

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

Epigenome Informatics Workshop
Bioinformatics Research Laboratory



Reference and Credits

Reference:

Human Epigenome Atlas and the Genboree Epigenomic Toolset for Comparative Epigenome Analysis
Coarfa C¹, Harris RA¹, Jackson AR¹, Pichot CS², Raghuraman S¹, Paithankar S¹, Lee AV³, McGuire SE², Milosavljevic A¹

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

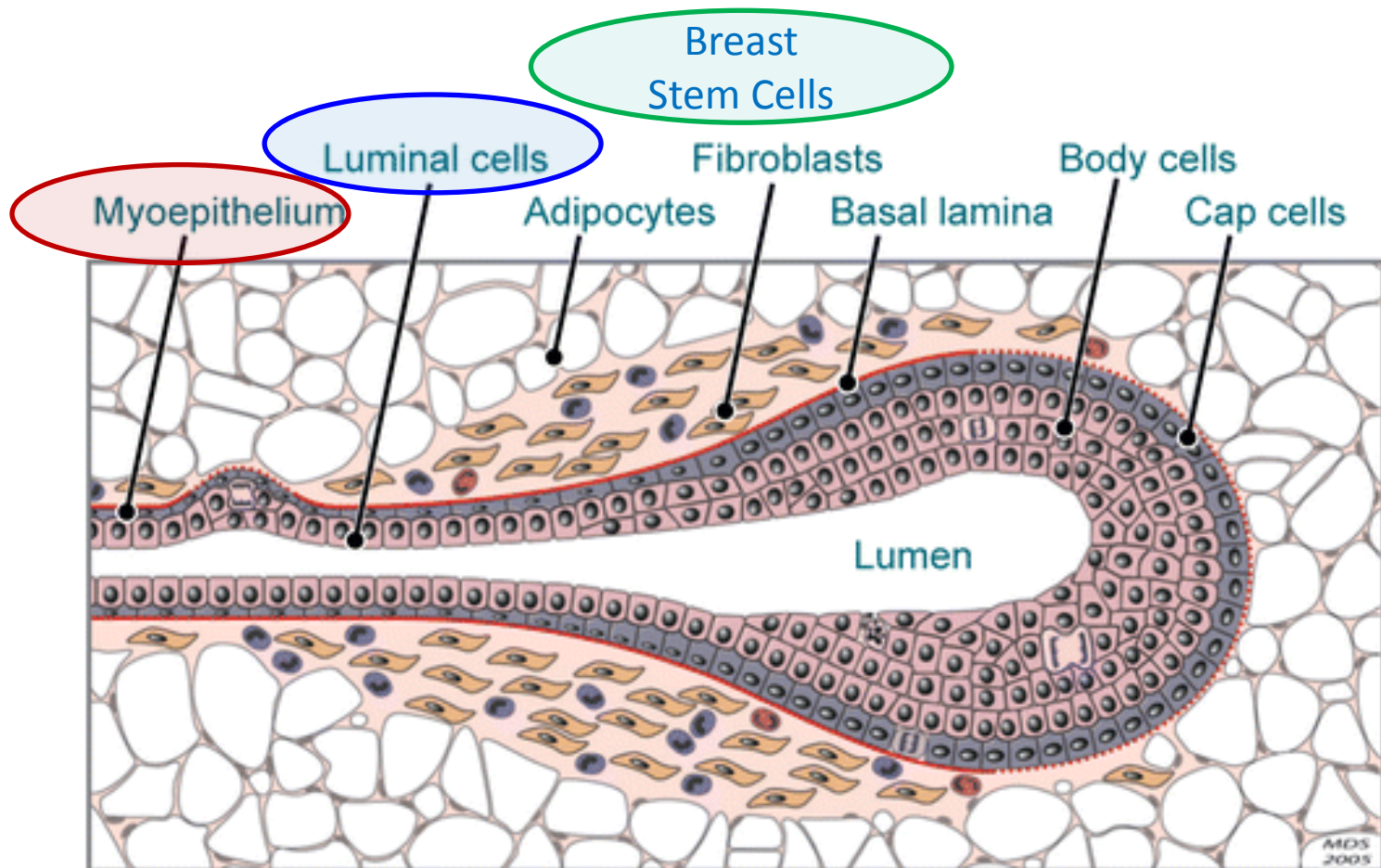
Promoter DNA Methylation in the Human Genome

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

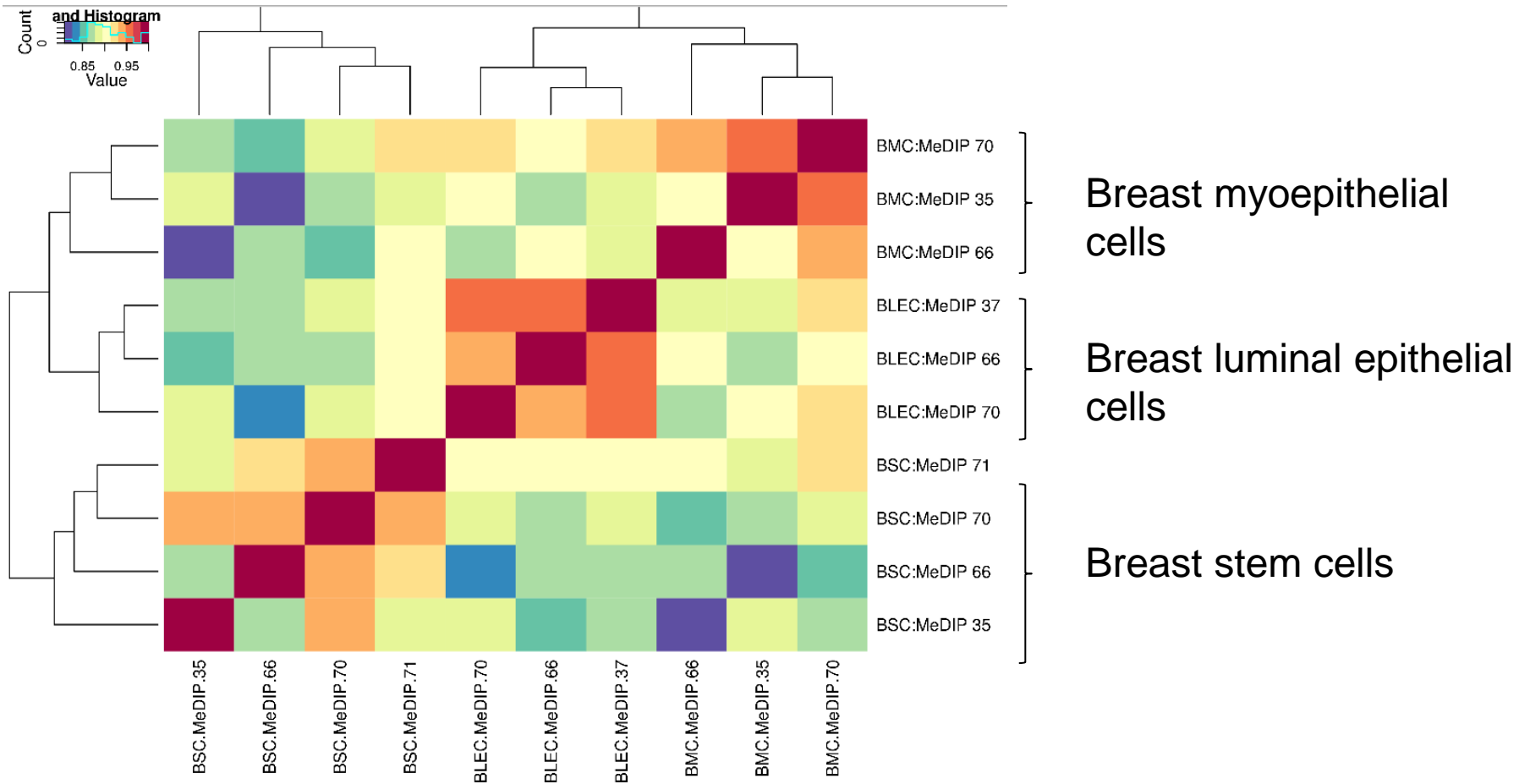
Breast Cell Type Differentiation



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)



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 active 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’

The “Grid Viewer” provides an easy way to visualize and select for analysis, only those tracks and assays from the large number that may be available. The grid partitions the tracks by the type of assay used to generate the track (i.e. MeDIP, etc)

Attribute	value
Group	Epigenome ToolSet Demo Input Data
Role	subscriber
Name	Breast
Description	Template for Human Genome, UCSC Build Hg19
Species	Homo sapiens
Version	hg19

Drag

Tool Settings

View Track Grid

Tool Overview

Databases with tracks of interest:

Database: *Breast* Group: *Epigenome ToolSet Demo Input Data*

Settings

X-axis attribute

Y-axis attribute

Page Title

Grid Title

X Label

Y Label

Advanced Settings:

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]

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.

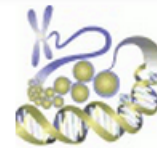
OK




Step 4. Select "Launch Grid Viewer" to select the samples (i.e. tracks of interest)



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

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".



- Select cells by clicking and dragging, then use the "View Selections in" pulldown
- NOTE: Some pages may not be accessible over low bandwidth internet connection

Tracks from Breast

Filter rows

eaSampleType

Breast Luminal Epithelial Cells

3

Breast Myoepithelial Cells

3

Breast Stem Cells

4

Save Track Selections

Choose a group and database to save selections in:

Select a Group:

ns will be saved

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_A"

OK

Enter a name to identify this set of selections

UseCase2_Breast_A

Save Selections

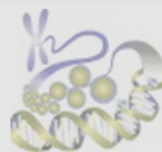
Cancel

← 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

- 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. T

Tracks from Breast

Filter rows: Selections ▾ Choose Databases

eaAssayType" →

eaSampleType

eaSampleType	MeDIP-Seq
Breast Luminal Epithelial Cells	3
Breast Myoepithelial Cells	3
Breast Stem Cells	4

Save Track Selections

Choose a group and database to save selections in:

Select a Group:

This is the group where your selections will be saved

GenboreeUser_group

Step 11. Select your user group (which you created)

Select a Database:

Choose a database within your group to save to

GenboreeUser_database

Step 12. Select your database (which you created)

Save Selection as:

Enter a name to identify this set of selections

UseCase2_Breast_B

Step 13. Name this list of tracks

Save Selections

Cancel

Step 14. Click "Save Selections"



- 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. T

Tracks from Breast

Filter rows:

eaSampleType

Breast Luminal Epithelial Cells

Breast Myoepithelial Cells

Breast Stem Cells

3

3

4

Save Track Selections

Choose a group and database to save selections in:

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"



Step 15. Click on "Workbench Data Selector" to the Data Selector by clicking on the link "Workbench Data Selector"

OK

Save Selections as:

Enter a name to identify this set of selections

UseCase2_Breast_B

Save Selections

Cancel

Welcome to the Genboree Workbench

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

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"

Track 2	BLEC:MeDIP 66
Track 3	BLEC:MeDIP 70
Track 4	BMC:MeDIP 35
Track 5	BMC:MeDIP 66
Track 6	BMC:MeDIP 70

Input Data

↑ ↓ × 📄

- UseCase2_Breast_A
- UseCase2_Breast_B

Output Targets

↑ ↓ × 📄

Drag

Welcome to the Genboree Workbench!

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

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.

BigBed	none
BioWin	none

Input Data

- UseCase2_Breast_A
- UseCase2_Breast_B
- Promoters:LCP**

Output Targets

-

Drag

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).

The screenshot displays the Genboree Workbench interface. At the top, a navigation bar includes tabs for System/Network, Data, QC and Pre-processing, Genome, Transcriptome, Cistrome, Epigenome (highlighted in green), and Metagenome. Below the navigation bar, a welcome message reads "Welcome to the Genboree Workbench! [Getting Started]".

The main interface is divided into several sections:

- Data Selector:** A tree view on the left showing a hierarchy of data sources. The "Use_Case_02_GU" project is selected.
- Details:** A panel on the right showing metadata for the selected project, including Attribute, View Link, Group, Name, and Refs.
- Epigenome Menu:** A dropdown menu is open, showing various tools. The "Compute Similarity Matrix (heatmap)" tool is highlighted in green. Other tools include Random Forest, QIIME, QC, Search for Similar Signals by Correlation, Analyze Signals, Slice Epigenomic Data, and Analyze Signals in the Context of Epigenome Atlas.
- Input Data:** A section below the menu showing the data sources used for the selected tool: UseCase2_Breast_A, UseCase2_Breast_B, and Promoters:LCP.
- Output Targets:** A section at the bottom showing the target outputs: GenboreeUser_database and Use_Case_02_GU.

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.

Tool Settings

Compute Similarity Matrix (heatmap)

Tool Overview

Input Entity Lists(s)/ROI-Track:

Items: UseCase2_Breast_A (Track Entity List)
UseCase2_Breast_B (Track Entity List)
Promoters:LCP (Track)

Output Database/Project:

Database/Projects Of Interest: GenboreeUser_database Group: GenboreeUser_group
Use_Case_02_GU Group: GenboreeUser_group

Epigenomic Experiment Heatmap Tool

Analysis Name EpigenomeExpHeatmap2013.

Remove No Data Regions?

Normalization Quantile

Aggregating Function Avg

Distance Function dist

Hierarchical Clustering Function hclust

Key

Key Size 0.75

Color Spectral

Height 8

Width 10

Trace None

Density Histogram

Dendograms Both

Submit **Cancel**

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:

The image shows a software interface with a modal dialog box titled "Job Submission Status". The dialog has a blue header and a light green content area. The title bar includes a close button (X). The main heading is "Compute Similarity Matrix (heatmap)". A green checkmark icon is next to the text "Job Id: *wbJob-epigenomicsHeatmap-PuHErD-9259*". Below this, a message states: "Your job has been successfully submitted. You will be notified by email when your job has completed." At the bottom of the dialog, there is an "OK" button. In the background, a table is partially visible with columns "Group", "Name", and "Refs". The "Group" column contains "GenboreeUser_group", "Name" contains "Use_Case_02_GU", and "Refs" contains "JobObject1". Below the dialog, the text "Output Targets" is visible, followed by "GenboreeUser_group" and "Use_Case_02_GU".

Group	Name	Refs
GenboreeUser_group	Use_Case_02_GU	JobObject1

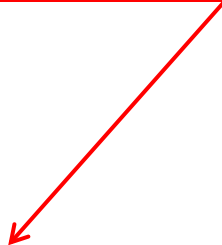
Output Targets

- GenboreeUser_group
- Use_Case_02_GU

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



The screenshot shows the Genboree project page. At the top left is the Genboree logo, which features the word "GENBOREE" in large, blue, stylized letters with a background of horizontal lines and a small image of a primate. To the right of the logo is the BCM logo, which consists of the letters "BCM" in a serif font, with "Baylor College of Medicine" written below it. Below the logos is a navigation menu with the following items: Home, Workbench, Browser, Profile, Groups, Projects, Databases, Tools, Log Out, and Help (with a question mark icon). In the top right corner, there is a button labeled "Edit Mode". The main content area features a large heading "Use_Case_02_GU" and a placeholder text "[[Put description for the project 'Use_Case_02_GU' here]]". Below this is a section titled "Project News:" followed by a news entry dated "2013/2/22:" which describes a user running an Epigenomic Heatmap Tool and provides a list of bullet points: "Study Name: EpigenomeExpHeatmap2013-02-22-12 02 09" and "Link to results".

GENBOREE

BCM
Baylor College of Medicine

Home Workbench Browser Profile Groups Projects Databases Tools Log Out Help ?

Edit Mode

Use_Case_02_GU

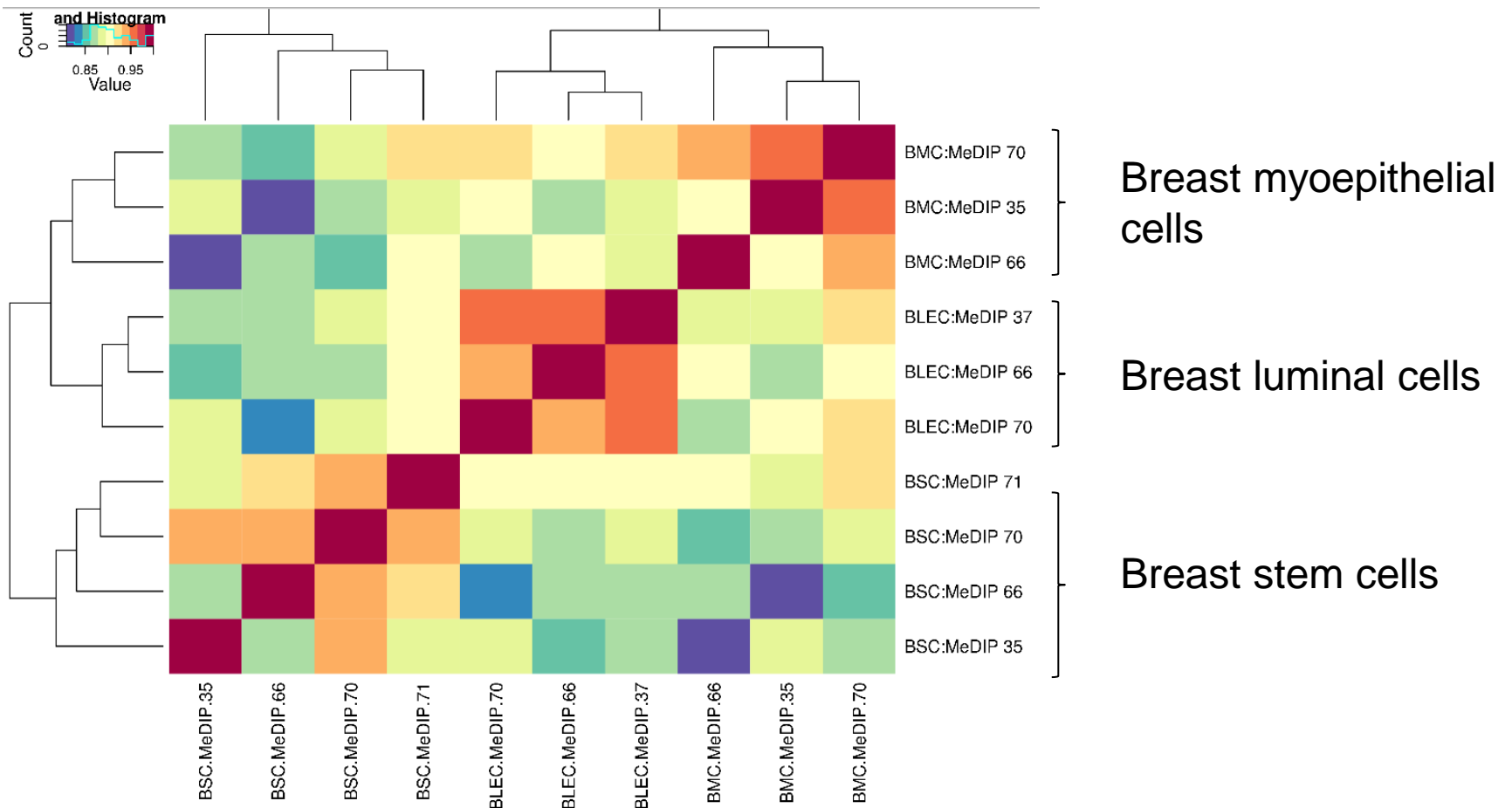
[[Put description for the project 'Use_Case_02_GU' here]]

Project News:

2013/2/22: Genboree User ran Epigenomic Heatmap Tool (EpigenomeExpHeatmap2013-02-22-12 02 09) and the results are available at the link below.

- **Study Name:** EpigenomeExpHeatmap2013-02-22-12 02 09
- **Link to results**

Use Case 2 Results: Breast Cell Types Cluster Based on MeDIP Profile (Epigenome Atlas and UCSF REMC data)



Data from: Epigenome Atlas, Release 5