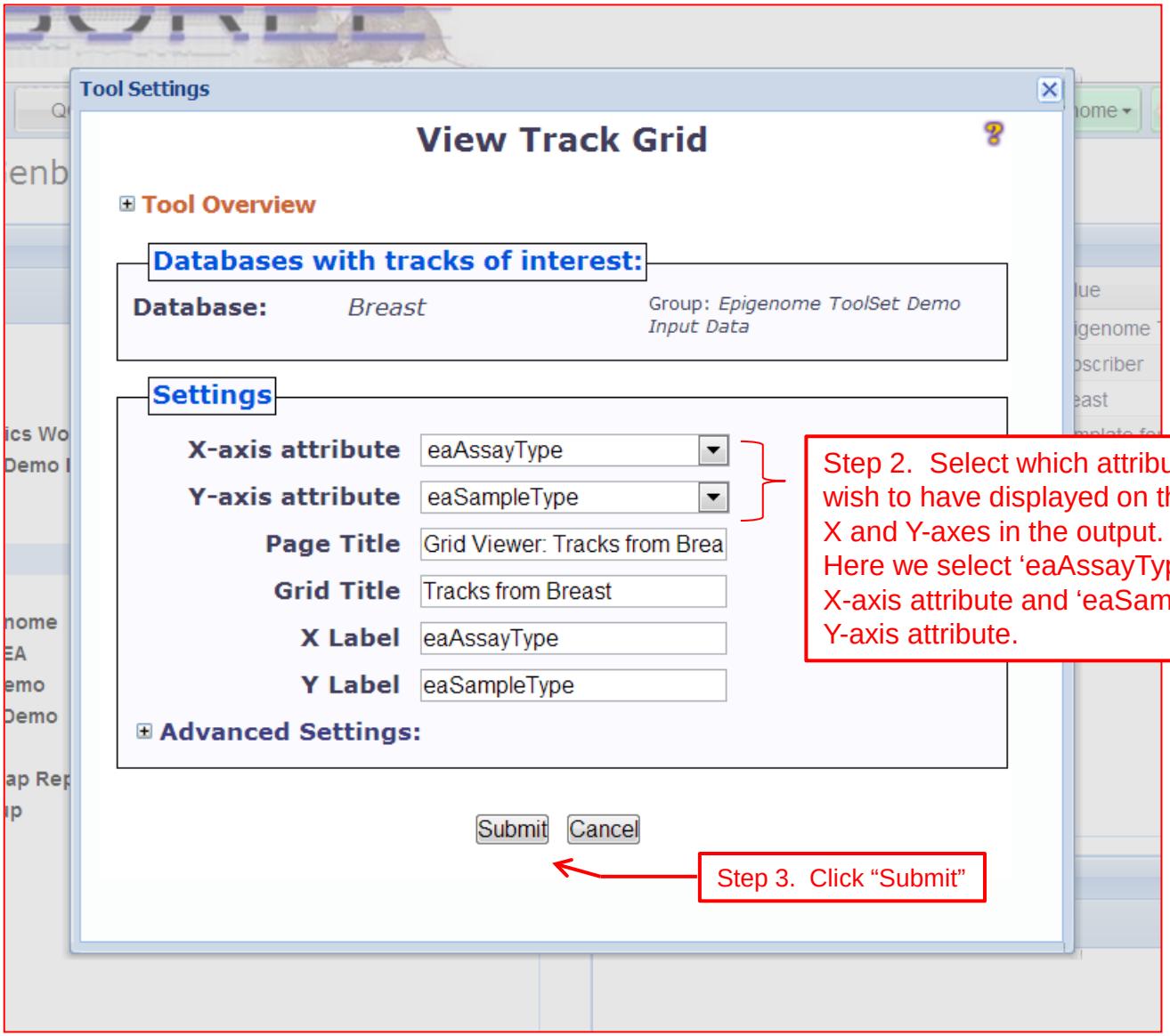
Since you already created a Project and Database in Use Case 1, you will not need to do that again. The results of this analysis will be part of the same Project and be deposited in the same Database.

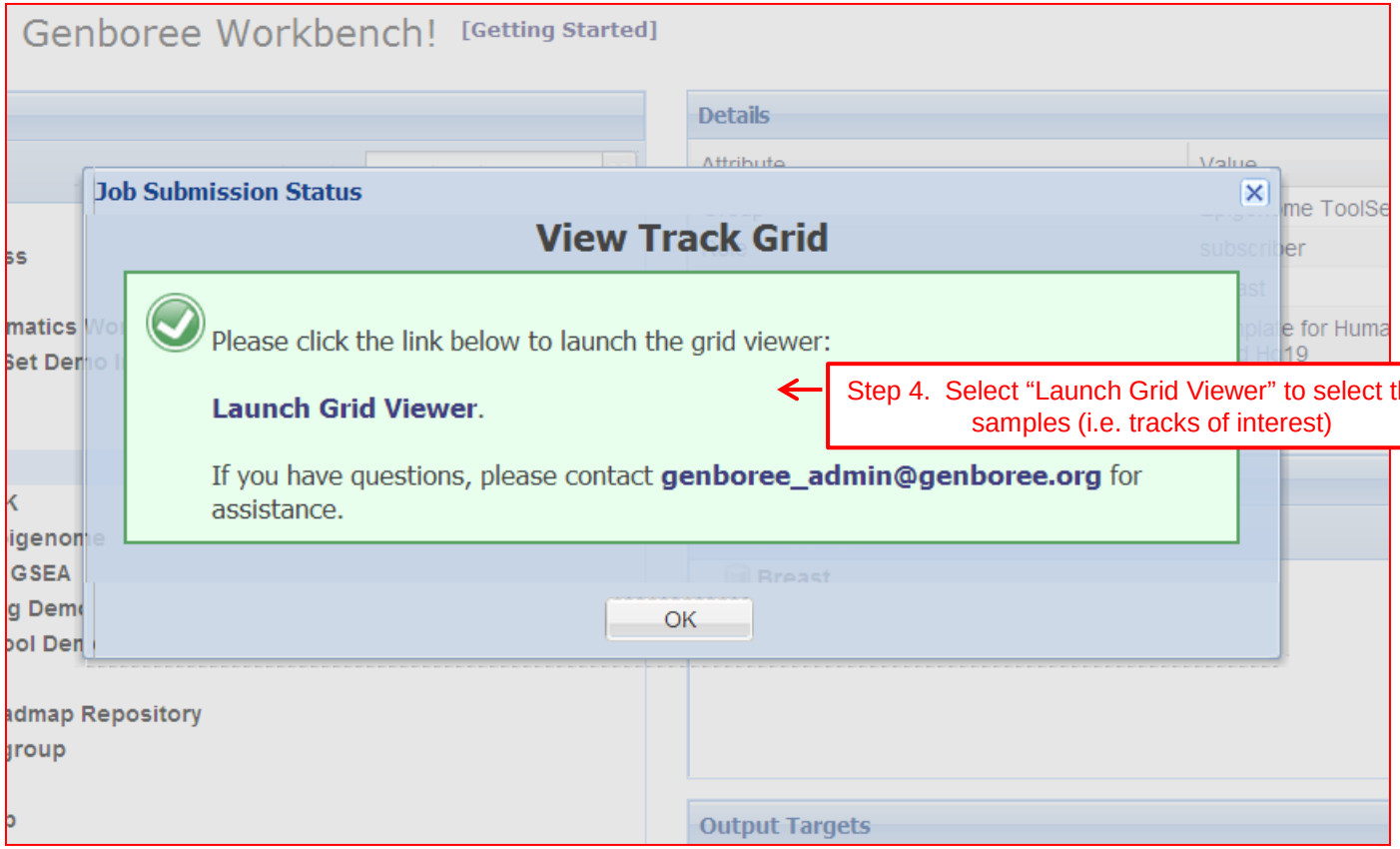
The next step is to select the samples to analyze.

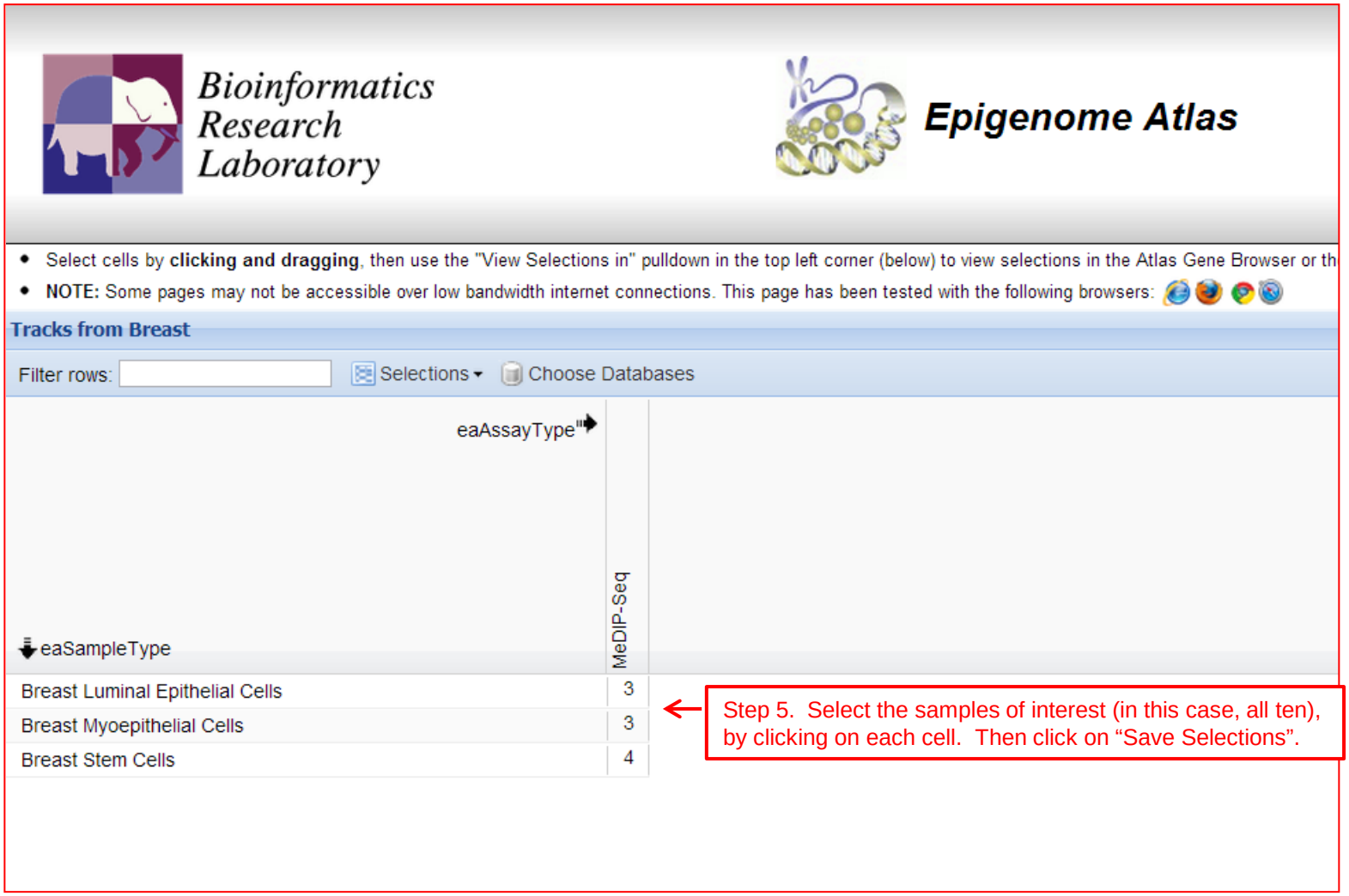
Step 1. Drag “Breast” database into the “Input Data” box.
This will cause the “Visualization” menu to turn green, meaning a tool(s) within that menu is active. A tool is acti
when “Input Data” and “Output Targets” have been populated with the appropriate data/tracks/files required for that tool to operate. Here, you will be using the “View Track Grid” tool.

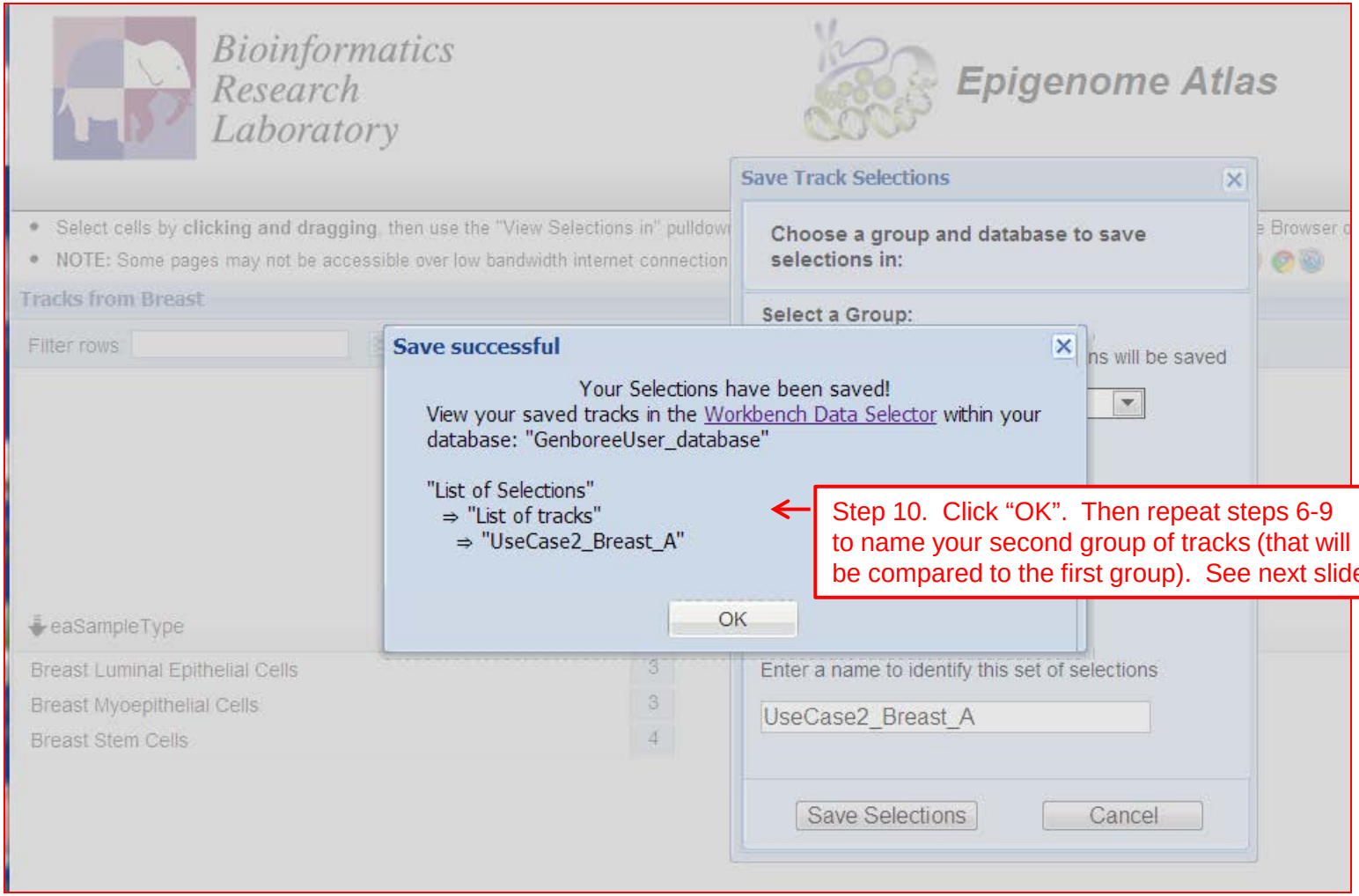
-Click ‘Visualization’ and then ‘View Track Grid’




Tool Settings
View Track Grid
Tool Overview
Databases with tracks of interest:
Database: Breast
Group: Epigenome ToolSet Demo Input Data
Settings
X-axis attribute eaAssayType
Y-axis attribute eaSampleType
Page Title Grid Viewer: Tracks from Brea
Grid Title Tracks from Breast
X Label eaAssayType
Y Label eaSampleType
Advanced Settings:
Submit
Cancel
Step 2. Select which attributes you wish to have displayed on the X and Y-axes in the output. Here we select 'eaAssayType' for the X-axis attribute and 'eaSampleType' for the Y-axis attribute.
Step 3. Click "Submit"

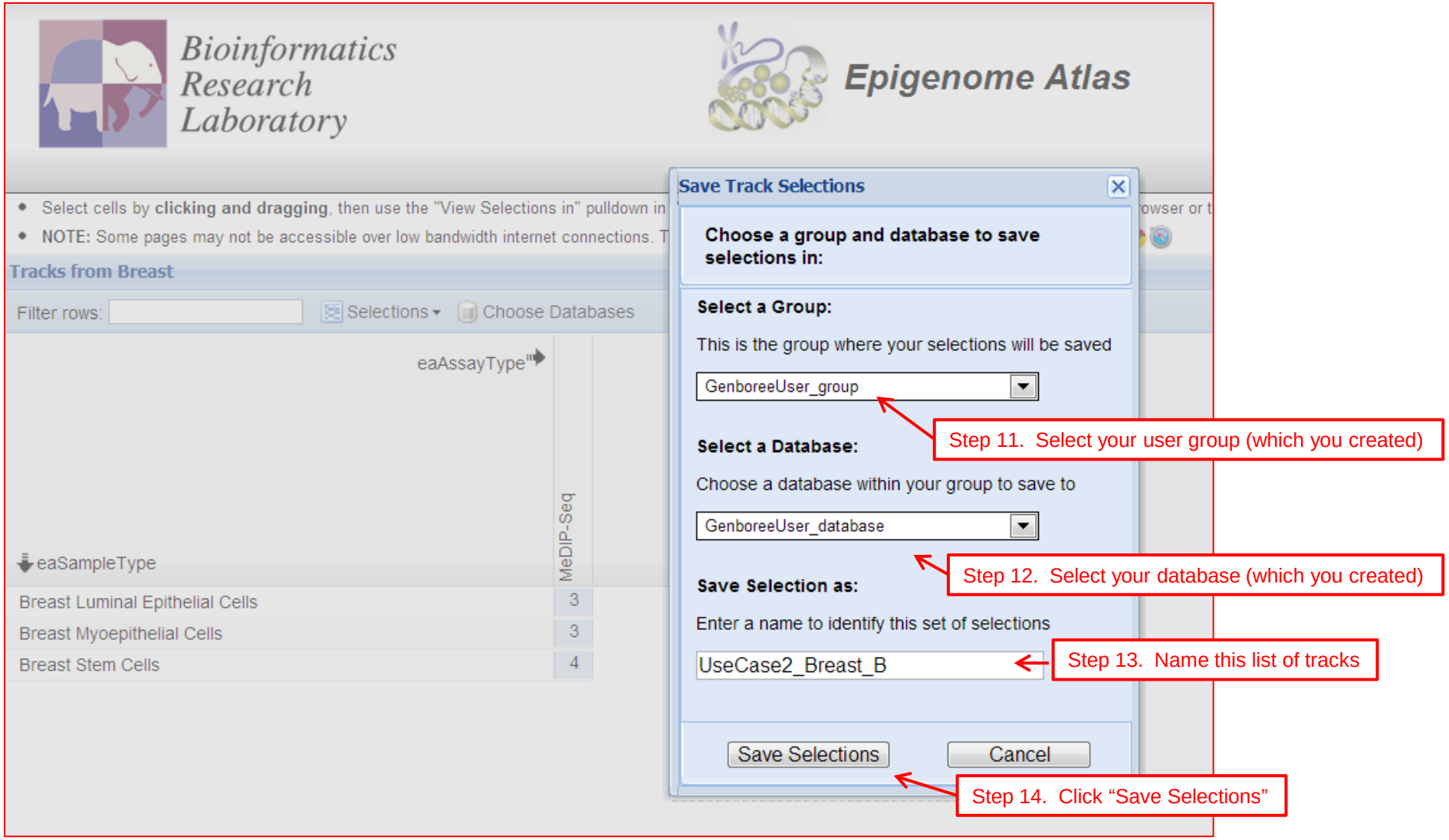

Genboree Workbench! [Getting Started]
Details
Job Submission Status
View Track Grid
Please click the link below to launch the grid viewer:
Launch Grid Viewer.
If you have questions, please contact genboree_admin@genboree.org for assistance.
OK
Step 4. Select "Launch Grid Viewer" to select the samples (i.e. tracks of interest)
Output Targets

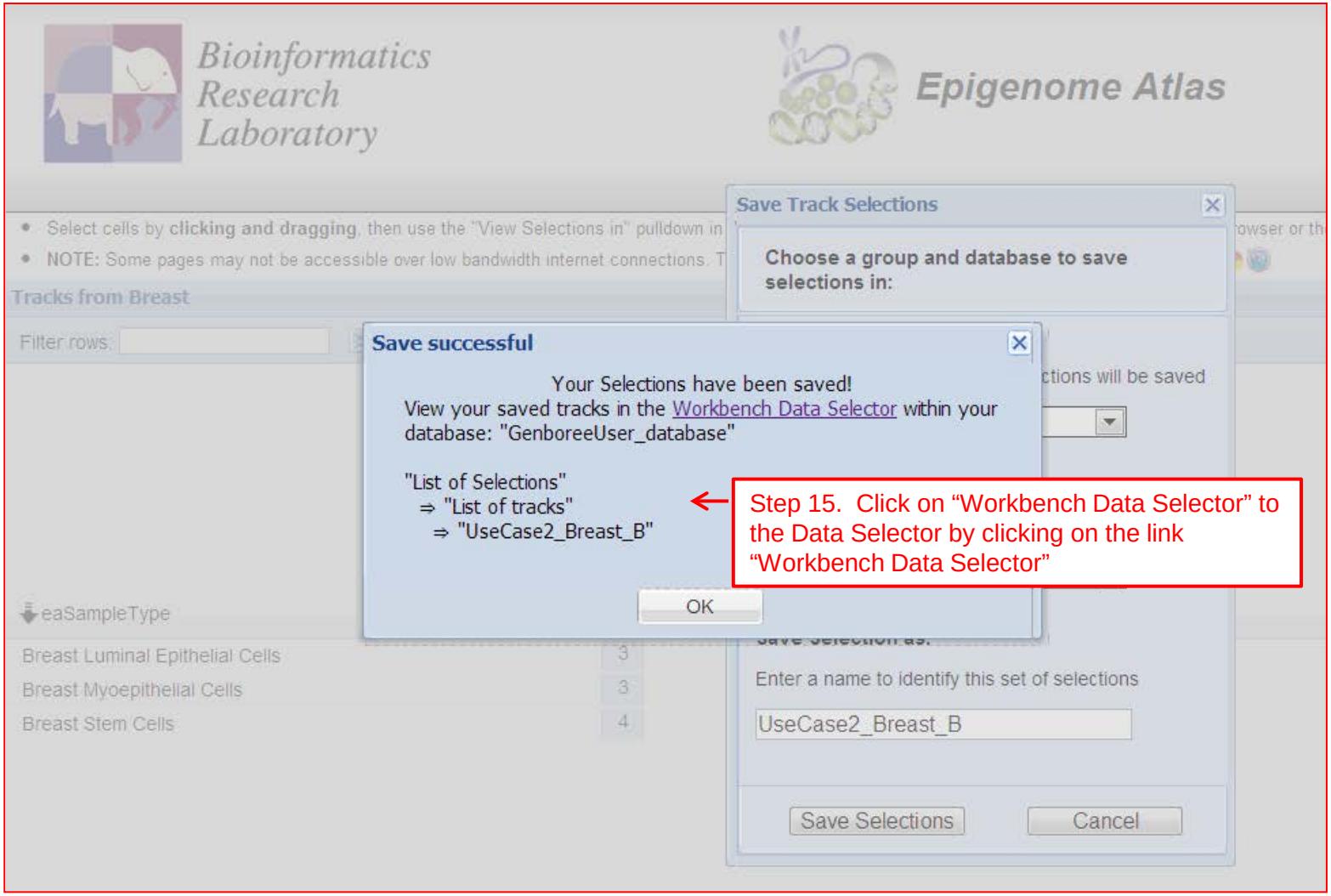

Bioinformatics Research Laboratory
Epigenome Atlas
Select cells by clicking and dragging, then use the "View Selections in" pulldown in the top left corner (below) to view selections in the Atlas Gene Browser or th
NOTE: Some pages may not be accessible over low bandwidth internet connections. This page has been tested with the following browsers:
Tracks from Breast
Filter rows:
Selections
Choose Databases
eaAssayType
eaSampleType
MeDIP-Seq
Breast Luminal Epithelial Cells 3
Breast Myoepithelial Cells 3
Breast Stem Cells 4
Step 5. Select the samples of interest (in this case, all ten), by clicking on each cell. Then click on "Save Selections".

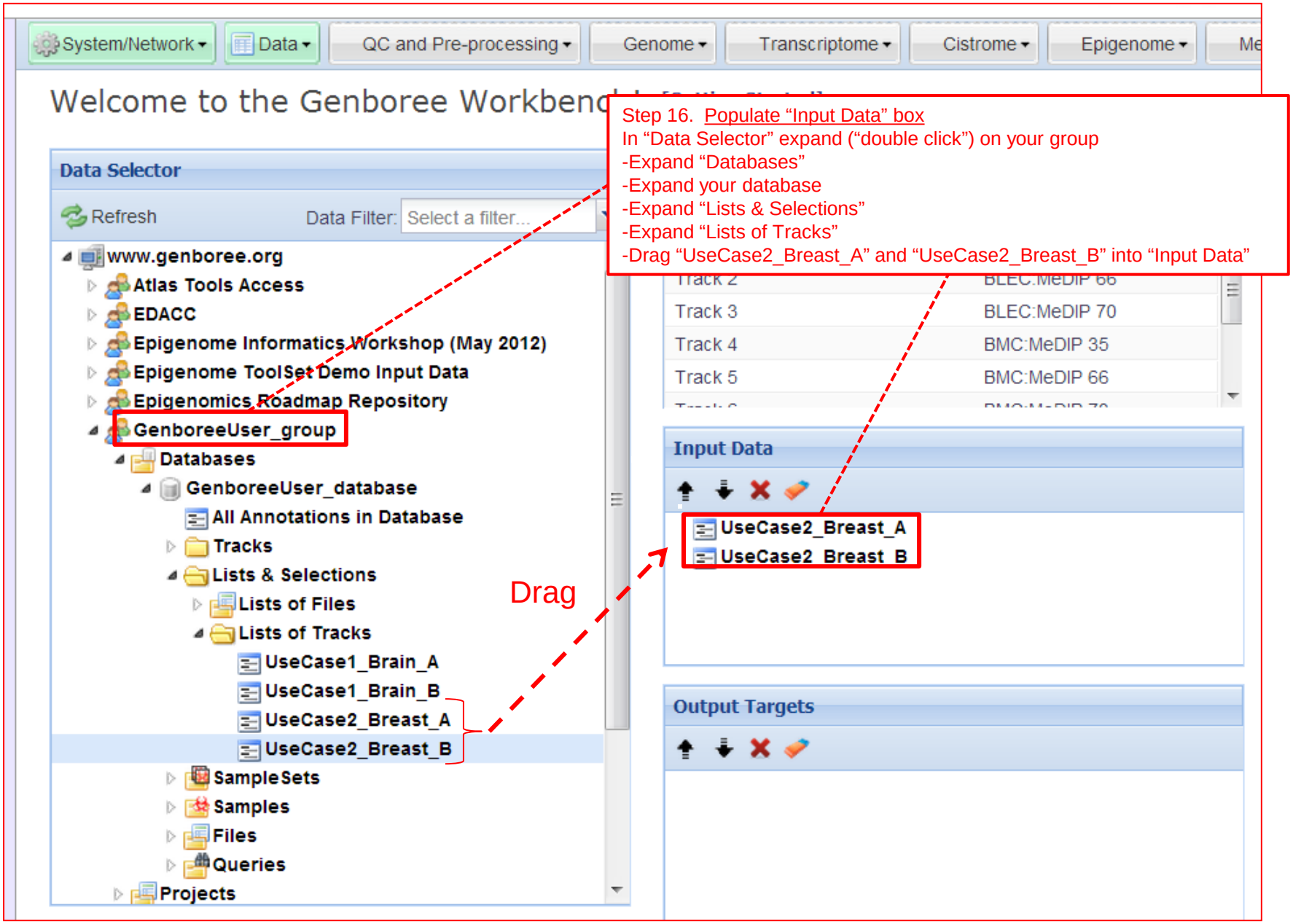


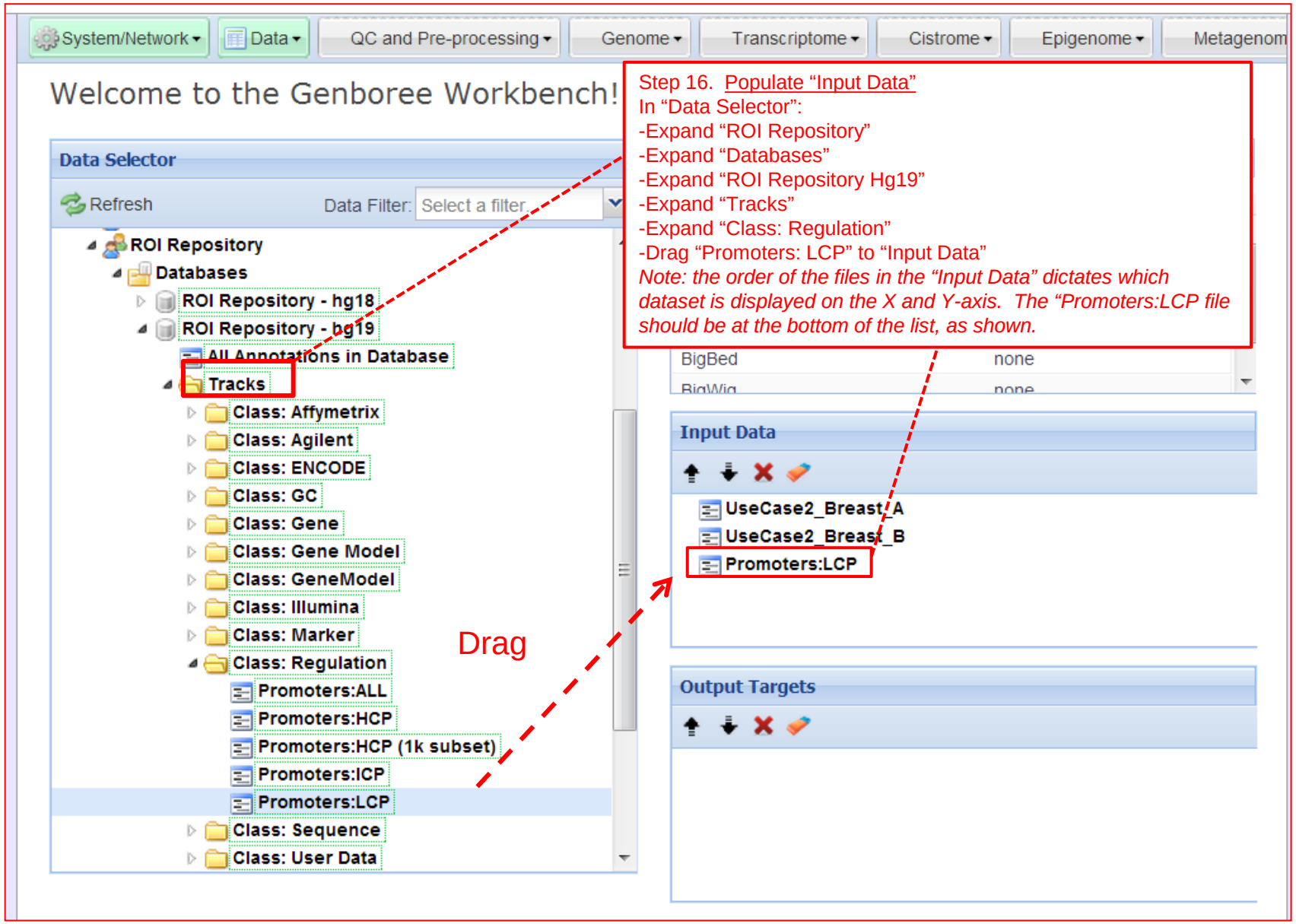
Step 10. Click "OK". Then repeat steps 6-9 to name your second group of tracks (that will be compared to the first group). See next slide.

Steps 6-10 are repeated here, but with the same set of tracks given a different name ("UseCase2_B"). We are using the same set of tracks for illustration purposes.




Bioinformatics Research Laboratory
Epigenome Atlas
Save Track Selections
Choose a group and database to save selections in:
Select cells by clicking and dragging, then use the "View Selections in" pulldown in
NOTE: Some pages may not be accessible over low bandwidth internet connections.
Tracks from Breast
Filter rows:
Save successful
Your Selections have been saved!
View your saved tracks in the Workbench Data Selector within your database: "GenboreeUser_database"
"List of Selections"
⇒ "List of tracks"
⇒ "UseCase2_Breast_B"
OK
Step 15. Click on "Workbench Data Selector" to the Data Selector by clicking on the link "Workbench Data Selector"
eaSampleType
Breast Luminal Epithelial Cells 3
Breast Myoepithelial Cells 3
Breast Stem Cells 4
Enter a name to identify this set of selections
UseCase2_Breast_B
Save Selections
Cancel


System/Network
Data
QC and Pre-processing
Genome
Transcriptome
Cistrome
Epigenome
Welcome to the Genboree Workbench
Step 16. Populate "Input Data" box
In "Data Selector" expand ("double click") on your group
-Expand "Databases"
-Expand your database
-Expand "Lists & Selections"
-Expand "Lists of Tracks"
-Drag "UseCase2_Breast_A" and "UseCase2_Breast_B" into "Input Data"
Data Selector
Refresh
Data Filter: Select a filter...
www.genboree.org
Atlas Tools Access
EDACC
Epigenome Informatics Workshop (May 2012)
Epigenome ToolSet Demo Input Data
Epigenomics Roadmap Repository
GenboreeUser_group
Databases
GenboreeUser_database
All Annotations in Database
Tracks
Lists & Selections
Lists of Files
Lists of Tracks
UseCase1_Brain_A
UseCase1_Brain_B
UseCase2_Breast_A
UseCase2_Breast_B
SampleSets
Samples
Files
Queries
Projects
Drag
Track 3
BLEC:MeDIP 70
Track 4
BMC:MeDIP 35
Track 5
BMC:MeDIP 66
Input Data
UseCase2_Breast_A
UseCase2_Breast_B
Output Targets


System/Network
Data
QC and Pre-processing
Genome
Transcriptome
Cistrome
Epigenome
Metagenom
Welcome to the Genboree Workbench!
Step 16. Populate "Input Data"
In "Data Selector":
-Expand "ROI Repository"
-Expand "Databases"
-Expand "ROI Repository Hg19"
-Expand "Tracks"
-Expand "Class: Regulation"
-Drag "Promoters: LCP" to "Input Data"
Note: the order of the files in the "Input Data" dictates which dataset is displayed on the X and Y-axis. The "Promoters:LCP file should be at the bottom of the list, as shown.
Data Selector
Refresh
Data Filter: Select a filter
ROI Repository
Databases
ROI Repository - hg18
ROI Repository - hg19
All Annotations in Database
Tracks
Class: Affymetrix
Class: Agilent
Class: ENCODE
Class: GC
Class: Gene
Class: Gene Model
Class: GeneModel
Class: Illumina
Class: Marker
Class: Regulation
Promoters:ALL
Promoters:HCP
Promoters:HCP (1k subset)
Promoters:ICP
Promoters:LCP
Class: Sequence
Class: User Data
Drag
BigBed
none
Input Data
UseCase2_Breast_A
UseCase2_Breast_B
Promoters:LCP
Output Targets

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

Welcome to the Genboree Workbench! [Getting Started]

Data Selector

Refresh Data Filter: Select a filter...

- www.genboree.org
  - Atlas Tools Access
  - EDACC
  - Epigenome Informatics Workshop (May 2012)
  - Epigenome ToolSet Demo Input Data
  - Epigenomics Roadmap Repository
  - GenboreeUser_group
    - Databases
      - GenboreeUser_database
    - 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
  - GMT_Tutorial
  - JonathanMill_Lab
  - paithank_group
  - Public

Details

Attribute	Value
View Link	Link to Project

Step 17. Populate "Output Targets"
In "Data Selector" expand ("double click") on your user group
-Expand "Databases"
-Drag your destination database to "Output Targets"
-Expand "Projects"
-Drag your project to "Output Targets"

Input Data

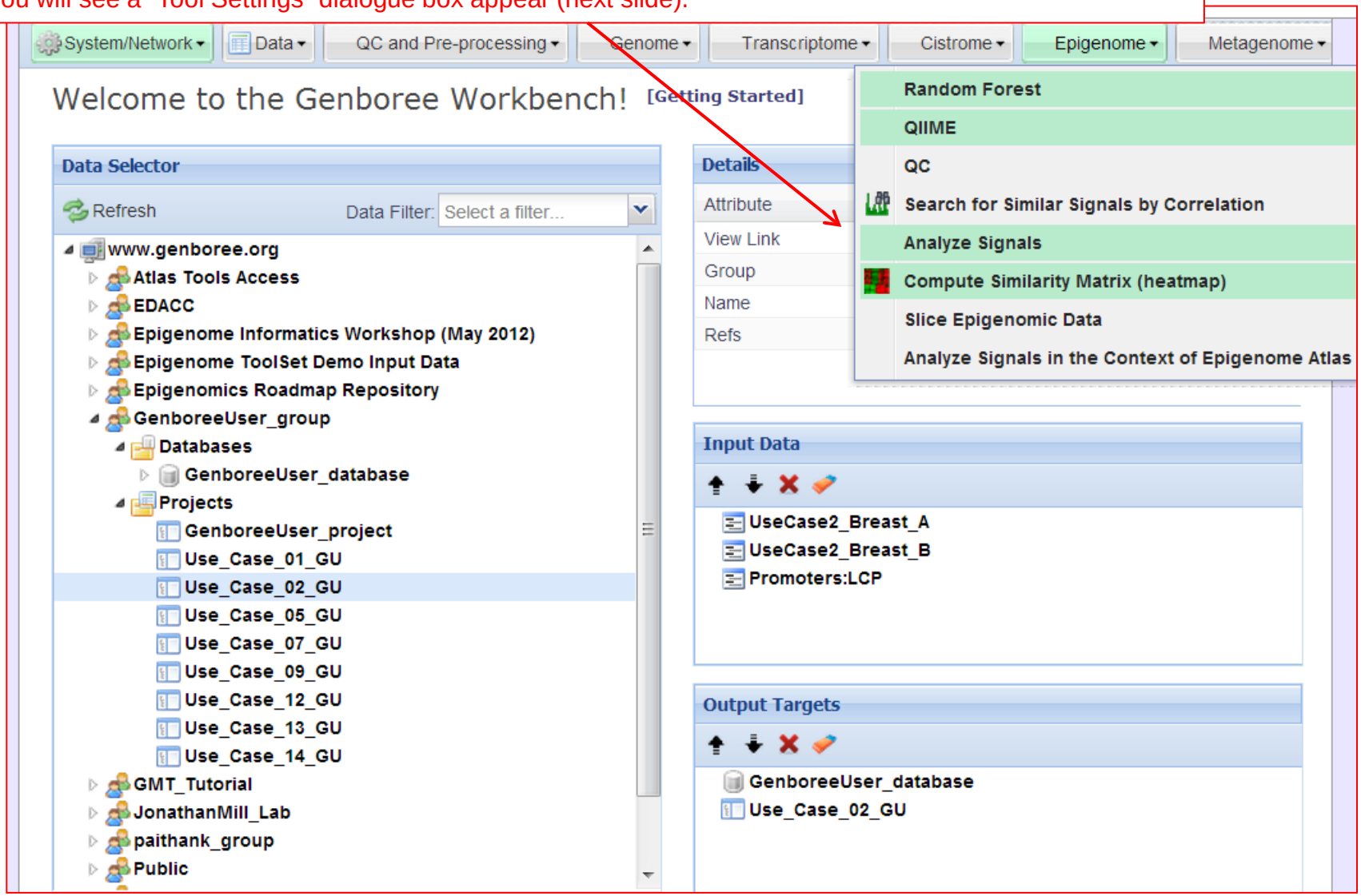
- UseCase2_Breast_A
- UseCase2_Breast_B
- Promoters:LCP

Output Targets

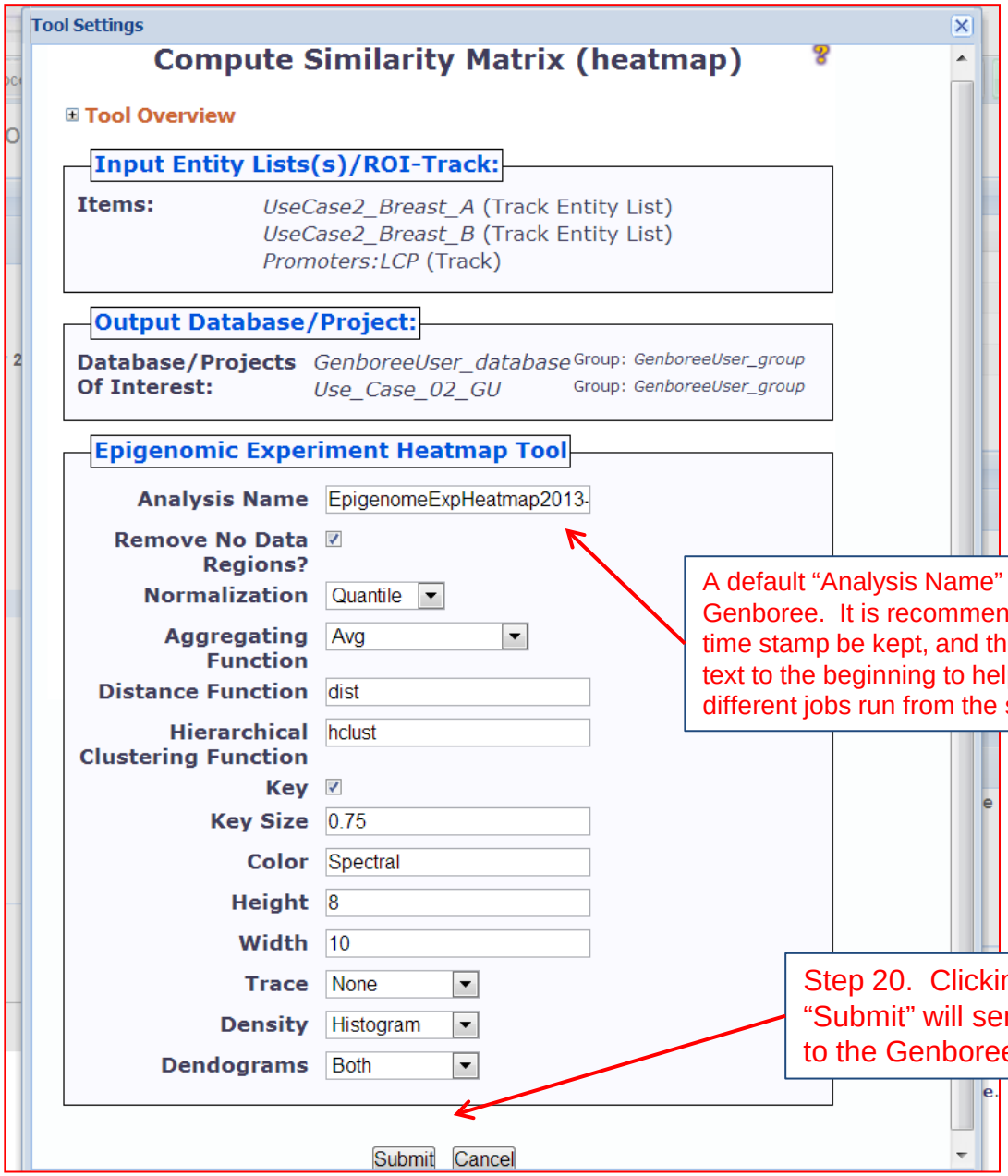
- GenboreeUser_database
- Use_Case_02_GU

Note the “Epigenome” menu turns green when “Input Data” and “Output Targets” are properly populated.

Step 18. Click on “Epigenome”
-Click on “Compute Similarity Matrix (heatmap)”
You will see a “Tool Settings” dialogue box appear (next slide).



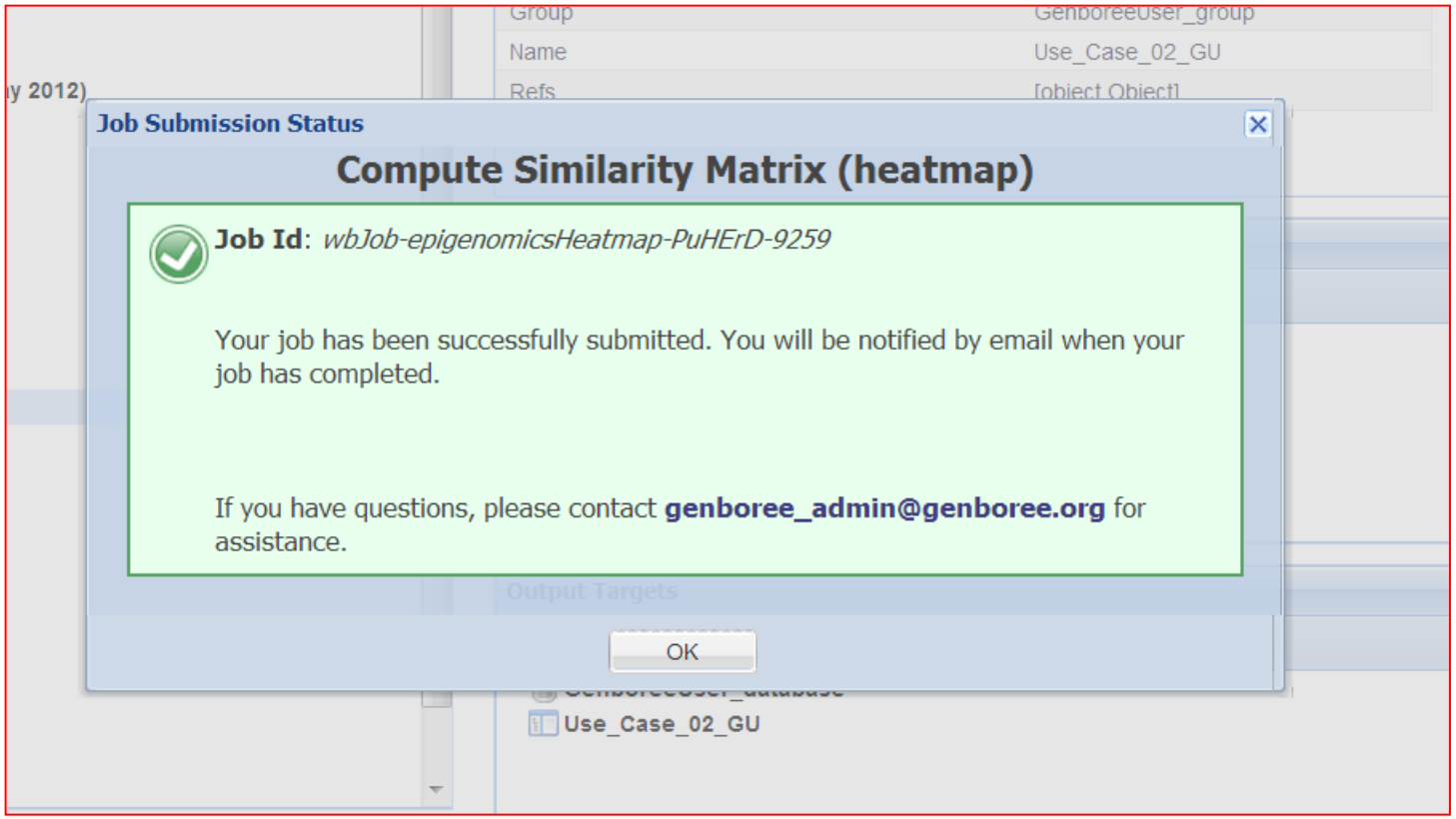
Step 19. Check that the “Input Files Directory” and “Output Database” and “Project” are correct (based on what you named them). Use the default parameters to begin with, and experiment with changing the parameters in subsequent jobs.



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

Step 20. Clicking on “Submit” will send the job to the Genboree cluster.

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



You will receive an email with the following message when your job is finished:

Hello Genboree User,

Your job completed successfully.

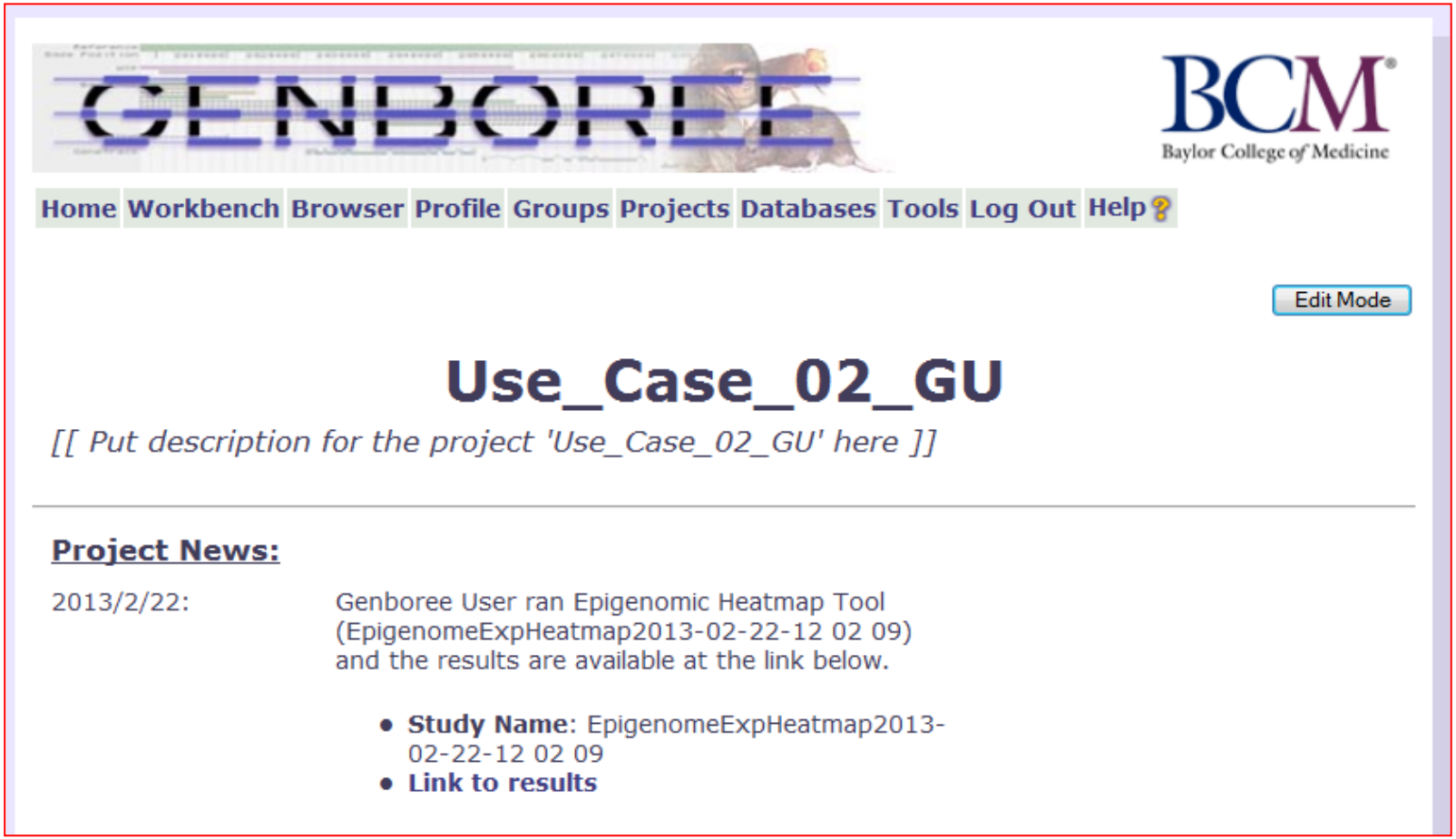
Job Summary:
JobID - wbJob-epigenomicsHeatmap-PuHErD-9259
Analysis Name - EpigenomeExpHeatmap2013-02-22-12:02:09
Inputs:
1. Entitylist - UseCase2_Breast_A
2. Entitylist - UseCase2_Breast_B
3. Trk - Promoters%3ALCP
Outputs:
1. Db - GenboreeUser_database
2. Prj - Use_Case_02_GU
Settings:
analysisName - EpigenomeExpHeatmap2013-02-22-12:02:09
color - Spectral
dendograms - both
density - histogram
distfun - dist
hclustfun - hclust
height - 8
key - TRUE
keySize - 0.75
normalization - quant
quantileNormalized - false
removeNoDataRegions - true
spanAggFunction - avg
trace - none
width - 10

- The Genboree Team

Result File Location in the Genboree Workbench:
http://www.genboree.org/java-bin/project.jsp?projectName=Use_Case_02_GU

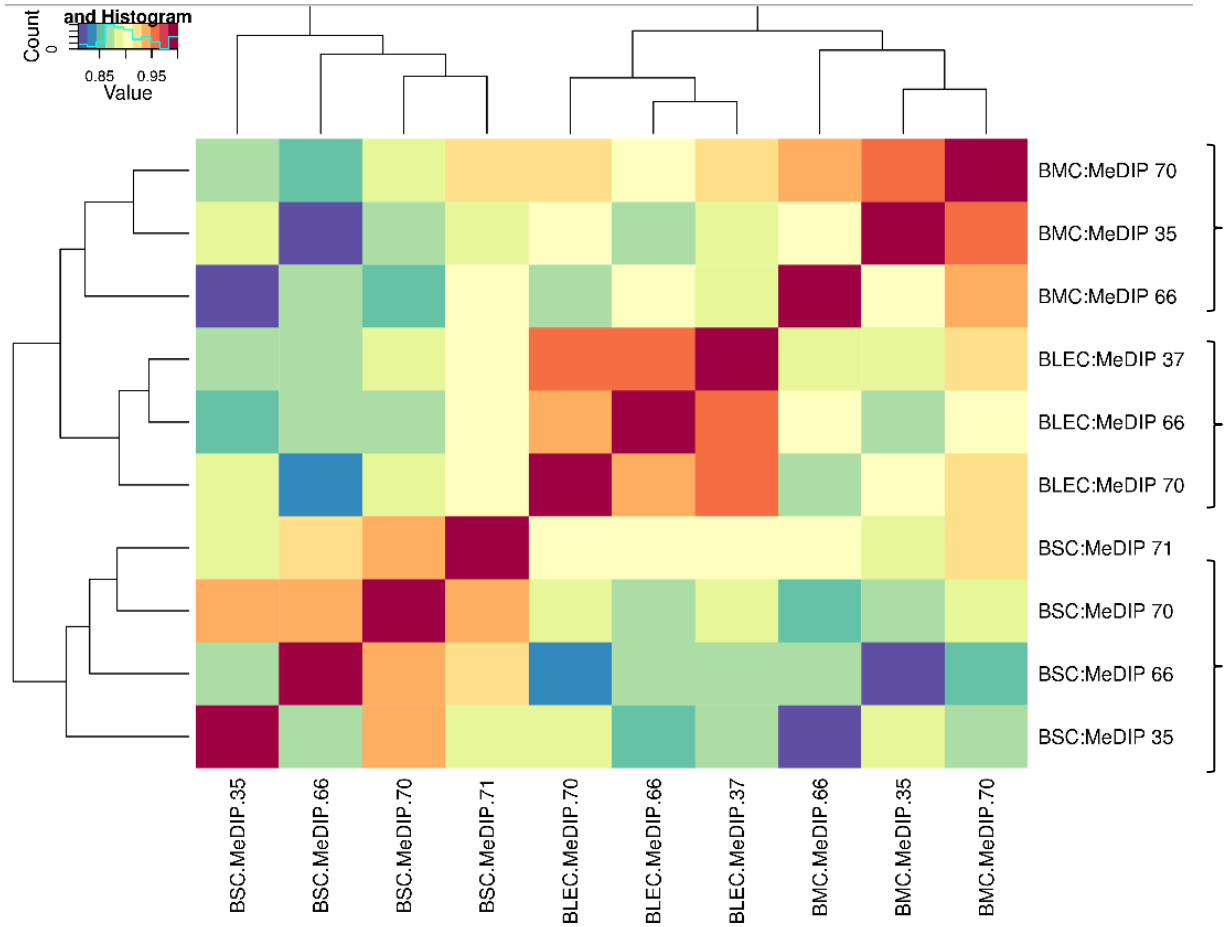
Clicking on the link will take you to the project page containing your results.

The Genboree Project Page



Use Case 2 Results: Breast Cell Types Cluster Based on MeDIP Profile

(Epigenome Atlas and UCSF REMC data)



Breast myoepithelial cells

Breast luminal cells

Breast stem cells

Data from: Epigenome Atlas, Release 5