SOMatic Tutorial

Camden Jansen

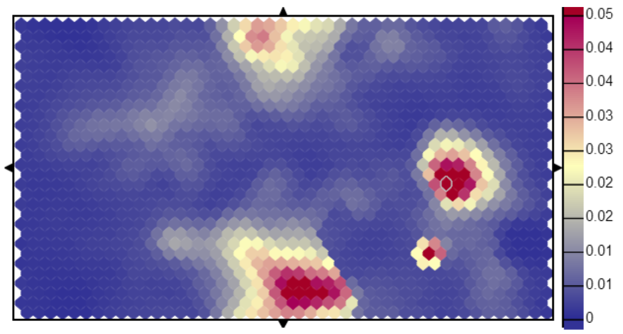
Mortazavi Lab

csjansen@uci.edu

University of California, Irvine

Presentation outline

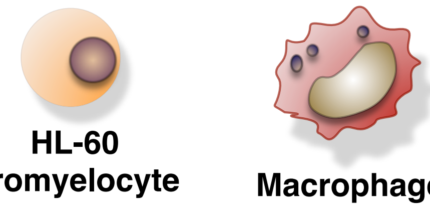
- Background on Self-Organizing Maps (SOMs)
- In-depth description of SOM training
- Using SOMatic to build your own SOM
- How to use the SOMatic viewer

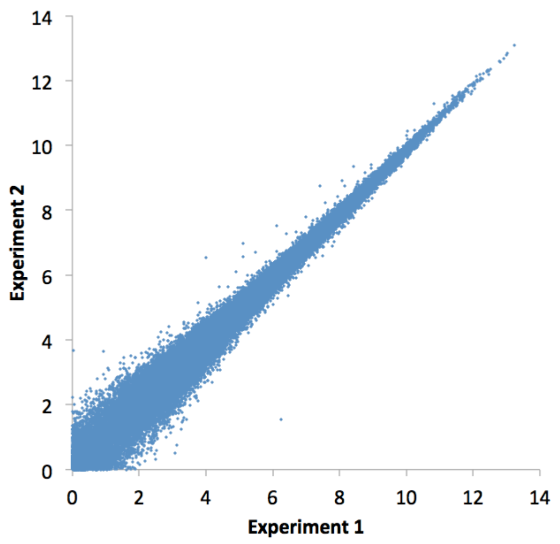


UCI University of California, Irvine

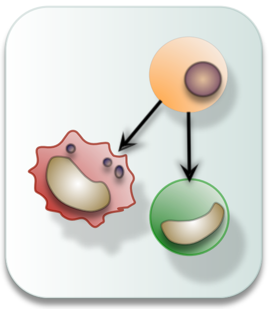
Data sets above 3 dimensions cannot be visualized easily

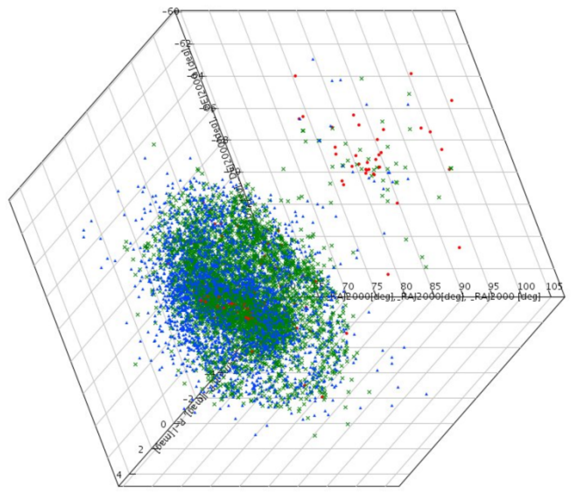
Example 1: 2 Experiments





Example 2: 3 Experiments

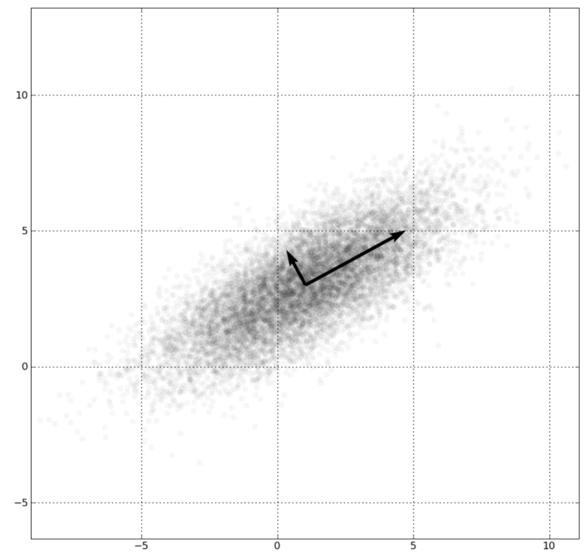




UCI University of California, Irvine

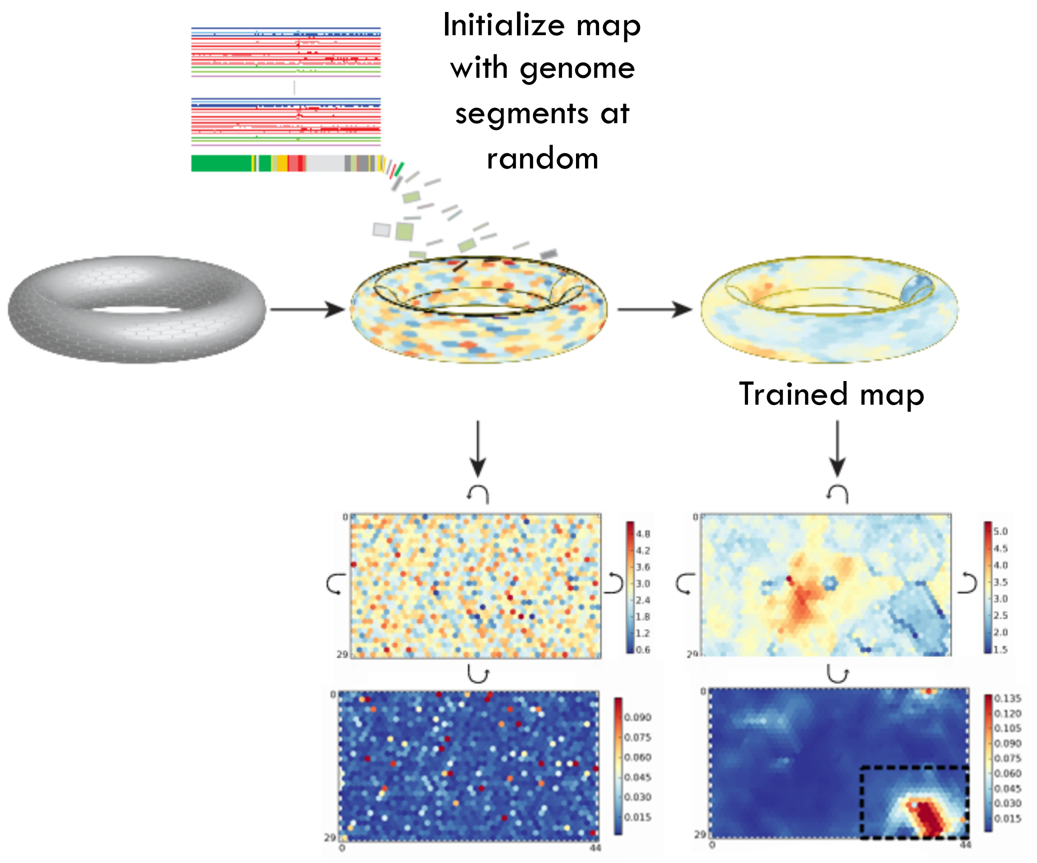
Principal component analysis (PCA) attempts to reduce the dimensions in a data set

- Principal Component Analysis
  - A linear transformation to a new coordinate system
  - Every dimension of this new system contains a decreasing amount of the variance.
- Pros
  - Can reduce a data set to fewer dimensions in a mathematically robust way (same result every time)
- Cons
  - Assumes a linear space
  - Loses spatial information with each dimension that you drop.



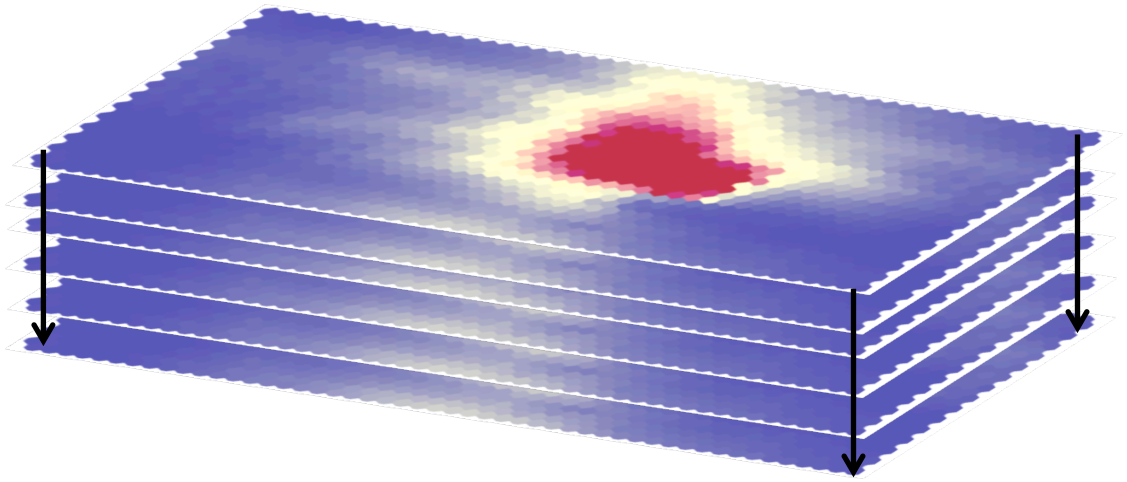
UCI University of California, Irvine

Self-organizing maps (SOMs) can reduce the dimensions in a data set in a non-linear way



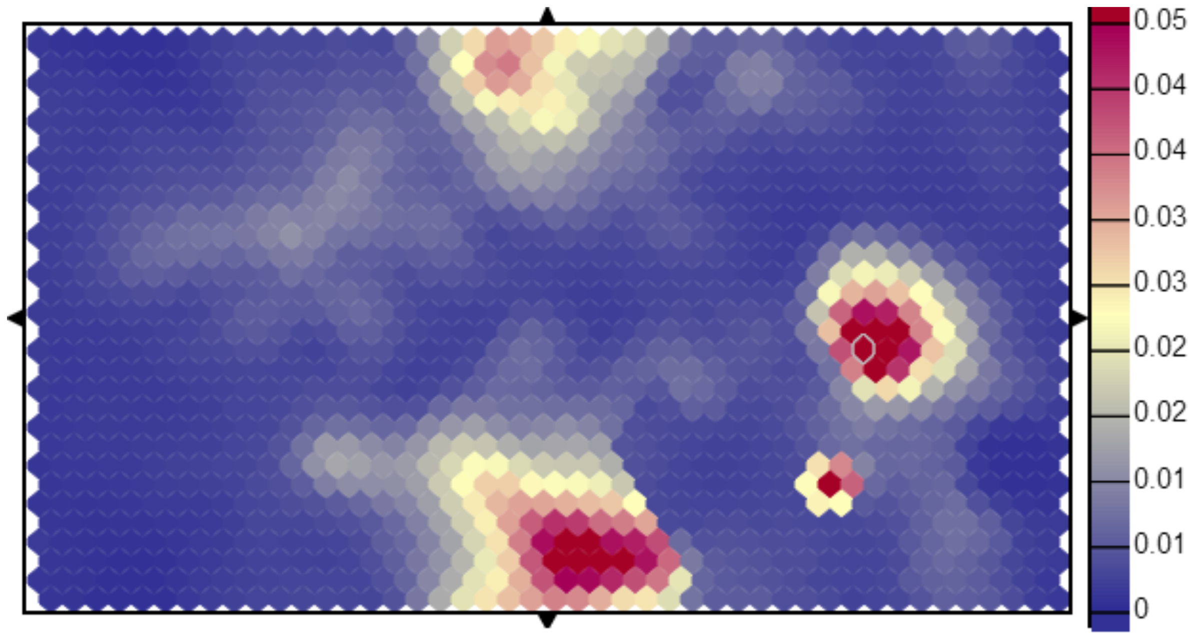
ENCODE Consortium, 2012

Each slice of a SOM represents a different experiment



UCI University of California, Irvine

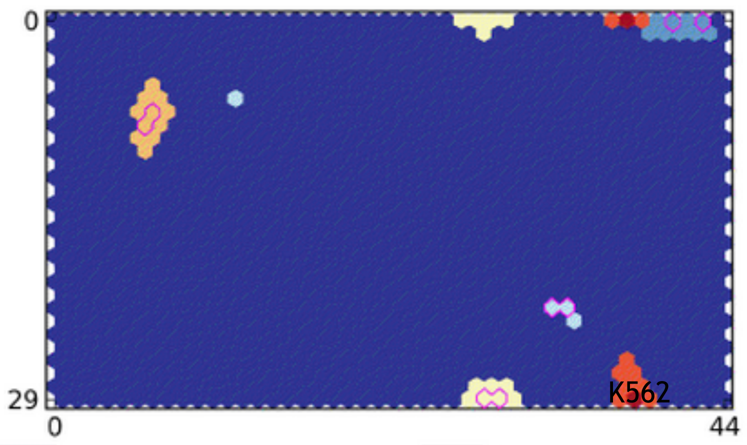
Each hexagon (unit) represents a cluster of genomic segments/genes/GO terms



University of California, Irvine

SOMs can be mined to find interesting regions

P300



	🟨 HepG2	
FOXA transcription factor networks		7.0e-46
FOXA2 & FOXA3 transcription factor networks		1.3e-45
Integrins in angiogenesis		6.4e-27

	🟦 GM12878	
Immune system		6.3e-140
Interferon Signaling		4.9e-91
Interferon alpha/beta Signaling		6.4e-27
Signaling		7.1e-80

	🟦 H1 hESC	
Wnt		1.3e-7

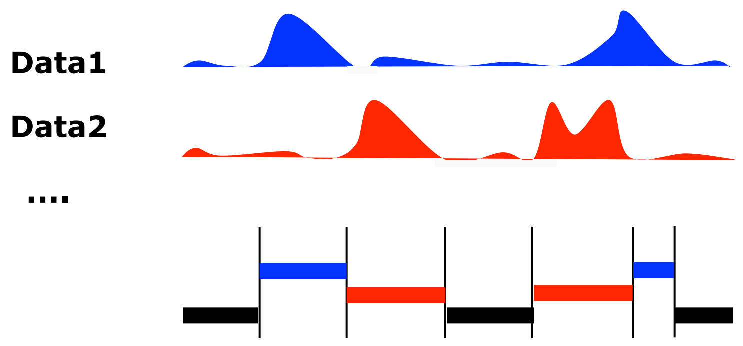
🟧	
IL2-mediated signaling events	2.0e-29
KitReceptor	1.5e-18
ATF-2 transcription factor network	6.45e-18

🟧🟥 Multiple / All	
Gene Expression	2.0e-29
Metabolism of RNA	1.1e-61
Metabolism of mRNA	6.45e-18

UCI University of California, Irvine

Self-organizing maps (SOMs) are unsupervised neural networks that must be trained on a data set

- Build your training matrix

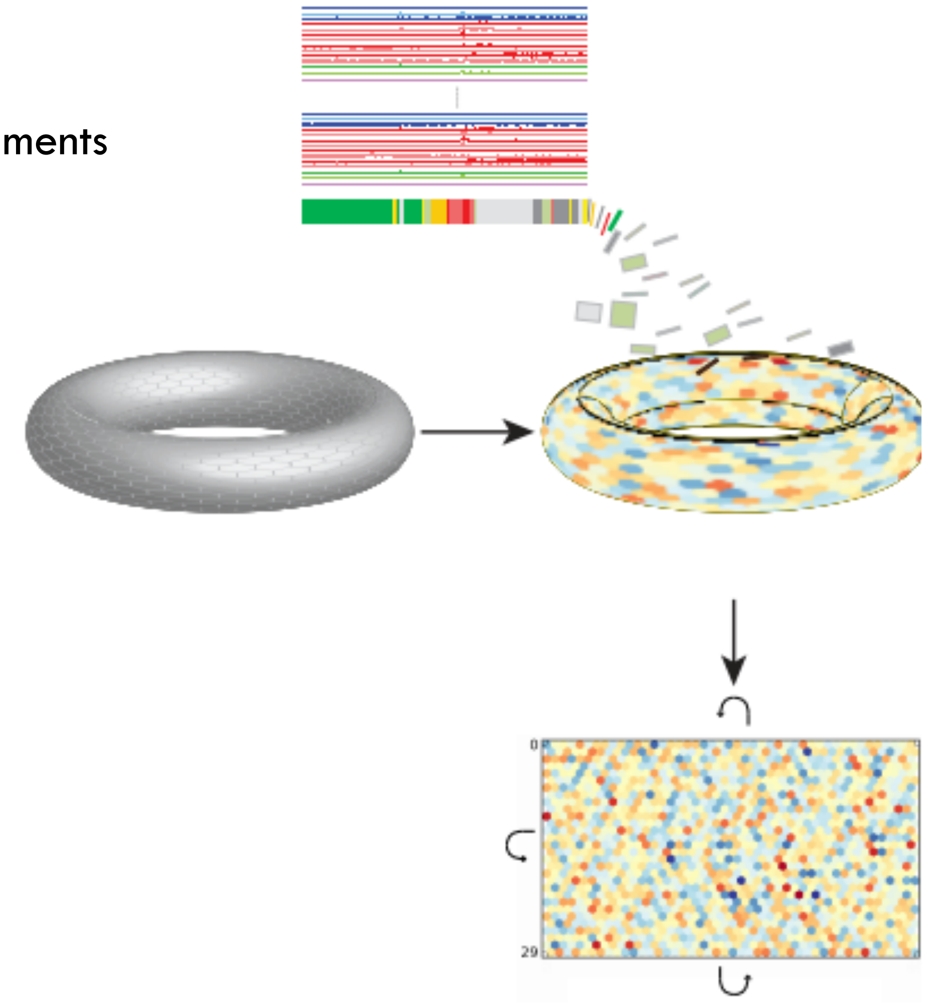


chromHMM-derived genome segmentation

segment	Data1	Data2	…	DataN
chrA:a-b	$rpkm_{11}$	$rpkm_{12}$	...	$rpkm_{1N}$
...	...	...	...	...
chrN:k-z	$rpkm_{i1}$	$rpkm_{i2}$	...	$rpkm_{iN}$

Self-organizing maps (SOMs) are unsupervised neural networks that must be trained on a data set

- Build your training matrix
- Initialize map with genome segments at random



ENCODE Consortium, 2012

Self-organizing maps (SOMs) are unsupervised neural networks that must be trained on a data set

- Build your training matrix
- Initialize map with genome segments at random
- Reorganize segments randomly
- Each time step:
  - Take a vector from the matrix

segment	Data1	Data2	…	DataN
chrA:a-b	$rpkm_{11}$	$rpkm_{12}$	…	$rpkm_{1N}$
…	…	…	…	…
chrN:k-z	$rpkm_{i1}$	$rpkm_{i2}$	…	$rpkm_{iN}$

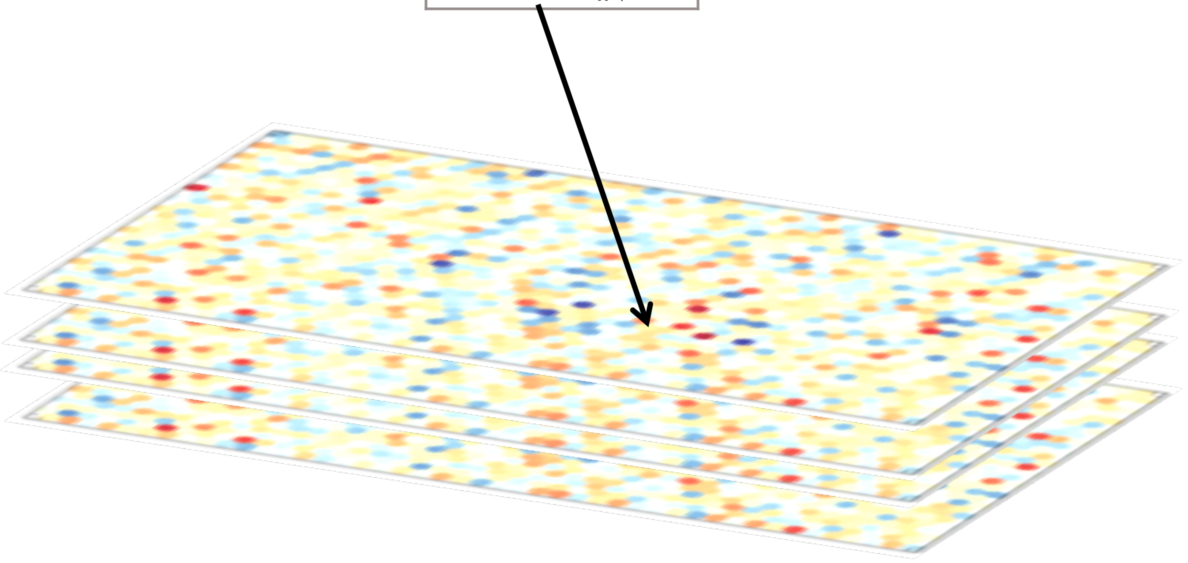
chrA:a-b
$rpkm_{i1}$
$rpkm_{i2}$
…
$rpkm_{iN}$

UCI University of California, Irvine

Self-organizing maps (SOMs) are unsupervised neural networks that must be trained on a data set

- Build your training matrix
- Initialize map with genome segments at random
- Reorganize segments randomly
- Each time step:
    - Take a vector from the matrix
    - Find the unit that's closest

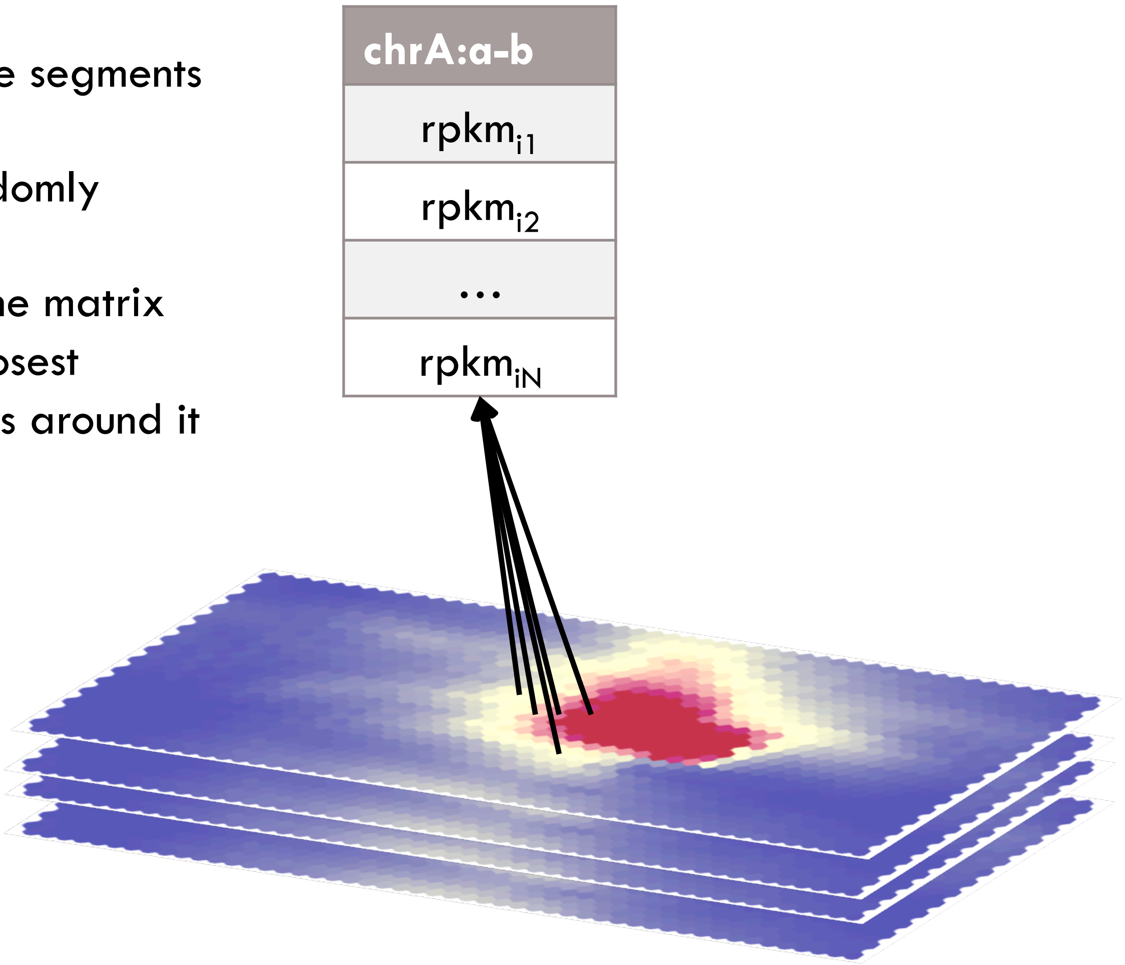
chrA:a-b
rpkm_{i1}
rpkm_{i2}
…
rpkm_{iN}



UCI University of California, Irvine

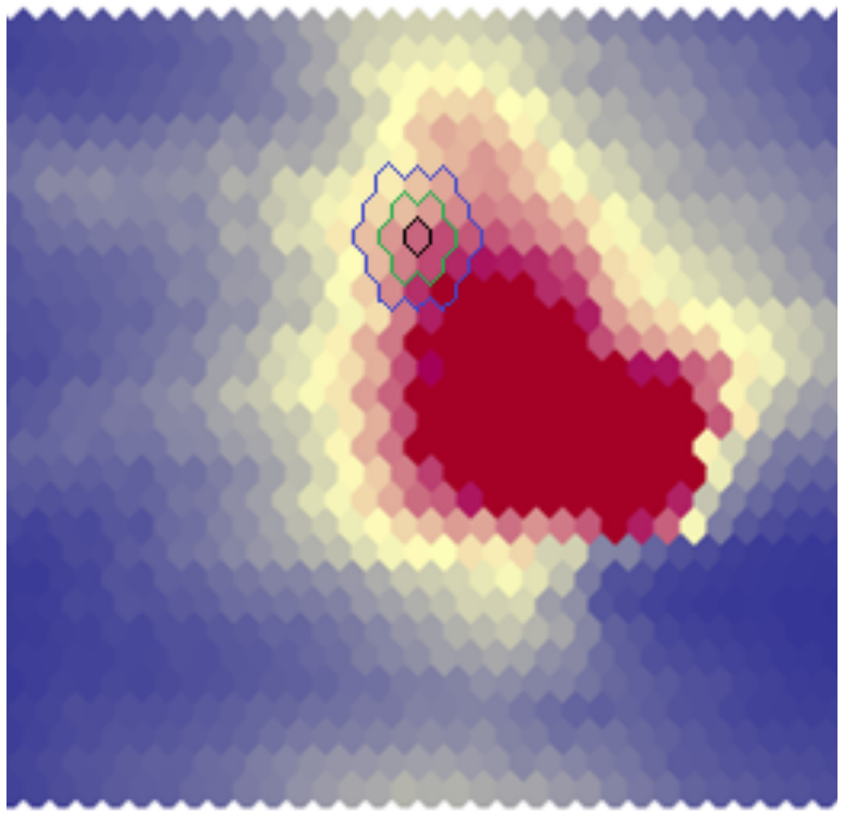
Self-organizing maps (SOMs) are unsupervised neural networks that must be trained on a data set

- Build your training matrix
- Initialize map with genome segments at random
- Reorganize segments randomly
- Each time step:
  - Take a vector from the matrix
  - Find the unit that's closest
  - Pull that unit and units around it closer to the vector



Self-organizing maps (SOMs) are unsupervised neural networks that must be trained on a data set

- Build your training matrix
- Initialize map with genome segments at random
- Reorganize segments randomly
- Each time step:
  - Take a vector from the matrix
  - Find the unit that's closest
  - Pull that unit and units around it closer to the vector
  - Reduce radius/learning rate

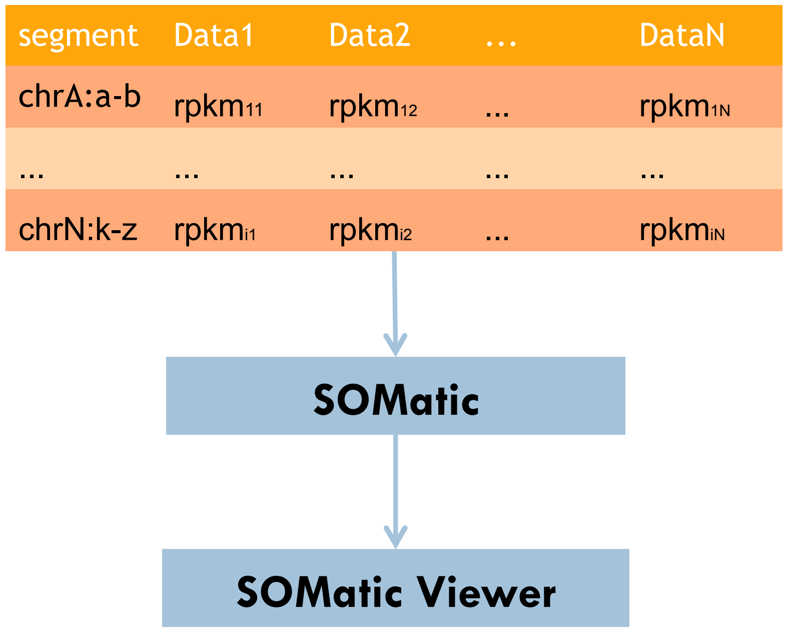


UCI University of California, Irvine

SOMatic: a tool for generating SOMs

- Build to be very general
    - Works for any coordinate system
        - Genome Coordinates (ChIP-seq, DNase-seq, ATAC-seq)
        - Genes (RNA-seq)

- Automatically, builds a website to explore your data space



segment	Data1	Data2	…	DataN
chrA:a-b	$rpkm_{11}$	$rpkm_{12}$	…	$rpkm_{1N}$
…	…	…	…	…
chrN:k-z	$rpkm_{i1}$	$rpkm_{i2}$	…	$rpkm_{iN}$

Requirements

- SOMatic has only been built/tested in a Linux environment
  - g++ version>2.8.2
    - Can be checked by running: `g++ --version`

- The SOMatic viewer needs to be placed on a web server
  - Has only been tested on an Apache server
    - Version > 2.4.10
    - Directory listings must be turned on.
      - This is done by adding "Indexes" to the options for the Directory in your httpd.conf file.

      - Example:

```
<Directory "/var/www/html">
   Options Indexes
   AllowOverride None
   Require all granted
</Directory>
```

UCI University of California, Irvine

Downloading/Installing SOMatic

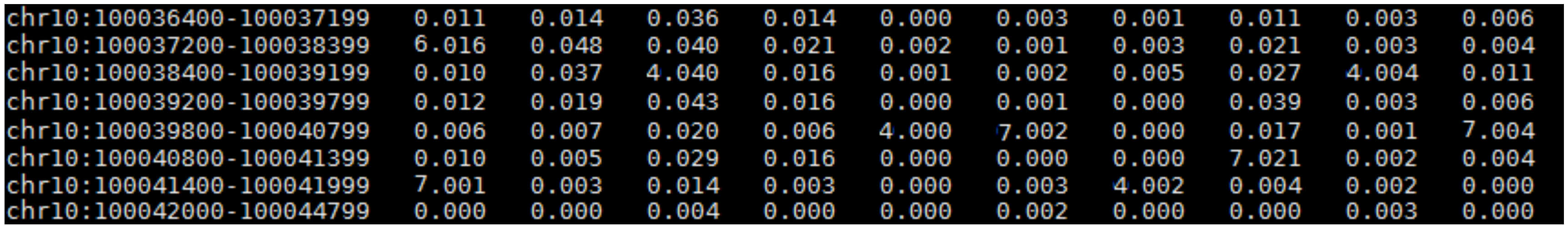
- Download the latest version:
```
$ wget http://crick.bio.uci.edu/SOMatic/SOMatic_Latest.tgz
```

- Installing:
  - Be sure that gcc version>2.8.2 is loaded by running:
```
$ g++ --version
```
  - Untar the SOMatic folder and go inside the bin directory:
```
$ tar –zxf SOMatic_Latest.tgz
$ cd SOMatic/bin
```
  - Run make:
```
$ make
```

Required files

- Training Matrix

  Segments RPKMs

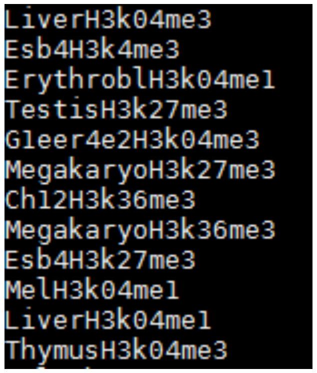


```
chr10:100036400-100037199  0.011  0.014  0.036  0.014  0.000  0.003  0.001  0.011  0.003  0.006
chr10:100037200-100038399  6.016  0.048  0.040  0.021  0.002  0.001  0.003  0.021  0.003  0.004
chr10:100038400-100039199  0.010  0.037  4.040  0.016  0.001  0.002  0.005  0.027  4.004  0.011
chr10:100039200-100039799  0.012  0.019  0.043  0.016  0.000  0.001  0.000  0.039  0.003  0.006
chr10:100039800-100040799  0.006  0.007  0.020  0.006  4.000  7.002  0.000  0.017  0.001  7.004
chr10:100040800-100041399  0.010  0.005  0.029  0.016  0.000  0.000  0.000  7.021  0.002  0.004
chr10:100041400-100041999  7.001  0.003  0.014  0.003  0.000  0.003  4.002  0.004  0.002  0.000
chr10:100042000-100044799  0.000  0.000  0.004  0.000  0.000  0.002  0.000  0.000  0.003  0.000
```

- There is an example training matrix at SOMatic/examples/example.matrix

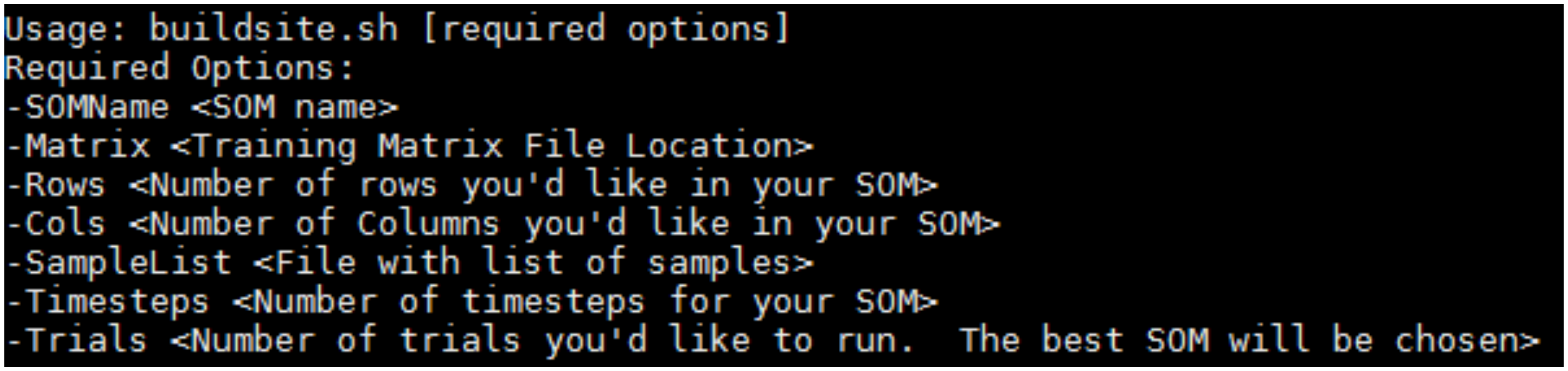
- Sample List
  - Rows in this file correspond to the RPKMs in the columns of the Training Matrix

  - There is an example sample list at SOMatic/examples/sample.list



```
LiverH3k04me3
Esb4H3k4me3
ErythroblH3k04me1
TestisH3k27me3
Gleer4e2H3k04me3
MegakaryoH3k27me3
Ch12H3k36me3
MegakaryoH3k36me3
Esb4H3k27me3
MelH3k04me1
LiverH3k04me1
ThymusH3k04me3
```

First step: buildsite.sh



```
Usage: buildsite.sh [required options]
Required Options:
-SOMName <SOM name>
-Matrix <Training Matrix File Location>
-Rows <Number of rows you'd like in your SOM>
-Cols <Number of Columns you'd like in your SOM>
-SampleList <File with list of samples>
-Timesteps <Number of timesteps for your SOM>
-Trials <Number of trials you'd like to run.  The best SOM will be chosen>
```

- To test the program, go to SOMatic/scripts and run the following:
  $ ./buildsite.sh -SOMName ExampleWebsite -Matrix ../examples/example.matrix -Rows 30 -Cols 50 -SampleList ../examples/sample.list -Timesteps 4000000 -Trials 3

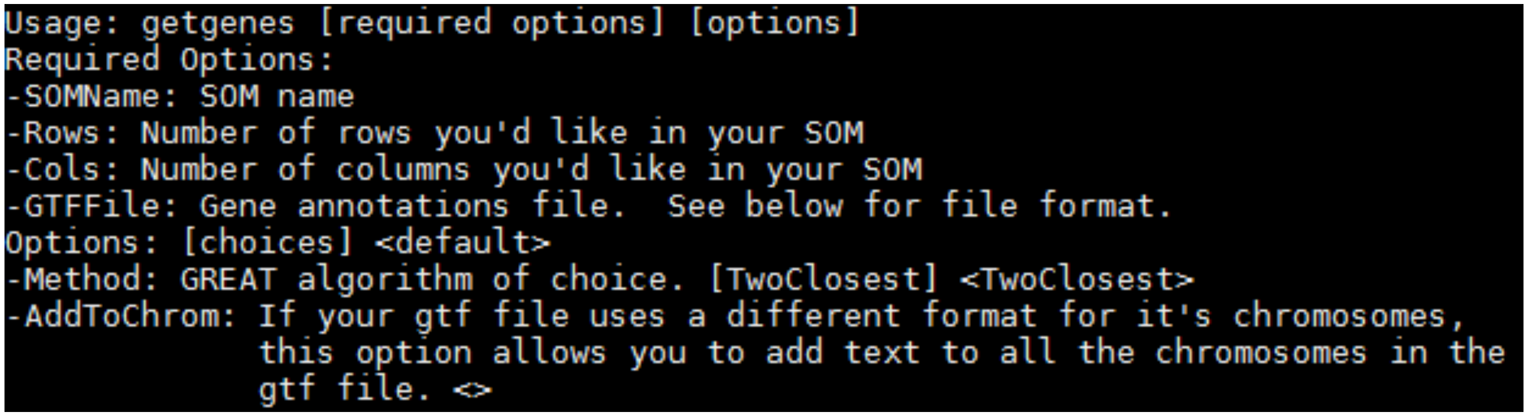
- This program runs the following steps of building your SOM automatically on the order of hours:
  - Training/Scoring SOM
  - Generating maps/summary
  - Building website

University of California, Irvine

(Optional) Add gene overlay

- If your training matrix uses genome segments (i.e. from ATAC-seq or DNase-seq), you can add a gene overlay in order to see which genes are in your unit of interest. This also allows you to add a GO term overlay and GO maps in the next step.

- We use the same algorithm for gene association as GREAT.



```
Usage: getgenes [required options] [options]
Required Options:
-SOMName: SOM name
-Rows: Number of rows you'd like in your SOM
-Cols: Number of columns you'd like in your SOM
-GTFFile: Gene annotations file.  See below for file format.
Options: [choices] <default>
-Method: GREAT algorithm of choice. [TwoClosest] <TwoClosest>
-AddToChrom: If your gtf file uses a different format for it's chromosomes,
             this option allows you to add text to all the chromosomes in the
             gtf file. <>
```

UCI University of California, Irvine

(Optional) Add gene overlay

- For this tutorial, start in the SOMatic directory, download the gtf file from Ensembl, and unzip it:

```
$ wget ftp://ftp.ensembl.org/pub/release-80/gtf/mus_musculus/
Mus_musculus.GRCm38.80.gtf.gz

$ gzip –d Mus_musculus.GRCm38.80.gtf.gz
```

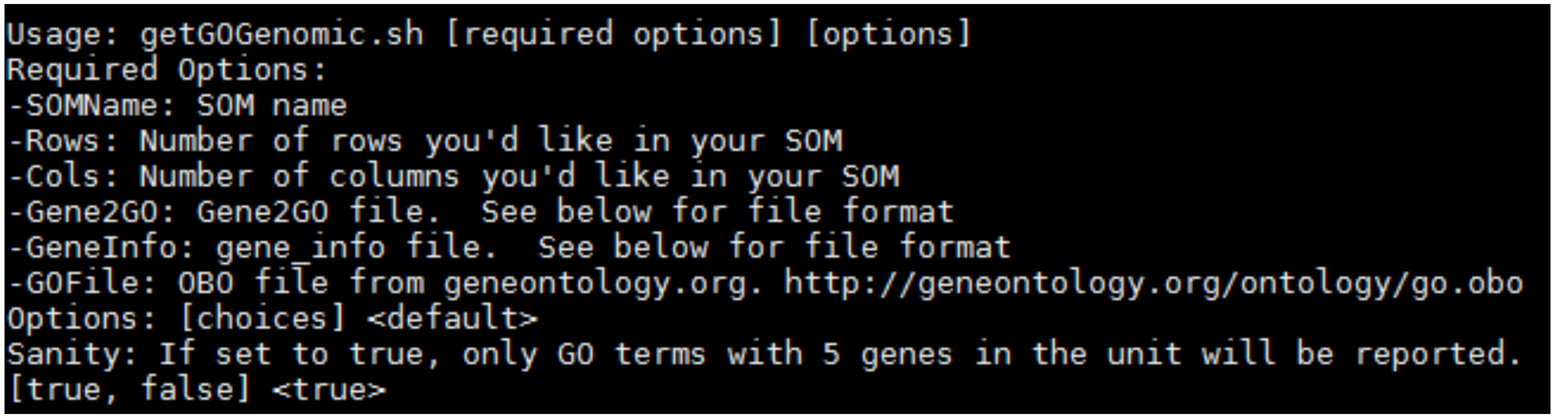
- This will allow us to run the following in the SOMatic/scripts directory:

```
$ ./getgenes.sh -SOMName ExampleWebsite -Rows 30        -Cols 50 -
GTFFile ../Mus_musculus.GRCm38.80.gtf        -AddToChrom chr
```

UCI University of California, Irvine

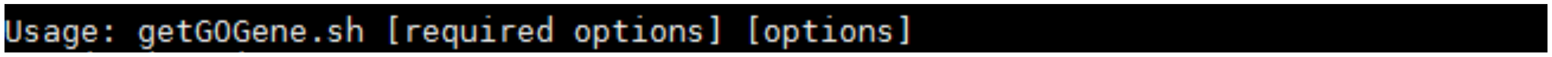
(Optional) Add GO overlay

- To see GO enrichments, run one of the two following scripts
- If your training matrix uses genome segments (i.e. from ATAC-seq or DNase-seq), you should use:



```
Usage: getGOGenomic.sh [required options] [options]
Required Options:
-SOMName: SOM name
-Rows: Number of rows you'd like in your SOM
-Cols: Number of columns you'd like in your SOM
-Gene2GO: Gene2GO file.  See below for file format
-GeneInfo: gene_info file.  See below for file format
-GOFile: OBO file from geneontology.org. http://geneontology.org/ontology/go.obo
Options: [choices] <default>
Sanity: If set to true, only GO terms with 5 genes in the unit will be reported.
[true, false] <true>
```

- If your training matrix uses genes (i.e. from RNA-seq), you should use:



```
Usage: getGOGene.sh [required options] [options]
```

(Optional) Add GO overlay

- For this tutorial, start in the SOMatic directory, download the gene2go and gene_info files from ncbi, and unzip them:

```
$ wget ftp://ftp.ncbi.nih.gov/gene/DATA/gene2go.gz
$ gzip –d gene2go.gz
$ wget ftp://ftp.ncbi.nih.gov/gene/DATA/GENE_INFO/ Mammalia/
Mus_musculus.gene_info.gz
$ gzip –d Mus_musculus.gene_info.gz
$ wget http://geneontology.org/ontology/go.obo
```

- This will allow us to run the following in the SOMatic/scripts directory:

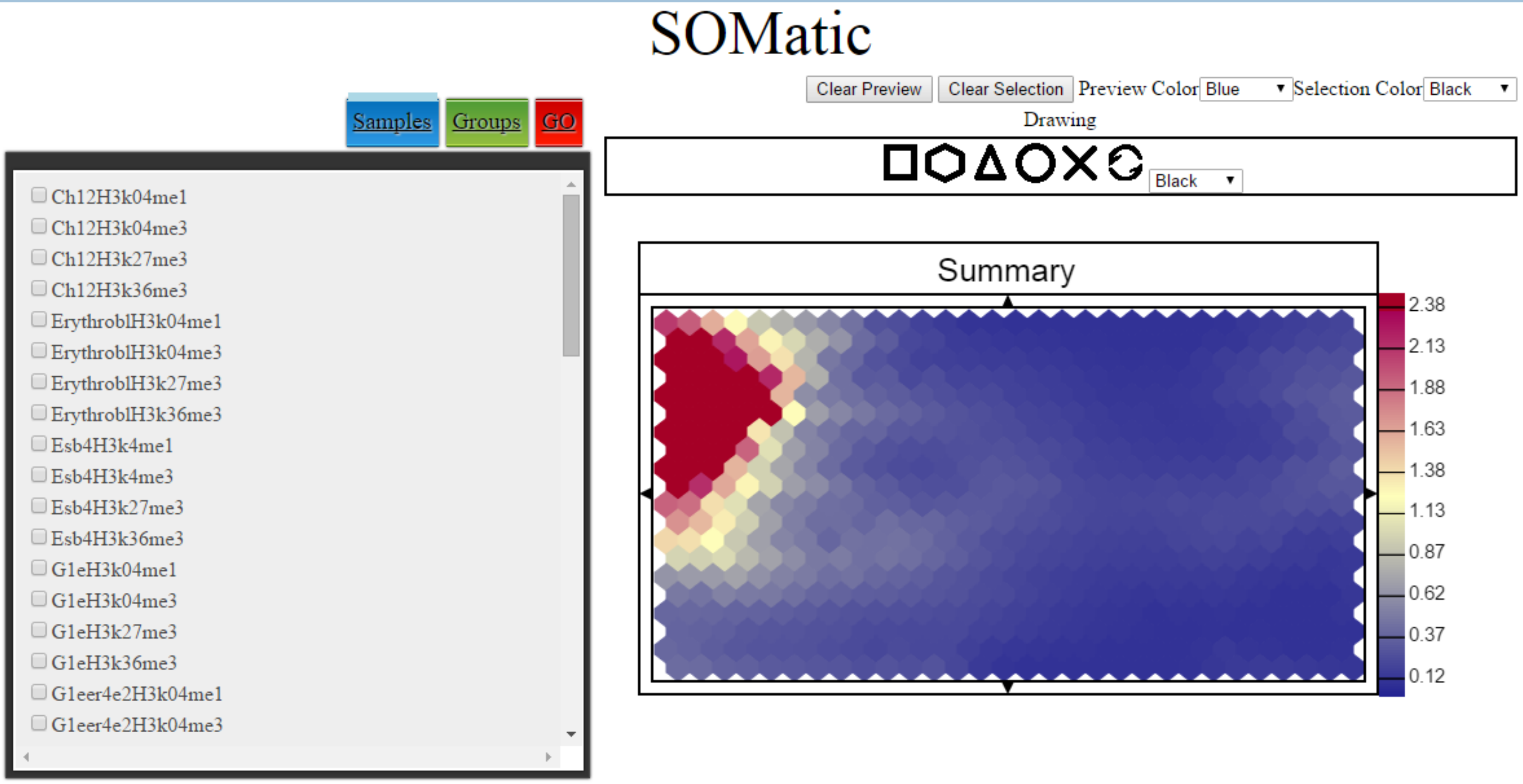
```
$ ./getGOGenomic.sh -SOMName ExampleWebsite -Rows 30    -Cols 50 -
Gene2GO ../gene2go -GeneInfo ../Mus_musculus.gene_info -GOFile ../
go.obo
```

SOMatic Viewer

Follow along at http://crick.bio.uci.edu/SOMatic/ExampleWebsite

UCI University of California, Irvine

Samples appear on the left under the samples tab



SOMatic

Clear Preview | Clear Selection | Preview Color: Blue | Selection Color: Black

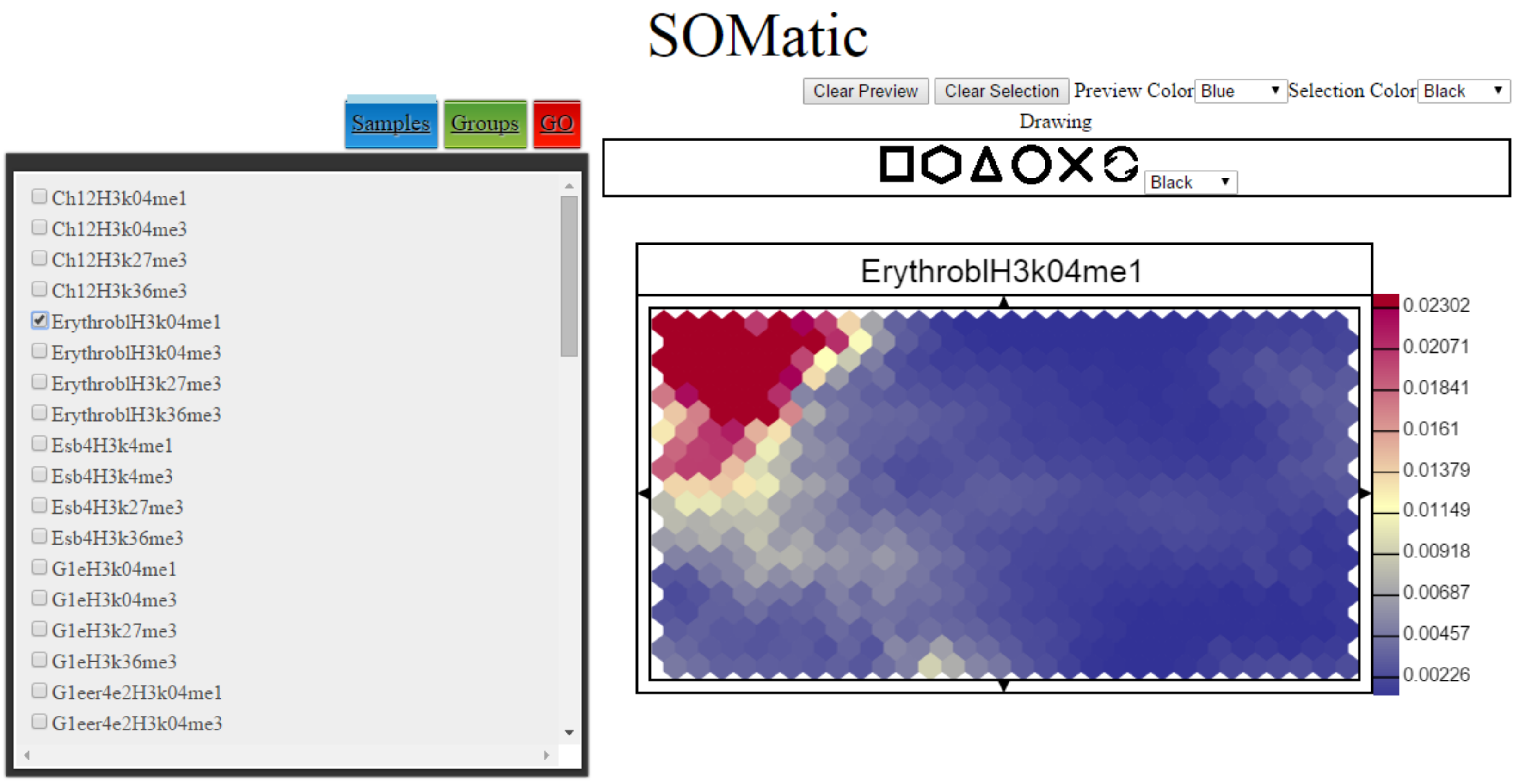
Drawing: Black

Samples | Groups | GO

- Ch12H3k04me1
- Ch12H3k04me3
- Ch12H3k27me3
- Ch12H3k36me3
- ErythroblH3k04me1
- ErythroblH3k04me3
- ErythroblH3k27me3
- ErythroblH3k36me3
- Esb4H3k4me1
- Esb4H3k4me3
- Esb4H3k27me3
- Esb4H3k36me3
- G1eH3k04me1
- G1eH3k04me3
- G1eH3k27me3
- G1eH3k36me3
- G1eer4e2H3k04me1
- G1eer4e2H3k04me3

UCI University of California, Irvine

Selected SOMs appear on the right



SOMatic

Clear Preview | Clear Selection | Preview Color: Blue | Selection Color: Black

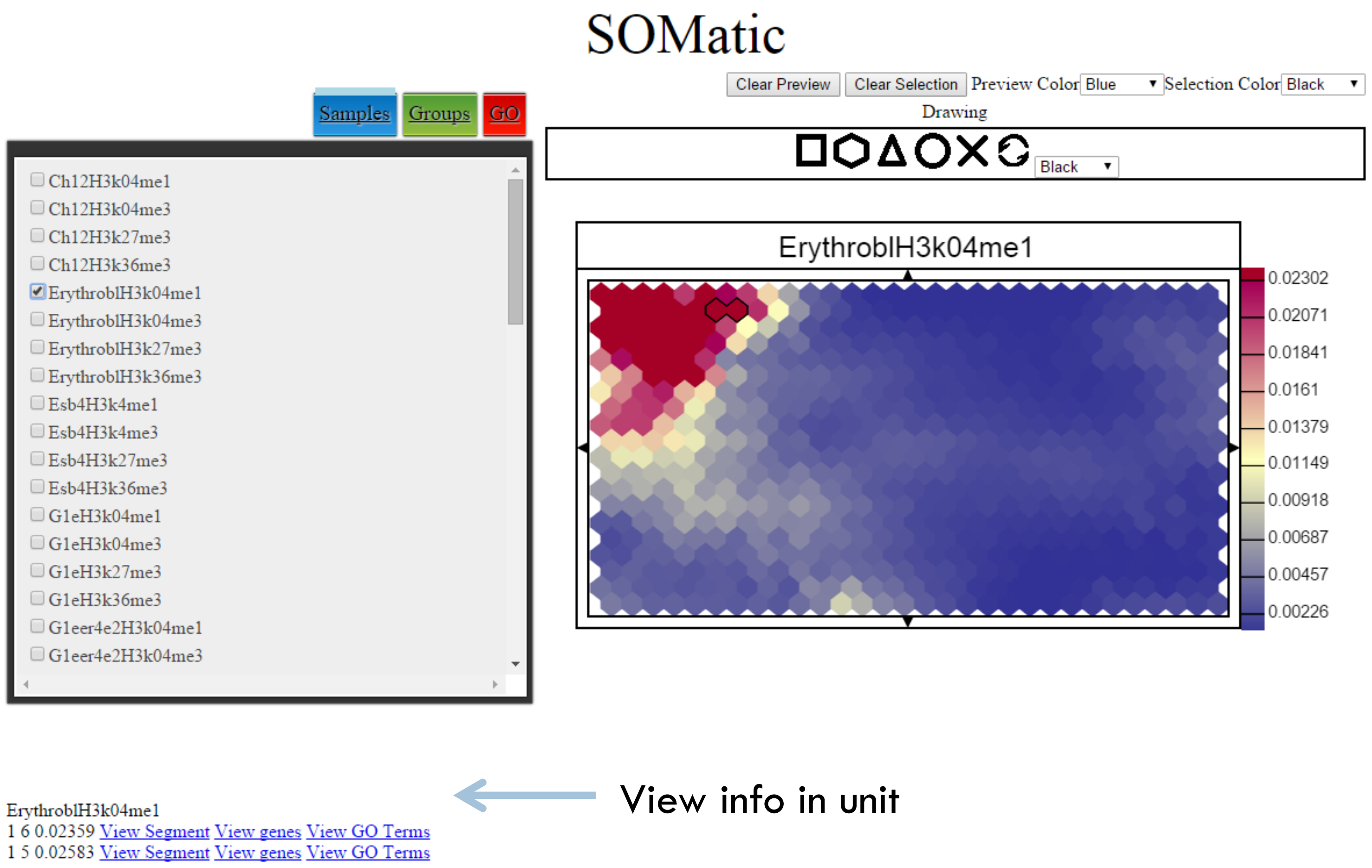
Drawing: Black

Samples | Groups | GO

- ☐ Ch12H3k04me1
- ☐ Ch12H3k04me3
- ☐ Ch12H3k27me3
- ☐ Ch12H3k36me3
- ☑ ErythroblH3k04me1
- ☐ ErythroblH3k04me3
- ☐ ErythroblH3k27me3
- ☐ ErythroblH3k36me3
- ☐ Esb4H3k4me1
- ☐ Esb4H3k4me3
- ☐ Esb4H3k27me3
- ☐ Esb4H3k36me3
- ☐ G1eH3k04me1
- ☐ G1eH3k04me3
- ☐ G1eH3k27me3
- ☐ G1eH3k36me3
- ☐ G1eer4e2H3k04me1
- ☐ G1eer4e2H3k04me3

UCI University of California, Irvine

Selected units appear below tabbed area



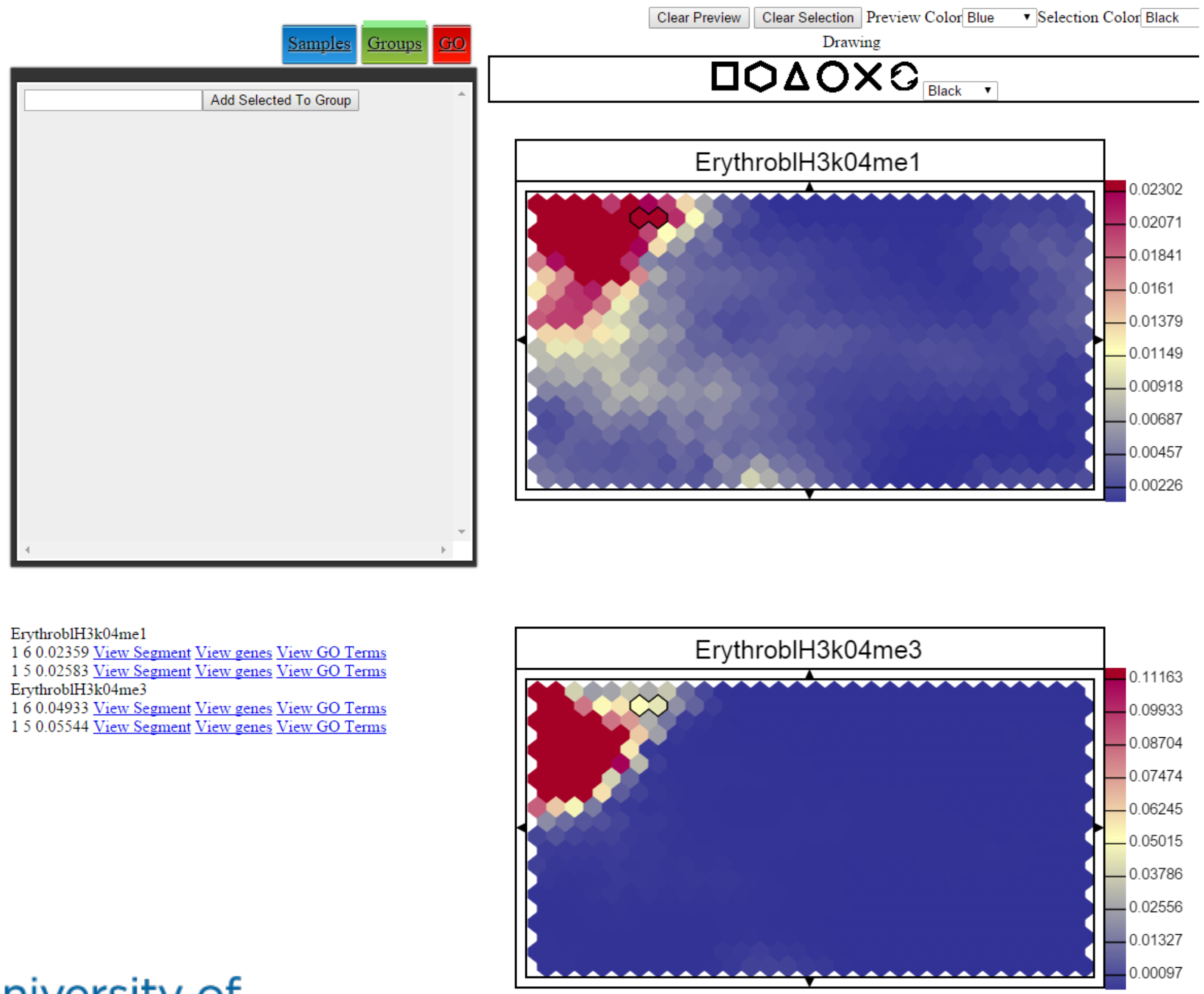
- ☐ Ch12H3k04me1
- ☐ Ch12H3k04me3
- ☐ Ch12H3k27me3
- ☐ Ch12H3k36me3
- ☑ ErythroblH3k04me1
- ☐ ErythroblH3k04me3
- ☐ ErythroblH3k27me3
- ☐ ErythroblH3k36me3
- ☐ Esb4H3k4me1
- ☐ Esb4H3k4me3
- ☐ Esb4H3k27me3
- ☐ Esb4H3k36me3
- ☐ G1eH3k04me1
- ☐ G1eH3k04me3
- ☐ G1eH3k27me3
- ☐ G1eH3k36me3
- ☐ G1eer4e2H3k04me1
- ☐ G1eer4e2H3k04me3

ErythroblH3k04me1
1 6 0.02359 View Segment View genes View GO Terms
1 5 0.02583 View Segment View genes View GO Terms

View info in unit

University of California, Irvine

Groups tab allows for grouping of maps



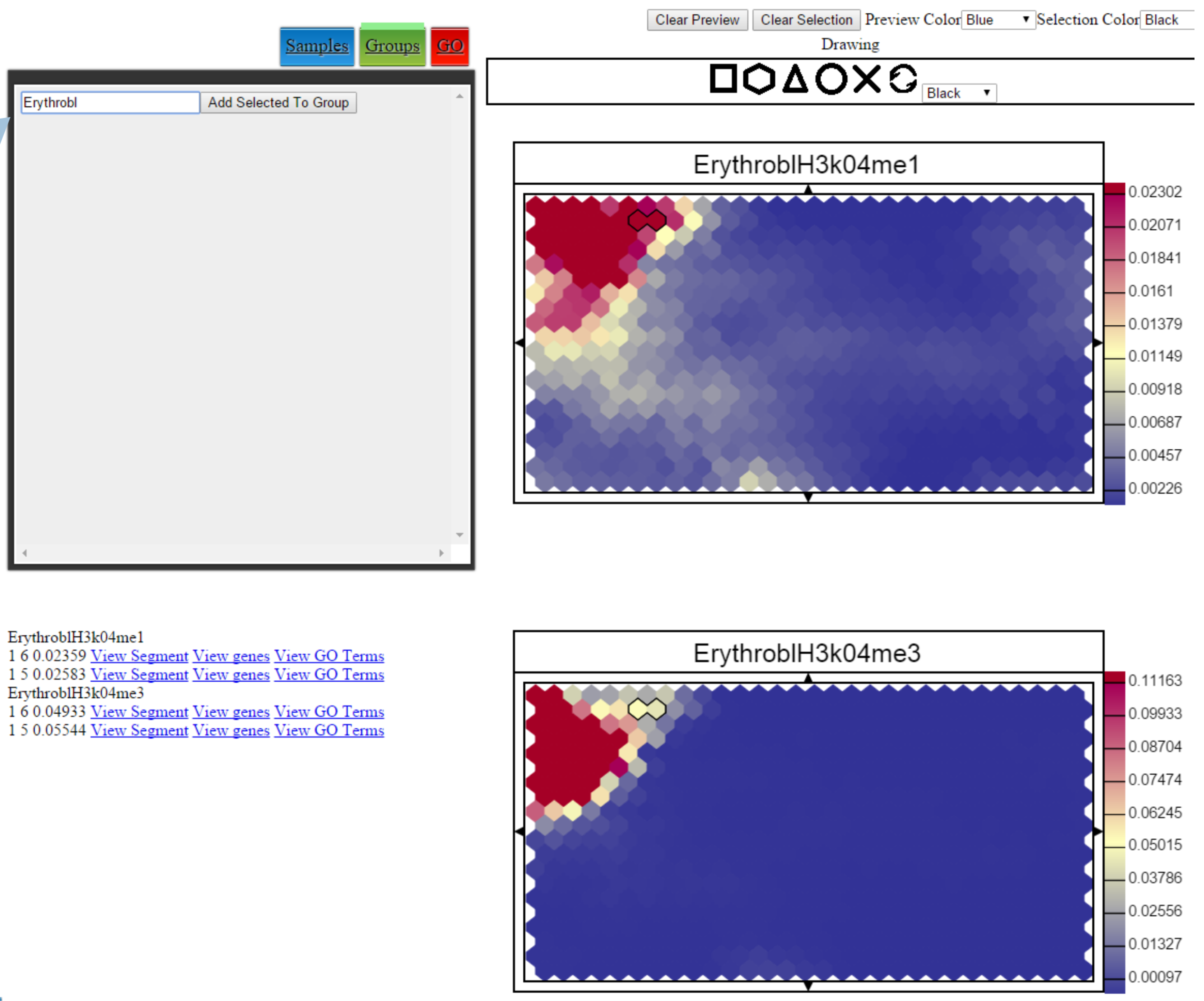
ErythroblH3k04me1
1 6 0.02359 View Segment View genes View GO Terms
1 5 0.02583 View Segment View genes View GO Terms
ErythroblH3k04me3
1 6 0.04933 View Segment View genes View GO Terms
1 5 0.05544 View Segment View genes View GO Terms

UCI University of California, Irvine

Groups tab allows for grouping of maps

Enter name for group

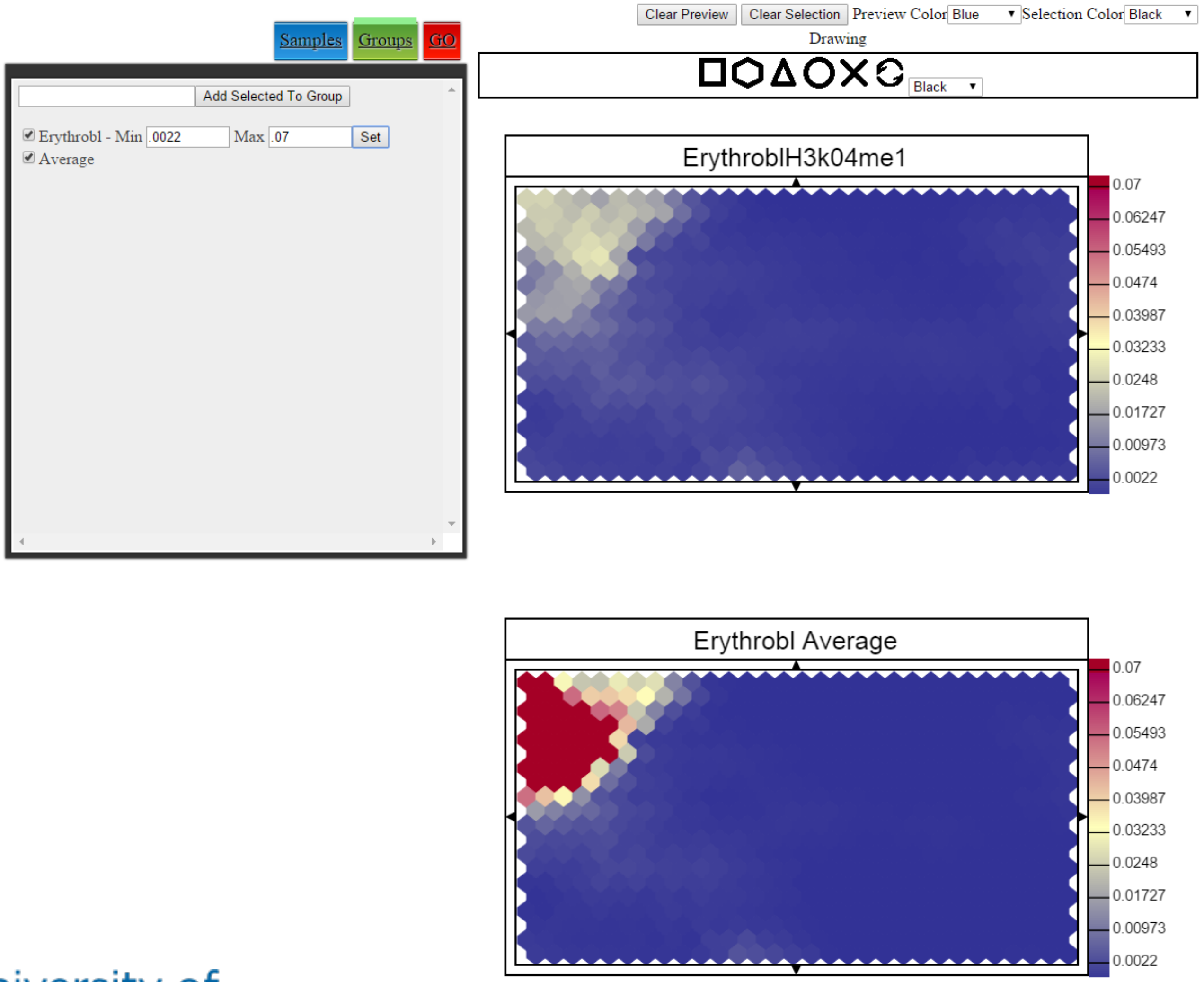
Press button to create a group



```
ErythroblH3k04me1
1 6 0.02359 View Segment View genes View GO Terms
1 5 0.02583 View Segment View genes View GO Terms
ErythroblH3k04me3
1 6 0.04933 View Segment View genes View GO Terms
1 5 0.05544 View Segment View genes View GO Terms
```

UCI University of California, Irvine

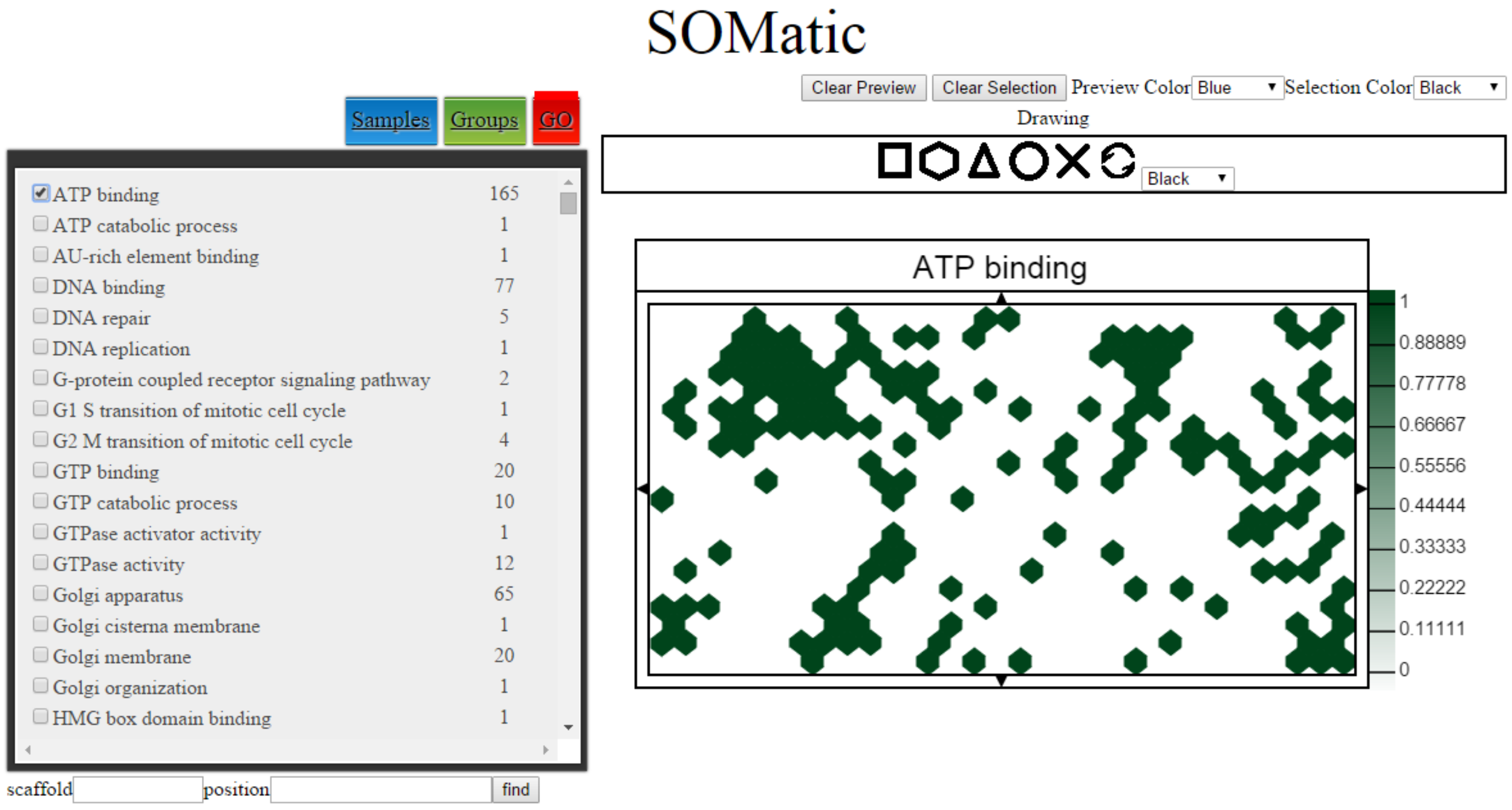
Groups allow setting maps to the same scale and viewing an average map



Erythrobl - Min .0022 Max .07 Set

Average

GO tab displays map GO enrichments calculated by binomial



SOMatic

Clear Preview | Clear Selection | Preview Color: Blue | Selection Color: Black

Drawing: Black

Samples | Groups | GO

✓ ATP binding	165
ATP catabolic process	1
AU-rich element binding	1
DNA binding	77
DNA repair	5
DNA replication	1
G-protein coupled receptor signaling pathway	2
G1 S transition of mitotic cell cycle	1
G2 M transition of mitotic cell cycle	4
GTP binding	20
GTP catabolic process	10
GTPase activator activity	1
GTPase activity	12
Golgi apparatus	65
Golgi cisterna membrane	1
Golgi membrane	20
Golgi organization	1
HMG box domain binding	1

scaffold | position | find

University of California, Irvine

Acknowledgments

- Mortazavi lab:
  - **Dr. Ali Mortazavi**
  - **Ricardo Ramirez**
  - **Dr. Weihua Zeng**
  - Rabi Murad
  - Dr. Eddie Park
  - Marissa Macchietto
  - Mandy Jiang
  - Nicole El-Ali

- HudsonAlpha-led ENCODE production group
- HPC

- SOMatic URL: http://crick.bio.uci.edu/SOMatic

UCI University of California, Irvine