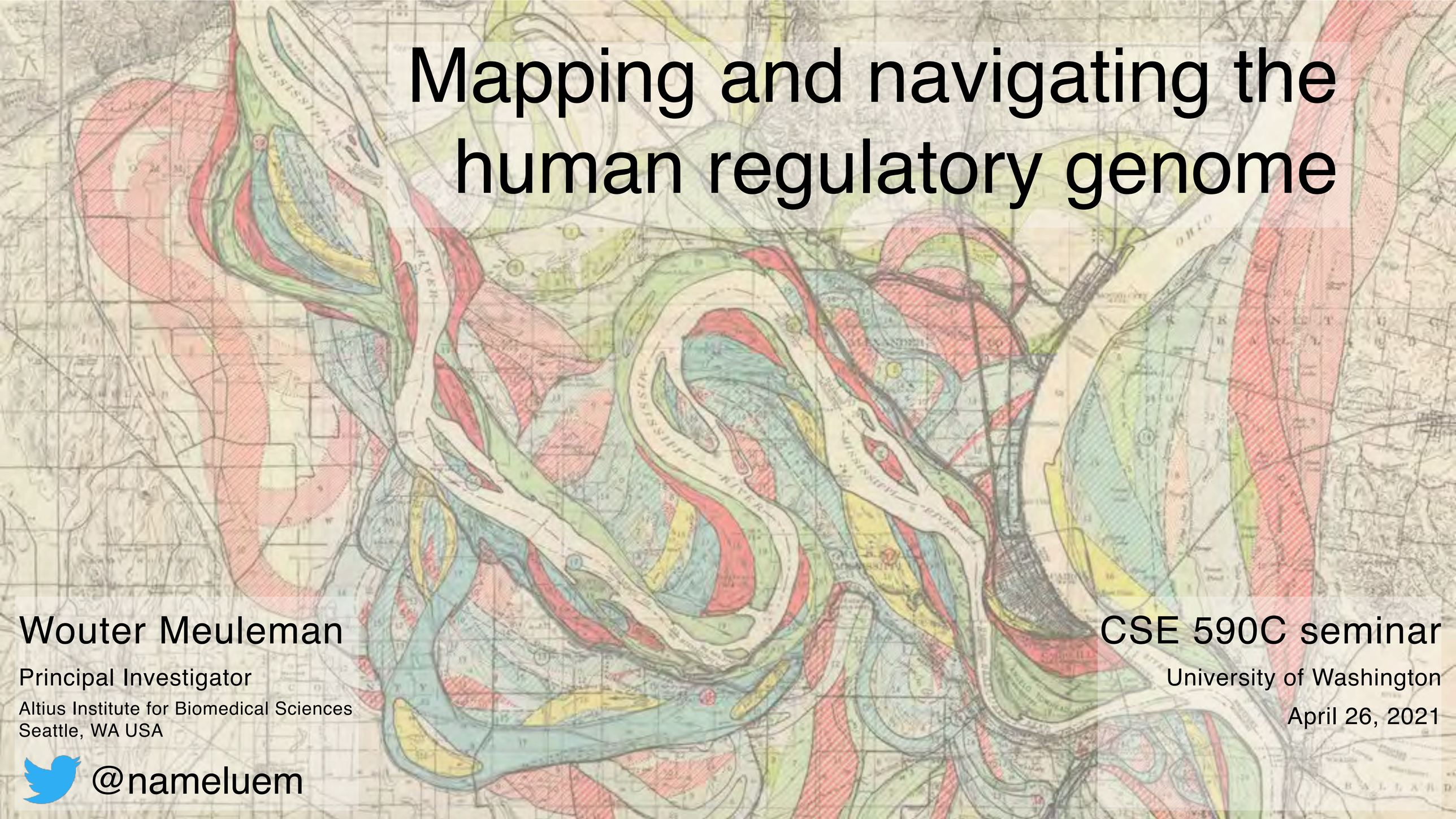
Mapping and navigating the human regulatory genome



Wouter Meuleman

Principal Investigator

Altius Institute for Biomedical Sciences
Seattle, WA USA

@nameluem

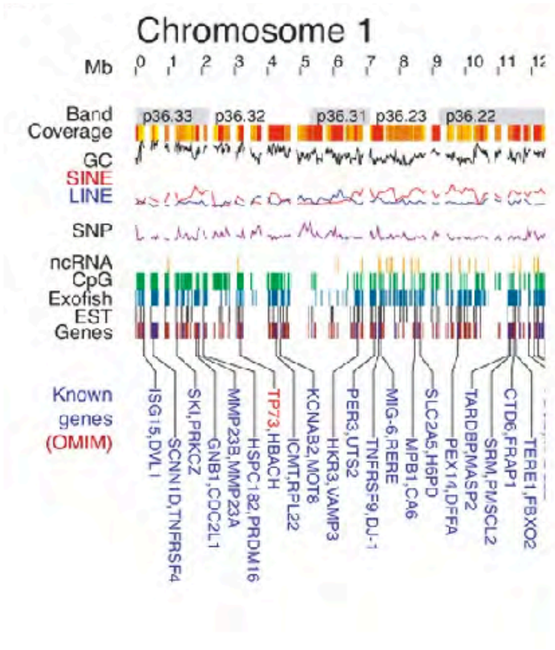
CSE 590C seminar

University of Washington

April 26, 2021

The Human Genome

The Map (2001)



Lander *et al.* Nature (2001)

20 years

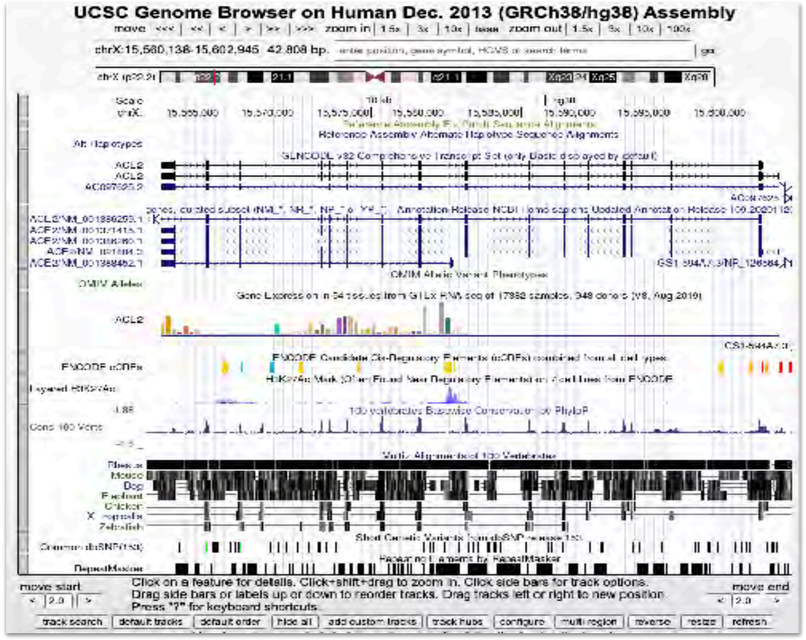
The Navigation System (1981)

m66roepers @ Flickr

Fast-forward 20 years!

The Human Genome

The Map (2021)



UCSC Genome Browser (2021)

20 years

The Navigation System (2001)

TomTom Navigator (2001)

From paper to screen, but still hard to access and interpret at scale

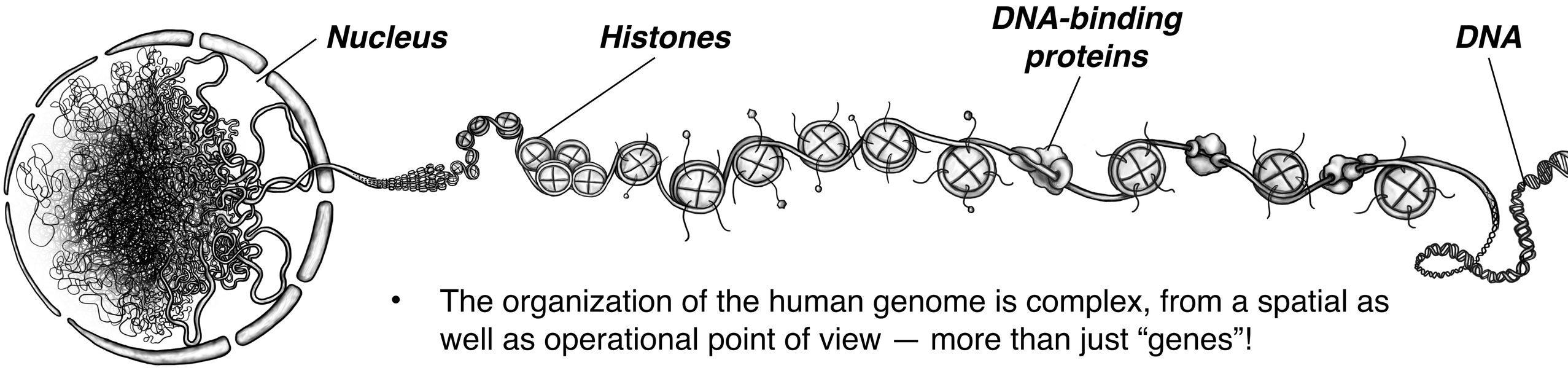
Maps (should) encourage exploration

Maps are useful because they are wrong*

- Maps provide summarized representations of reality, highlighting only the most relevant information
- What is considered '**relevant**' depends on e.g. the mapped subject matter, map resolution, and context
- Towards a Disney Map of Genomics, we need to **annotate** the most 'exciting attractions' of the genome

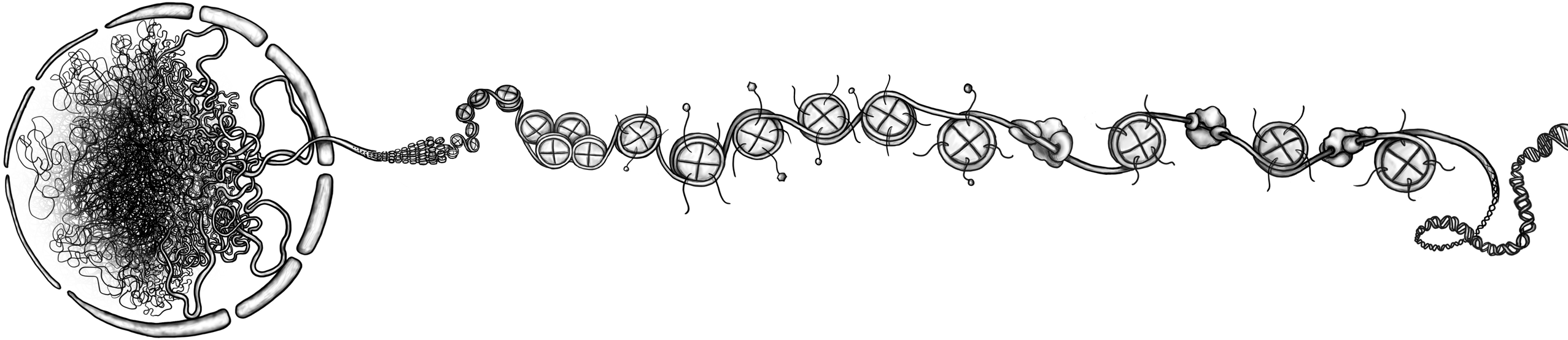
*: Paraphrasing/misquoting George Box (1987)

In search of ‘relevance’: annotating the (regulatory) human genome



- The organization of the human genome is complex, from a spatial as well as operational point of view — more than just “genes”!
- Protein-coding regions make up at most a few % of the genome, with regulatory elements hidden in its vast non-coding portion.
- To interrogate the (non-coding) genome, many experimental methods are available, most utilizing high-throughput sequencing.
- The resulting genome-wide datasets can be hard to interpret on their own, but offer lots of opportunities for creating useful annotations.

In search of ‘relevance’: annotating the (regulatory) human genome



Chromatin domains

Scale: 10kbp-Mbp

Guelen *et al.*, Nature (2008)
Peric-Hupkes, Meuleman *et al.* Mol. Cell (2010)
Meuleman *et al.*, Genome Res. (2013)

Chromatin states

Scale: 200bp-1kbp

Kundaje, Meuleman *et al.*, Nature (2015)
Claussnitzer *et al.*, NEJM (2015)
Marco, Meuleman *et al.*, Nature Comm. (2017)

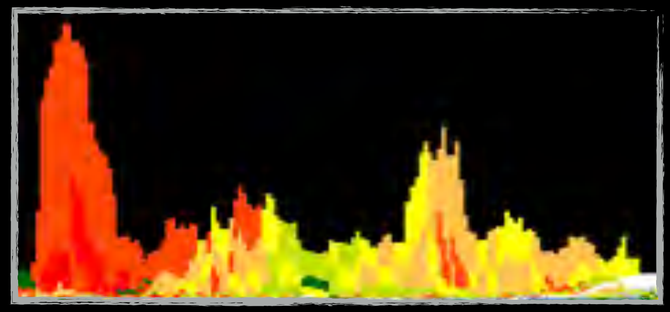
Chromatin accessibility

Scale: <200bp

Meuleman *et al.*, Nature (2020)
Vierstra *et al.*, Nature (2020)
Boix *et al.*, Nature (2021)

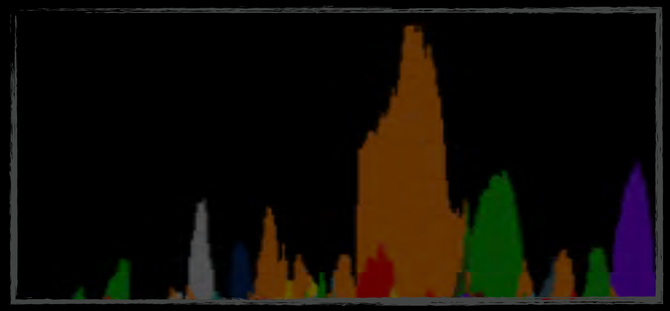
In search of ‘relevance’: two types of genomic annotations

- **Chromatin states (epilogos):**
  “What type of functionality does a genomic region encode?”
  (e.g. promoter, enhancer, repressor)

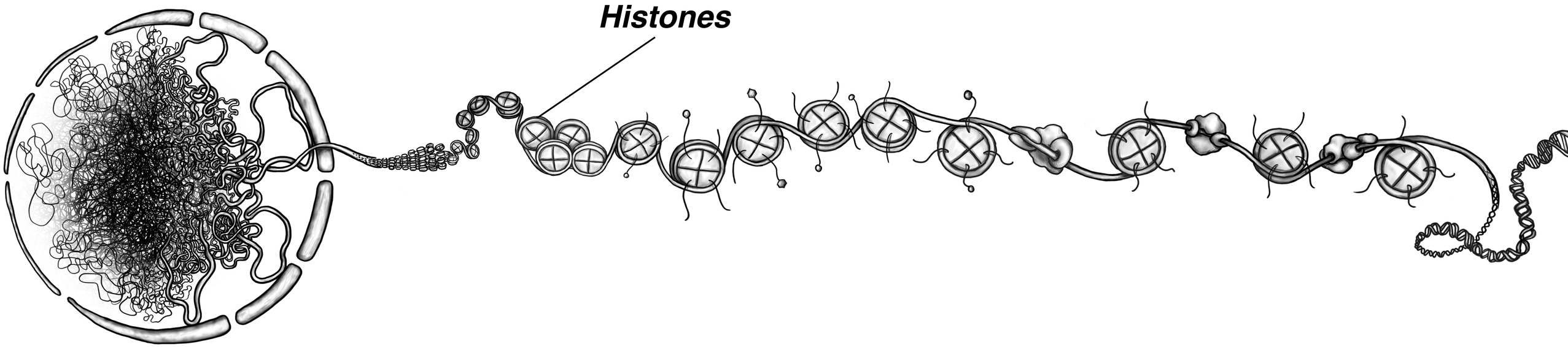


https://epilogos.net

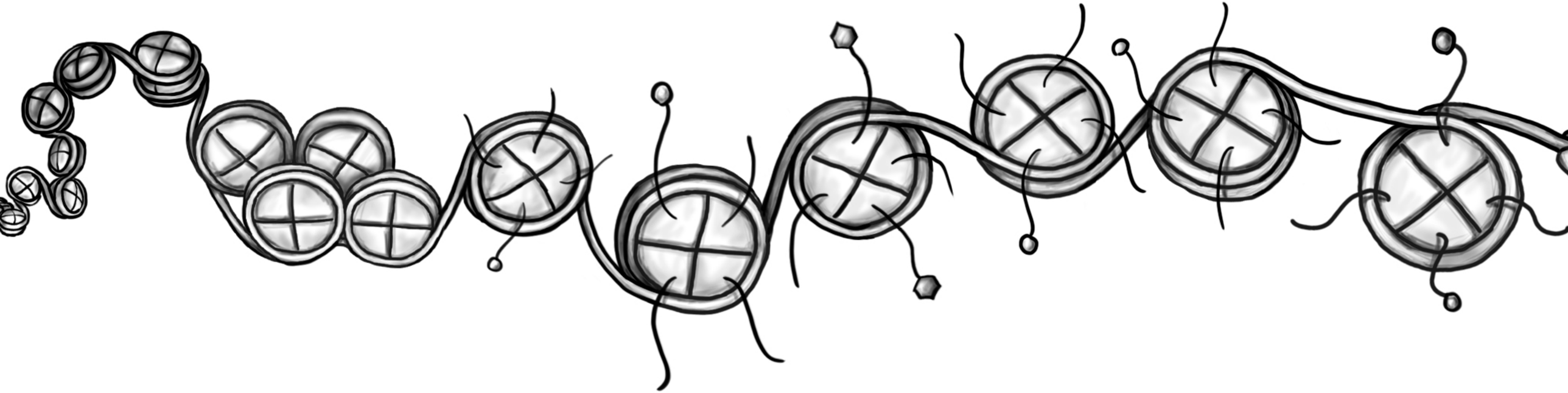
- **Chromatin accessibility (DHS Index)**:
  “In which cellular contexts are regulatory regions utilized?”
  (e.g. cardiac, lymphoid, neural)



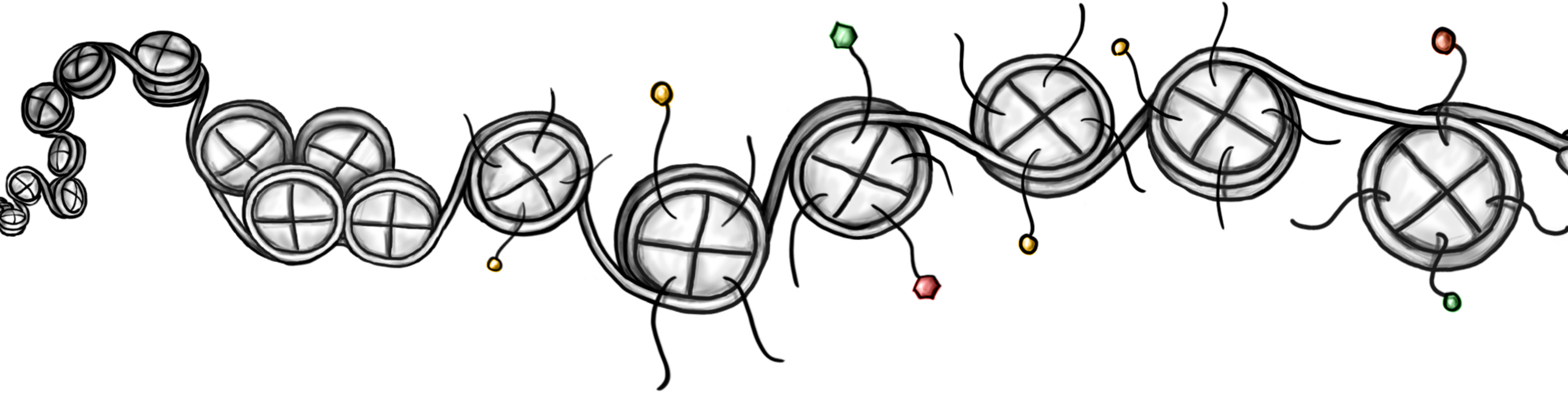
Meuleman *et al.*, 2020 & ongoing

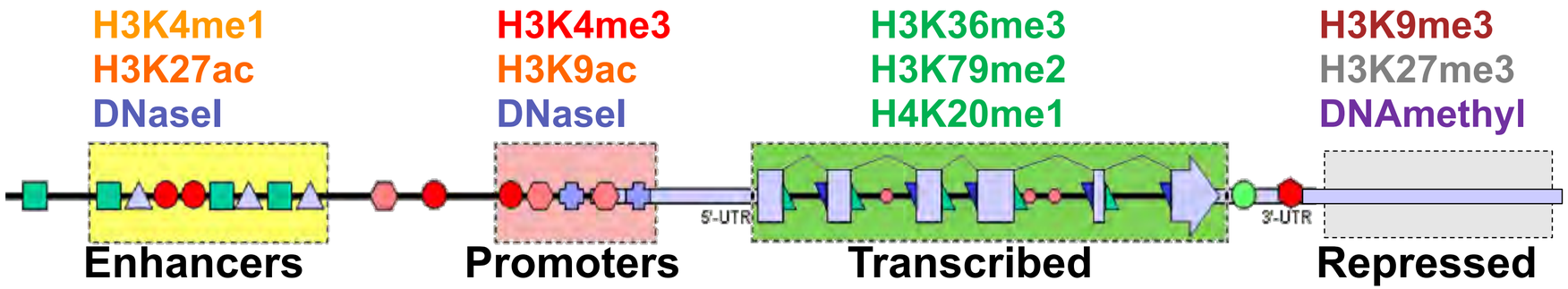

Histones

Histone tails can be chemically tagged with ***epigenomic marks***



These epigenomic marks are associated with ***functional elements***





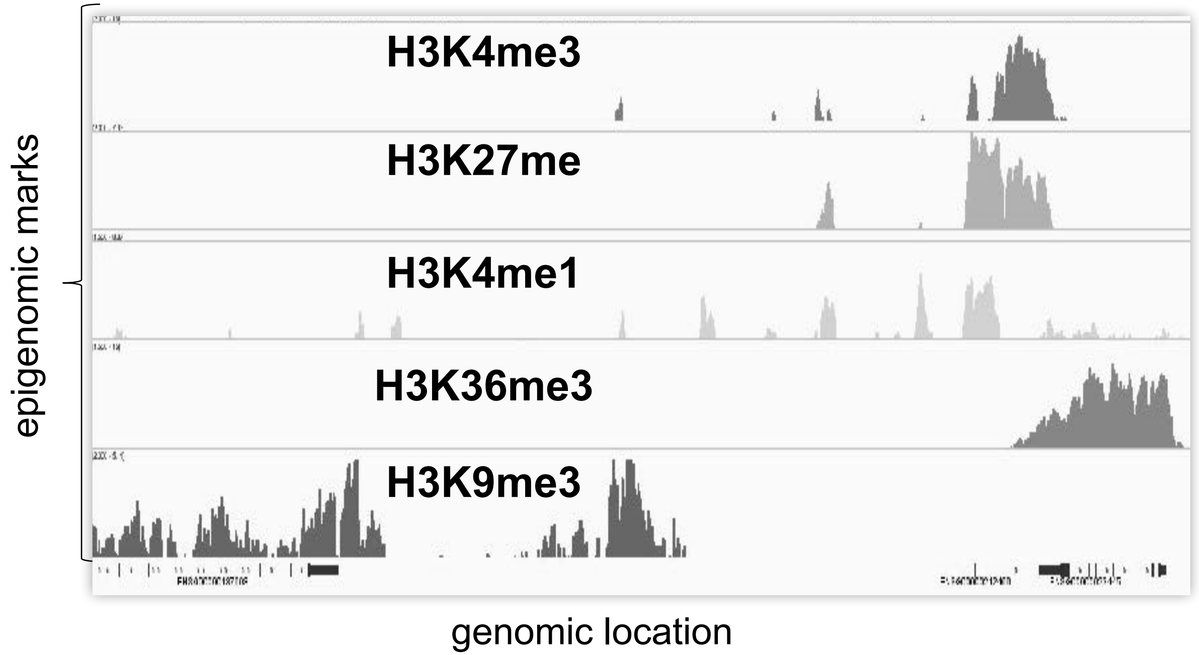
The NIH Roadmap Epigenomics Project (2008-2017)

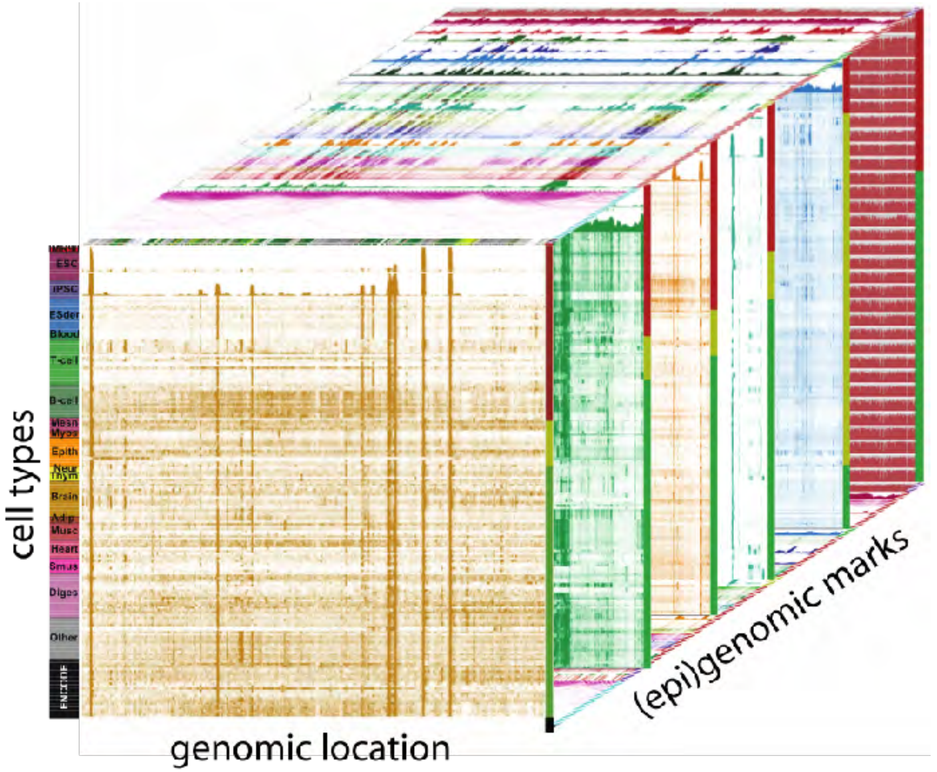
Goal: create reference maps of a wide variety of epigenomic marks across many cell types from healthy individuals.

- 4 Reference Epigenome Mapping Centers (REMCs)
- Central data repository and read mapping at Baylor
- Uniform processing and integrative analysis at MIT

Genome-wide profiling of epigenomic marks has resulted in giant data cubes

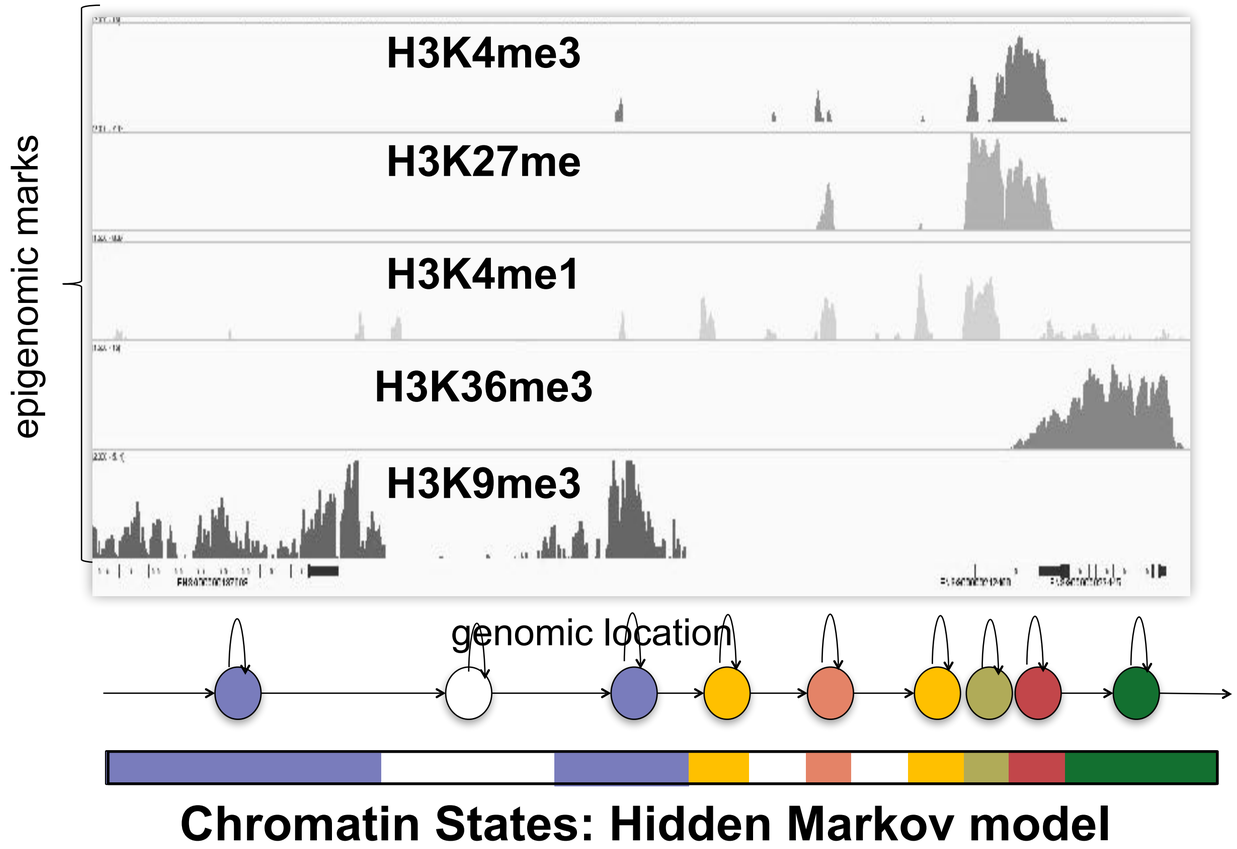
Chromatin immunoprecipitation sequencing (ChIP-seq)

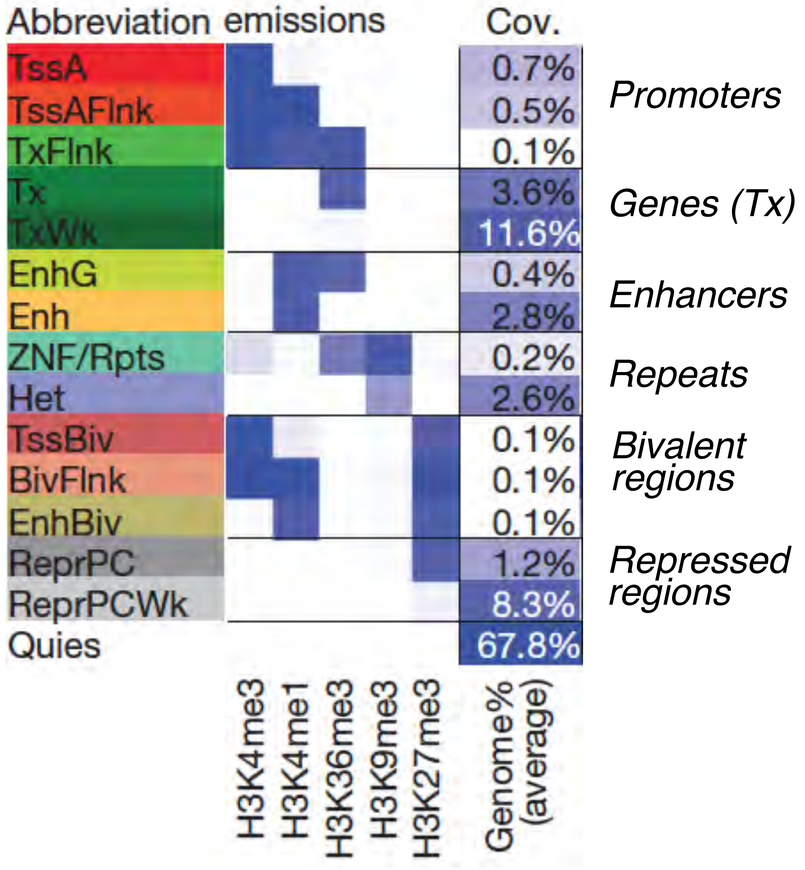




Kundaje*, Meuleman*, *et al.*, Nature (2015)

Data can be summarized by learning a limited number of ***chromatin states***



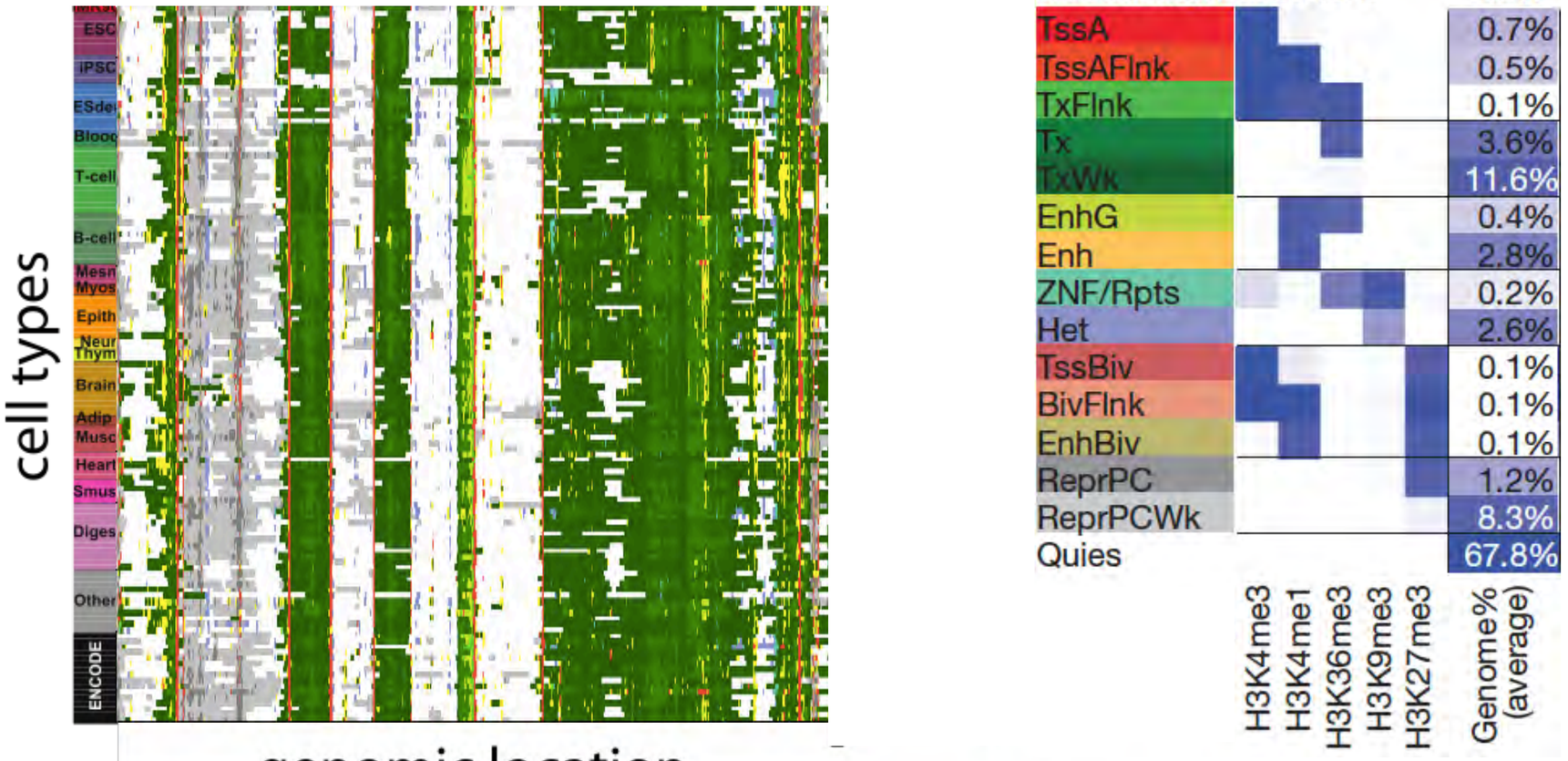


Kundaje*, Meuleman*, *et al.*, Nature (2015)

This allows us to transform the 3D cube into a 2D matrix:

A color here…

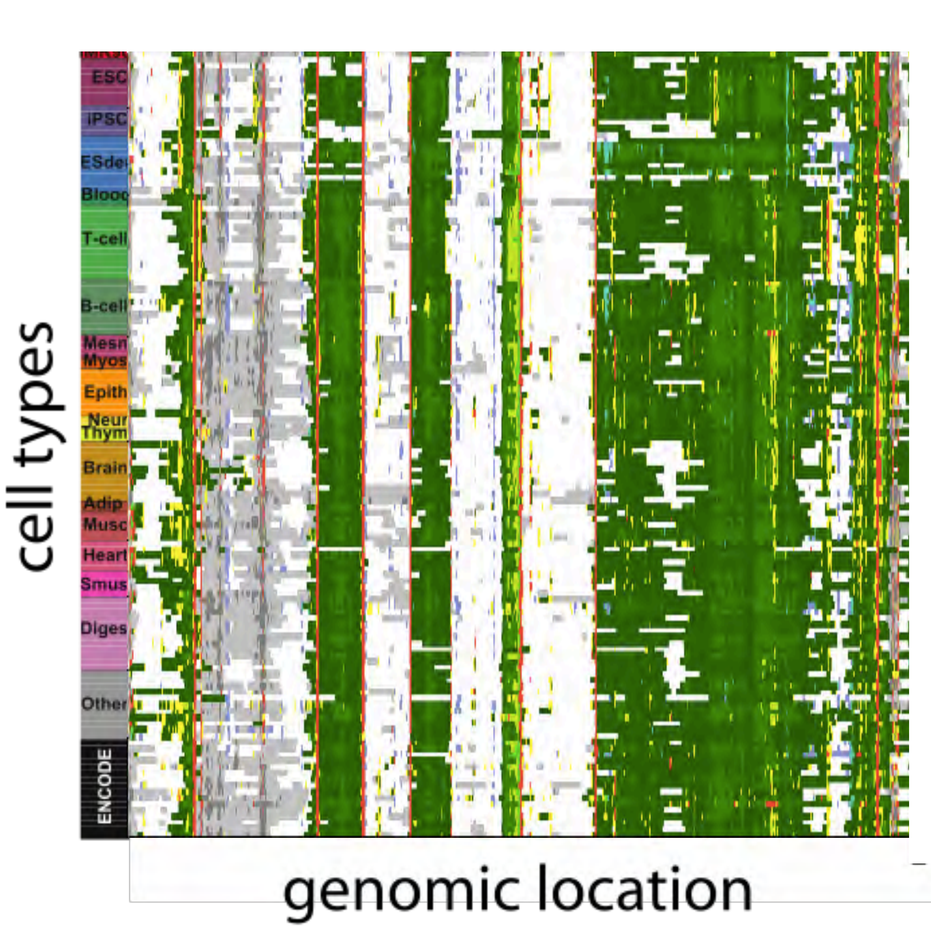
…corresponds to a state here:

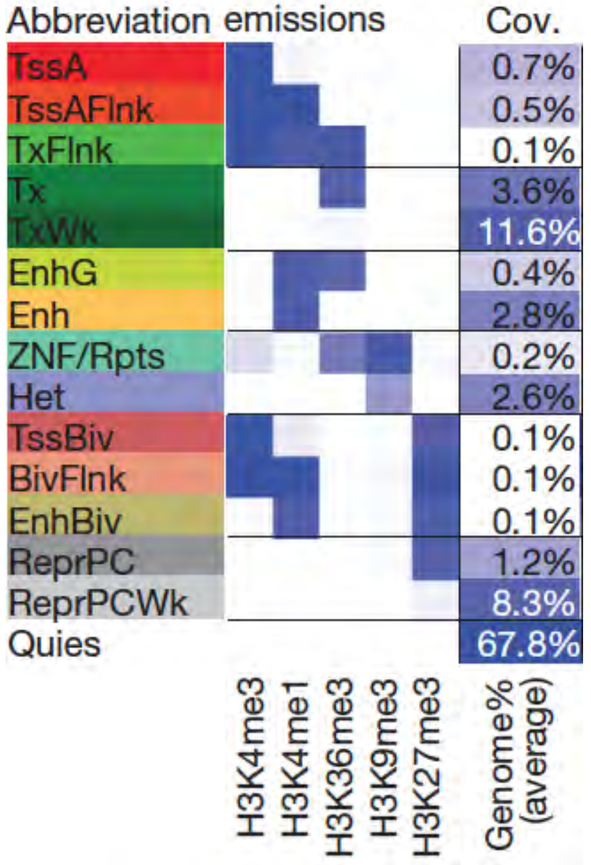


Kundaje*, Meuleman*, *et al.*, Nature (2015)

A reference map of chromatin states across 127 epigenomes

Documents the dynamics of chromatin states
between cell types, e.g. during cell differentiation

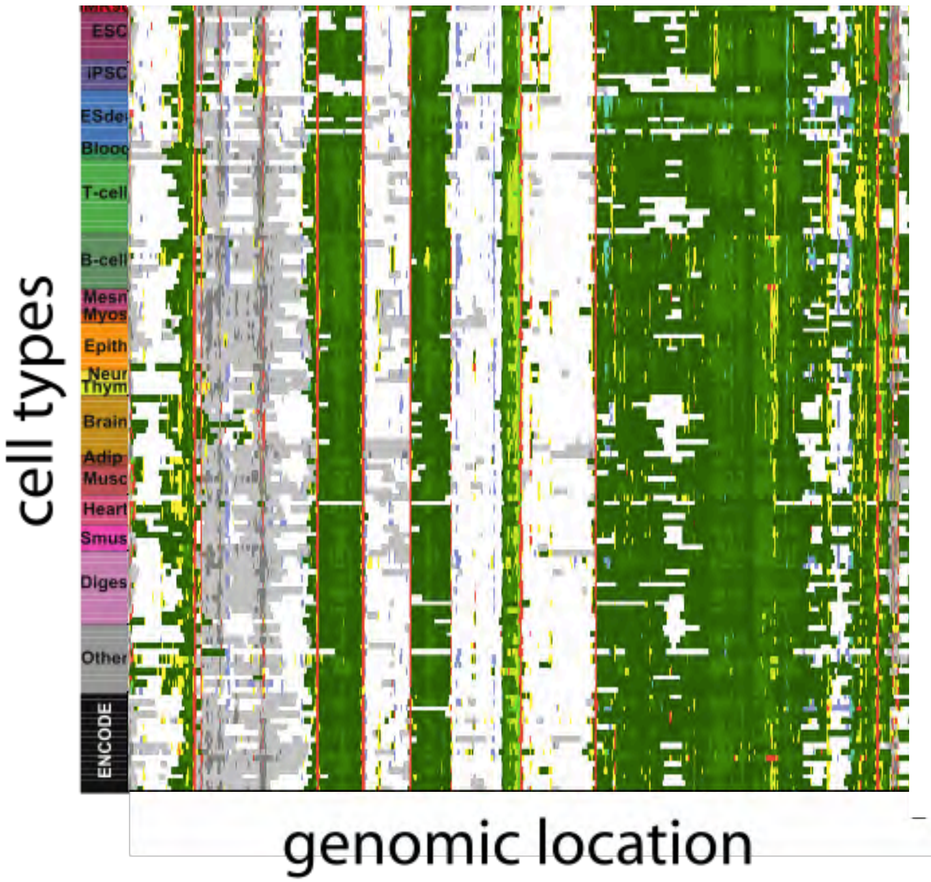




Kundaje*, Meuleman*, *et al.*, Nature (2015)

New methods are needed to navigate these maps

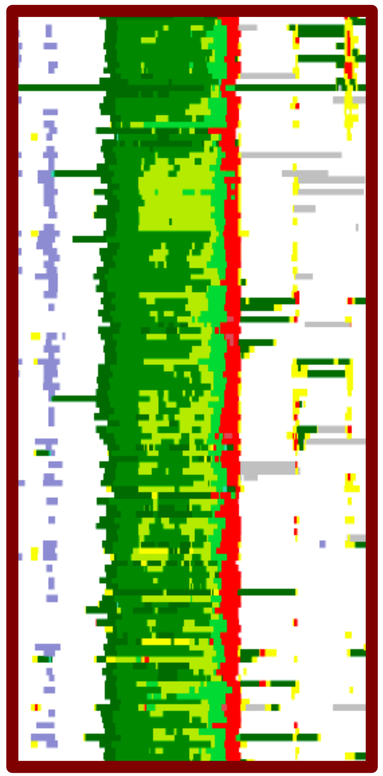
Many more epigenomes
are being profiled
(cell types, disease states,
personal epigenomics, etc)



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

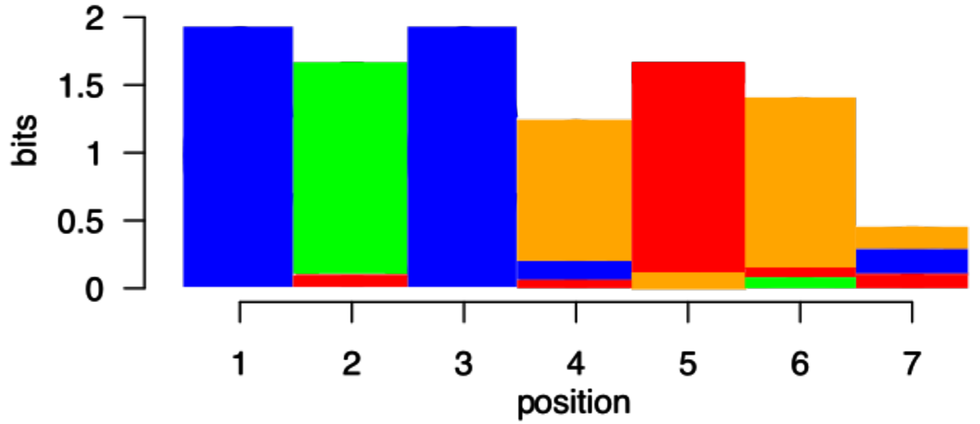
Chromatin states across many epigenomes: analogy with sequence motifs

Alignments of multiple sequences



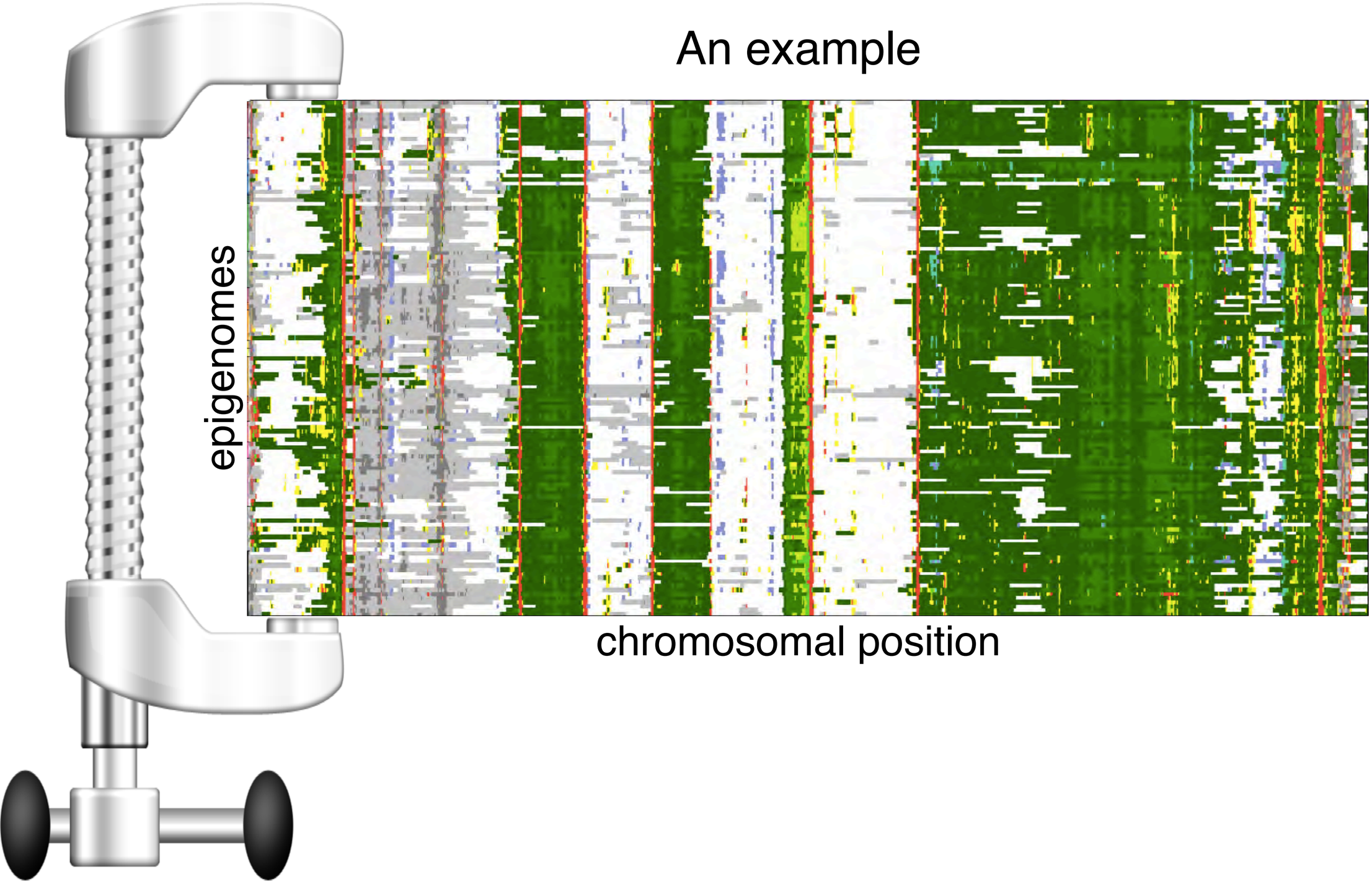
~

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

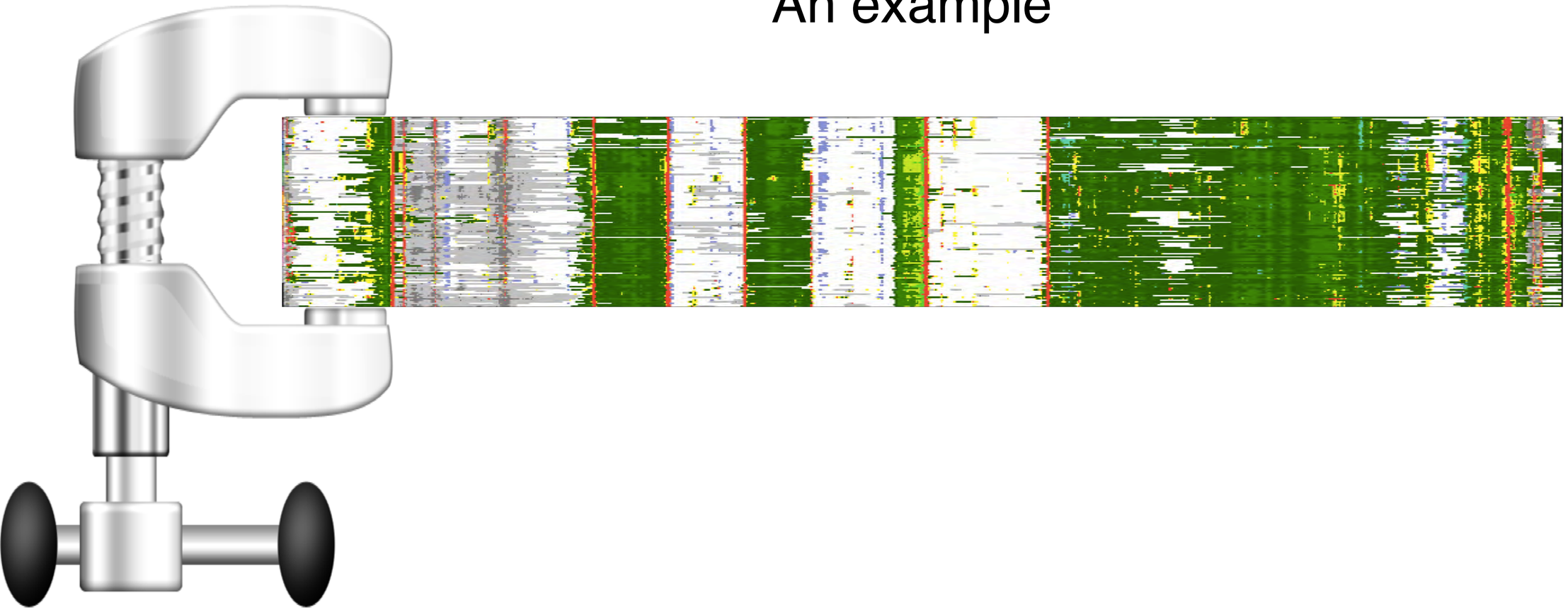


There are good ways of modeling such alignments: logos!
Information content of a region, considering background

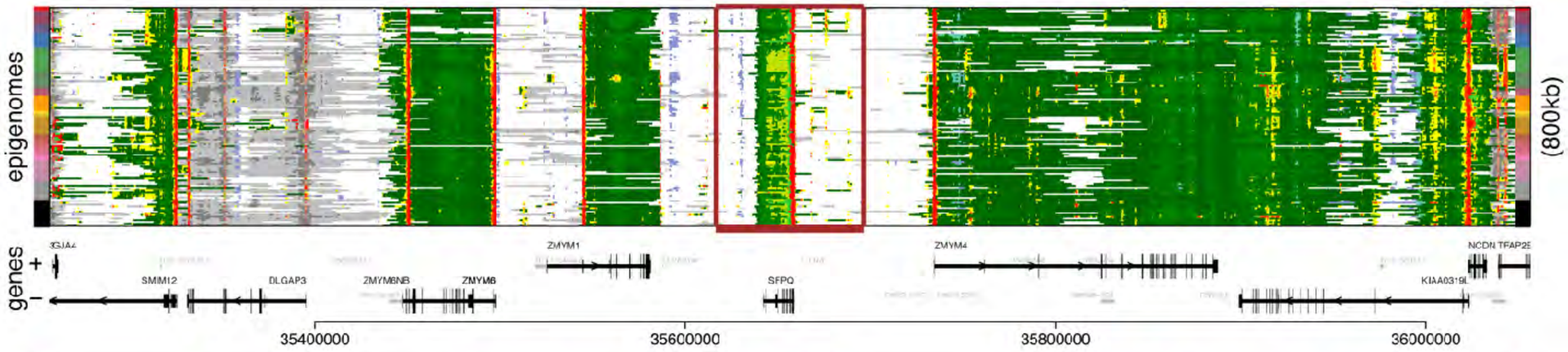
An example



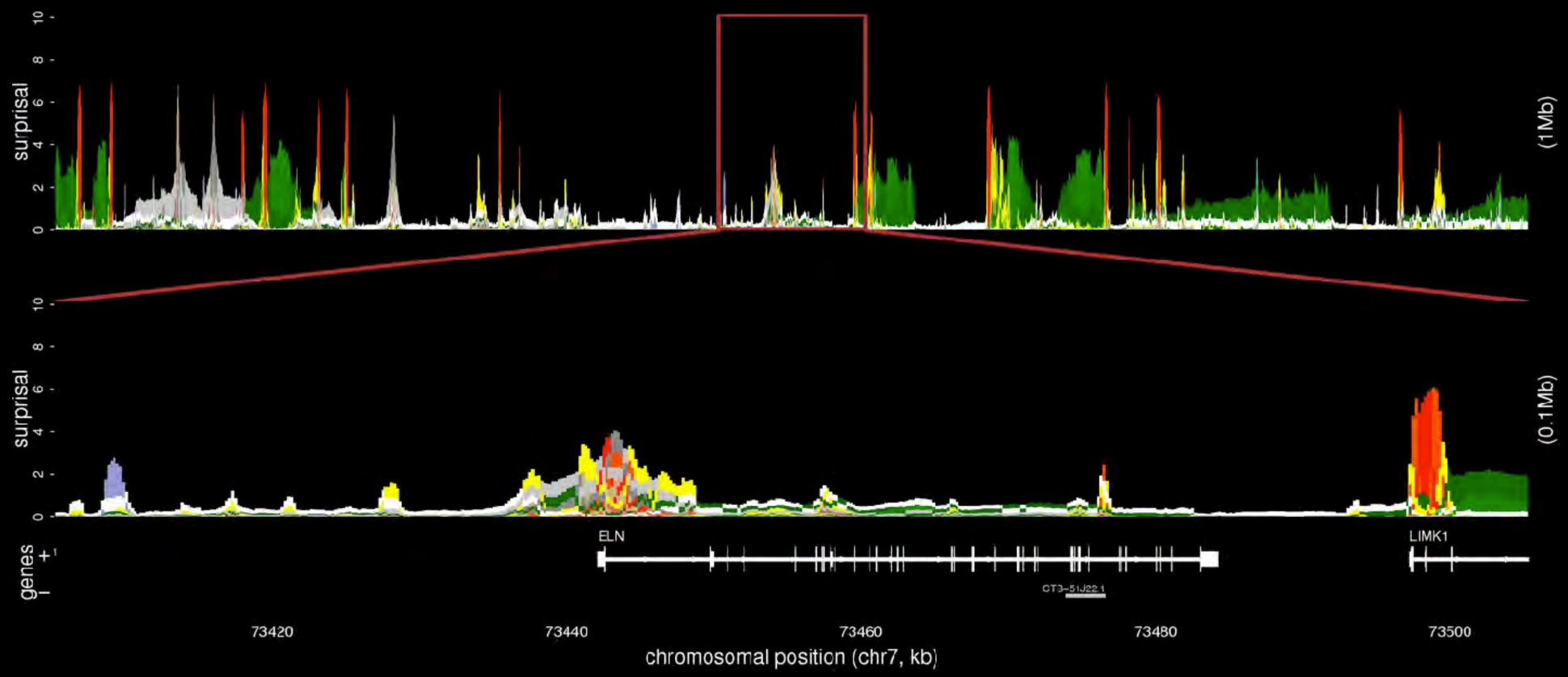
An example



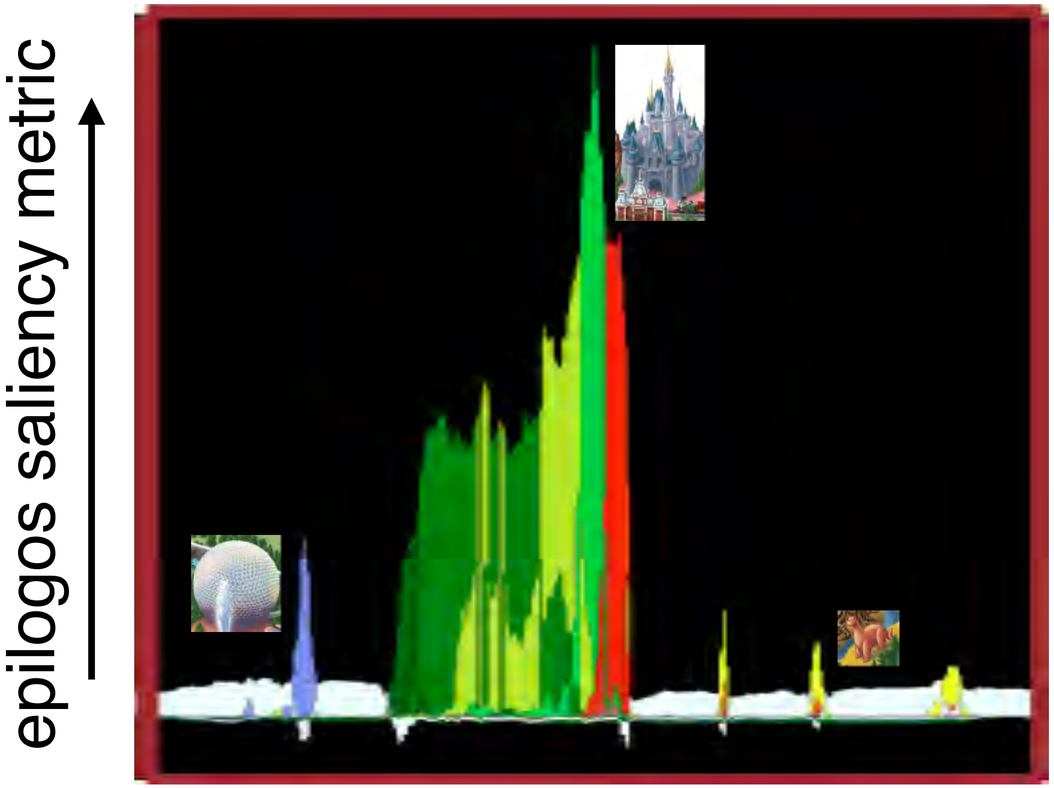
epilogos

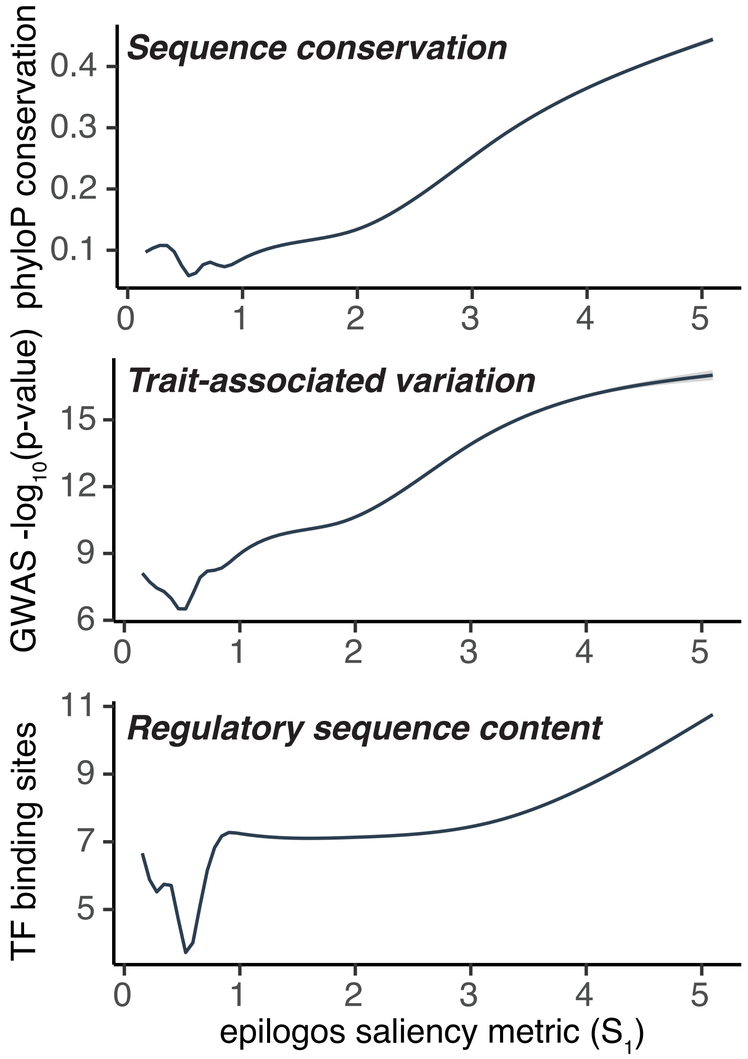

epigenomes
(800kb)
genes
+
−
SMIM12
DLGAP3
ZMYM6NB
ZMYM6
ZMYM1
SFPQ
ZMYM4
KIAA0319L
35400000
35600000
35800000
36000000

epilogos


surprisal
0
2
4
6
8
10
(1Mb)
surprisal
0
2
4
6
8
10
(0.1Mb)
genes
1
+
-
ELN
LIMK1
CT3-51J22.1
73420
73440
73460
73480
73500
chromosomal position (chr7, kb)

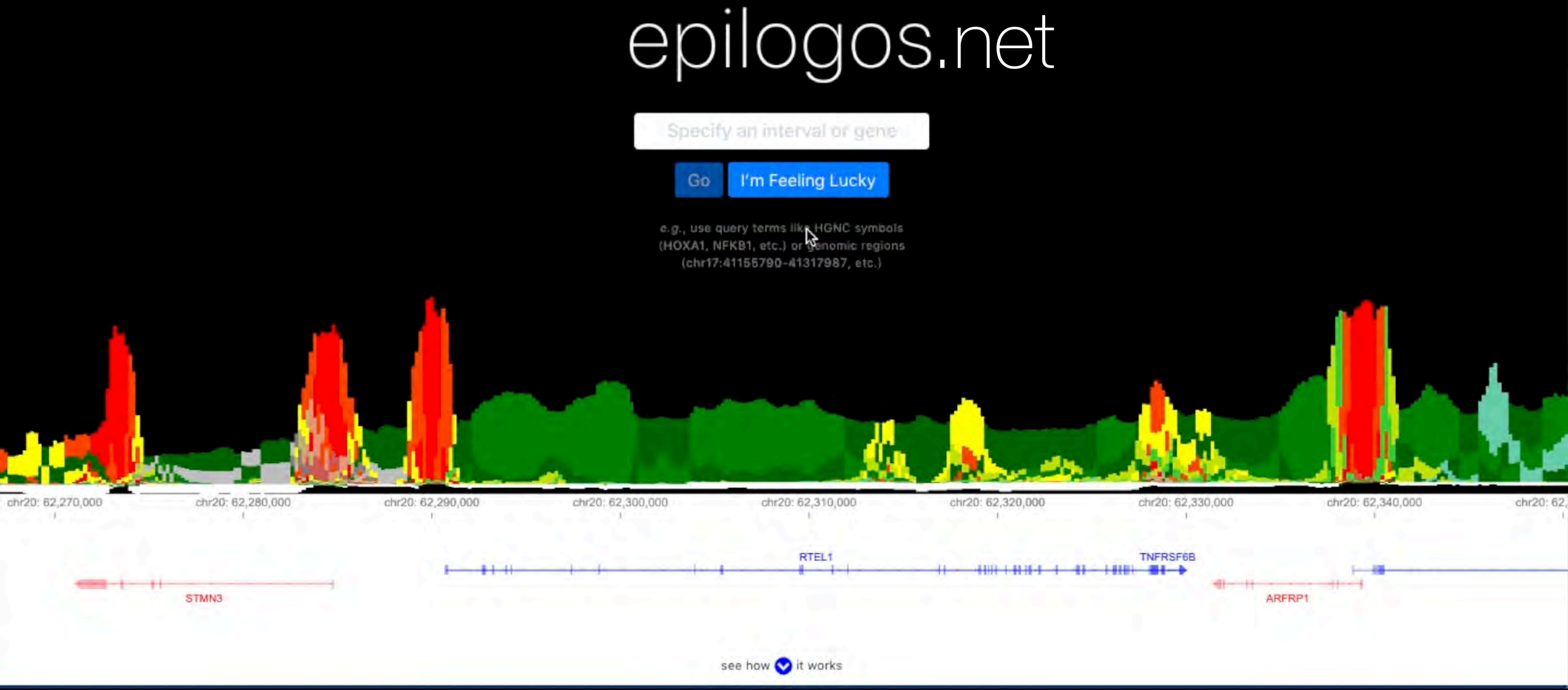
Epilogos saliency metrics enrich for functionally relevant regions





Work with Nasi Teodosiadis

You’re looking at a summarization of ~5,000 genome-wide datasets

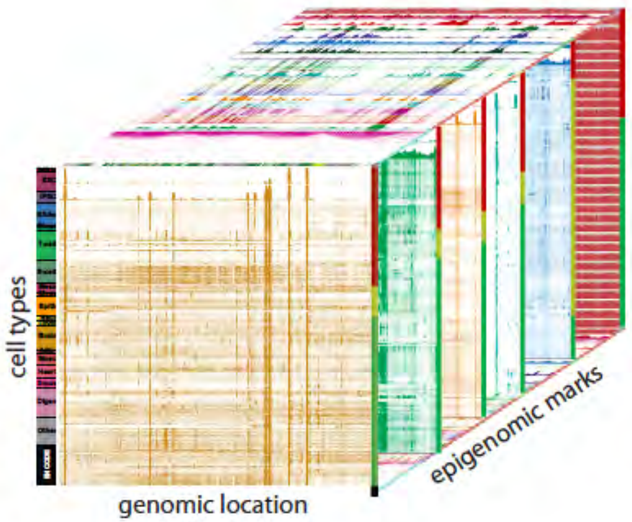


epilogos.net

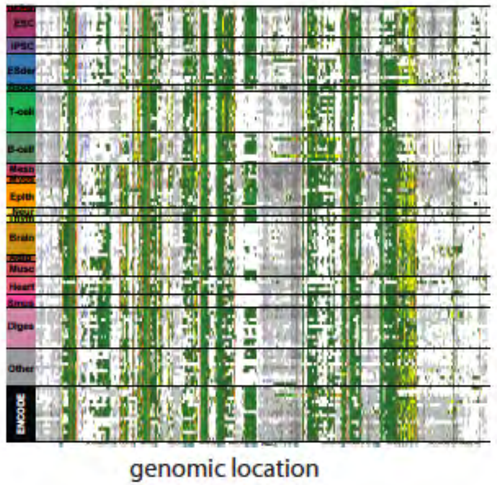
epilogos

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

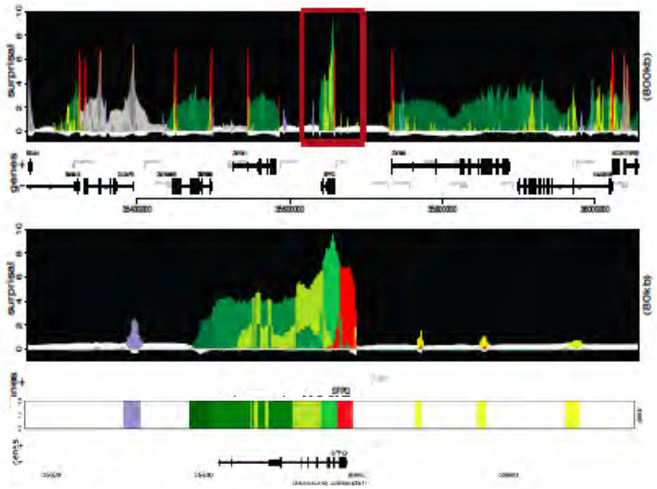
3D



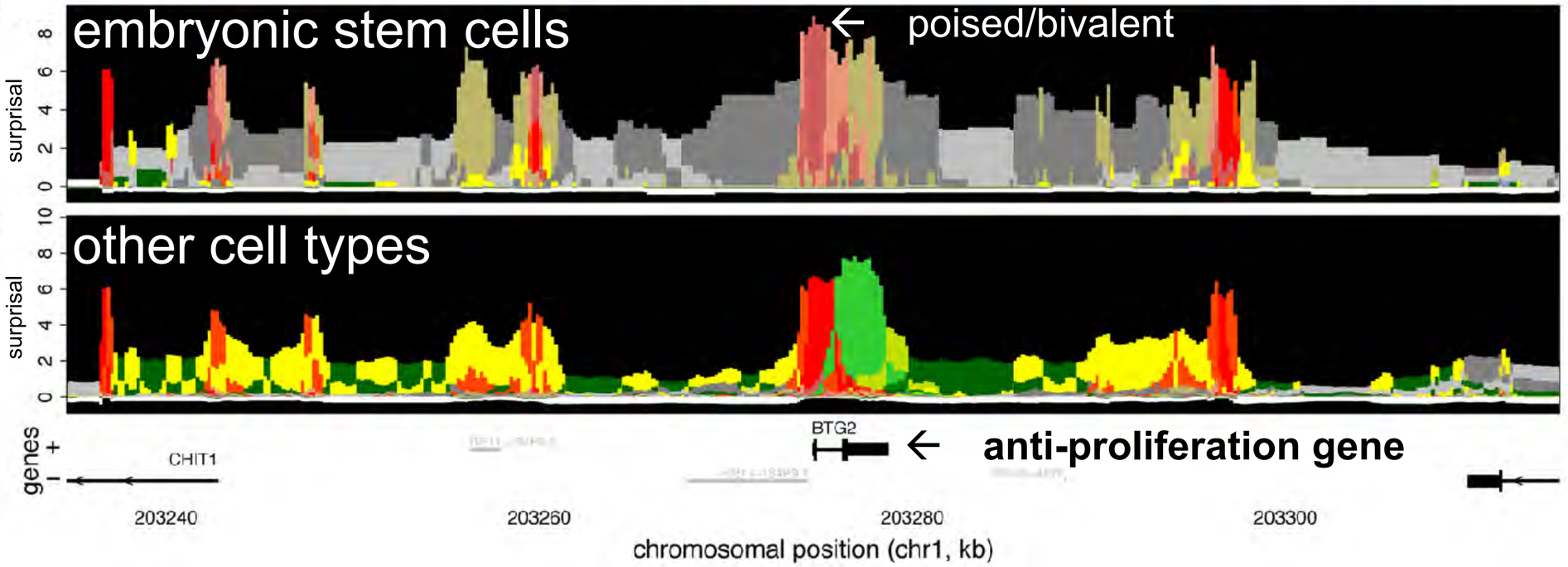
2D



1D

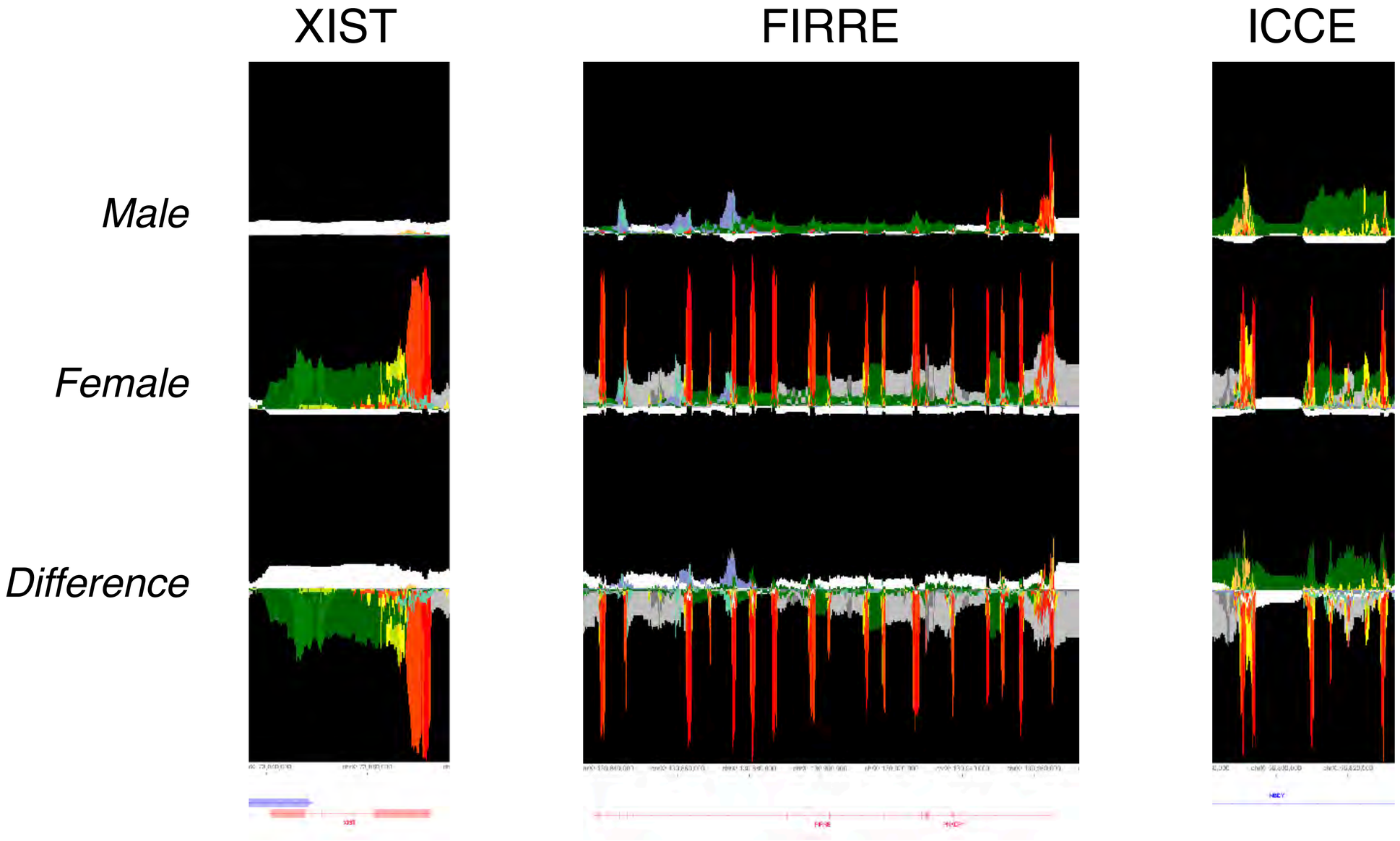


Pairwise comparison of groups of biosamples or interest



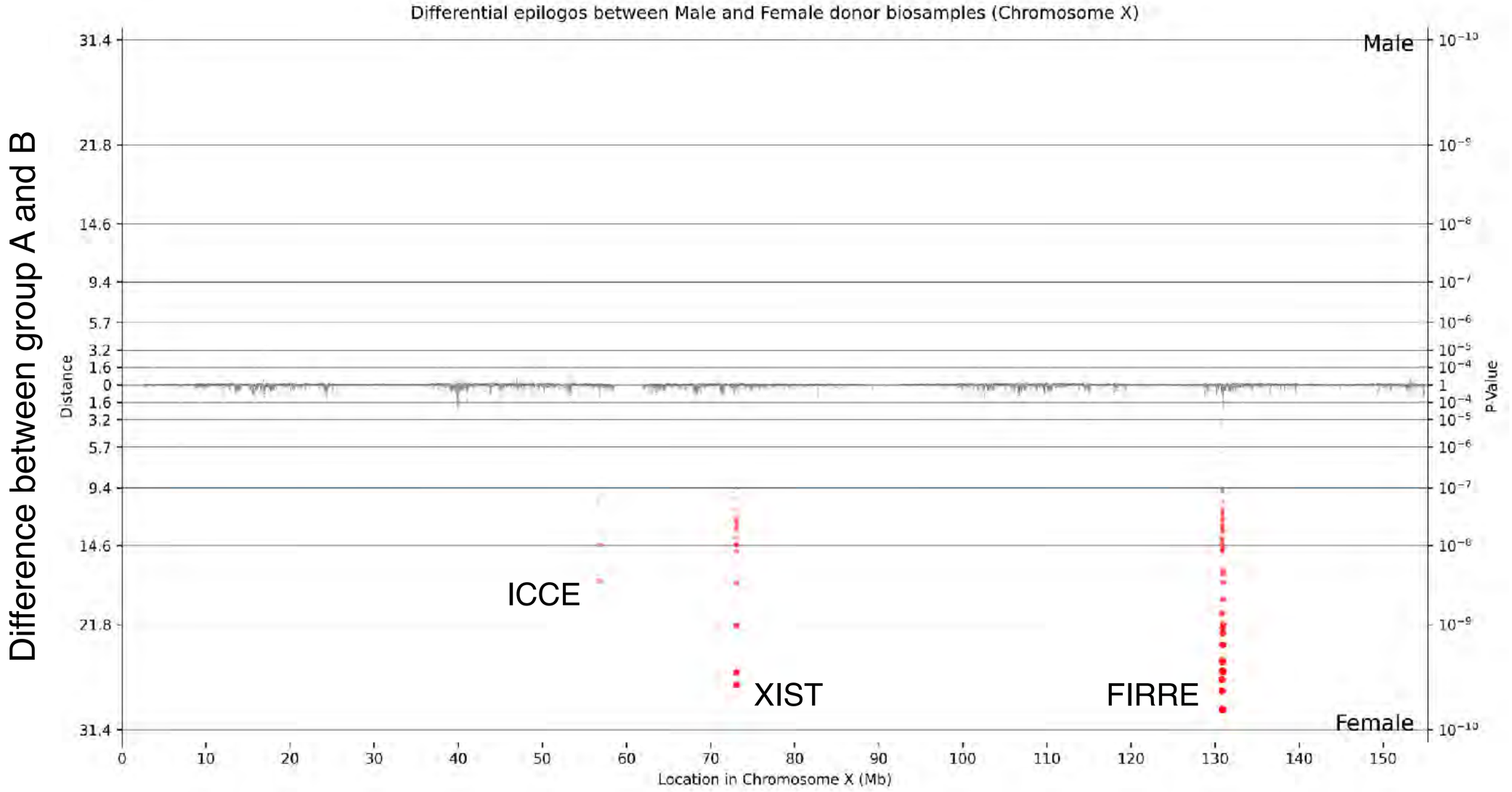
Work with Jacob Quon

A comparison of male vs. female donors



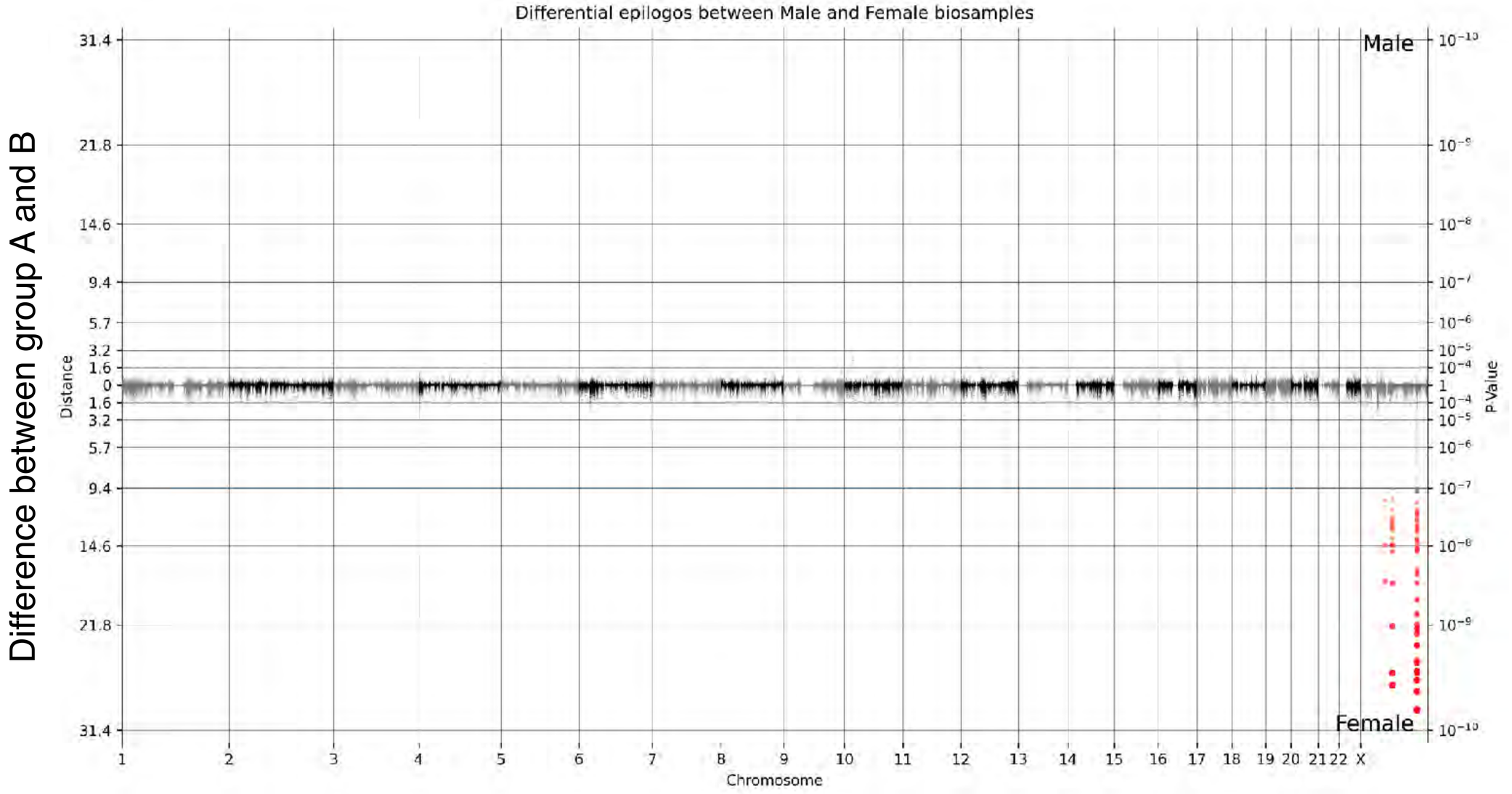
Work with Jacob Quon

Male vs. female donors (total of 684 biosamples)



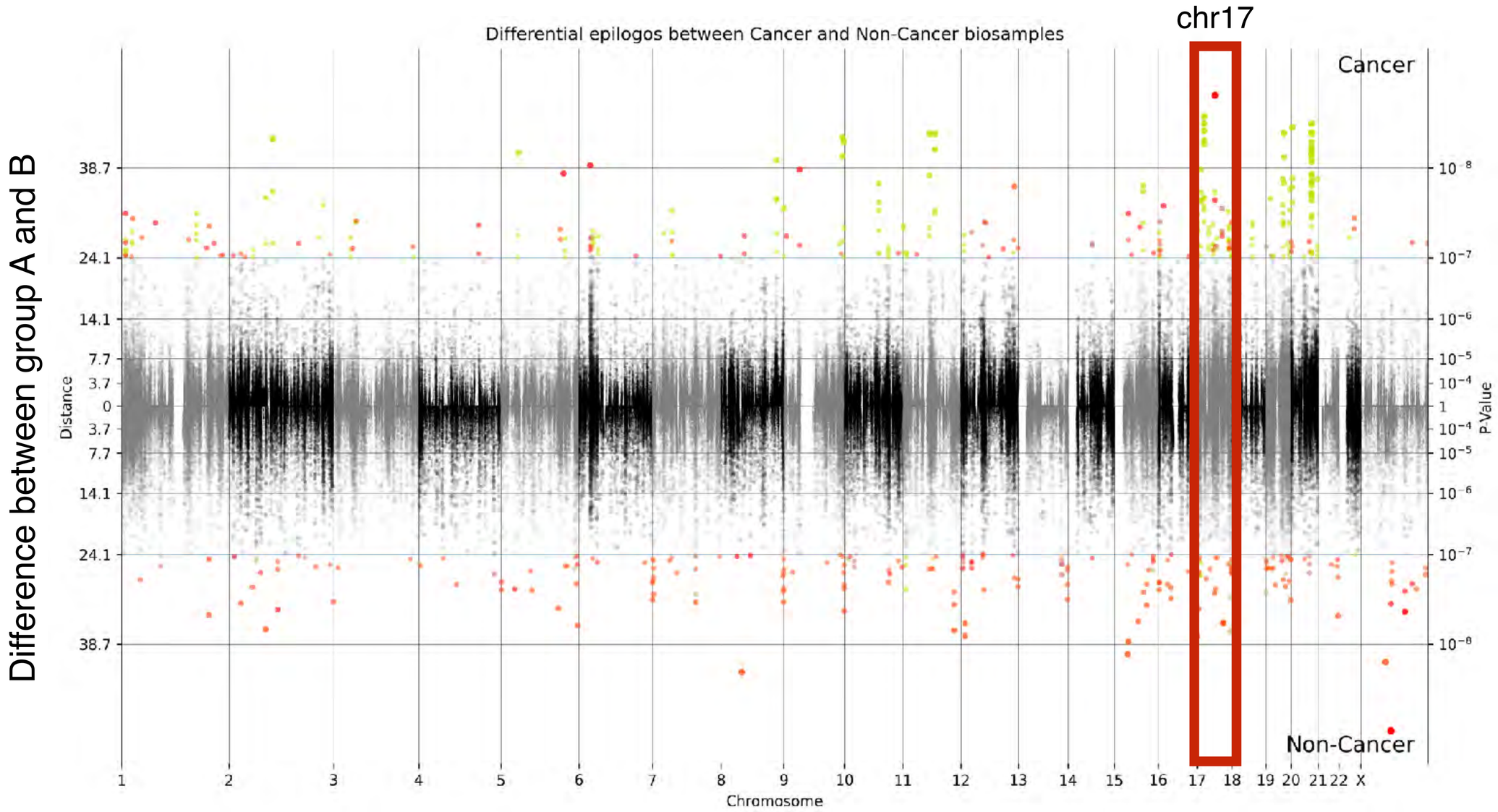
1% FWER

Male vs. female donors (total of 684 biosamples)



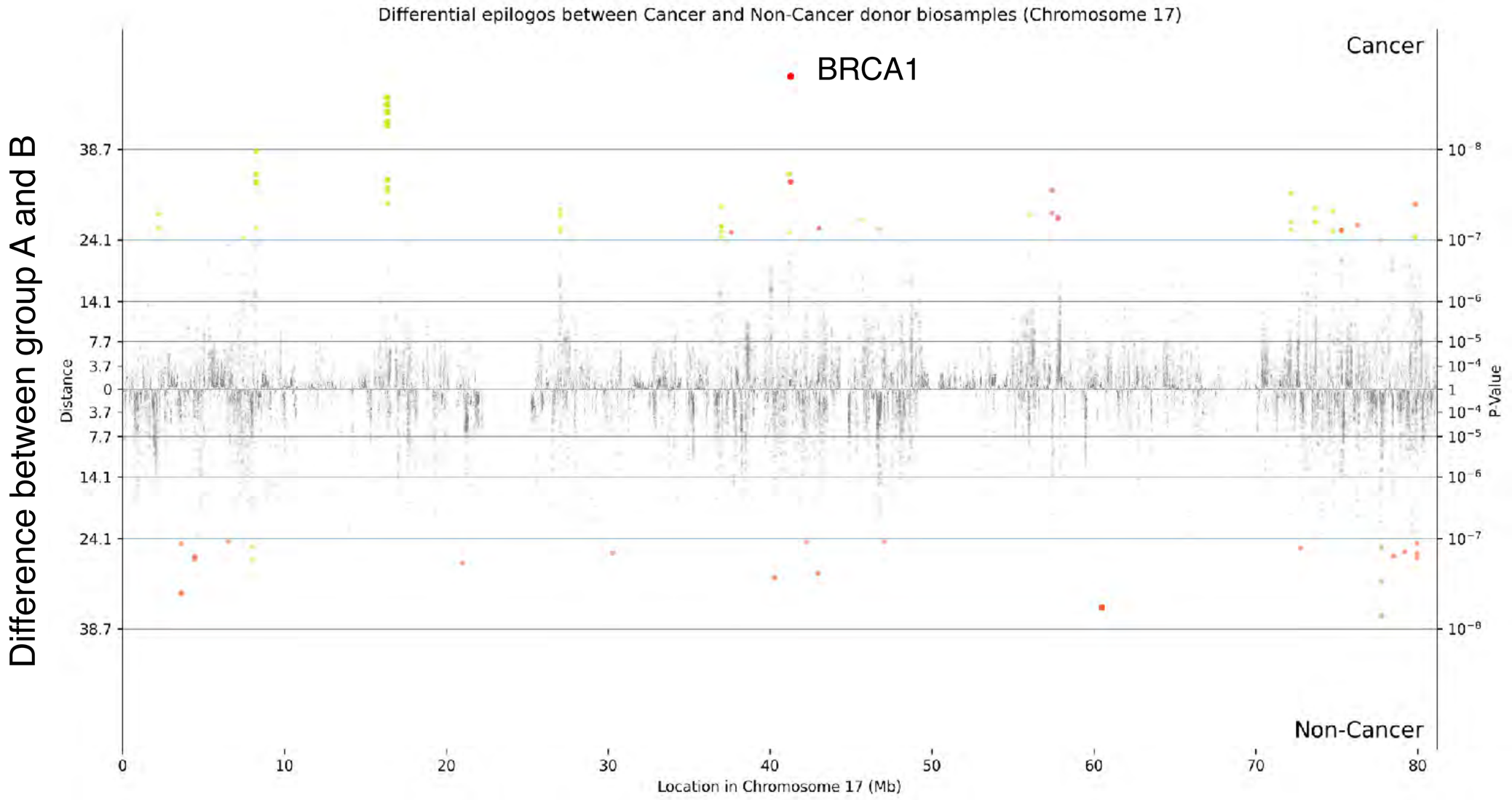
1% FWER

Cancer vs. non-cancer biosamples



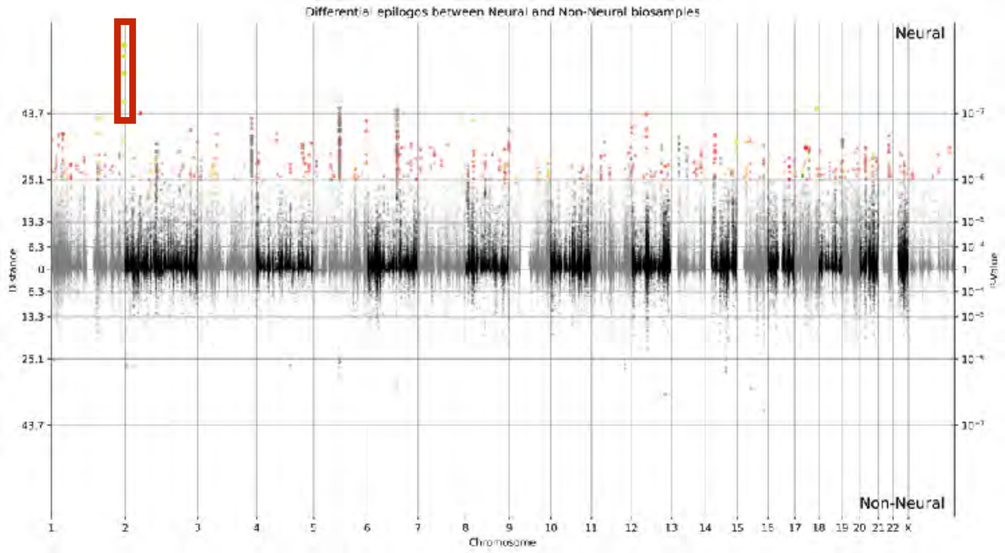
1% FWER

Cancer vs. non-cancer biosamples (chr17)



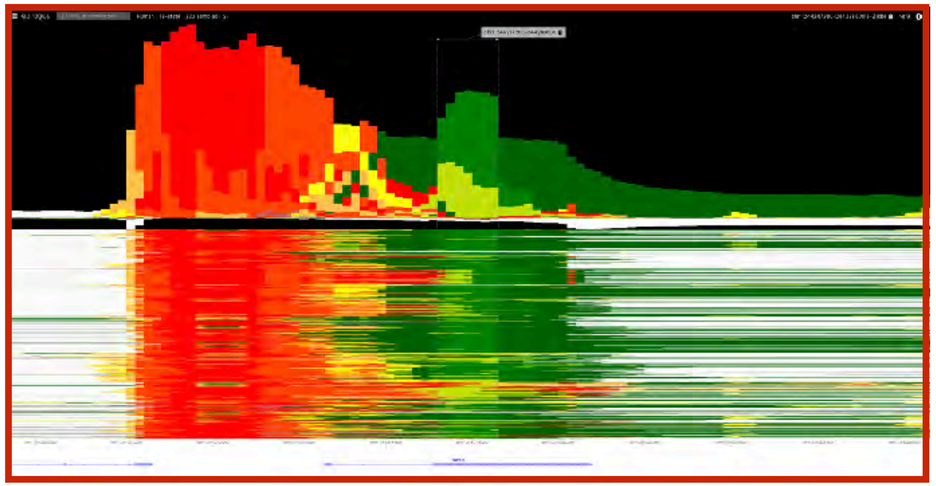
1% FWER

neural vs. non-neural

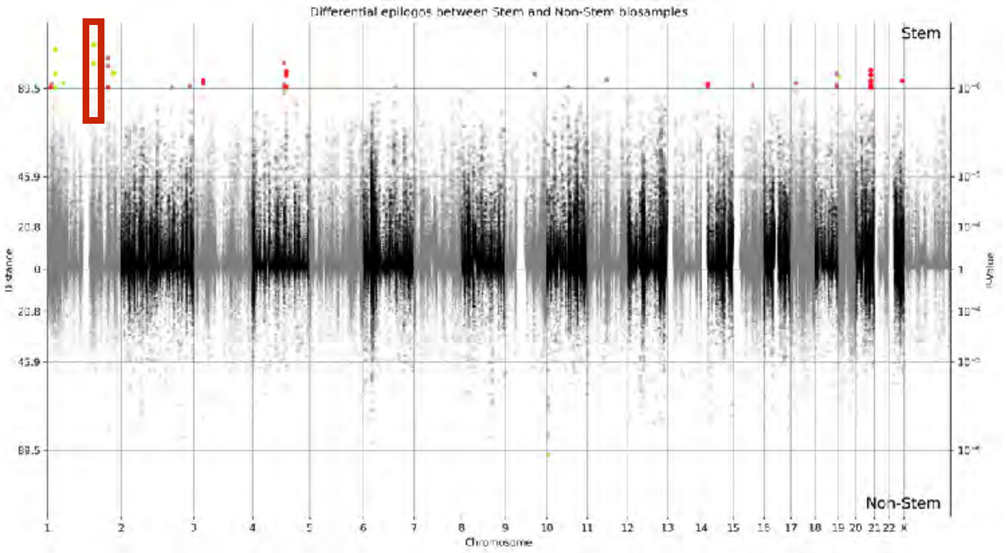


ZBTB18

transcriptional repressor involved in neuronal development

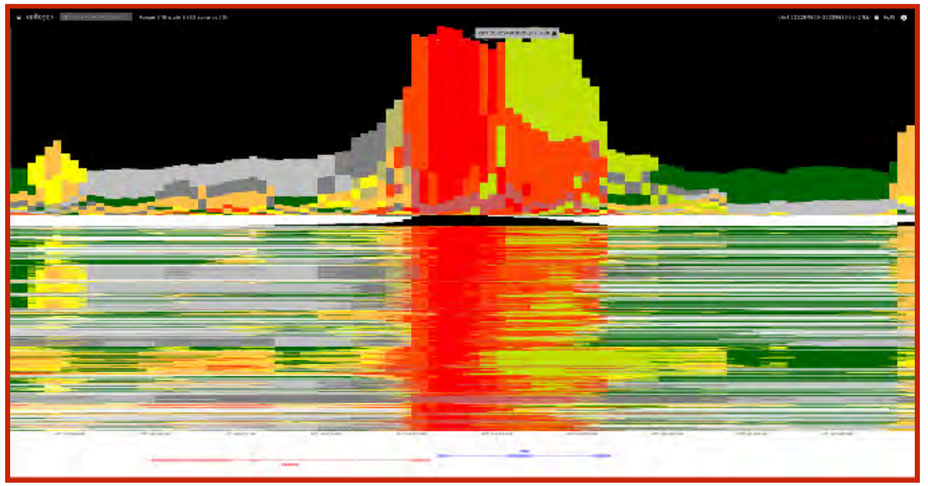


stem cell-like vs. other



BTG2

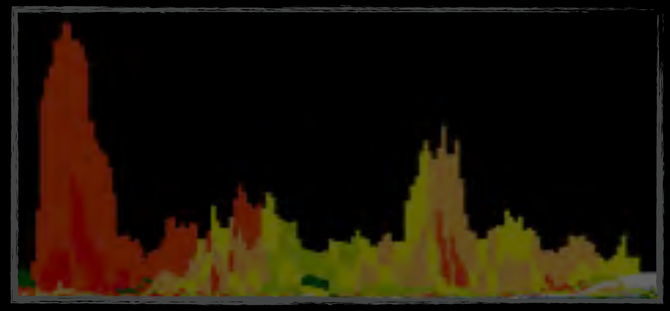
regulation of the G1/S transition of the cell cycle



10% FWER

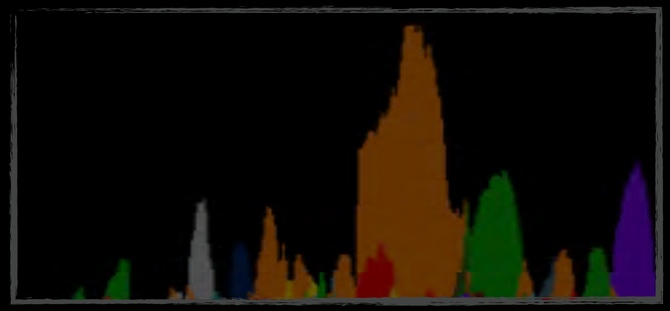
In search of ‘relevance’: two types of genomic annotations

- **Chromatin states (epilogos):**
  “What type of functionality does a genomic region encode?”
  (e.g. promoter, enhancer, repressor)

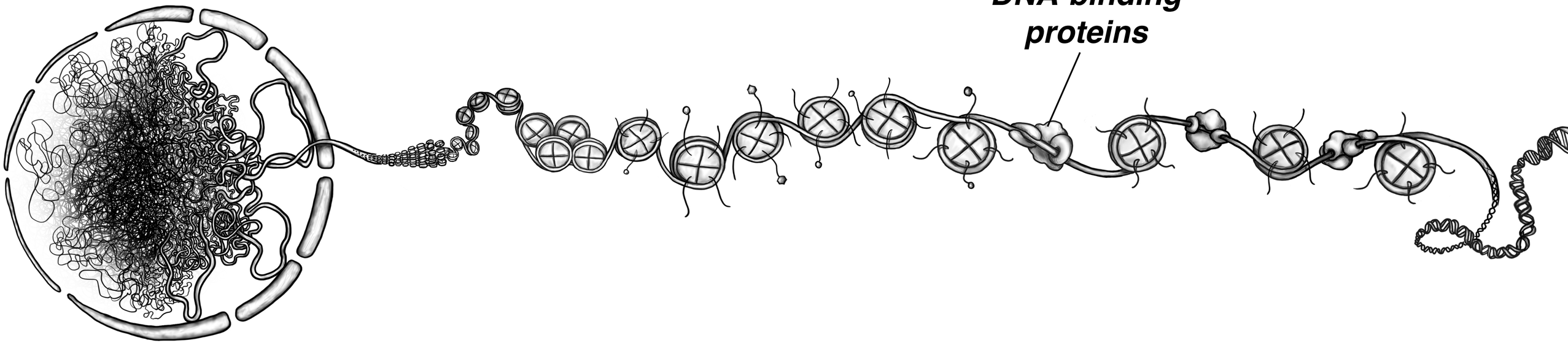


https://epilogos.net

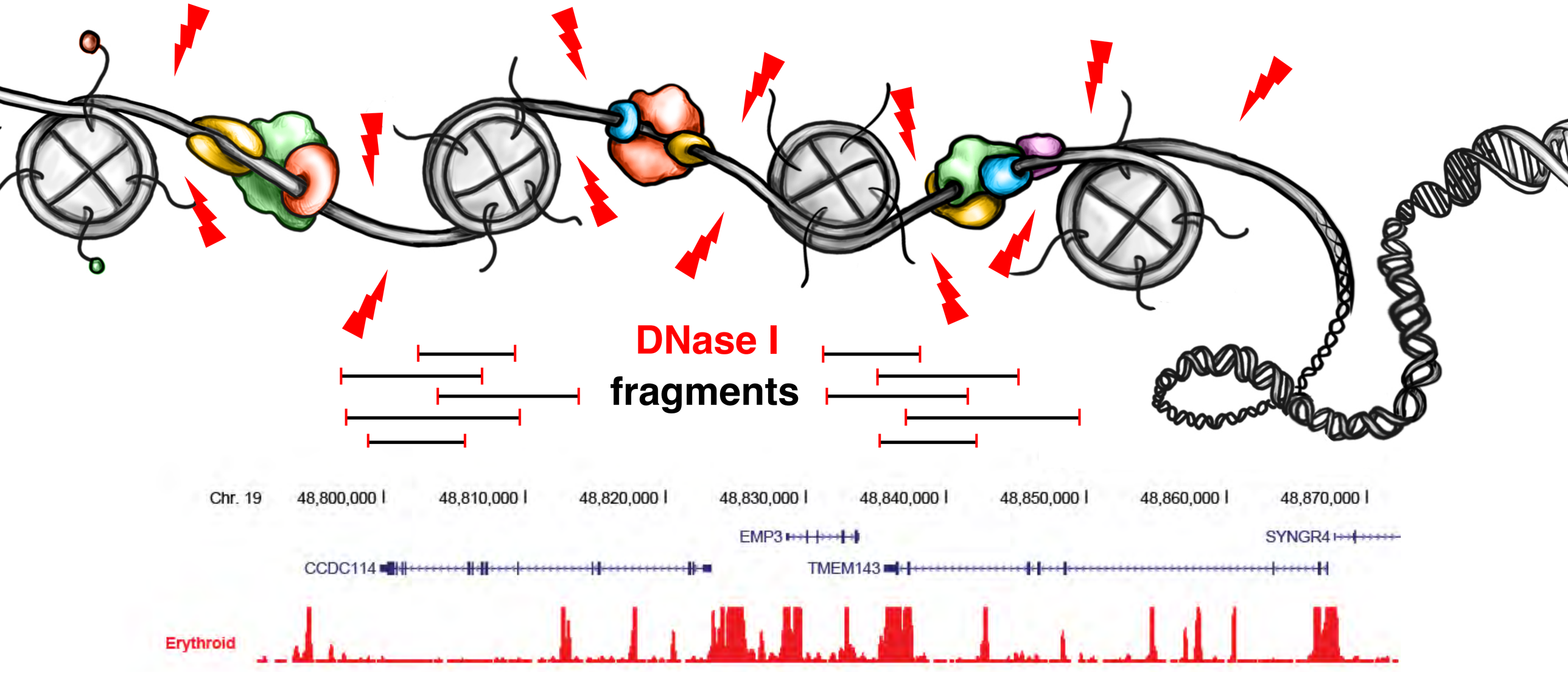
- **Chromatin accessibility (DHS Index)**:
  “In which cellular contexts are regulatory regions utilized?”
  (e.g. cardiac, lymphoid, neural)



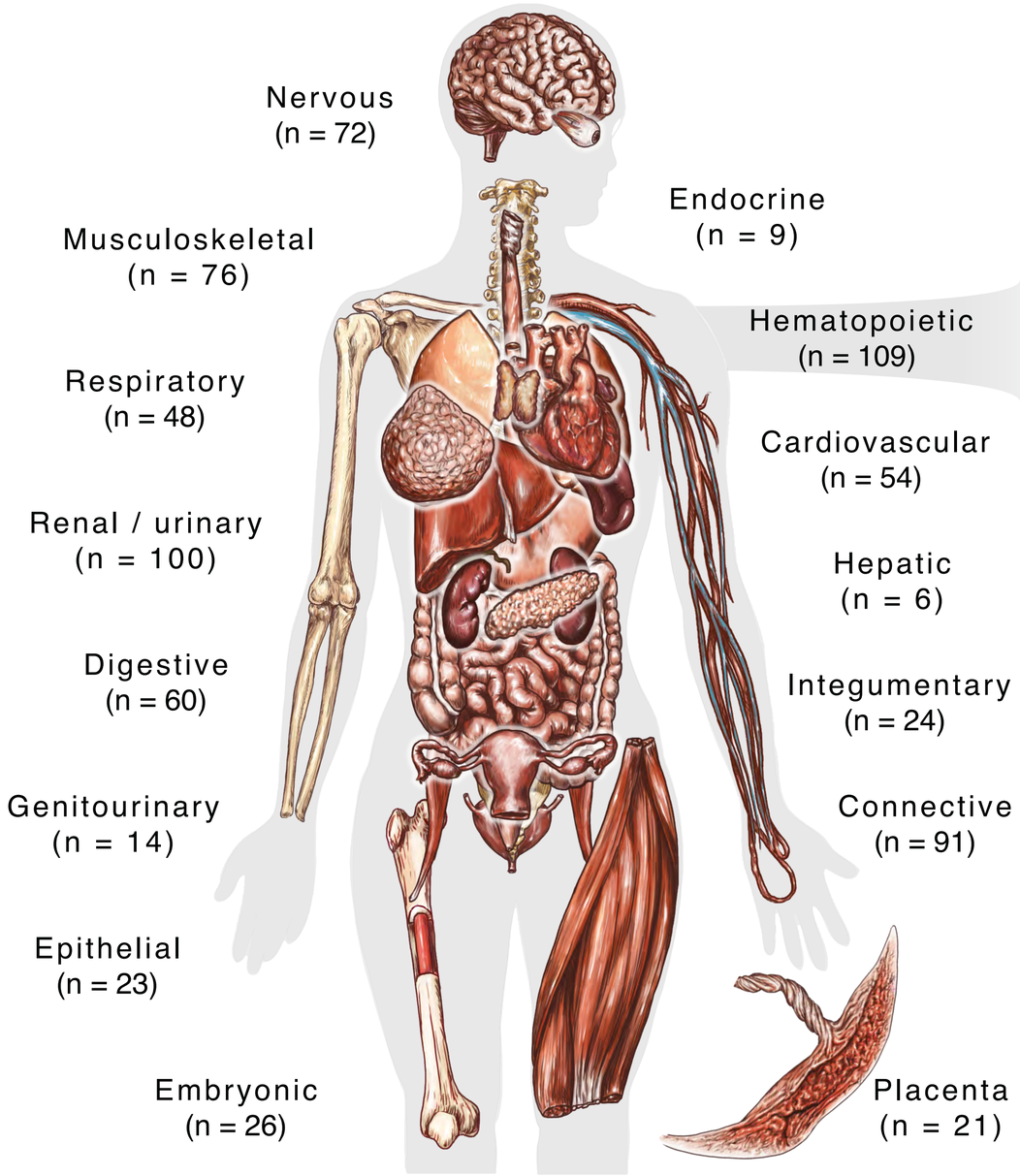
Meuleman *et al.*, 2020 & ongoing

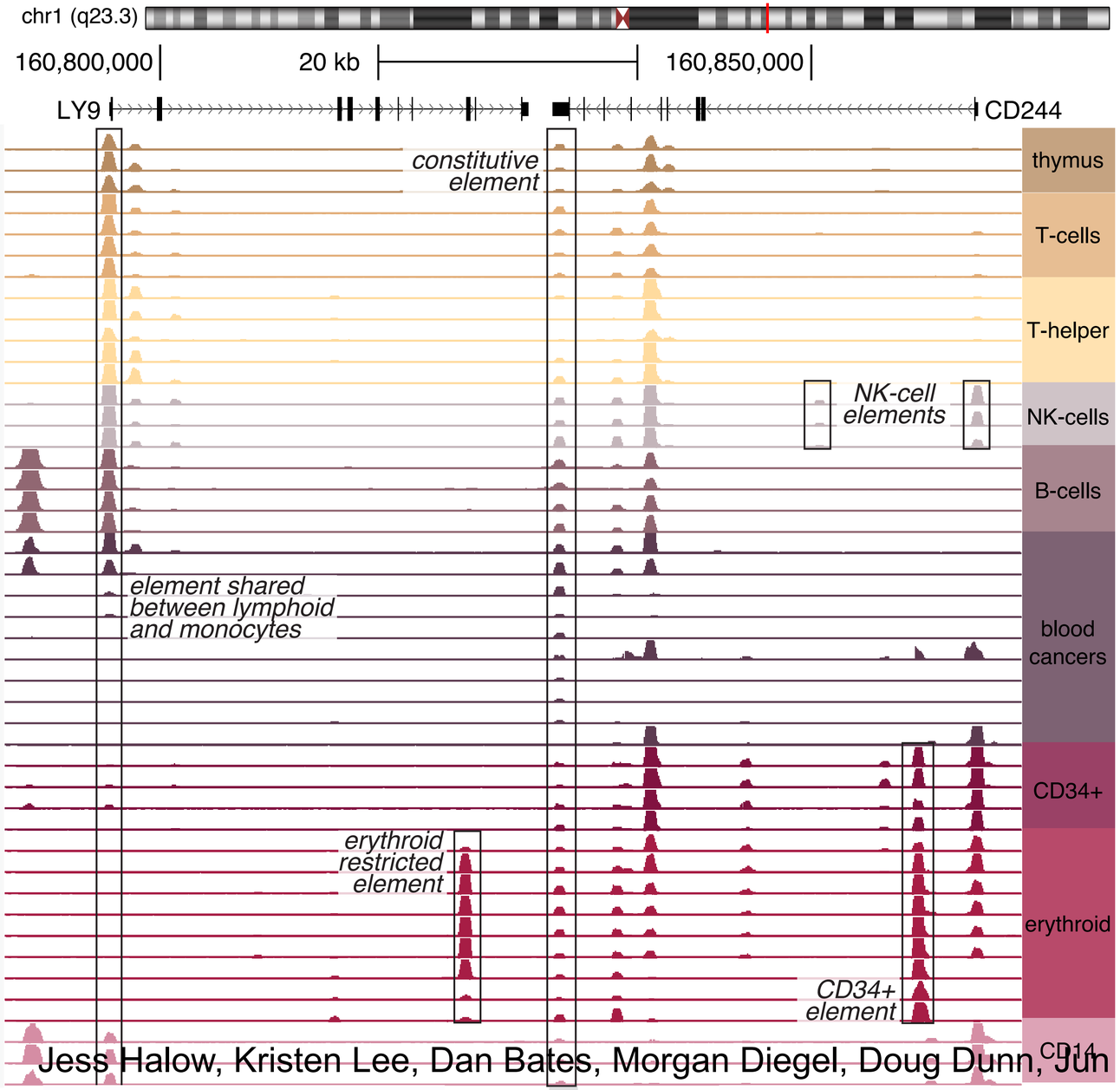

DNA-binding
proteins

The regulatory genome can be mapped using DNase I digestion



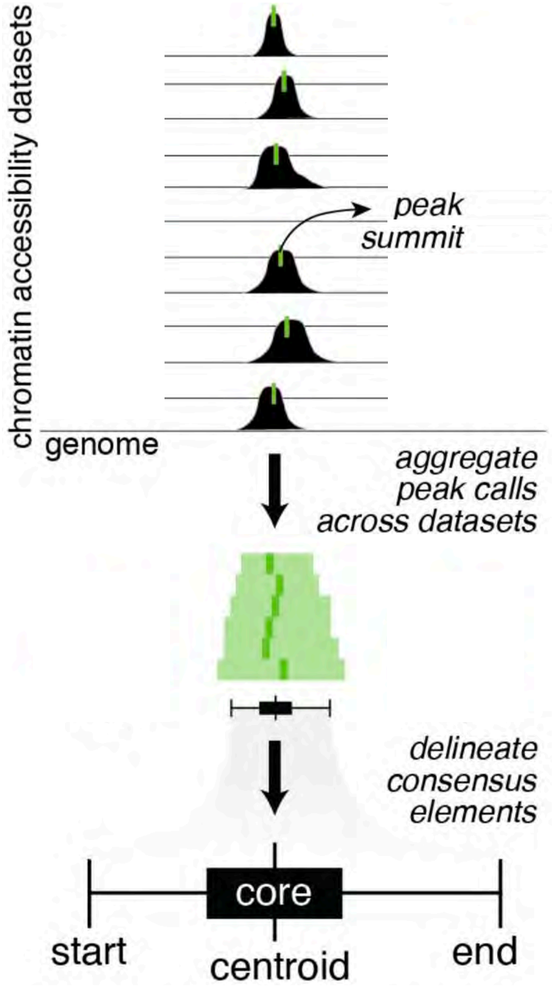
A survey of chromatin accessibility across 400+ cell types and states



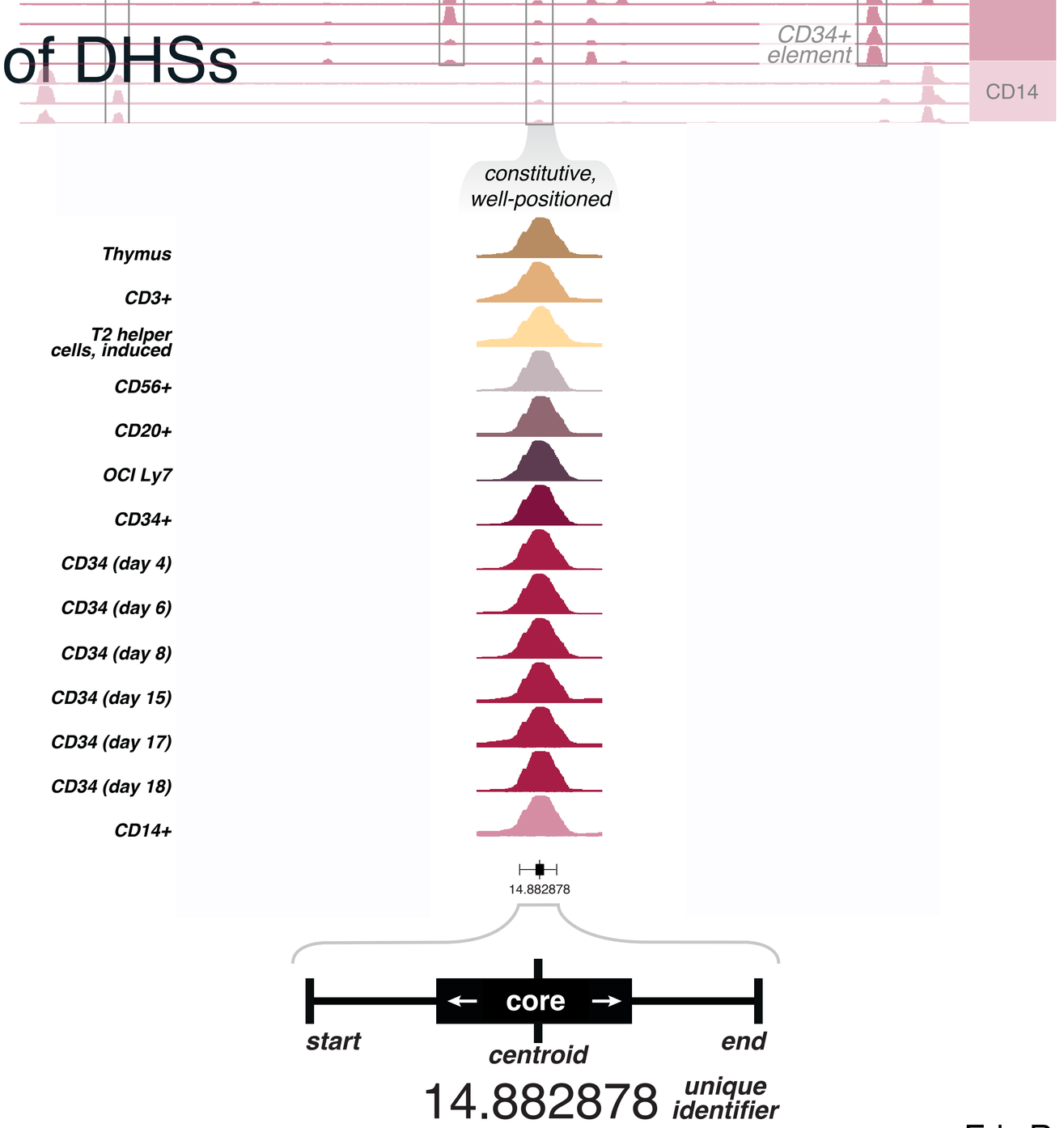


Jess Halow, Kristen Lee, Dan Bates, Morgan Diegel, Doug Dunn, Jun Neri

A common coordinate system of DHSs

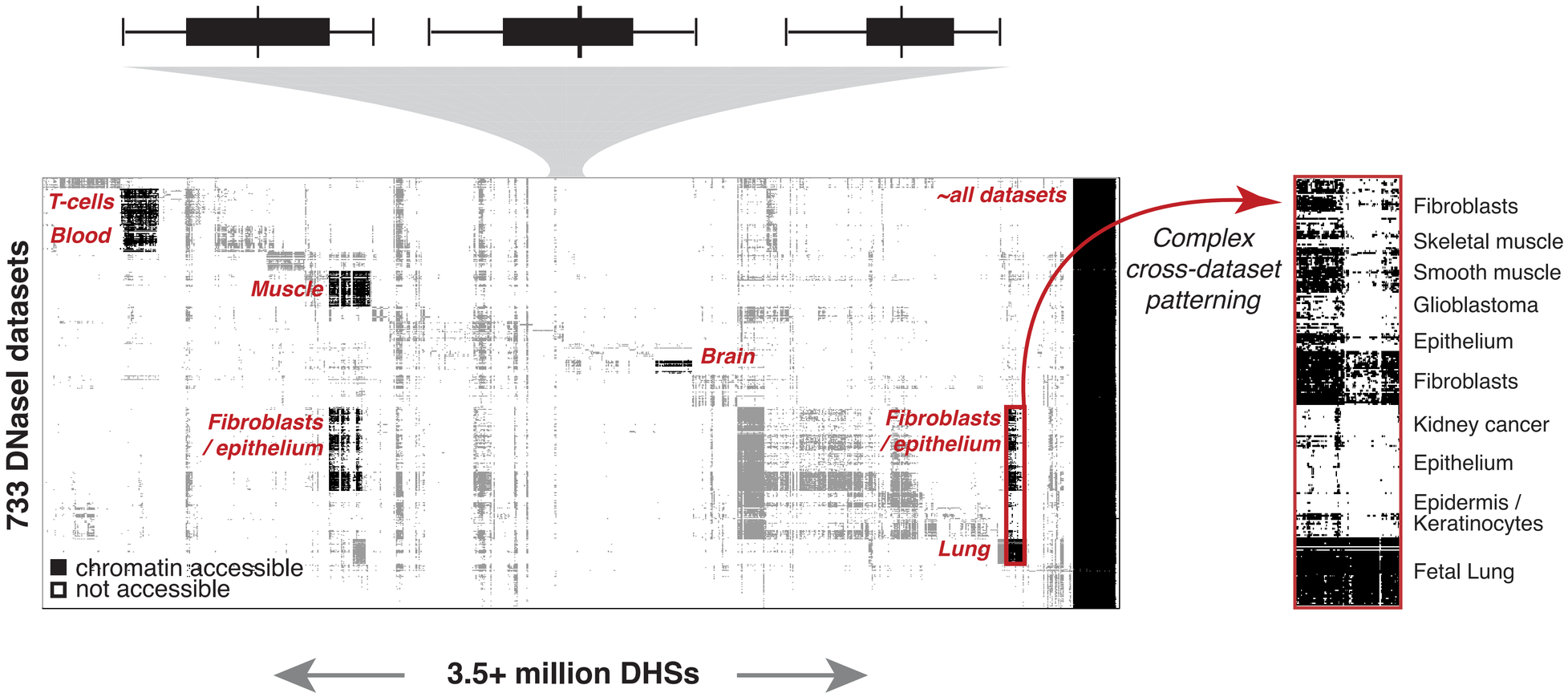


3.5M+ consensus elements



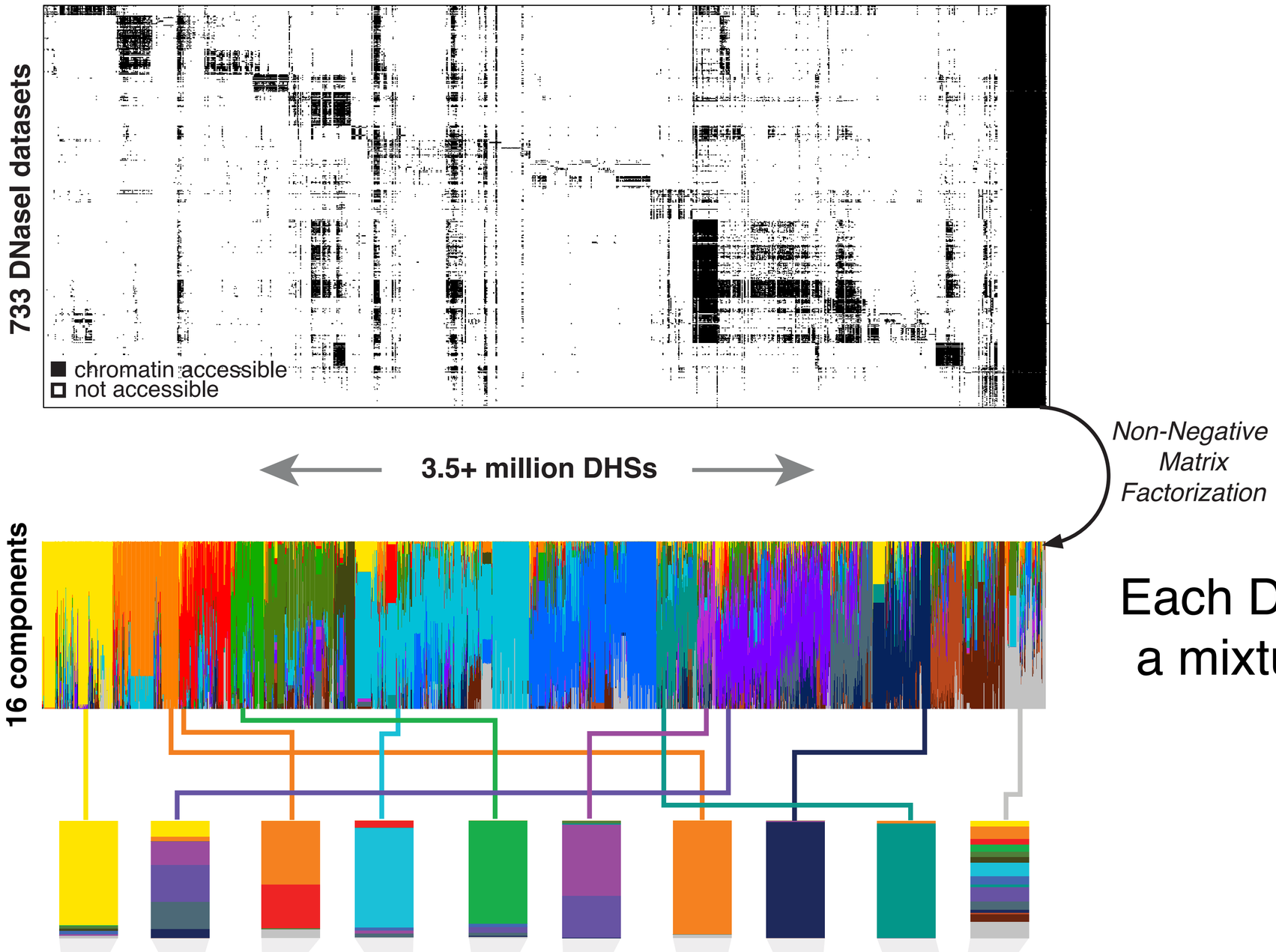
Eric Rynes

Complex DHS patterning across cell types and states



Although generally quite cell type/state selective,
DHSs are often shared between broader cellular contexts

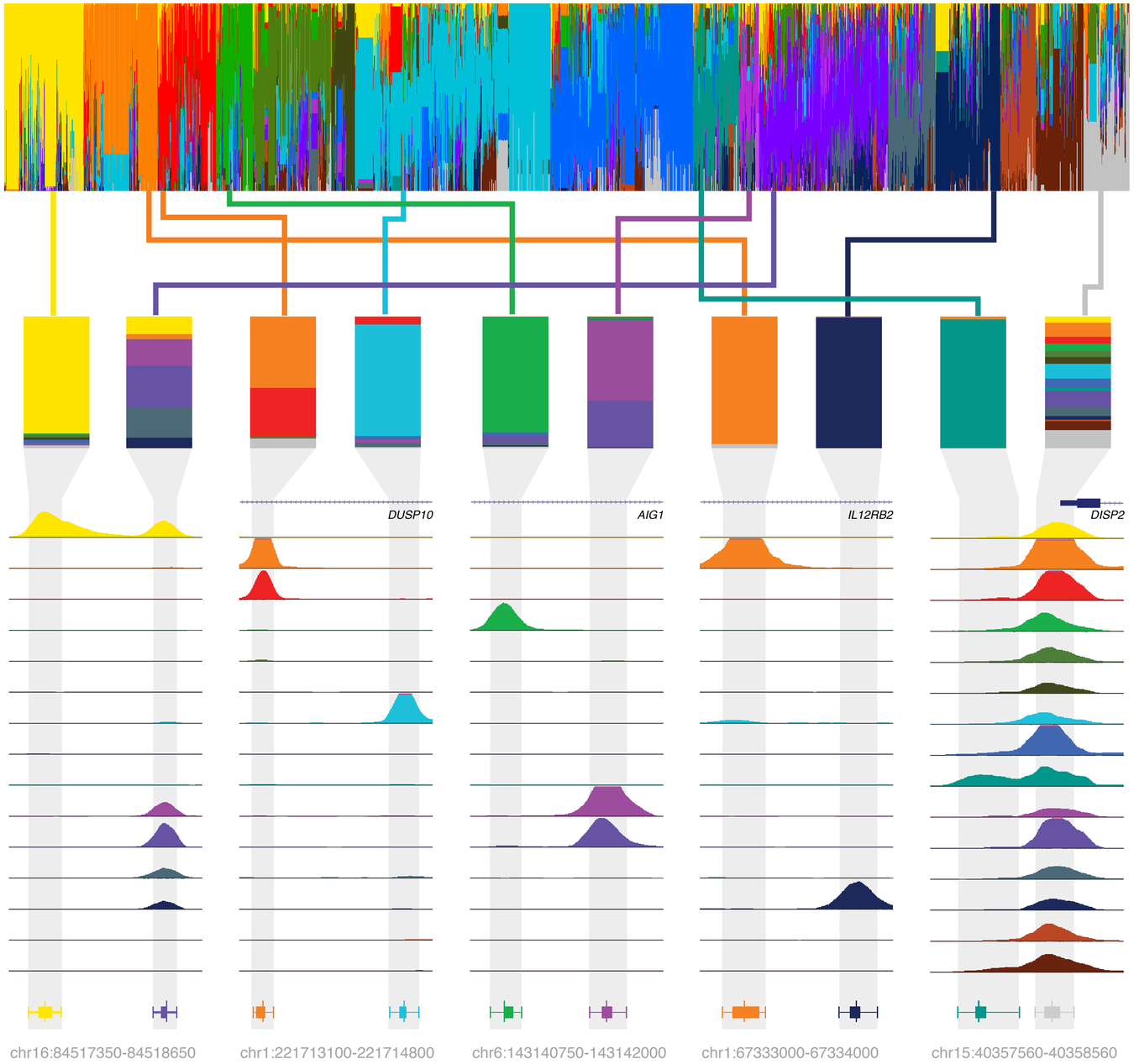
DHS patterns can be decomposed into *components*



Each DHS is described by a mixture of components

Sasha Muratov

Components reflect distinct biological contexts and regulatory signal



Datasets

- Trophoblasts
- CD8 cells
- CD34 cells
- Fetal heart
- Fetal muscle (leg)
- HMVEC
- hESC
- Fetal brain
- Large intestine
- Skin fibroblasts (leg)
- PGP1 fibroblasts
- Renal cell carcinoma
- Esophageal epithelial
- Fetal lung
- Fetal kidney

Motifs

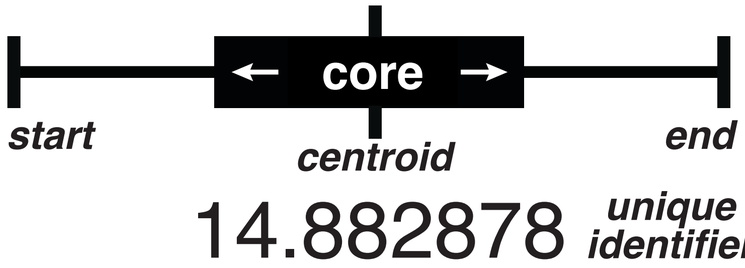
GCM1	
IRF	
ETS/SPI1	
MEF2	
E-box	
ETS/ERG	
Oct-4	
NeuroD	
HNF4A	
AP-1	
AP-1	
HNF1	
TP63	
FOX	
PAX2	

DHS Vocabulary

- **Placental**
- **Lymphoid**
- **Myeloid / erythroid**
- **Cardiac**
- **Musculoskeletal**
- **Vascular / endothelial**
- **Embryonic / primitive**
- **Neural**
- **Digestive**
- **Stromal A**
- **Stromal B**
- **Renal / cancer**
- **Cancer / epithelial**
- **Pulmonary devel.**
- **Organ devel. / renal**
- **Tissue invariant**

DHS Index and Vocabulary: a novel annotation of regulatory DNA

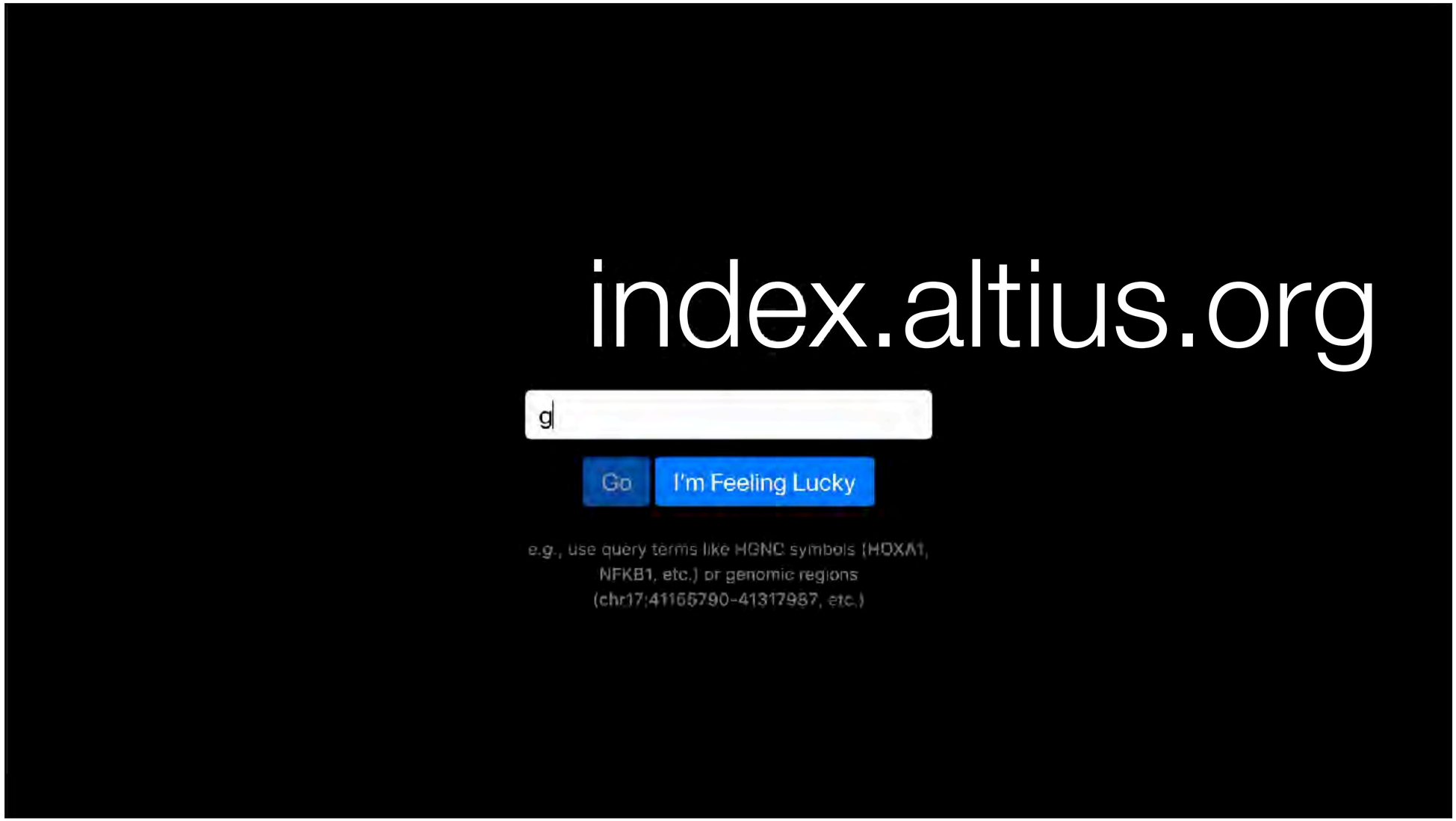
DHS Index



3.5M+ DHSs
Richly annotated and
indexed across cell types

DHS Vocabulary

- Placental
- Lymphoid
- Myeloid / erythroid
- Cardiac
- Musculoskeletal
- Vascular / endothelial
- Embryonic / primitive
- Neural
- Digestive
- Stromal A
- Stromal B
- Renal / cancer
- Cancer / epithelial
- Pulmonary devel.
- Organ devel. / renal
- Tissue invariant

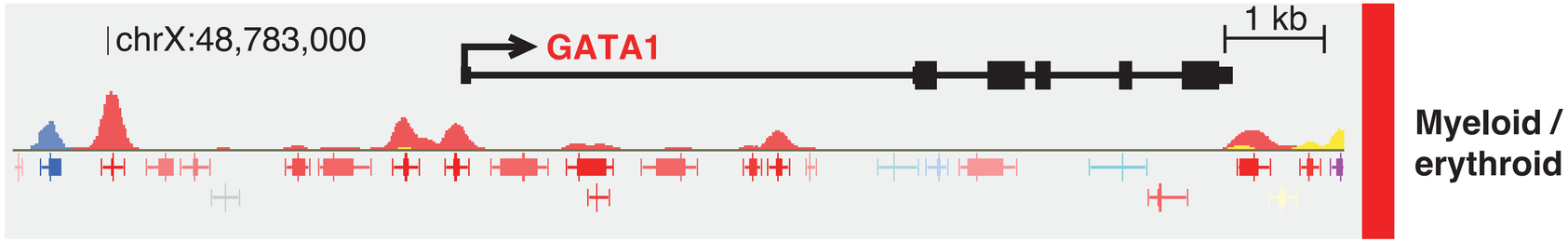


How can we use these data to further identify regions of “relevance”?

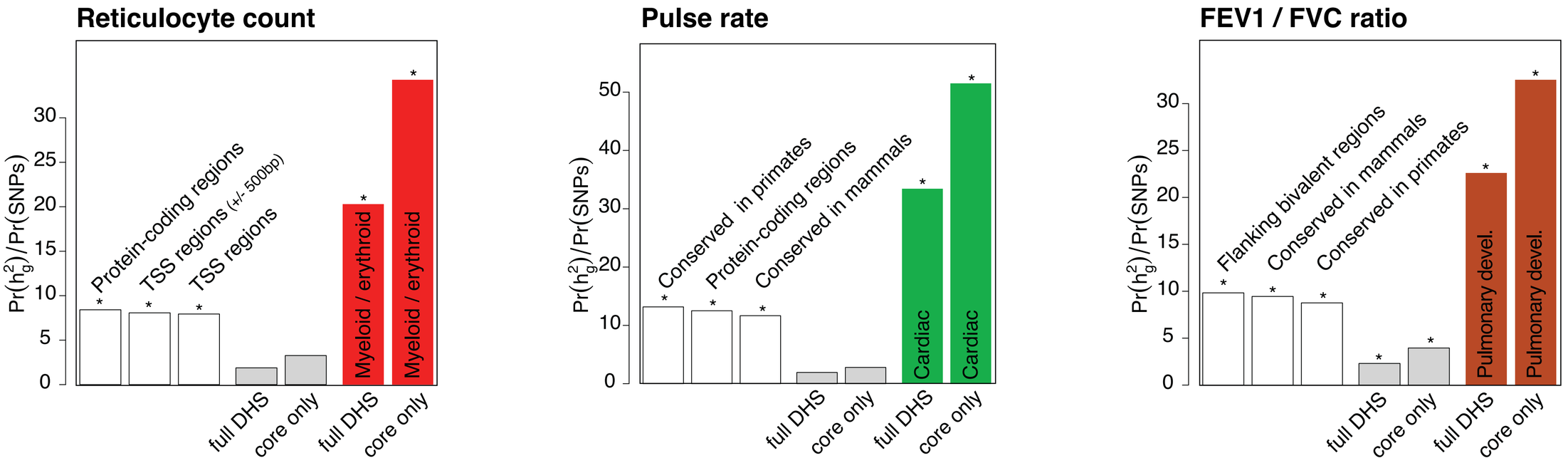
powered by HiGlass

Work with Alex Reynolds

Regions around genes show component-specific patterning



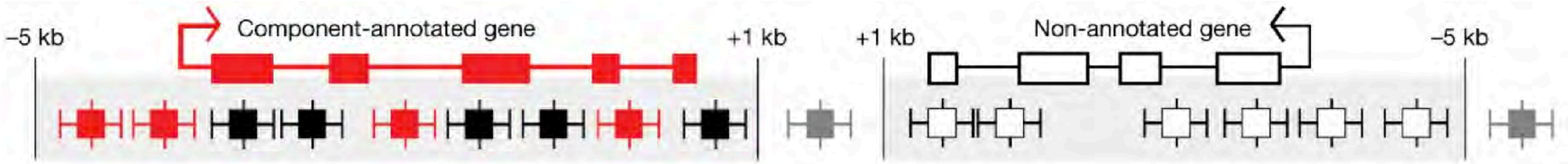
GWAS signal is strongly enriched in relevant component-associated DHSs

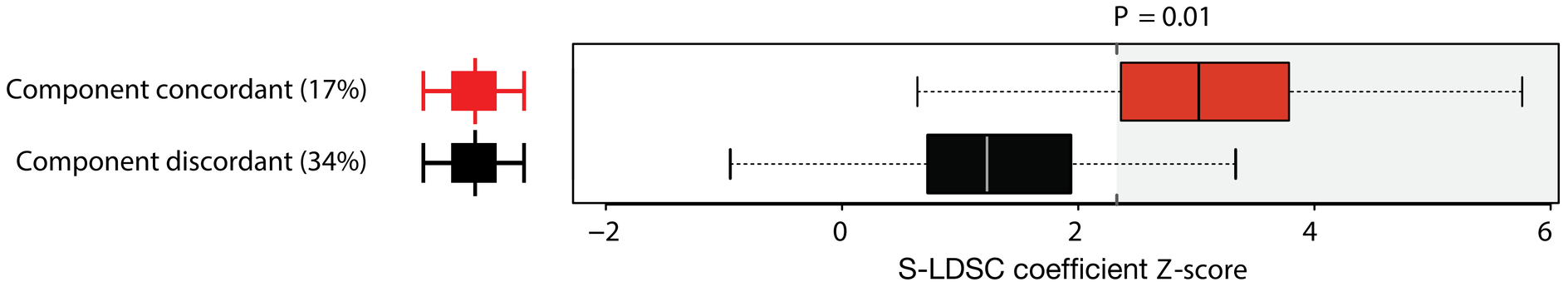


Strong heritability enrichment in relevant DHS components, relative to all DHSs or 85 other genome-wide annotations

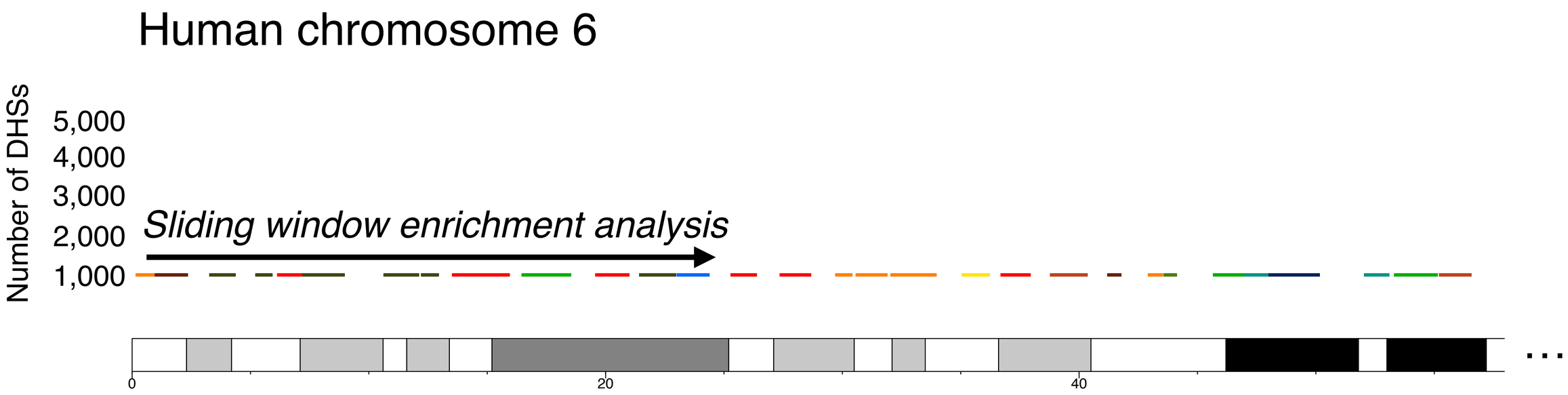
UK Biobank traits
LD score regression with partitioned heritability

GWAS signal is spread across congruently annotated genic DHSs



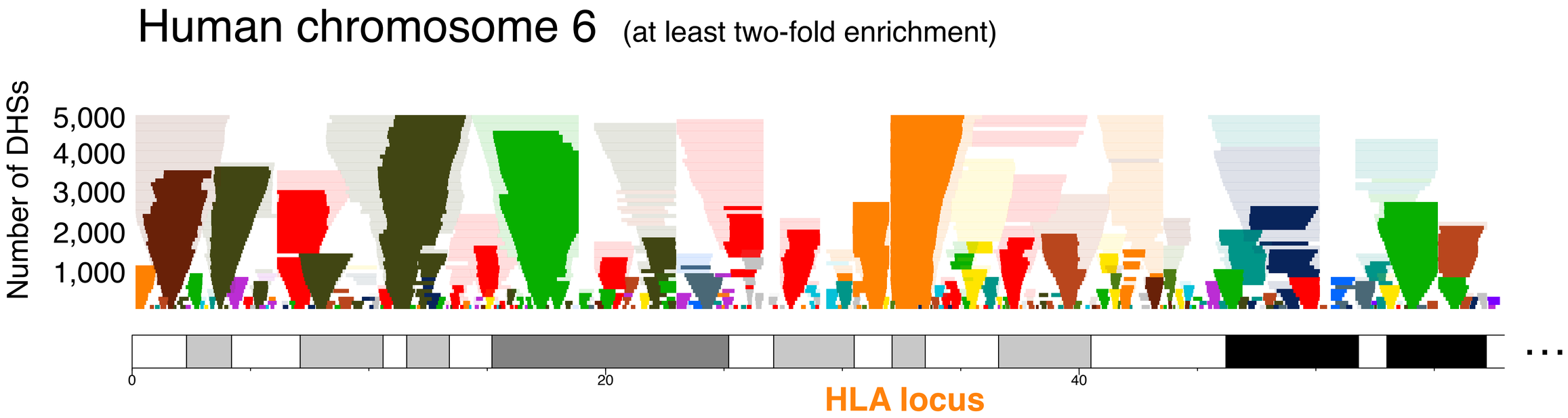


Domain-level organization of component-associated DHSs



Work with Nasi Teodosiadis

Domain-level organization of component-associated DHSs



The human leukocyte antigen (HLA) super-locus is a genomic region on chromosome 6 that encodes genes with important roles in the regulation of the immune system

Work with Nasi Teodosiadis

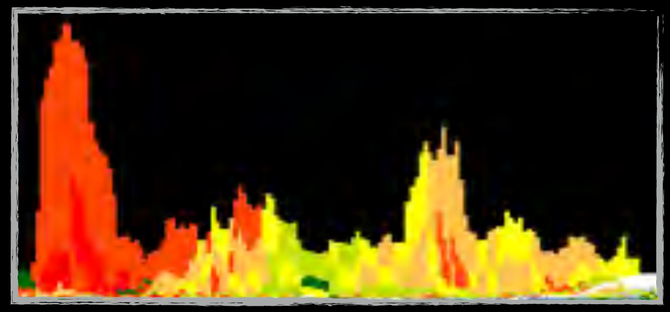
Domain-level organization of component-associated DHSs

100s of multi-DHS “tornado” domains genome-wide

Work with Nasi Teodosiadis

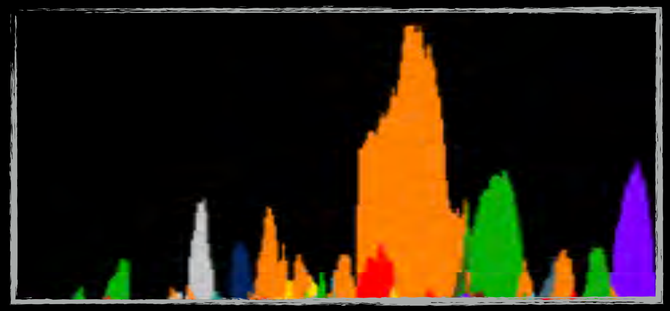
In search of ‘relevance’: two types of genomic annotations

- **Chromatin states (epilogos):**
  “What type of functionality does a genomic region encode?”
  (e.g. promoter, enhancer, repressor)



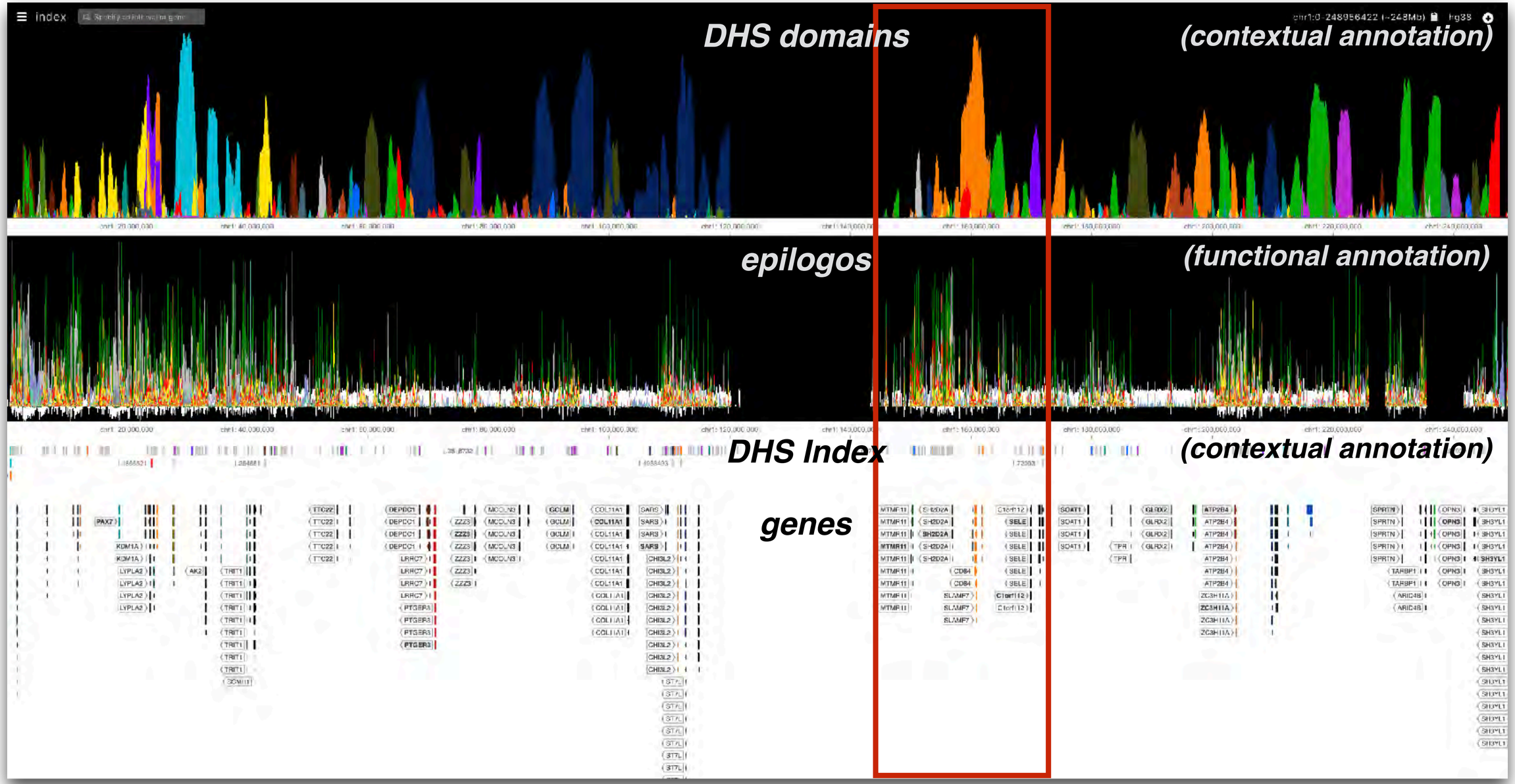
https://epilogos.net

- **Chromatin accessibility (DHS Index)**:
  “In which cellular contexts are regulatory regions utilized?”
  (e.g. cardiac, lymphoid, neural)



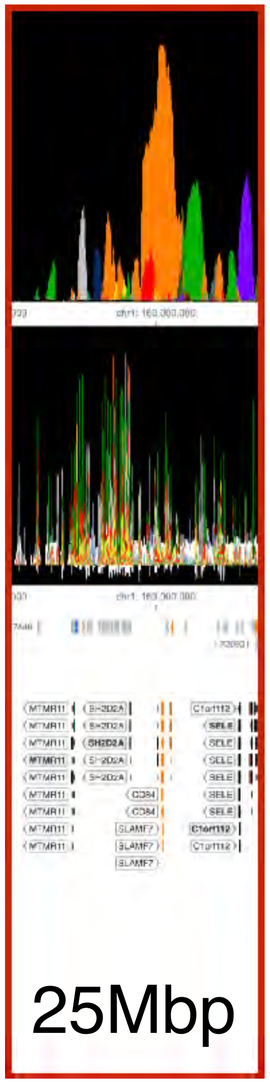
Meuleman *et al.*, 2020 & ongoing

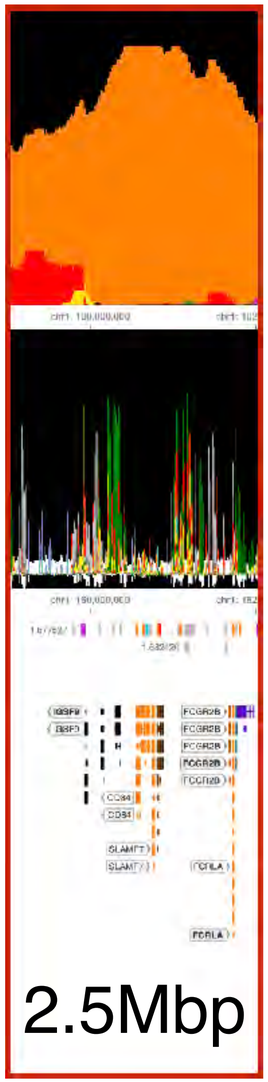
Towards multi-scale regulatory reference annotations

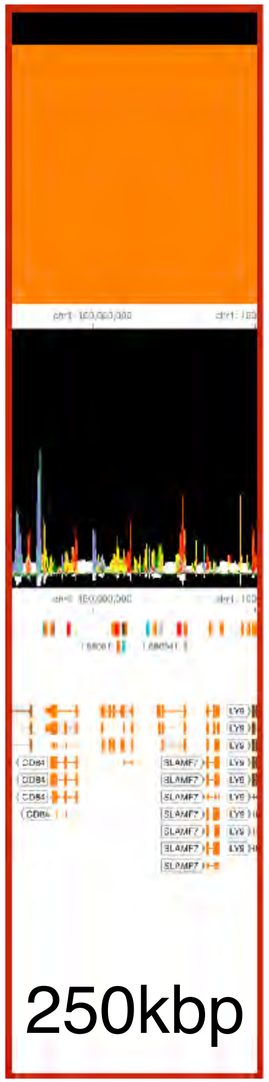


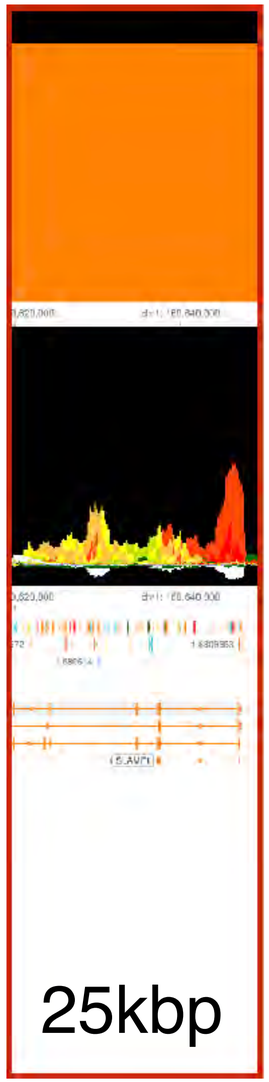
Entirety of human chromosome 1 (~250Mbp)

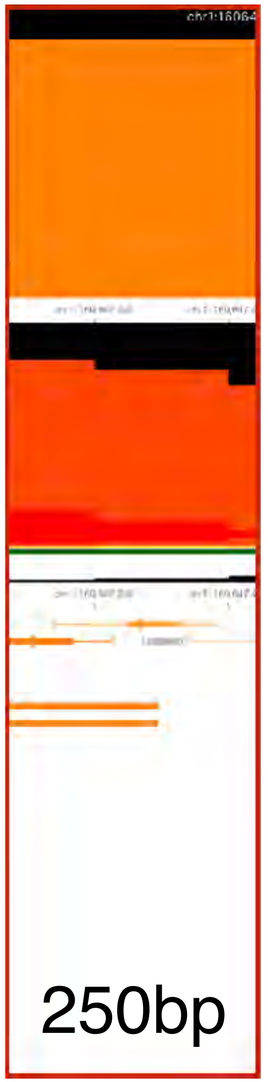
Different annotations for different scales











…and beyond
(footprints, motifs, genetic variation)

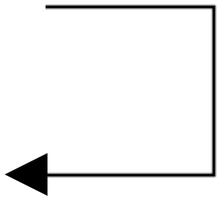
DHS domains

epilogos

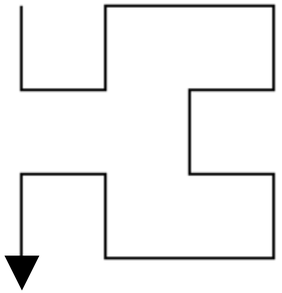
DHS Index

“But where’s my Disney map?”

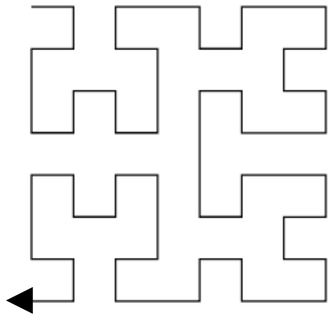
Mapping the linear 1D genome to 2D using Hilbert curves



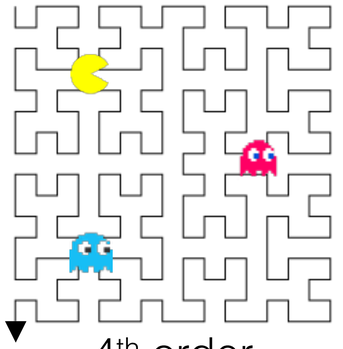
1st order



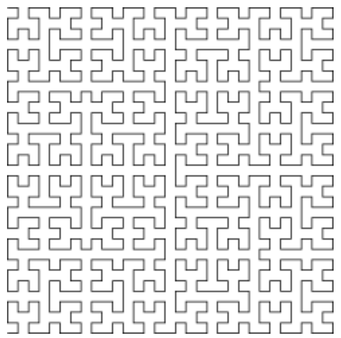
2nd order



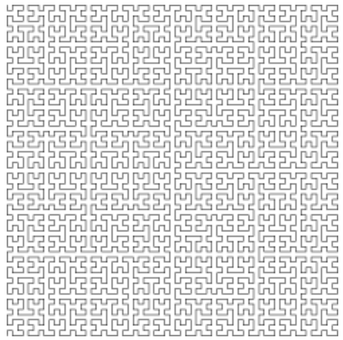
3rd order



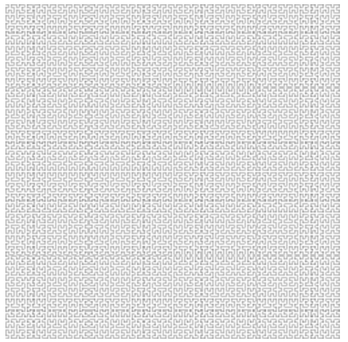
4th order



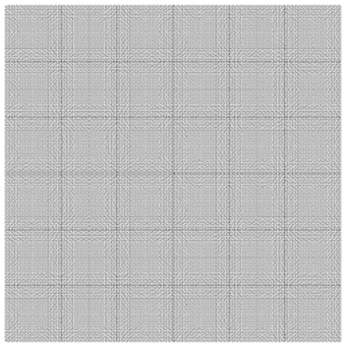
5th order



6th order



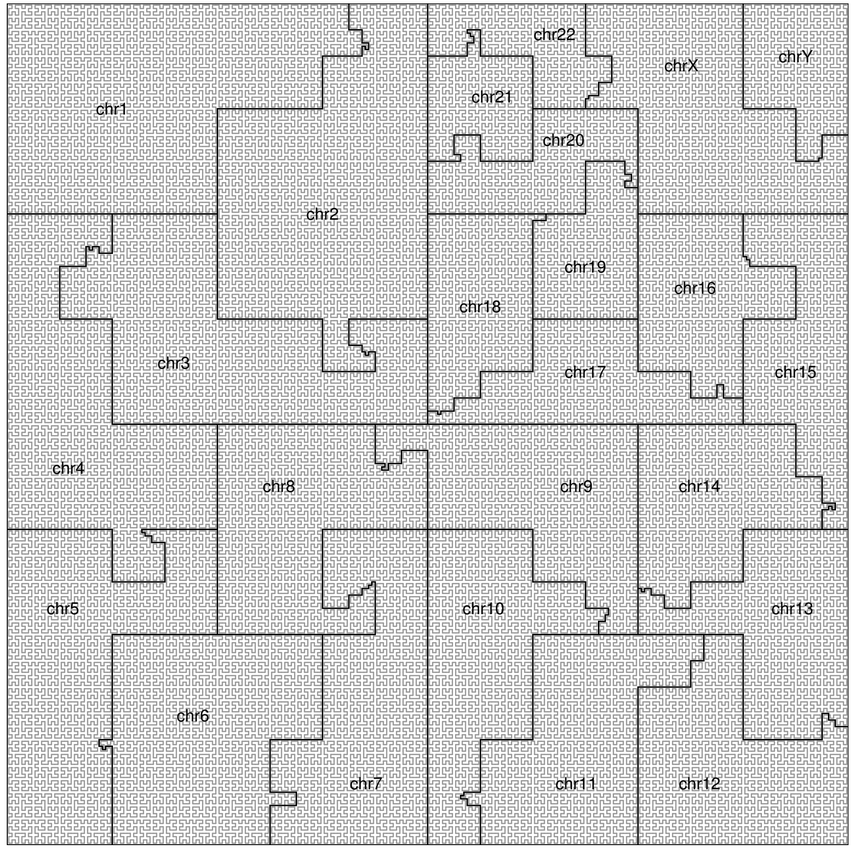
7th order



8th order

Hilbert, Mathematische Annalen (1891)

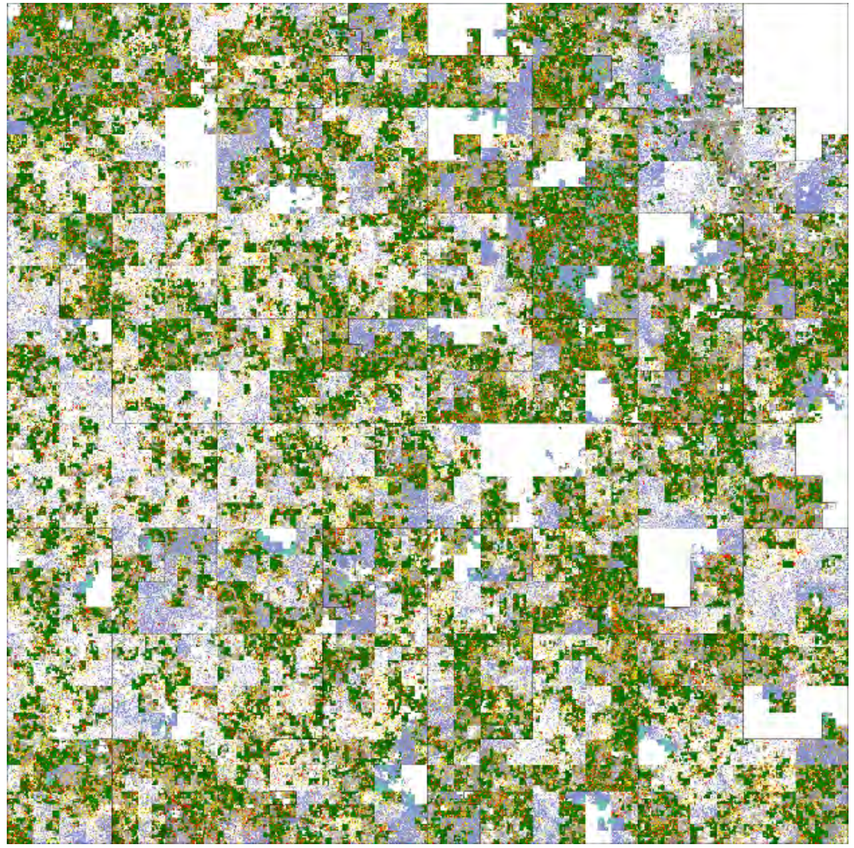
Hilbert curve of the human genome



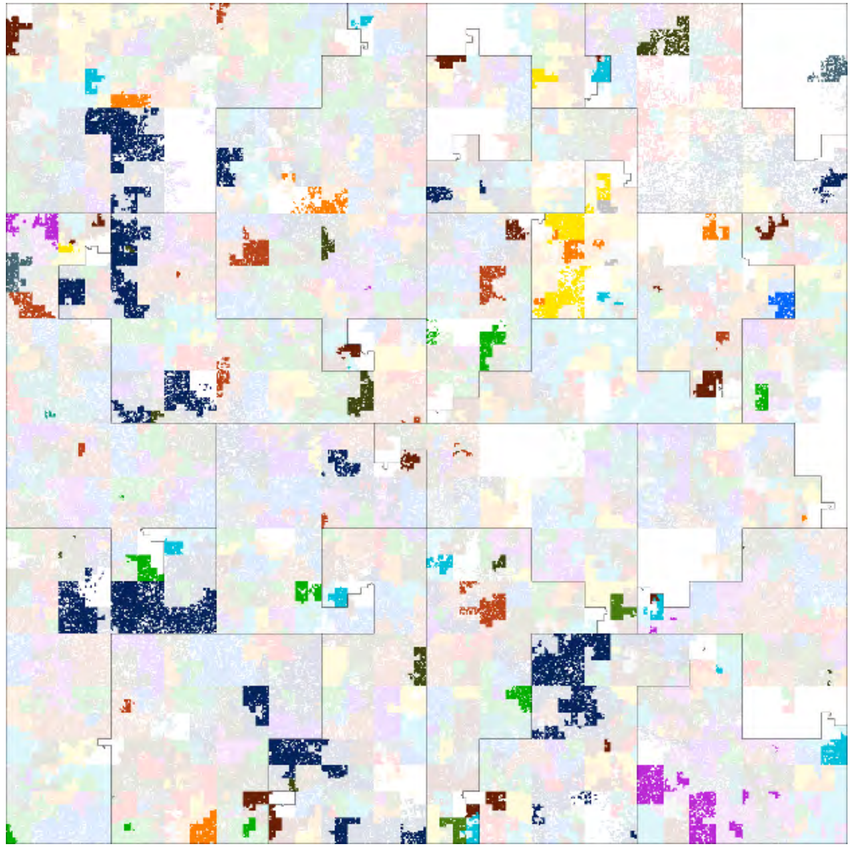
- Regions close in 1D are close in 2D
- The genome provides a (fixed) scaffold to project annotations on
- Full coverage across the genome and not limited by resolution

Gu *et al.*, Bioinformatics (2016)

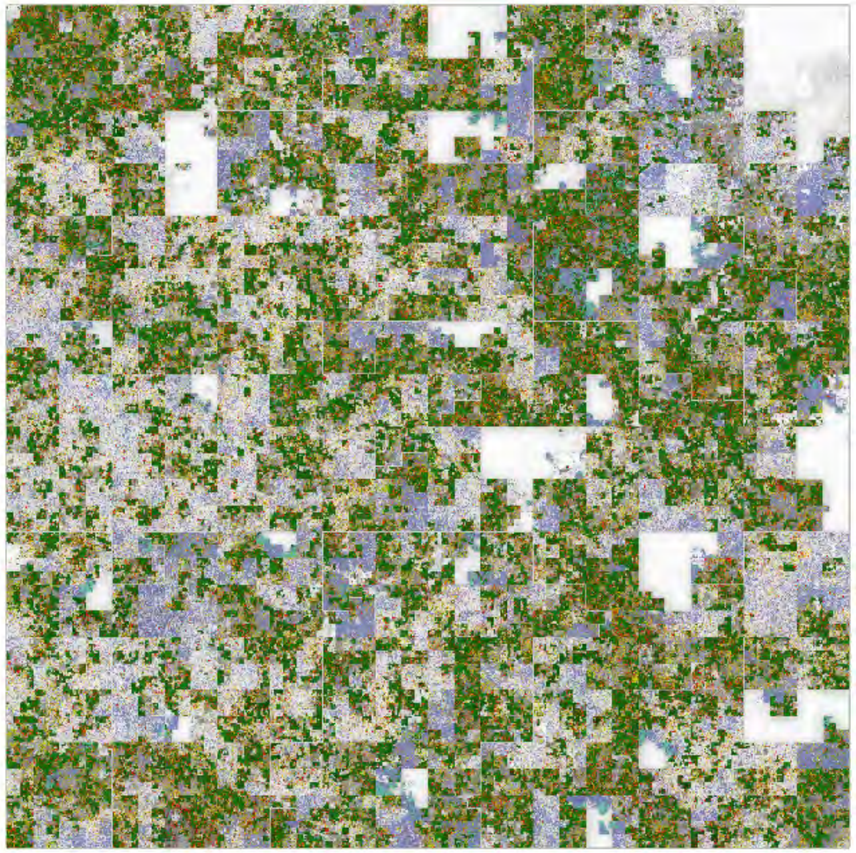
Hilbert curves of **functional** and **contextual** annotations



Functional annotation
(Chromatin states and epilogos)

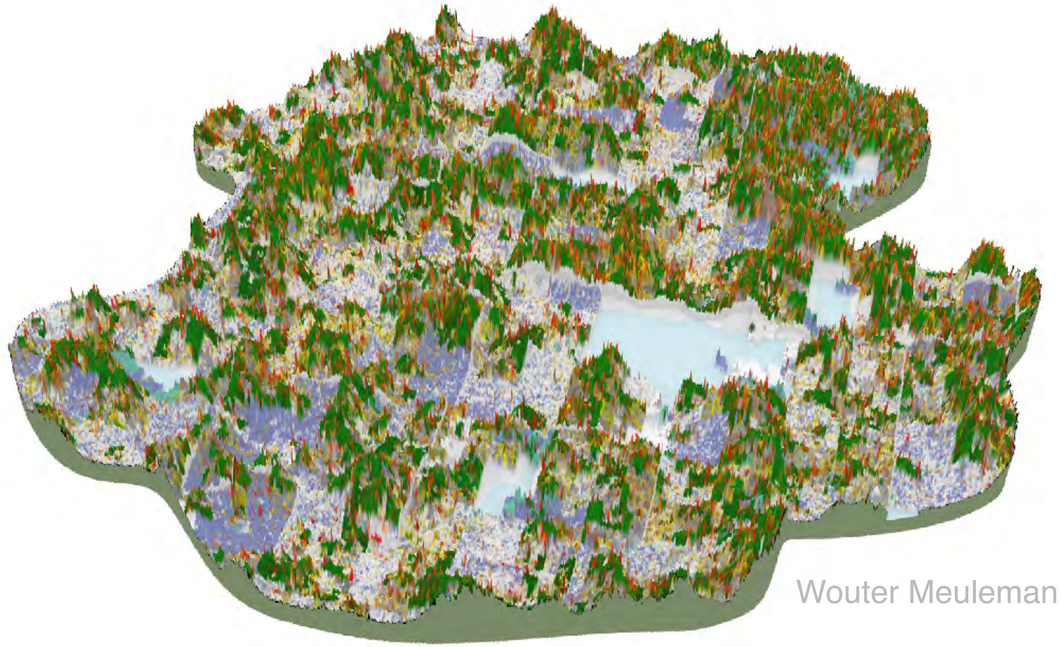


Contextual annotation
(DHS Index and domains)

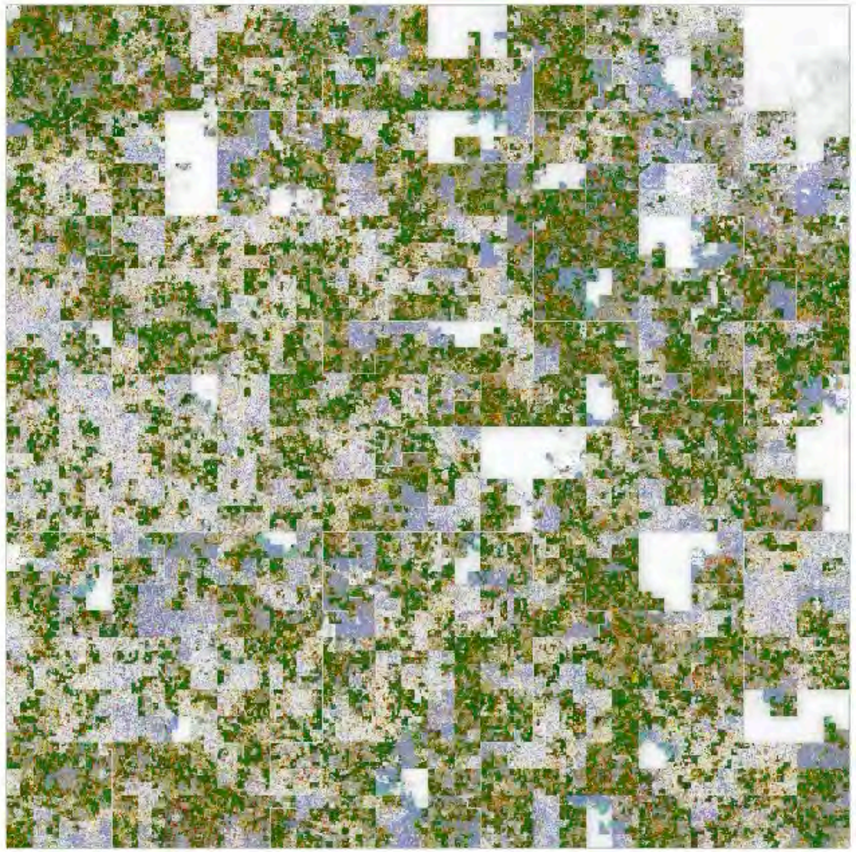


Integrative annotation
(functional+contextual+more)

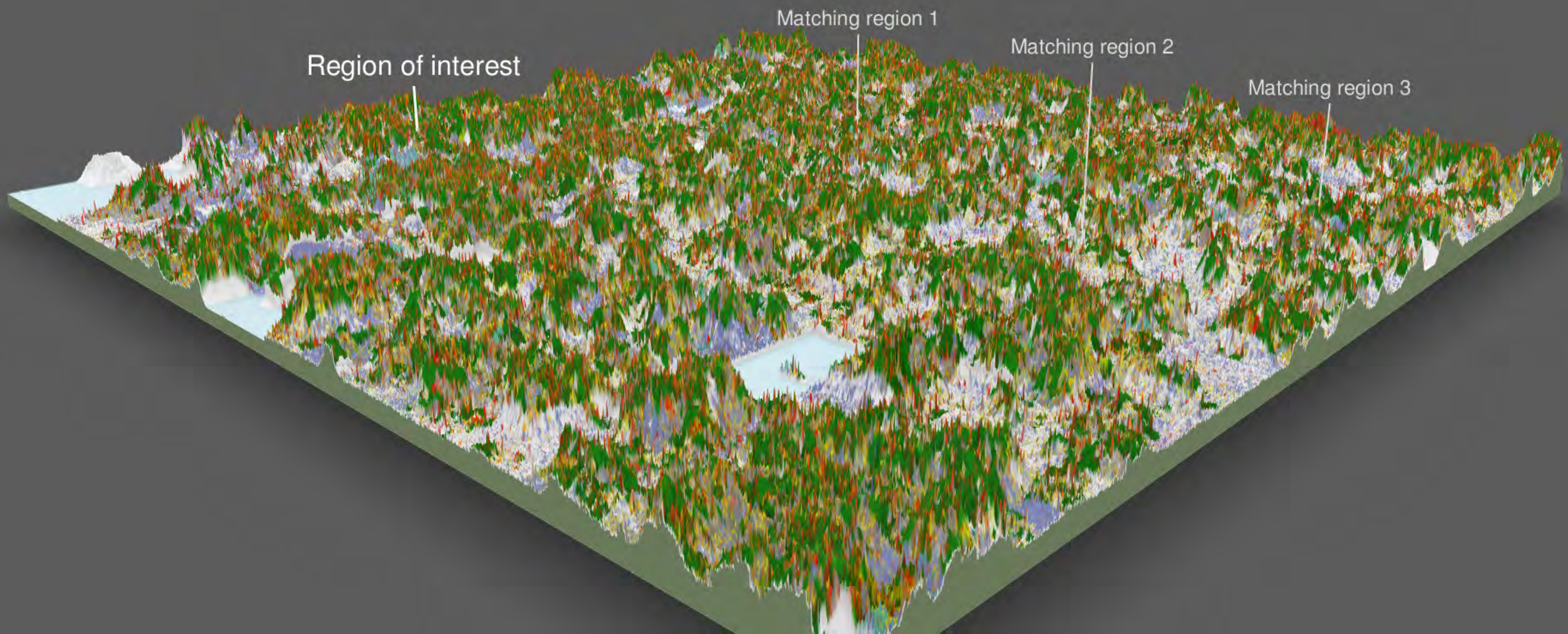
“A Disney map for Genomics”



Maps encourage exploration



Many opportunities for data driven exploration of these maps



“Man ~~versus~~ ***with*** Machine”

- Show only “relevant” information to humans…
- …while machines provide full data-driven guidance
- Human decisions get *augmented* by machines

We need better Navigation Systems 

I consider these efforts part of a new field: "**augmented genomics**", in which the work of genome scientists is supplemented – not replaced! – by data-driven machine intelligence

@nameluem
www.meuleman.org/**hiring**

Now hiring!

Looking for students, postdocs and alternatively experienced folks

Are you curious about the regulatory genome and how it is organized in a cell nucleus? Do you have affinity with squeezing information out of large datasets? Want to have an impact in next generation regulatory annotations?

Positions available immediately, until filled

National Human Genome Research Institute

@nameluem
www.meuleman.org/**hiring**

Acknowledgements

Data analysis
Sasha Muratov
Eric Rynes
Alex Reynolds
Jacob Quon
Nalu Tripician
Nasi Teodosiadis
Eric Haugen

Miscellaneous
Chad Lundberg
Rae Senarighi
Tim Mercer
Jeff Vierstra
John Stam.

Data generation
Jessica Halow
Kristen Lee
Daniel Bates
Morgan Diegel
Douglass Dunn
Fidencio Neri

Data coordination
Richard Sandstrom
Audra Johnson
Jemma Nelson
Mark Frerker
Michael Buckley
Rajinder Kaul

National Human Genome Research Institute

@nameluem
www.meuleman.org/**hiring**