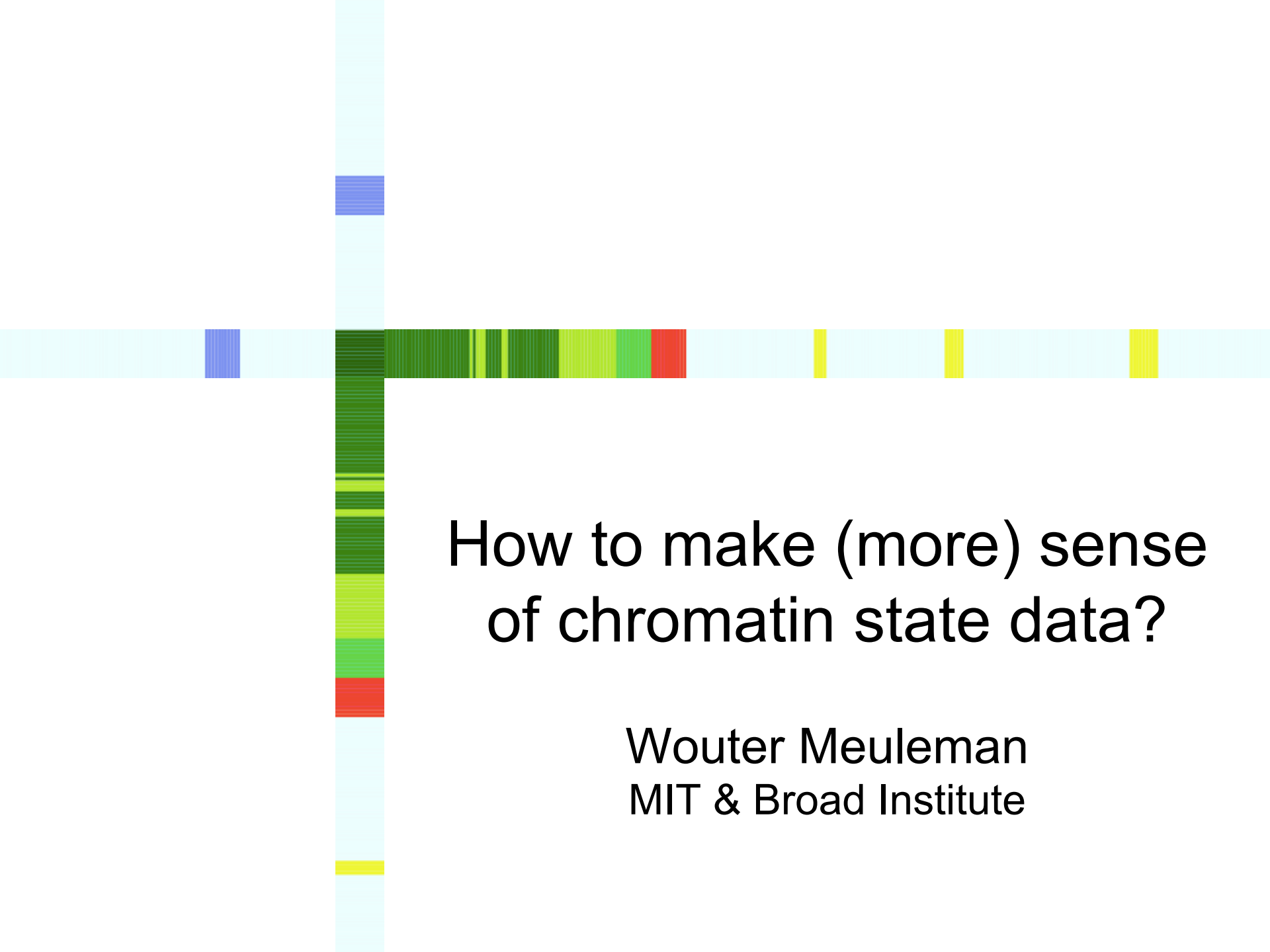


How to make (more) sense of chromatin state data?

Wouter Meuleman
MIT & Broad Institute

ChromHMM

NATURE METHODS | CORRESPONDENCE

ChromHMM: automating chromatin-state discovery and characterization

Jason Ernst & Manolis Kellis

Affiliations | **Corresponding author**

Nature Methods **9**, 215–216 (2012) | doi:10.1038/nmeth.1906
Published online 28 February 2012

Segway

NATURE METHODS | BRIEF COMMUNICATION

Unsupervised pattern discovery in human chromatin structure through genomic segmentation

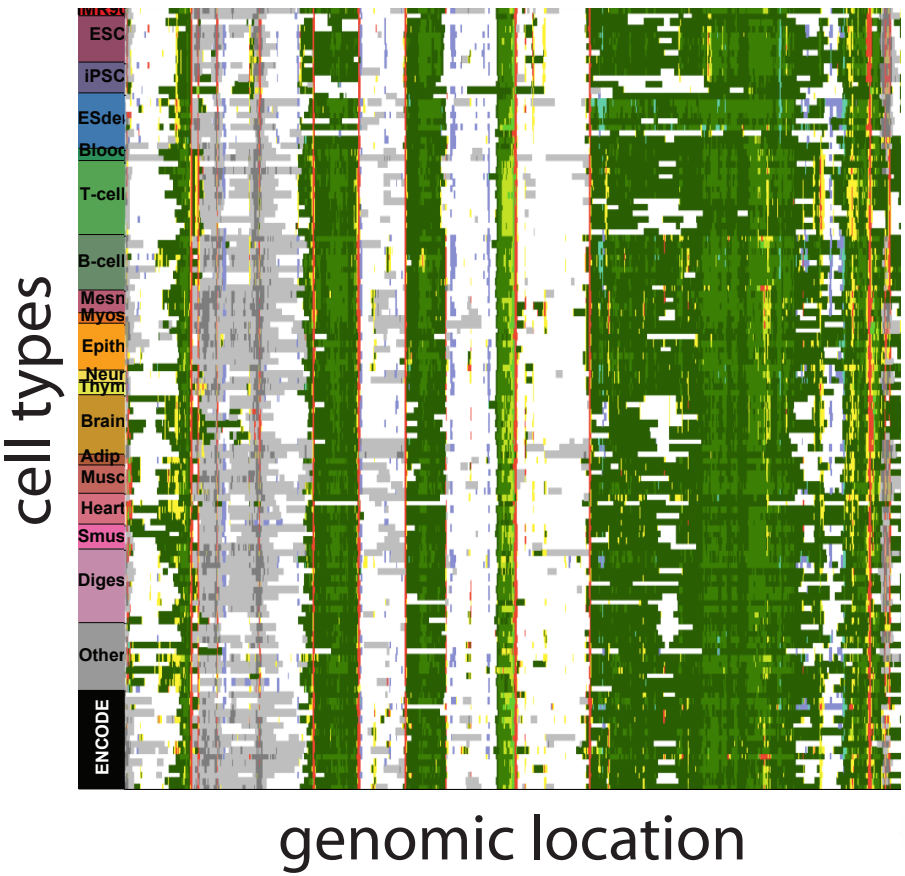
Michael M Hoffman, Orion J Buske, Jie Wang, Zhiping Weng, Jeff A Bilmes & William Stafford Noble

Affiliations | **Contributions** | **Corresponding author**

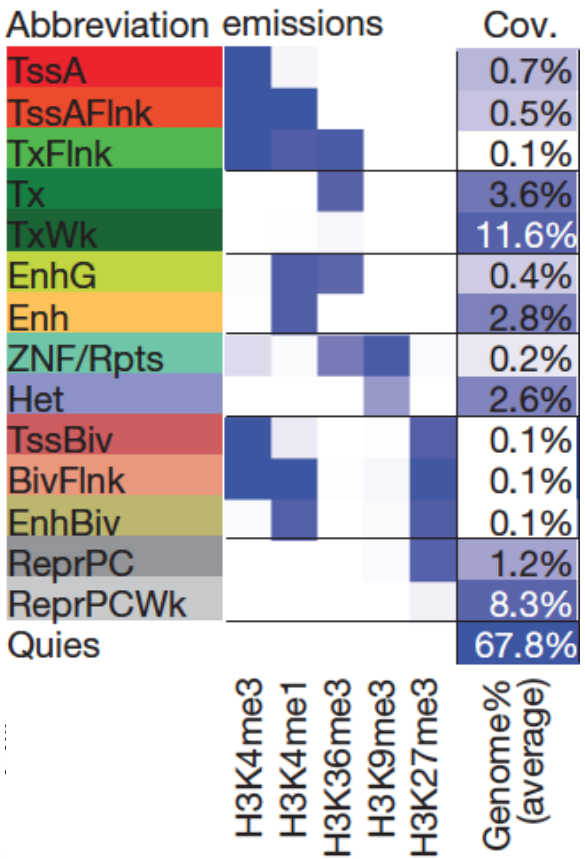
Nature Methods **9**, 473–476 (2012) | doi:10.1038/nmeth.1937
Received 01 July 2011 | Accepted 14 February 2012 | Published online 18 March 2012

These methods allow us to summarize our “raw” data:

A color here…

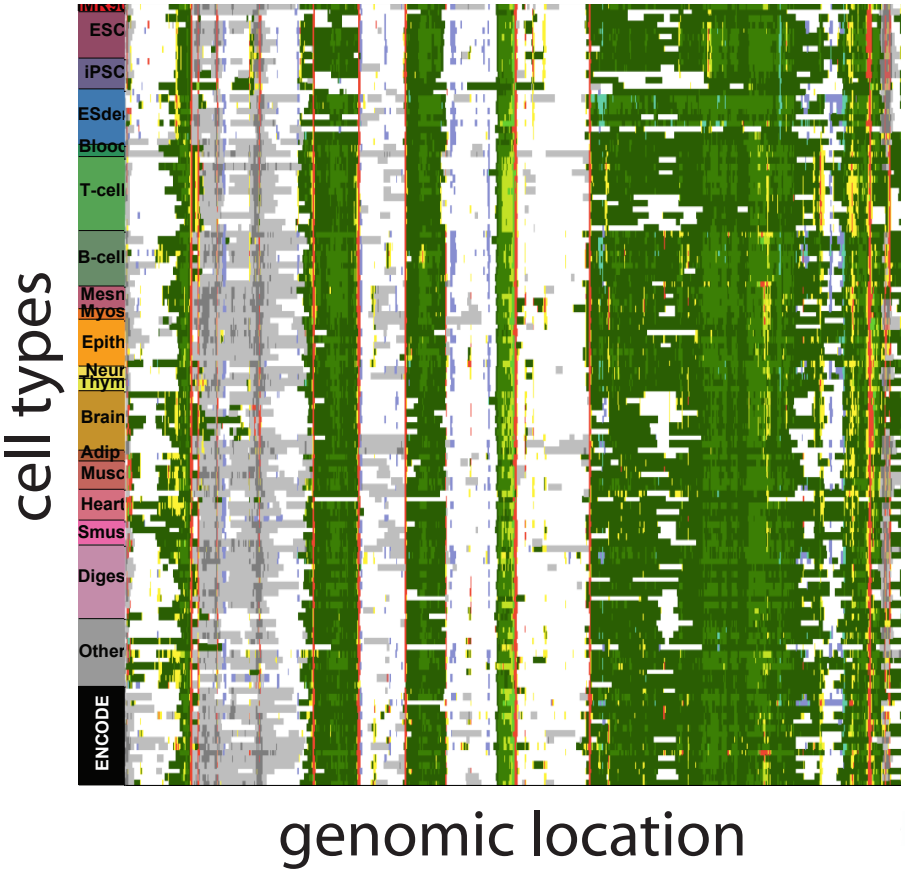


…corresponds to a state here:



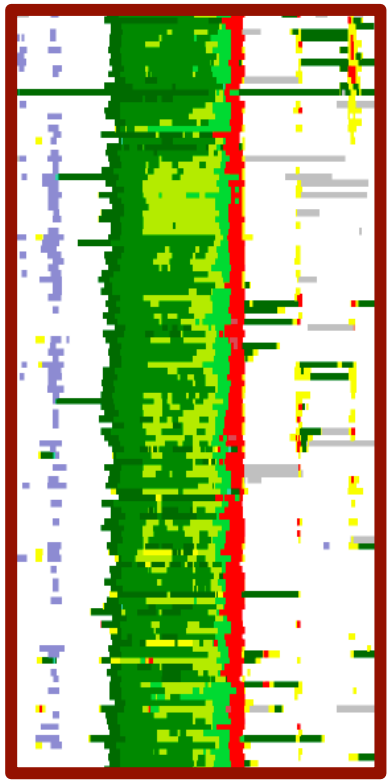
Going forward…

Many more samples
are being profiled
(cell types, disease vs. control,
personal epigenomics, etc)



The genome is large
(shown here is only
0.0267% of the genome)

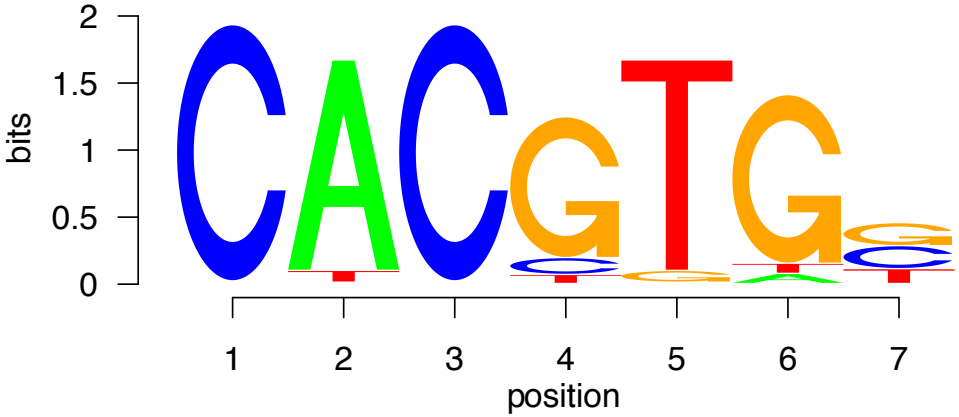
Chromatin state calls across many samples can be viewed as an alignment of sequences with a finite alphabet



~

```
CTCTTAT
CACGTGC
CACGTGC
CACGTGG
CACGTGG
CACGTGG
CACGTGC
CACGTGG
CACGTGT
CACGTGC
CACGTGT
CACGTGC
CACCTGT
CACCGTC
CACGTGC
CACGTGC
CACGTGG
CACGTGT
CACGTGG
CACGTGG
CACGTGG
```

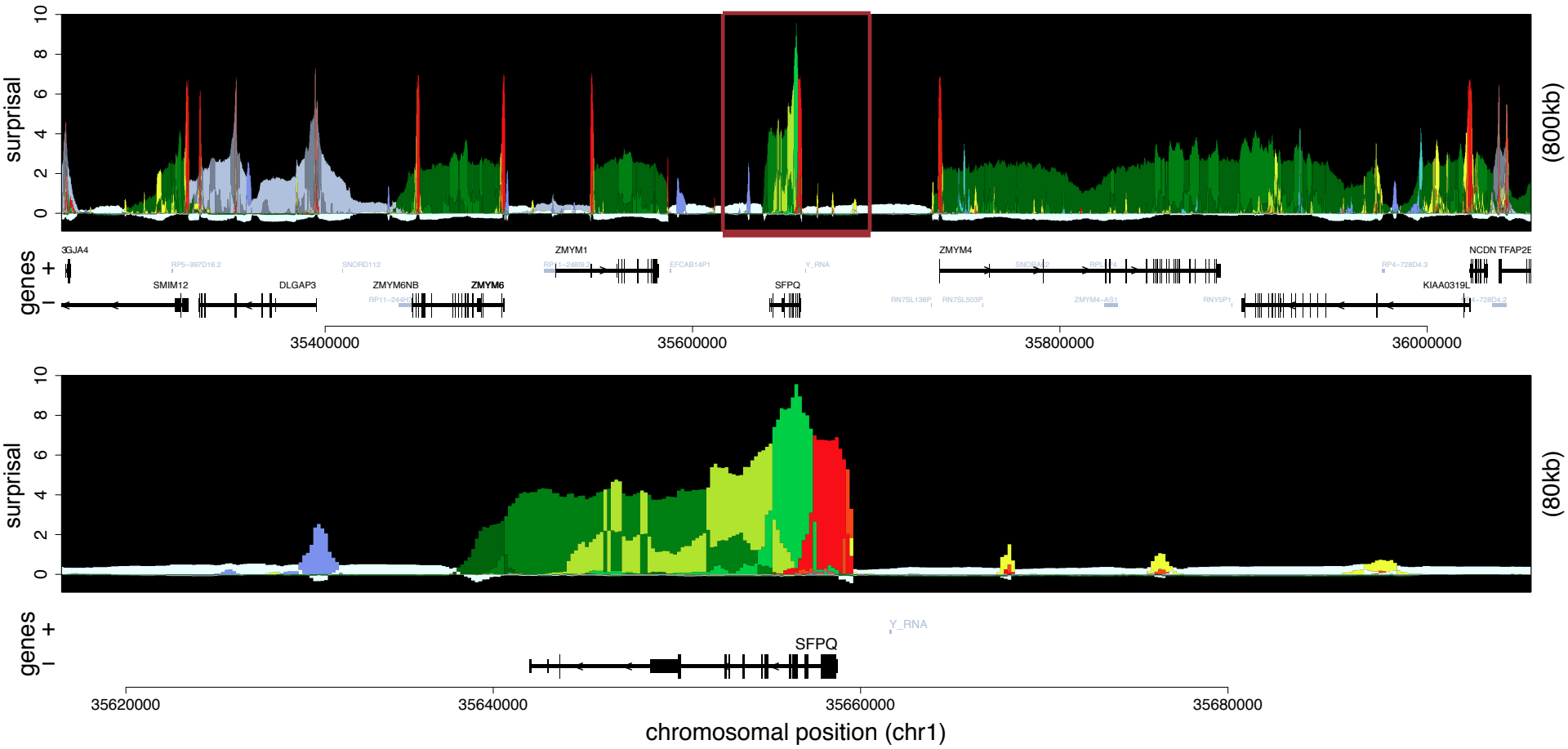
}



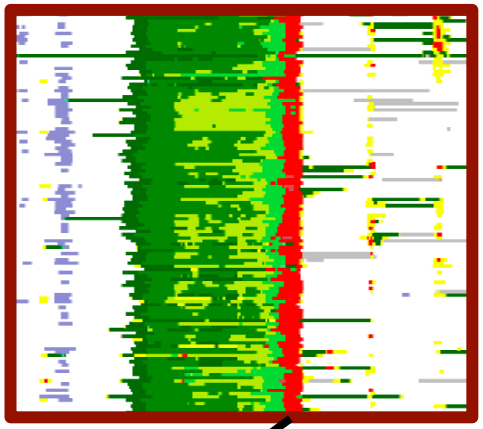
There are good ways of modeling such alignments: logos!

Information content of a region, considering background

epilogos

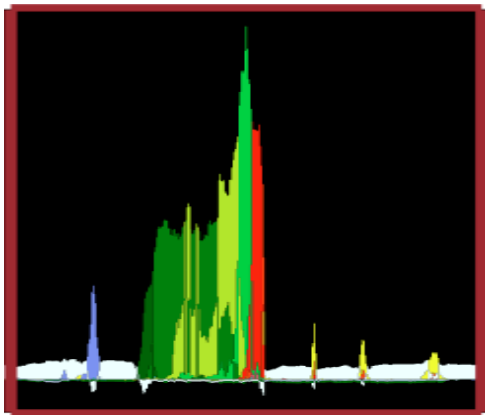

surprisal
0
2
4
6
8
10
(800kb)
genes
+
−
3GJA4
SMIM12
DLGAP3
ZMYM6NB
ZMYM6
ZMYM1
SFPQ
ZMYM4
NCDN
TFAP2E
KIAA0319L
35400000
35600000
35800000
36000000
surprisal
0
2
4
6
8
10
(80kb)
genes
+
−
Y_RNA
SFPQ
35620000
35640000
35660000
35680000
chromosomal position (chr1)

tinyurl.com/epilogos-feedback



Kullback-Leibler divergence
(relative entropy)

$$D_{KL}(P||Q) = \sum_i p(i) \log_2 \left(\frac{p(i)}{q(i)} \right)$$



P = observed chromatin state frequencies at any single genomic location, e.g.:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
100	0	0	0	0	0	0	0	0	0	0	0	0	0	0

local chromatin state frequency (%)

Q = expected frequencies based on genome-wide state occurrences:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0.7	0.5	0.1	3.6	11.6	0.4	2.8	0.2	2.6	0.1	0.1	0.1	1.2	8.3	67.8

epigenome–wide chromatin state frequency (%)

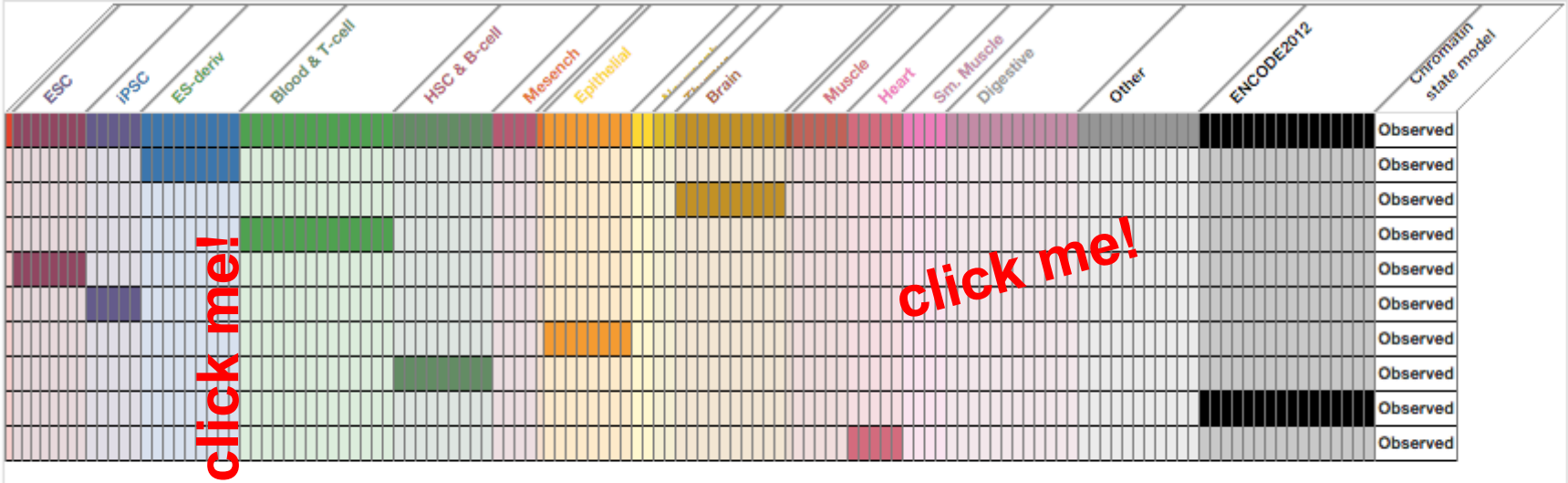
http://epilogos.broadinstitute.org

epilogos Index About

epilogos

visualization and analysis of chromatin state model data

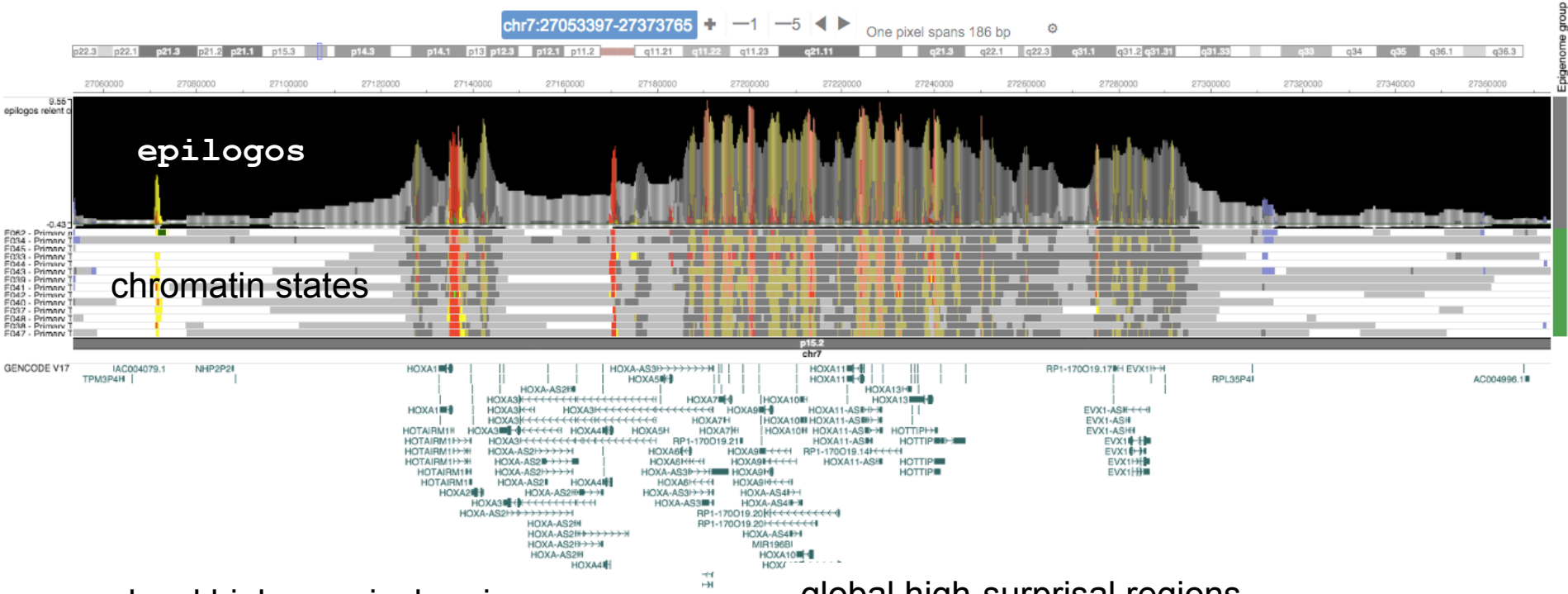
Fastest option: choose some popular pre-computed example epilogos:
(Each row depicts a single epilogos computation, with bright colors indicating the selected epigenomes)



epilogos Index About

your epilogos computation

If you do not see the result you were expecting, please try reloading the page. If that does not resolve the issue please contact us.



local high-surprisal regions

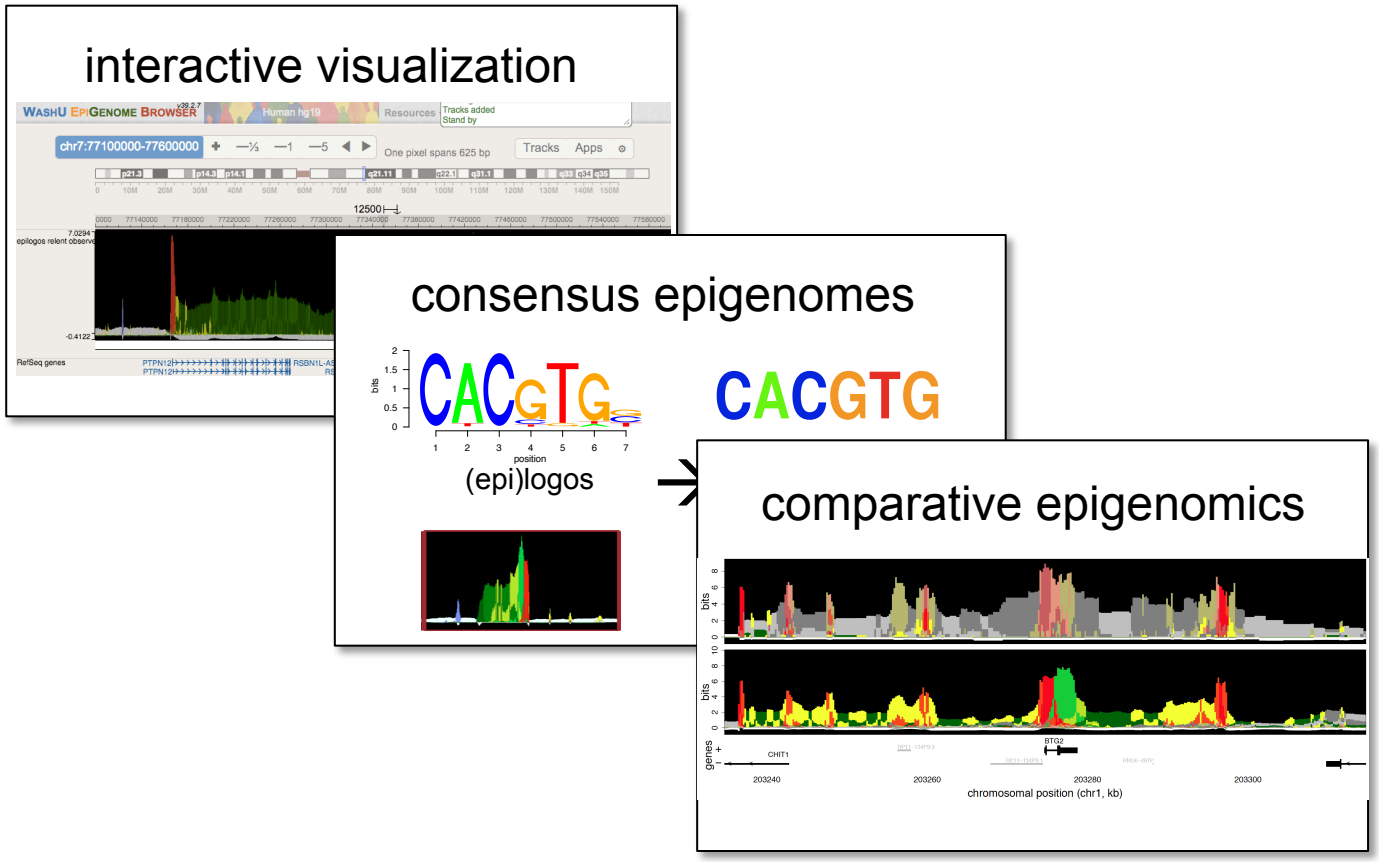
Heterochromatin	chr7:25483000-25483200
Enhancers	chr7:25660600-25660800
Active TSS	chr7:25990800-25991000
Transcr at gene 5' and 3'	chr7:26236800-26237000
Repressed PolyComb	chr7:26415000-26415200
Active TSS	chr7:26704600-26704800
Flanking Active TSS	chr7:26904800-26905000
Flanking Bivalent TSS/Enh **[Current Window]**	chr7:27053397-27373765
Active TSS	chr7:27702600-27702800
Flanking Active TSS	chr7:28011800-28012000
Flanking Active TSS	chr7:28218800-28219000
Flanking Bivalent TSS/Enh	chr7:28449000-28449200
Heterochromatin	chr7:28699600-28699800
Flanking Bivalent TSS/Enh	chr7:28893400-28893600

global high-surprisal regions

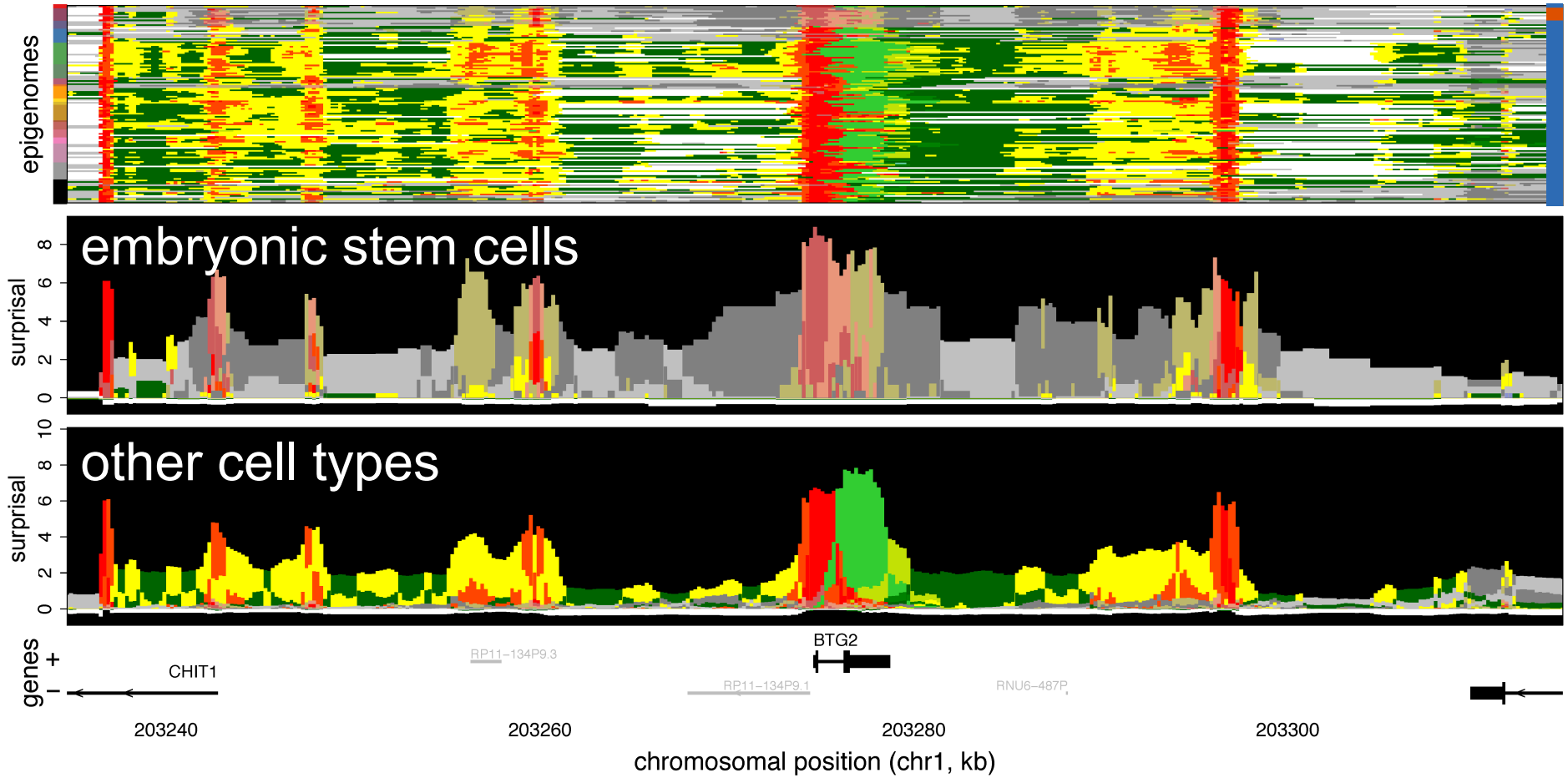
Select/Deselect All

1	Bivalent/Poised TSS	chr6:137365600-137365800
2	Bivalent/Poised TSS	chr9:77112200-77112400
3	Bivalent Enhancer	chr12:57636000-57636200
4	Bivalent Enhancer	chr11:69453000-69453200
5	Bivalent Enhancer	chr10:8084600-8084800
6	Bivalent Enhancer	chr17:35300400-35300600
7	Bivalent Enhancer	chr17:58220600-58220800
8	Bivalent Enhancer	chr16:3016000-3016200
9	Bivalent Enhancer	chr16:68680000-68680200
10	Bivalent Enhancer	chr19:18062400-18062600
11	Bivalent Enhancer	chr19:18336200-18336400
12	Bivalent Enhancer	chr19:46475400-46475600
13	Bivalent Enhancer	chr19:47975400-47975600
14	Bivalent Enhancer	chr19:55952200-55952400

Some potential uses of `epilogos`

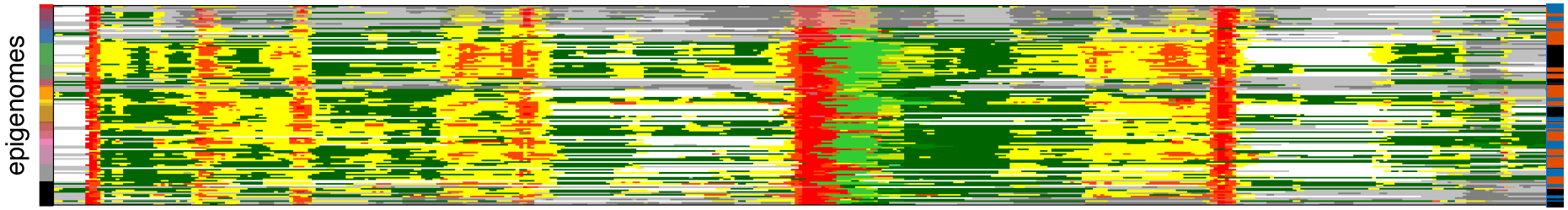


Comparative epigenomics of ESC vs. “rest”



Select two subsets of epigenomes,
build separate **`epilogos`** and compare them

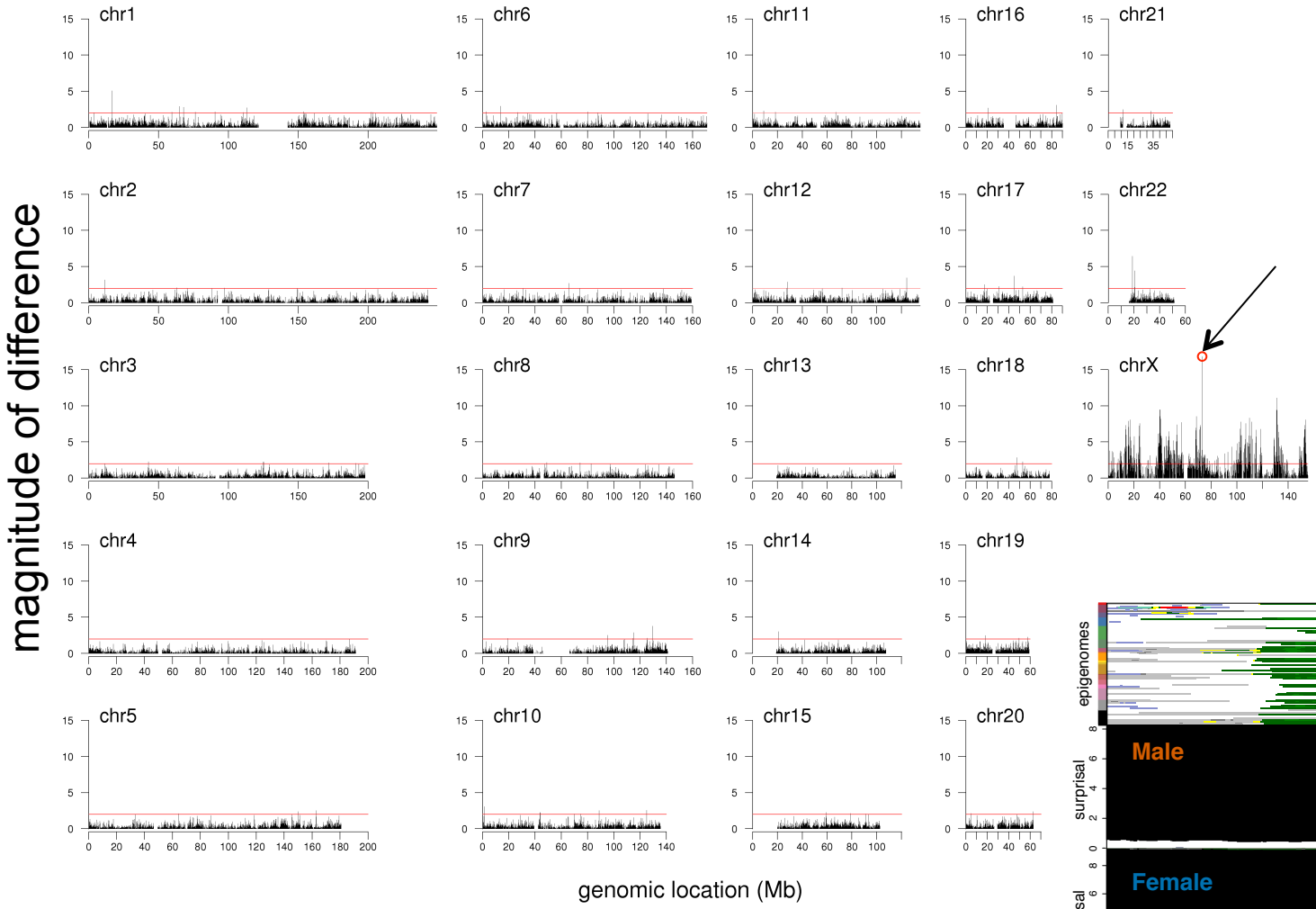
Comparative epigenomics of ???

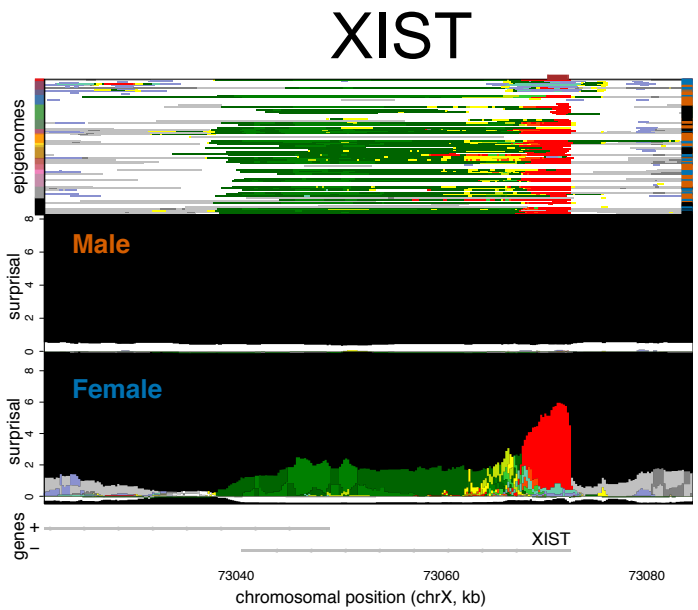


Arbitrary subsets of epigenomes,
such as these ones (red vs. blue)

Select two subsets of epigenomes,
build separate **`epilogos`** and compare them

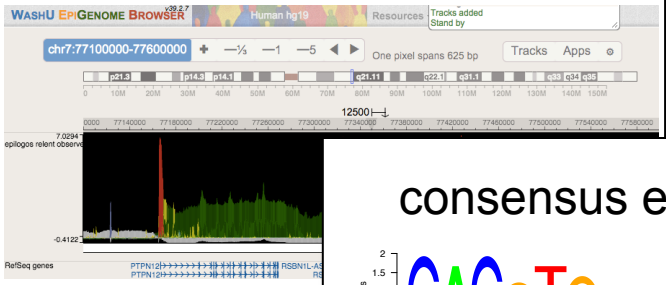
Comparative epigenomics of male vs. female



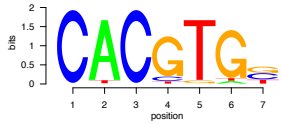


Some potential uses of **epilogos**

interactive visualization

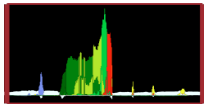


consensus epigenomes

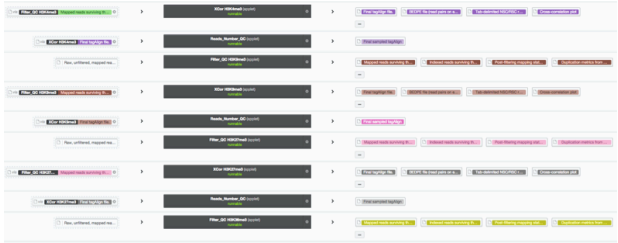


CACGTG

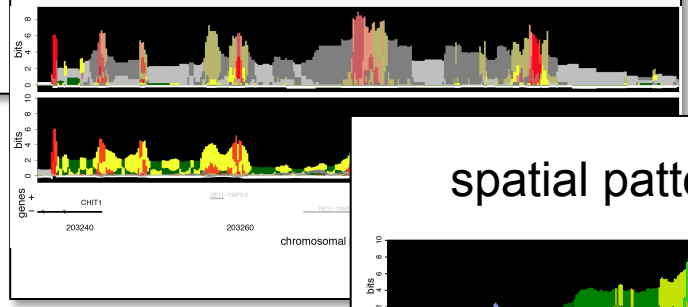
(epi)logos



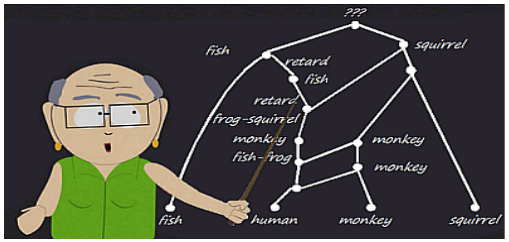
use with third-party data



comparative epigenomics

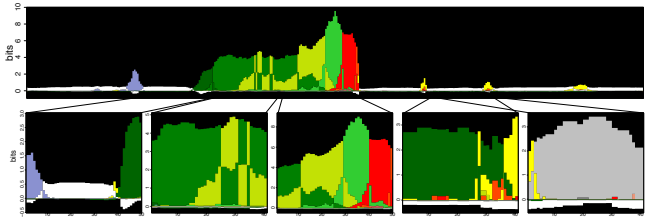


temporal pattern analysis



("phyloepigenomic" modeling)

spatial pattern analysis



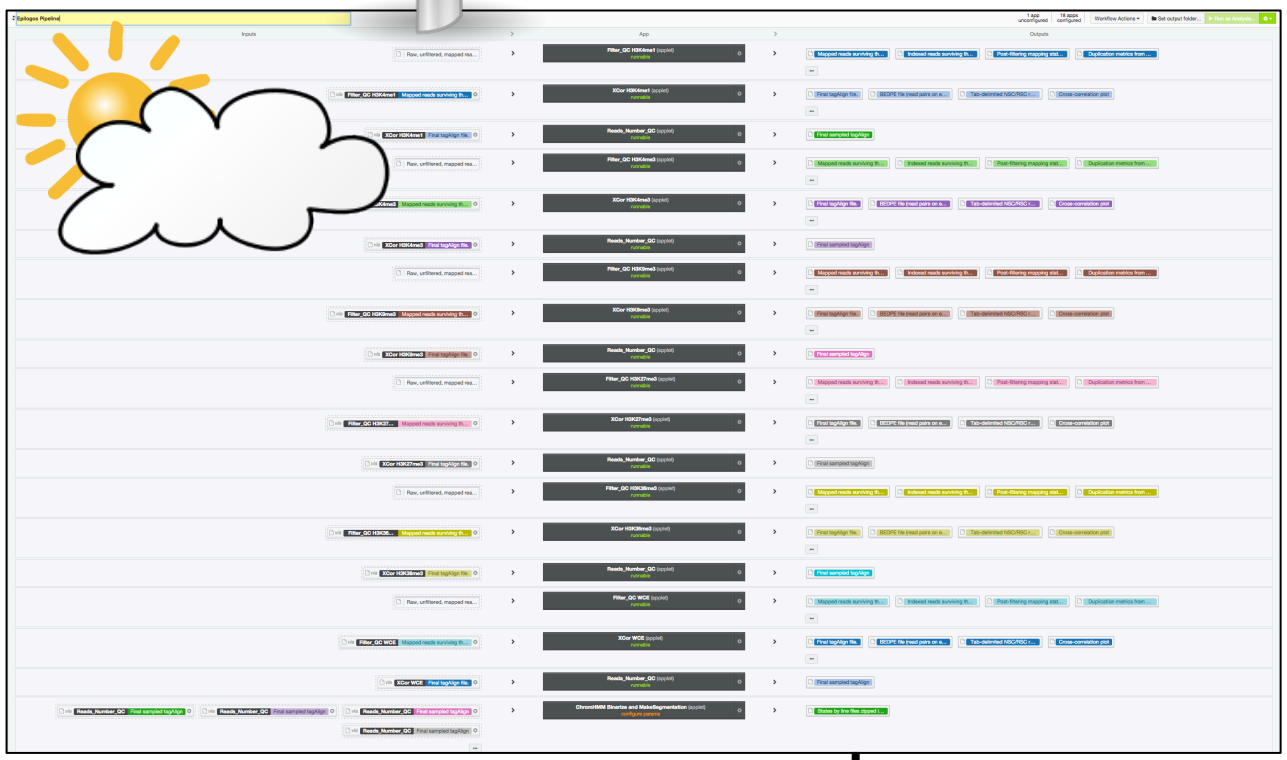
(*de novo* motif discovery & search)

Use with third-party data

input: mapped ChIP-seq reads
of histone tail modifications

~10GB

Actual integration with Roadmap reference epigenomes



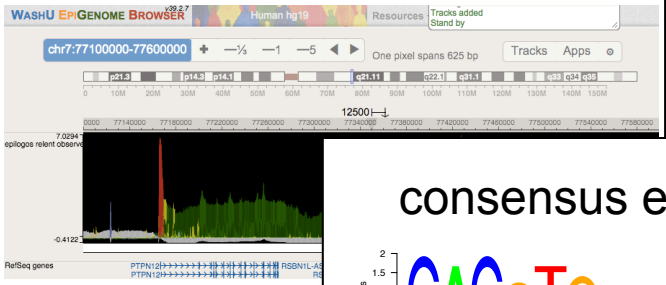
Group Select	EID	Epigenome name
IMR90	E017	IMR90 fetal lung fibroblasts Cell Line
ESC	E002	ES-WA7 Cells
	E008	H9 Cells
	E001	ES-I3 Cells
	E015	HUES6 Cells
	E014	HUES48 Cells
	E016	HUES64 Cells
	E003	H1 Cells
	E024	ES-UCSF4 Cells
iPSC	E020	iPS-20b Cells
	E019	iPS-18 Cells
	E018	iPS-15b Cells
	E021	iPS DF 6.9 Cells
	E022	iPS DF 19.11 Cells
ES-deriv	E007	H1 Derived Neuronal Progenitor Cultured Cells
	E009	H9 Derived Neuronal Progenitor Cultured Cells
	E010	H9 Derived Neuron Cultured Cells
	E013	hESC Derived CD56+ Mesoderm Cultured Cells
	E012	hESC Derived CD56+ Ectoderm Cultured Cells
	E011	hESC Derived CD184+ Endoderm Cultured Cells
	E004	H1 BMP4 Derived Mesendoderm Cultured Cells
	E005	H1 BMP4 Derived Trophoblast Cultured Cells
	E006	H1 Derived Mesenchymal Stem Cells
	your sample?	

~1MB

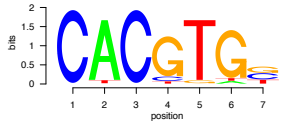
output: chromatin state calls
according to Roadmap models

Some potential uses of **epilogos**

interactive visualization

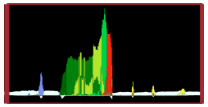


consensus epigenomes

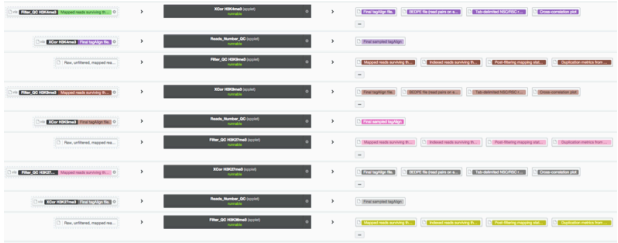


CACGTG

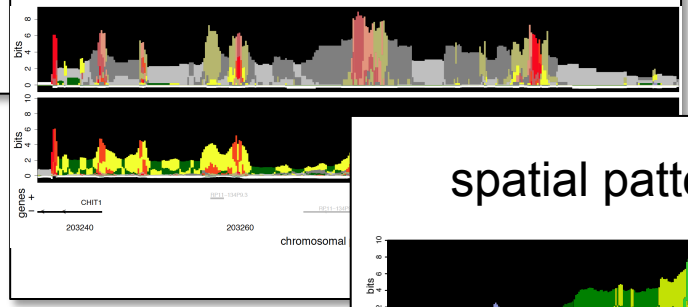
(epi)logos



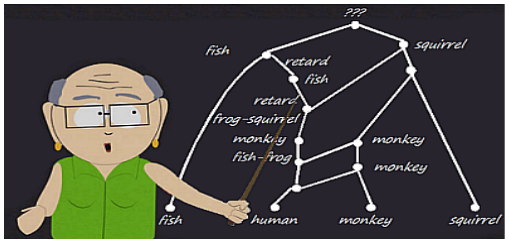
use with third-party data



comparative epigenomics

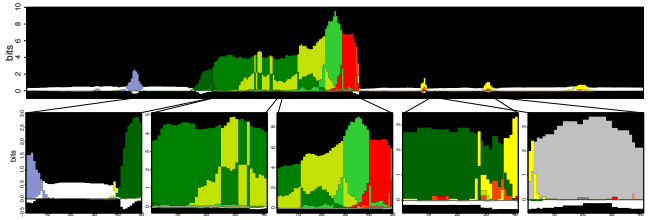


temporal pattern analysis



("phyloepigenomic" modeling)

spatial pattern analysis



(*de novo* motif discovery & search)

THANKS!

ENCODE & Roadmap Epigenomics Consortia (data)

Soheil Feizi (information theory)

Apostolos Papadopoulos, Terrance Liang, Kevin Liu, Tiffany Chen & Miguel Medrano (web-application & more)

Ting Wang, Xin Zhou & Daofeng Li (WashU browser integration)

Luca Pinello & Nezar Abdennur (website technology)

YOU!

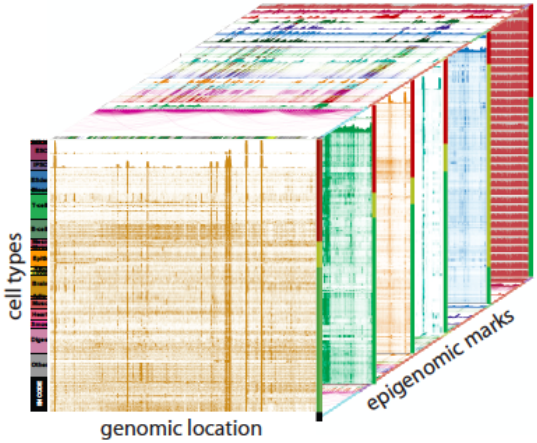
→ tinyurl.com/epilogos-feedback

→ epilogos.broadinstitute.org

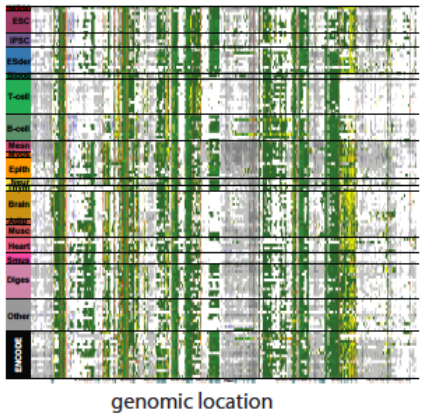
epilogos

Interpretation of large-scale (epi)genomic datasets through information-based dimensionality reduction

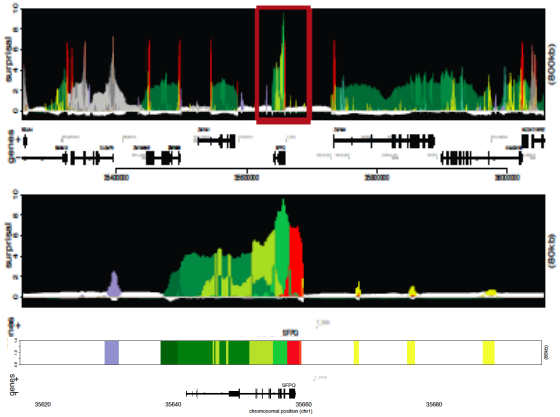
3D



2D



1D



Questions?