RETHINKING GENOME BROWSING:

NAVIGATION BY FUNCTION NOT POSITION

CYDNEY NIELSEN
Postdoctoral Fellow
February 17, 2012

Mardis *Genome Medicine* 2010, **2**:84
http://genomemedicine.com/content/2/11/84

Genome **Medicine**

MUSINGS

The $1,000 genome, the $100,000 analysis?

Elaine R Mardis*

The challenges of having so much data

Most challenging of all, it will be of critical importance to develop meta-analyses and statistical analysis tools that integrate across disparate data types, … … thereby enable researchers to collectively interpret these data for all samples in a study and to form testable hypotheses from this discovery phase.

- Elaine Mardis
  Anticipating the $1,000 genome

The challenges of having so much data

Most challenging of all, it will be of critical importance to develop meta-analyses and statistical analysis tools that integrate across disparate data types, …
… thereby enable researchers to collectively interpret these data for all samples in a study and to form testable hypotheses from this discovery phase.

- Elaine Mardis
Anticipating the $1,000 genome

Data visualization will be a key player in this domain

Why visualization?

Data landscape is unknown

Why visualization?

Data landscape is unknown

- In the discovery phase, not yet clear where the interesting features lie or what they look like
- Features are not sufficiently well defined to be extracted in a purely automated fashion
- Visualization is a powerful approach in such cases:

  Exploit our visual system and knowledge to identify biologically interesting data patterns and subsequently generalize

Why visualization?

Improves data accessibility to biological community

Why visualization?

Improves data accessibility to biological community

- More and more data are being produced from large consortia (ENCODE, Epigenome Roadmap Project, etc.)

- Download portals are valuable, but of primary use to computational experts

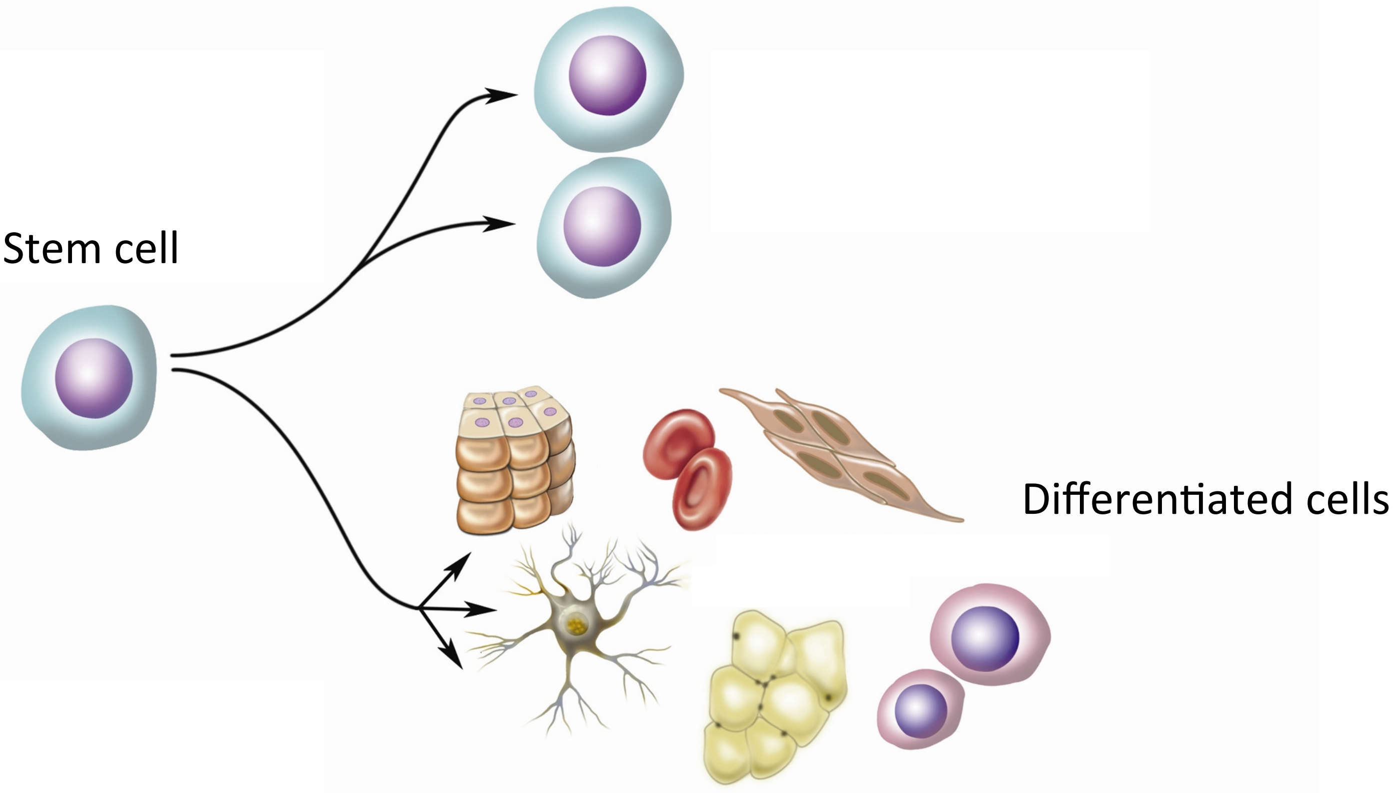
- If we are to make the most of these large datasets, need to bridge the computational gap between primary data and biological community – visualization can play a key role

Outline

- Exploring genome-wide datasets
  - Computational and visual methods to investigate ChIP-Seq data
- Spark
  - A navigational paradigm for interactive genome exploration
- Demo
- Future work

ROADMAP epigenomics PROJECT

Many cell types with same DNA sequence but different morphologies

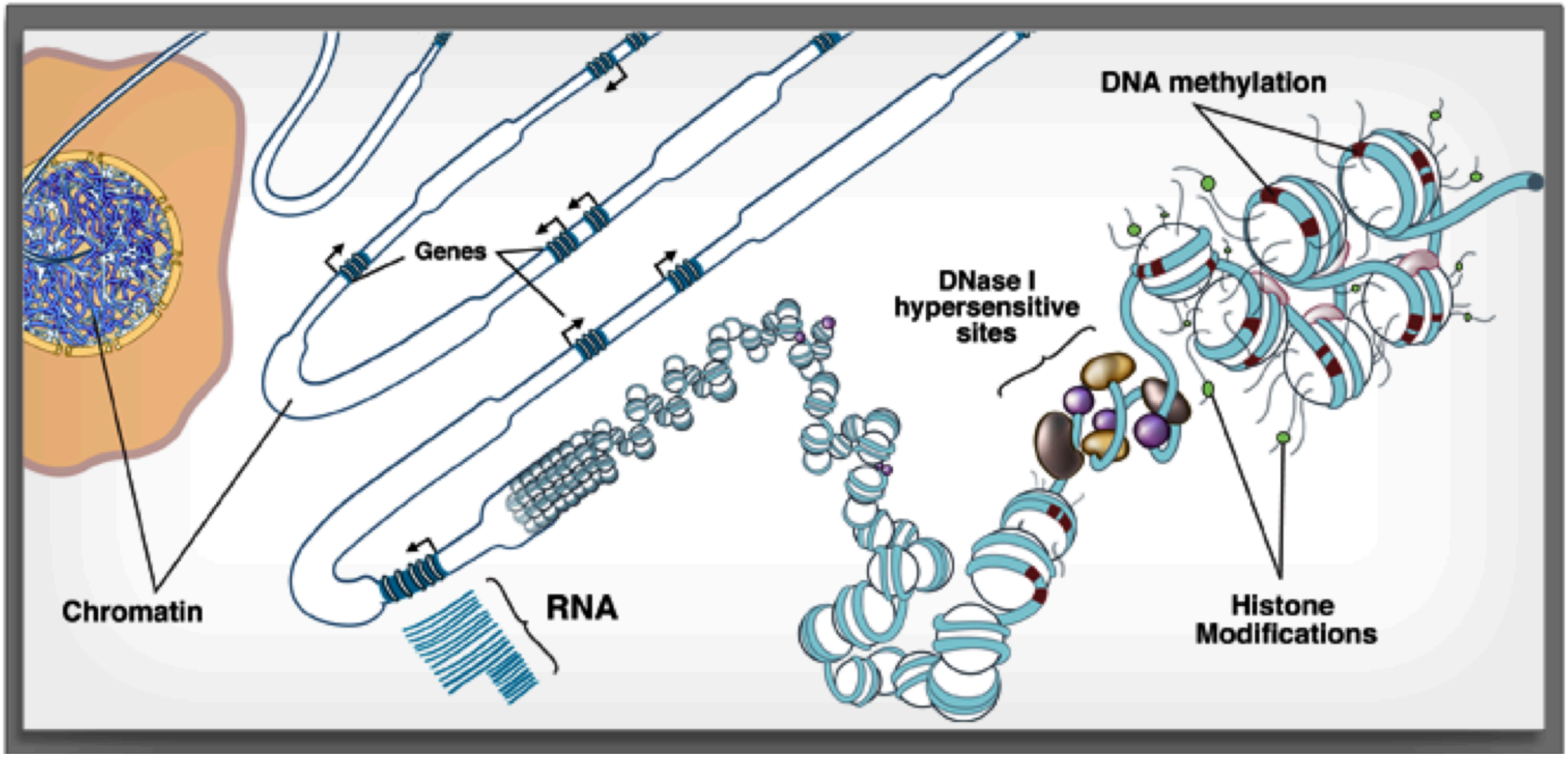


www.nationalacademies.org/stemcells

ROADMAP epigenomics PROJECT

Epigenetics - the study of changes in the regulation of gene activity that are not dependent on gene sequence

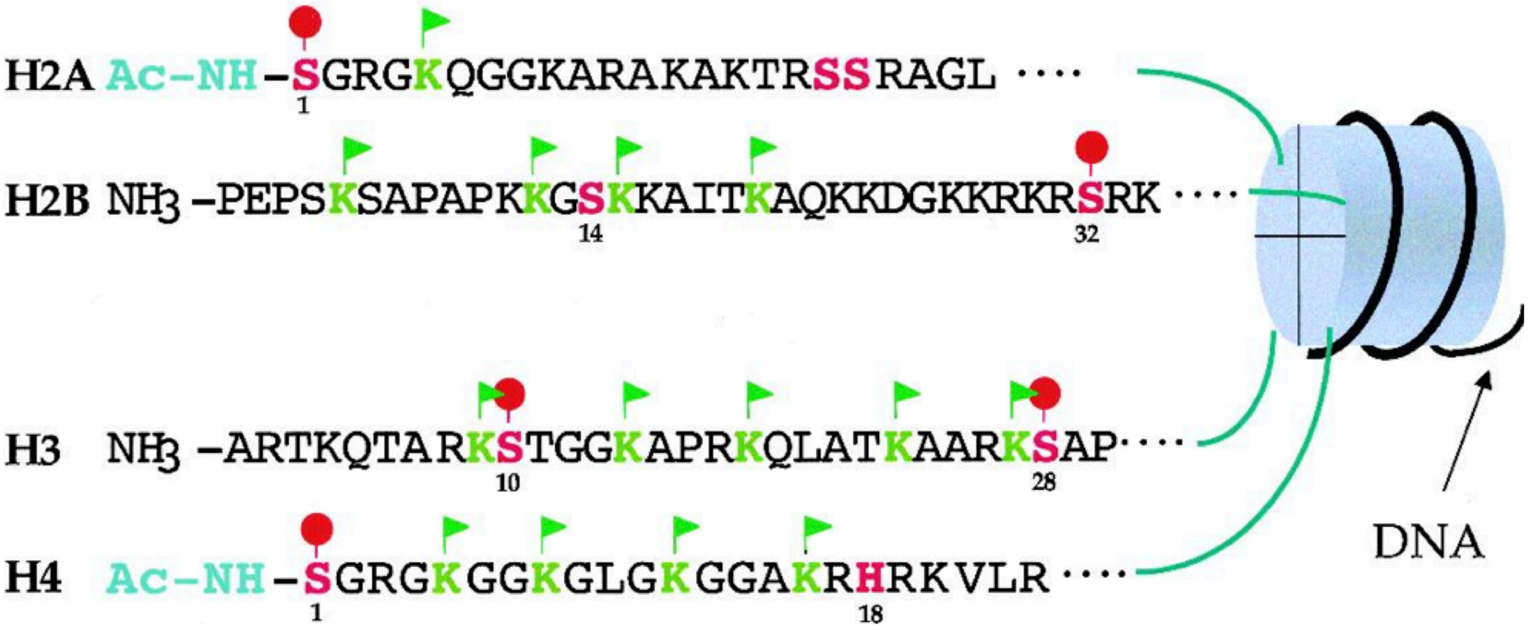
epi - (above) – genetics



ROADMAP epigenomics PROJECT

Epigenetics - the study of changes in the regulation of gene activity that are not dependent on gene sequence

epi - (above) – genetics

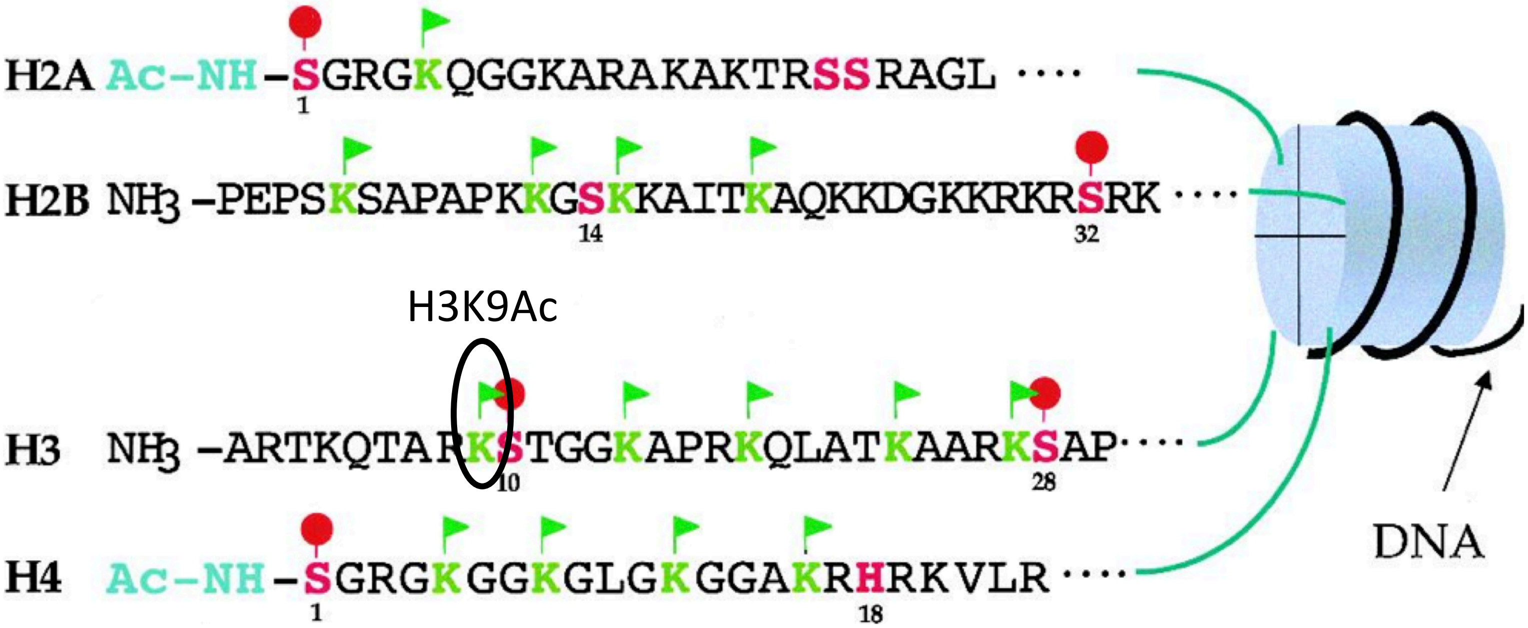

H2A Ac–NH –SGRGKQGGKARAKAKTRSSRAGL ····
1
H2B NH3 –PEPSKSAPAPKKGSKKAITKAQKKDGKKRKRSRK ····
14
32
H3 NH3 –ARTKQTARKSTGGKAPRKQLATKAARKSAP····
10
28
H4 Ac–NH –SGRGKGGKGLGKGGAKRHRKVLR ····
1
18
DNA

http://bricker.tcnj.edu/Amb/amble9.html

ROADMAP epigenomics PROJECT

Epigenetics - the study of changes in the regulation of gene activity that are not dependent on gene sequence

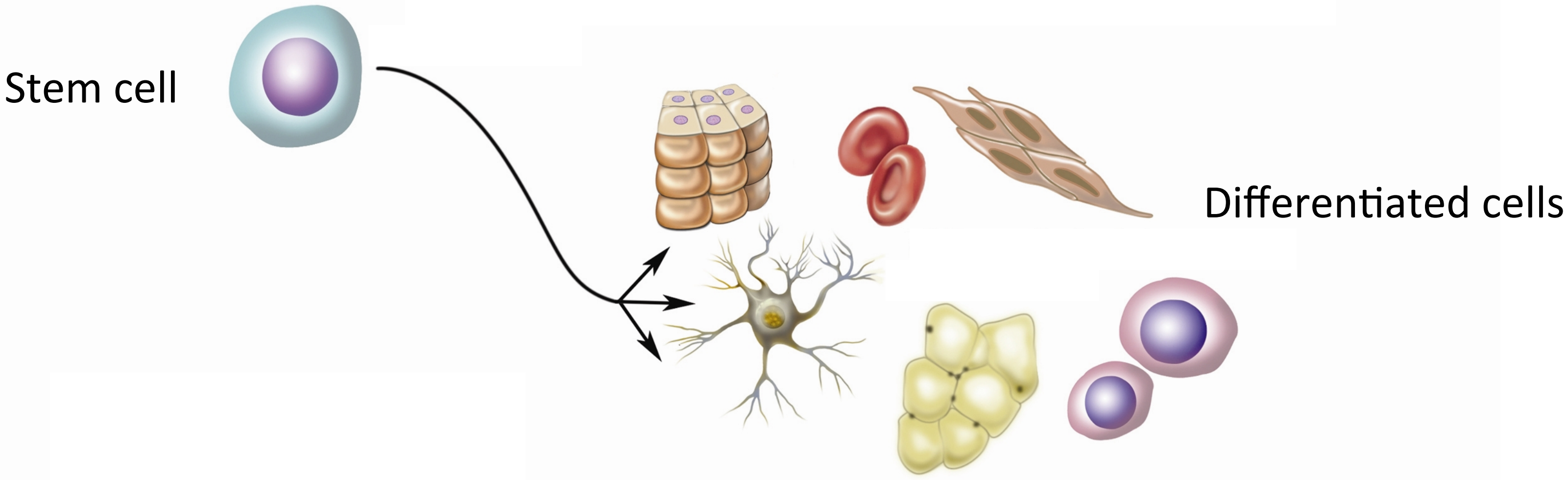
epi - (above) – genetics


H2A Ac–NH–SGRGKQGGKARAKAKTRSSRAGL ····
1
H2B NH3 –PEPSKSAPAPKKGSKKAITKAQKKDGKKRKRSRK ····
14
32
H3K9Ac
H3 NH3 –ARTKQTARKSTGGKAPRKQLATKAARKSAP····
10
28
H4 Ac–NH–SGRGKGGKGLGKGGAKRHRKVLR ····
1
18
DNA

http://bricker.tcnj.edu/Amb/amble9.html

ROADMAP epigenomics PROJECT

A Roadmap Project goal is to create reference epigenomic maps of many different human cell types (i.e. a map of histone modifications across the genome)

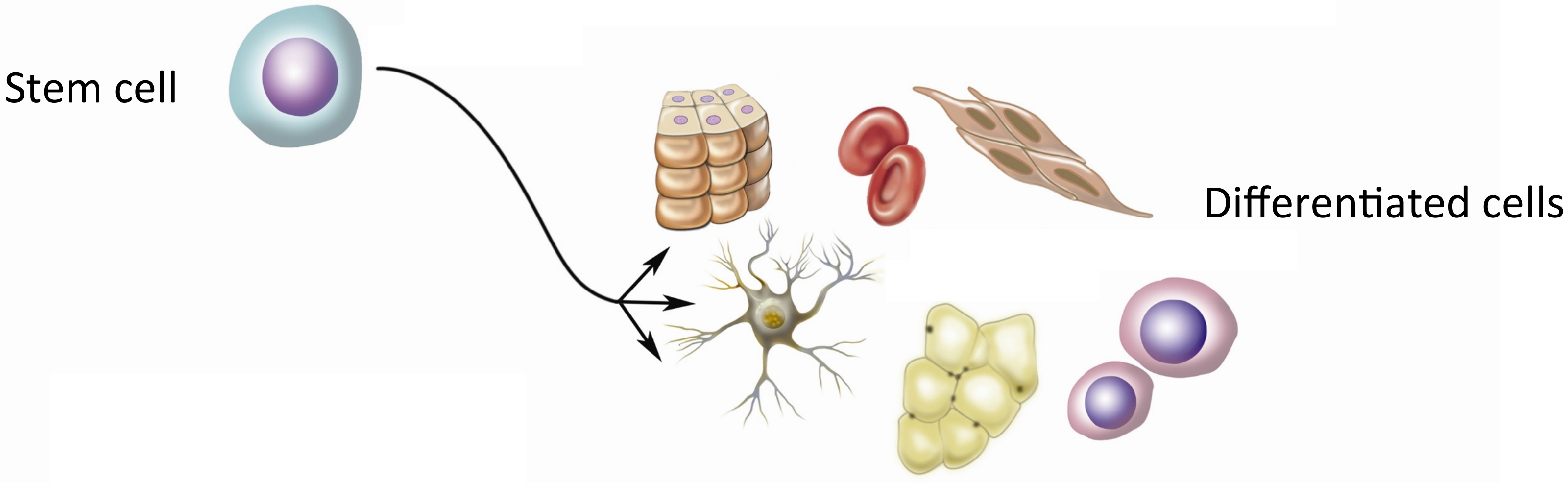


Stem cell

Differentiated cells

ROADMAP epigenomics PROJECT

A Roadmap Project goal is to create reference epigenomic maps of many different human cell types (i.e. a map of histone modifications across the genome)



Stem cell

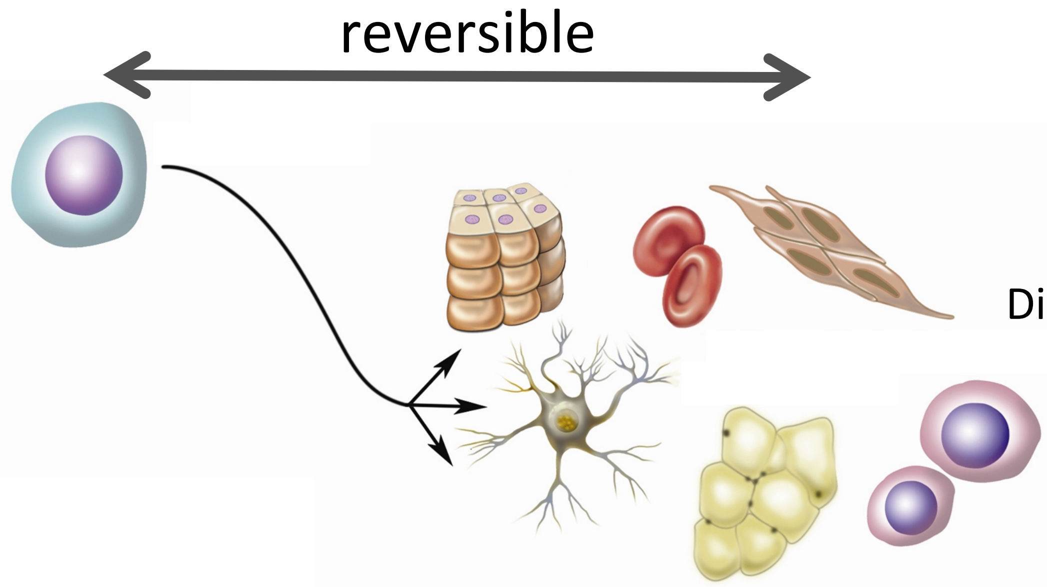
Differentiated cells

Developmentally important genes are “poised”:
H3K4me3 (active) **AND** **H3K27me3** (inactive)

Resolve to either
H3K4me3 OR H3K27me3

ROADMAP epigenomics PROJECT

A Roadmap Project goal is to create reference epigenomic maps of many different human cell types (i.e. a map of histone modifications across the genome)



reversible

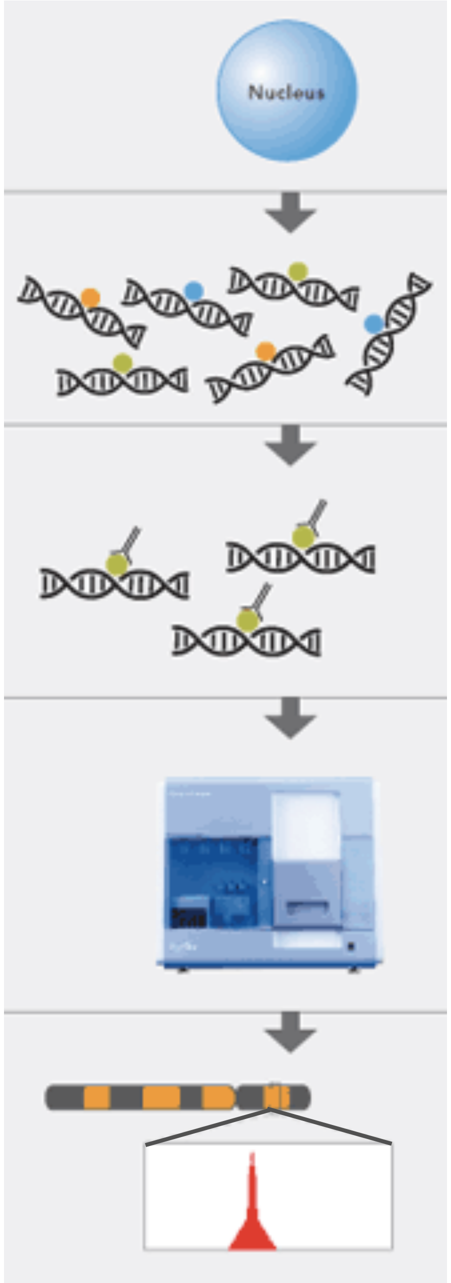
Stem cell

Differentiated cells

Developmentally important genes are “poised”:
H3K4me3 (active) **AND** H3K27me3 (inactive)

Resolve to either
H3K4me3 OR H3K27me3

ChIP-Seq | Chromatin Immunoprecipitation and Sequencing



Cross-link and fractionate chromatin

ChIP - enrich for targeted binding sites

Sequence

Map sequence to reference genome

Peak = putative binding site

illumina

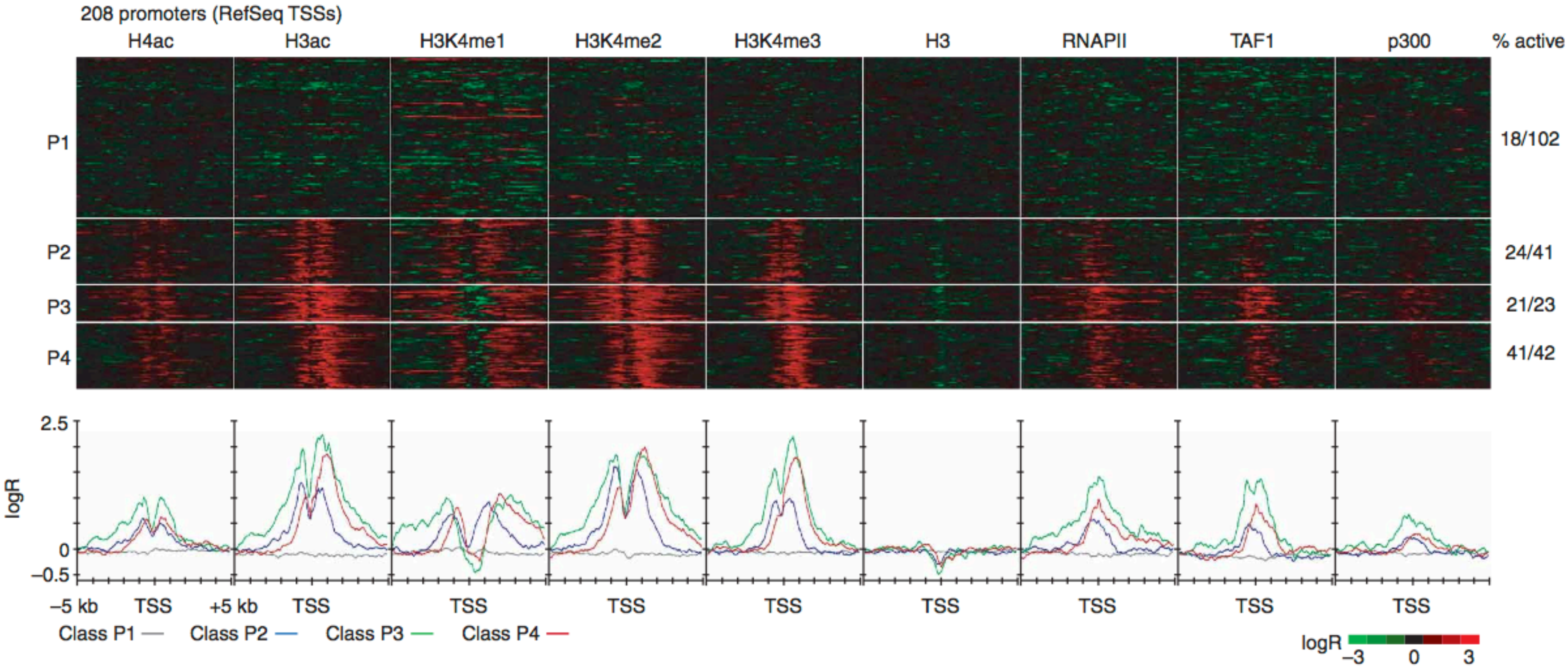
Current computational techniques

Heintzman *et al*. Nature Genetics, 2007

Motivation – "…the distinguishing chromatin features of promoters and enhancers have yet to be determined, hindering our understanding of a predictive histone code for different classes of regulatory elements."

Current computational techniques

Heintzman *et al*. Nature Genetics, 2007

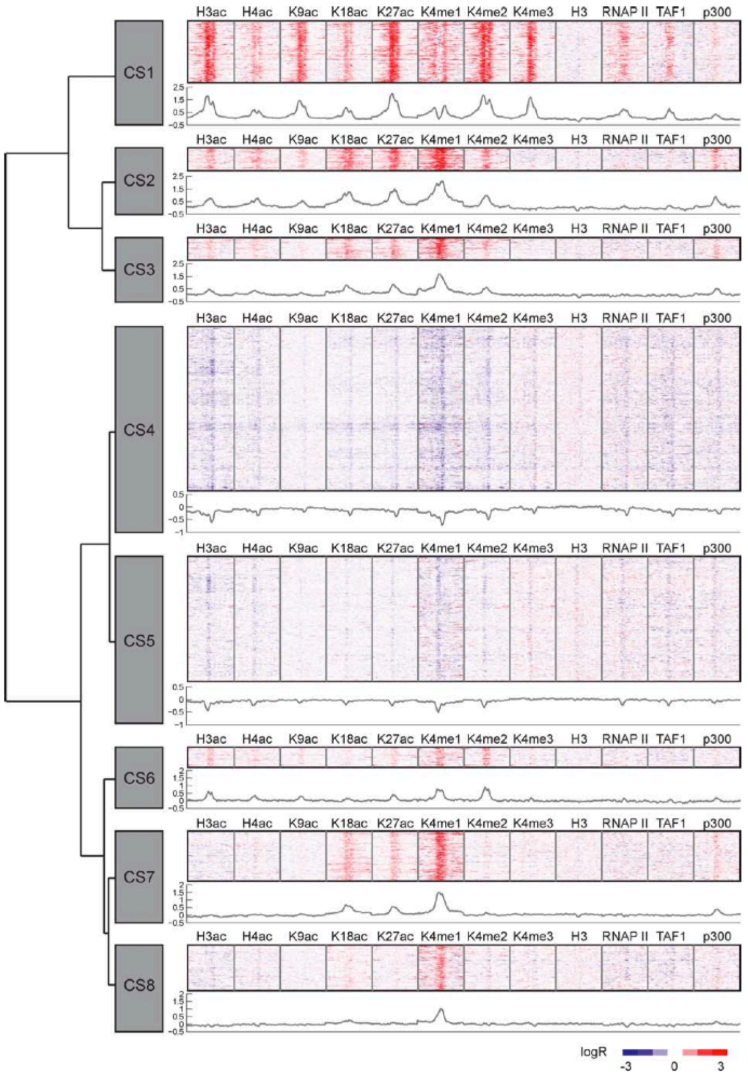


Clustering data from well annotated regions

Current computational techniques

Hon *et al.*
PLoS Comput. Biol., 2008

ChromaSig: a probabilistic method that enables discovery of chromatin signatures de novo (no dependence on annotation)

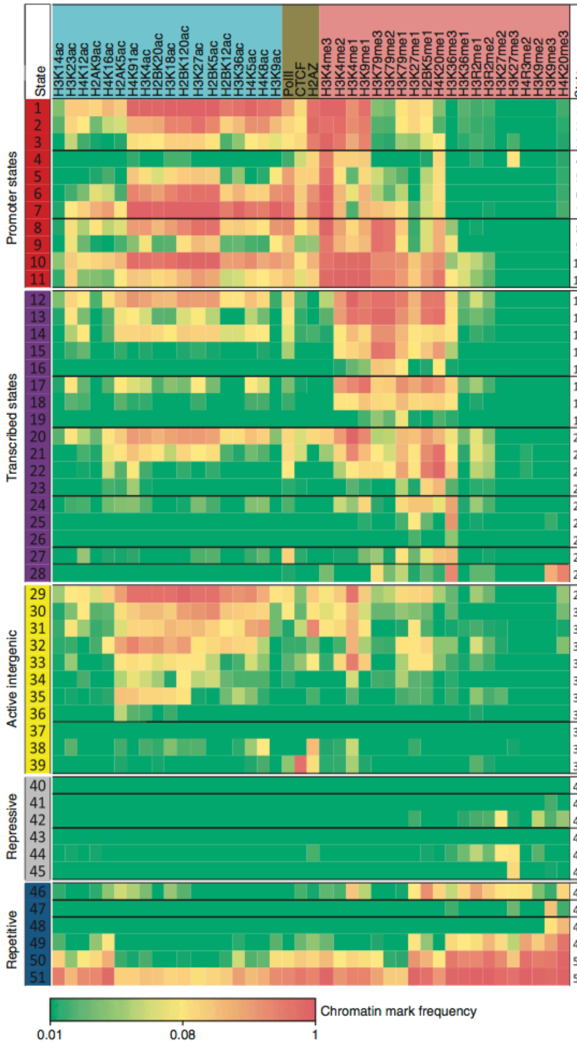


Current computational techniques

Ernst and Kellis Nature Biotechnol. 2010

Hidden Markov Model to reveal 'chromatin states' in human T cells

State display: abstracted away from all detail of primary data



State	Percent of genome	% +-2kb TSS	Percent of TSS	xF TSS exact	% RefSeq gene	Expression level	xF ZNF gene	xF 5' UTR	xF All exons	xF Spliced exons	xF 3' UTR	xF TES	xF Conserved	xF DNaseI	xF TF binding	xF CpG island	% GC	% Lamina	% Repeat	
1	0.11	51	2.8	26	45	0.19	2.9	3.6	4.2	1.5	1.5	2.3	2.3	38	6.3	29	53	13	30	Promoter upstream high expr; potential enh looping
2	0.14	41	2.2	16	44	0.04	2.5	2.9	3.2	1.7	1.6	2.3	2.3	34	5.8	18	50	17	36	Promoter upstream med expr; potential enh looping
3	0.16	52	4.6	28	49	-0.15	2.6	3.7	4.9	2.2	1.7	2.5	2.8	32	7.1	37	53	16	35	Promoter upstream low expr; potential enh looping
4	0.19	57	8.5	44	52	-0.53	1.1	4.5	7.6	2.5	2.0	2.5	3.9	19	5.7	69	61	23	33	Repressed promoter
5	0.11	74	14	125	63	0.36	2.9	7.6	11	1.9	1.8	3.4	4.4	36	9.4	76	64	9.7	49	TSS low-med expr; most GC rich
6	0.11	78	12	109	61	0.41	4.1	7.7	12	1.8	1.9	3.6	4.1	39	10	78	61	9.5	27	TSS med expr
7	0.08	89	9.0	115	70	0.66	5.3	9.5	14	1.2	1.6	1.9	4.3	39	11	86	61	8.7	16	TSS high expr
8	0.07	72	3.0	40	87	0.96	1.8	8.1	7.6	1.9	1.4	1.6	3.7	35	8.6	44	54	10	26	Transcribed promoter; highest expr, TSS for active genes
9	0.10	41	1.0	10	89	1.04	1.3	6.0	3.6	2.2	2.6	3.1	2.8	20	3.3	11	42	13	39	Transcribed promoter; highest expr, downstream
10	0.08	48	0.6	7.3	87	0.75	2.5	5.7	4.0	2.4	1.3	1.3	2.4	33	6.7	17	53	7.1	29	Transcribed promoter; high expr, near TSS
11	0.09	54	0.4	4.3	92	0.85	2.4	5.4	4.0	3.2	2.4	3.4	2.6	25	5.8	10	50	7.2	34	Transcribed promoter; high expr, downstream
12	0.14	8.3	0.1	0.7	93	0.82	0.7	5.2	1.6	1.6	1.1	0.5	1.5	9.1	3.1	0.5	51	7.9	40	Transcribed 5′ proximal, higher expr, open chr, TF binding
13	0.39	6.3	0.2	0.5	95	0.77	1.8	5.8	1.2	1.3	1.0	0.5	1.3	3.3	1.9	0.7	46	6.0	52	Transcribed 5′ proximal, higher expr, open chr
14	0.22	9.2	0.1	0.7	82	0.70	1.4	4.4	1.4	1.5	1.6	1.2	1.5	6.2	1.7	0.2	42	12	55	Transcribed 5′ proximal, high expr, open chr
15	0.78	3.4	0.2	0.3	95	0.67	2.7	5.9	0.9	0.9	0.9	0.4	1.4	1.4	1.0	0.4	40	6.6	64	Transcribed 5′ proximal, high expr
16	0.90	3.7	0.1	0.2	92	0.53	3.2	5.4	0.5	0.5	1.0	0.3	1.0	0.9	0.7	0.1	41	8.5	79	Transcribed 5′ proximal, med expr; Alu repeats
17	0.25	6.6	0.1	0.5	94	0.49	3.2	3.8	1.6	1.6	1.4	0.8	1.3	2.9	2.0	0.1	46	10	52	Transcribed less 5′ proximal, med expr, open chr
18	0.79	2.7	0.2	0.3	93	0.43	2.8	3.5	1.3	1.4	1.1	0.7	1.3	1.1	1.0	0.1	42	10	60	Transcribed less 5′ proximal, med expr
19	2.02	2.1	0.4	0.2	89	0.32	2.6	3.5	1.0	1.0	1.2	0.8	1.1	0.5	0.6	0.1	40	11	73	Transcribed less 5′ proximal, lower expr; Alu repeats
20	0.14	21	0.4	2.6	70	0.39	2.4	2.9	3.5	3.5	3.0	3.4	1.9	20	5.2	4.1	52	9.2	39	Candidate strong enhancer in transcribed regions
21	0.21	28	0.7	3.5	83	0.51	2.8	2.7	7.8	8.1	4.8	8.4	2.4	8.5	4.9	9.1	55	3.9	36	Spliced exons/GC rich; open chr, TF binding
22	0.42	4.4	0.4	1.0	92	0.54	3.3	1.8	8.2	9.2	3.5	4.5	2.2	0.9	1.2	6.4	55	3.5	34	Spliced exons/GC rich
23	0.66	10	1.3	2.0	81	0.29	2.7	1.8	6.1	6.5	4.6	7.5	1.7	1.8	1.4	7.7	54	3.2	51	Spliced exons/GC rich; Alu repeats
24	0.76	1.3	0.3	0.4	90	0.40	2.9	0.8	5.1	5.7	3.3	2.9	1.8	0.8	1.0	0.3	46	13	48	Transcribed 5′ distal; exons
25	1.34	0.6	0.3	0.2	91	0.32	3.6	0.4	6.4	7.3	4.8	5.0	2.0	0.4	0.6	0.1	42	19	52	Transcribed further 5′ distal; exons
26	5.28	1.0	1.0	0.2	86	0.20	2.1	0.7	2.7	3.1	3.1	3.2	1.4	0.4	0.5	0.2	40	17	66	Transcribed 5′ distal; Alu repeats
27	0.39	3.2	0.4	1.0	88	0.92	1.8	0.5	8.6	9.7	8.9	12	2.2	1.5	0.9	0.8	44	6.8	48	End of transcription; exons; high expr
28	0.17	4.2	0.2	1.4	77	0.09	112	3.2	5.4	5.9	5.0	5.9	0.8	0.2	0.1	0.3	41	23	67	ZNF genes; KAP-1 repressed state
29	0.14	11	0.2	1.3	44	0.18	0.8	1.4	1.7	1.8	2.0	1.7	1.5	22	4.7	0.9	49	15	44	Cand strong distal enh; higher open chr; higher target expr
30	0.18	15	0.3	1.9	46	0.33	1.2	1.6	1.7	1.7	2.0	2.8	1.7	18	3.5	1.3	43	19	54	Cand strong distal enh; high open chr; higher target expr
31	0.17	15	0.4	2.2	32	-0.27	0.6	1.1	1.2	1.1	1.1	1.2	1.7	16	4.0	1.3	46	24	47	Intergenic H2AZ with open chr/TF binding. Cand. distal enh
32	0.24	8	0.3	1.1	42	-0.05	0.6	1.2	2.1	2.2	1.8	1.9	1.3	5.5	2.7	0.8	51	15	48	Candidate weak distal enhancer
33	0.34	14	0.6	1.9	51	0.02	1.3	1.8	1.9	1.9	2.0	2.6	1.6	11	3.2	1.6	48	15	54	Candidate distal enhancer
34	0.70	13	0.7	1.0	52	0.20	1.3	1.8	1.7	1.8	2.2	2.9	1.3	4.8	2.1	0.9	44	16	65	Proximal to active enhancers; Alu repeats
35	0.75	5.3	0.8	1.1	34	-0.26	0.5	0.8	1.7	1.8	1.5	1.7	1.3	1.7	2.2	0.4	49	18	52	Active intergenic regions not enhancer specific
36	3.61	4.3	2.7	0.8	33	-0.32	0.5	0.9	1.3	1.3	1.2	1.5	1.1	0.8	1.5	0.5	46	20	66	Active intergenic further from enhancers; Alu repeats
37	11.0	2.6	5.4	0.5	38	-0.38	0.6	1.0	1.0	1.0	1.1	1.1	1.0	0.5	0.9	0.5	42	26	71	Non-repressive intergenic domains; Alu repeats
38	0.84	8.4	1.1	1.3	32	-0.49	0.4	1.0	1.0	0.9	0.8	0.9	1.5	3.6	2.2	0.6	43	31	56	H2AZ specific state
39	0.17	4.1	0.3	1.5	37	-0.32	0.4	0.8	1.5	1.6	1.5	2.2	2.5	19	2.1	2.1	44	26	45	CTCF island; candidate insulator
40	13.2	0.9	4.2	0.3	9.7	-0.41	0.0	0.3	0.3	0.2	0.5	0.4	0.3	0.1	0.1	0.3	40	50	33	Unmappable
41	23.3	0.3	2.6	0.1	22	-0.80	0.4	0.5	0.2	0.2	0.4	0.2	0.8	0.1	0.3	0.1	37	66	66	Heterochr; nuclear lamina; most AT rich
42	1.55	0.5	0.4	0.2	23	-0.77	0.4	0.6	0.2	0.2	0.4	0.2	0.9	0.1	0.5	0.2	39	68	63	Heterochr; nuclear lamina; ERVL repeats
43	22.3	1.3	8.5	0.4	33	-0.69	0.4	0.8	0.6	0.6	0.6	0.6	1.1	0.2	1.0	0.2	41	48	66	Heterochr; lower gene depletion
44	1.95	1.5	1.1	0.5	34	-0.67	0.5	0.8	0.8	0.8	0.7	0.6	1.1	0.3	1.2	0.3	44	47	58	Heterochr; ERVL repeats: lower gene/exon depletion
45	1.59	11	4.2	2.6	34	-0.67	0.5	1.4	1.2	1.0	1.0	1.4	1.7	0.9	1.6	4.9	45	42	54	Specific repression
46	0.09	4.2	0.0	0.6	42	-0.43	0.9	1.0	1.5	1.5	1.0	1.1	0.7	1.2	2.1	2.2	51	24	69	Simple repeats (CA)n, (TG)n
47	1.09	1.6	0.4	0.4	29	-0.53	0.8	0.8	0.4	0.4	0.7	0.8	0.7	0.1	0.5	0.3	38	46	79	L1/LTR repeats
48	0.67	2.7	0.5	0.8	17	-0.22	11	0.7	0.9	0.7	1.6	1.8	0.3	0.2	0.4	1.3	41	57	83	Satellite repeat
49	0.06	1.9	0.1	0.9	8.7	-0.43	4.2	0.3	0.6	0.4	1.6	0.8	0.3	1.8	0.6	1.3	40	61	83	Satellite repeat; moderate mapping bias
50	0.02	0.9	0.0	0.2	5.8	-0.03	1.2	0.1	0.2	0.2	0.4	0.0	0.2	6.6	0.5	1.0	41	53	88	Satellite repeat; high mapping bias
51	0.01	1.1	0.0	0.8	6.1	0.05	0.0	0.1	0.4	0.4	0.0	0	0.5	20	15	3.5	43	52	88	Satellite repeat/rRNA; extreme mapping bias
	100	2.7	100	0.1	36	-0.16	0.5	5.1	3.2	2.8	1.5	0.1	15	2.5	N/A	0.9	41	43	61	Genome total/average

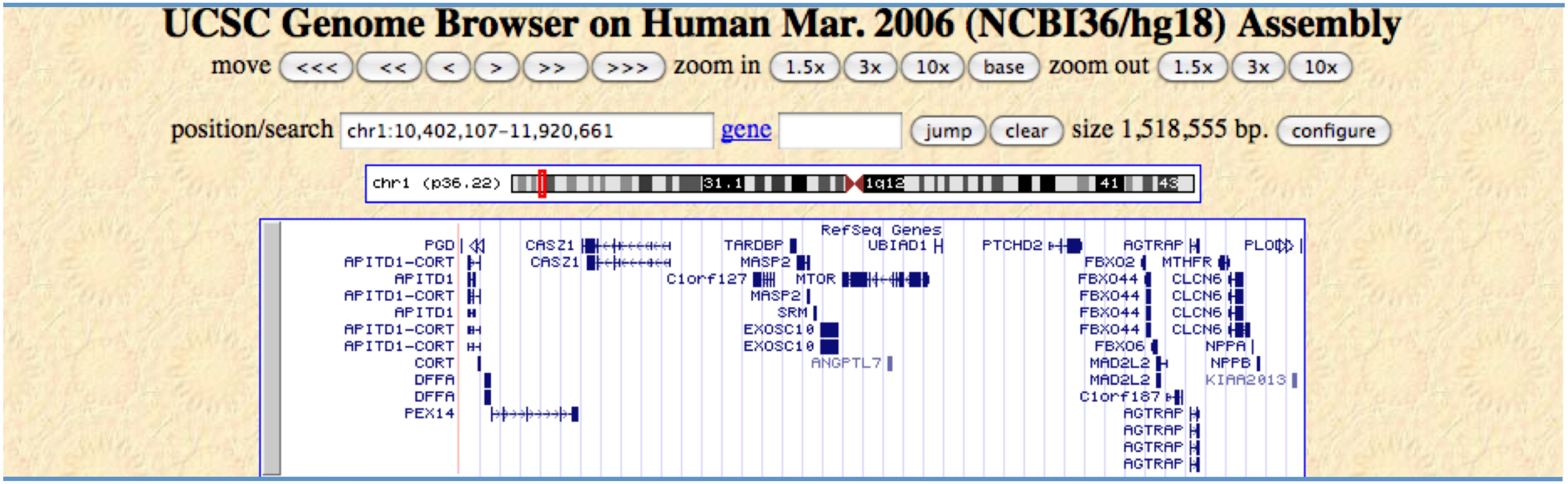
Current computational techniques

- Require significant computational skill to use (only ChromaSig provides an implementation)

- All produce static overview images and do not support interactive data exploration of individual genes within a signature class

- No integration with downstream processing (e.g. gene ontology enrichments, etc.)

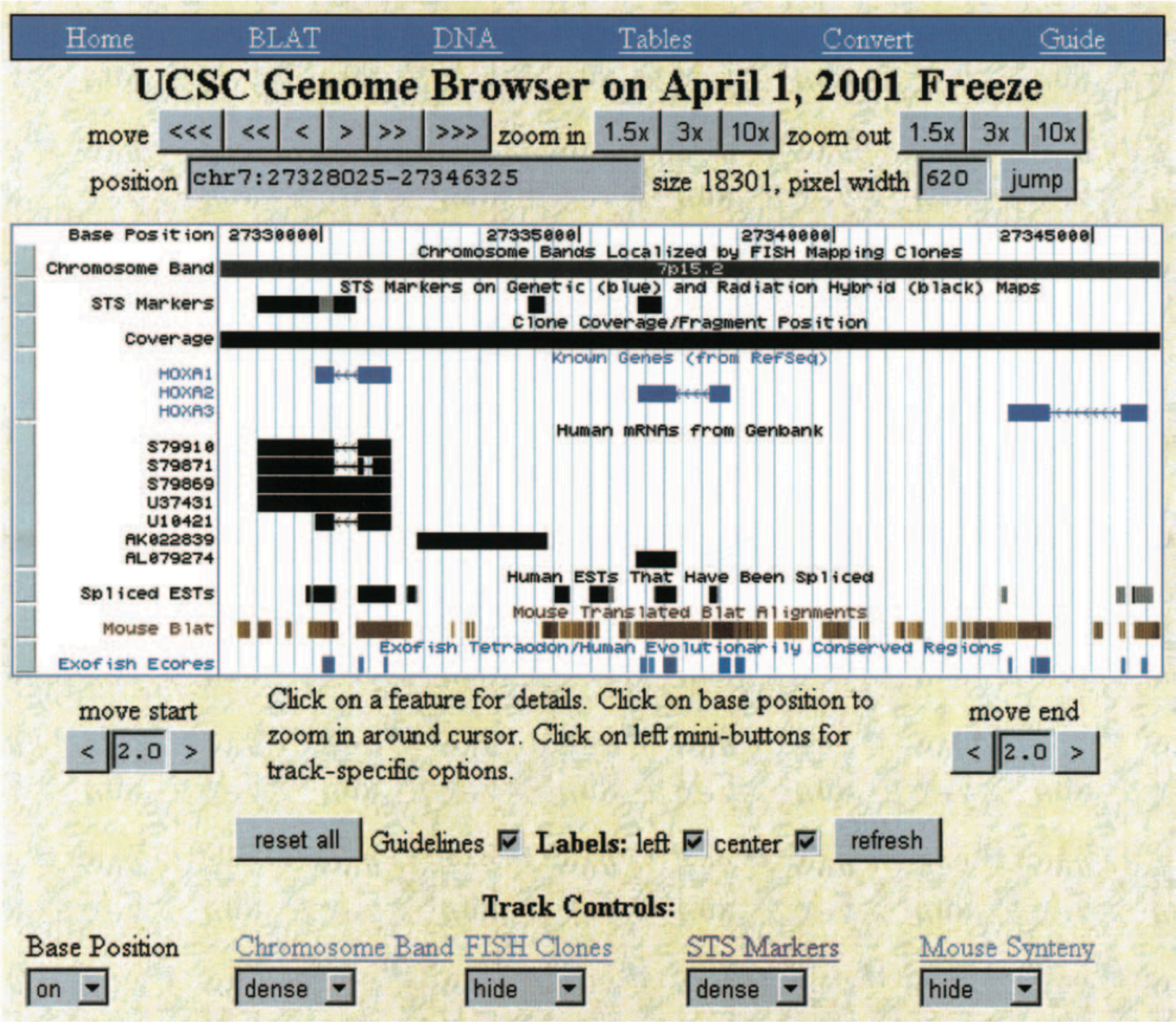
Current visualization techniques



UCSC Genome Browser on Human Mar. 2006 (NCBI36/hg18) Assembly

position/search chr1:10,402,107-11,920,661

size 1,518,555 bp.



UCSC Genome Browser on April 1, 2001 Freeze

position chr7:27328025-27346325 size 18301, pixel width 620

move start < 2.0 >

Click on a feature for details. Click on base position to zoom in around cursor. Click on left mini-buttons for track-specific options.

move end < 2.0 >

Guidelines ☑ **Labels:** left ☑ center ☑

Track Controls:

Base Position: on

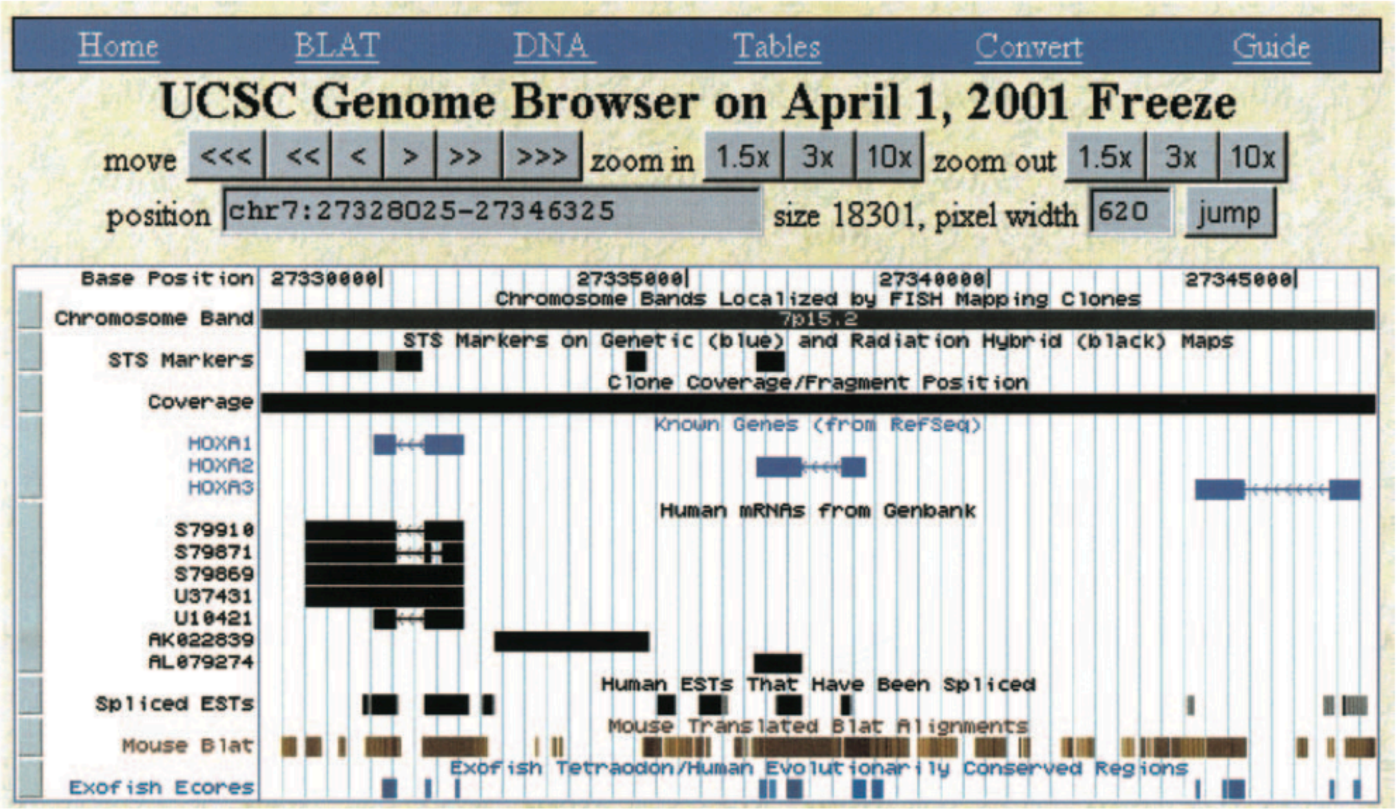
Chromosome Band: dense

FISH Clones: hide

STS Markers: dense

Mouse Synteny: hide

Kent *et al*. Genome Research, 2002



UCSC Genome Browser on April 1, 2001 Freeze

position chr7:27328025-27346325

size 18301, pixel width 620

Home
BLAT
DNA
Tables
Convert
Guide
move
<<<
<<
<
>
>>
>>>
zoom in
1.5x
3x
10x
zoom out
1.5x
3x
10x
jump
Base Position
27330000
27335000
27340000
27345000
Chromosome Bands Localized by FISH Mapping Clones
Chromosome Band
7p15.2
STS Markers on Genetic (blue) and Radiation Hybrid (black) Maps
STS Markers
Clone Coverage/Fragment Position
Coverage
Known Genes (from RefSeq)
HOXA1
HOXA2
HOXA3
Human mRNAs from Genbank
S79910
S79871
S79869
U37431
U10421
AK022839
AL079274
Human ESTs That Have Been Spliced
Spliced ESTs
Mouse Translated Blat Alignments
Mouse Blat
Exofish Tetraodon/Human Evolutionarily Conserved Regions
Exofish Ecores

“The UCSC browser had humble origins. The code originated with a small script in the C programming language, which displayed a splicing diagram for a gene prediction from the nematode *C. elegans* (Kent and Zahler 2000). This web-based splicing display later acquired tracks for mRNA alignments and for homology with the related nematode *Caenorhabditis briggsae*. This was published as the tracks display at http://www.cse.ucsc.edu/~kent/intronerator”

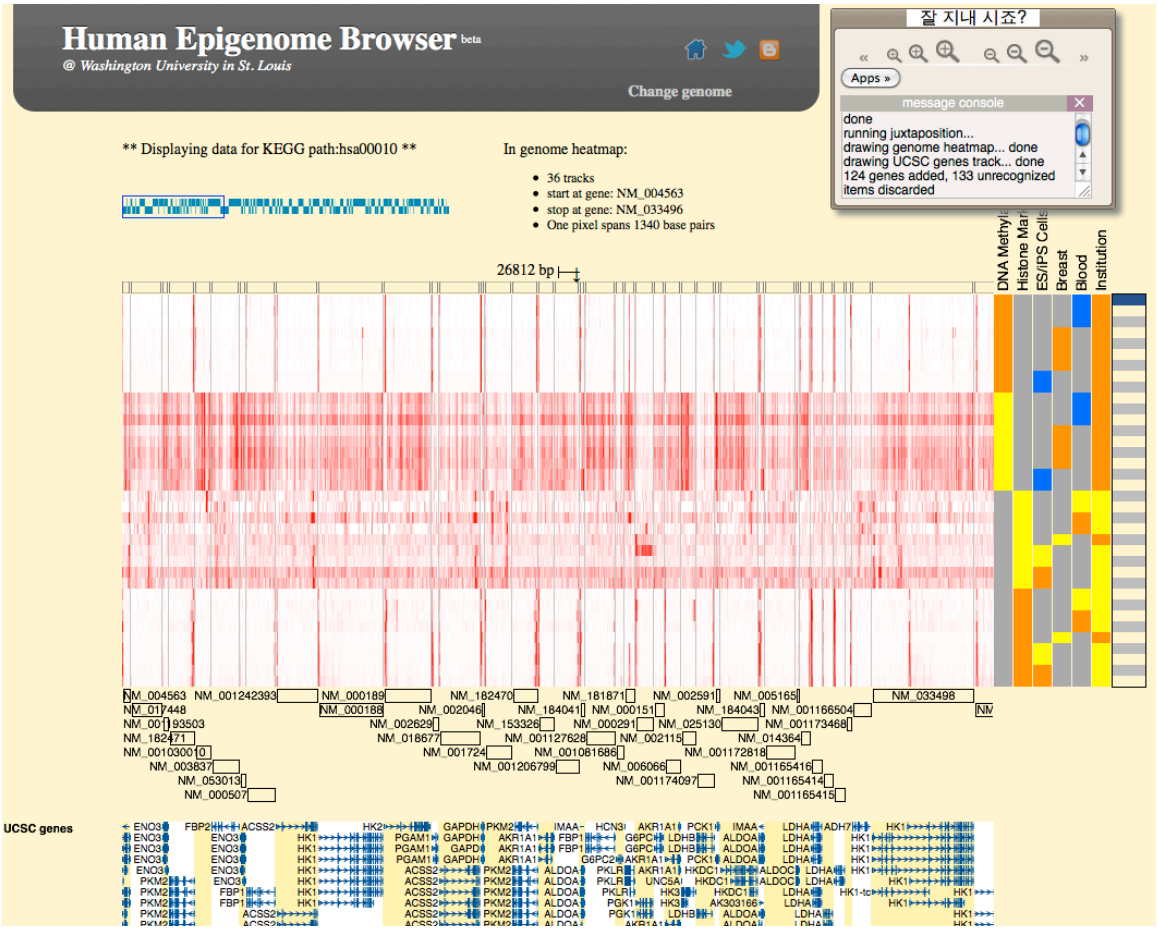
Kent *et al*. Genome Research, 2002

Current visualization techniques

Genome Browsers

- Intended to provide a local view of a genomic region
- In the absence of other tools, many biologists attempt to informally use them to gather a global overview of data patterns
- In these cases, there is a mismatch between the level of data abstraction at which a biologist reasons about the problem and the level provided by the browser
- Functional similarity should drive navigation not genomic position

Current visualization techniques



Human Epigenome Browser beta

@ Washington University in St. Louis

Change genome

** Displaying data for KEGG path:hsa00010 **

In genome heatmap:

- 36 tracks
- start at gene: NM_004563
- stop at gene: NM_033496
- One pixel spans 1340 base pairs

잘 지내 시죠?

Apps »

message console

```
done
running juxtaposition...
drawing genome heatmap... done
drawing UCSC genes track... done
124 genes added, 133 unrecognized
items discarded
```

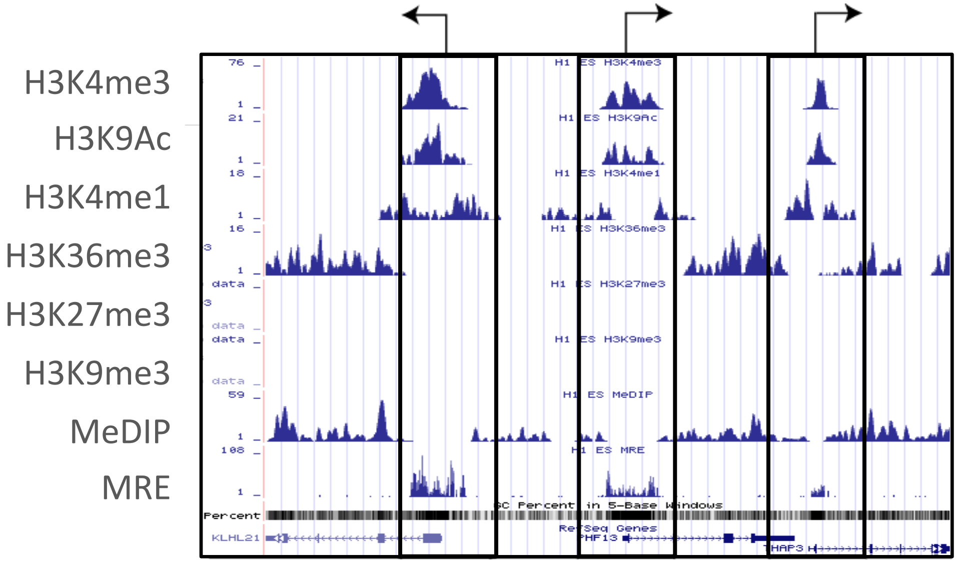
Zhou *et al*. Nature Methods, 2011

Can filter genomic x-axis to just display genes from a pathway of interest (by KEGG ID)

Spark – A Discovery Tool

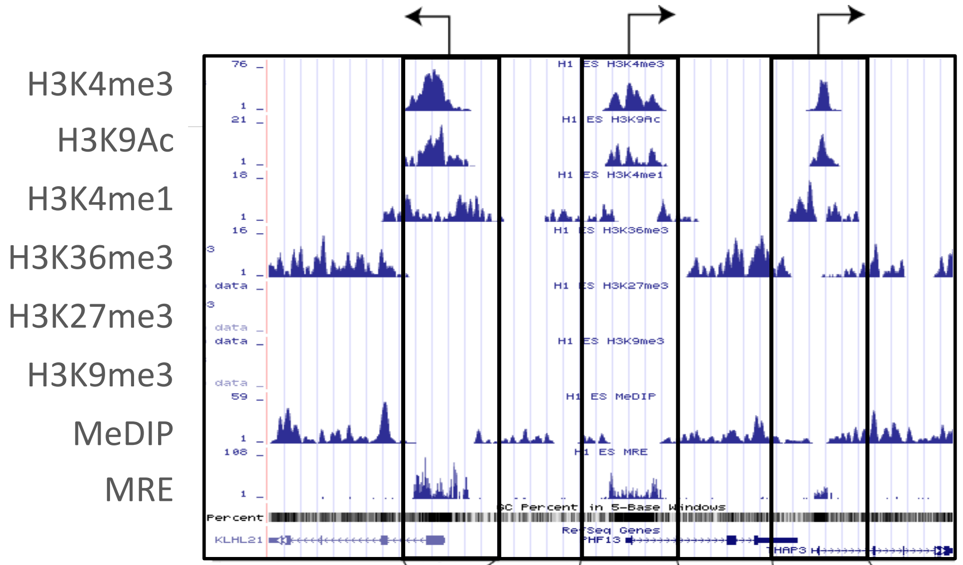
Spark

1. Focus on regions of interest (e.g. transcriptional start sites (TSSs) +/- 3000 nt)



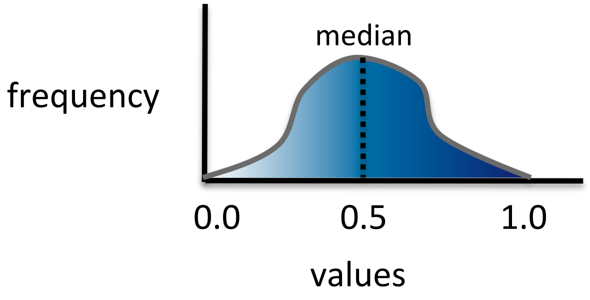
Spark

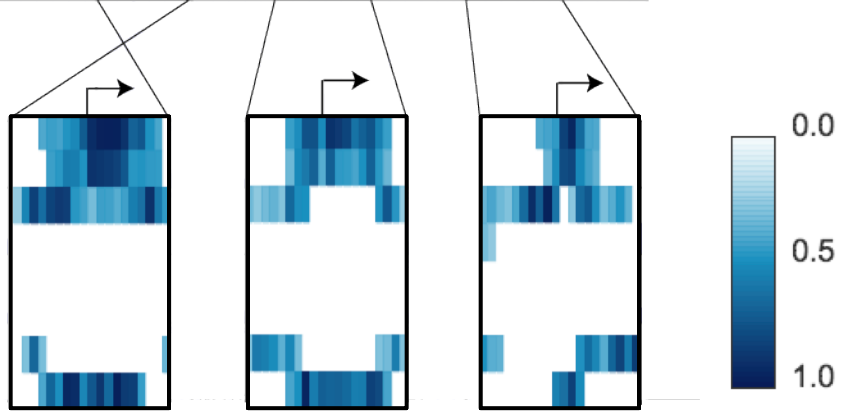
1. Focus on regions of interest (e.g. transcriptional start sites (TSSs) +/- 3000 nt)



2. Extract data matrices

Normalization for bin *i*, sample *h*:

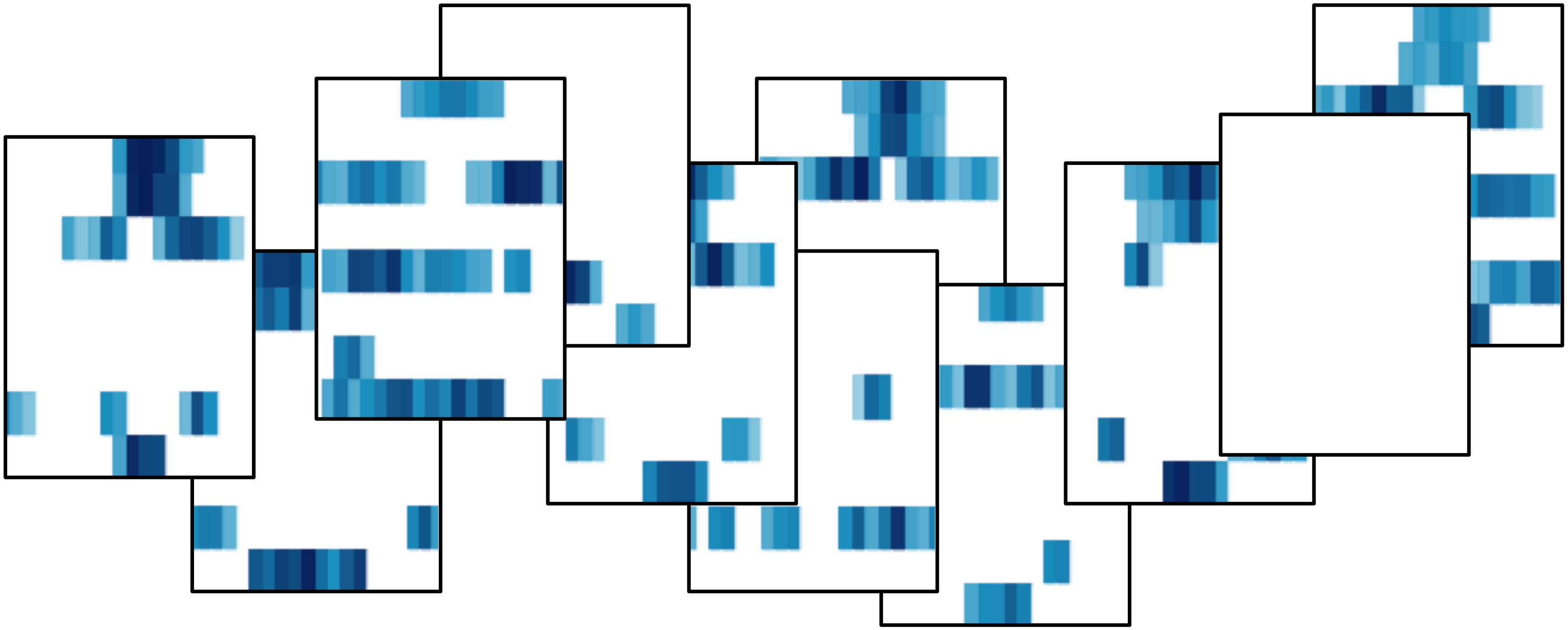




$$x'_{h,i} = \frac{1}{1+e^{-(x_{h,i}-median(x_h))/std(x_h)}}$$

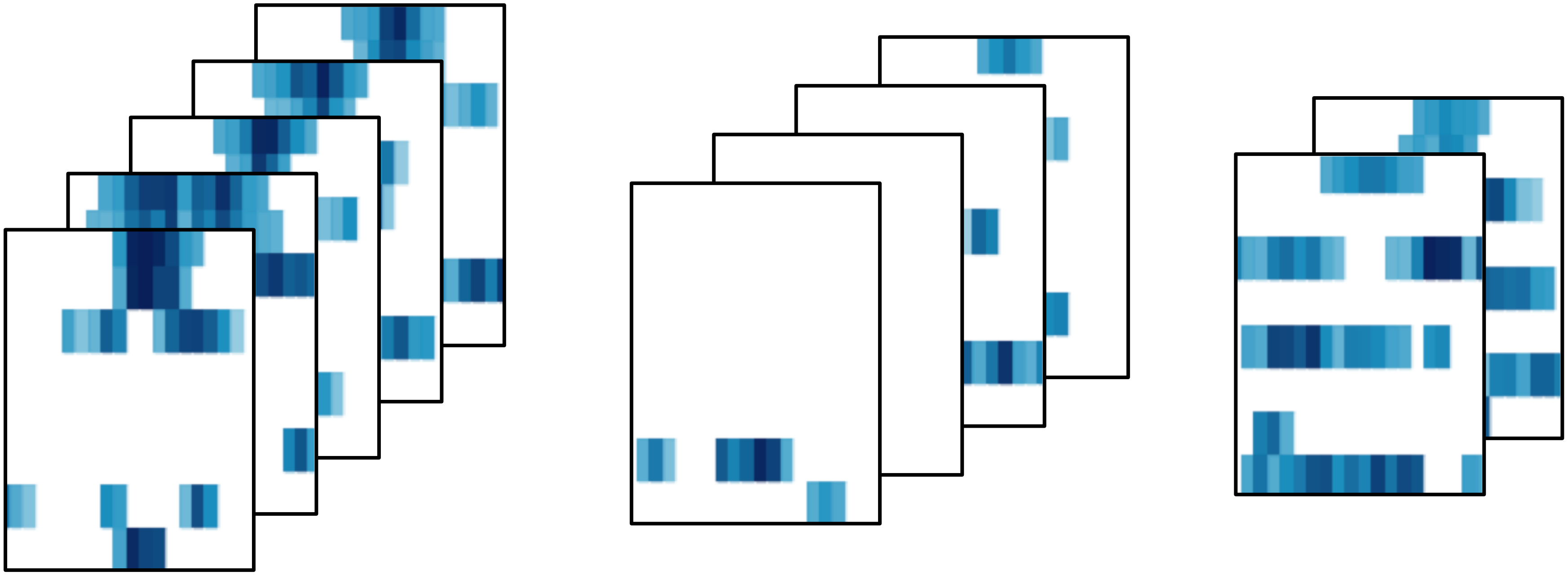
Spark

3. Cluster matrices (*k*-means)



Spark

3. Cluster matrices (*k*-means) $k = 3$



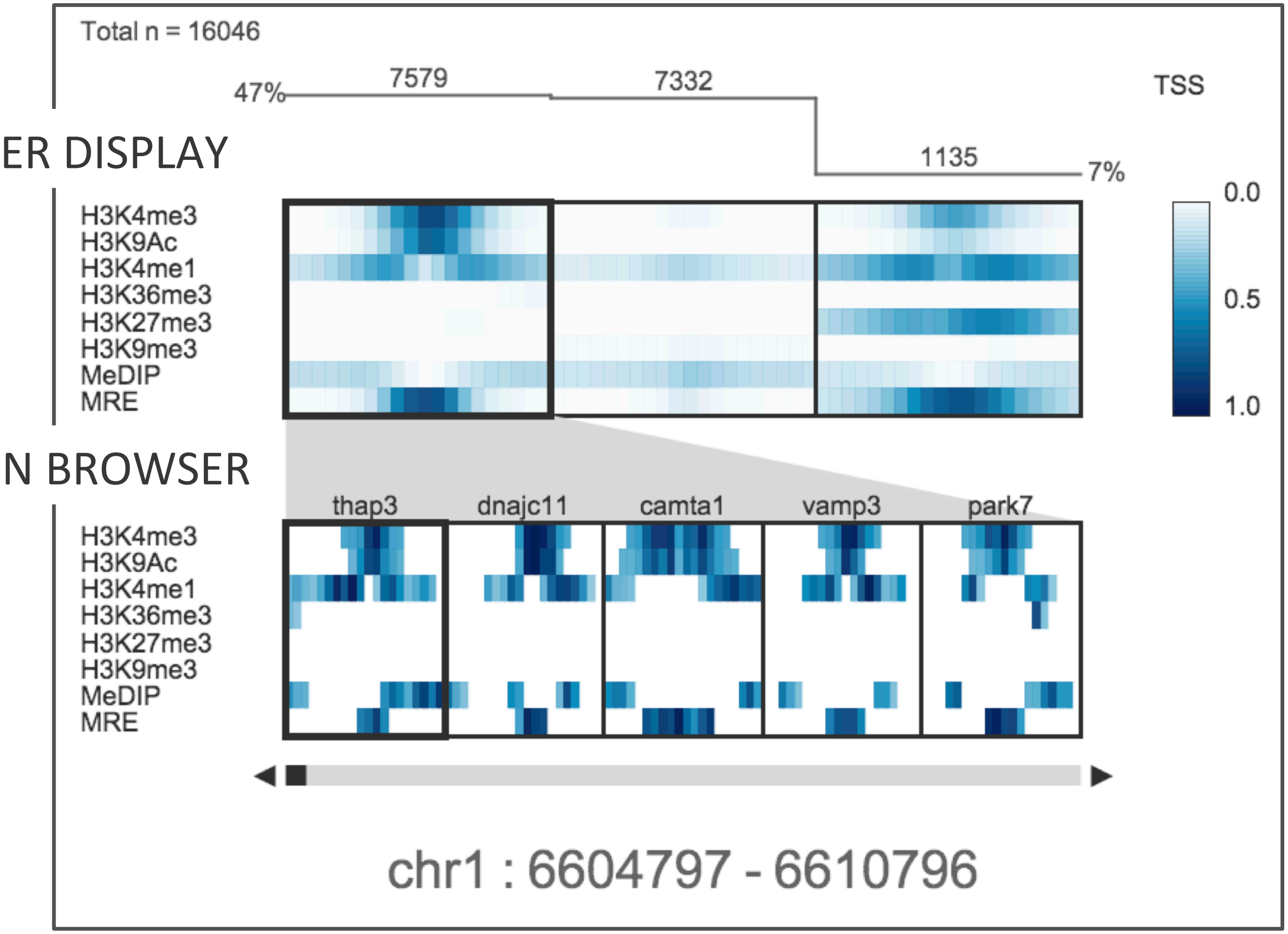
Cluster 1

Cluster 2

Cluster 3

Spark

4. Interactive cluster visualization - data from human H1 embryonic stem cells



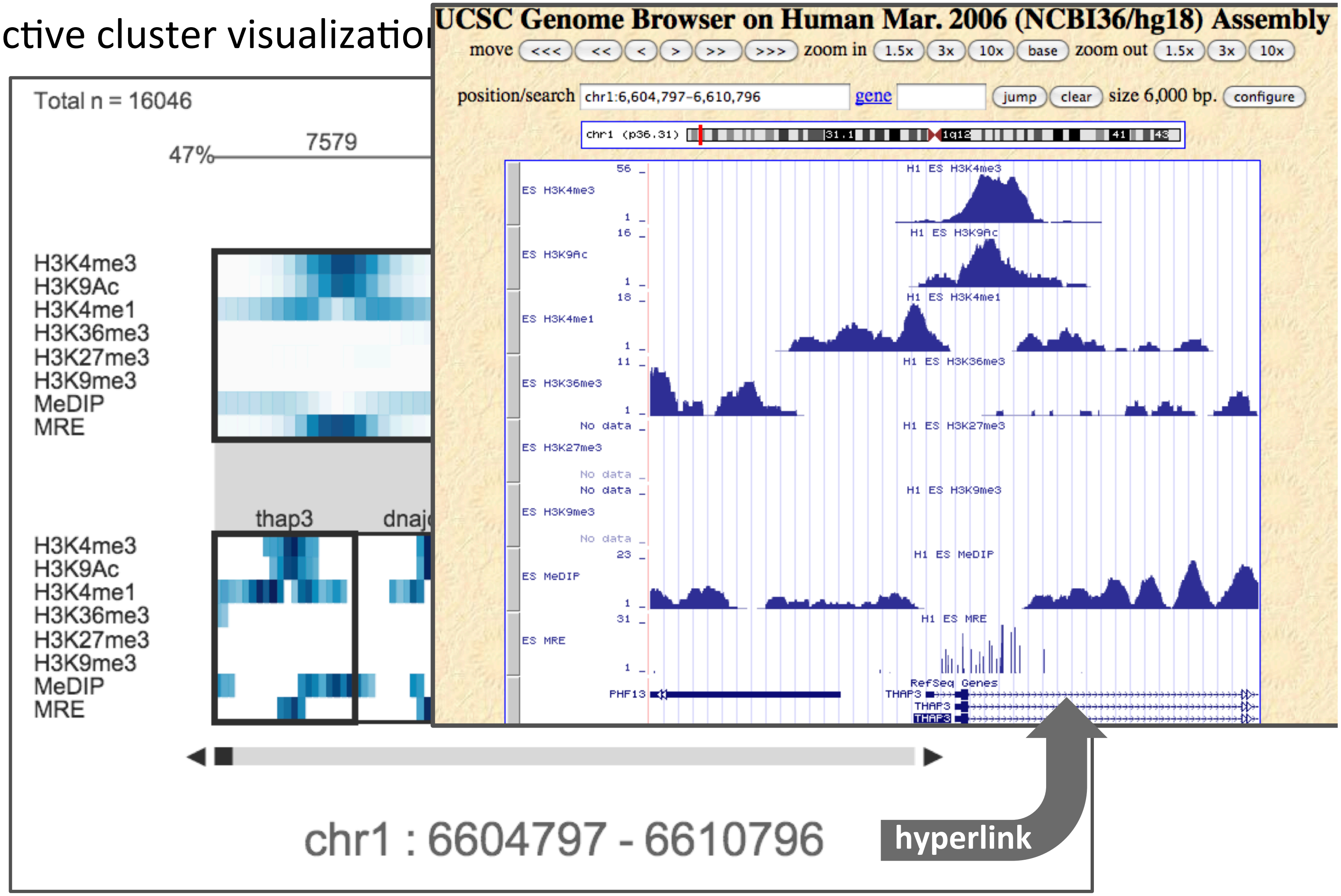
CLUSTER DISPLAY

REGION BROWSER

screenshot

Spark

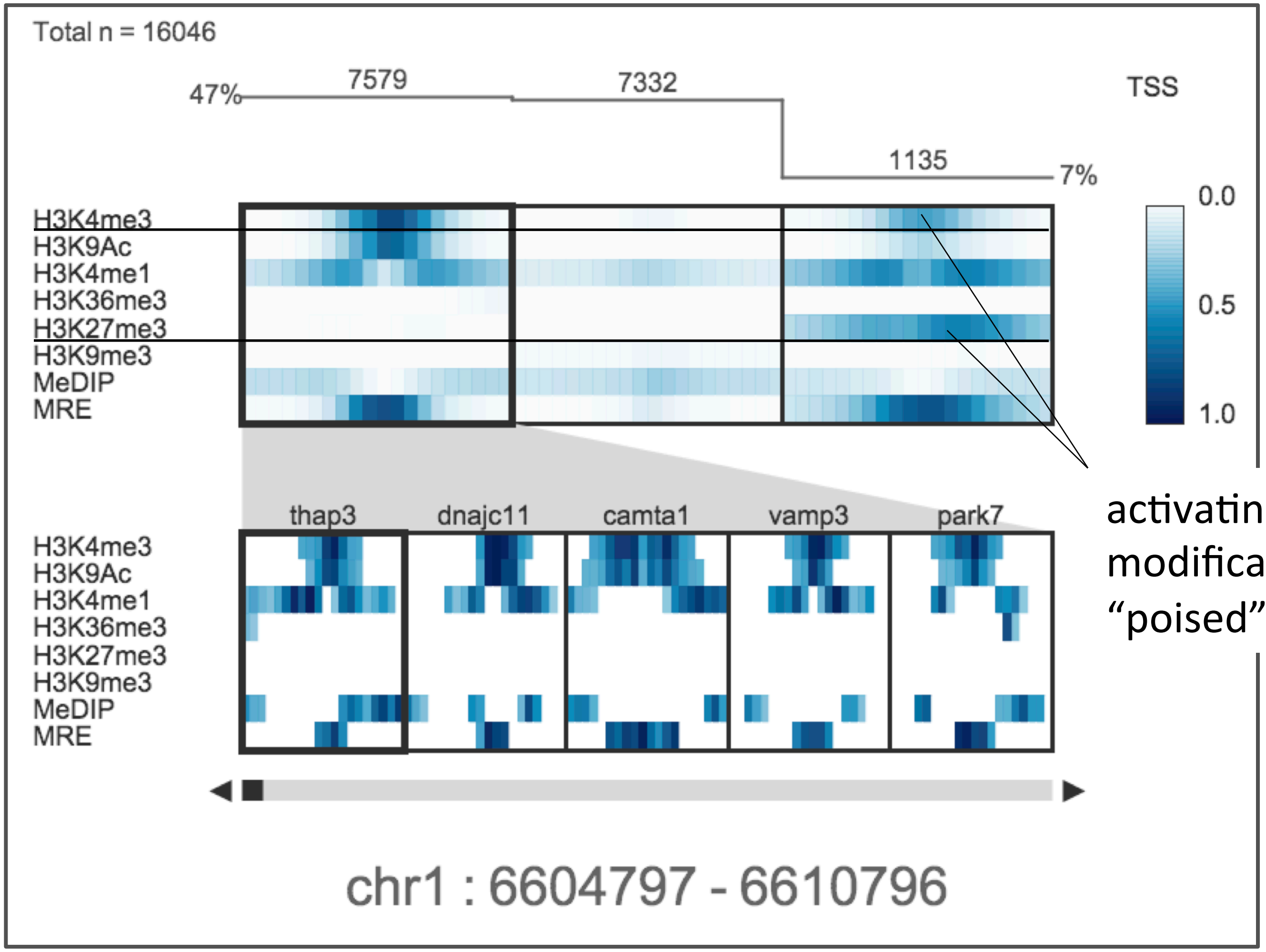
4. Interactive cluster visualizatio



screenshot

Spark

4. Interactive cluster visualization - data from human H1 embryonic stem cells



activating + silencing modifications: “poised”

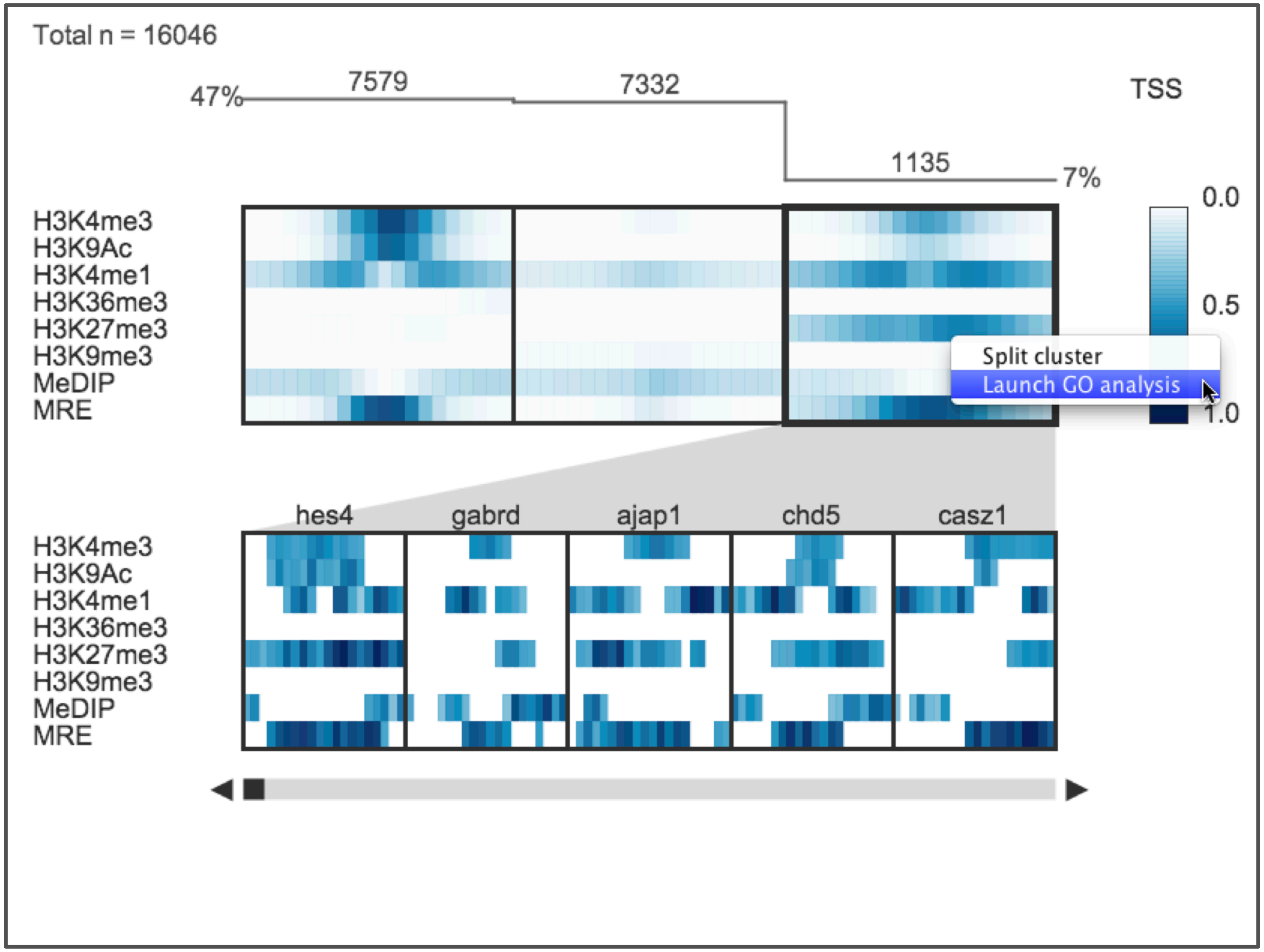
screenshot

Spark

Do these clusters have functional meaning?

Spark

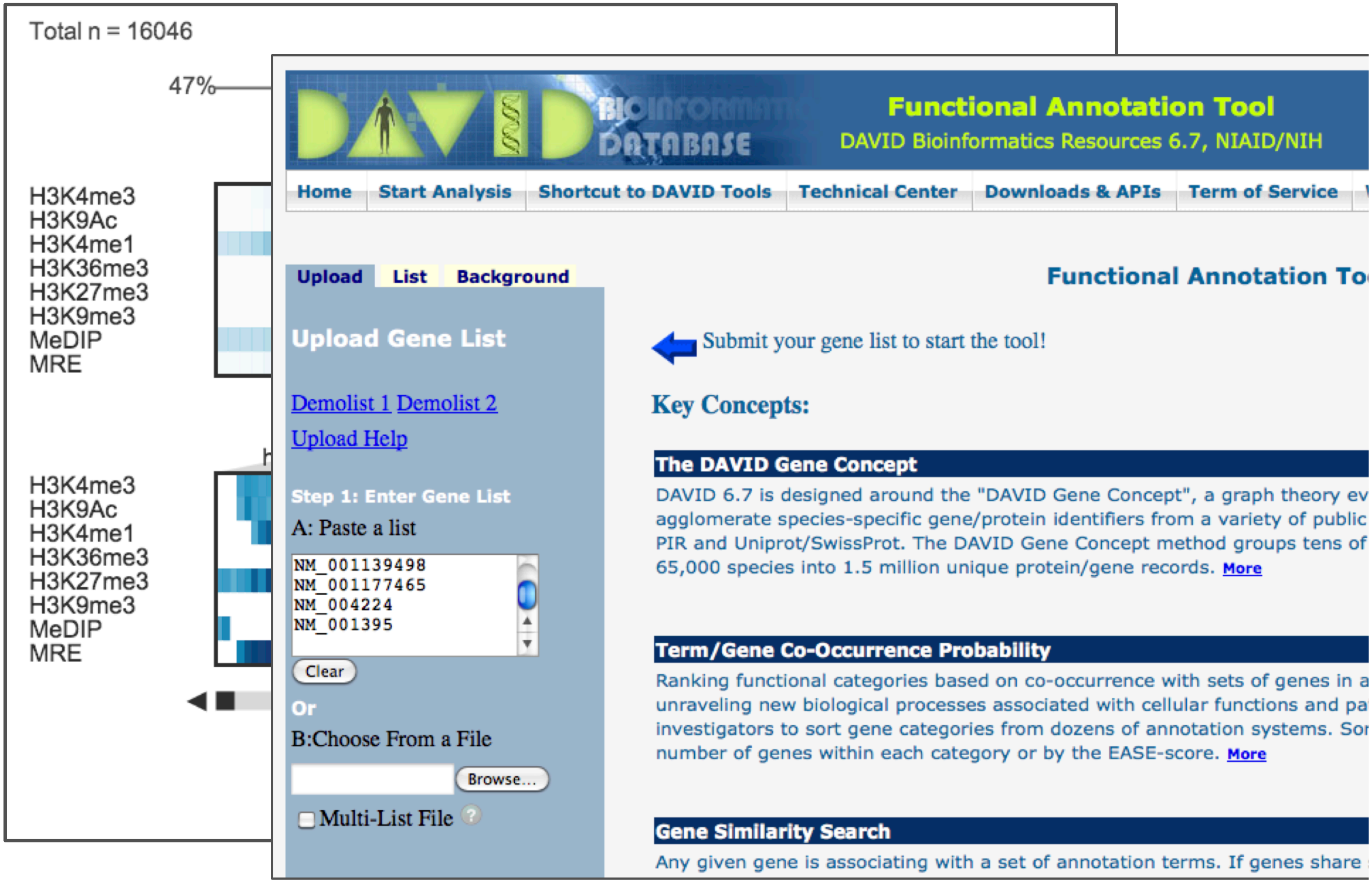
5. Interactive gene ontology analysis



screenshot

Spark

5. Interactive gene ontology analysis



Functional Annotation Tool

DAVID Bioinformatics Resources 6.7, NIAID/NIH

Upload Gene List

Demolist 1 Demolist 2

Upload Help

Step 1: Enter Gene List

A: Paste a list

```
NM_001139498
NM_001177465
NM_004224
NM_001395
```

Or

B:Choose From a File

□ Multi-List File

Functional Annotation To

Submit your gene list to start the tool!

Key Concepts:

The DAVID Gene Concept

DAVID 6.7 is designed around the "DAVID Gene Concept", a graph theory ev agglomerate species-specific gene/protein identifiers from a variety of public PIR and Uniprot/SwissProt. The DAVID Gene Concept method groups tens of 65,000 species into 1.5 million unique protein/gene records. More

Term/Gene Co-Occurrence Probability

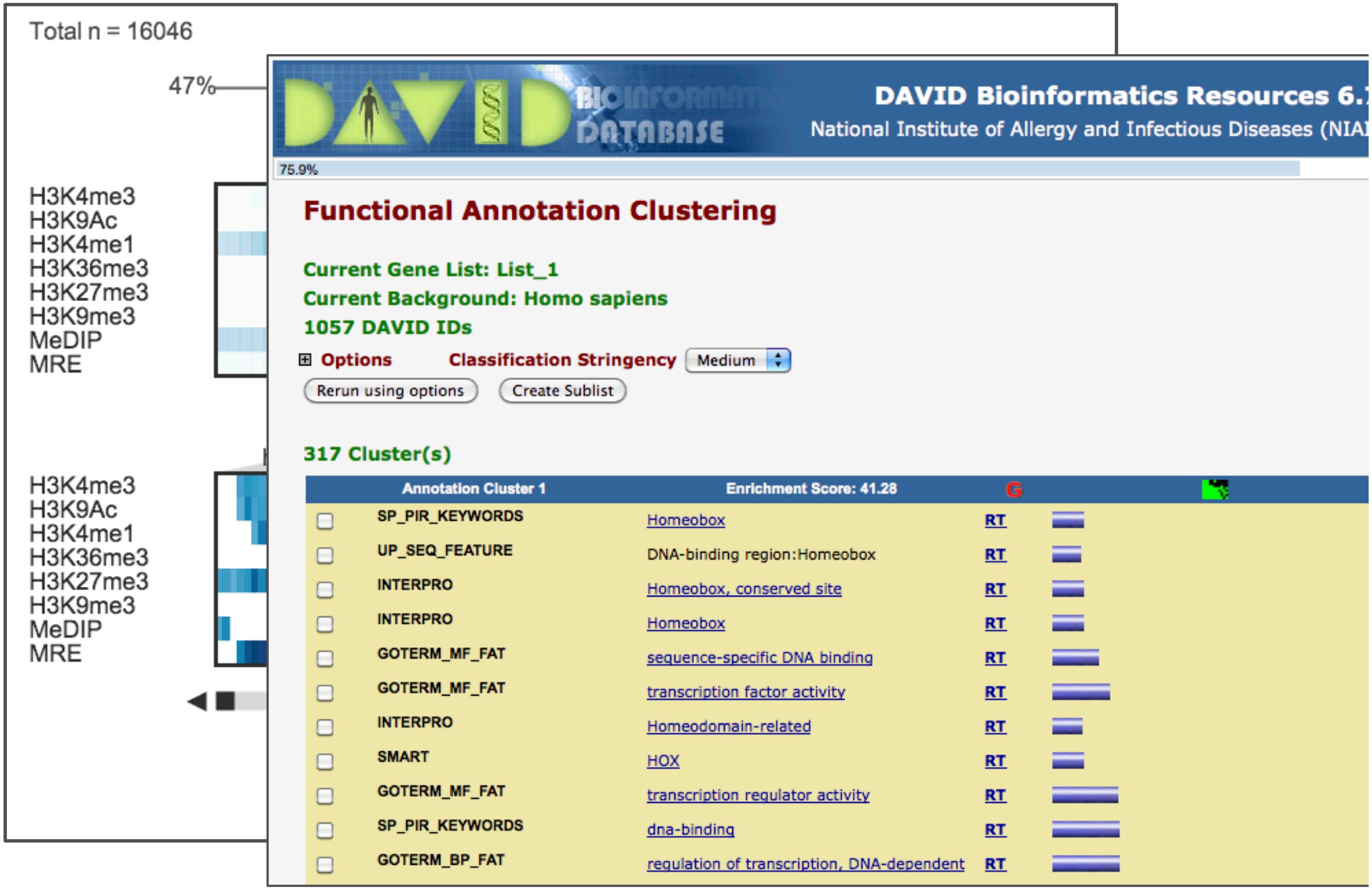
Ranking functional categories based on co-occurrence with sets of genes in a unraveling new biological processes associated with cellular functions and pa investigators to sort gene categories from dozens of annotation systems. So number of genes within each category or by the EASE-score. More

Gene Similarity Search

Any given gene is associating with a set of annotation terms. If genes share

Spark

5. Interactive gene ontology analysis



DAVID Bioinformatics Resources 6.

National Institute of Allergy and Infectious Diseases (NIA

Functional Annotation Clustering

Current Gene List: List_1

Current Background: Homo sapiens

1057 DAVID IDs

⊞ **Options** **Classification Stringency** Medium

Rerun using options | Create Sublist

317 Cluster(s)

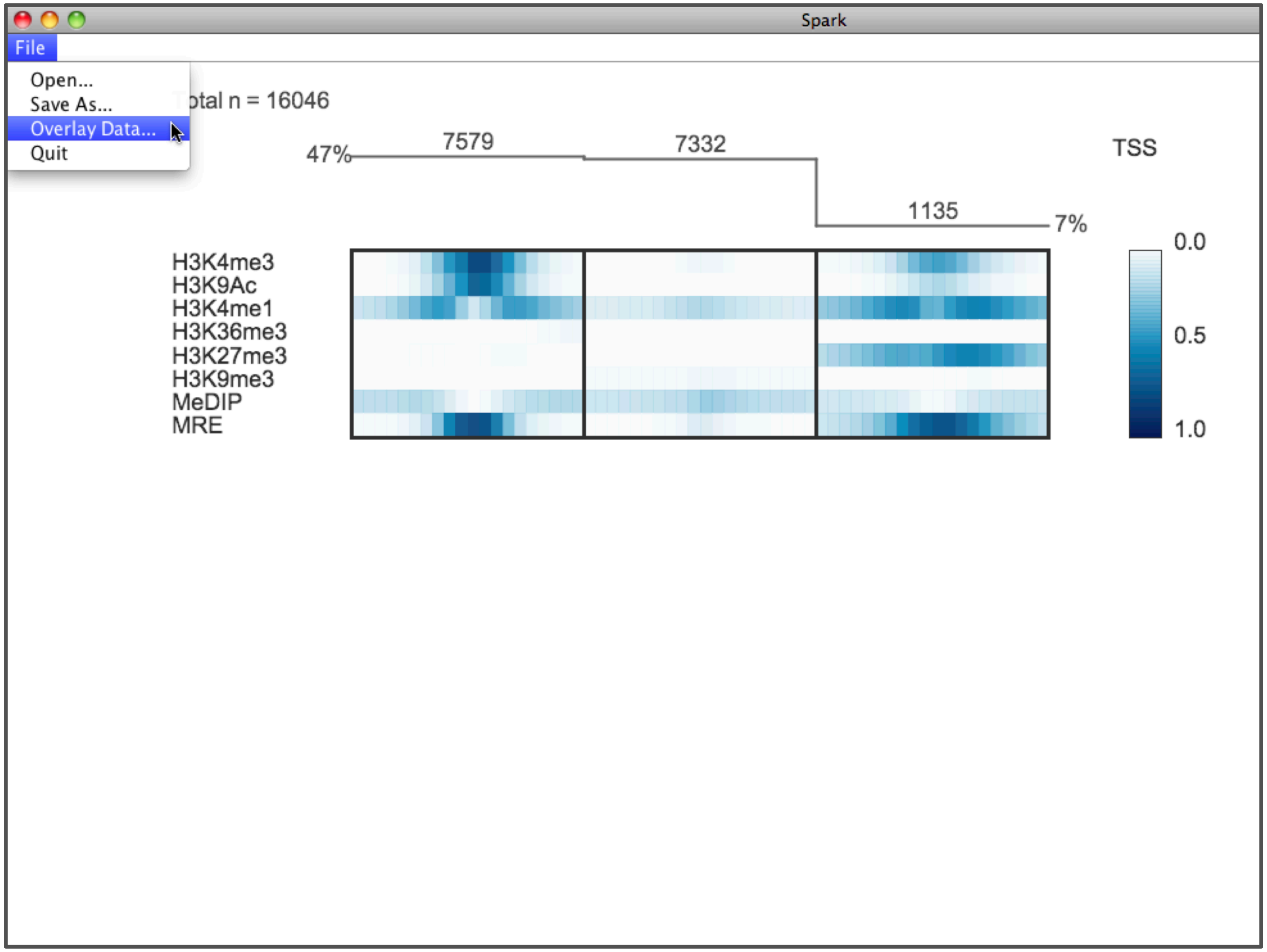
	Annotation Cluster 1	Enrichment Score: 41.28	G	
□	SP_PIR_KEYWORDS	Homeobox	RT	▬
□	UP_SEQ_FEATURE	DNA-binding region:Homeobox	RT	▬
□	INTERPRO	Homeobox, conserved site	RT	▬
□	INTERPRO	Homeobox	RT	▬
□	GOTERM_MF_FAT	sequence-specific DNA binding	RT	▬
□	GOTERM_MF_FAT	transcription factor activity	RT	▬
□	INTERPRO	Homeodomain-related	RT	▬
□	SMART	HOX	RT	▬
□	GOTERM_MF_FAT	transcription regulator activity	RT	▬
□	SP_PIR_KEYWORDS	dna-binding	RT	▬
□	GOTERM_BP_FAT	regulation of transcription, DNA-dependent	RT	▬

Spark

Do these clusters have distinct gene expression patterns?

Spark

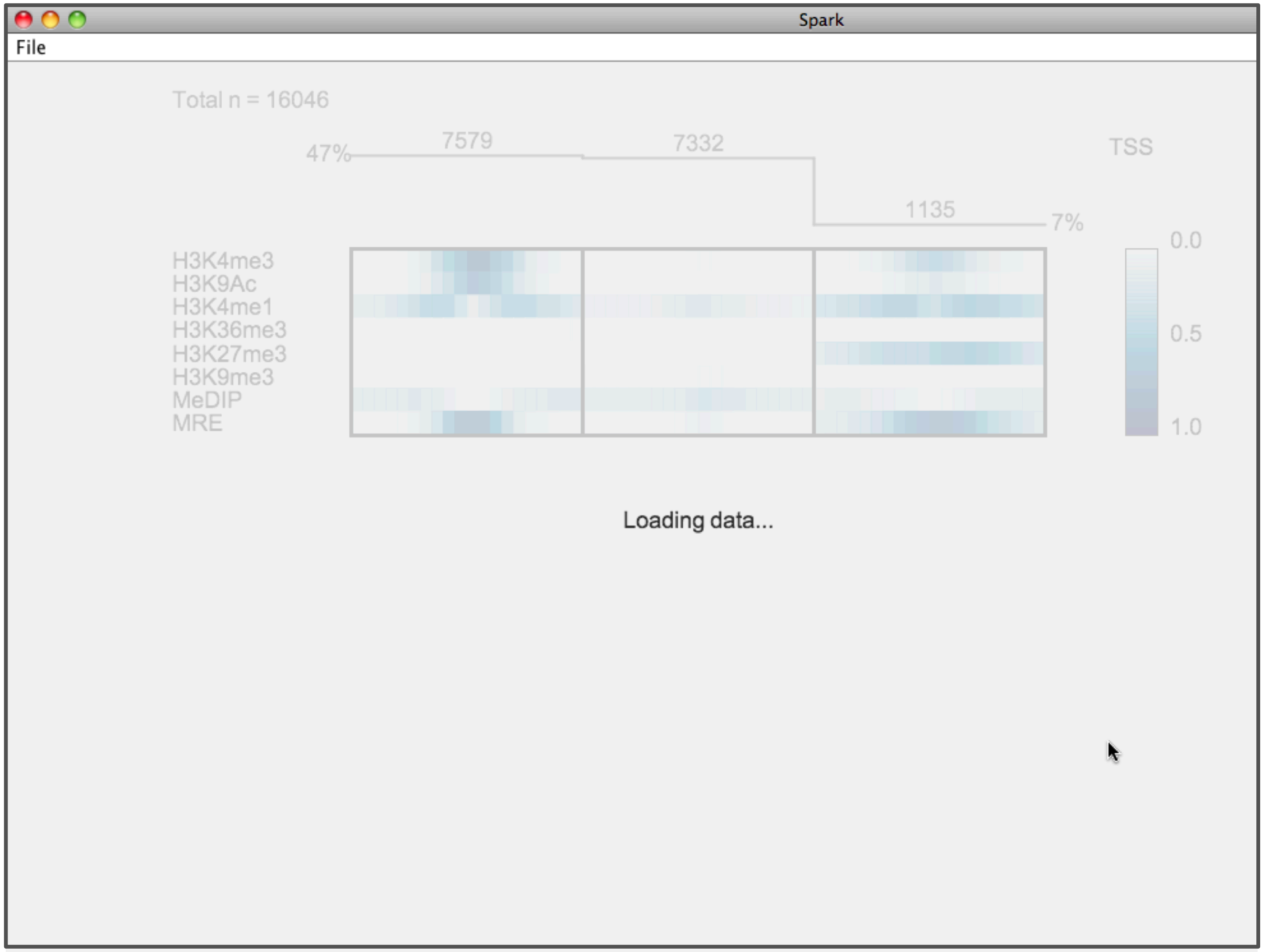
6. Interactive data overlay



screenshot

Spark

6. Interactive data overlay

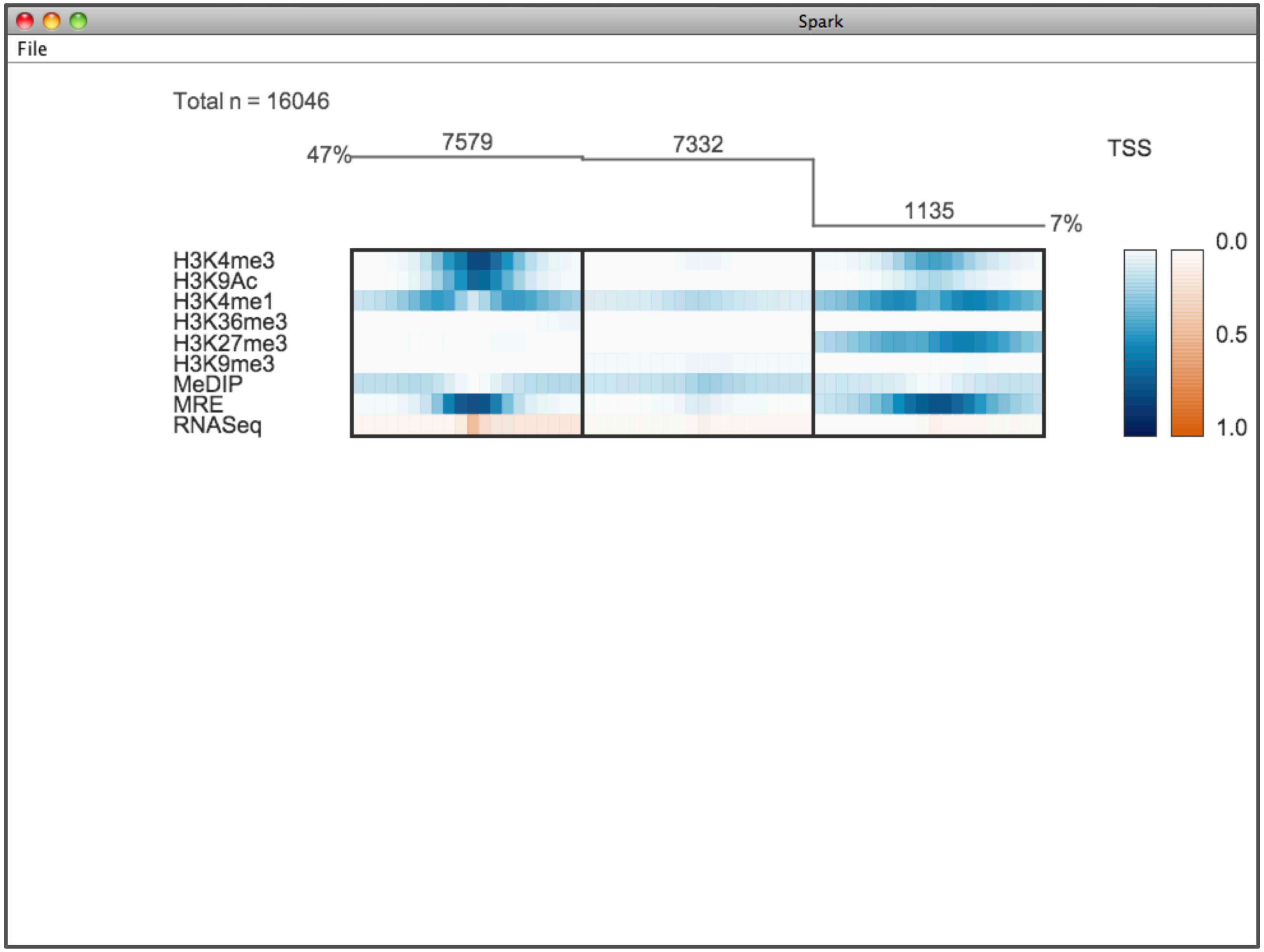


Loading data...

screenshot

Spark

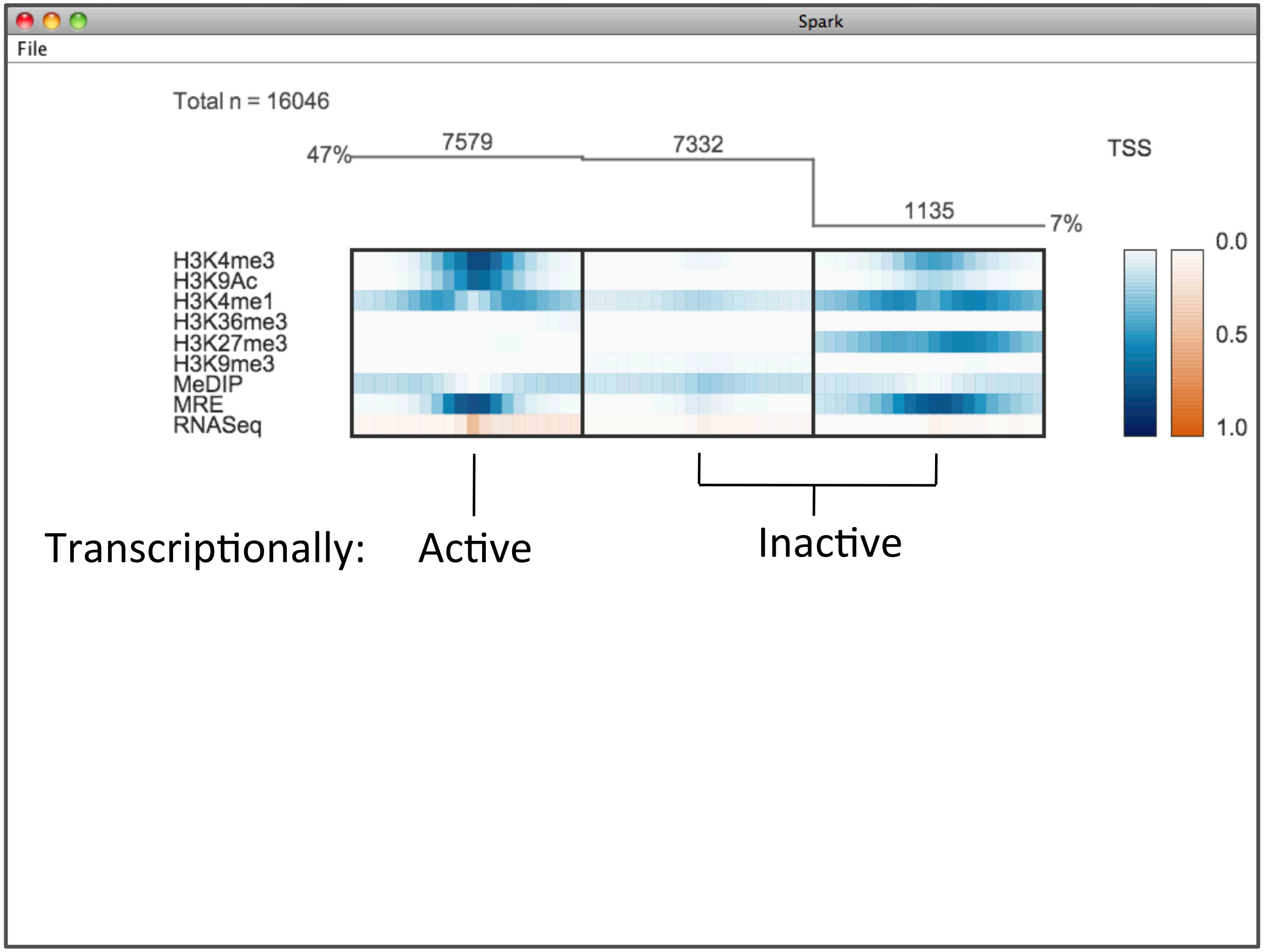
6. Interactive data overlay



screenshot

Spark

6. Interactive data overlay



Transcriptionally: Active Inactive

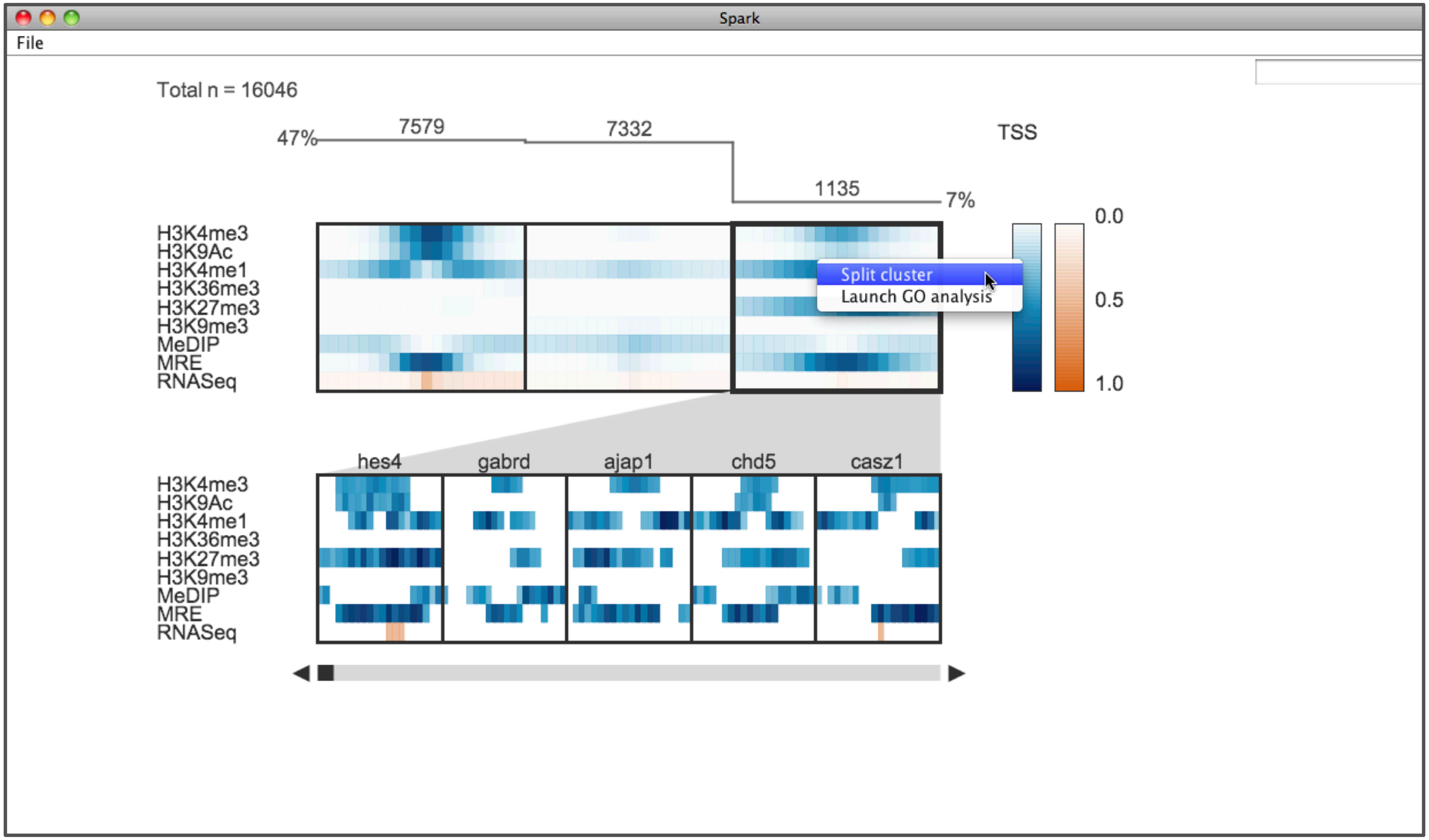
screenshot

Spark

How many clusters should I generate?

Spark

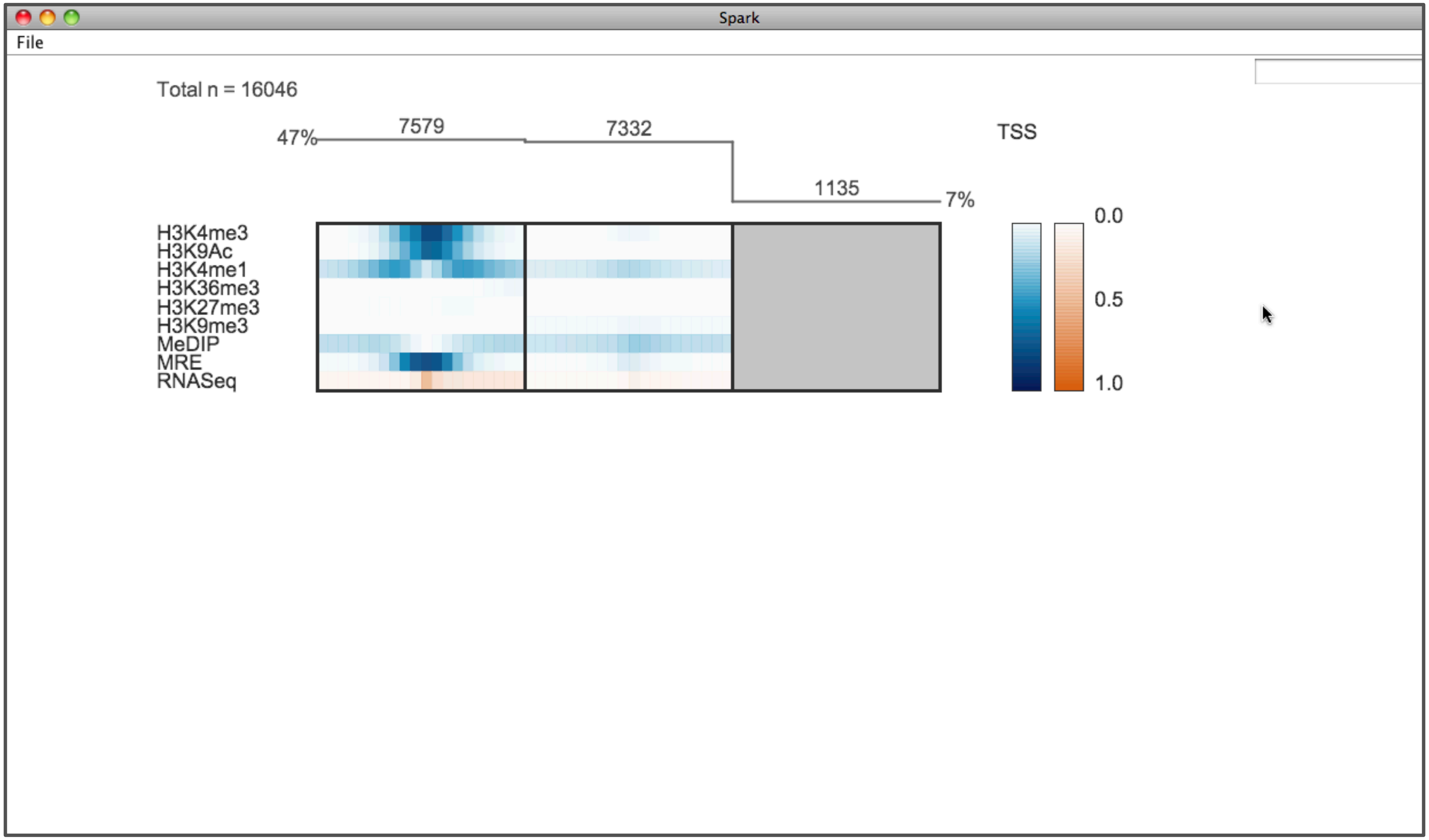
7. Interactive cluster splitting



screenshot

Spark

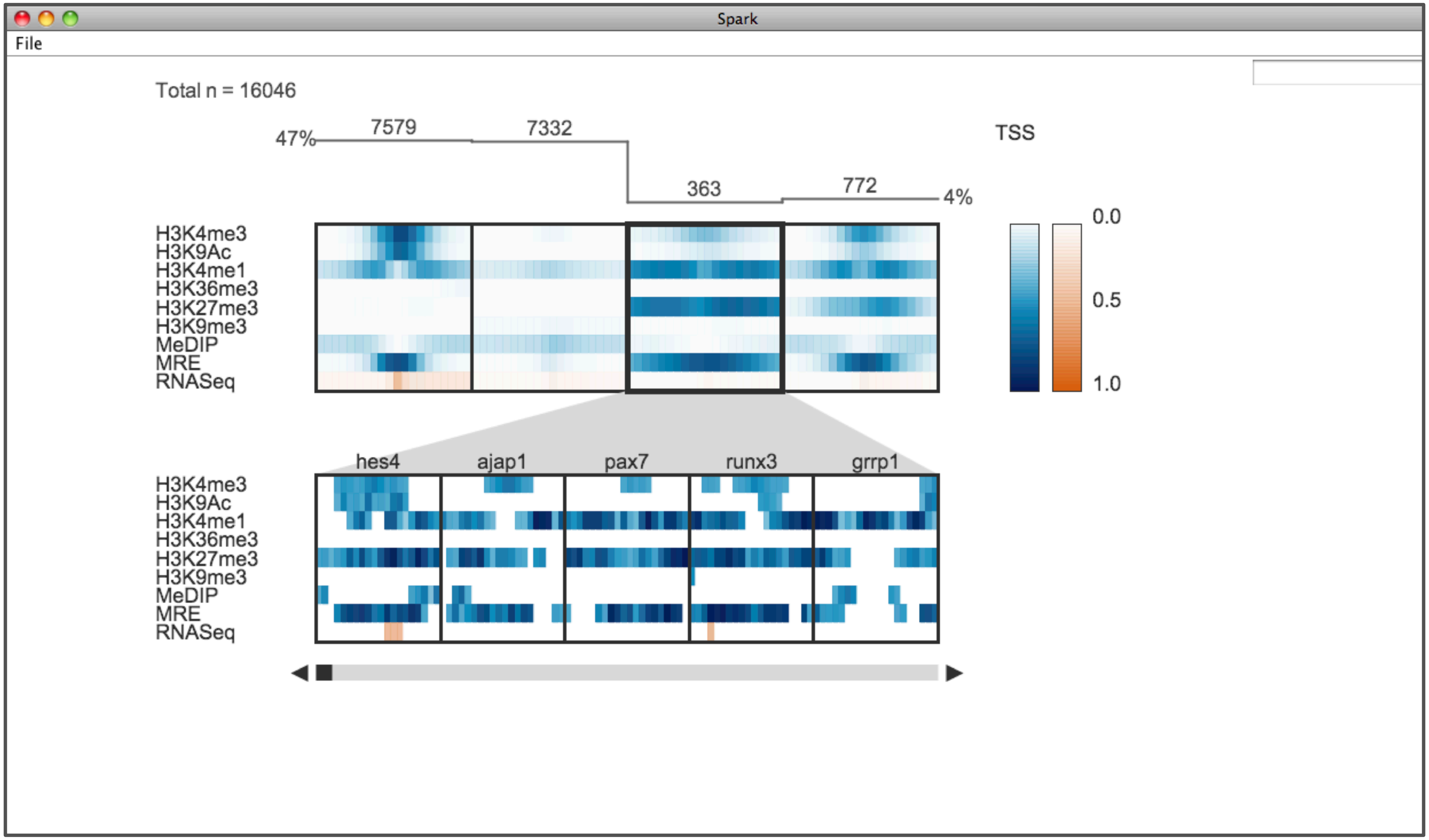
7. Interactive cluster splitting



screenshot

Spark

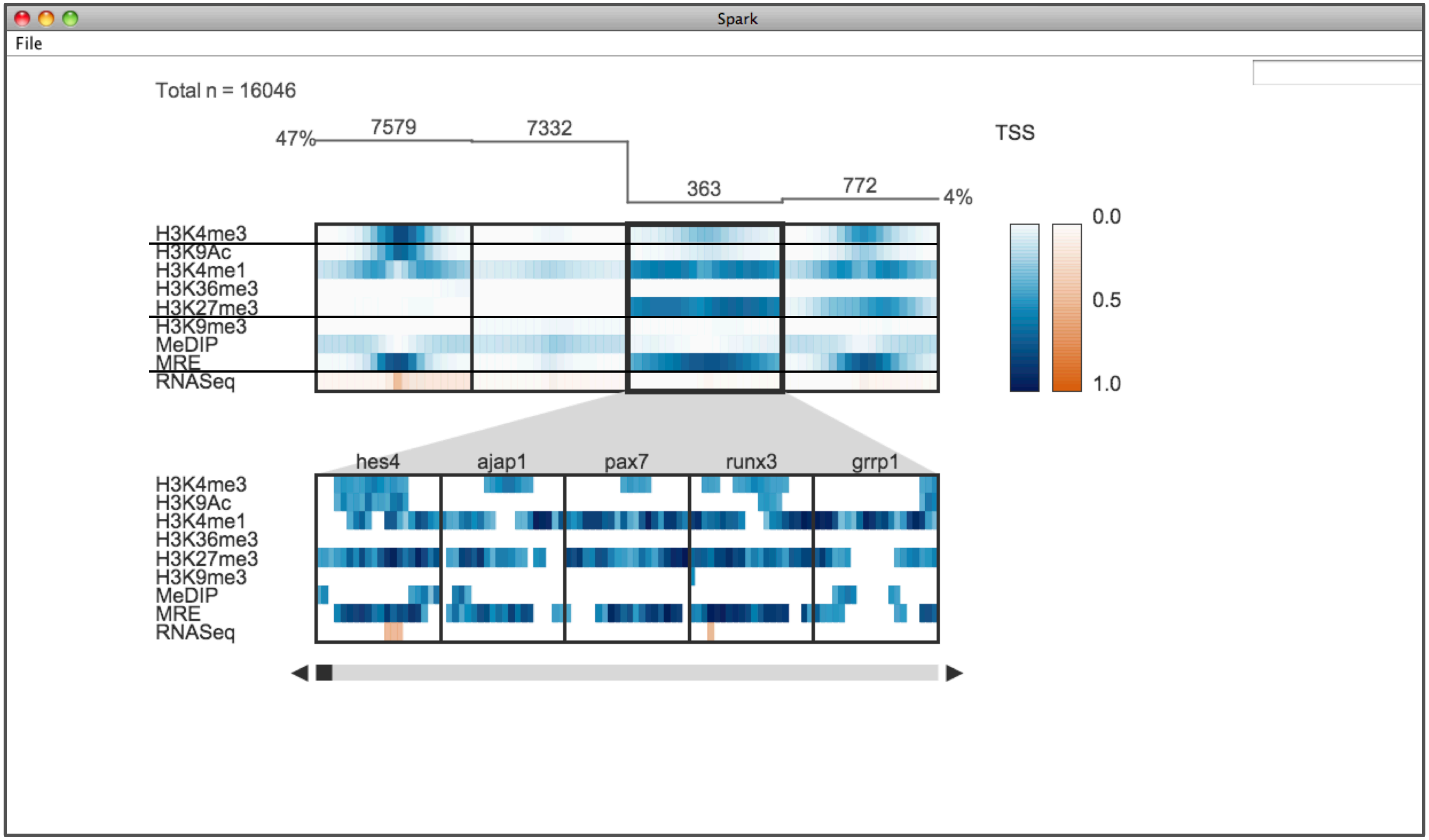
7. Interactive cluster splitting



screenshot

Spark

7. Interactive cluster splitting



screenshot

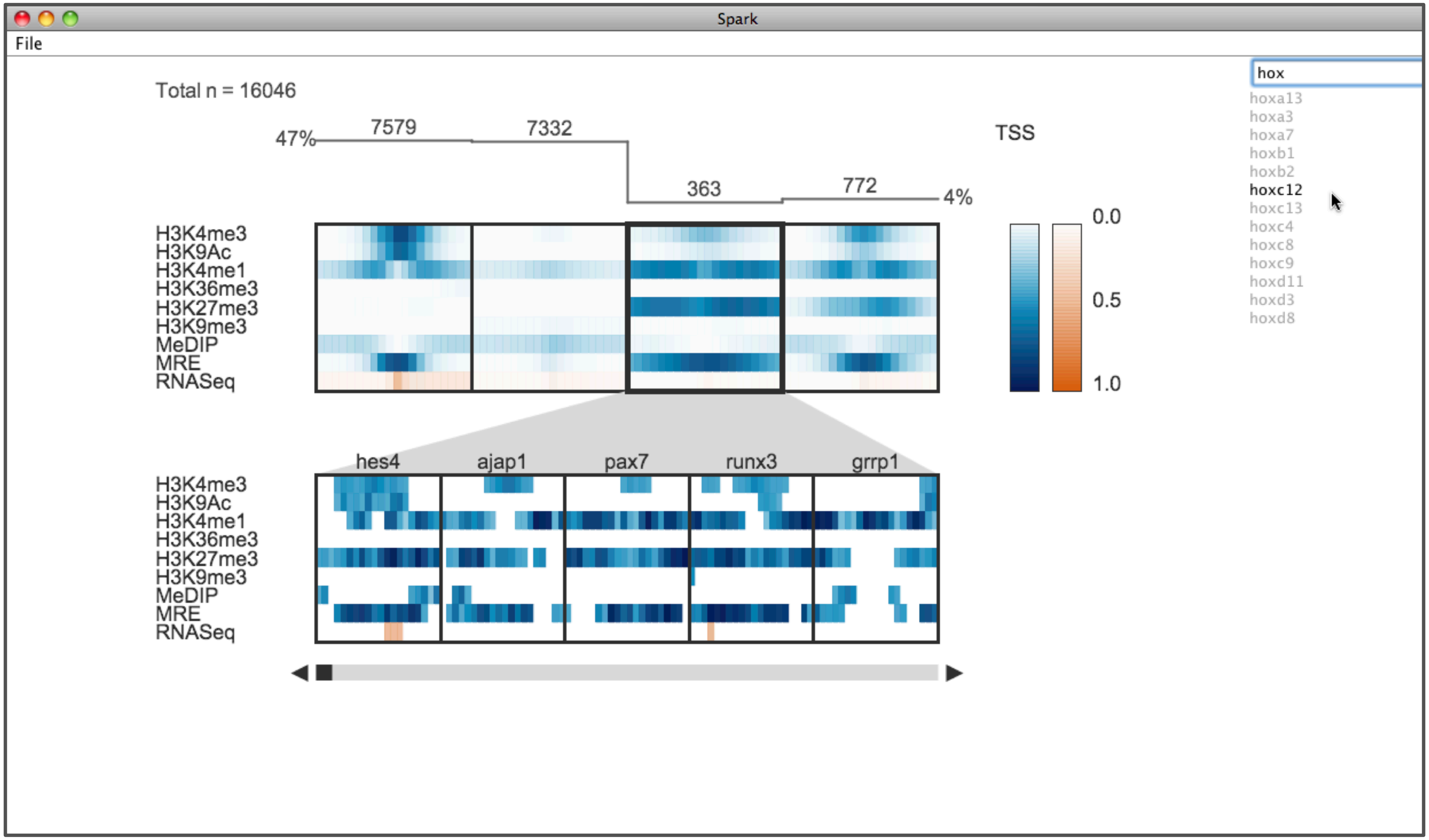
Spark

Where is my favorite gene?

What other genes have similar data patterns?

Spark

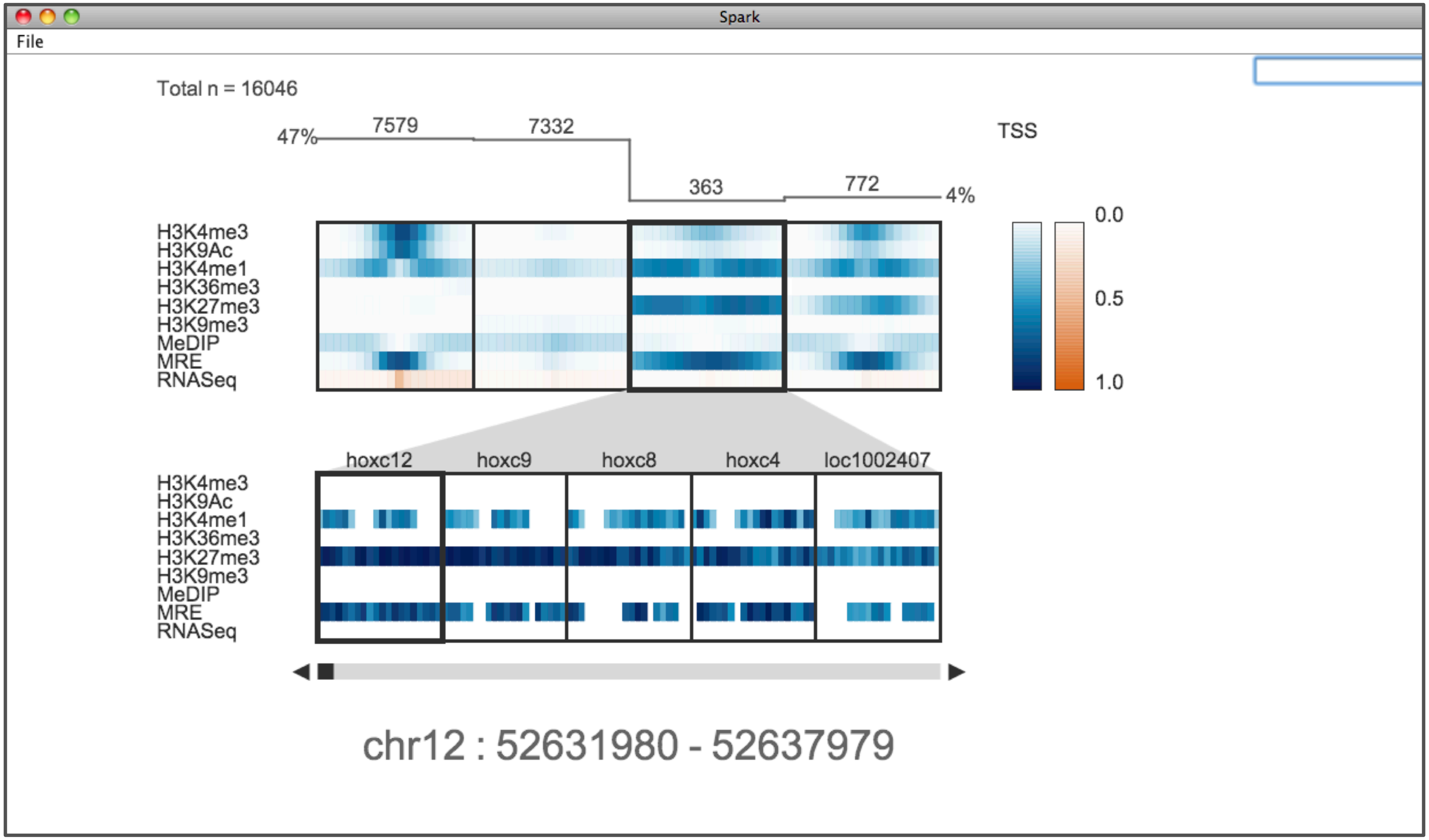
8. Finding genes of interest



screenshot

Spark

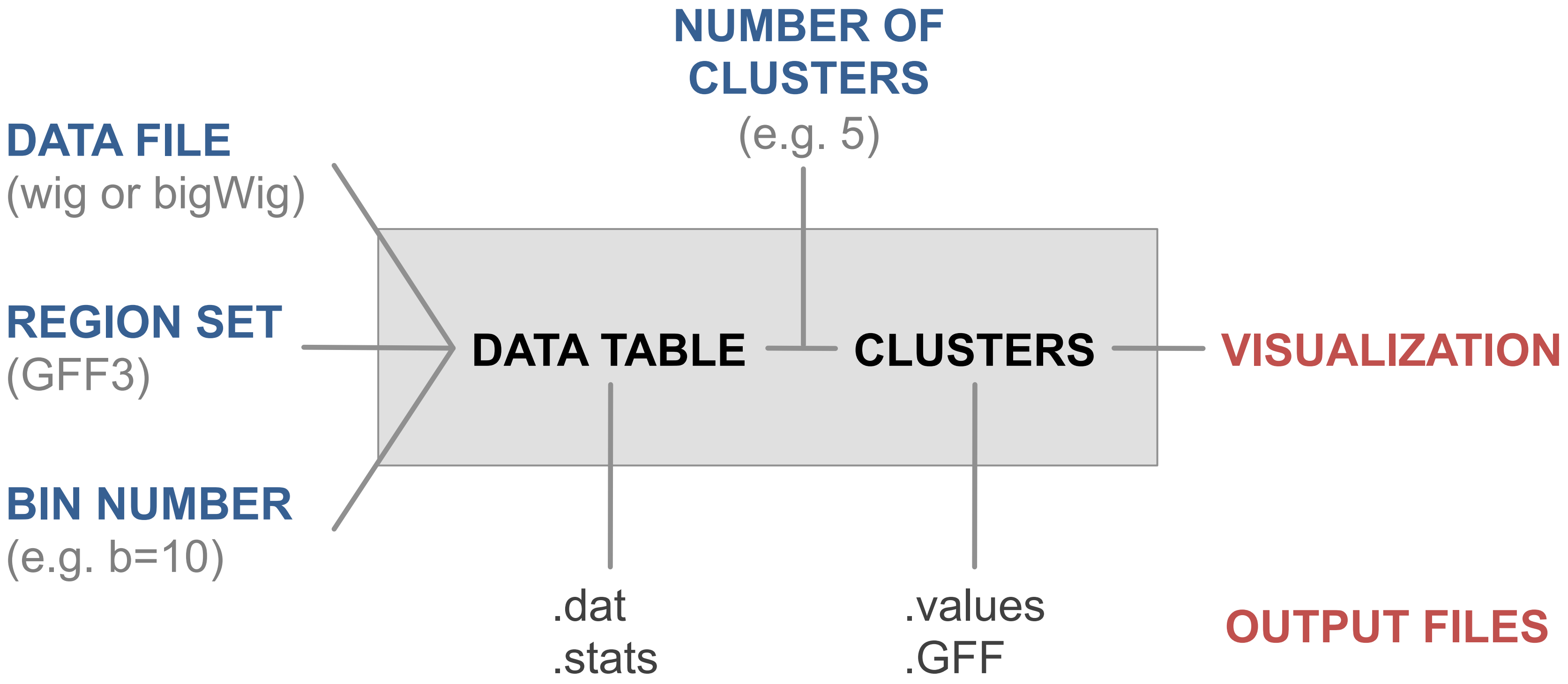
8. Finding genes of interest



screenshot

Spark - technical details

Spark - technical details



NUMBER OF CLUSTERS

(e.g. 5)

DATA FILE

(wig or bigWig)

REGION SET

(GFF3)

BIN NUMBER

(e.g. b=10)

DATA TABLE

CLUSTERS

VISUALIZATION

.dat

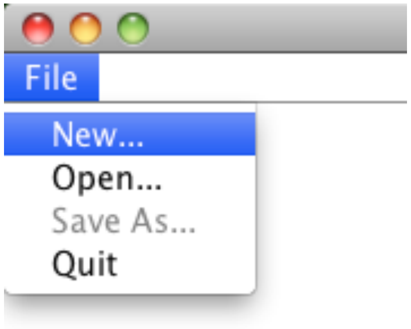
.stats

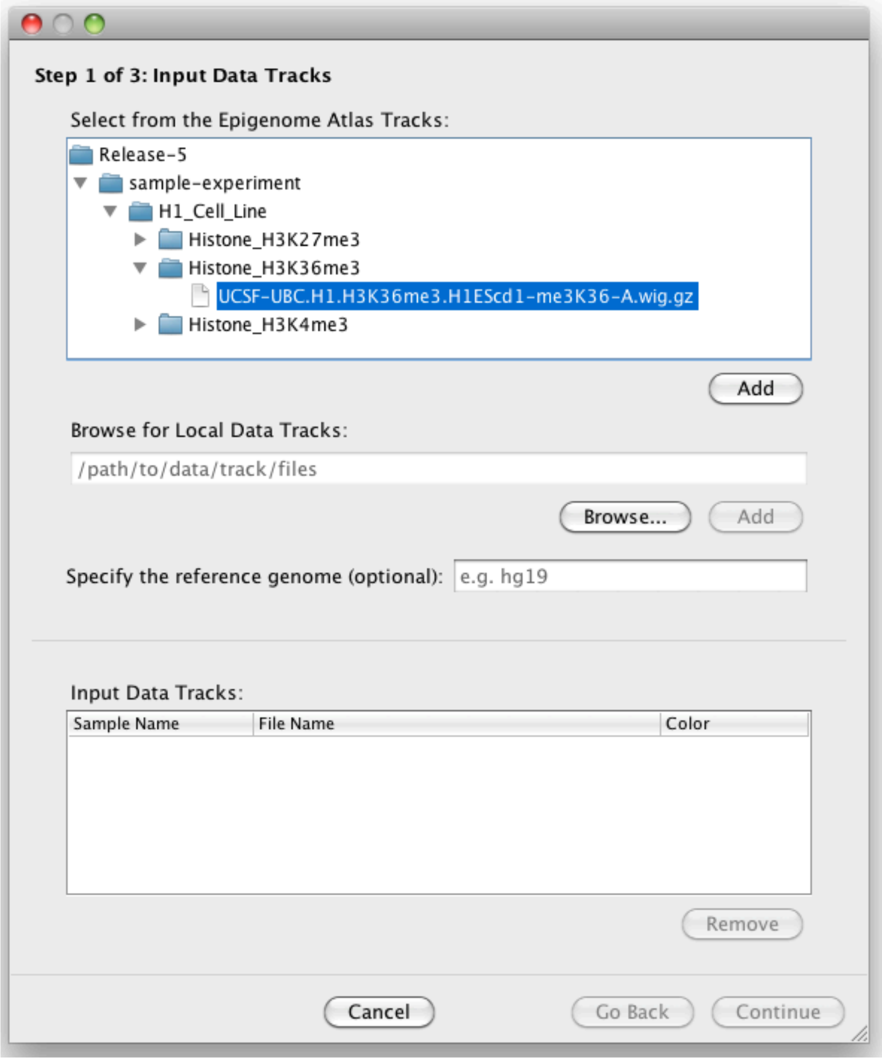
.values

.GFF

OUTPUT FILES

Spark - technical details





Step 1 of 3: Input Data Tracks

Select from the Epigenome Atlas Tracks:

```
Release-5
▼ sample-experiment
  ▼ H1_Cell_Line
    ▶ Histone_H3K27me3
    ▼ Histone_H3K36me3
        UCSF-UBC.H1.H3K36me3.H1EScd1-me3K36-A.wig.gz
    ▶ Histone_H3K4me3
```

Browse for Local Data Tracks: /path/to/data/track/files

Specify the reference genome (optional): e.g. hg19

Input Data Tracks:

Sample Name	File Name	Color

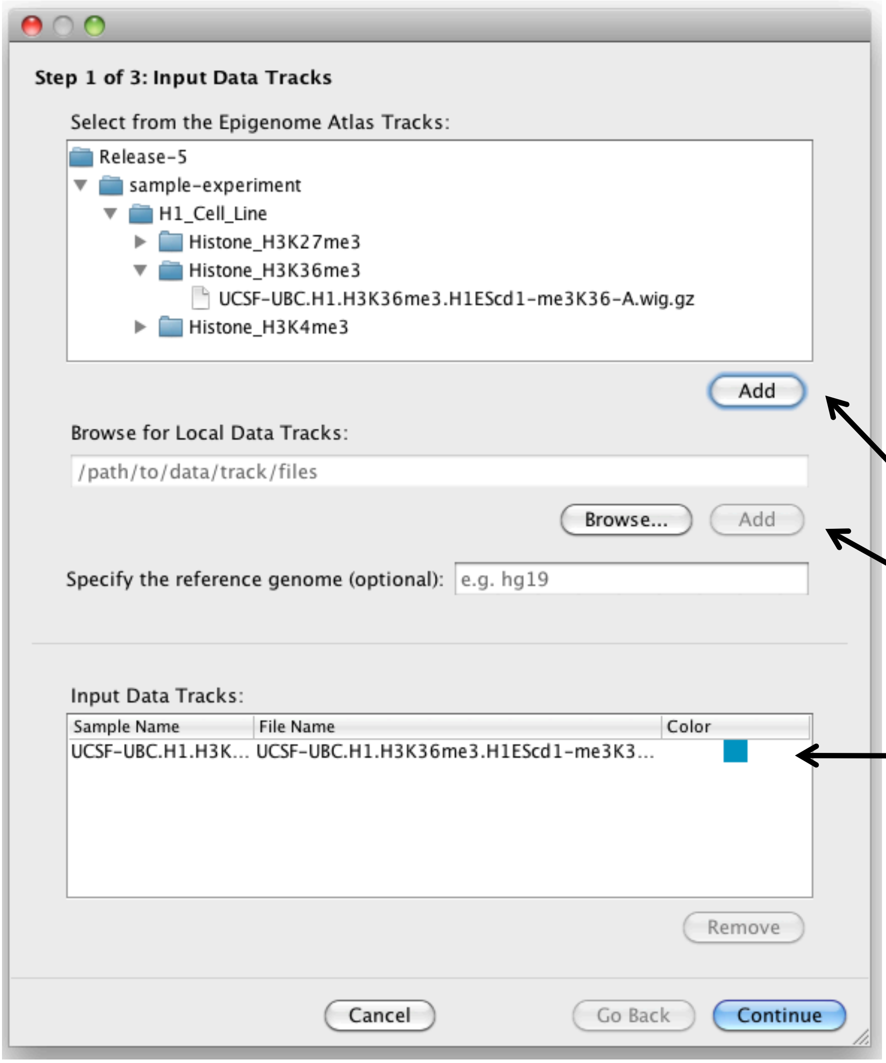
Step 1

← Select from the Epigenome Atlas data tracks

← Or browse for local wig files or paste a URL

This Epigenome Atlas Tracks tree above is populated from an XML file containing the directory structure from www.genboree.org/epigenomeatlas and therefore can very easily be updated to reflect new data releases

Spark - technical details



Step 1 of 3: Input Data Tracks

Select from the Epigenome Atlas Tracks:

```
Release-5
▼ sample-experiment
  ▼ H1_Cell_Line
    ▶ Histone_H3K27me3
    ▼ Histone_H3K36me3
        UCSF-UBC.H1.H3K36me3.H1EScd1-me3K36-A.wig.gz
    ▶ Histone_H3K4me3
```

Browse for Local Data Tracks: /path/to/data/track/files

Specify the reference genome (optional): e.g. hg19

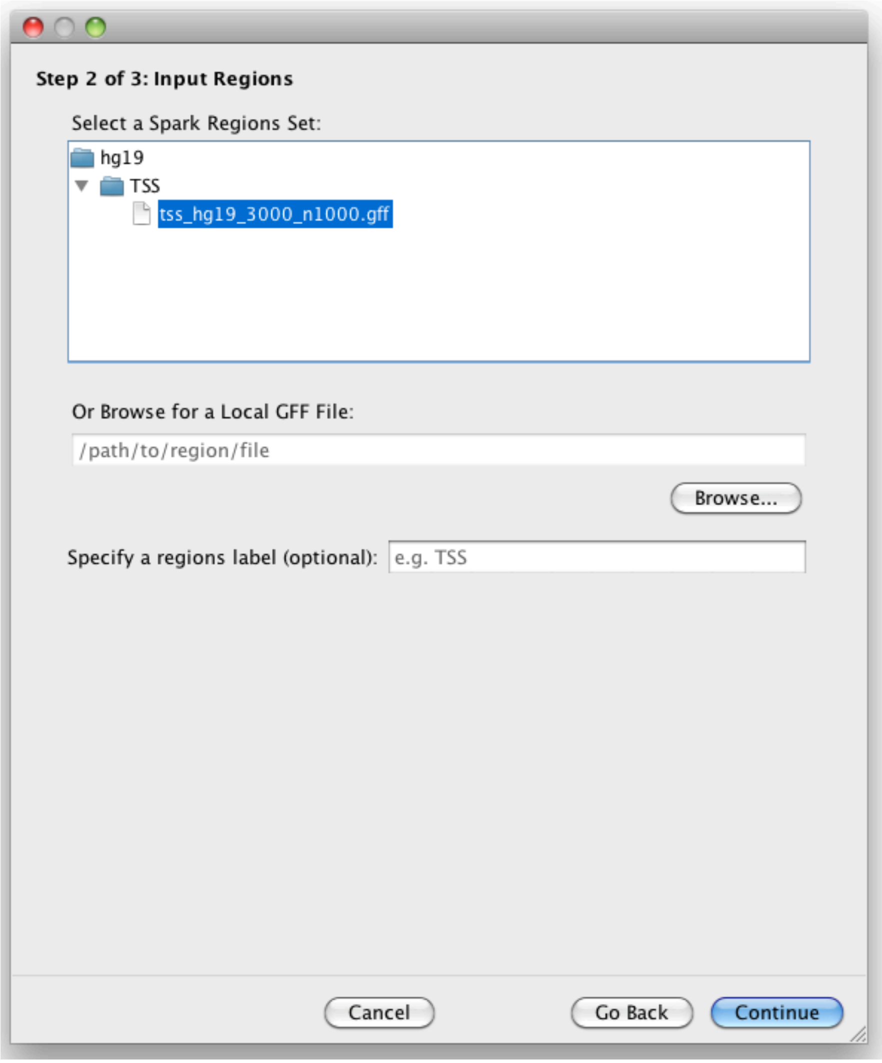
Input Data Tracks:

Sample Name	File Name	Color
UCSF-UBC.H1.H3K...	UCSF-UBC.H1.H3K36me3.H1EScd1-me3K3...	🟦

Step 1

Your selections can then be added to the input data table below, in which you can edit the sample name and display colour.

Spark - technical details



Step 2 of 3: Input Regions

Select a Spark Regions Set:

Or Browse for a Local GFF File: /path/to/region/file

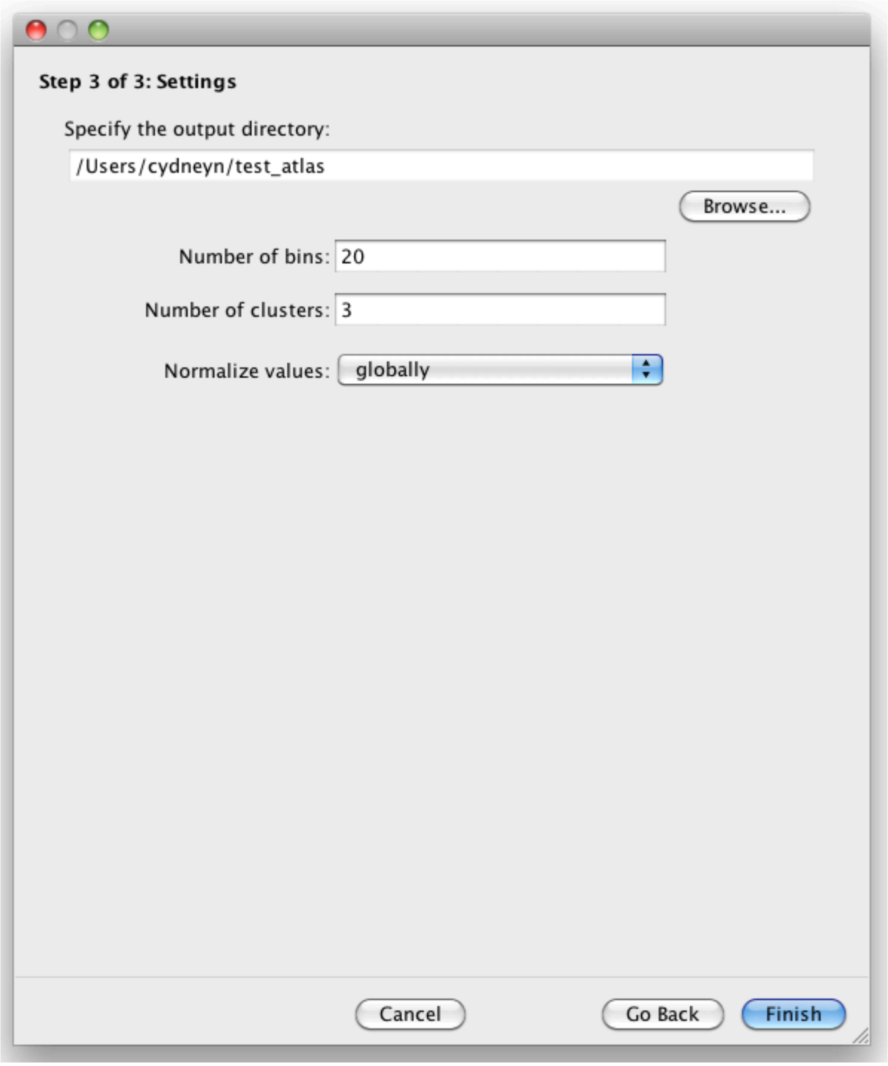
Specify a regions label (optional): e.g. TSS

Step 2

← Select from a provided set of region files

← Or browse for local GFF file

Spark - technical details



Step 3 of 3: Settings

Specify the output directory: /Users/cydneyn/test_atlas

Number of bins: 20

Number of clusters: 3

Normalize values: globally

Step 3

Specify where to save your analysis

Adjust clustering parameters or use provided defaults

In cases where the user selects an Epigenome Atlas data track and one of the provided region sets, Spark makes use of pre-computed data files that greatly increase performance.

If you want to use a custom region set with an Epigenome Atlas data track, Spark handles all of the data file downloading and caching to keep things simple.

You're done!
Spark will generate and display your clustering

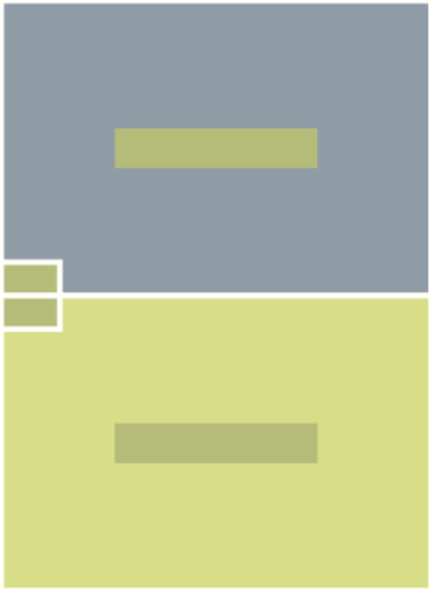
Demo

Future work

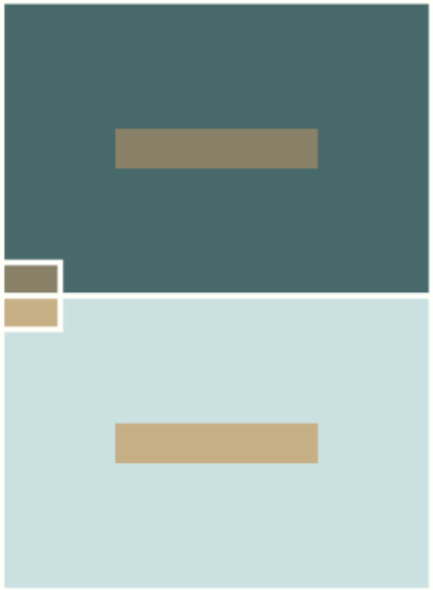
- Heatmaps are not the best representation

Future work

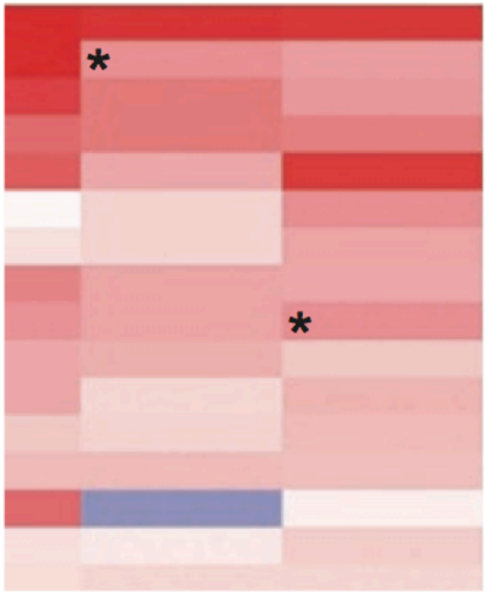
- Heatmaps are not the best representation



Same colour looks different



Different colour looks the same

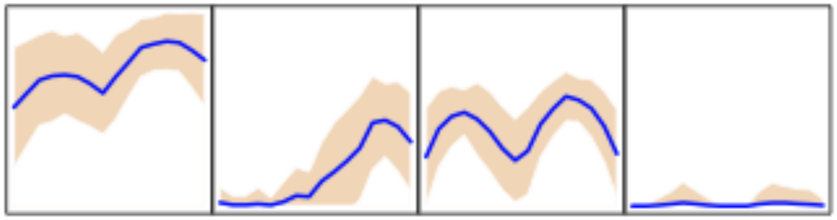


* These rectangles have the same colour but look different

Bang Wong, Nature Methods, 2010

Future work

- Heatmaps are not the best representation
  - Histograms would be better (also capture variation)



Future work

- Heatmaps are not the best representation
  - Histograms would be better (also capture variation)
- Greater functionality for manipulating clusters
  - Merge clusters
  - Edit current settings

Future work

- Heatmaps are not the best representation
  - Histograms would be better (also capture variation)
- Greater functionality for manipulating clusters
  - Merge clusters
  - Edit current settings
- Input files likely require filtering
  - could filter data based on an input peak set

Future work

- Heatmaps are not the best representation
  - Histograms would be better (also capture variation)
- Greater functionality for manipulating clusters
  - Merge clusters
  - Edit current settings
- Input files likely require filtering
  - could filter data based on an input peak set
- More directly support data driven region selection

Future work

- Heatmaps are not the best representation
  - Histograms would be better (also capture variation)
- Greater functionality for manipulating clusters
  - Merge clusters
  - Edit current settings
- Input files likely require filtering
  - could filter data based on an input peak set
- More directly support data driven region selection
- Spark may not perform well in detected a small subpopulation with a particular pattern

Spark releases

Public release (old)
http://www.bcgsc.ca/platform/bioinfo/software/spark

Current version (significantly updated; in testing)
http://www.bcgsc.ca/downloads/spark/v1.1.0/start.jnlp

Acknowledgements

BCGSC
Steve Jones (advisor)
Marco Marra
Martin Hirst
Misha Bilenky
Gordon Robertson
Yongjun Zhao
Martin Krzywinski

Sequencing Team
Bioinformatics Group
LIMS

UCSF
Joe Costello
Chibo Hong
Ravi Nagarajan
Thea Tlsty
Philippe Gascard
Mahvash Sigaroudinia
Art Weiss
Terri Kaldecek
Michael McManus
Hunter Richards
Yun Choi
Susan Fisher
Olga Genbacev

UC Davis
Peggy Farnham
Henny O'Geen
Lorigail Echipare
Vitalina Komashko
Kimberly Blahnik

UCSC
David Haussler
Tracy Ballinger

Washington Univ.
Ting Wang

CIHR IRSC
Canadian Institutes of Health Research
Instituts de recherche en santé du Canada

Michael Smith Foundation for Health Research

National Institutes of Health

ROADMAP epigenomics PROJECT