

# Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Start by constructing a gene list</b>	<b>2</b>
2.1	Using predefined gene list(s)	2
2.2	Search for genes of interest using entrez id, symbol or description.	5
2.3	Paste a list of genes in the box provided.	5
2.4	Find predefined gene list(s) for your choice of genes.	6
2.5	Find genes with a phrase in their RIFs.	7
2.6	Find biogrid gene pairs for your gene(s).	8
<b>3</b>	<b>Start by selecting an experiment</b>	<b>8</b>
3.1	Search for an experiment	9
3.2	Query experiment	10
3.3	Cluster experiment	10
<b>4</b>	<b>Miscellaneous modules</b>	<b>11</b>
4.1	Filter samples and select sample grouping for analysis	11
<b>5</b>	<b>Interpreting Results</b>	<b>12</b>
5.1	Interactive Treeview Browsing	12
5.2	Static heatmaps	13
5.2.1	Statistical Analysis	14
5.3	Gene List Statistics	14
5.3.1	Predictive Ability Pvalue (LR)	14
5.4	Kegg Pathways for submitted genes	15
<b>6</b>	<b>Case Study: Characterizing experimentally identified proliferation signature</b>	<b>15</b>
6.1	Gene Expression data	16
6.1.1	Select a dataset from the portal	16
6.1.2	Paste a query gene list	16
6.1.3	Select Sample Grouping	16
6.1.4	Results	17
6.2	ChIP-seq data for different transcription factors	17
6.2.1	Select a data set from the portal	18
6.2.2	Paste a query gene list	18
6.2.3	Select Sample Grouping	18
6.2.4	Results	18
6.3	Tri-methylation of histone across 5 human cell lines	19
6.3.1	Select a data set from the portal	19
6.3.2	Paste a query gene list	20
6.3.3	Select Sample Grouping	20
6.3.4	Results	20

## 1 Introduction

GenomicsPortals is a web-based integrative computational platform for the analysis and mining of genomics data. We aim to integrate the primary genomics data, functional knowledge base and analytical tools within a single framework.

Genomics datasets are organized thematically into different portals. Different portals can contain datasets related to different diseases (eg Breast Cancer and Prostate Cancer), specific types of genomics data (eg Epigenomics and Transcription Factors), or different biological processes (eg Development). The same dataset can be assigned to different portals.

A typical analysis starts by constructing a list of genes by either using the predefined lists or pasting a gene list of interest, querying one of the databases with genome-scale data and producing analysis summaries. One can also start by searching for dataset of interest, and then constructing the query gene lists. In this case, one can also construct gene lists by browsing pre-computed clustering results.

We would like to note that we have designed the layout with the font size of 16 as a reference. If required, this default font size can be changed in the browser to increase the readability. In certain cases, simply “zoming in” will also make the text easier to read without pictures going out of focus.

## 2 Start by constructing a gene list

There are many ways to construct a gene list.

- Use a predefined gene list.
- Search for genes of interest using entrez id, symbol or description.
- Paste a list of genes in the box provided.
- Find predefined gene list(s) for your choice of genes.
- Find genes with a phrase in their RIFs.
- Find biogrid gene pairs for your gene(s).

The above list depicts various starting points to generate a list of genes of your interest. Rest of the work-flow is quite similar no matter how one selects a gene list.

### 2.1 Using predefined gene list(s)

Figure1 shows the interface to select a predefined gene list. Clicking on “Gene List” tab in the left menu would get this page. The lists are organized in different categories and we are constantly adding new lists and categories. Let’s say we are interested in gene lists in category “KEGG” with keywords “cell cycle”. Click

### Construct a gene list to query genomics data

Genes lists can be constructed by combining one or more **pre-defined gene list** (Gene Ontologies, KEGG pathways, Mouse Phenotypes, L2L and disease-related gene lists) from the database, or by searching **gene RIF's** and querying the database of all known **protein-protein Interaction**

**1) Select a predefined gene list.** [help](#)

- Transcription factor binding and epigenomics gene lists
- Cancer related gene lists
- GO gene lists
- KEGG gene lists

Keyword    
leave blank for all the lists

- L2L gene lists
- Disease gene lists based on text mining Gene RIFs
- MirBase lists
- MousePhenotype lists
- Transcription factor targets gene lists

Figure 1: Search a predefined gene list

on the link “KEGG gene list” to expand the search box as shown below. Type cell cycle in the text box and click submit.

This takes us to the following screen (figure2). Here we see a list of gene list returned for the keywords. Select one or multiple lists using the check boxes and click submit.

About

	Name	Description
<input checked="" type="checkbox"/>	<a href="#">hsa04110</a> <a href="#">view genes</a>	Cell cycle
<input type="checkbox"/>	<a href="#">mmu04110</a> <a href="#">view genes</a>	Cell cycle
<input type="checkbox"/>	<a href="#">mo04110</a> <a href="#">view genes</a>	Cell cycle

Query by:  Gene List  Experiments

User:   
 Password:

Figure 2: Select gene list

Now we see summary for the submitted lists. Note that below the summary table, we are asked to select union or intersection of the gene lists selected. In this case, because we submitted only one gene list, both union and intersection are identical. Let's select union and proceed to select an experiment for analysis.

Experiments are organized into different portals. If we know the portal the experiment of our interest belongs to, we can click on the portal name to list all the experiments in that portal. Since this is not the case most times, a search function is provided to look for experiments of our interest. In the “Filter experiments” box shown below, select organism, sample type, portal of your interest and type in keyword. Keyword could be left blank.

Let's click on portal name “Breastcancer” to proceed with our example.

Figure5 is a part of the screen showing experiments in the portal Breast-

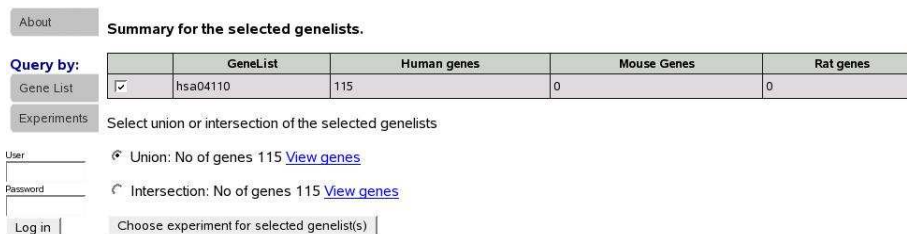


Figure 3: Summary of gene list

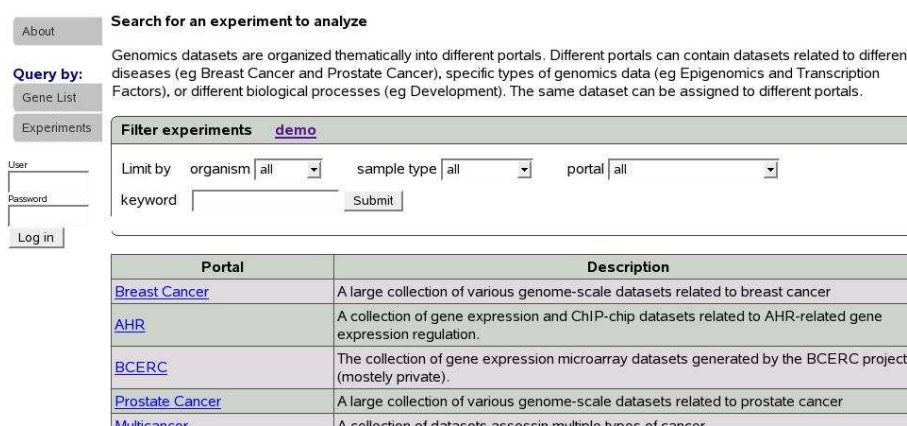


Figure 4: Select or search for experiments

cancer. If we had used the search function to look for an experiment instead, a similar screen would be shown. Below you would see a list of experiments. Let's select first experiment (GSE10797) and scroll down and click submit.

At this point, data is retrieved from database for the selected experiment and gene list(s) as shown in figure6.

One could download the data for his/her own analysis either as a tab delimited file or an eset for analysis in R. In this example, we have 66 samples, 209 probes and 103 genes. If we want to analyze only a subset of samples, we could select samples using the "Step 1" shown above. This step is optional and default is to select all the samples. Next we select a sample grouping for the analysis. We could choose to cluster on genes, samples, both genes and samples or none using the combo box shown above. Check the box "Compute LR" to compute predictive ability pvalue. Let's select "CellType" sample grouping, leave step 1 as it is to select all the samples, cluster on "none" and click Analyze button.

If we had checked the compute LR box we would see an additional column "Gene list Statistics" with the computed pvalue in the results tables as shown in figure8.

Figure 8depicts a typical summary table details of which can be found in

**Breast Cancer Genomics Database**

This is a web-portal for accessing our (soon to be) comprehensive database of public microarray datasets related to breast cancer. For a few examples on how to query data check out the [poster](#) we presented at Third Annual Early Environmental Exposures Meeting, Berkeley, CA, 2006. ([Berkeley-11-1-06Final.pdf](#))

**Query by:**  
 Gene List  
 Experiments

User: \_\_\_\_\_  
 Password: \_\_\_\_\_  
 Log in

**Filter experiments** [demo](#)

Limit by organism | all | sample type | all | portal | all |  
 keyword | \_\_\_\_\_ | Submit

**Public datasets**

	Description	Reference
<input checked="" type="radio"/>	66 Affy U133A 2.0 microarrays hybridized with total RNA from epithelial and stromal cells that were laser captured from normal breast tissue (n=5) and invasive breast cancer (n=28). <a href="#">GSE10797</a>	Casey T, Bond J, Tighe S, Hunter T et al. Molecular signatures suggest a major role for stromal cells in development of invasive breast cancer. <i>Breast Cancer Res Treat</i> 2008 Mar 29. <a href="#">Pub Med</a>
<input checked="" type="radio"/>	66 Affy U133A 2.0 microarrays hybridized with total RNA from epithelial and stromal cells that were laser captured from normal breast tissue (n=5) and invasive breast cancer (n=28). <a href="#">GSE10797</a> <a href="#">Entrez</a>	Casey T, Bond J, Tighe S, Hunter T et al. Molecular signatures suggest a major role for stromal cells in development of invasive breast cancer. <i>Breast Cancer Res Treat</i> 2008 Mar 29. <a href="#">Pub Med</a>
<input checked="" type="radio"/>	18 Affy U133A microarrays hybridized with mRNA from estrogen receptor-positive MCF-7 cells, stably transfected with the aromatase gene (known as MCF-7aro cells), after treatment with testosterone, 17β. <a href="#">GSE2225</a>	Itoh T, Karlsberg K, Kijima I, Yuan YC et al. Letrozole-, anastrozole-, and tamoxifen-responsive genes in MCF-7aro cells: a microarray approach. <i>Mol Cancer Res</i> 2005 Apr;3(4):203-18. <a href="#">Pub Med</a>
<input checked="" type="radio"/>	12 Affy U133A microarrays hybridized with mRNA from MDA-MB-231 Breast Cancer Cells expressing either wild-type estrogen receptor or the mutant estrogen receptor L540Q when treated with estradiol for 1 <a href="#">GSE1045</a>	Acevedo et al. (2004) Selective Recognition of Distinct Classes of Coactivators by a Ligand-Inducible Activation Domain. <i>Molecular Cell</i> 13: 725-738, 2004 <a href="#">Pub Med</a>

Figure 5: show experiments

Section Interpreting Results’.

## 2.2 Search for genes of interest using entrez id, symbol or description.

This section describes how to search for genes in the database and proceed with the analysis of genes found in search results.

Figure9 shows gene search page. This can be retrieved by clicking on the “Gene List” tab in the left menu. Genes could be searched by one of the three parameters: Gene ID (Entrez ID) e.g. 2099, symbol e.g. “ESR” or description e.g “estrogen”. Type the value in the text box shown above. We can limit the search to a specific organism if required e.g. human, mouse or rat. For this example let’s search for symbol “ESR”. To do this, first select “Symbol” radio in the left column, type “ESR” in the text box. Select human from the Organism combo box (default is to look across all organisms) and click submit.

Figure10 shows the search result. Now we have all the genes with symbol “ESR” or similar symbol names. To analyze gene(s) select all the genes of interest and then click submit. Now we are presented a screen to select an experiment to analyze. From this point on, we proceed as explained in previous section “Using predefined gene lists”.

## 2.3 Paste a list of genes in the box provided.

This section describes how to submit your own list of genes for analysis.

Figure11 shows the screen to submit your own list of genes. You can use either entrez ids (e.g. 2099) or symbols (e.g. esr1). As shown in the figure,

**Query by:**

View genes found in the platform

#Samples	#Probes	#Genes	Data download
66	209	103	<a href="#">Tabular format (xls)</a> <a href="#">RData (eset)</a>

User: \_\_\_\_\_  
 Password: \_\_\_\_\_

**Step 1 (optional) Select samples for analysis. [demo](#)**

include  exclude  
 Disease  
 CellType  
 Replicate

**Step 2) Select sample grouping for analysis**

none  
 Disease  
 CellType  
 Replicate

Cluster on:   Compute LR

Figure 6: Get data

**Query by:**

#Samples	#Probes	#Genes	Data download
66	209	103	<a href="#">Tabular format (xls)</a> <a href="#">RData (eset)</a>

User: \_\_\_\_\_  
 Password: \_\_\_\_\_

Interactive Treeview Browsing	Static Heatmaps(pdf)	Data Download
<a href="#">* Original Data</a> <a href="#">Centered Data</a>	<a href="#">Legend for all the heatmaps</a> <a href="#">Statistical Analysis †</a> <a href="#">Original Data</a> <a href="#">Centered data</a>	<a href="#">Tabular format (xls)</a> <a href="#">RData (eset)</a>

\* Please note that unless the original data is in the form of log-transformed ratios, Non-centered treeview might not be very informative  
 † Values represented by heatmaps correspond to average expression levels for the same sample type. Red box in the left color box indicates pvalue less than 0.05

Figure 7: Results

type/paste a list in the box. We could optionally select an organism (human, mouse or rat) to filter these genes. By default, all the genes are submitted. Let's genes 10,12,2099 in the box and click submit.

At this point our database is searched for all the genes submitted and figure12 shows the list of genes found.

Now we can proceed as explained in the previous section 2.2“Search for genes of interest using entrez id, symbol or description”.

## 2.4 Find predefined gene list(s) for your choice of genes.

This section describes how to find predefined gene lists that contain genes of interest.

There are two links beside title “demo” and “demo gene list”. Clicking on demo gene list link shows a few sample genes that we are going to use for the purpose of this demo. The demo genes are as follows:

79575

About		Description	Reference
Query by:		66 Affy U133A 2.0 microarrays hybridized with total RNA from epithelial and stromal cells that were laser captured from normal breast tissue (n=5) and invasive breast cancer (n=28). <a href="#">GSE10797</a>	Casey T, Bond J, Tighe S, Hunter T et al. Molecular signatures suggest a major role for stromal cells in development of invasive breast cancer. Breast Cancer Res Treat 2008 Mar 29. <a href="#">Pub Med</a>
Gene List		#Samples	#Probes
Experiments		#Genes	Data download
		66	209
		103	<a href="#">Tabular format (xls)</a> <a href="#">RData (eset)</a>

Interactive Treeview Browsing	Static Heatmaps(pdf)	Gene list Statistics	Data Download
* <a href="#">Original Data</a> <a href="#">Centered Data</a>	<a href="#">Legend for all the heatmaps</a> <a href="#">Statistical Analysis</a> † <a href="#">Original Data</a> <a href="#">Centered data</a>	Predictive Ability pvalue (LR) 0.0	<a href="#">Tabular format (.xls)</a> <a href="#">RData (eset)</a>

\* Please note that unless the original data is in the form of log-transformed ratios, Non-centered treeview might not be very informative  
† Values represented by heatmaps correspond to average expression levels for the same sample type. Red box in the left color box indicates pvalue less than 0.05

Figure 8: Results with LR

**2) Search for genes** [demo](#)

Using  Organism

Gene ID  
 Symbol  
 Description

Figure 9: Search genes

10096  
9447

Now copy and paste these genes in the text box above. The radio buttons provide option of how we want to search for the gene lists in the database. “Match any gene” would find all the gene lists that contain any of the genes we input whereas “Match all genes” would find only the gene lists that contain all of the genes.

We also have option of selecting which categories of predefined gene lists to search.

Figure14 shows all the categories of predefined gene lists that visible after clicking “Search in following lists”.

By default all the lists are selected.

Let’s proceed with our example using “Match all genes” option and default case for searching lists (search all lists).

Figure15 shows the result of our query. It shows all the gene lists found along with their description.

Let’s select first list (NFkB) and submit.

Figure16 shows the resultant screen which shows the summary of gene lists selected. You might recall that this is similar to the screen shown in first section 2.1“Using predefined gene lists” and rest of the analysis is as described in that section.

## 2.5 Find genes with a phrase in their RIFs.

This section describes how to find genes based on their RIFs.

Please check genes you want to analyze.

	GeneID	Symbol	Description	Entrez
<input checked="" type="checkbox"/>	<a href="#">2099</a>	ESR1	estrogen receptor 1	<a href="#">Entrez</a>
<input checked="" type="checkbox"/>	<a href="#">2100</a>	ESR2	estrogen receptor 2 (ER beta)	<a href="#">Entrez</a>
<input checked="" type="checkbox"/>	<a href="#">2101</a>	ESRRA	estrogen-related receptor alpha	<a href="#">Entrez</a>
<input checked="" type="checkbox"/>	<a href="#">2103</a>	ESRRB	estrogen-related receptor beta	<a href="#">Entrez</a>
<input checked="" type="checkbox"/>	<a href="#">2104</a>	ESRRG	estrogen-related receptor gamma	<a href="#">Entrez</a>
<input checked="" type="checkbox"/>	<a href="#">144847</a>	ESRRAP	estrogen-related receptor alpha pseudogene	<a href="#">Entrez</a>

Figure 10: Gene search result

**3) Paste a list of genes in the box below** [demo](#) [demo gene list](#)

10  
12  
2099

Organism

All

Figure 11: Submit a custom list of genes

The figure17 is self-explanatory. Let's type "argyrophilic grain disease" in the box and click submit.

Figure18 shows the result of our search. Select genes of interest and submit. Now we are presented a screen to select an experiment for analysis. Proceed as explained in previous section 2.2"Search for genes of interest using entrez id, symbol or description".

## 2.6 Find biogrid gene pairs for your gene(s).

This section describes how to find biogrid gene pairs for your genes.

Let's use genes 79575,10096,9447 we used in previous examples. Select "Match any genes" option and click submit.

Figure20 shows result for our search. Click submit and proceed with analysis as explained in section 2.2"Search for genes of interest using entrez id, symbol or description".

## 3 Start by selecting an experiment

If one is interested in a particular experiment, it is useful to locate the experiment first and then proceed with the analysis.

This section describes how to do this. At the time of this writing, there are 1904 experiments in the database and this number is continuously growing. Experiments are organized into different portals. If we know the portal the experiment of our interest belongs to, we can click on the portal name to list all

Please check genes you want to analyze.

	GeneID	Symbol	Description	Entrez
<input checked="" type="checkbox"/>	10	NAT2	N-acetyltransferase 2 (arylamine N-acetyltransferase)	<a href="#">Entrez</a>
<input checked="" type="checkbox"/>	12	SERPINA3	serpin peptidase inhibitor, clade A (alpha-1 antitrypsin), member 3	<a href="#">Entrez</a>
<input checked="" type="checkbox"/>	2099	ESR1	estrogen receptor 1	<a href="#">Entrez</a>

Figure 12: Submitted genes

**4) Find predefined gene lists containing your choice of gene(s).** [demo](#) [demo gene list](#)

Genes 79575 10096 9447	<input type="radio"/> Match any gene <input checked="" type="radio"/> Match all genes	Search in following <input type="checkbox"/> lists	<input type="button" value="Submit"/>
---------------------------------	--	---	---------------------------------------

Figure 13: Find predefined gene lists

the experiments in that portal. Since this is not the case most times, a search function is provided to look for experiments of our interest.

Start by clicking “Experiments” tab on the left menu.

Figure 21 shows the experiment tab. The “Filter experiments” box on top provides search functionality.

The table below shows various portals and their descriptions. Clicking on a portal name shows the experiments belonging to that portal.

### 3.1 Search for an experiment

Search is a very important part of this portal given the number of experiments we have. To locate experiments of interest, a simple and effective search functionality is provided.

Figure22 shows the screen to filter experiments.

Following are the components of this module:

1) Organism: Experiments could be filtered by selecting one of the organism from the combo box named “organism”. Three options proved are human, mouse and rat.

2) Sample type: Three sample types are provided for selection. Tissue, cell line and motif score. Select appropriate from the combo box.

3) Data type: Six data types are available for selection from the combo box.

4) Portal: All the available portals are listed here. Select a portal if you want to limit your search to that particular portal.

5) Keyword: This could be a name of experiment, a word in description or reference.

Let’s search for experiments with keyword “miller” across all organisms, sample types and portals as an example.

4) Find predefined gene lists containing your choice of gene(s). [demo](#) [demo gene list](#)

Search in following lists

- GO
- BCERC
- MirBase
- L2L
- Kegg
- TF binding and epigenomics gene list
  - Cancer related
  - Transcription factor
  - Mouse Phenotype
  - ORegAnno

Submit

Genes

75575  
10096  
9447

Match any gene  
 Match all genes

Figure 14: Select categories of predefined gene lists to search

Total lists found: 3

	Name	Description	Table
<input checked="" type="checkbox"/>	NFkB	The study performs genome scale location analysis to profile DNA binding of NF-KappaB proteins in U937 cells. It identifies novel NF-kappaB target genes.	ChipchipLists
<input type="checkbox"/>	GO:0008150	biological_process	GOlists
<input type="checkbox"/>	all	all	GOlists

Submit

Figure 15: Result for finding predefined gene lists

Figure 23 shows a part of the result page. As shown above, all the experiments found for the search criteria are listed in a table. Notice the “Query” and “Cluster” buttons in the last two columns of the table. These buttons provide a way to analyze the experiments and are explained in detail in following sections.

### 3.2 Query experiment

Clicking “Query” button shown in figure 23 opens a new window for that particular experiment.

e.g. Figure 24 shows the query page for experiment “GSE1045” in portal Breastcancer. The top table provides a summary of the experiment and bottom table lists all the properties (sample subgroupings) available for this experiment. These are useful for analysis.

These two tables are followed by all the options to construct a gene list shown in “Gene List” tab.

The procedure for analysis is similar to what was described in section 2 “Start by constructing gene list” except that the step to select experiment is skipped as we already have an experiment to work with.

### 3.3 Cluster experiment

Clicking “Cluster” button shown in figure 23 opens a new window for that particular experiment.

**Summary for the selected genelists.**

	GeneList	Human genes	Mouse Genes	Rat genes
<input checked="" type="checkbox"/>	NFKB	283	1	0

Select union or intersection of the selected genelists

Union: No of genes 285 [View genes](#)

Intersection: No of genes 285 [View genes](#)

Figure 16: Summary of gene list

**5) Find genes with a phrase (e.g. "breast cancer") in their RIFs**

Search phrase

Figure 17: Find genes from their RIFs

## 4 Miscellaneous modules

### 4.1 Filter samples and select sample grouping for analysis

Note: This step is optional and default is to select all the available samples for analysis.

An experiment may have a number of samples which are organized in different groups (sample subgroupings).

One may wish to restrict analysis only to a subset of all the available samples for an experiment.

This section describes how this is achieved.

Figure 26 shows the screen to filter samples for experiment GSE10797. We can choose to either include or exclude all the samples that satisfy the criteria we are going to define by selecting appropriate option using the radio button.

All the sample subgroupings are listed in this box. When we click on a sample subgrouping, the link expands to show all the unique values for the same as shown in figure 27.

Let's say, we want to include only the samples for which Disease is cancer and CellType is epithelial.

As shown in figure 28, select include from the radio button, and check cancer box under Disease and epithelial box under CellType. When you click Analyze, only the samples for this criteria will be used for analysis.

Next step is to select sample grouping for analysis.

Please check genes you want to analyze.

	GeneID	Symbol	Description	Entrez
<input checked="" type="checkbox"/>	2	A2M	alpha-2-macroglobulin	<a href="#">Entrez</a>
<input checked="" type="checkbox"/>	<a href="#">4137</a>	MAPT	microtubule-associated protein tau	<a href="#">Entrez</a>

Figure 18: RIF search result

**6) Find biogrid gene pairs for your gene(s).**

Genes  
79575  
10096  
9447

Match any gene  
 Match all genes

Figure 19: BioGrid

## 5 Interpreting Results

This section will describe the results page in detail. For illustration purpose we will take all genes with “stem cell” keyword in GO category as shown in figure 30.

Check all the genelists on the resultant page and proceed as explained in section ‘Using predefined gene list(s)’. Select experiment “GSE2225” in portal “Breast cancer” and use sample subgrouping “Treatment” and cluster on “genes” to obtain the results shown in figure 31.

The results page structure is as follows. The first table gives a brief description of the selected experiment. The second table summarizes the data retrieved for the analysis of submitted query gene list and provides links for download. Data can be retrieved in the form of spreadsheet and R data object. The third table gives the analysis results which are explained in detail as follows.

### 5.1 Interactive Treeview Browsing

Unsupervised clustering of the query data was performed using the Bayesian model-based procedures [1] as well as simple hierarchical clustering. The functional annotation of the clustering structures was performed using the CLEAN framework [2], the integrative browsing of the data and functional annotations is facilitated through the Functional TreeView (FTreeView) which is a Java web-start based clustering browser [2]. Using FTreeView, one can identify clusters of genes based on their data profile and correlation with specific functional categories and use such gene lists to query and analyzed genomics data in other datasets.

We would like to note that in the case where no clustering option (on the genes as well as samples) is chosen, the TreeView application would show the heatmap with no dendrograms on either sides. This might make the heatmap

Total genes found: 10

	GeneID	Symbol	Description	Entrez
<input checked="" type="checkbox"/>	<a href="#">2017</a>	CTTN	cortactin	<a href="#">Entrez</a>
<input checked="" type="checkbox"/>	<a href="#">3059</a>	HCLS1	hematopoietic cell-specific Lyn substrate 1	<a href="#">Entrez</a>
<input checked="" type="checkbox"/>	<a href="#">10093</a>	ARPC4	actin related protein 2/3 complex, subunit 4, 20kDa	<a href="#">Entrez</a>
<input checked="" type="checkbox"/>	<a href="#">10094</a>	ARPC3	actin related protein 2/3 complex, subunit 3, 21kDa	<a href="#">Entrez</a>
<input checked="" type="checkbox"/>	<a href="#">10095</a>	ARPC1B	actin related protein 2/3 complex, subunit 1B, 41kDa	<a href="#">Entrez</a>
<input checked="" type="checkbox"/>	<a href="#">10097</a>	ACTR2	ARP2 actin-related protein 2 homolog (yeast)	<a href="#">Entrez</a>
<input checked="" type="checkbox"/>	<a href="#">10109</a>	ARPC2	actin related protein 2/3 complex, subunit 2, 34kDa	<a href="#">Entrez</a>
<input checked="" type="checkbox"/>	<a href="#">10163</a>	WASF2	WAS protein family, member 2	<a href="#">Entrez</a>
<input checked="" type="checkbox"/>	<a href="#">10289</a>	EIF1B	eukaryotic translation initiation factor 1B	<a href="#">Entrez</a>
<input checked="" type="checkbox"/>	<a href="#">10810</a>	WASF3	WAS protein family, member 3	<a href="#">Entrez</a>

Figure 20: BioGrid search result

Genomics datasets are organized thematically into different portals. Different portals can contain datasets related to different diseases (eg Breast Cancer and Prostate Cancer), specific types of genomics data (eg Epigenomics and Transcription Factors), or different biological processes (eg Development). The same dataset can be assigned to different portals.

**Filter experiments** [demo](#)

Limit by organism  sample type  portal

keyword

Portal	Description
<a href="#">Breast Cancer</a>	A large collection of various genome-scale datasets related to breast cancer
<a href="#">AHR</a>	A collection of gene expression and ChIP-chip datasets related to AHR-related gene expression regulation.
<a href="#">BCERC</a>	The collection of gene expression microarray datasets generated by the BCERC project (mostly private).
<a href="#">Prostate Cancer</a>	A large collection of various genome-scale datasets related to prostate cancer

Figure 21: Experiments tab

incomprehensible at first. However, one can click on any of the genes or group of genes and the corresponding gene annotations will be displayed in the right-most window. The scenario is depicted in 33 where genes and samples are not clustered.

## 5.2 Static heatmaps

In addition to interactive treeview interface, Cluster analysis results are also available as static annotated heatmaps saved in pdf files. The values represented by heatmaps correspond to log transformed ratios.

Figure 34 illustrates static heatmap clustered on selected stem cell genes across 6 treatment types. These sample annotations are provided separately in the link “legend for all the heatmaps” as shown in figure 35.

**Filter experiments** [demo](#)

Limit by

organism  sample type  data type  portal

keyword

Figure 22: Search for experiments

**Filter experiments** [demo](#)

Limit by

organism  sample type  data type  portal

keyword

Description	Reference		
251 Affy U133A microarrays hybridized with mRNA from primary breast tumors. <a href="#">GSE3494Entrez</a>	Miller,L.D. et al. (2005) From The Cover: An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. Proceedings of the Natl <a href="#">Pub Med</a>	<input type="button" value="Query"/>	<input type="button" value="Cluster"/>
116 Affy U133A microarrays hybridized with mRNA extracted from the paired biopsies taken from the same subjects both pretreatment and after 10-14 days Letrozol, 2.5 mg/day, oral <a href="#">GSE5462</a>	Miller WR, Lanonov AA, Renshaw L, Anderson TJ et al. Changes in breast cancer transcriptional profiles after treatment with the aromatase inhibitor, letrozole. Pharmacogenet Genomics 2007 Oct;17(10): <a href="#">Pub Med</a>	<input type="button" value="Query"/>	coming soon
251 Affy U133A microarrays hybridized with mRNA from primary breast tumors. <a href="#">GSE3494</a>	Miller,L.D. et al. (2005) From The Cover: An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. Proceedings of the National Academy of Sciences 102[38], 13550-13555 <a href="#">Pub Med</a>	<input type="button" value="Query"/>	<input type="button" value="Cluster"/>
18 Affymetrix GeneChip Mouse Genome 430 2.0 microarrays hybridized with mRNA generated during the differentiation of	Christoforou N, Miller RA, Hill CM, Jie CC et al. Mouse ES cell-derived cardiac precursor cells are multipotent and facilitate identification of novel cardiac genes. J Clin Invest	<input type="button" value="Query"/>	coming soon

Figure 23: Filter experiment result

### 5.2.1 Statistical Analysis

For the selected samples in the dataset, we can identify differentially expressed significant genes. Values represented by heatmaps correspond to average expression levels for the same sample subgrouping. Red box in the left sidebar indicates pvalue less than 0.05.

## 5.3 Gene List Statistics

### 5.3.1 Predictive Ability Pvalue (LR)

To assess the predictive ability of the selected sample grouping (in this example “treatment”), we select random genes of the same length as that of query gene list from the particular platform. The enrichment of the statistically significant genes in the query list was then assessed using logistic regression [3].

Description	Reference
12 Affy U133A microarrays hybridized with mRNA from MDA-MB-231 Breast Cancer Cells expressing either wild-type estrogen receptor or the mutant estrogen receptor L540Q when treated with estradiol for 1 <a href="#">GSE1045</a>	Acevedo et al. (2004) Selective Recognition of Distinct Classes of Coactivators by a Ligand-inducible Activation Domain. Molecular Cell 13: 725-738, 2004 <a href="#">Pub Med</a>

Properties stored for the experiment:

genotype  
estrodial  
duration  
sample

Figure 24: Query experiment

Description	Reference
18 Affy U133A microarrays hybridized with mRNA from estrogen receptor-positive MCF-7 cells, stably transfected with the aromatase gene (known as MCF-7aro cells), after treatment with testosterone, 17 $\beta$ -estradiol, two aromatase inhibitors (letrozole and anastrozole), an anti-estrogen (tamoxifen), and control. <a href="#">GSE2225</a>	Itoh T, Karlsberg K, Kijima I, Yuan YC et al. Letrozole-, anastrozole-, and tamoxifen-responsive genes in MCF-7aro cells: a microarray approach. Mol Cancer Res 2005 Apr;3(4):203-18. <a href="#">PubMed</a>

To view result of the cluster analysis for this dataset using the basic infinite mixture model (IMM) click on the button below. You need to have the Java Virtual Machine 1.42 installed on your computer to view the clustering. The results will open in a JavaTreeView viewer ([Saldanha A.J. Bioinformatics 20: 3246-3248, 2004.](#)). Cluster analysis was performed using [gimmR package](#) and the basic IMM model ([Medvedovic M and Sivaganesan S. Bioinformatics 18: 1194-1206, 2002.](#)). Genes were first pre-filtered based on variation and reliability of measurements in at least 10% of samples which resulted in 8,070 probes representing 6,160 unique genes. We selected best representative probes for each gene based on quality control considerations and clustered them using IMM. After clustering, we selected subclusters with average linkage posterior probability of clustering of at least 0.9. Samples were clustered using average linkage and the Pearson's correlation coefficient.

In order to run the applet below, you will need Java Runtime Environment 1.5 installed ().

Clustering Algorithm	No of Genes	Gene selection method	
DCGimm	5000	Copa	<input type="button" value="Cluster"/>

Figure 25: Cluster experiment

## 5.4 Kegg Pathways for submitted genes

Query gene lists are incorporated into KEGG pathway images. One can click on a Pathway ID to view graphical representation of the pathway. Significantly expressed genes are painted yellow and other genes that were found in that particular pathway but are not significantly expressed are painted blue.

## 6 Case Study: Characterizing experimentally identified proliferation signature

We demonstrate the utility of the Genomics Portals through a case study investigating proliferation gene expression signature in rat mammary epithelium induced by different fatty acid diets [4]. This study established the increased proliferation of mammary epithelium as a consequence of several different dietary regiments in virgin female Sprague-Dawley rats. The study also identified a set of 85 genes whose expression levels were correlated with the increased proliferation.



Figure 26: Select samples for analysis



Figure 27: Sample selection expanded

## 6.1 Gene Expression data

We used Genomics Portals to study the functional importance of these 85 genes in five different biological processes examined in 4 gene expression datasets [5, 6, 7, 8] which are available in the portal. Here, we present step-by-step instructions for reproducing the results using Miller et.al. [5] dataset which comprises of 251 primary human breast tumors. This dataset was re-processed and curated before being deposited into the back-end databases under the id “GSE3494Entrez”. The comparison of interest in this case was between the largest (top quartile) and smallest (bottom quartile) tumor with the assumption that large tumors are “more proliferative” than small tumors.

### 6.1.1 Select a dataset from the portal

Go to “Experiments” tab and type ‘GSE3494Entrez’ in the keyword field of ‘Filter experiments’ option. You can also find this experiment under ‘Breast Cancer’ portal. Press “submit”. This will fetch the corresponding experiment and then press “Query” button.

### 6.1.2 Paste a query gene list

Paste a list of Entrez ids of 85 up regulated proliferation genes found at <http://eh3.uc.edu/documentation/upregulatedDietsGenes.txt> in the box (option 3) and press “submit”.

### 6.1.3 Select Sample Grouping

This page provides collective information about the selected dataset, gene list submitted (and the actual number of probes found on this platform) as well as sample groupings associated with this dataset. In this example, select “tumorSize-quartiles” as sample grouping in step 2. We do not want to filter any samples

Step 1 (optional) Select samples for analysis. [demo](#)

- include  exclude
- Disease
  - cancer
  - normal
- CellType
  - epithelial
  - stroma
- Replicate

Figure 28: Filter sample example

Step 2) Select sample grouping for analysis

- none
- Disease
- CellType
- Replicate

Figure 29: Select sample group

hence we can skip step 1. Also, select “computeLR” and press “Analyze” button. Figure 37 depicts the snapshot of this step.

#### 6.1.4 Results

Click on the “statistical Analysis” link and you will get a heatmap as shown in 39. The corresponding legends can be found by clicking on the link “legend for all the heatmaps” as shown in 38. One can see that indeed the genes in the query list are up-regulated in large tumors (quart-4) and are enriched for differentially expressed gene (LRpath p-value<10<sup>-9</sup>).

Similar analysis could be performed on the other 3 datasets using the same list of 85 up regulated proliferation genes. We have established the universality of the proliferation signature identified in the rat dietary studies across four very different biological systems using the Genomics Portals interface. The entire process of querying and generating results can be completed in less than 10 minutes. More details could be obtained from the manuscript.

## 6.2 ChIP-seq data for different transcription factors

In addition to using gene expression data, we further characterize our proliferation signature using ChIP-seq data for E2F1 transcription factor (TF) [9]. In the original paper, an extended set of genes identified through cluster analysis was linked to regulatory domain of E2F transcription factors by examining the overlap with E2F targets established in ChIP-chip [10] and global expression profiling [11] experiments, and computationally predicted E2F targets. Here, we used Genomics Portals to examine the newer ChIP-seq dataset assessing DNA binding of 15 different transcription factors, including E2F1, in mouse embryonic stem cells. Following steps can be conducted to obtain the respective heatmaps.

1) Select a predefined gene list. [help](#)

- Transcription factor binding and epigenomics gene lists
- Cancer related gene lists
- GO gene lists
- Keyword:  leave blank for all the lists
- KEGG gene lists
- L2L gene lists
- Disease gene lists based on text mining Gene RIFs
- MirBase lists
- MousePhenotype lists
- Transcription factor targets gene lists

Figure 30: Select gene list with “stem cell” keyword

### 6.2.1 Select a data set from the portal

Go to “Experiments” tab and type ‘GSE11431peaks’ in the keyword field of ‘Filter experiments’ option. You can also find this experiment under ‘Transcription Factors’ portal. Press “submit”. This will fetch the corresponding experiment and then press “Query” button.

### 6.2.2 Paste a query gene list

Paste a list of Entrez ids of 85 up regulated proliferation genes in the box (option 3) and press “submit”.

### 6.2.3 Select Sample Grouping

Select “Transcriptionfactor” as sample grouping in step 2. We do not want to filter any samples hence we can skip step 1 and then press “Analyze”

### 6.2.4 Results

Click on the link “Centered data” under static heatmap column of the result table. Figure 41 shows heatmap of 15 Tfs and figure 40 displays corresponding legends for each of the TFs.

We can see that in addition to most of the genes having a ChIP-seq peak for E2F1 within the regulatory region examined (-4kb to +1kb around TSS marked by 0), there were several other transcription factors such as N-myc, Tcfp2l1, c-Myc etc. that seem to have unusually many peaks for these gene. We can then focus on each of the TFs separately to take a closer look. We will illustrate the case using n-Myc TF.

We can select n-Myc TF out of 15 Tfs using “select sample” option in step 1 as shown in figure 42. Expand “Transcription Factor” and select n-Myc TF and click radio button “include” to select this sample. Then select “TranscriptionFactor” in step 2. select Cluster on “Genes” and “compute LR” options and click “Analyze”.

Description			Reference
18 Affy U133A microarrays hybridized with mRNA from estrogen receptor-positive MCF-7 cells, stably transfected with the aromatase gene (known as MCF-7aro cells), after treatment with testosterone, 177- <a href="#">GSE2225</a>			Itoh T, Karlsberg K, Kijima I, Yuan YC et al. Letrozole-, anastrozole-, and tamoxifen-responsive genes in MCF-7aro cells: a microarray approach. Mol Cancer Res 2005 Apr;3(4):203-18. <a href="#">Pub Med</a>
#Samples	#Probes	#Genes	Data download
18	35	16	<a href="#">Tabular format (xls)</a> <a href="#">RData (eset)</a>
Interactive Treeview Browsing	Static Heatmaps(pdf)	Gene list Statistics	Data Download
* <a href="#">Original Data</a> <a href="#">Centered Data</a>	<a href="#">Legend for all the heatmaps</a> <a href="#">Statistical Analysis †</a> <a href="#">Original Data</a> <a href="#">Centered Data</a>	Predictive Ability pvalue (LR) 0.13	<a href="#">Tabular format (.xls)</a> <a href="#">RData (eset)</a>
<small>* Please note that unless the original data is in the form of log-transformed ratios, Non-centered treeview might not be very informative  † Values represented by heatmaps correspond to average expression levels for the same sample type. Red box in the left color box indicates pvalue less than 0.05</small>			
<a href="#">Kegg Pathways for submitted genes</a>			
<b>Step 1 (optional) Select samples for analysis.</b> <a href="#">demo</a>			
<input type="radio"/> include <input checked="" type="radio"/> exclude <input checked="" type="checkbox"/> Treatment <input checked="" type="checkbox"/> Replicate			
<b>Step 2) Select sample grouping for analysis</b>			
<input checked="" type="radio"/> none <input type="radio"/> Treatment <input type="radio"/> Replicate			
Cluster on <input type="text" value="none"/> <input type="checkbox"/> Compute LR <input type="button" value="Analyze"/>			

Figure 31: Results of stem cell gene list query

Then click on the link “Centered data” under static heatmap column of the result table. Figure 43 shows increased binding around TSS of the promoter region (-4kb to +1kb in this case) for some of these genes.

Here, we used the comparison to “random” sample by LRpath. Instead of the p-values, in this situation Genomics Portals by default uses the maximum “peak intensity” calculated for each gene across its whole regulatory region. Such statistical analysis confirmed that in addition to E2F1 (p-value < 10<sup>-14</sup>), n-Myc (p-value < 10<sup>-7</sup>), Tcfp2l1 (p-value < .001), c-Myc (p-value < .01), and Klf4 (p-value < 0.01) all show signs of increased binding to regulatory regions of these genes.

### 6.3 Tri-methylation of histone across 5 human cell lines

We performed similar analysis on two epigenomics histone marks, H3k4me3 and H3k27me3 across five human cell line at different “differentiation” stages [12]. Following steps can be conducted to obtain the respective heatmaps.

#### 6.3.1 Select a data set from the portal

Go to “Experiments” tab and type ‘GSE11074’ in the keyword field of ‘Filter experiments’ option. You can also find this experiment under ‘Epigenomics’

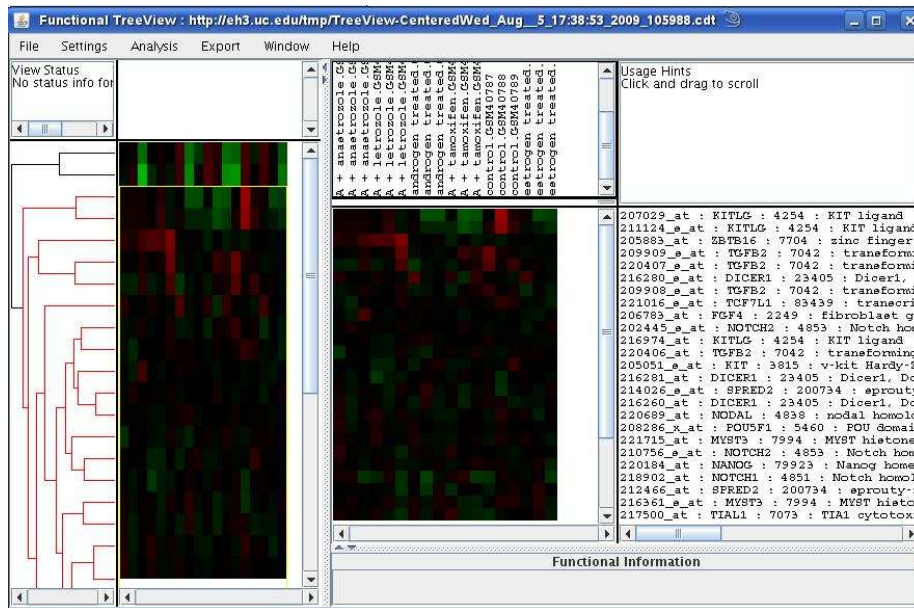


Figure 32: TreeView

portal. Press submit. This will fetch the corresponding experiment and then press “Query” button.

### 6.3.2 Paste a query gene list

Paste a list of Entrez ids of 85 up regulated proliferation genes in the box (option 3) and press “submit”.

### 6.3.3 Select Sample Grouping

In step1 (select samples for analysis), click on sample grouping name 'Histone'. This will show 2 options namely H3k4me3 and H3K27me3. We want to analyze the 2 histones separately. Choose H3K4me3 first by checking radio button 'include'. This step will filter samples in the analysis. In this case it will include only one type of selected histone. In step2, select “cell” as sample grouping for further analysis. Then choose clustering on “Genes” and press “Analyze”. Figure 44 depicts the snapshot of this step.

### 6.3.4 Results

Click on “Centered data” link in the static heatmap column of the summary results table. Similar steps could be performed for other histone type. Figure 46 and figure 47 show heatmaps of the 2 histones respectively. Figure 45 shows legend for 5 cell types.



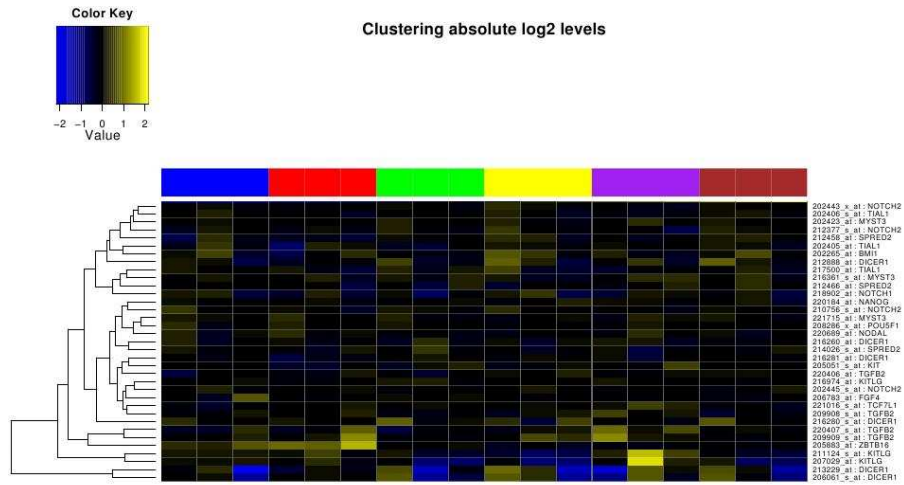


Figure 34: Static heatmap for stem cell genes

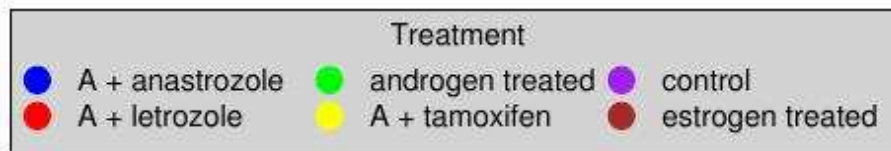


Figure 35: legend for 6 treatment types

p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. PNAS 2005, 102:13550-13555.

- [6] Fournier MV, Martin KJ, Kenny PA, Khaja K, Bosch I, Yaswen P, Bissell MJ: Gene Expression Signature in Organized and Growth-Arrested Mammary Acini Predicts Good Outcome in Breast Cancer. *Cancer Res* 2006, 66:7095-7102.
- [7] Herschkowitz J, Simin K, Weigman V, Mikaelian I, Usary J, Hu Z, Rasmussen K, Jones L, Assefnia S, Chandrasekharan S et al.: Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. *Genome Biology* 2007, 8:R76.
- [8] Moggs JG, Murphy TC, Lim FL, Moore DJ, Stuckey R, Antrobus K, Kimber I, Orphanides G: Anti-proliferative effect of estrogen in breast cancer cells that re-express ER{alpha} is mediated by aberrant regulation of cell cycle genes. *J Mol Endocrinol* 2005, 34:535-551.
- [9] Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, Wong E, Orlov YL, Zhang W, Jiang J et al.: Integration of External Signaling Pathways with

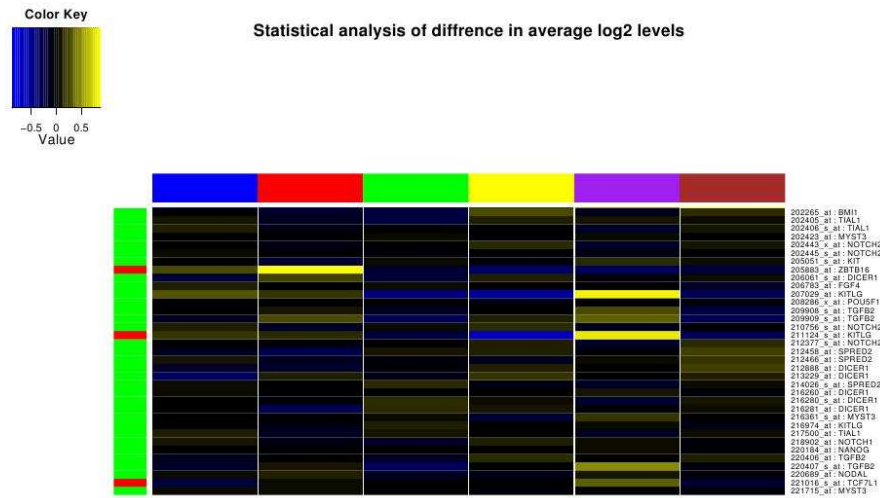


Figure 36: Statistical Analysis of stem cell genes

the Core Transcriptional Network in Embryonic Stem Cells. *Cell* 2008, 133:1106-1117.

- [10] Xu X, Bieda M, Jin VX, Rabinovich A, Oberley MJ, Green R, Farnham PJ: A comprehensive ChIP-chip analysis of E2F1, E2F4, and E2F6 in normal and tumor cells reveals interchangeable roles of E2F family members. *Genome Res* 2007, 17:1550-1561.
- [11] Kalma Y, Marash L, Lamed Y, Ginsberg D: Expression analysis using DNA microarrays demonstrates that E2F-1 up-regulates expression of DNA replication genes including replication protein A2 3. *Oncogene* 2001, 20:1379-1387.
- [12] Mikkelsen TS, Hanna J, Zhang X, Ku M, Wernig M, Schorderet P, Bernstein BE, Jaenisch R, Lander ES, Meissner A: Dissecting direct reprogramming through integrative genomic analysis 2. *Nature* 2008, 454:49-55.

Description	Reference
251 Affy U133A microarrays hybridized with mRNA from primary breast tumors. <a href="#">GSE3494Entrez</a>	Miller, L. D. et al. (2005) From The Cover: An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. Proceedings of the Natl <a href="#">Pub Med</a>

View genes found in the platform

#Samples	#Probes	#Genes	Data download
251	69	69	<a href="#">Tabular format (xls)</a> <a href="#">RData (eset)</a>

**Step 1 (optional) Select samples for analysis. [demo](#)**

include  exclude  
 ID  
 p53seg  
 p53DLDA  
 DLDA  
 Elston  
 ER  
 PgR  
 age-quartiles  
 tumorSize-quartiles  
 Lymph  
 DSSTime-quartiles  
 DSSEvent

**Step 2) Select sample grouping for analysis**

none  
 ID  
 p53seg  
 p53DLDA  
 DLDA  
 Elston  
 ER  
 PgR  
 age-quartiles  
 tumorSize-quartiles  
 Lymph  
 DSSTime-quartiles  
 DSSEvent

Cluster on   Compute LR

Figure 37: Proliferation genes on Miller dataset

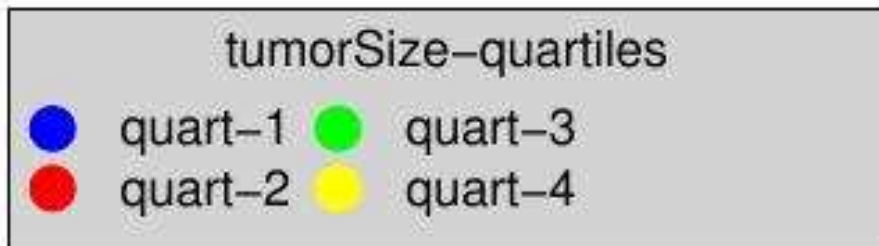


Figure 38: Legend for Tumor Size grade

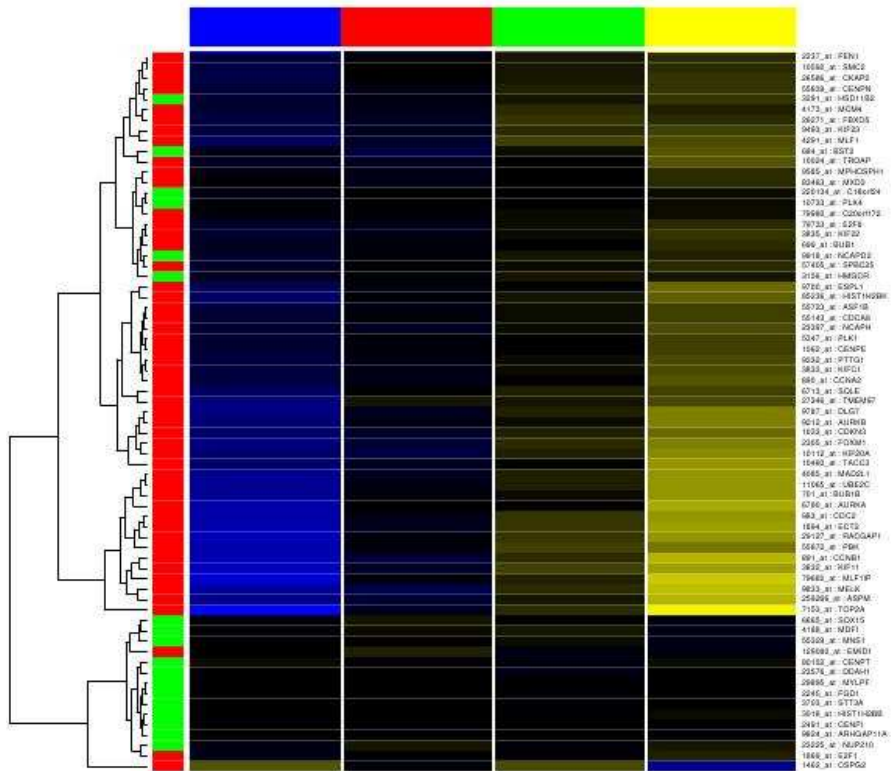


Figure 39: Statistical significance of up regulated genes

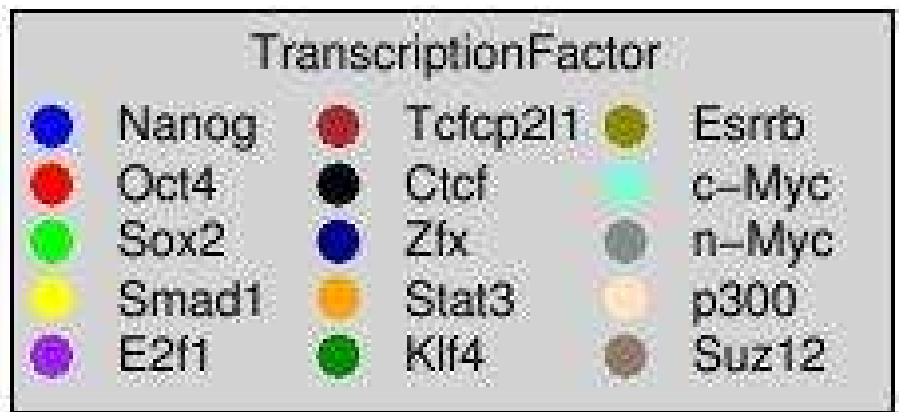


Figure 40: Legend for 15 TFs



Description	Reference
ChIP-seq experiment to map the locations of thirteen sequence specific transcription factors (Nanog, Oct4, STAT3, Smad1, Sox2, Zfx, c-Myc, n-Myc, Klf4, Esrrb, Tcfcp2l1, E2f1 and CTCF) and two transcription regulators (p300 and Suz12). Promoter regions, starting from 4000 bps upstream and ending 1000 bps downstream of the TSS, are truncated into 50 bp regions and intensity of each region is summarized as the average of all intensities within the region. Results are log2 transformed and only peaks are included. <a href="#">GSE11431peaks</a>	Chen X, Xu H, Yuan P, et al. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell. 2008 Jun 13;133(6):1106-17. <a href="#">PubMed</a>

View genes found in the platform

#Samples	#Probes	#Genes	Data download
1500	95	85	<a href="#">Tabular format (xls)</a> <a href="#">RData (reset)</a>

#### Step 1 (optional) Select samples for analysis. [demo](#)

include  exclude  
 sample.distance  
 sample  
    TranscriptionFactor
    

- Nanog
- Oct4
- Sox2
- Smad1
- E2f1
- Tcfcp2l1
- Ctcf
- Zfx
- Stat3
- Klf4
- Esrrb
- c-Myc
- n-Myc
- p300
- Suz12

 distance

#### Step 2) Select sample grouping for analysis

none  
 sample.distance  
 sample  
 TranscriptionFactor  
 distance

Cluster on   Compute LR

Figure 42: Select n-Myc TF

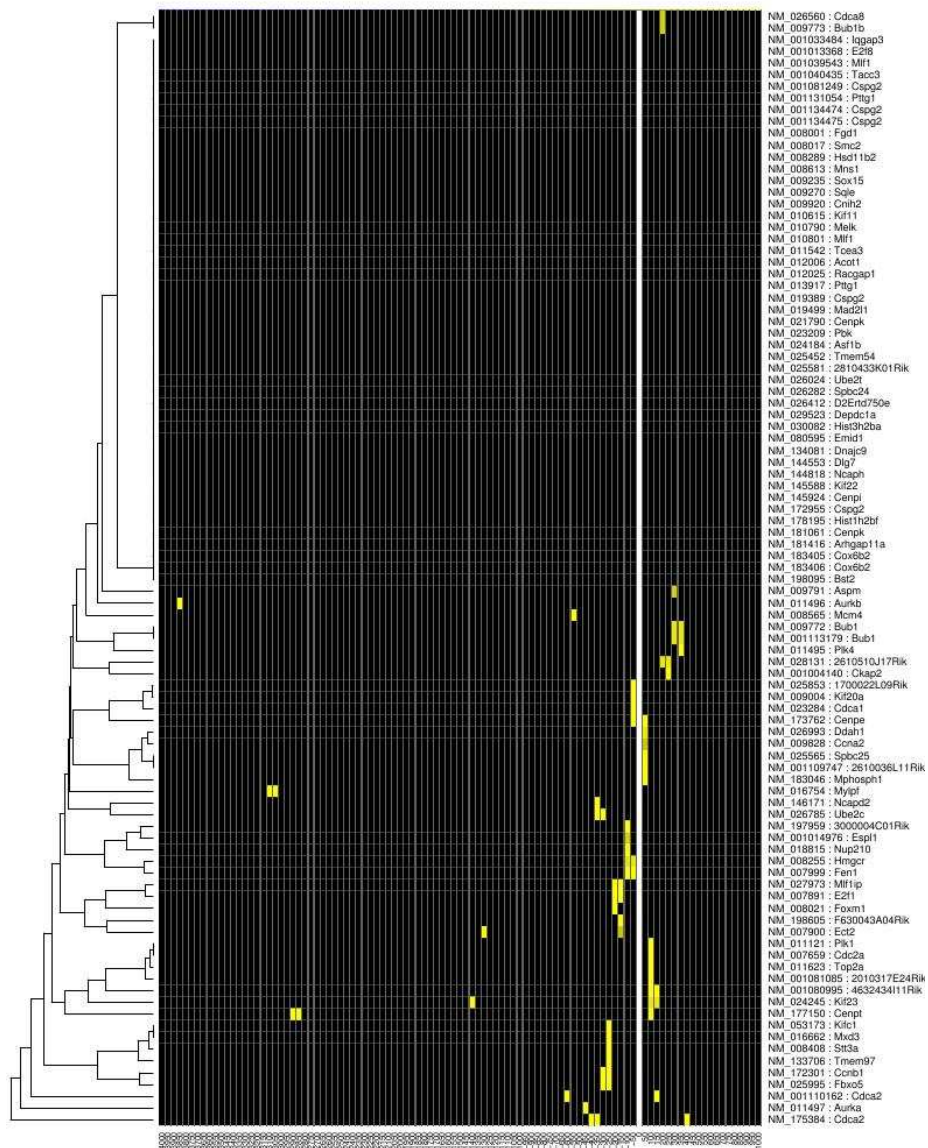


Figure 43: n-Myc TF heatmap

Description			Reference
Chromatin state maps (H3K4me3 and H3K27me3) from partially and fully reprogrammed mouse cell lines. Promoter regions, starting from 4000 bps upstream and ending 1000 bps downstream of the TSS, are truncated into 50 bps regions and intensity of each region is summarized as the average of all raw signal intensities within the region. <a href="#">GSE11074</a>			Mikkelsen TS, Hanna J, Zhang X, Ku M et al. Dissecting direct reprogramming through integrative genomic analysis. Nature 2008 Jul 3;454(7200):49-55. <a href="#">Pub Med</a>

View genes found in the platform

#Samples	#Probes	#Genes	Data download
1000	95	85	<a href="#">Tabular format (xls)</a> <a href="#">RData (eset)</a>

**Step 1 (optional) Select samples for analysis. [demo](#)**

include  exclude

- Sample
- Distance
- Cell
  - Histone
    - H3K4me3
    - H3K27me3
- Cell.Histone
- Sample.Distance
- Cell.Distance
- Histone.Distance

**Step 2) Select sample grouping for analysis**

- none
- Sample
- Distance
- Cell
- Histone
- Cell.Histone

Cluster on Genes  Compute LR Analyze

Figure 44: Filter H3k4me3 histone samples

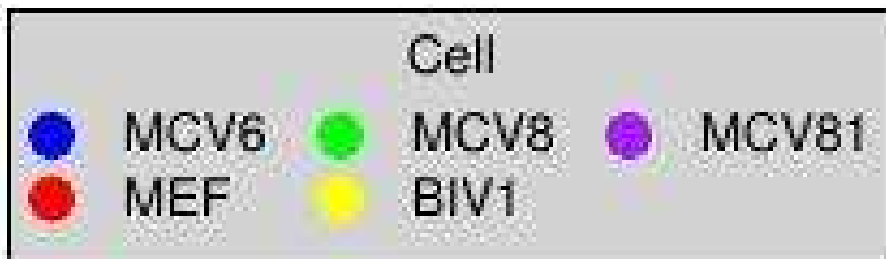


Figure 45: Legend for 5 cell types



