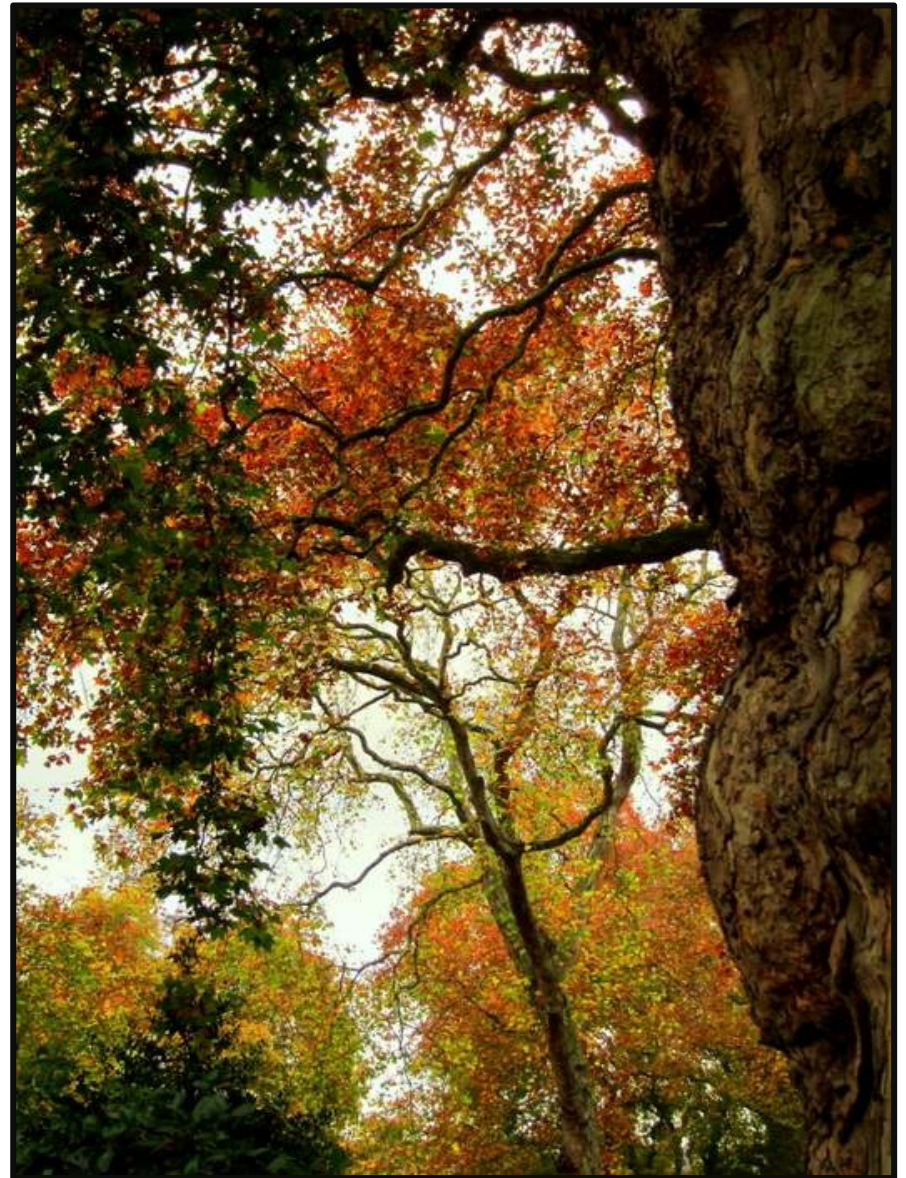


Overview of Genome Annotation & ENCODE Elements

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?

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

cede 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

Color is Function

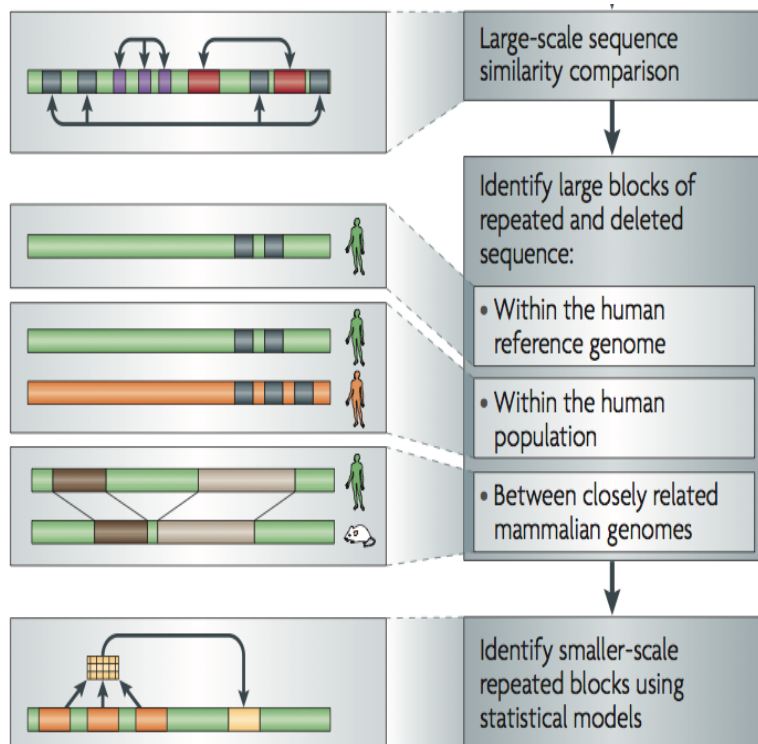
Lines are Similarity

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

Non-coding Annotations: Overview

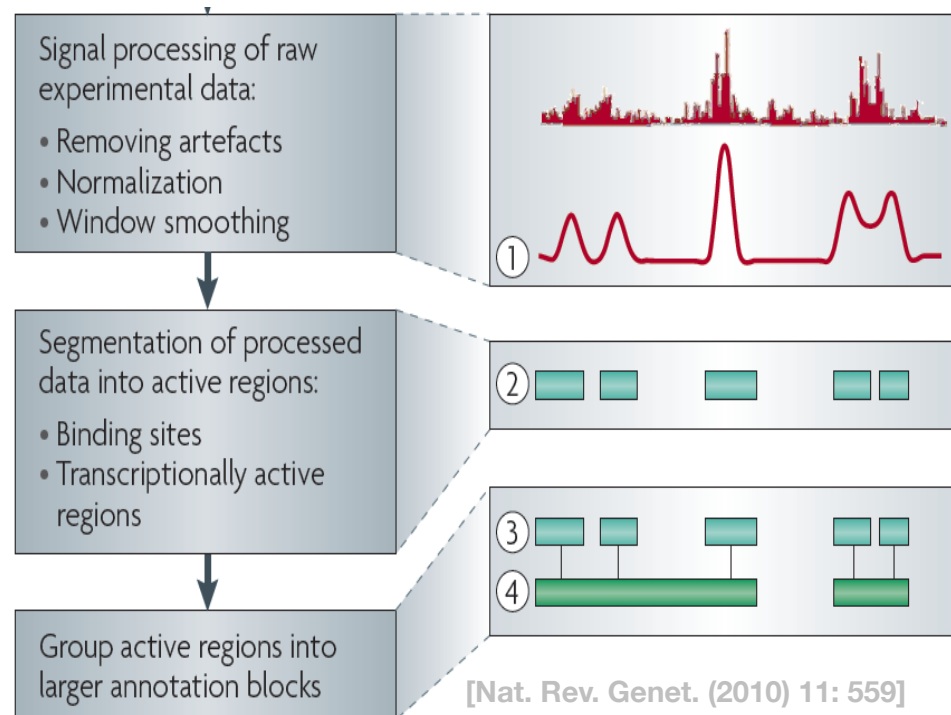
There are several collections of information "tracks" related to non-coding features

Sequence features, incl. Conservation

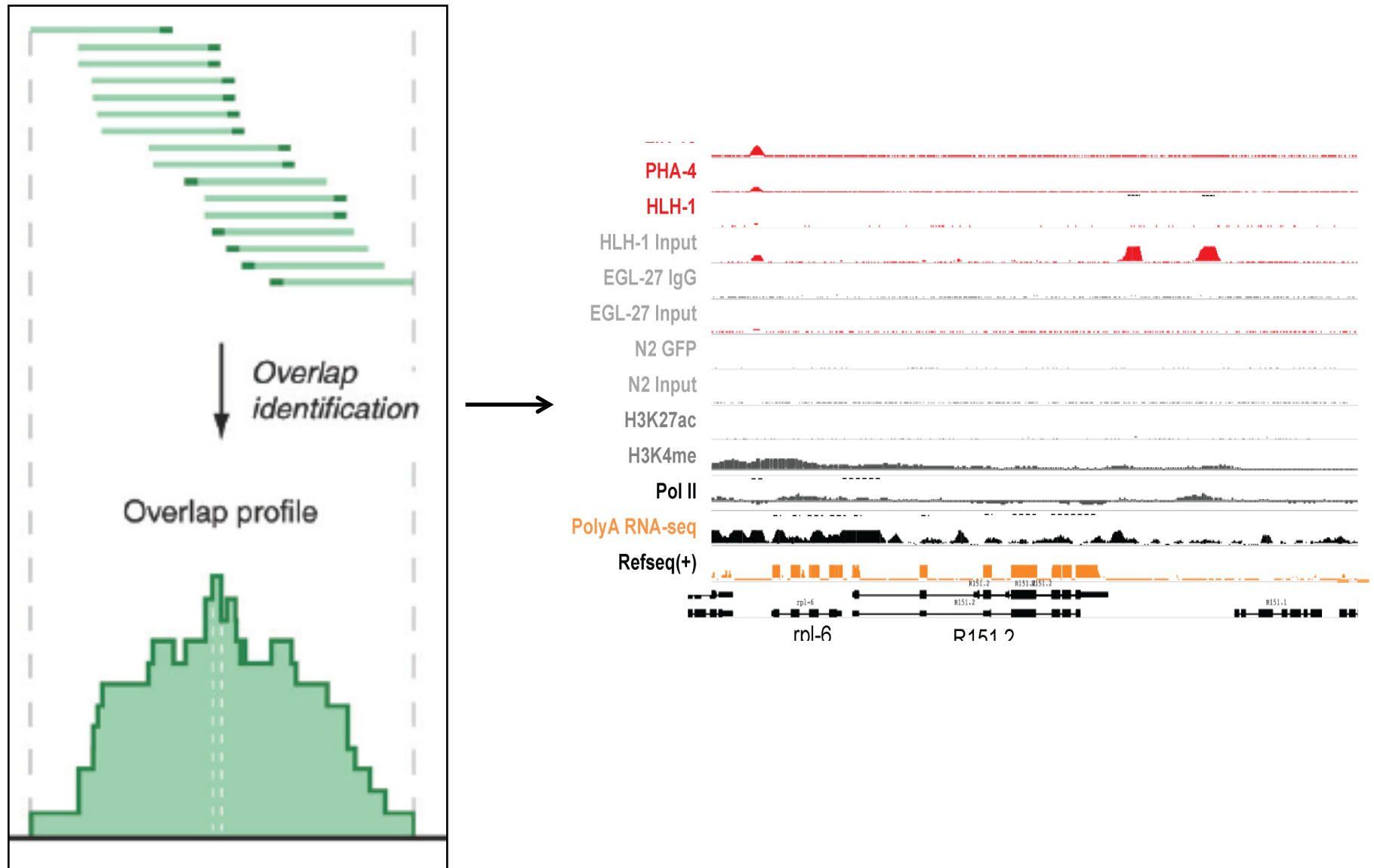


Functional Genomics

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



Signal Track for RNA-seq & ChIP-seq



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

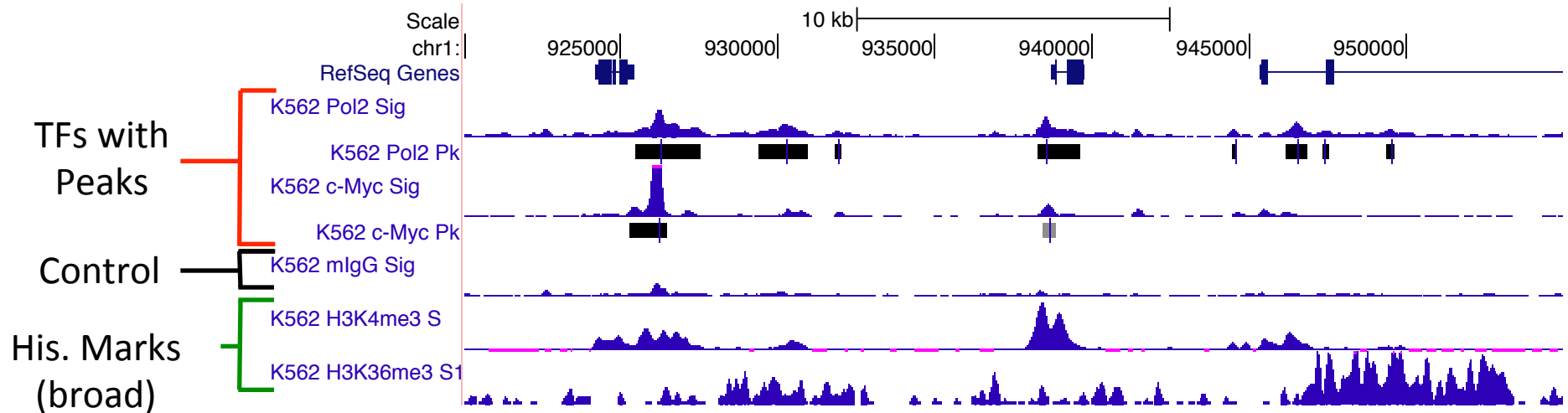
H) RNA

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

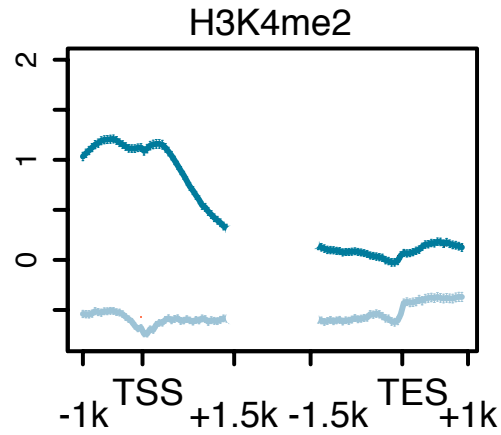
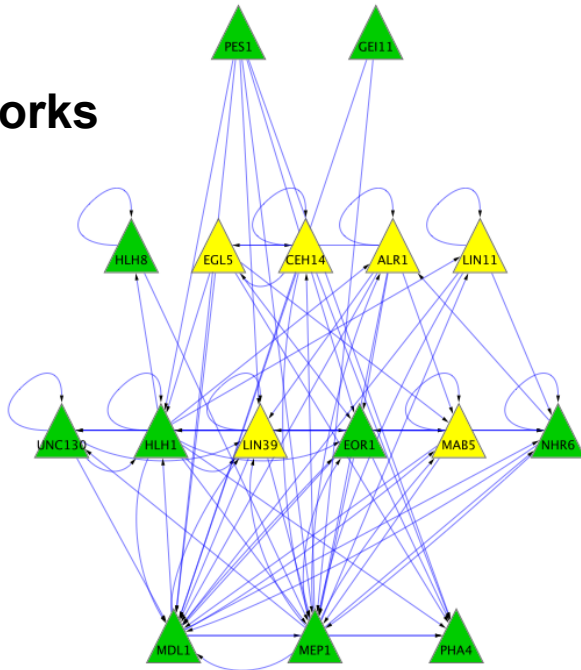
I) OTHER

1. List of Allelic SNPs & Regions
2. Models

Higher level Information from ChIP-seq



Networks



Aggregations

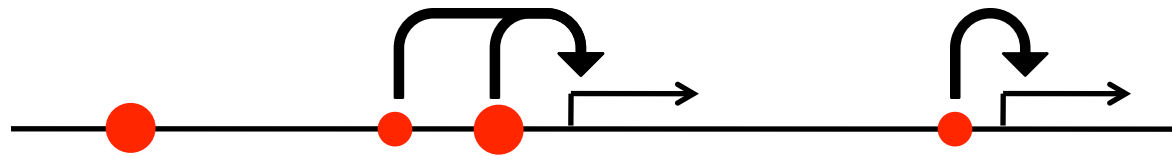
[*Science* 330: 1775
+ ENCODE Data Sources
TFs & Control: Yale
HMs: UW & Broad]

Data Flow: peaks to proximal & distal networks

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



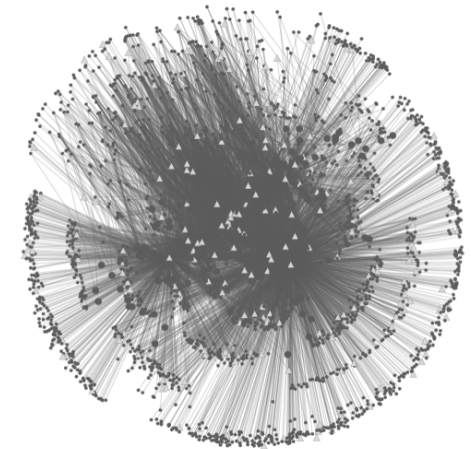
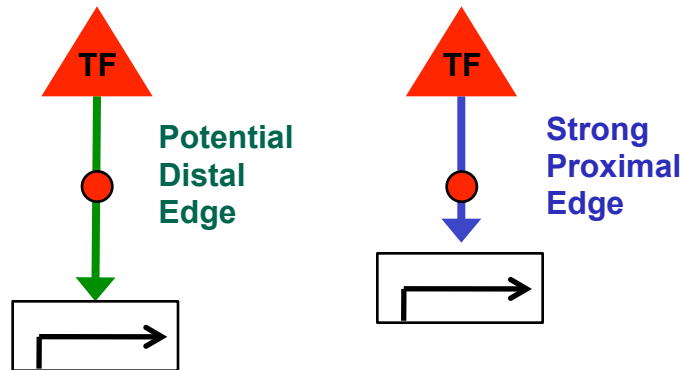
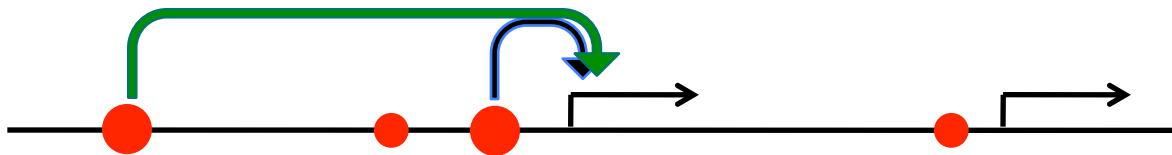
Assigning TF binding sites to targets



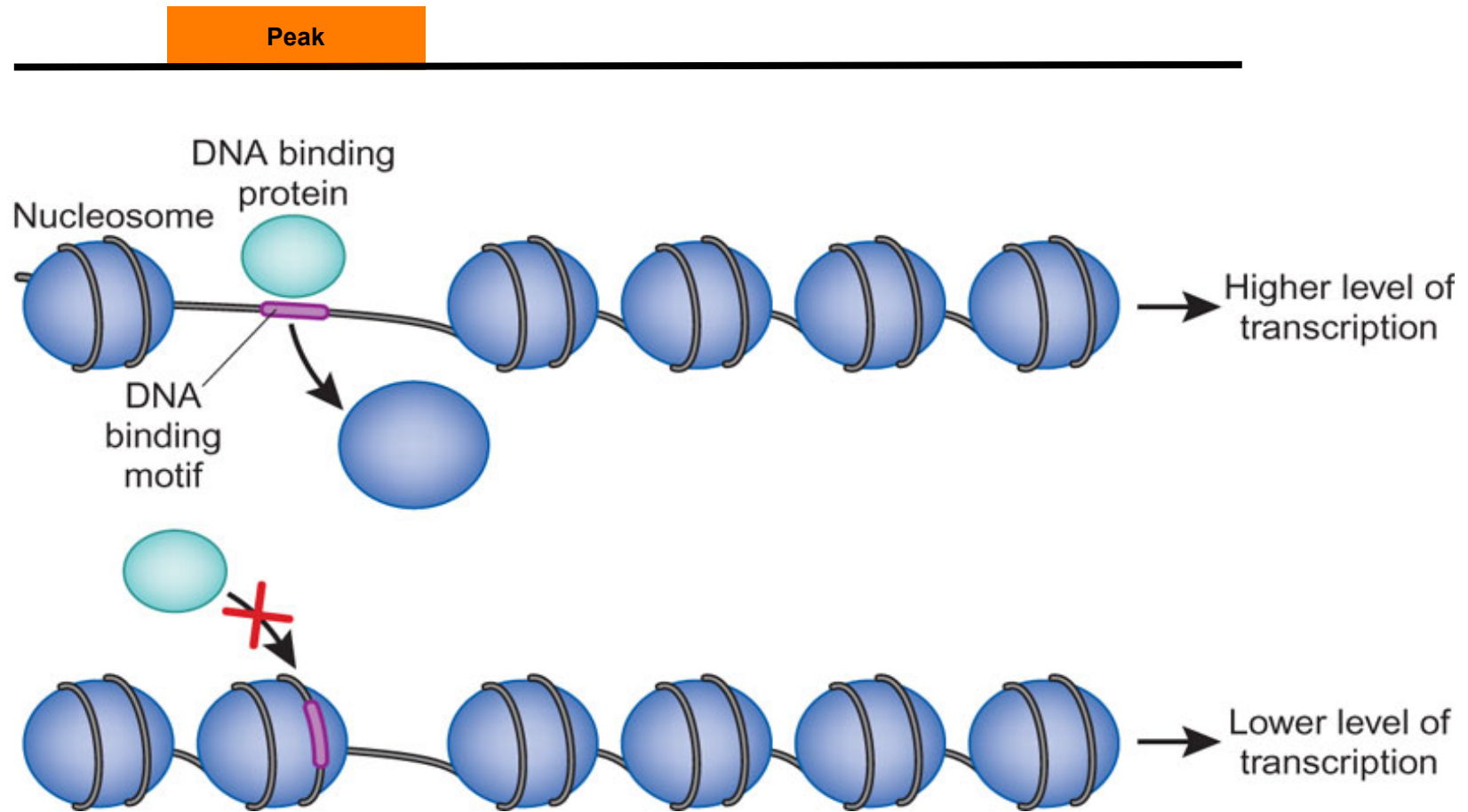
~500K Edges

Filtering high confidence edges & distal regulation

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

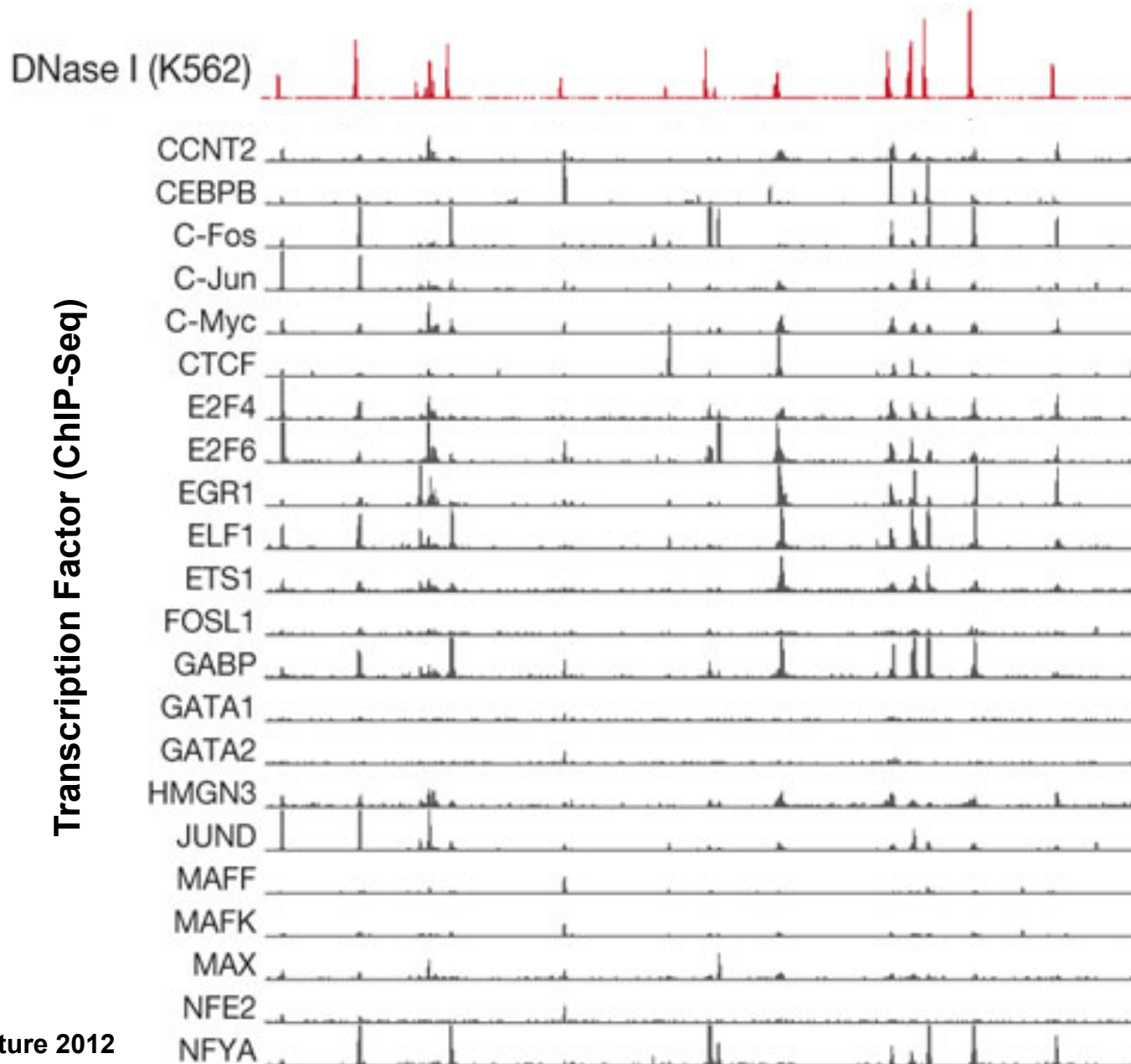


DNase Peaks & Open Chromatin

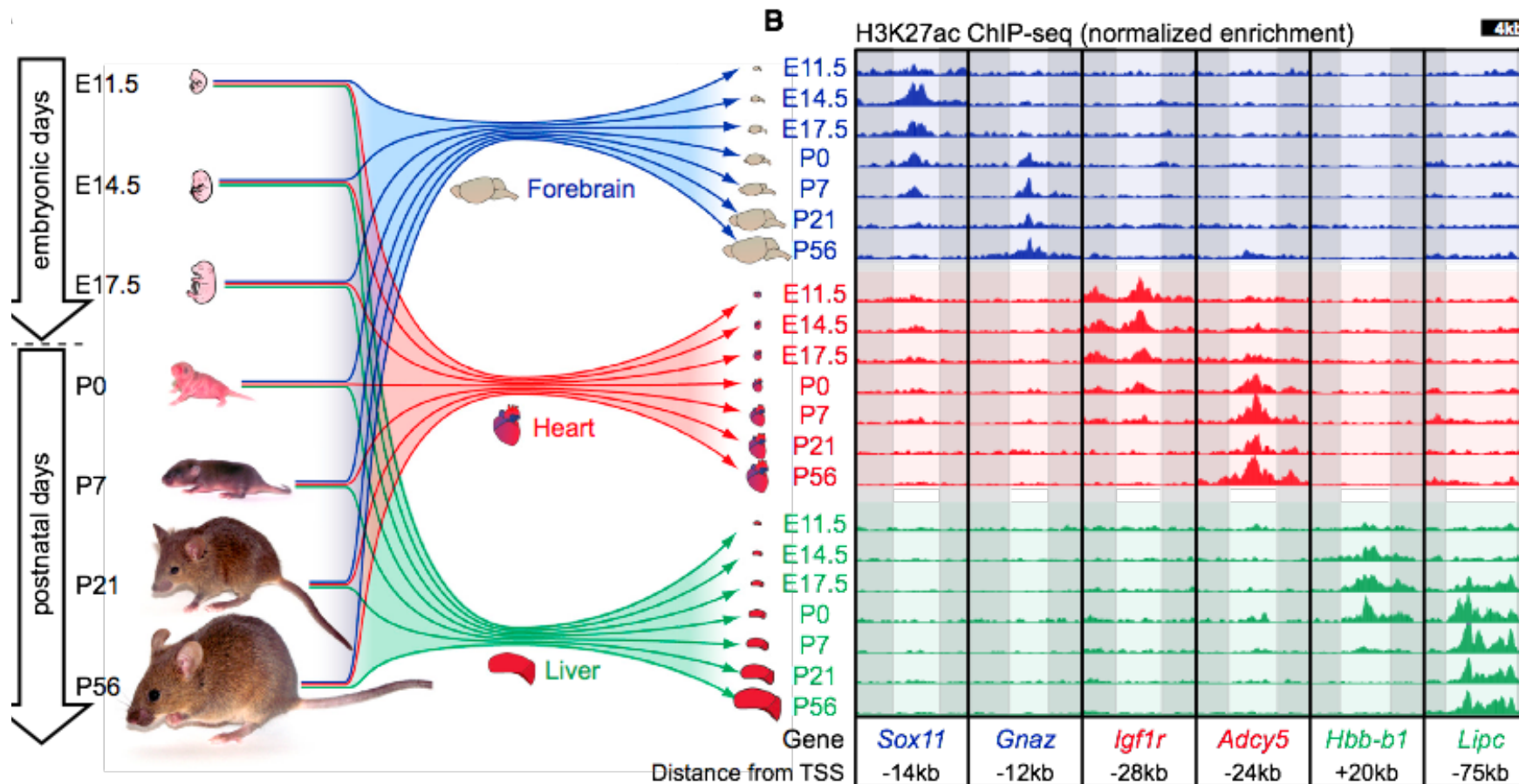


Group L. Nature genetics, 2010, 42(3): 190-192.

DNase hypersensitivity as a mark of functionality

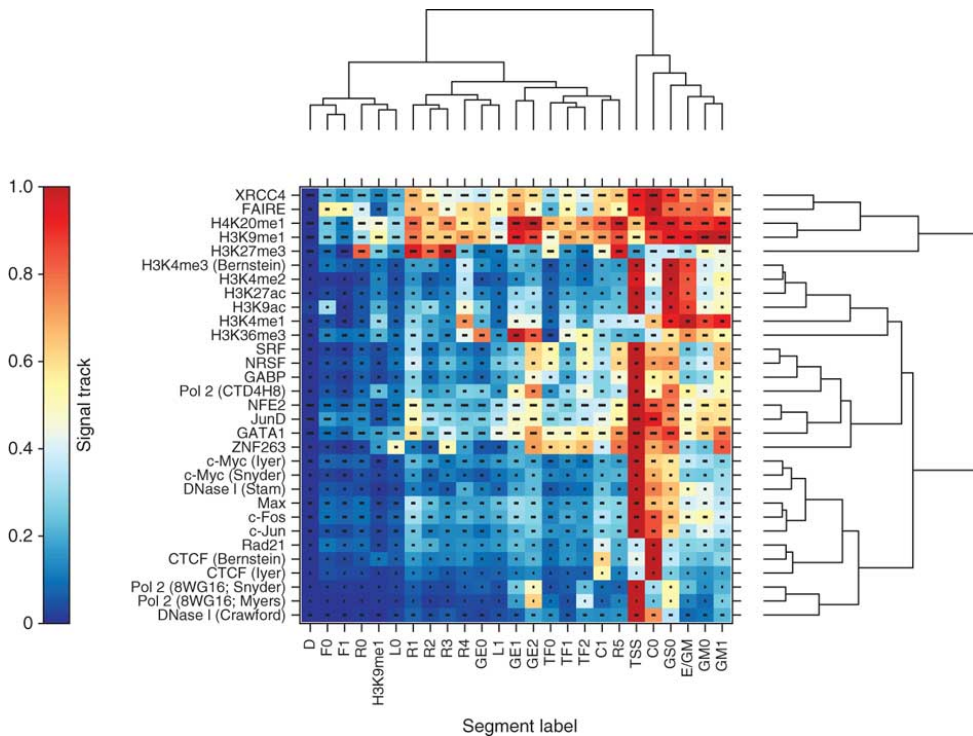
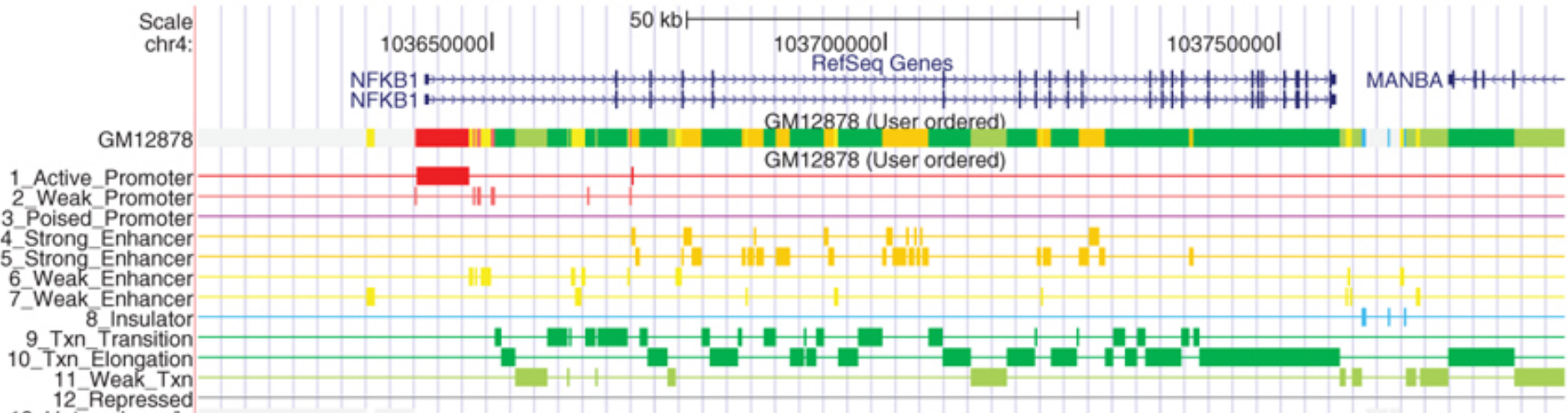


H3K27ac is an important mechanism to regulate the activity of enhancers in different developmental stages



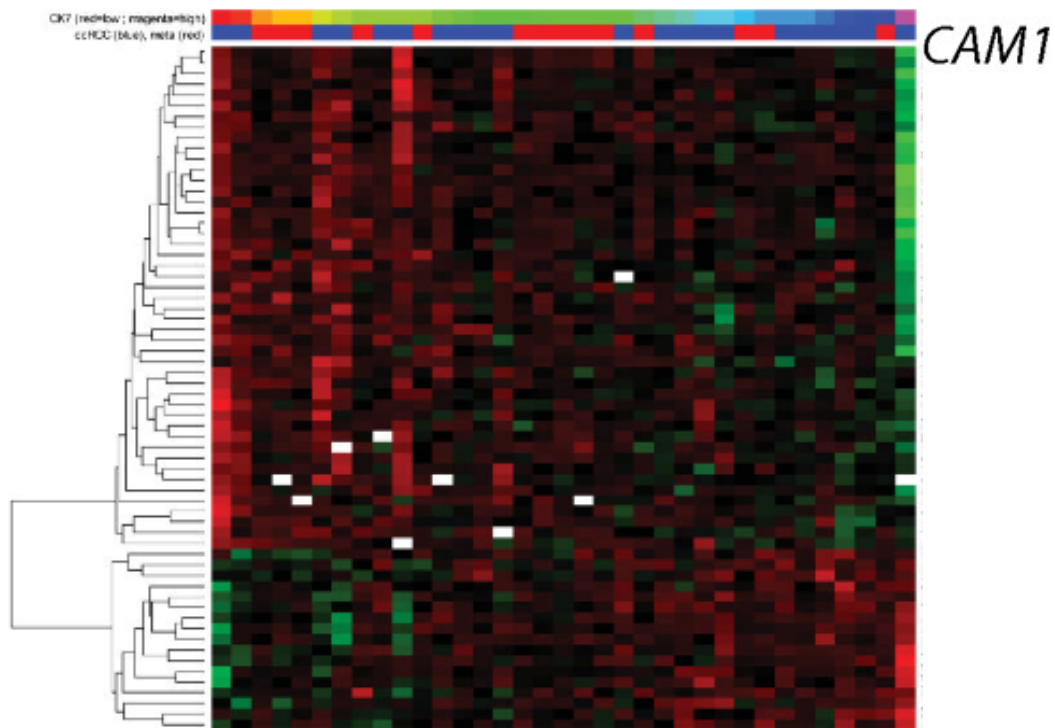
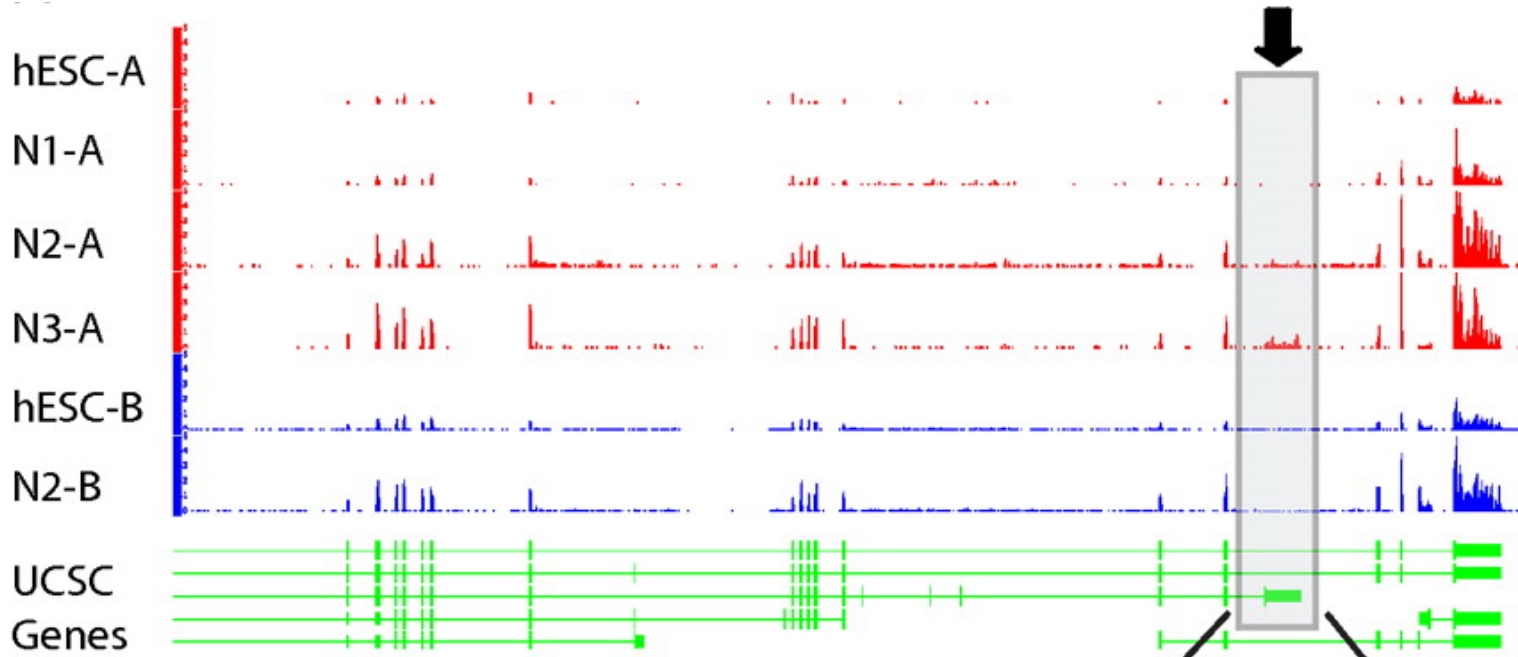
Epigenetically, H3K27ac marks are present near active enhancers.

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" (RPKM)

[PNAS 4:107: 5254 ; JIC 123:569]

Genomic annotations

Introduction

The ENCODE Project provides a set of candidate genomic regions that can serve as predictions for further investigation. This page provides links to download a set of candidate genomic regions as well as a list of publications that contain additional data.

Candidate genomic regions

- Gene expression matrix over ENCODE2 cell lines (~60 cell lines in total) in GENCODE 19 [Download data | Download methods]
- GENCODE v19 TSS list stratified by Fantom5 CAGE data [View README]
 - Strict CAGE clusters [Download]
 - Robust CAGE clusters [Download]
 - Permissive CAGE clusters [Download]
- Candidate enhancers based on DNase hypersensitivity and H3K27ac and annotated with TF-ChIP peaks as well as candidate promoters annotated with TF-ChIP peaks. [Visualize data | Download methods]
 - Distal DNase peaks [Download]
 - Proximal DNase peaks [Download]
 - H3K27ac annotations [Download]
 - Distal TF binding sites [Download]
 - Proximal TF binding sites [Download]

Additional annotations

Papers previously published by the ENCODE Consortium contain data files that include additional genomic annotations. Search for all publications with ENCODE element data

Peaks

Peaks are enriched regions of the genome corresponding to either sites of transcription factor binding or DNase hypersensitivity identified during various functional genomic assays. In this section, we provide a list of peaks in various cell lines using both DNase-Seq and ChIP-Seq assays. [View publications.](#)

RNAs

RNA represents the direct readout of the genetic information encoded by genomes and a significant proportion of a cell's regulatory capabilities are focused on its synthesis, processing, transport, modification and translation. A catalogue of the RNA species made inside the cell and the amount of RNA from each of these loci across various cell lines is provided in this section. [View publications.](#)

Promoters

The promoter is the region proximal to the transcription start site of a gene that regulates its transcription using transcription factor binding sites. These transcription factors recruit RNA polymerase after binding to the promoter and initiate transcription of the gene. [View publications.](#)

Simplified

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

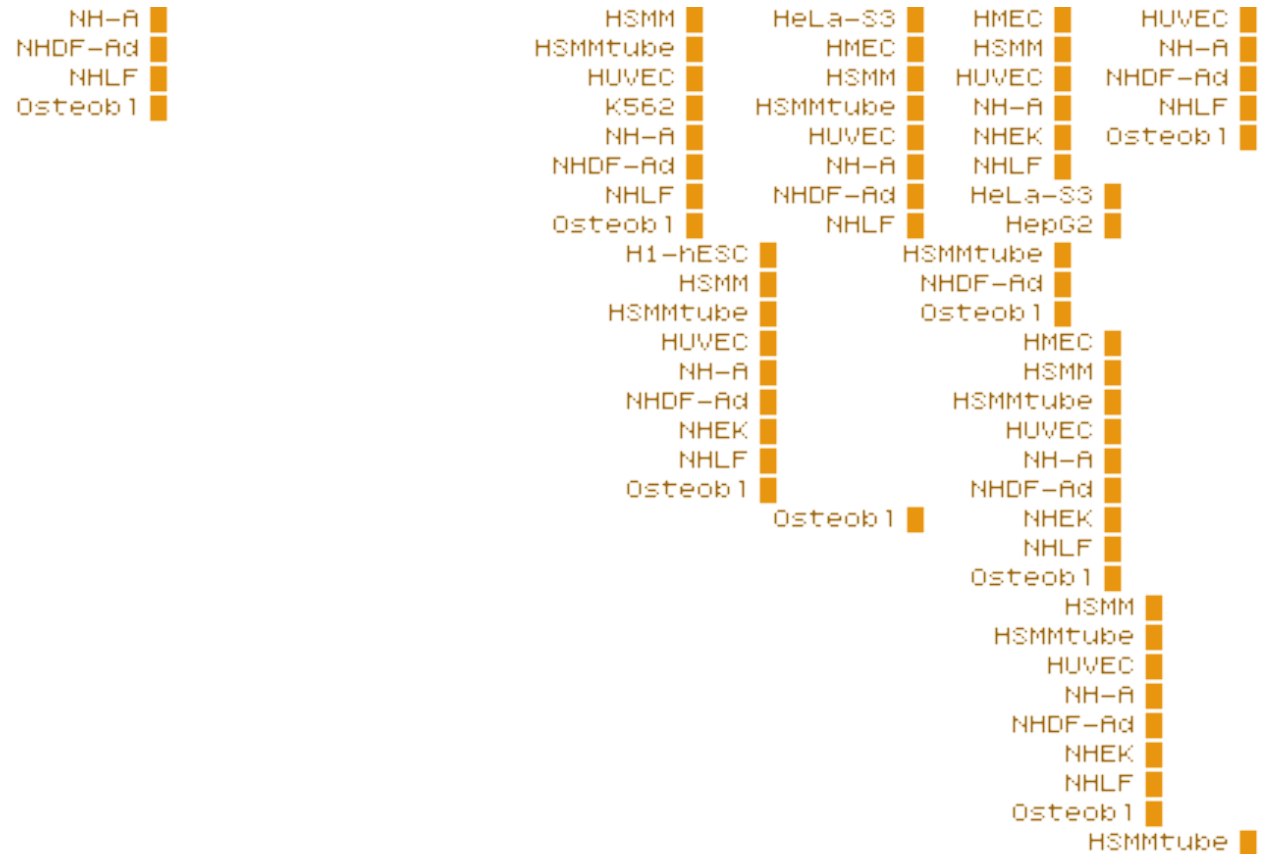
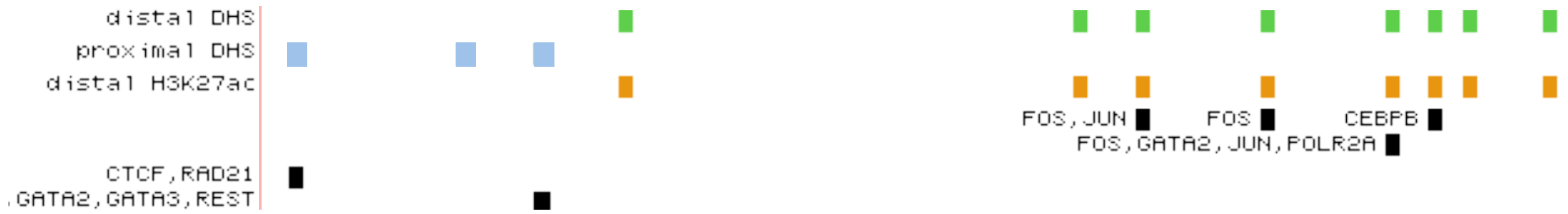
"Simplified" Annotation

- "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.



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

The screenshot shows the ENCODE website interface. The top navigation bar includes 'ENCODE', 'Data', 'Methods', 'About ENCODE', and 'Help'. A search bar and a 'Sign in' link are also present. The 'Data' menu is open, showing options for 'Assays', 'Biosamples', 'Antibodies', 'Annotations', and 'Release policy'. The main content area features the title 'ENCODE Encyclopedia of DNA Elements' and a diagram illustrating the relationship between various genomic annotations and their functional elements. The diagram shows a DNA strand with several regions: 'Long-range regulatory elements (enhancers, repressors/silencers, insulators)', 'Promoters', and 'Transcripts'. Above the DNA, various assays are shown in boxes: '3C ChIA-PET', 'DNase-seq FAIRE-seq', 'ChIP-seq', 'WGBS RRBS methyl450k', 'Computational predictions and RT-PCR', 'RNA-seq', and 'CLIP-seq RIP-seq'. Arrows indicate how these assays map to the functional elements. For example, '3C ChIA-PET' and 'DNase-seq FAIRE-seq' map to 'Long-range regulatory elements'. 'ChIP-seq' and 'WGBS RRBS methyl450k' map to 'Promoters'. 'RNA-seq' and 'CLIP-seq RIP-seq' map to 'Transcripts'. The diagram also shows 'Hypersensitive Sites' and 'RNA polymerase' associated with the DNA structure.

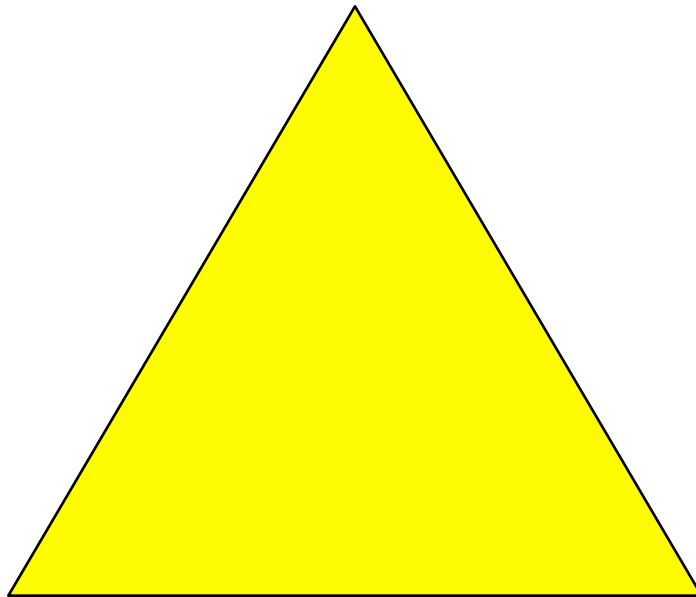
The ENCODE (Encyclopedia of DNA Elements) Consortium is an international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI). The goal of ENCODE is to build a comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active.

Image credits: Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI)

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

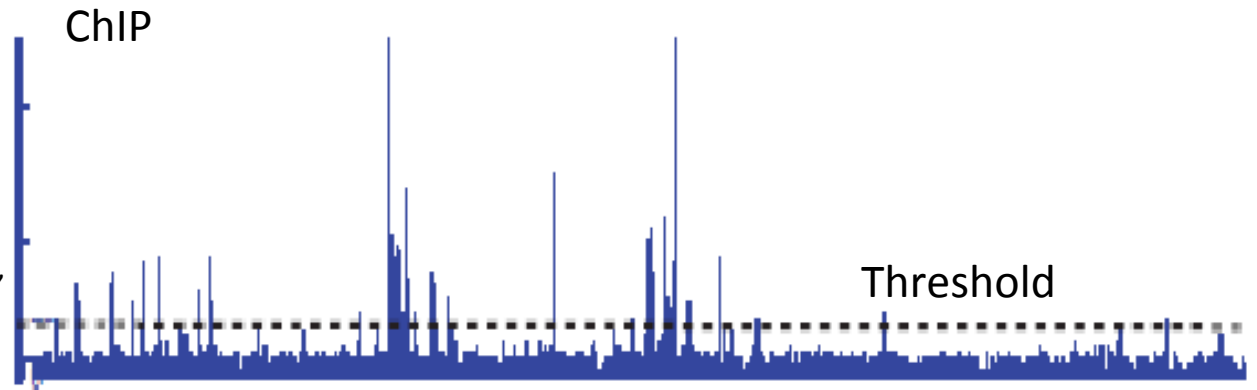
TF annotation

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

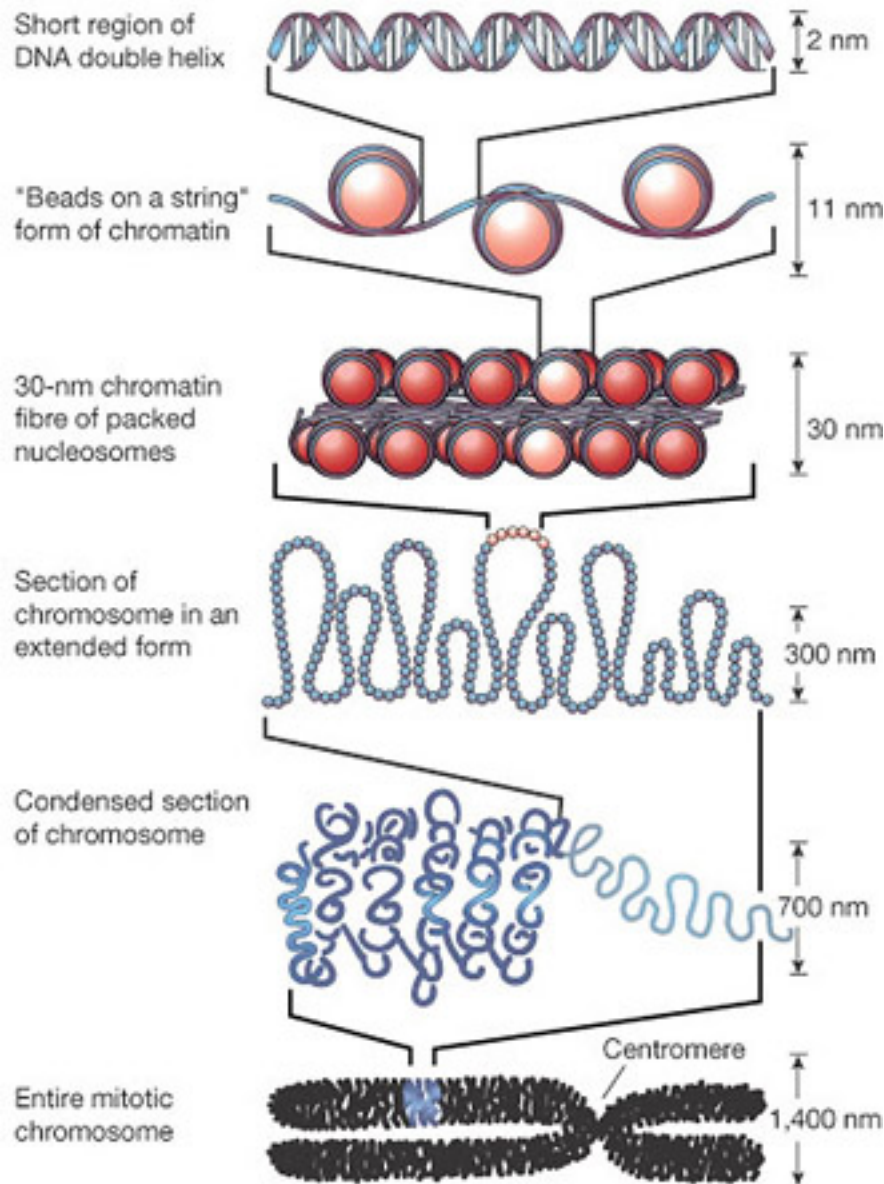


- Score against the control

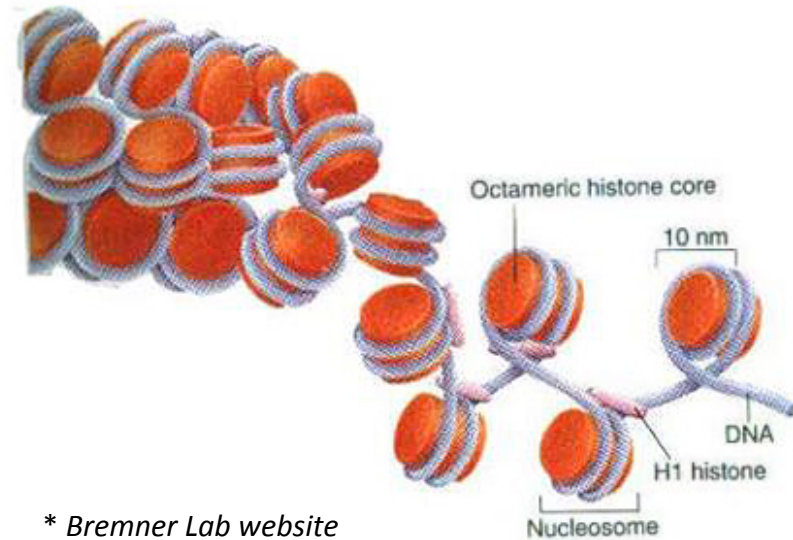


Significantly Enriched targets





Chromatin structure



* *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.