Overview of Genome Annotation & ENCODE Elements

M Gerstein
on behalf of

Slides freely downloadable from
Lectures.GersteinLab.org
& “tweetable” (via @markgerstein).
See last slide for references & more info.

How might we annotate a human text?

Color is Function

Lines are Similarity

[B Hayes, Am. Sci. (Jul.- Aug. '06)]

The Semicolon Wars

Brian Hayes

If you want to be a thorough-going world traveler, you need to learn 6,912 ways to say "Where is the toilet, please?" That's the number of languages known to be spoken by the peoples of planet Earth, according to Ethnologue.com.

If you want to be the complete polyglot programmer, you also have quite a challenge ahead of you, learning all the ways to say:

```
printf("hello, world\n");
```

(This one is in C.) A catalog maintained by Bill Kinnersley of the University of Kansas lists about 2,500 programming languages. Another survey, compiled by Diarmuid Piggott, puts the total even higher, at more than 8,500. And keep in mind that whereas human languages have had millennia to evolve and diversify, all the computer languages have sprung up in just 50 years. Even by the more-conservative standards of the Kinnersley count, that means we've been inventing one language a week, on average, ever since Fortran.

For ethnologists, linguistic diversity is a cultural resource to be nurtured and preserved, much like biodiversity.

Every programmer knows there is one true programming language. A new one every week

a good-enough notation—for expressing an algorithm or defining a data structure.

There are programmers of my acquaintance who will dispute that last statement. I expect to hear from them. They will argue—zealously, ardently, vehemently—that we have indeed found the right programming language, and for me to claim otherwise is willful ignorance. The one true language may not yet be perfect, they'll concede, but it's built on a sound foundation and solves the main problems, and now we should all work together to refine and improve it. The catch, of course, is that each of these friends will decide which end of a boiled egg to crack. This famous tempest in an egg cup was replayed 250 years later by designers of computer hardware and communications protocols. When a block of data is stored or transmitted, either the least-significant bit or the most-significant bit can go first. Which way is better? It hardly matters, although life would be easier if everyone made the same choice. But that's *not* what has happened, and so quite a lot of hardware and software is needed just to swap ends at boundaries between systems.

This modern echo of Swift's Endian wars was first pointed out by Danny Cohen of the University of Southern California in a brilliant 1980 memo, "On holy wars and a plea for peace." The memo, subsequently published in *Computer*, was widely read and admired; the plea for peace was ignored.

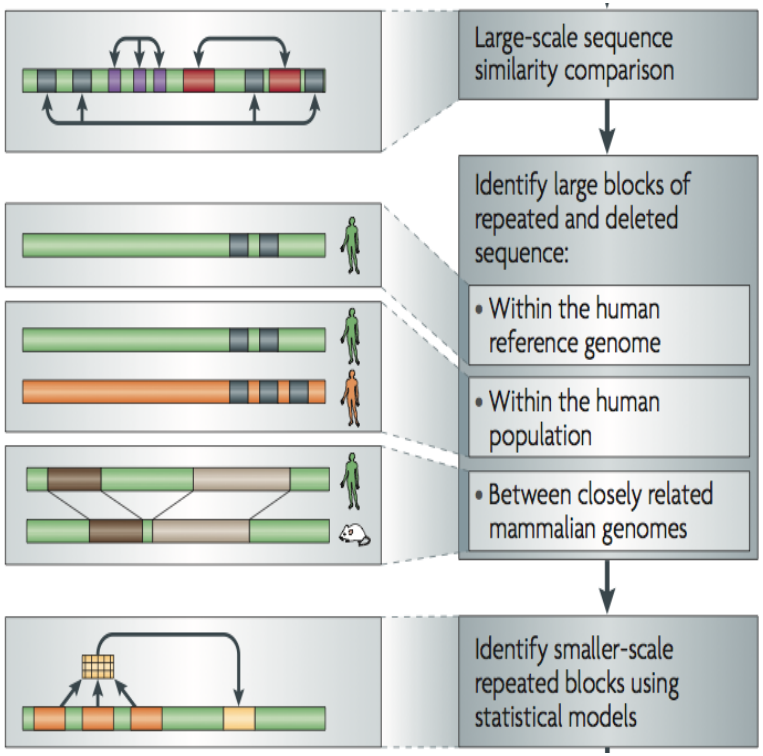
Another feud—largely forgotten, I think, but never settled by truce or treaty—focused on the semicolon. In Algol and Pascal, program statements have to be separated by semicolons. For example, in `x:=0; y:=x+1; z:=2` the semicolons tell the compiler where one statement ends and the next begins. C

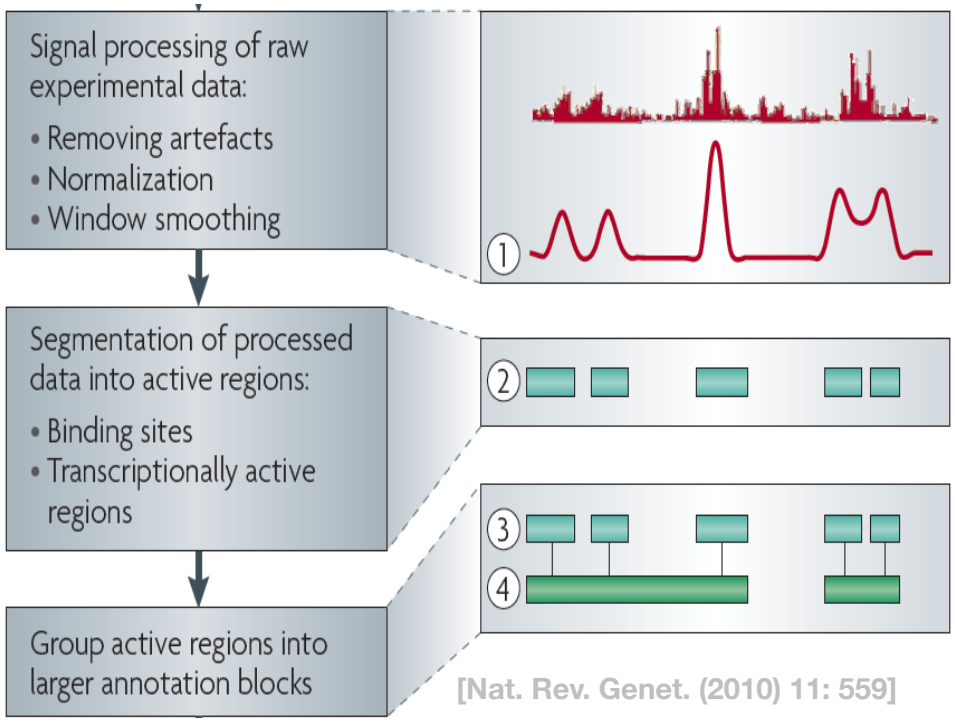
Non-coding Annotations: Overview

There are several collections of information "tracks" related to non-coding features

Sequence features, incl. **Conservation**

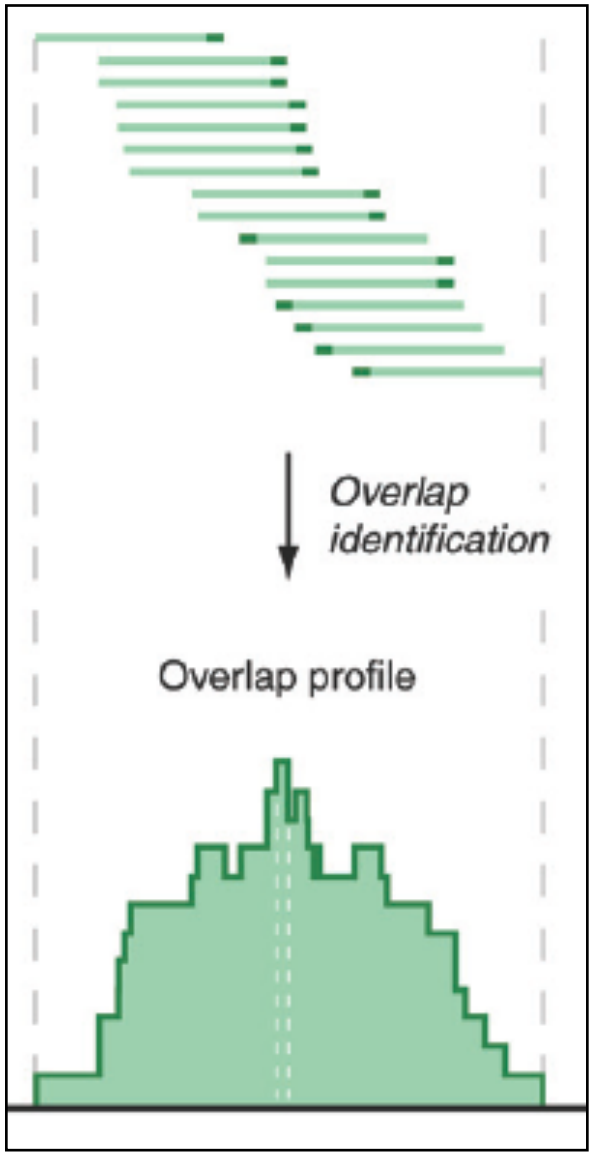
Functional Genomics
ChIP-seq (Epigenome & seq. specific TF) and ncRNA & un-annotated transcription

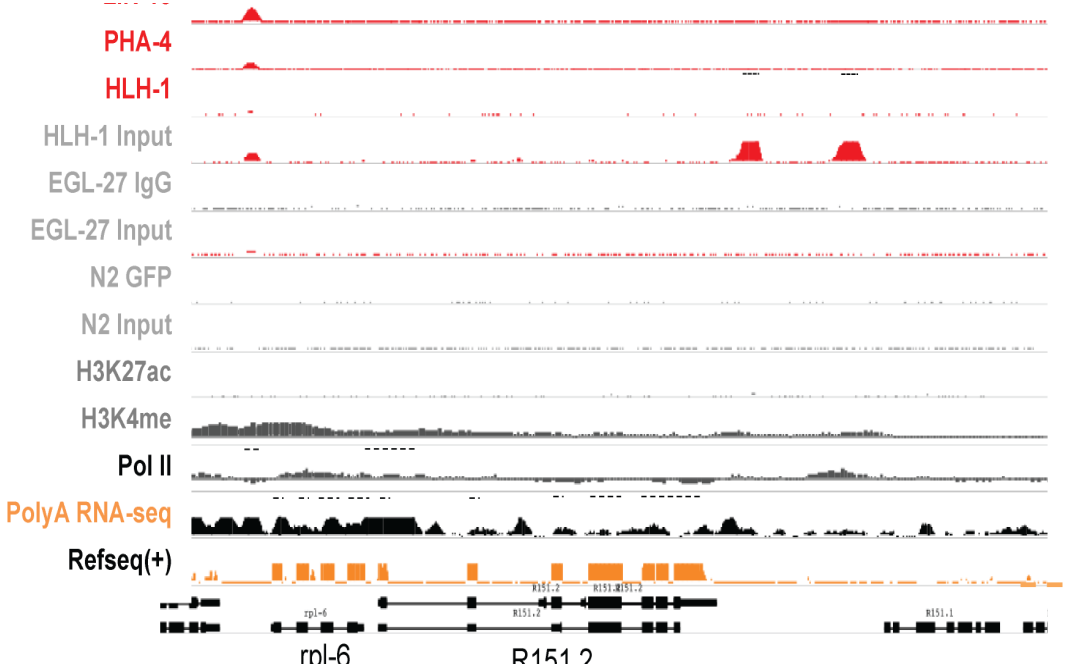




[Nat. Rev. Genet. (2010) 11: 559]

Signal Track for RNA-seq & ChIP-seq





[*PLOS CB* 4:e1000158]

Functional Genomics Annotations

A) PEAKS

1. DNase peaks at the UCSC genome browser {on many cell lines}

2. The regulation track at the UCSC genome browser, with compilation of TF ChIP-seq peaks from uniform processing (individual peaks are annotated with TF and cell line)

3. Blacklist Regions

B) PROMOTERS

Annotated GENCODE TSSes (also, TSSes with FANTOM CAGE support)

C) ENHANCERS (Supervised)

D) UNSUPERVISED SEGMENTATIONS, INCLUDING ENHANCERS

ChromHMM, SegWay, HiHMM....

E) HOT/LOT REGIONS

F) CONNECTIVITY

1. Enhancer-target gene connection
2. TF-target network connectivity
3. TADs: Topologically Associated Domain

G) MOTIFS

for TF binding

H) RNA

1. A matrix of expression data of known genes (or exons) for protein-coding genes & known ncRNAs {on many cell lines}

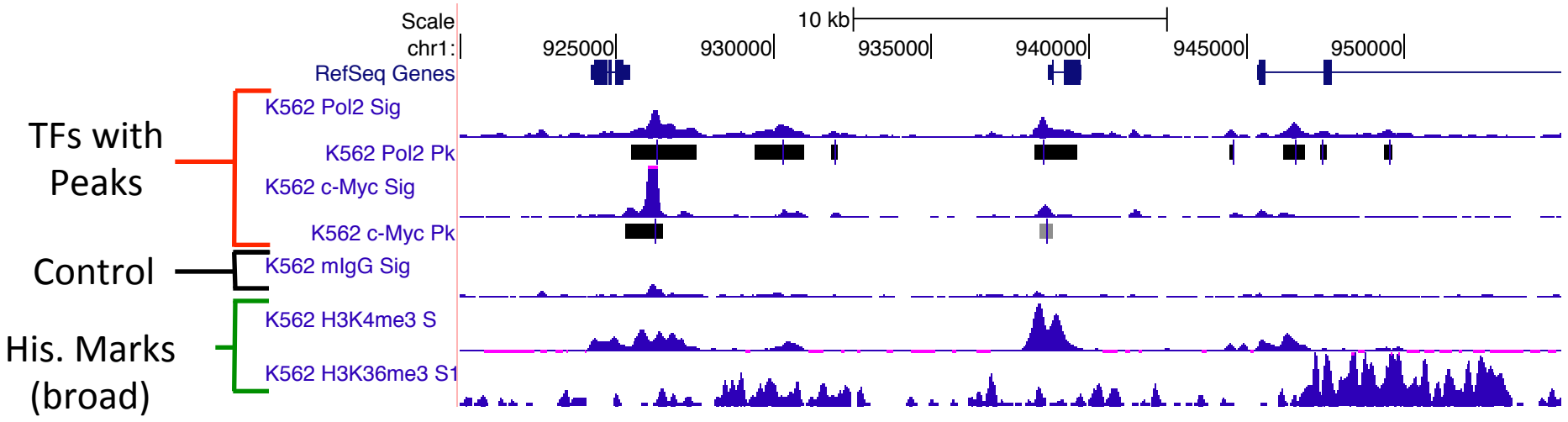
2. Novel RNA contigs track, i.e., possible novel transcripts (ie Transcriptionally Active Regions or TARs)

3. Novel junctions

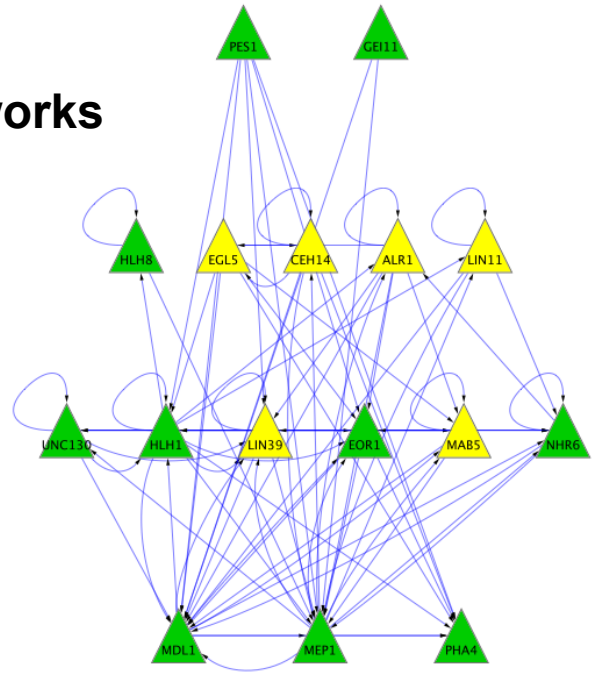
I) OTHER

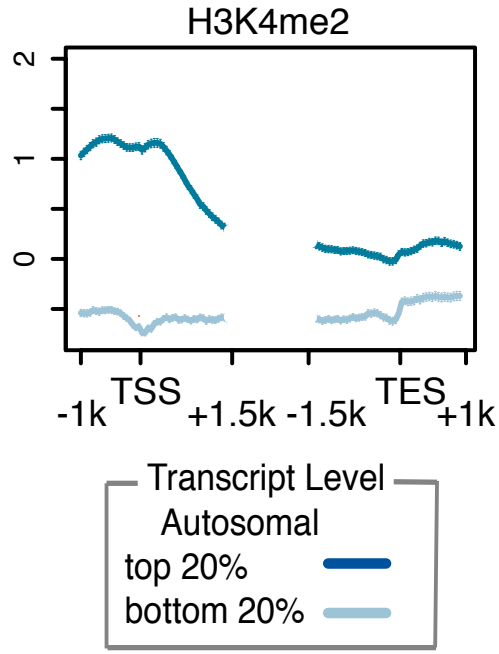
1. List of Allelic SNPs & Regions
2. Models

Higher level Information from ChIP-seq



Networks



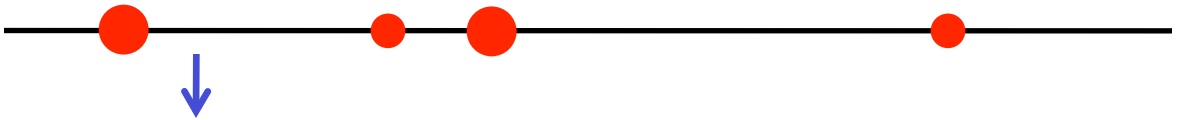


Aggregations

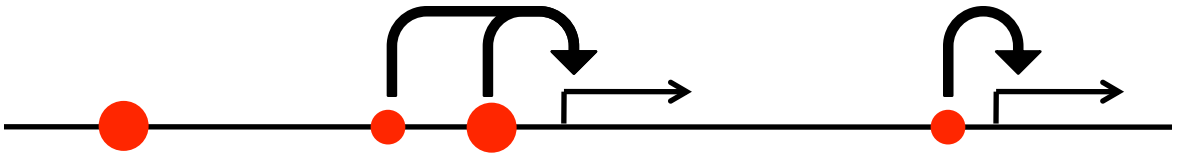
[*Science* 330: 1775
+ ENCODE Data Sources
TFs & Control: Yale
HMs: UW & Broad]

Data Flow: peaks to proximal & distal networks

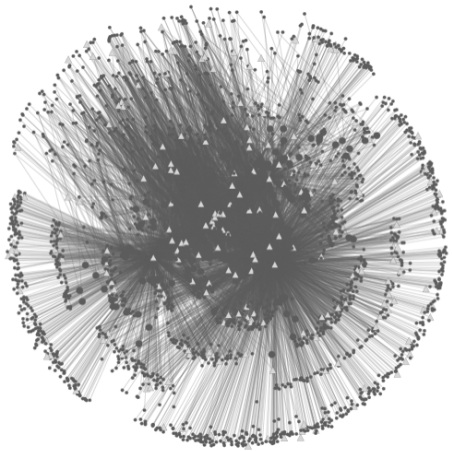
Peak Calling



Assigning TF binding sites to targets



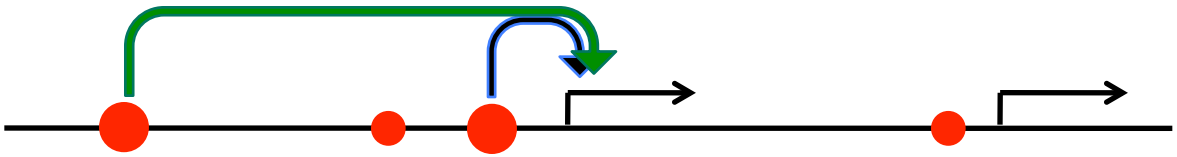
~500K Edges

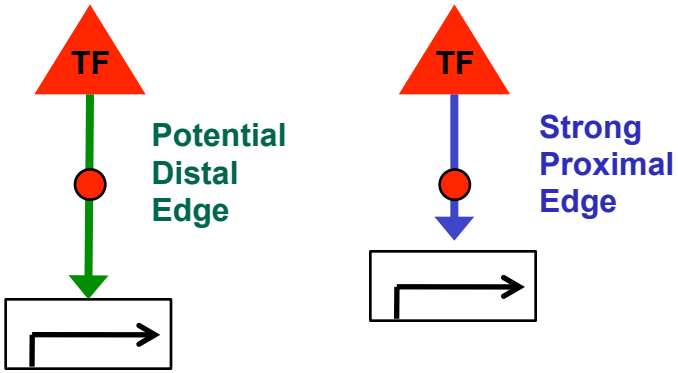


[Cheng et al., *Bioinfo.* ('11); *Nature* 489:91 ('12), doi:10.1038/nature11245; Yip et al., *GenomeBiology* ('12)]

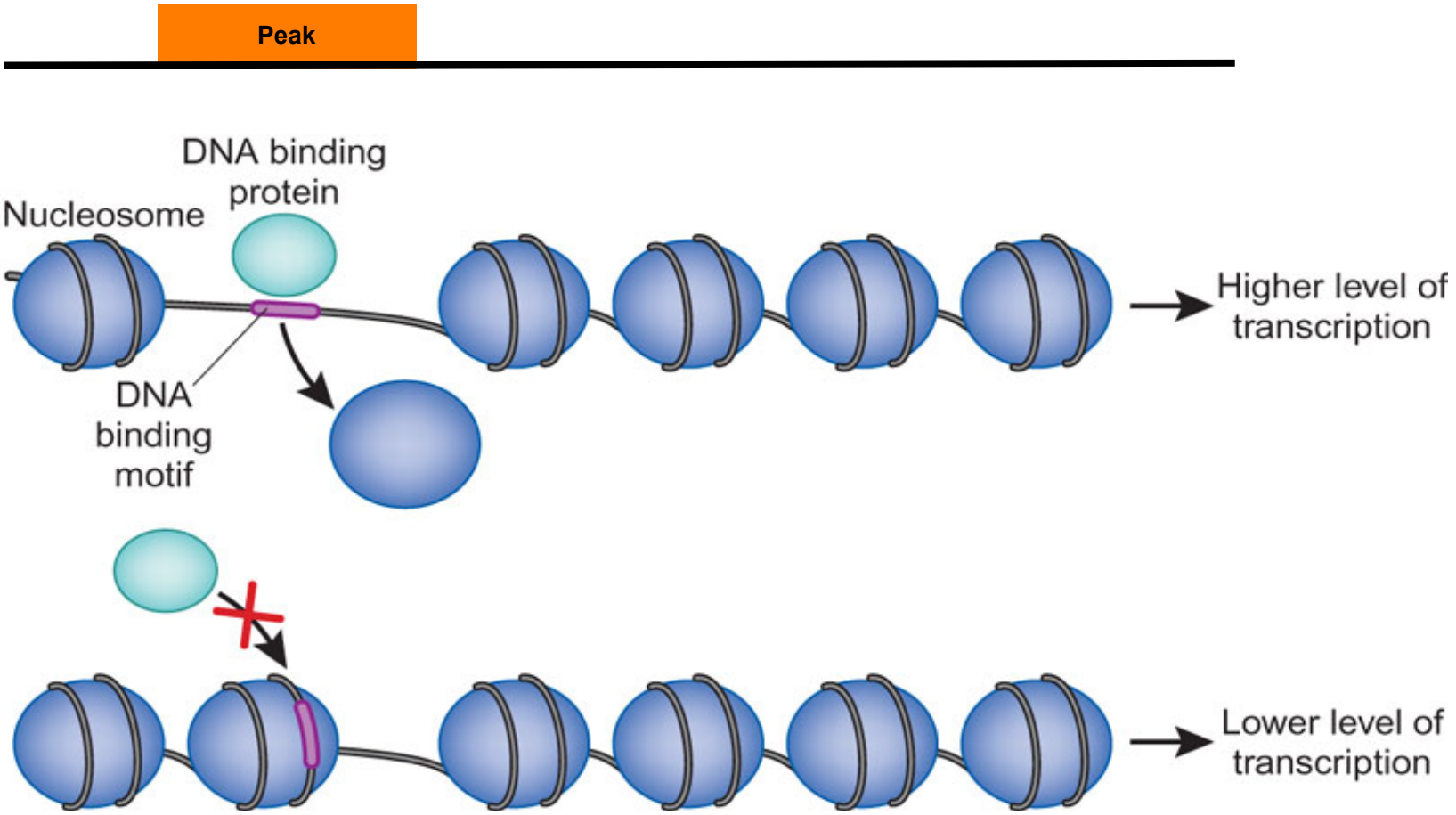
Filtering high confidence edges & distal regulation

Based on stat. model combining
signal strength & location relative to typical binding



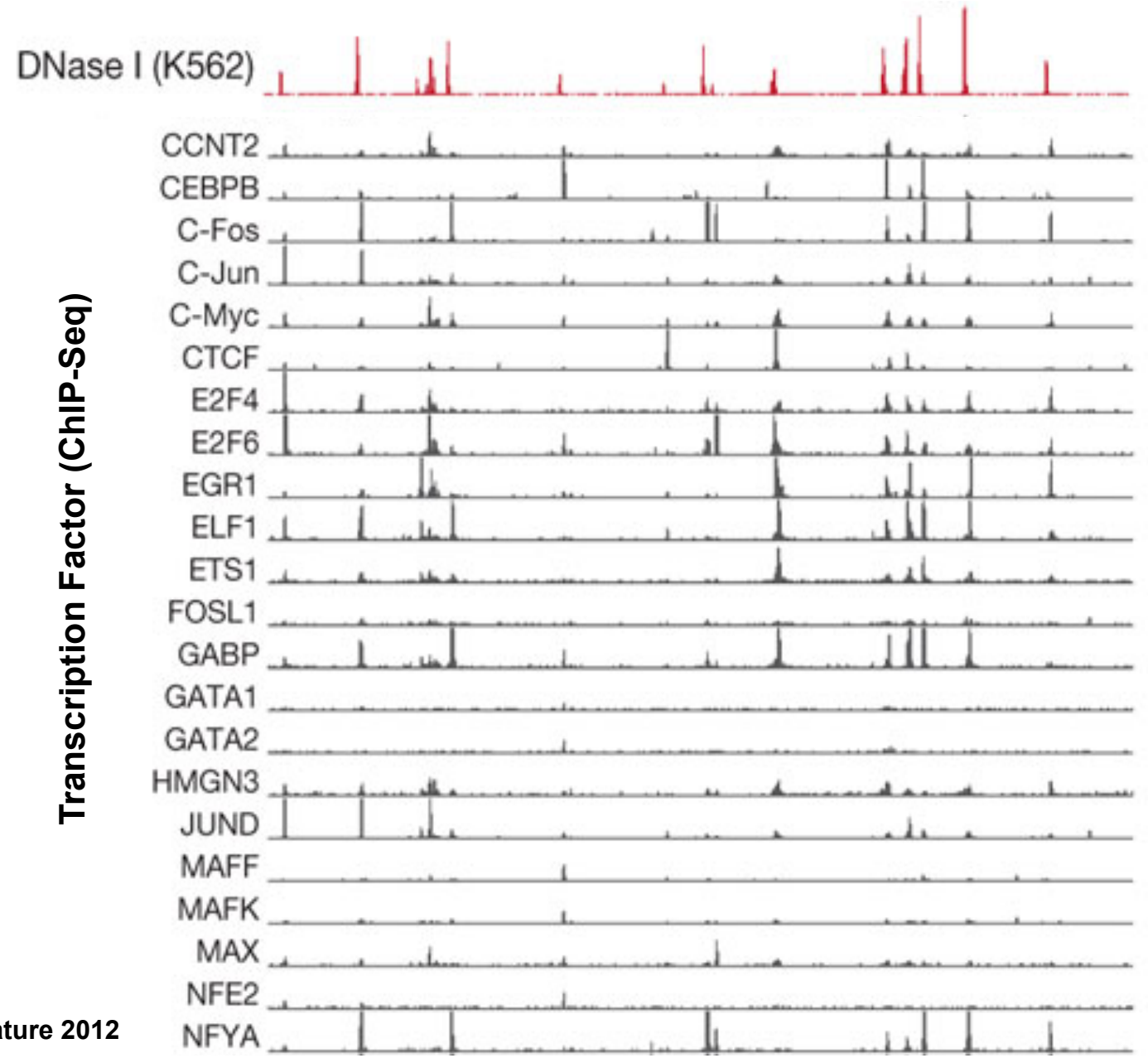


DNase Peaks & Open Chromatin



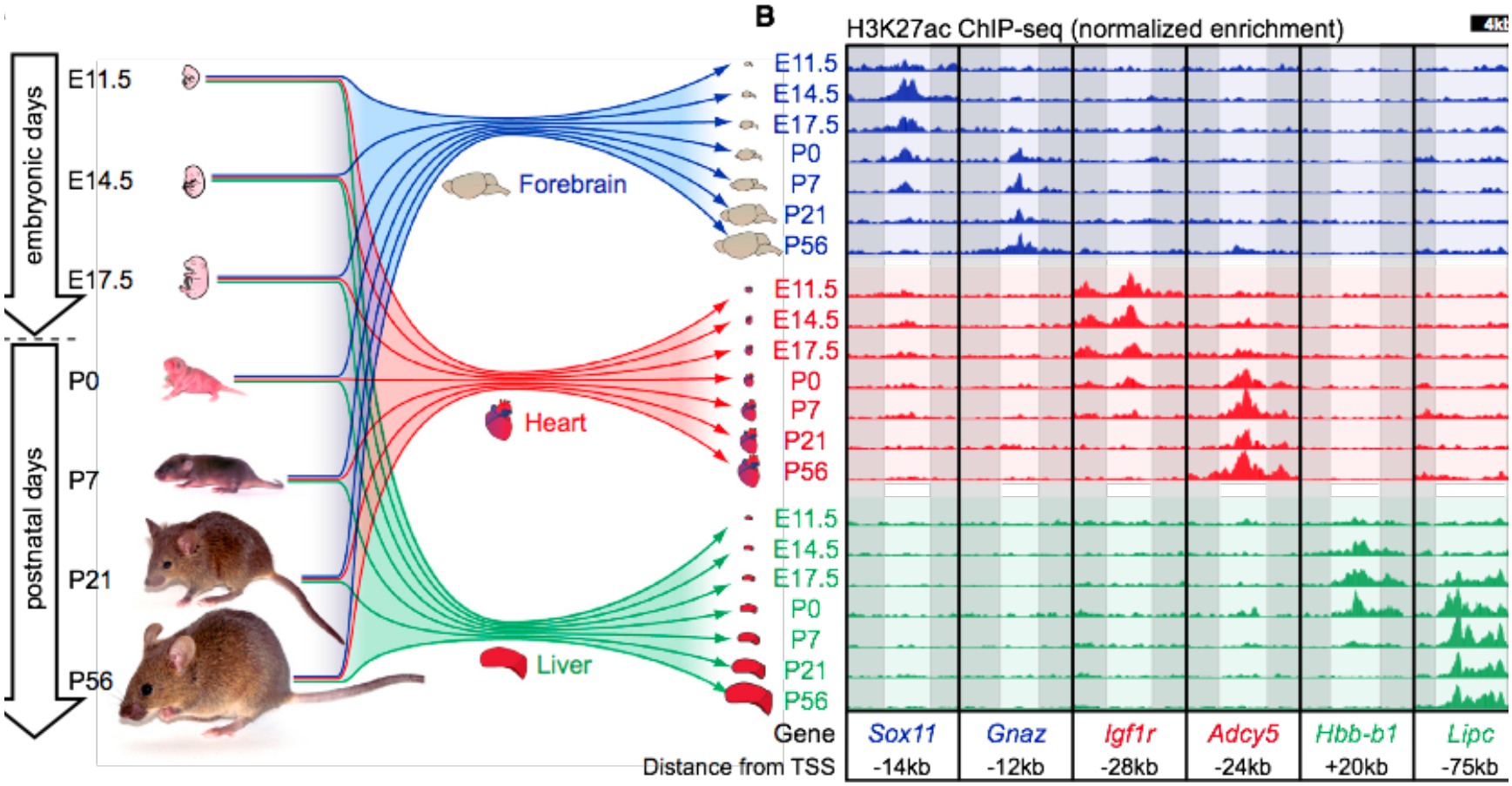
Groop L. Nature genetics, 2010, 42(3): 190-192.

DNase hypersensitivity as a mark of functionality



Thurman et al Nature 2012

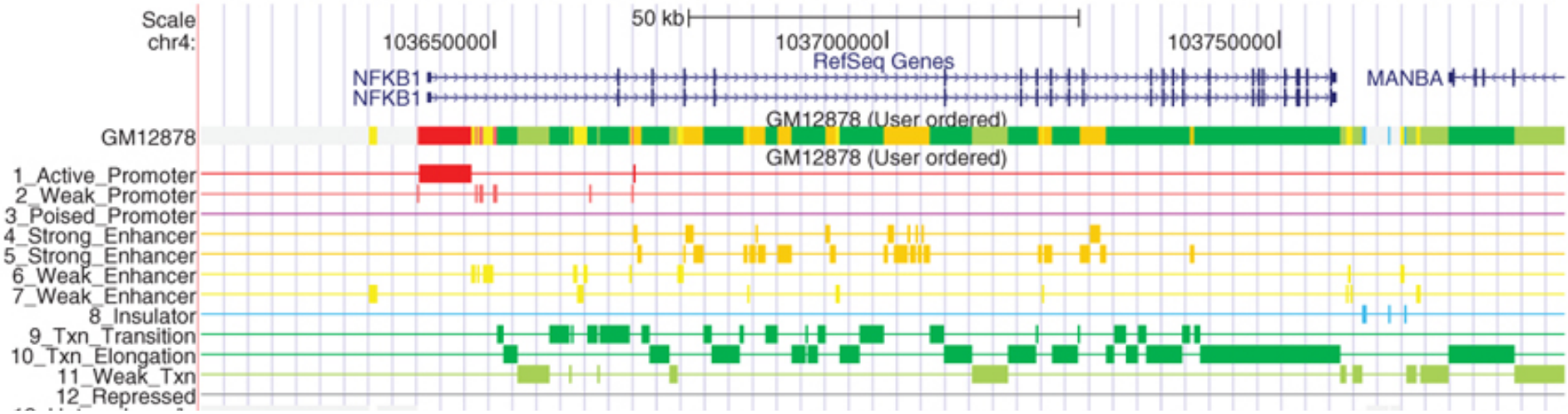
H3K27ac is an important mechanism to regulate the activity of enhancers in different developmental stages

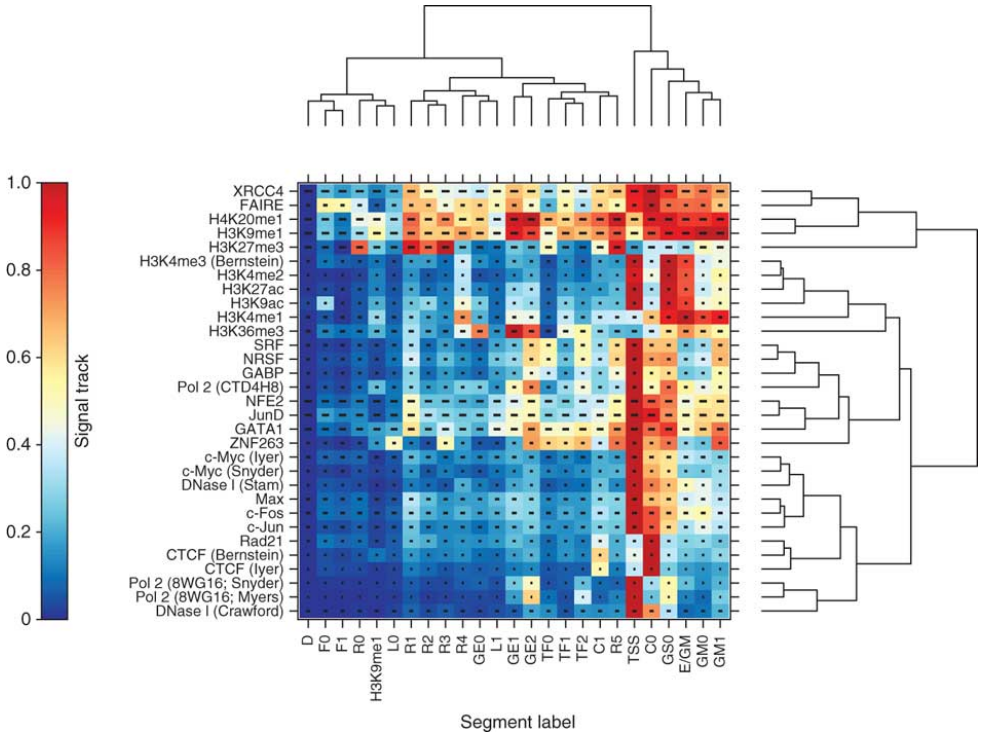


Epigenetically, H3K27ac marks are present near active enhancers.

Nord, et al., Cell, 2013

Genome Segmentations

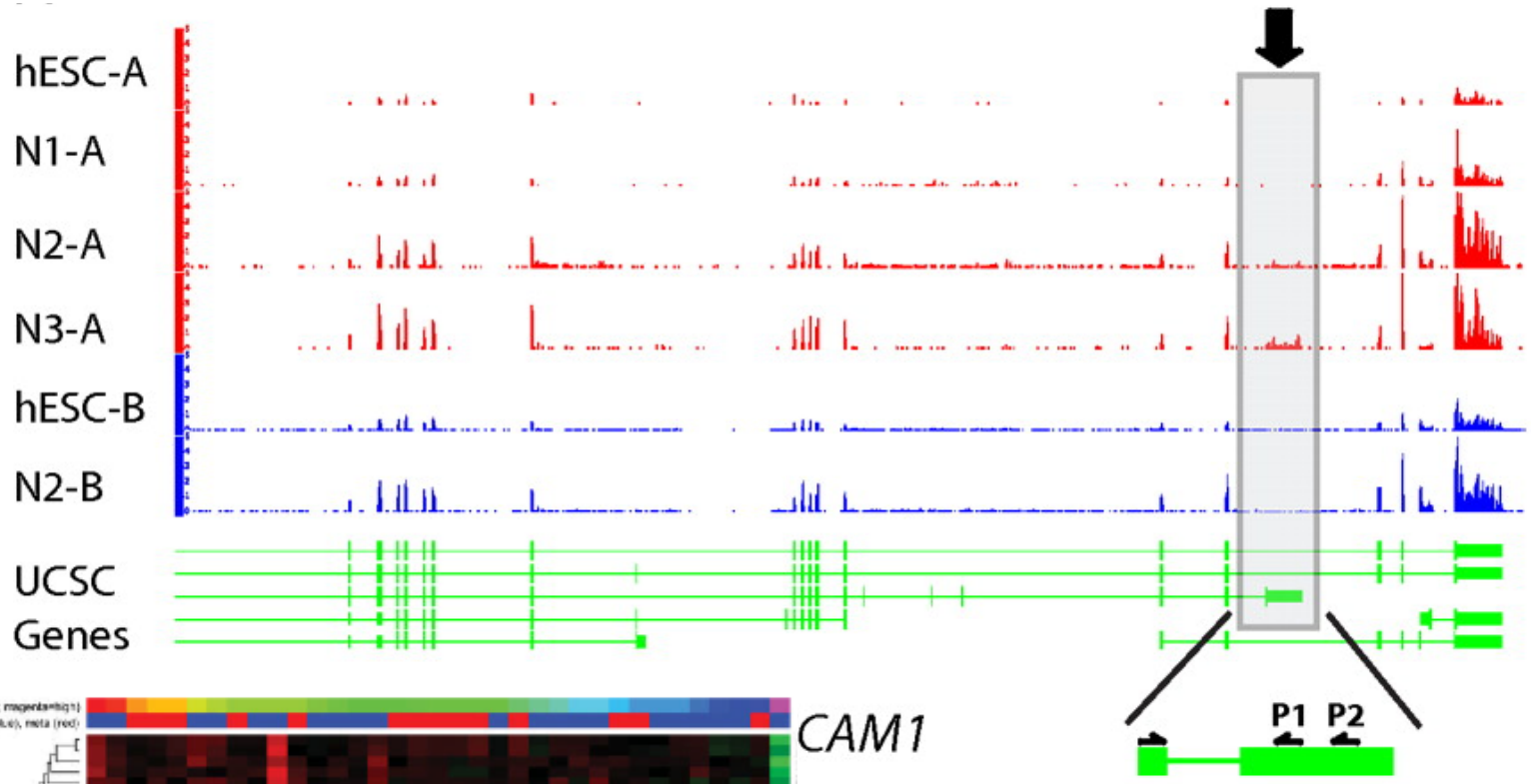


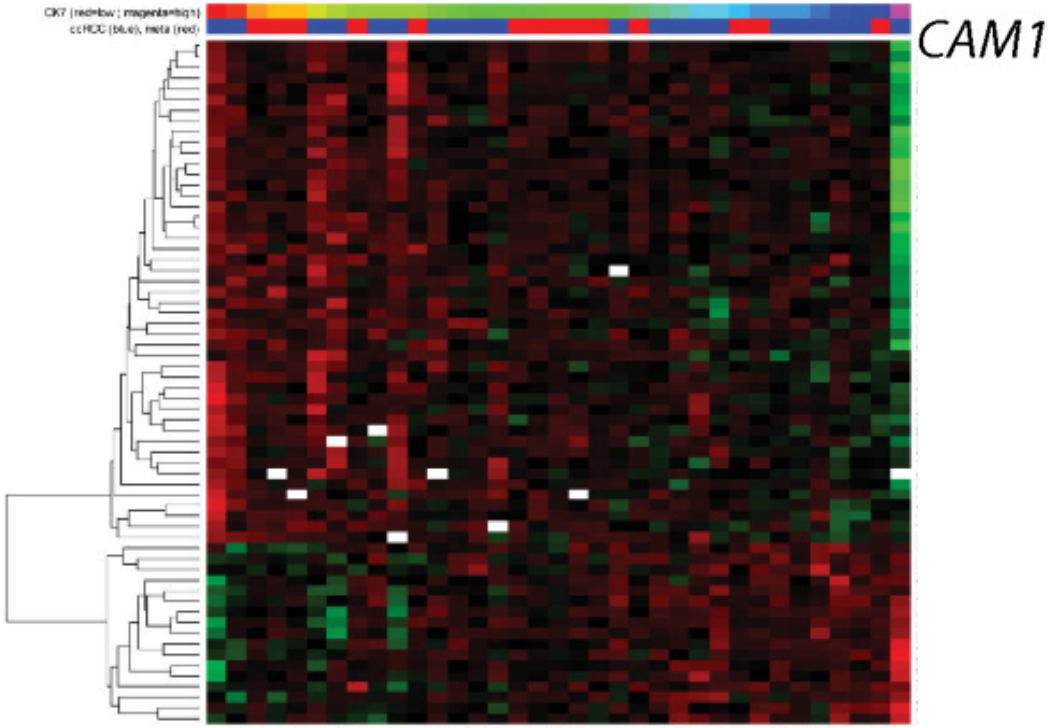


Unsupervised segmentation of chromatin features groups regions with similar patterns and labels each pattern, thus, annotating the genome.

Hoffman, et al, Nature Methods, 2012

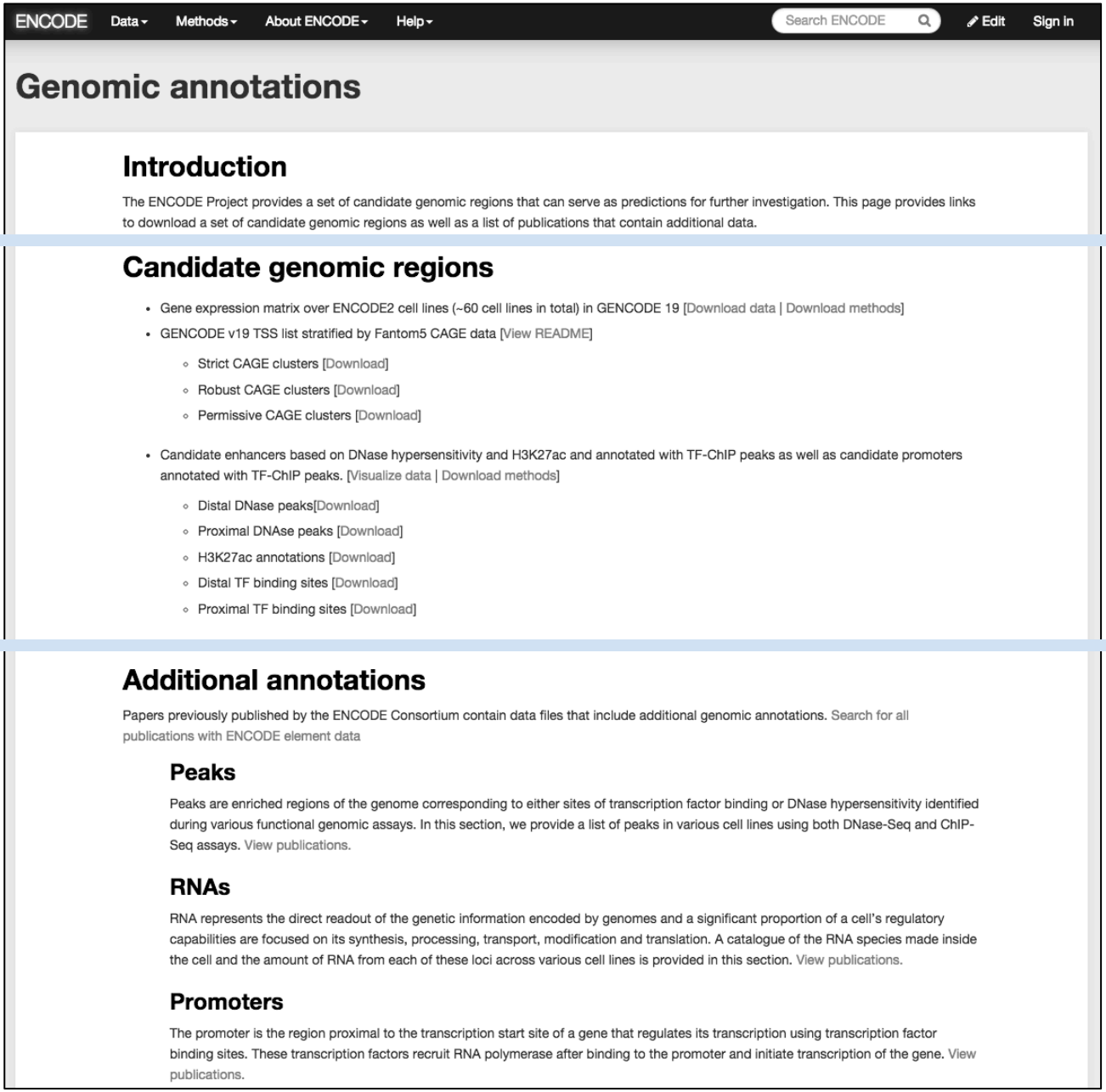
Ernst & Kellis, Nature Methods, 2012.





Higher level Information from RNA-seq:
Avg. signal at exons & "TARs" (RPKMs)

[*PNAS* 4:107: 5254 ; *IJC* 123:569]



Simplified

Comprehensive

(published annotation, mostly in '12 & '14 rollouts)

"Simplified" Annotation

- "Slice" through the ENCODE, providing close-to-data subset of the annotations

- Gene expression matrix
  - over ENCODE2 cell lines (~60 cell lines in total) in GENCODE 19
- TSS list
  - GENCODE v19

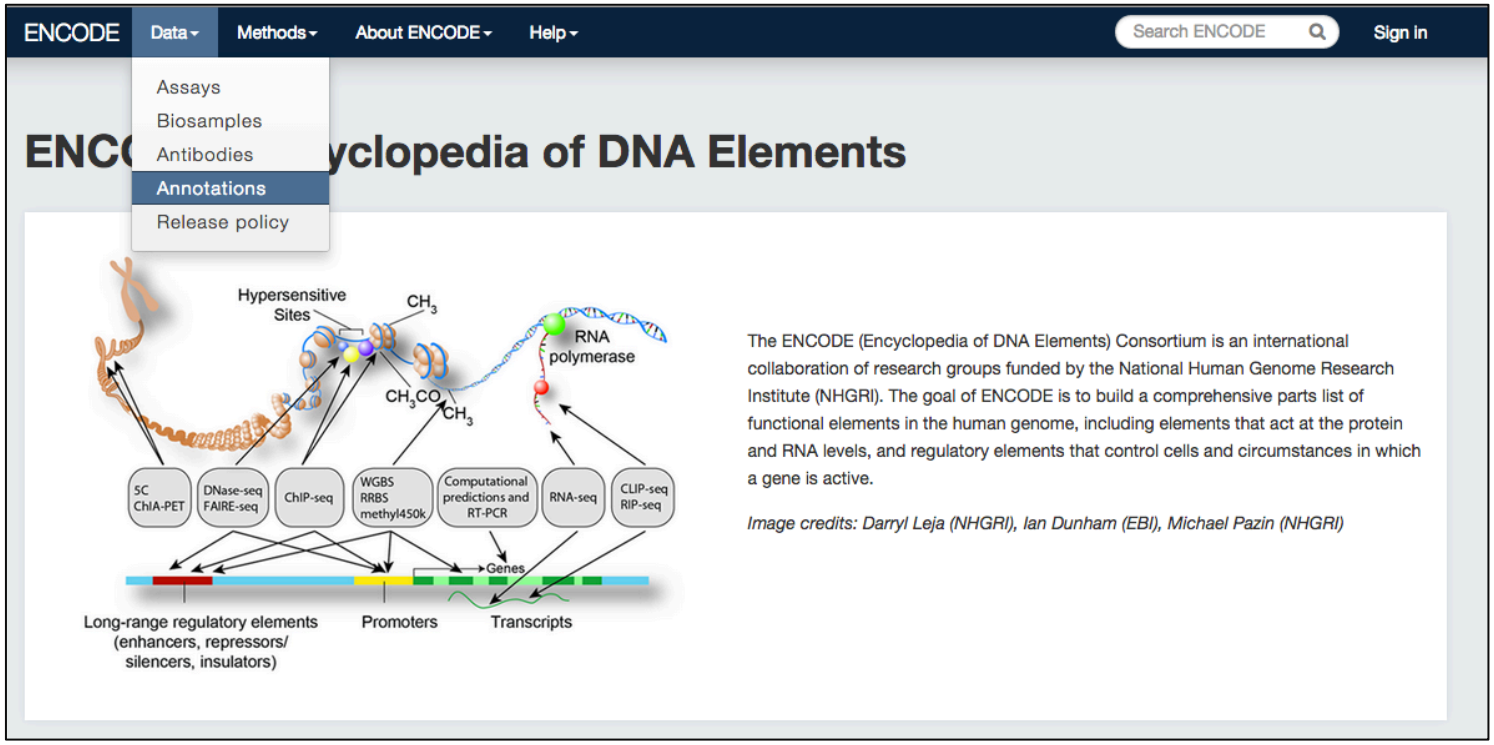
- “Tissue type” facet for the cell lines (DCC)

Simplified regulatory sites

- Candidate enhancers: The master list of TSS-distal DHS peaks annotated with
  - H3K27ac enrichment (percentile over background) in a cell-type-specific manner.
  - TF ChIP-seq peaks across cell-types
- Candidate promoters: The master list of TSS-proximal DHS peaks annotated with TF ChIP-seq peaks across cell types.


distal DHS
proximal DHS
distal H3K27ac
CTCF,RAD21
,GATA2,GATA3,REST
FOS,JUN
FOS
CEBPB
FOS,GATA2,JUN,POLR2A
NH-A
NHDF-Ad
NHLF
Osteobl
HSMM
HSMMtube
HUVEC
K562
NH-A
NHDF-Ad
NHLF
Osteobl
HeLa-S3
HMEC
HSMM
HSMMtube
HUVEC
NH-A
NHDF-Ad
NHLF
HMEC
HSMM
HUVEC
NH-A
NHEK
NHLF
HeLa-S3
HepG2
HUVEC
NH-A
NHDF-Ad
NHLF
Osteobl
H1-hESC
HSMM
HSMMtube
HUVEC
NH-A
NHDF-Ad
NHEK
NHLF
Osteobl
HSMMtube
NHDF-Ad
Osteobl
HMEC
HSMM
HSMMtube
HUVEC
NH-A
NHDF-Ad
NHEK
NHLF
Osteobl
Osteobl
HSMM
HSMMtube
HUVEC
NH-A
NHDF-Ad
NHEK
NHLF
Osteobl
HSMMtube

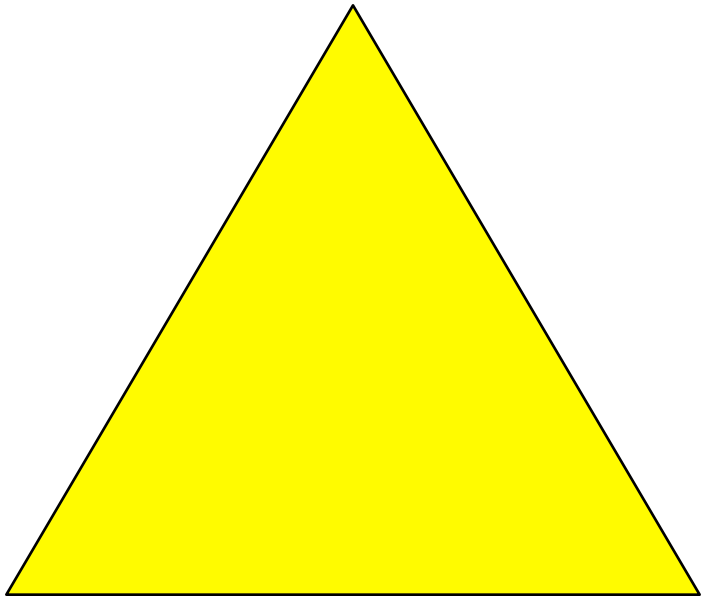
Access candidate genomic annotations via encodeproject.org on the "Data" menu bar



encodeproject.org/data/annotations

Default Theme

- Default Outline Level 1
  - Level 2



Details of DNase peaks, H3K27ac annotation and TF ChIP-seq annotations

DNase peak detail

Item: re9.73027455

Cell lines merged: HMEC,Osteobl,HepG2,HSMMtube,GM12878,HSMM,HUVEC
Number of cell lines merged: 7
Number of peaks merged: 7
DNase scores of peaks merged: 52,32,9,5,23,37,23

H3K27ac annotation

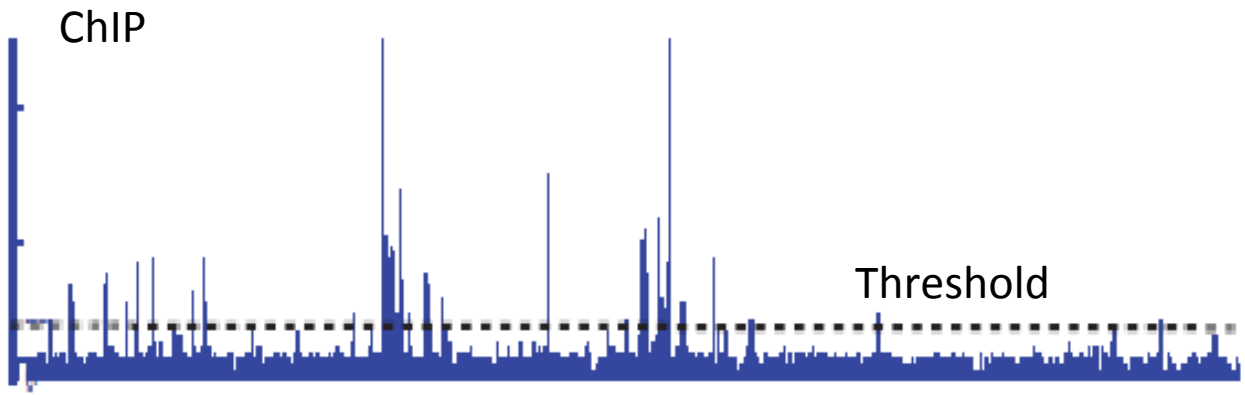
Item: H1-hESC
H3K27ac percentile over background: 99.53

TF annotation

Item: FOS,SPI1
Transcription factors (cell lines) represented by interval: FOS(HUVEC),SPI1(GM12878),SPI1(GM12891)

Peak Calling

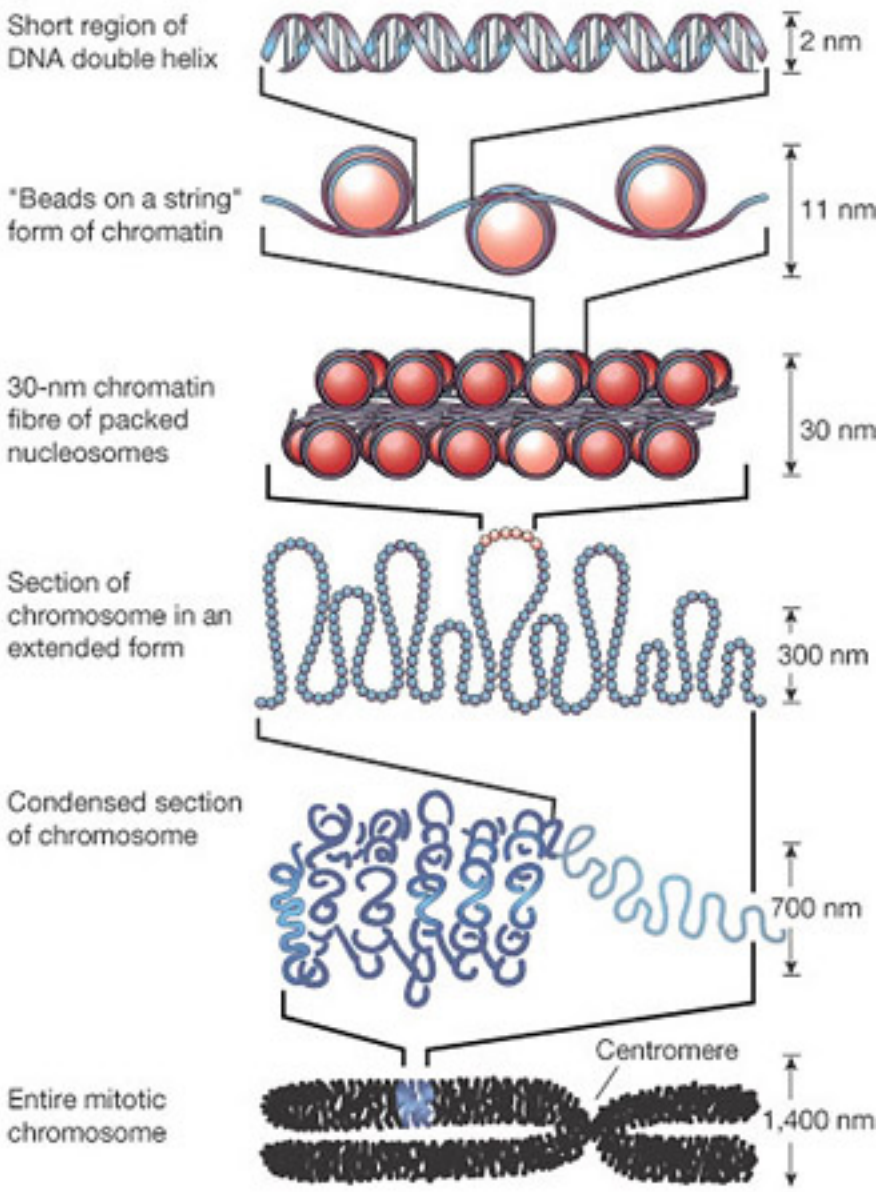
- Generate and threshold the signal profile and identify candidate target regions
  - Simulation (PeakSeq),
  - Local window based Poisson (MACS),
  - Fold change statistics (SPP)

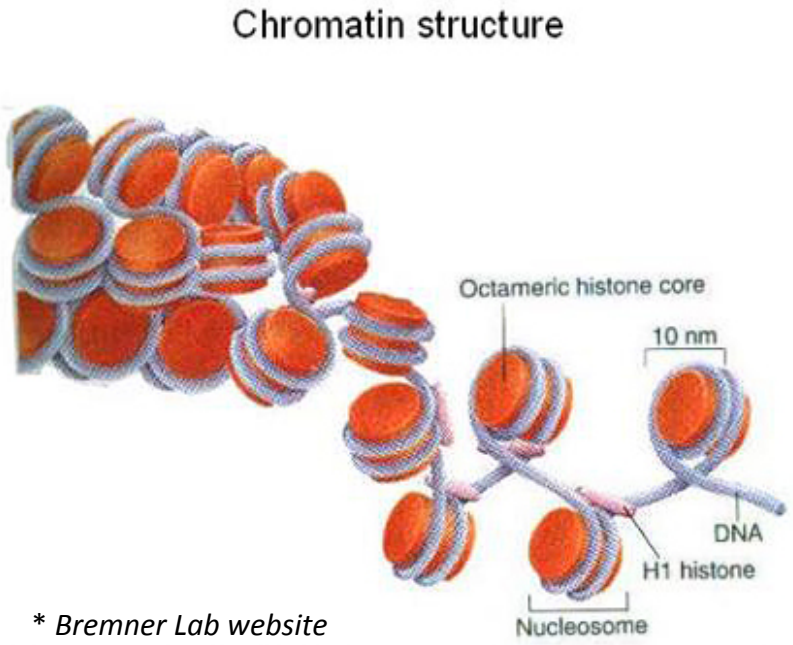


Potential Targets →

- Score against the control

Significantly Enriched targets →





** Bremner Lab website*

Chromatin is the combination or complex of DNA and proteins that make up the contents of the nucleus of a cell. The basic repeat element of chromatin is the nucleosome, interconnected by sections of linker DNA.

http://www.integratedhealthcare.eu/1/en/histones_and_chromatin/