

Use Case 15: Epigenomic data exploration of ENCODE and Human Roadmap Epigenome project with SPARK

Epigenome Informatics Workshop Bioinformatics Research Laboratory



SPARK and GREAT tools References

Nielsen, C., Younesy, H., O'Geen, H., Xu, X., Jackson, A., Milosavljevic, A., Wang, T., Costello, J., Hirst, M., Farnham, P., et al. (2012). Spark: a navigational paradigm for genomic data exploration. *Genome Research* 22, 2262–2269.

McLean, C., Bristor, D., Hiller, M., Clarke, S., Schaar, B., Lowe, C., Wenger, A., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnology* 28, 495–501.

Summary of Nielsen et al Manuscript

Background: Biologists possess the detailed knowledge critical for extracting biological insight from genome-wide data resources, and yet they are increasingly faced with nontrivial computational analysis challenge posed by genome-scale methodologies. To lower this computational barrier, Nielsen et al developed an interactive pattern discovery and visualization tool, Spark, designed with epigenomic data in mind. For instance, Spark can be used to reveal epigenetic signatures or patterns at user specified regions of genomic coordinates e.g., TSS or ChIP-seq of transcription factor.

In this use case, we will demonstrate how SPARK can be used to explore ENCODE and Human Roadmap Epigenome project datasets.

Tool: SPARK Workflow

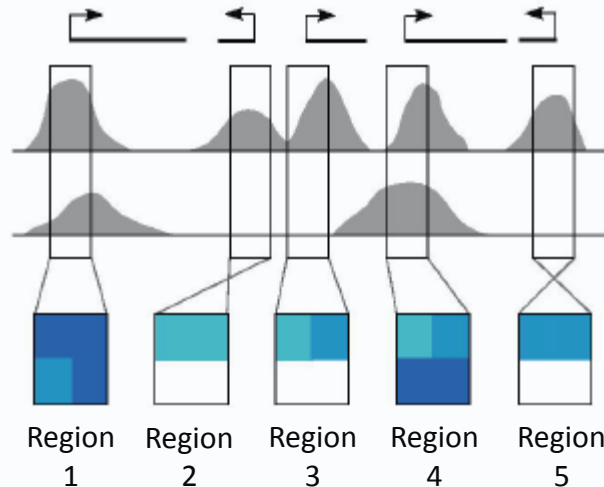
Step 1: Preprocessing

ChIP-seq –Txn Factor1

ChIP-seq –Txn Factor2

Txn Factor1

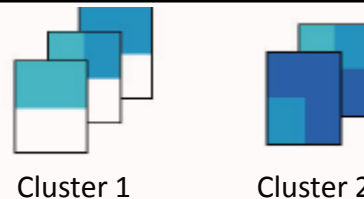
Txn Factor2



User specified regions of genomic coordinates
eg.
5 TSS Regions

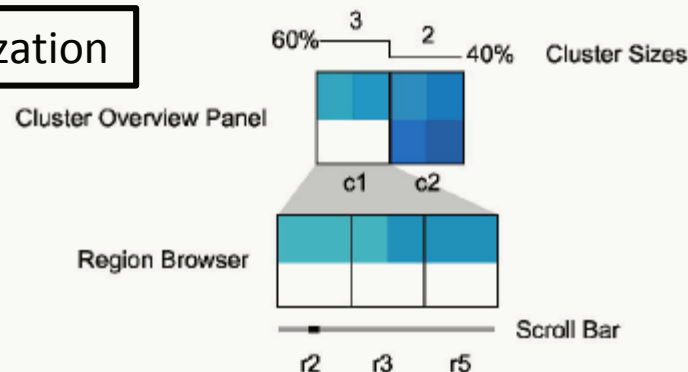
Color represents signal intensity at these regions

Step 2: Clustering



Regions are clustered through *k*-means clustering

Step 3: Interactive Visualization

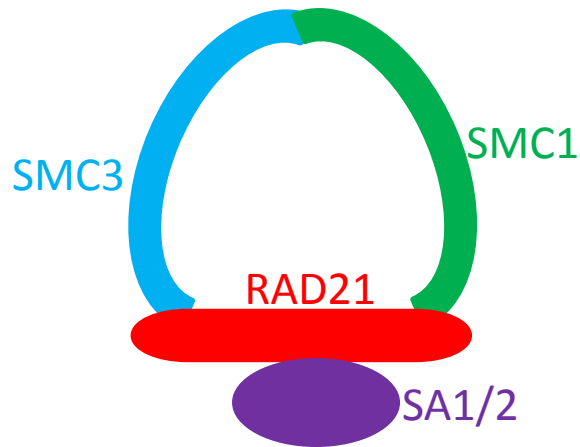


SPARK tool for epigenomic data exploration

Objective:

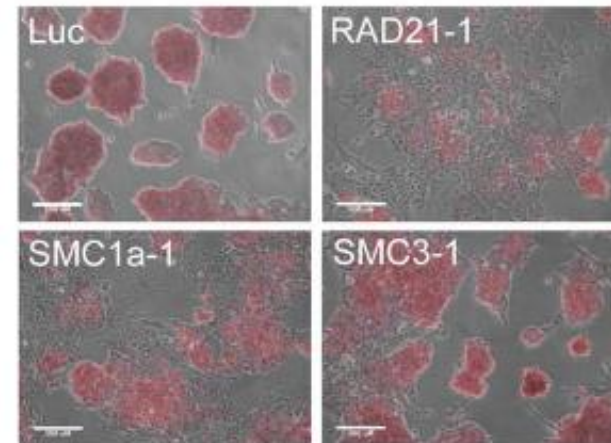
We will study biology of cohesion in human embryonic stem cells (ESCs). Subunit of cohesion, RAD21, predominantly binds together with CTCF. However, subset of RAD21 binding sites are independent of CTCF in ESCs. Surprisingly these regions (CTCF independent RAD21 binding sites) are co-localized with pluripotent transcription factors (NANOG, OCT4, KLF4) and therefore RAD21 are important in maintaining stem cell self-renewal. However its unclear mechanism by which RAD21 play role in ESCs self-renewal. Therefore, objective of this use case is to identify role of CTCF independent RAD21 sites in ESCs.

Cohesion subunit is implicated to play role in maintenance and self-renewal of Embryonic Stem Cells



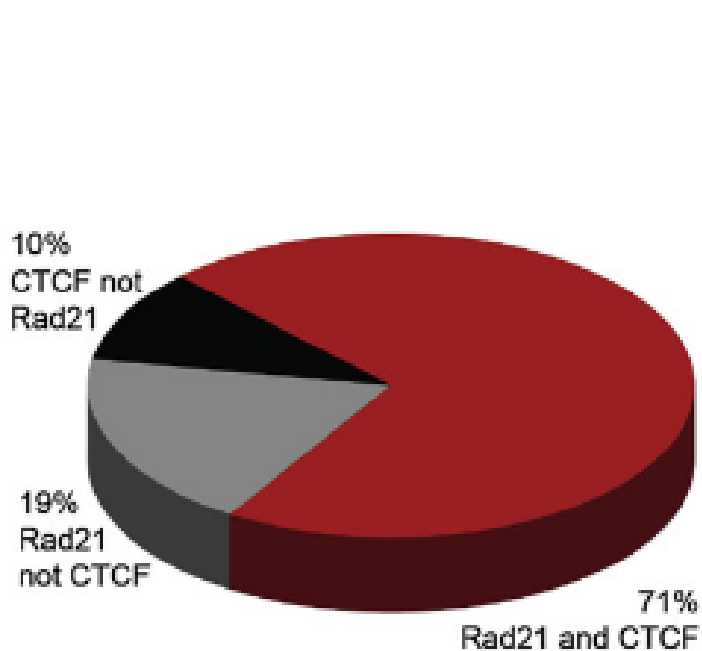
Vertebrate Cohesin comprises of a ring made of three subunits – SMC3, SMC1 and RAD21, and an additional protein SA1 or SA2

Alkaline phosphatase staining

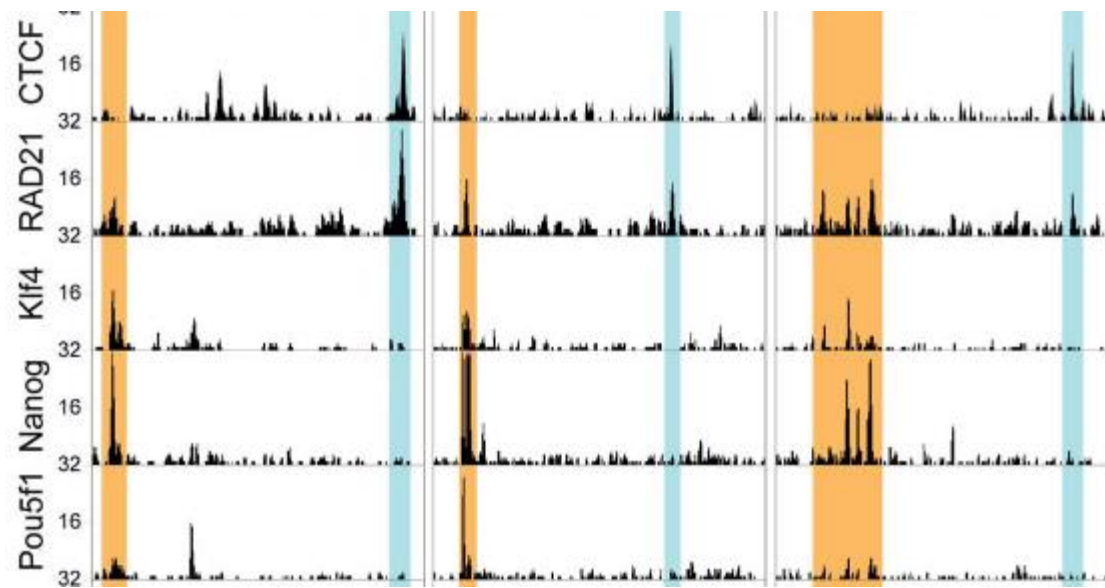


Depletion of RAD21 and SMC1a leads to differentiation of ESCs shown by reduced alkaline phosphatase staining

RAD21 colocalizes with pluripotency related transcription factors at CTCF-independent sites



- CTCF independent RAD21 sites preferentially co-localized with pluripotent transcription factors
- CTCF dependent RAD21 sites



Objective is to identify role of CTCF independent RAD21 sites in embryonic stem cells

CTCF-independent RAD21 binding sites preferentially co-localize with key pluripotency related transcription factors

RAD21 binding sites are used as region of interests

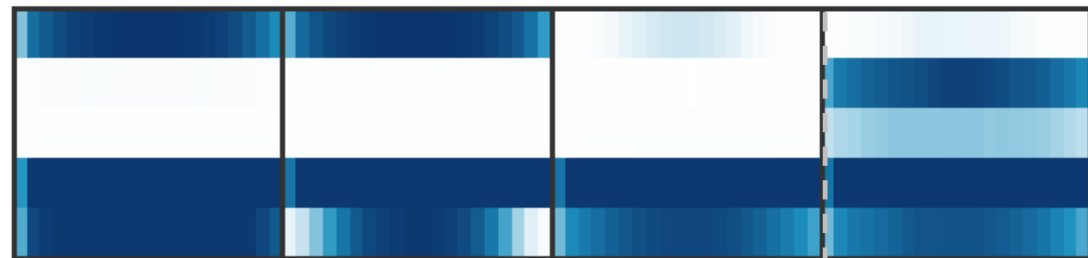
Cluster sizes (55674 regions clustered)



ENCODE dataset

Cluster profiles

H1hes: Ctfsc5916_V0416102
H1hes: Nanogsc33759_V0416102
H1hes: Pou5f1sc9081_V0416102
H1hes: Rad21_Iggrab
H1hes: Rad21_V0416102



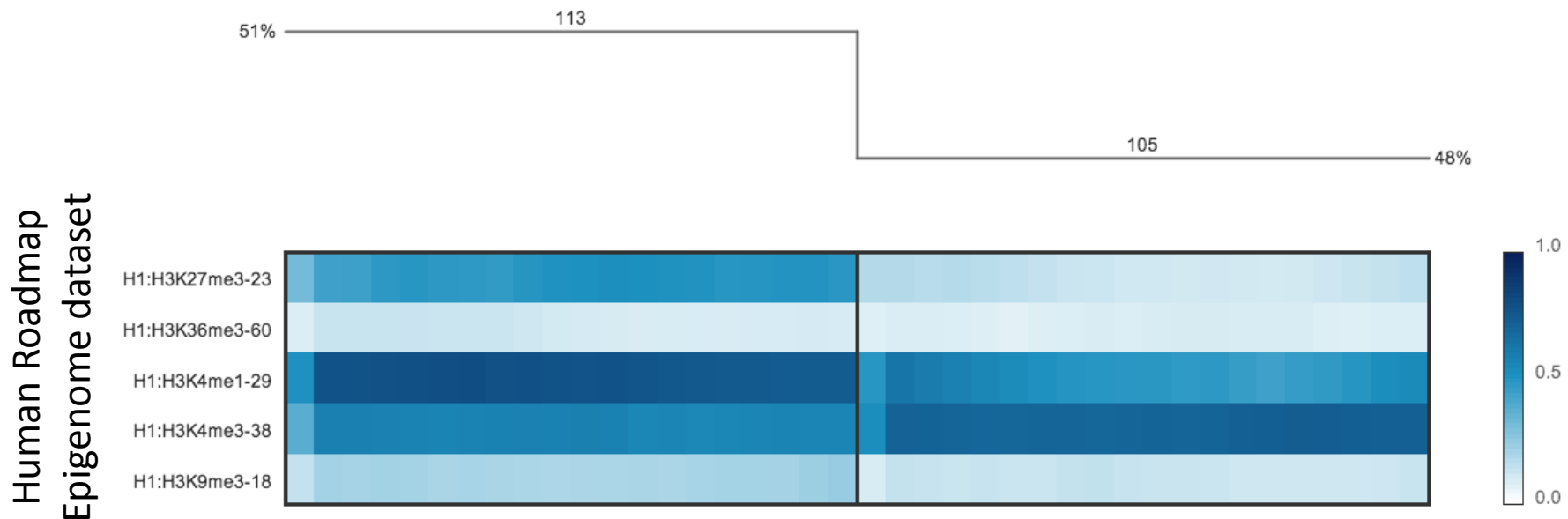
CTCF **dependent**
RAD21 regions

CTCF **independent**
RAD21 regions

RD_N regions are composed of distal cis-regulatory elements and promoters based on enriched H3K4me1 and H3K4me3 signals

RAD21 and Nanog co-occupancy regions = RD_N regions

Cluster sizes (218 regions clustered)



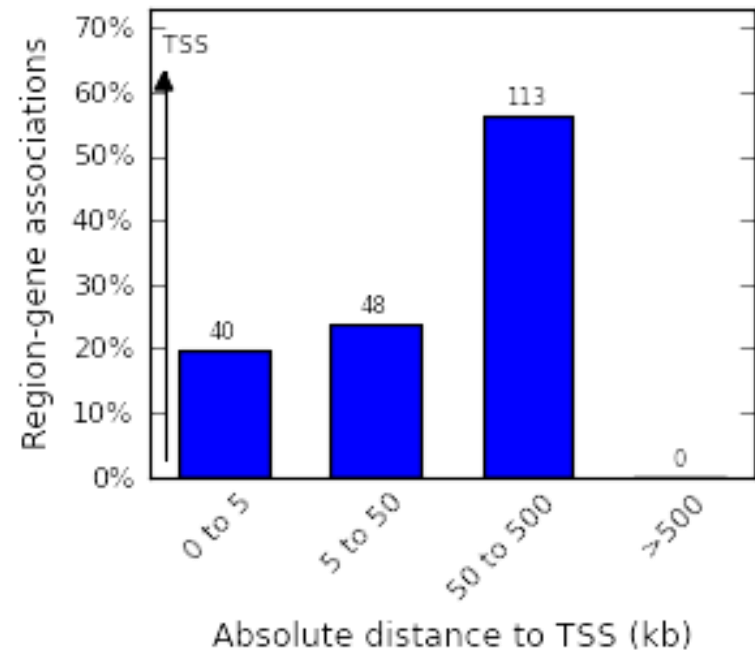
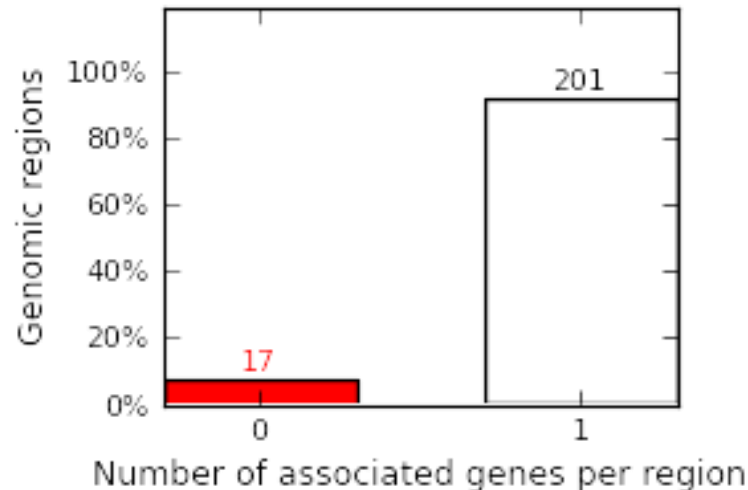
H3K4me1 signal is associated with enhancers and distal cis-regulatory element

H3K4me3 signal is associated with promoter

50% of RD_N regions are distal cis-regulatory elements

GREAT tool predicts functions of cis-regulatory regions by assigning genomic regions to nearby genes

GREAT Tool analysis of 218 genomic regions



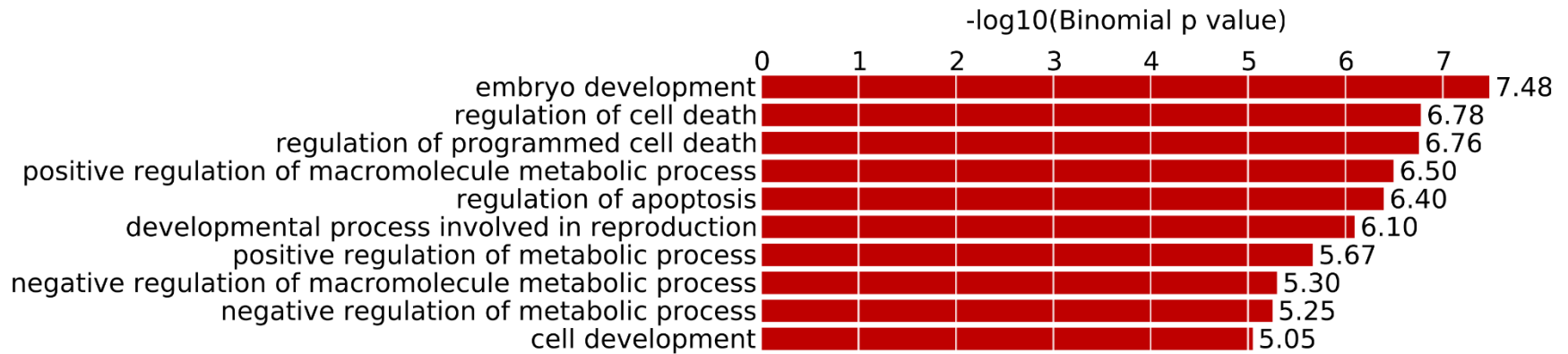
GREAT tool region-gene associations correlates well with epigenomic predictions on number of distal cis-regulatory elements and promoter

Genes associated with RD_N regions enrich GO terms such as 'Nanog targets' in ESCs

GREAT tools assigns biological meaning to the cis-regulatory associated genes by looking for enrichment of these gene sets in GO databases

# Term Name	Binom Rank	Binom Raw P-Value	Binom FDR Q-Val	Binom Fold Enrichment
Set 'Nanog targets': genes upregulated and identified by ChIP on chip as Nanog [Gene ID=79923] transcription factor targets in human embryonic stem cells.	1	2.21207e-10	5.25145e-7	2.4664
Set 'NOS targets': genes upregulated and identified by ChIP on chip as targets of the transcription factors NANOG [Gene ID=79923], OCT4[Gene ID=5460], and Sox2 [Gene ID=6657] (NOS) in human embryonic stem cells.	2	9.85145e-7	1.16937e-3	3.4715
Set 'Sox2 targets': genes upregulated and identified by ChIP on chip as SOX2 [Gene ID=6657] transcription factor targets in human embryonic stem cells.	3	1.36210e-5	1.07787e-2	2.1341
Set 'Oct4 targets': genes upregulated and identified by ChIP on chip as OCT4 [Gene ID=5460] transcription factor targets in human embryonic stem cells.	4	3.37562e-5	2.00343e-2	2.5278
Genes down-regulated in mice with skin specific knockout of RB1 [Gene ID=5925] by Cre-lox.	5	4.39191e-5	2.08528e-2	2.3275

GO Biological process enrichment analysis show that RD_N region-genes are associated with cell death and apoptosis



This suggests that CTCF independent RAD21 colocalized with Nanog (RD_N regions) are involved in regulating genes related to programmed cell death/apoptosis in ESCs and thus help maintain ESCs self-renewal

The following slides will walk you through the process of reproducing the results showed in previous slides.

Additionally there is a short tutorial describing usage of SPARK in Genboree.

<http://vimeo.com/48404125>

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

Welcome to the Genboree Workbench! [\[Getting Started\]](#)

Data Selector

Refresh Data Filter: Select a filter

- ☐ H1hesc:Ctcfsc5916_V0416102
- ☐ H1hesc:Egr1_V0416102
- ☐ H1hesc:Fos1sc183_V0416102
- ☐ H1hesc:Gapb_Pcr1x
- ☐ H1hesc:Gtf2f1_Iggrab
- ☐ H1hesc:Hdac2sc6296_V0416102
- ☐ H1hesc:Jund_Iggrab
- ☐ H1hesc:Jund_V0416102
- ☐ H1hesc:Mafk_Iggrab
- ☐ H1hesc:Mxi1_Iggrab
- ☐ H1hesc:Nanogsc33759_V0416102
- ☐ H1hesc:Nrf1_Iggrab
- ☐ H1hesc:Nrsf_V0416102
- ☐ H1hesc:P300_V0416102
- ☐ H1hesc:Pol24h8_V0416102
- ☐ H1hesc:Pol2_V0416102
- ☐ H1hesc:Pou5f1sc9081_V0416102
- ☐ H1hesc:Rad21_Iggrab
- ☐ H1hesc:Rad21_V0416102
- ☐ H1hesc:Rfx5200401194_Iggrab
- ☐ H1hesc:Rxra_V0416102
- ☐ H1hesc:Sin3ak20_Pcr1x

Input Data

- ☐ H1hesc:Ctcfsc5916_V0416102
- ☐ H1hesc:Nanogsc33759_V0416102
- ☐ H1hesc:Pou5f1sc9081_V0416102
- ☐ H1hesc:Rad21_V0416102
- ☐ H1hesc:Rad21_Iggrab

Output Targets

Drag

Step 1. Populate "Input Data"

In "Data Selector"

Expand "Epigenome ToolSet Demo Input Data" > "Databases"

Expand "Binding Sites Demo" > "Tracks" > "Class: ENCODE – T.f. Binding Site Data"

Scroll down till you see tracks begin with "H1hesc:___"

Drag following five datasets

"H1hesc:Ctcfsc5916_V0416102", "H1hesc: Nanogsc33759_V0416102",

"H1hesc:Pou5f1sc9081_V0416102", "H1hesc:Rad21_V0416102",

"H1hesc:Rad21_Iggrab"

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

Welcome to the Genboree Workbench! [Getting Started]

Data Selector

Refresh Data Filter: Select a filter

- H1hesc:Hdac2sc6296V0416102
- H1hesc:Jundlggrab
- H1hesc:JundV0416102
- H1hesc:MaxUcd
- H1hesc:Mxi1lggrab
- H1hesc:Nanogsc33759V0416102
- H1hesc:Nrf1lggrab
- H1hesc:Nrf1V0416102
- H1hesc:P300V0416102
- H1hesc:Pol24h8V0416102
- H1hesc:Pol2V0416102
- H1hesc:Pou5f1sc9081V0416102
- H1hesc:Rad21lggrab
- H1hesc:Rad21V0416102
- H1hesc:Rfx5200401194lggrab
- H1hesc:RxraV0416102
- H1hesc:Sin3ak20Pcr1x
- H1hesc:Sin3anb6001263lggrab
- H1hesc:Six5Pcr1x
- H1hesc:Sp1Pcr1x
- H1hesc:Sp2V0422111

Input Data

- H1hesc:Nanogsc33759_V0416102
- H1hesc:Pou5f1sc9081_V0416102
- H1hesc:Rad21_V0416102
- H1hesc:Rad21lggrab
- H1hesc:Rad21_Iggrab

Output Targets

- GenboreeUser_database

Drag

Step 2. ADD region of interest (ROI)
In "Data Selector"
Expand "ROI Repository" > "Databases"
Expand "ROI Repository – hg19" > "Tracks" > "Class: ENCODE – T.f. Binding Site Data"
Scroll down till you see tracks begin with "H1hesc:___"
Drag following ROI
"H1hesc:Rad21lggrab"

Step 3. Drag your database (i.e. "GenboreeUser_database") to "Output Targets"

Data dragged in "Input Data" earlier (slide 15) were high density tracks – dataset with scores
Now we dragged region of interests (ROIs) which are BED files

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

Welcome to the Genboree Workbench! [Getting Started]

Data Selection

Refresh Data Filter: Select a filter... Attribute

Databases

- ROI Repository - hg18
- ROI Repository - hg19
- All Annotations in Database
- Tracks
 - Class: Affymetrix
 - Class: Agilent
 - Class: ENCODE
 - Class: ENCODE - T.f. Binding Sites
 - A549:Atf3V0422111Etoh02
 - A549:Bcl3V0422111Etoh02
 - A549:Bhlhe40lggrab
 - A549:Cebpblggrab
 - A549:Creb1sc240V0416102Dex100nm
 - A549:Ctcfsc5916Pcr1x100nm
 - A549:Ctcfsc5916Pcr1xEtoh02
 - A549:CtcfStd
 - A549:Elf1V0422111Etoh02
 - A549:Ets1V0422111Etoh02
 - A549:FosI2V0422111Etoh02
 - A549:Foxa1V0416102Dex100nm
 - A549:GabpV0422111Etoh02
 - A549:Gata1V0422111Etoh02

Find Differences By Regression

Cluster by Spark

Cluster by Spark

Perform the preprocessing and clustering steps of Cydney Nielson's Spark tool on signal data found in tracks or files you specify.

Once complete, view the results in Spark's stand-alone GUI.

Random Forest

QIIME

QC

Search for Similar Signals by Correlation

Analyze Signals

Compute Similarity Matrix (heatmap)

Genboree User Database

Output Targets

GenboreeUser_database

Step 4. Cluster by Spark

Click on "Epigenome" > "Analyze Signals" > "Cluster by Spark"

Cluster by Spark (Analyze Signals)

Tool Overview

Inputs:

Data	H1hesc:Ctcfsc5916_V0416102	Group: Epigenome ToolSet Demo Input Data, Database: Binding Sites Demo
Tracks/Files:	H1hesc:Nanogsc33759_V0416102	Group: Epigenome ToolSet Demo Input Data, Database: Binding Sites Demo
	H1hesc:Pou5f1sc9081_V0416102	Group: Epigenome ToolSet Demo Input Data, Database: Binding Sites Demo
	H1hesc:Rad21_V0416102	Group: Epigenome ToolSet Demo Input Data, Database: Binding Sites Demo
	H1hesc:Rad21Iggrab	Group: ROI Repository, Database: ROI Repository - hg19
	H1hesc:Rad21_Iggrab	Group: Epigenome ToolSet Demo Input Data, Database: Binding Sites Demo

Output Database:

Database: GenboreeUser_database Group: GenboreeUser_group

Spark Analysis Settings

Analysis Name Rad21_H1

Select ROI Track
H1hesc:Nanogsc33759_V0416102
H1hesc:Pou5f1sc9081_V0416102
H1hesc:Rad21_Iggrab
H1hesc:Rad21_V0416102
H1hesc:Rad21Iggrab

Region Label MyROIs

Statistics Type global

of Clusters 3

of Bins: 20

Data Track Colors:

H1hesc:Ctcfsc5916_V0416102	blue
H1hesc:Nanogsc33759_V0416102	blue
H1hesc:Pou5f1sc9081_V0416102	blue
H1hesc:Rad21_Iggrab	blue
H1hesc:Rad21_V0416102	blue
H1hesc:Rad21Iggrab	blue

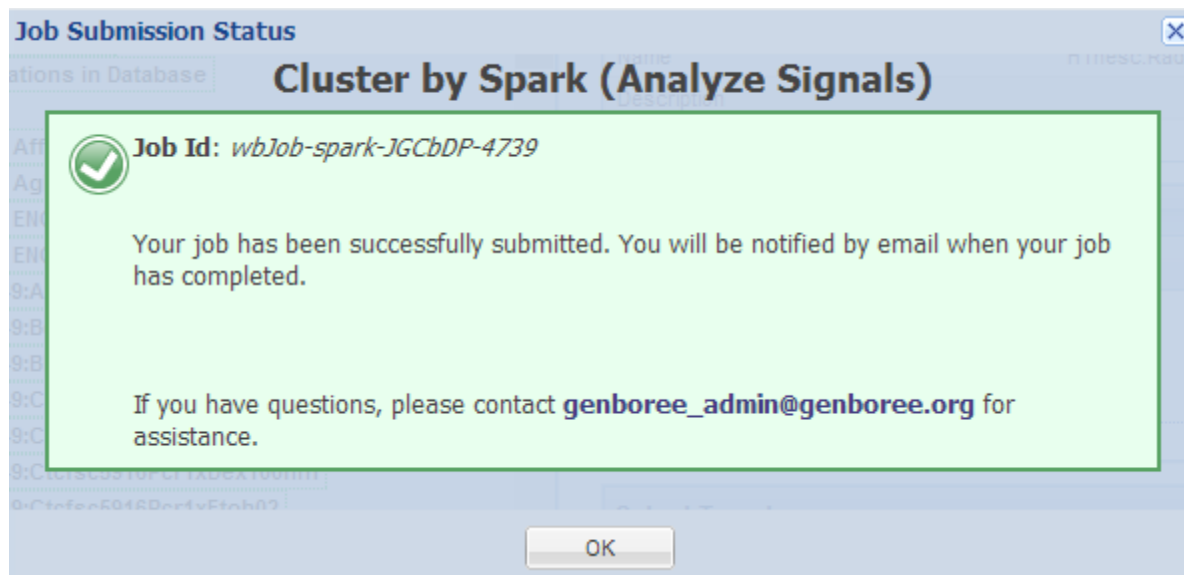
Step 5. Type in the Analysis Name "Rad21_H1"

Step 6. In Select ROI Track click on
"H1hesc:Rad21Iggrab"
DO NOT Select "H1hesc:Rad21_Iggrab" this is
score track not ROI

Step 7. Click on "Submit"

Submit Cancel

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



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

Your Spark job completed successfully.

Job Summary:

JobID - wbJob-spark-KMq1HG-5703

Analysis Name - Rad21_H1

Inputs:

of Data Tracks - 5

ROI Track - H1hesc:Rad21lggrab

Outputs:

Output DB - Dummy

Output Host - genboree.org

Settings:

k - 3

normType - exp

numBins - 20

regionLabel -

statsType - global

Additional Info:

To view your results in the Spark GUI:

(a) download and unzip the results archive and then

(b) launch Spark via Java Web Start and open the analysis folder.

Spark Java Web Start Link:

<http://www.bcgsc.ca/downloads/spark/current/start.jnlp>

Step 8. Click on the link to Download SPARK GUI.
Make sure your Java is updated

- The Genboree Team

Result File Location in the Genboree Workbench:

(Direct links to files are at the end of this email)

Host: genboree.org

Grp: vamin_group

Db: Dummy

Files Area:

* Spark - Results/

* Rad21_H1/

* ./Rad21_H1.zip

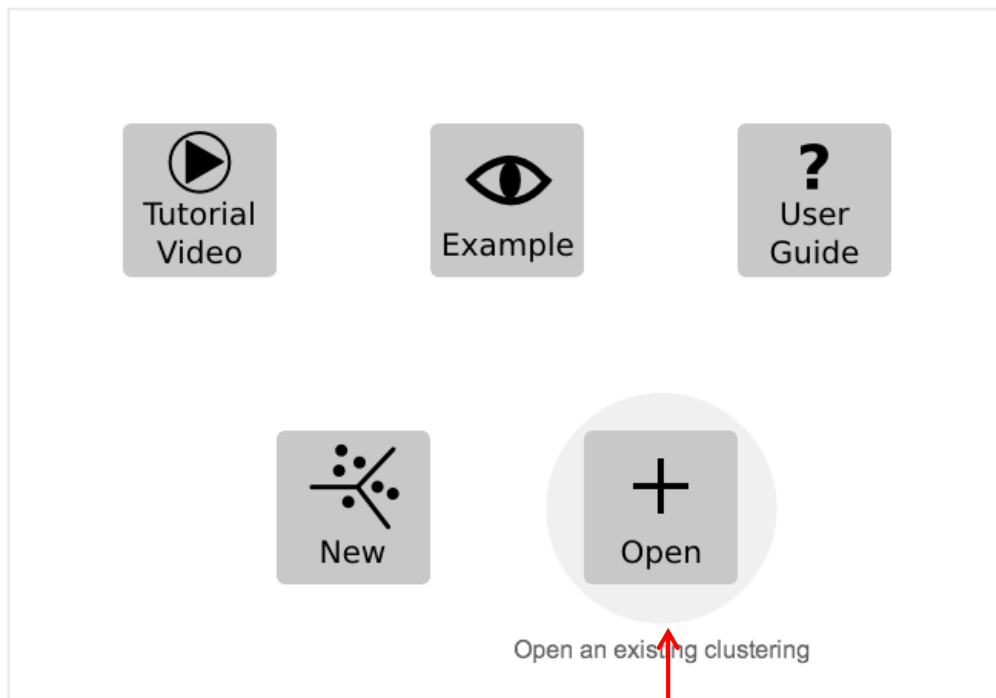
Step 9. Click on the link to Download "Rad21_H1.zip" folder
and extract all to designated location in your computer

Result File URLs (click or paste in browser to access file):

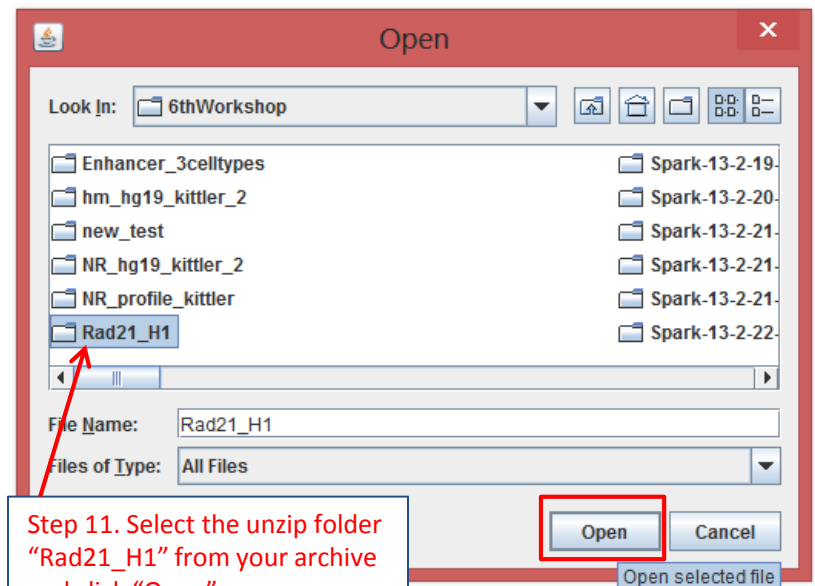
FILE: Rad21_H1.zip

URL:

http://genboree.org/java-bin/apiCaller.jsp?rsrcPath=http%3A%2F%2Fgenboree.org%2FREST%2Fv1%2Fgrp%2Fvamin_group%2Fdb%2FDummy%2Ffile%2FSpark%2520-%2520Results%2FRad21_H1%2FRad21_H1.zip%2Fdata%3F&fileDownload=true&promptForLogin=true&errorFormal=html

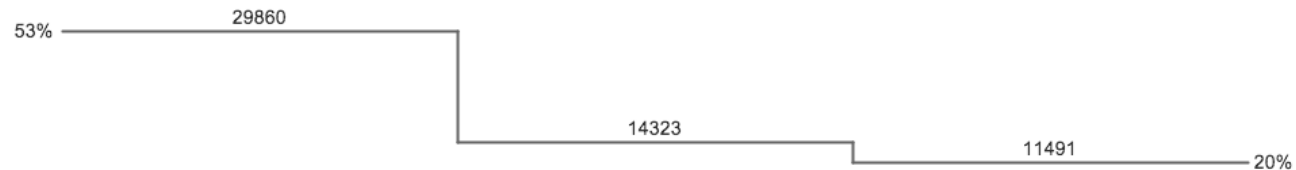


Step 10. Launch SPARK via Java Web Start



Step 11. Select the unzip folder "Rad21_H1" from your archive and click "Open"

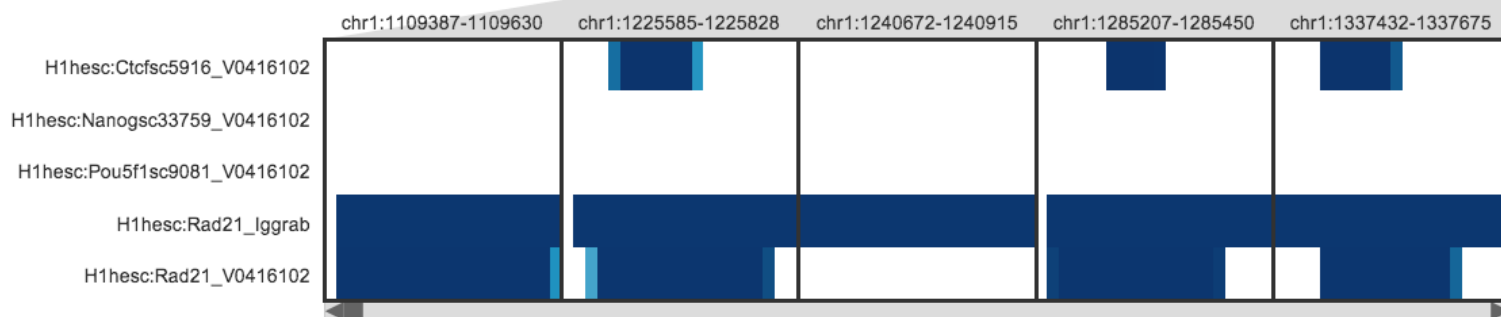
Cluster sizes (55674 regions clustered)



Cluster profiles



Region profiles



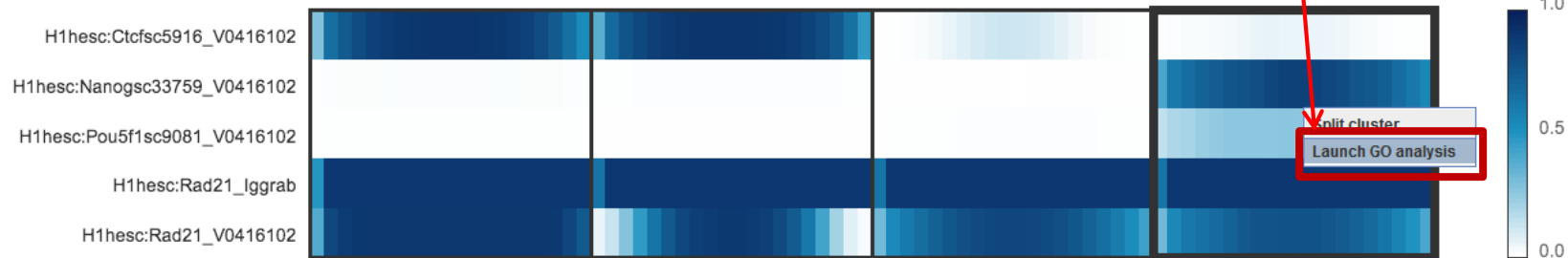
This allows you to visualize CTCF independent RAD21 binding sites co-localized with NANOG and POU5F1 (OCT4) regions

Here are steps to obtain CTCF independent RAD21 binding sites co-localized with NANOG and POU5F1 (OCT4) regions as bed file

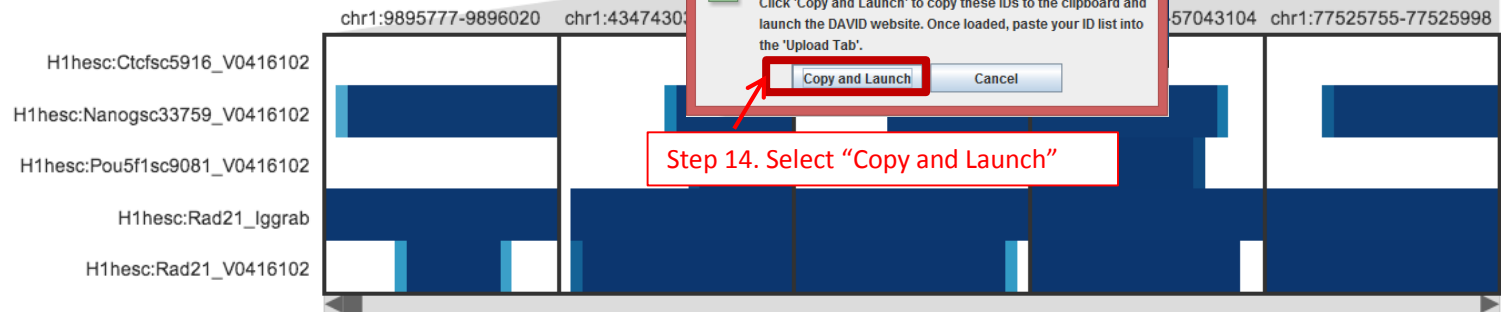
Cluster sizes (55674 regions clustered)



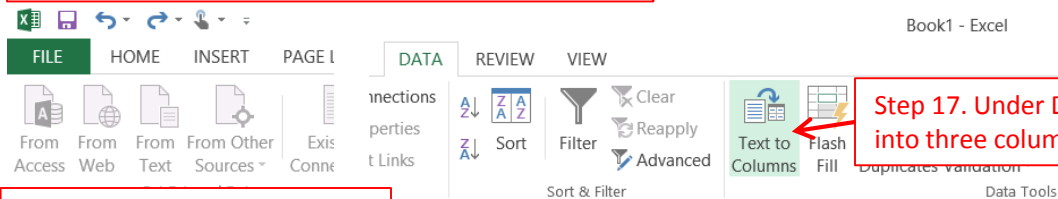
Cluster profiles



Region profiles

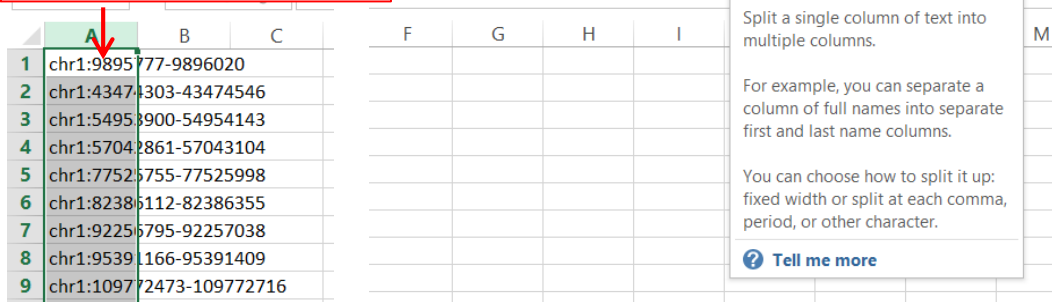


Step 15. Open Excel and paste regions

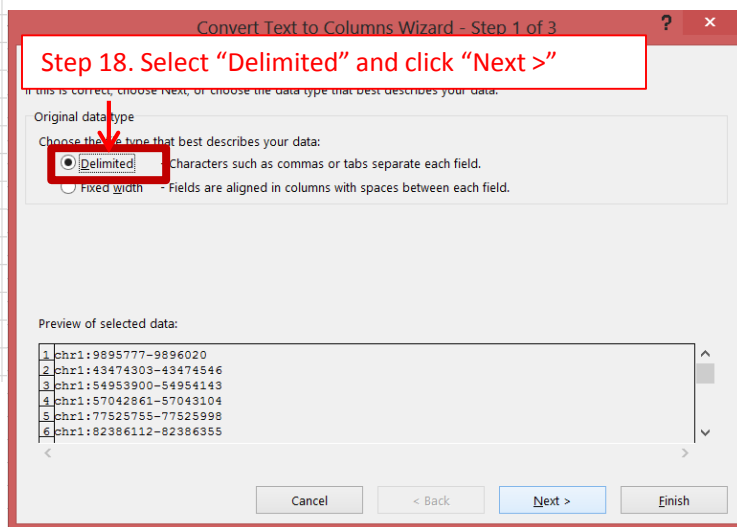


Step 17. Under DATA select "Text to Columns" to split file into three columns Chr#, start, and end

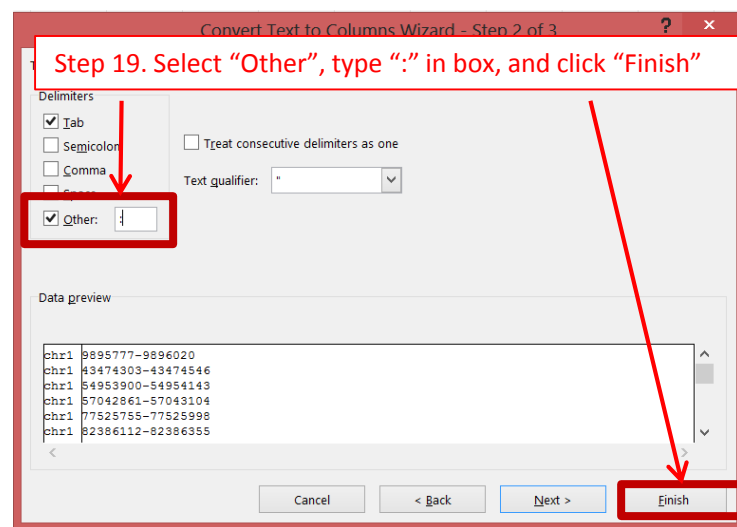
Step 16. Select entire column

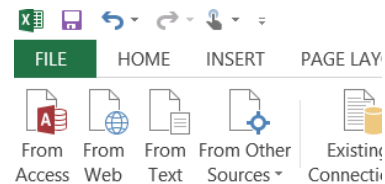


Step 18. Select "Delimited" and click "Next >"



Step 19. Select "Other", type ":" in box, and click "Finish"





Book1 - Excel

Step 20. Select entire column to split start and end sites

	A	B	C
1	chr1	9895777-9896020	
2	chr1	43474303-43474546	
3	chr1	54953900-54954143	
4	chr1	57042861-57043104	
5	chr1	77525755-77525998	
6	chr1	82386112-82386355	
7	chr1	92256795-92257038	
8	chr1	95391166-95391409	
9	chr1	109772473-109772716	
10	chr1	118385538-118385781	
11	chr1	145096240-145096483	
12	chr1	150540199-150540442	
13	chr1	164571887-164572130	
14	chr1	201979675-201979918	
15	chr1	203954281-203954524	
16	chr1	204063520-204063763	
17	chr1	205270600-205270843	
18	chr1	213843640-213843883	
19	chr1	214856989-214857232	
20	chr1	218759297-218759540	
21	chr1	218847862-218848105	
22	chr1	223101662-223101905	
23	chr2	116227171-11622960	

Step 21. Under DATA select "Text to Columns"

Text to Columns

Split a single column of text into multiple columns.

For example, you can separate a column of full names into separate first and last name columns.

You can choose how to split it up: fixed width or split at each comma, period, or other character.

[Tell me more](#)

Step 22. Select "Delimited" and click "Next >"

If this is correct, choose Next, or choose the data type that best describes your data.

Original data type

Choose the data type that best describes your data:

- ☒ **Delimited** - Characters such as commas or tabs separate each field.
☐ Fixed width - Fields are aligned in columns with spaces between each field.

Preview of selected data:

```
1 9895777-9896020
2 43474303-43474546
3 54953900-54954143
4 57042861-57043104
5 77525755-77525998
6 82386112-82386355
```

Cancel

< Back

Next >

Finish

Convert Text to Columns Wizard - Step 2 of 3

Step 23. Select "Other", type "-", and click "Finish"

Delimiters

- ☒ Tab
☐ Semicolon
☐ Comma
☐ Space
☒ Other: -
- ☐ Treat consecutive delimiters as one
 Text qualifier: "

Data preview

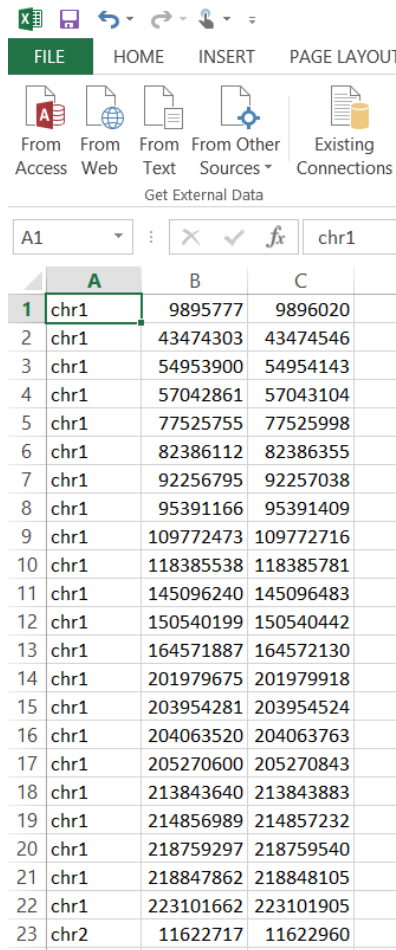
```
9895777 9896020
43474303 43474546
54953900 54954143
57042861 57043104
77525755 77525998
82386112 82386355
```

Cancel

< Back

Next >

Finish



	A	B	C
1	chr1	9895777	9896020
2	chr1	43474303	43474546
3	chr1	54953900	54954143
4	chr1	57042861	57043104
5	chr1	77525755	77525998
6	chr1	82386112	82386355
7	chr1	92256795	92257038
8	chr1	95391166	95391409
9	chr1	109772473	109772716
10	chr1	118385538	118385781
11	chr1	145096240	145096483
12	chr1	150540199	150540442
13	chr1	164571887	164572130
14	chr1	201979675	201979918
15	chr1	203954281	203954524
16	chr1	204063520	204063763
17	chr1	205270600	205270843
18	chr1	213843640	213843883
19	chr1	214856989	214857232
20	chr1	218759297	218759540
21	chr1	218847862	218848105
22	chr1	223101662	223101905
23	chr2	11622717	11622960

Step 24. This will generate three columns – Chr#, Start position, and End position which are minimum requirement for generating bed file.

Save the file as **Text (Tab delimited) (*.txt)**

Do make sure that column B and C are numbers and not scientific or other format.

Next we want to upload file in Genboree

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

Welcome to Genboree! [Getting Started]

Data Selector

Refresh

genboree.org

- Atlas Tool
- EDACC
- Epigenome
- Epigenome ToolSet Demo Input Data
- Epigenomics Roadmap Repository
- GenboreeUser_group
 - Databases
 - GenboreeUser_database
 - Projects
 - GMT_Tutorial
 - JonathanMill_Lab
 - paithank_group
 - Public
 - ROI Repository
 - Targeted Atlases
 - vamin_group

Import

- Array Data
- Track Metadata
- Upload Track Annotations

Upload Track Annotations

Import track data into a Genboree database.

Output Targets

GenboreeUser_database

Step 26. Upload Track Annotations
Click on "Data" > "Tracks" > "Import" > "Upload Track Annotations"

Step 25. Populate "Output Targets"
In "Data Selector" expand ("double click") on your user group
-Expand "Databases"
-Drag your database (i.e. "GenboreeUser_database") to "Output Targets"

Drag

Tool Settings

Upload Track Annotations

Tool Overview

Input Data:

Data File: *n/a*

Output Location:

Database: *GenboreeUser_database* Group: *GenboreeUser_group*

Settings

Select File **Choose File** Rad21_Nano...endent.txt

Input Format **Bed**

Track Class **SPARK**

Track Name **ESCs** : **Rad21_Nanog**

☒ Skip non-assembly chromosomes

☒ 0 based and half open

☐ 1 based and fully closed

Submit **Cancel**

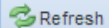
Step 27. Check the "Output Location" is correct
Click "Choose File" and upload bed file that was generated
Select "Input Format" as "Bed"
Specify "Track Class" name. Here its specified as "SPARK"
Specify "Track Name". Here its specified as "ESCs:Rad21_Nanog"

Keep everything else as default and click on "Submit"

You will see the message upon successful submission of your job

You will receive an email when your job is finished

Data Selector



Data Filter:

- ☐ H1:H3K4me3 98
- ☐ H1:H3K4me3 A
- ☐ H1:H3K4me3 C
- ☐ H1:H3K56ac 00
- ☐ H1:H3K56ac 11
- ☐ H1:H3K79me1 73
- ☐ H1:H3K79me1 82
- ☐ H1:H3K79me1 92
- ☐ H1:H3K79me2 04
- ☐ H1:H3K79me2 80
- ☐ H1:H3K9ac 26
- ☐ H1:H3K9ac 40
- ☐ H1:H3K9ac 62
- ☐ H1:H3K9ac 68
- ☐ H1:H3K9ac A
- ☐ H1:H3K9me3 18
- ☐ H1:H3K9me3 25

- GenboreeUser_group
 - Databases
 - GenboreeUser_database
 - All Annotations in Database
 - Tracks
 - Class: Class
 - Class: Gene
 - Class: High Density Score Data
 - Class: Marker
 - Class: Sequence
 - Class: SPARK
 - ESCs:Rad21_Nanog

Step 28. Pouplate "Input Data"

In "Data Selector"

Expand "Epigenomics Roadmap Repository" > "Databases"

Expand "Release 8 Repository" > "Tracks" > "Class: High Density Score Data"

Scroll down till you see tracks begin with "H1:___"

Drag following five histone modification tracks

"H1:H3K4me1 29", "H1: H3K36me3 60", "H1:H3K27me3 23", "H1:H3K9me3 18", "H1:H3K4me3 38"

Drag

Drag

Input Data

- ☐ H1:H3K4me1 29
- ☐ H1:H3K4me3 38
- ☐ H1:H3K36me3 60
- ☐ H1:H3K27me3 23
- ☐ H1:H3K9me3 18

Output Targets

- ☐ GenboreeUser_database

Step 29. Drag uploaded "ESCs:Rad21_Nanog" ROI track from your database into "Input Data"

Step 30. Drag your database (i.e. "GenboreeUser_database") to "Output Targets"

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

Welcome to the Genboree Workbench! [Getting Started]

Data Selector

Refresh Data Filter: Select a filter...

- genboree.org
 - Atlas Tools Access
 - EDACC
 - Epigenome Informatics Workshop (May 2012)
 - Epigenome ToolSet Demo Input Data
 - Epigenomics Roadmap Repository
 - Databases
 - Projects
 - GenboreeUser_group
 - Databases
 - GenboreeUser_database
 - All Annotations in Database
 - Tracks
 - Class: Class
 - Class: Gene
 - Class: High Density Score Data
 - Class: Marker
 - Class: Sequence
 - Class: SPARK
 - ESCs:Rad21_Nanog
 - Lists & Selections
 - SampleSets

Find Differences By Regression

Cluster by Spark

Cluster by Spark

Perform the preprocessing and clustering steps of Cydney Nielson's Spark tool on signal data found in tracks or files you specify.

Once complete, view the results in Spark's stand-alone GUI.

Random Forest

QIIME

QC

Search for Similar Signals by Correlation

Analyze Signals

Compute Similarity Matrix (heatmap)

Genomic Data

Genes in the Context of Epigenome Atlas

H1:H3K4me1 29

H1:H3K4me3 38

H1:H3K36me3 60

H1:H3K27me3 23

H1:H3K9me3 18

Output Targets

GenboreeUser_database

Step 31. Cluster by Spark
Click on "Epigenome" > "Analyze Signals" > "Cluster by Spark"

Tool Settings

Cluster by Spark (Analyze Signals)

Tool Overview

Inputs:

Data Tracks/Files:	Group:
H1:H3K4me1 29	Epigenomics Roadmap Repository, Database: Release 8 Repository
H1:H3K4me3 38	Epigenomics Roadmap Repository, Database: Release 8 Repository
H1:H3K36me3 60	Epigenomics Roadmap Repository, Database: Release 8 Repository
H1:H3K27me3 23	Epigenomics Roadmap Repository, Database: Release 8 Repository
ESCs:Rad21_Nanog	GenboreeUser_group, Database: GenboreeUser_database
H1:H3K9me3 18	Epigenomics Roadmap Repository, Database: Release 8 Repository

Output Database:

Database: GenboreeUser_database Group: GenboreeUser_group

Spark Analysis Settings

Analysis Name: Rad21_Nanog_H1_Epigenome

Select ROI Track: ESCs:Rad21_Nanog

Region Label: MyROIs

Statistics Type: global

of Clusters: 2

of Bins: 20

Data Track Colors:

ESCs:Rad21_Nanog	blue
H1:H3K27me3 23	blue
H1:H3K36me3 60	blue
H1:H3K4me1 29	blue
H1:H3K4me3 38	blue
H1:H3K9me3 18	blue

Submit **Cancel**

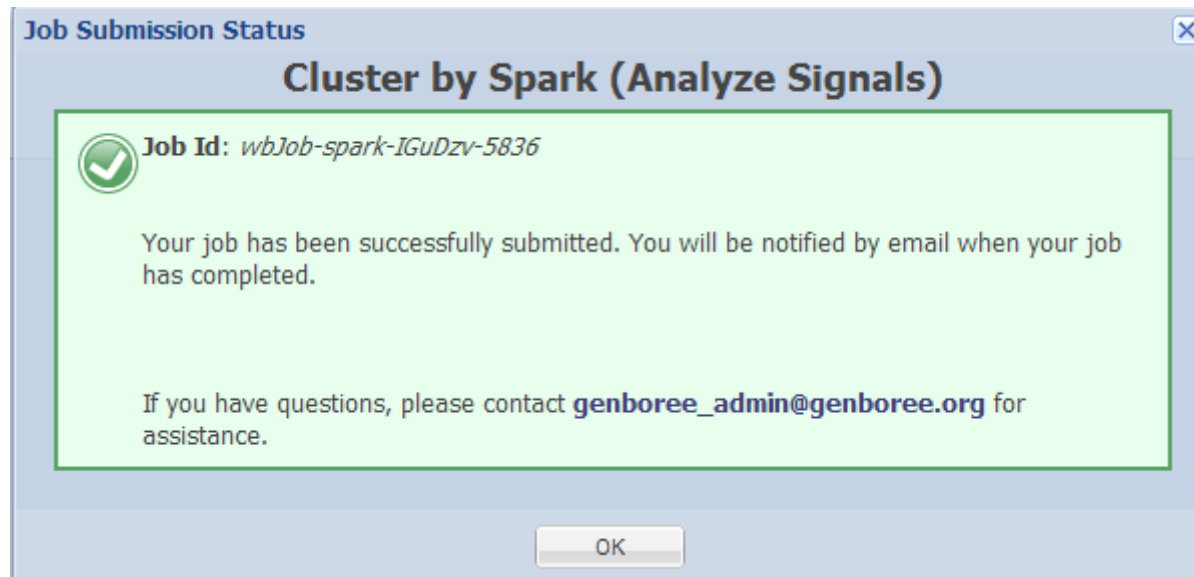
Step 32. Type in the Analysis Name
"Rad21_Nanog_H1_Epigenome"

Step 33. In Select ROI Track click on
"ESCs:Rad21_Nanog"

Step 34. In # of Clusters, type "2"

Step 35. Click on "Submit"

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



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

Your Spark job completed successfully.

Job Summary:

JobID - wbJob-spark-KMq1HG-5703

Analysis Name - Rad21_H1

Inputs:

of Data Tracks - 5

ROI Track - H1hesc:Rad21lggrab

Outputs:

Output DB - Dummy

Output Host - genboree.org

Settings:

k - 3

normType - exp

numBins - 20

regionLabel -

statsType - global

Additional Info:

To view your results in the Spark GUI:

(a) download and unzip the results archive and then

(b) launch Spark via Java Web Start and open the analysis folder.

Spark Java Web Start Link:

<http://www.bcgsc.ca/downloads/spark/current/start.jnlp>

- The Genboree Team

Result File Location in the Genboree Workbench:

(Direct links to files are at the end of this email)

Host: genboree.org

Grp: vamin_group

Db: Dummy

Files Area:

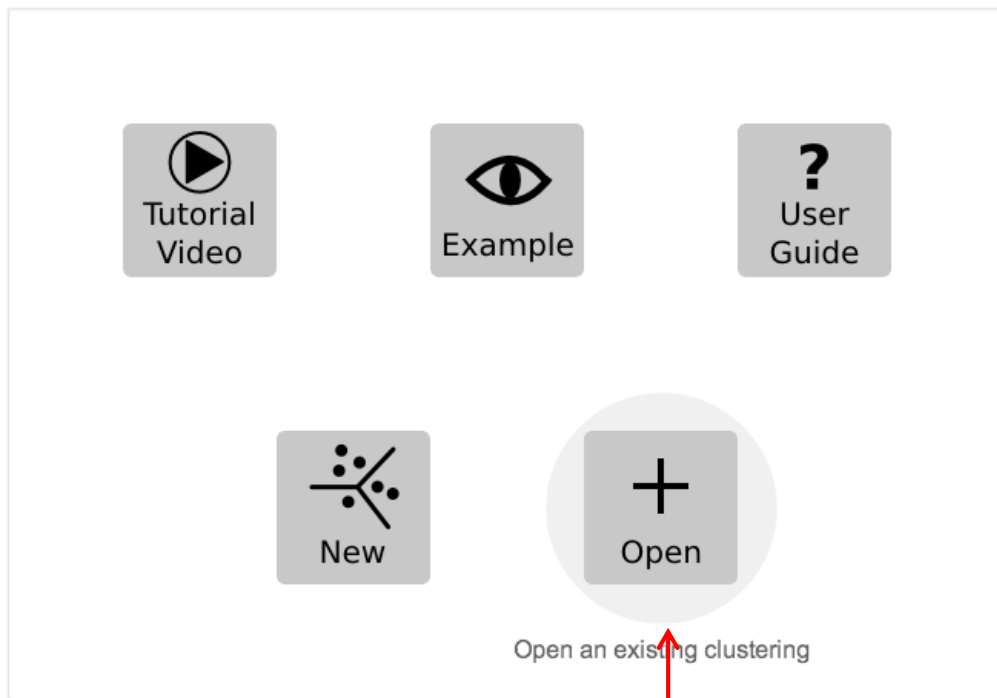
* Spark - Results/

* Rad21_H1/

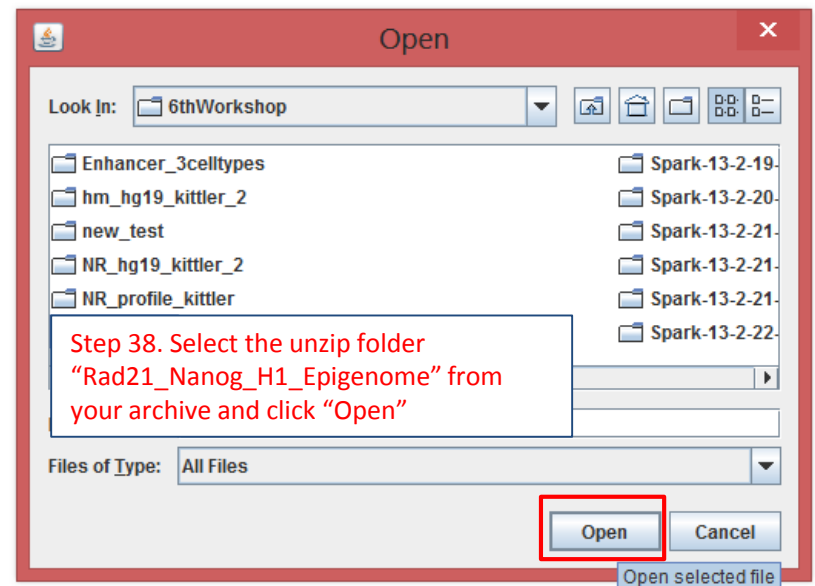
* ./Rad21_H1.zip

Step 36. Click on the link to Download "Rad21_Nanog_H1_Epigenome" folder and extract all to designated location in your computer

http://genboree.org/java-bin/apiCaller.jsp?rsrcPath=http%3A%2F%2Fgenboree.org%2FREST%2Fv1%2Fgrp%2Fvamin_group%2Fdb%2FDummy%2Ffile%2FSpark%2520-%2520Results%2FRad21_H1%2FRad21_H1.zip%2Fdata%3F&fileDownload=true&promptForLogin=true&errorFormal=html

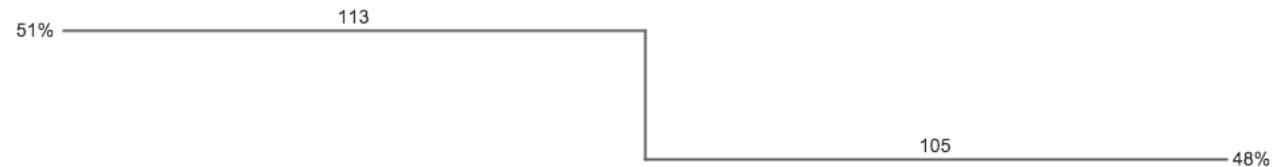


Step 37. Launch SPARK via Java Web Start

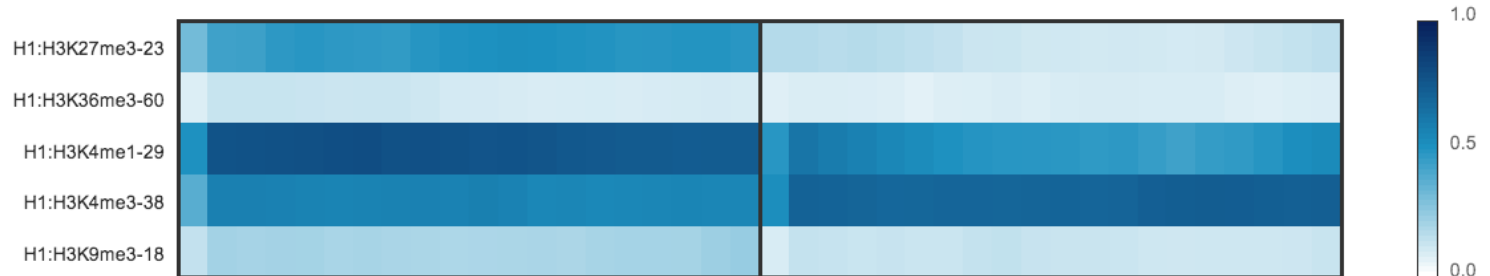


Rad21 and Nanog co-localized regions are composed of distal cis-regulatory elements and promoters based on enriched H3K4me1 and H3K4me3 signals

Cluster sizes (218 regions clustered)



Cluster profiles





GREAT predicts functions of cis-regulatory regions.

Many coding genes are well annotated with their biological functions. Non-coding regions typically lack such annotation. GREAT assigns biological meaning to a set of non-coding genomic regions by analyzing the annotations of the nearby genes. Thus, it is particularly useful in studying cis functions of sets of non-coding genomic regions. Cis-regulatory regions can be identified via both experimental methods (e.g. ChIP-seq) and by computational methods (e.g. [comparative genomics](#)). For more see our [Nature Biotech Paper](#).

News

- Apr 3, 2012: GREAT version 2.0 [adds new annotations to human and mouse ontologies](#) and visualization tools for data exploration.
- Feb 18, 2012: The [GREAT forums](#) are released, allowing increased user-to-user interaction

[More news items...](#)

Species Assembly

☒ Human: GRCh37 (UCSC hg19, Feb/2009)

☐ Human: NCBI build 36.1 (UCSC hg18, Mar/2006)

☐ Mouse: NCBI build 37 (UCSC mm9, Jul/2007)

☐ Zebrafish: Wellcome Trust Zv9 (danRer7, Jul/2010)



Zebrafish CNE set

[Can I use a different species or assembly?](#)

Select "Human: GRCh37 (UCSC hg19, Feb 2009)" as Species Assembly

Test regions

☒ BED file: Rad21_ES...dent.txt

☐ BED data:

[What should my test regions file contain?](#)

[How can I create a test set from a UCSC Genome Browser annotation track?](#)

Upload BED file that was generated

Background regions

☒ Whole genome

☐ BED file: No file chosen

☐ BED data:

[When should I use a background set?](#)
[What should my background regions file?](#)

Association rule settings

[Show settings »](#)

Expand "Show settings"
Select "Single nearest gene"
Enter within "500" kb

Associating genomic regions with genes

GREAT calculates statistics by associating genomic regions with nearby genes and applying the gene annotations to the regions. Association is a two step process. First, every gene is assigned a regulatory domain. Then, each genomic region is associated with all genes whose regulatory domain it overlaps.

☐ Basal plus extension

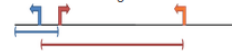
Proximal: kb upstream, kb downstream, plus Distal: up to kb



Gene regulatory domain definition: Each gene is assigned a basal regulatory domain of a minimum distance upstream and downstream of the TSS (regardless of other nearby genes). The gene regulatory domain is extended in both directions to the nearest gene's basal domain but no more than the maximum extension in one direction.

☐ Two nearest genes

within kb



Gene regulatory domain definition: Each gene is assigned a regulatory domain that extends in both directions to the nearest gene's TSS but no more than the maximum extension in one direction.

☒ Single nearest gene

within kb



Gene regulatory domain definition: Each gene is assigned a regulatory domain that extends in both directions to the midpoint between the gene's TSS and the nearest gene's TSS but no more than the maximum extension in one direction.

Gene Transcription Start Site (TSS)

☒ Include curated regulatory domains

[What are curated regulatory domains?](#)

Click "Submit"

GREAT provides region-gene association graphs and searches for Gene Ontology (GO) enrichment terms from various databases of the associated gene sets. This allows to make biologically meaningful predictions about the role of these cis-regulatory elements

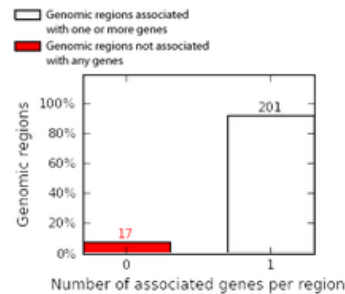
Job Description

Region-Gene Association Graphs

What do these graphs illustrate?

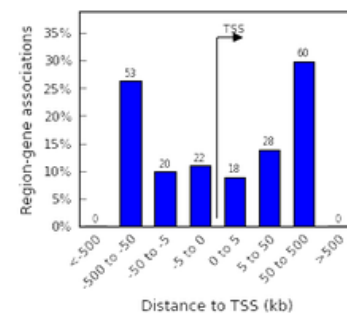
Number of associated genes per region

Download as PDF.



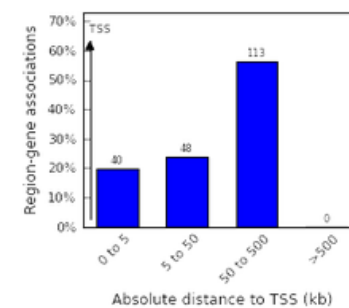
Binned by orientation and distance to TSS

Download as PDF.



Binned by absolute distance to TSS

Download as PDF.



Global Controls

Global Export



Which data is exported by each option?

GO Molecular Function (no terms)

Global controls

GO Biological Process (10+ terms)

Global controls

Table controls:

Export

Shown top rows in this table: 10

Set

Term annotation count: Min: 1

Max: Inf

Set

Visualize this table:



[select one]

Term Name	Binom Rank	Binom Raw P-Value	Binom FDR Q-Val	Binom Fold Enrichment	Binom Observed Region Hits	Binom Region Set Coverage	Hyper Rank	Hyper FDR Q-Val	Hyper Fold Enrichment	Hyper Observed Gene Hits	Hyper Total Genes	Hyper Gene Set Coverage
embryo development	1	3.3062e-8	2.8965e-4	2.8188	35	16.06%	1	1.1670e-6	3.7013	32	799	16.67%
regulation of cell death	3	1.6709e-7	4.8795e-4	2.5879	36	16.51%	24	7.3441e-5	2.6729	33	1,141	17.19%
regulation of programmed cell death	4	1.7500e-7	3.8330e-4	2.6296	35	16.06%	33	9.6038e-5	2.6571	32	1,113	16.67%
positive regulation of macromolecule metabolic process	6	3.1825e-7	4.6470e-4	2.2436	44	20.18%	7	2.0295e-5	2.6033	40	1,420	20.83%
regulation of apoptosis	7	4.0220e-7	5.0338e-4	2.5836	34	15.60%	45	1.8257e-4	2.5950	31	1,104	16.15%
developmental process involved in reproduction	10	8.0092e-7	7.0168e-4	3.5174	21	9.63%	26	7.1070e-5	4.3208	18	385	9.38%
positive regulation of metabolic process	13	2.1500e-6	1.4490e-3	2.0893	44	20.18%	17	6.9519e-5	2.4083	40	1,535	20.83%
negative regulation of macromolecule metabolic process	21	5.0220e-6	2.0951e-3	2.3415	33	15.14%	15	6.4829e-5	2.8005	32	1,056	16.67%
negative regulation of metabolic process	22	5.6157e-6	2.2363e-3	2.2914	34	15.60%	18	6.6468e-5	2.7181	33	1,122	17.19%
cell development	30	8.9351e-6	2.6094e-3	2.1181	38	17.43%	8	2.1497e-5	2.8880	34	1,088	17.71%

in ESCs maintain self-renewal by regulating genes related to programmed cell death/apoptosis

