Computational analysis of the ENCODE datasets

and other related epigenetic explorations

Ved Topkar

Harvard College class of 2016

Gunawardena Lab
Harvard Medical School
Department of Systems Biology
13 August 2013

Presentation Goals

- FULL understanding of discussed material
- Ask questions along the way!

Outline

Outline

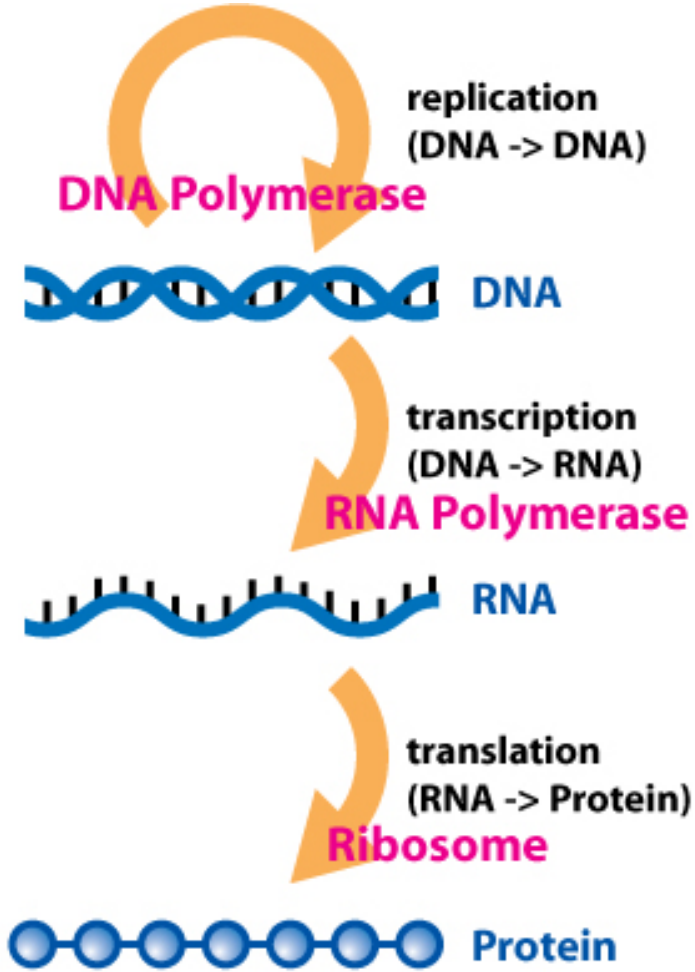
1. Molecular biology in a jiffy
2. A case study
   - Hypothesis formulation
   - Analyzing data
3. More examples

Outline

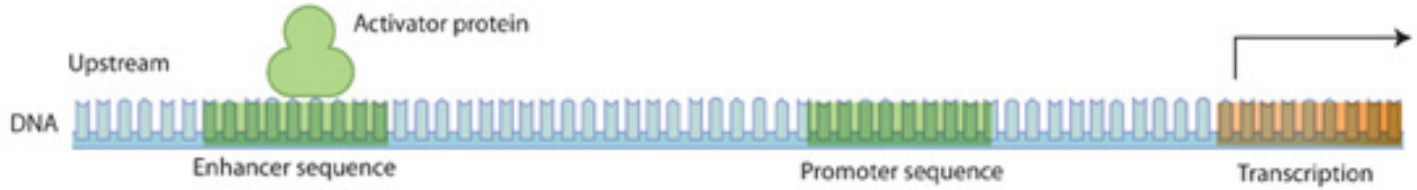
1. Molecular biology in a jiffy
2. A case study
   - Hypothesis formulation
   - Analyzing data
3. More examples

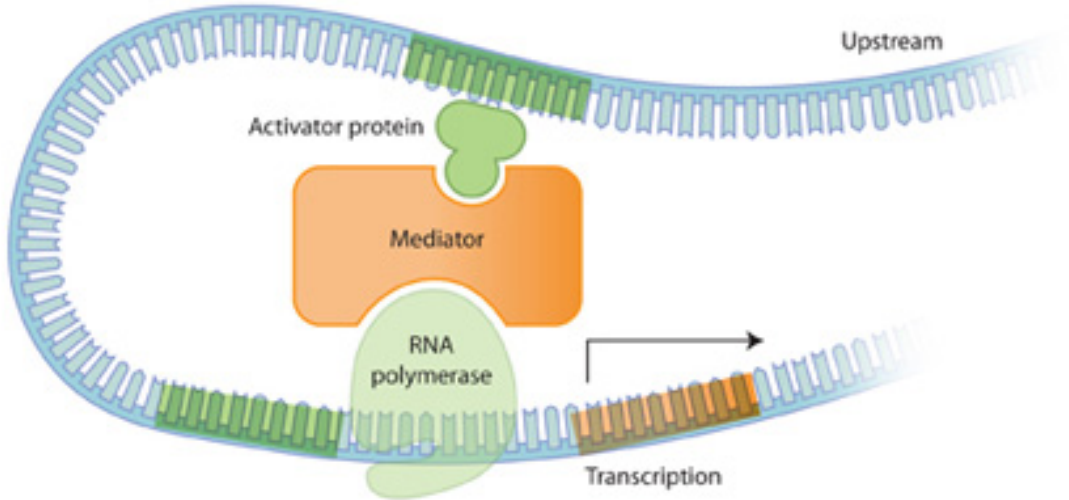
The Cell

The Central Dogma

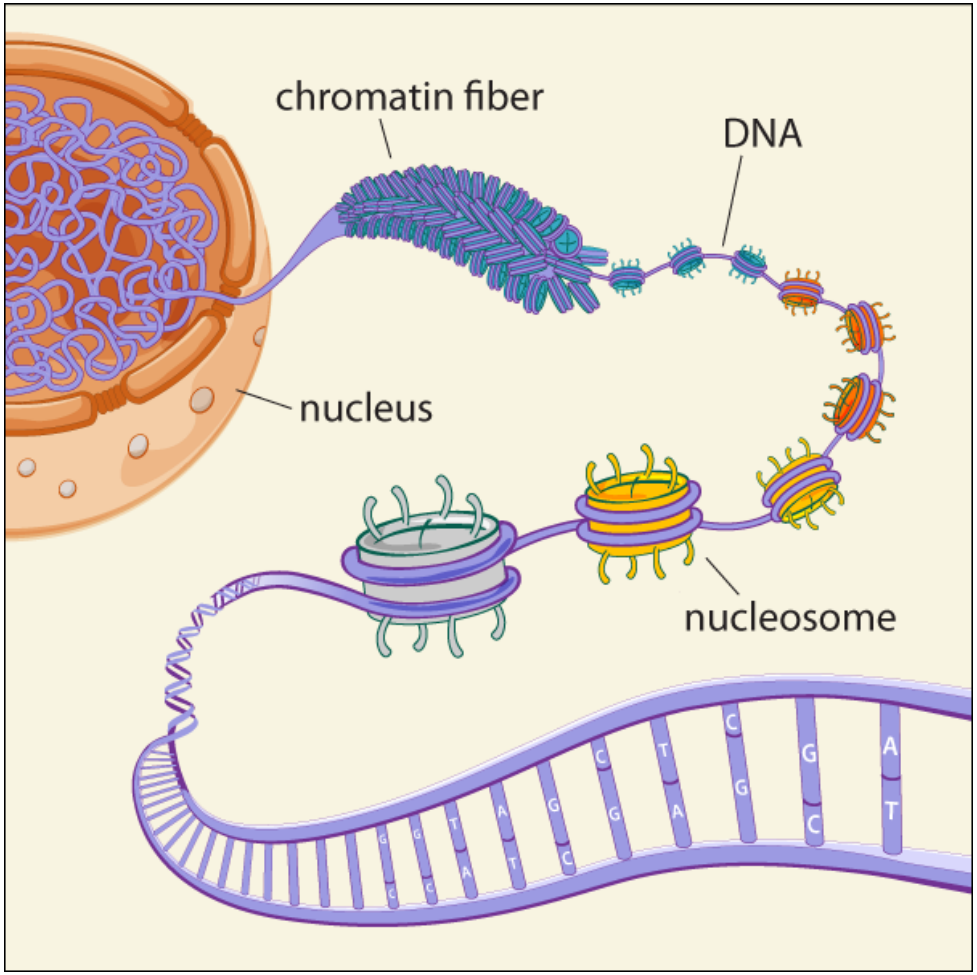


Transcriptional Regulation





Transcriptional Access



Epigenetics and Gene Expression

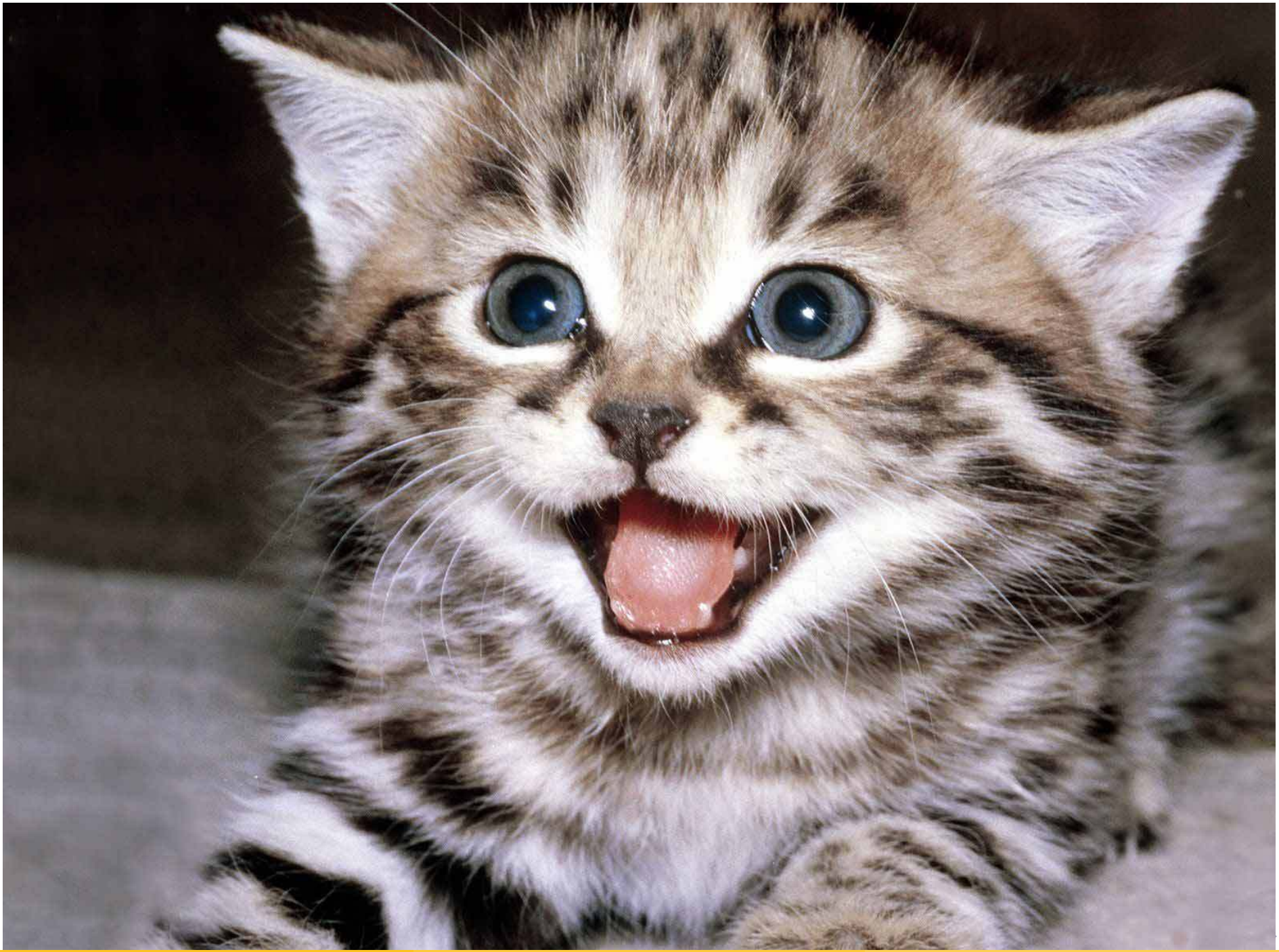
Things beyond just the base pairs in DNA matter $\rightarrow$ gene expression

The Question

Analyze the ENCODE dataset

The Question

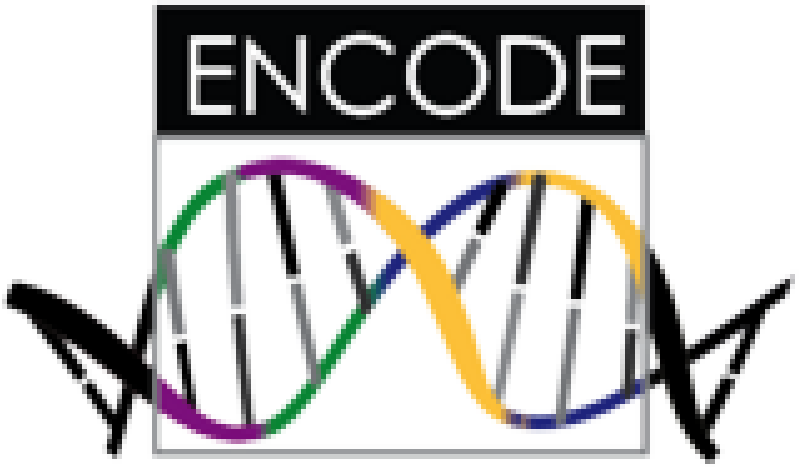
Analyze the ENCODE dataset



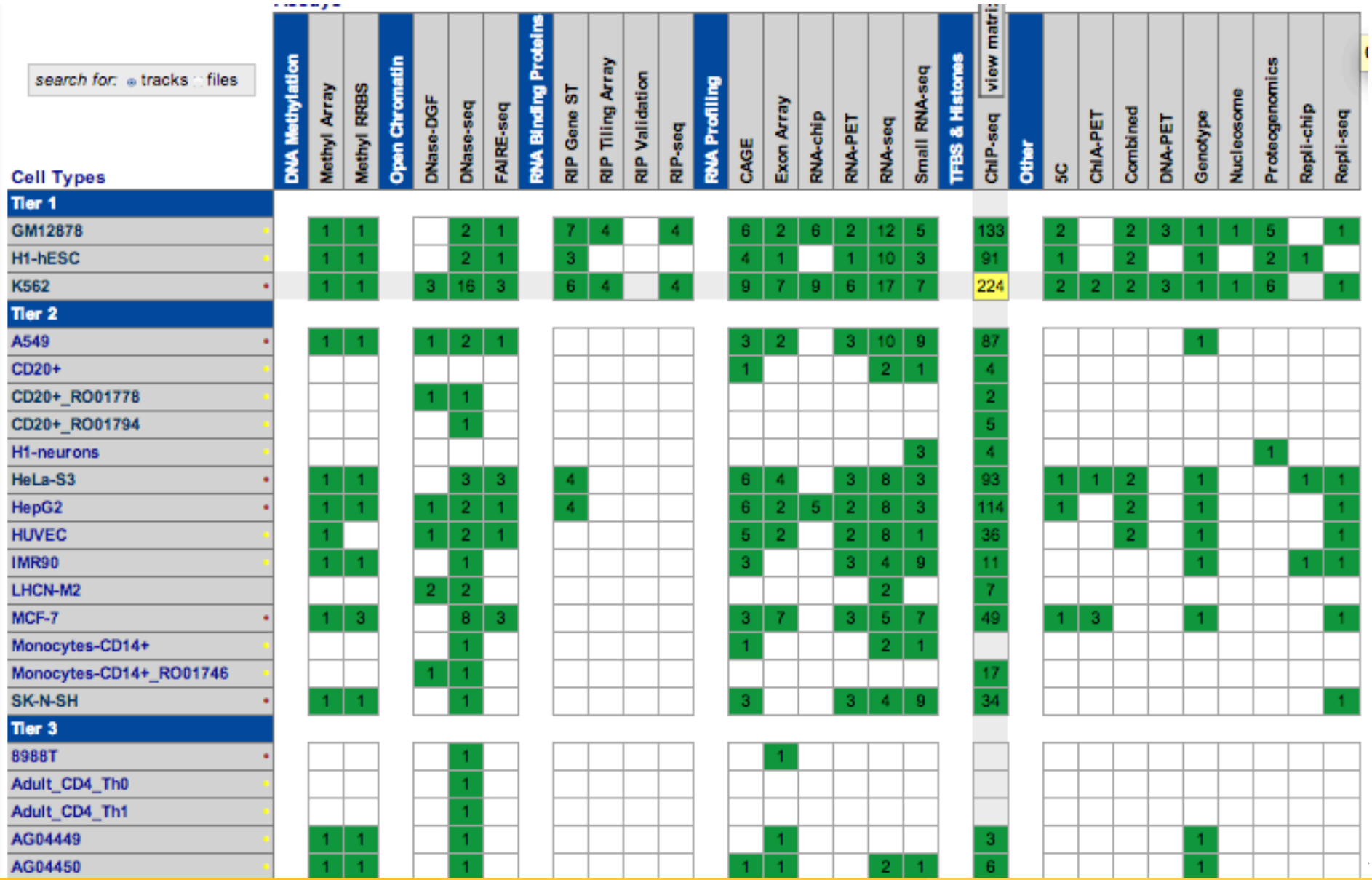
ENCODE (Overview)

Overview

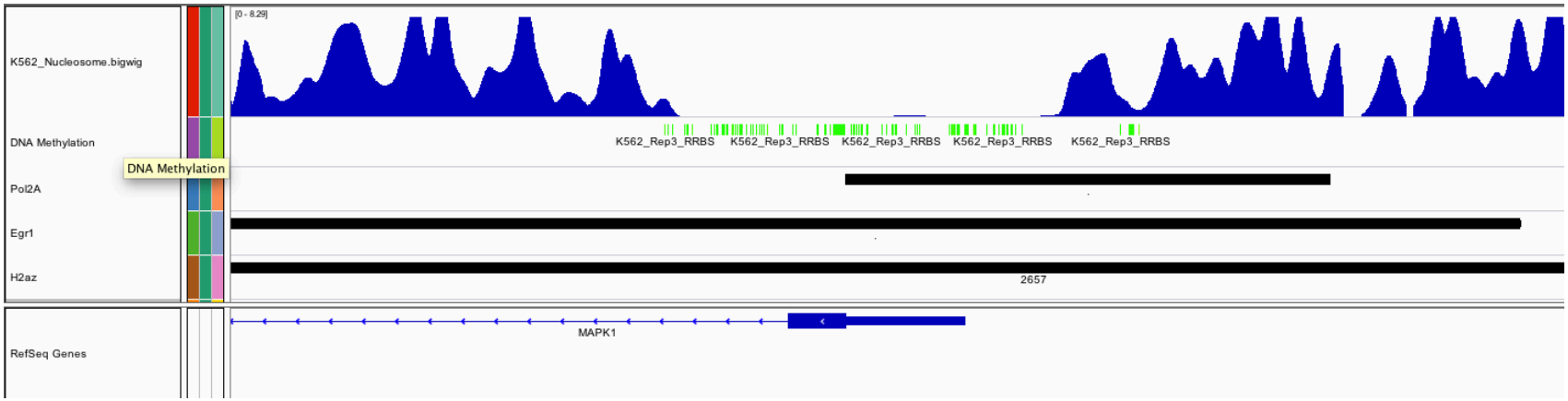
- National Human Genome Institute: Encyclopedia of DNA Elements (ENCODE)
- Nearly 600 collaborating labs post HGP



The Data Set


search for: tracks files
DNA Methylation
Methyl Array
Methyl RRBS
Open Chromatin
DNase-DGF
DNase-seq
FAIRE-seq
RNA Binding Proteins
RIP Gene ST
RIP Tiling Array
RIP Validation
RIP-seq
RNA Profiling
CAGE
Exon Array
RNA-chip
RNA-PET
RNA-seq
Small RNA-seq
TFBS & Histones
ChIP-seq
view matrix
Other
5C
ChIA-PET
Combined
DNA-PET
Genotype
Nucleosome
Proteogenomics
Repli-chip
Repli-seq
Cell Types
Tier 1
GM12878 1 1 2 1 7 4 4 6 2 6 2 12 5 133 2 2 3 1 1 5 1
H1-hESC 1 1 2 1 3 4 1 1 10 3 91 1 2 1 2 1
K562 1 1 3 16 3 6 4 4 9 7 9 6 17 7 224 2 2 2 3 1 1 6 1
Tier 2
A549 1 1 1 2 1 3 2 3 10 9 87 1
CD20+ 1 2 1 4
CD20+_RO01778 1 1 2
CD20+_RO01794 1 5
H1-neurons 3 4 1
HeLa-S3 1 1 3 3 4 6 4 3 8 3 93 1 1 2 1 1 1
HepG2 1 1 1 2 1 4 6 2 5 2 8 3 114 1 2 1 1
HUVEC 1 1 2 1 5 2 2 8 1 36 2 1 1
IMR90 1 1 1 3 3 4 9 11 1 1 1
LHCN-M2 2 2 2 7
MCF-7 1 3 8 3 3 7 3 5 7 49 1 3 1 1
Monocytes-CD14+ 1 1 2 1
Monocytes-CD14+_RO01746 1 1 17
SK-N-SH 1 1 1 3 3 4 9 34 1
Tier 3
8988T 1 1
Adult_CD4_Th0 1
Adult_CD4_Th1 1
AG04449 1 1 1 1 3 1
AG04450 1 1 1 1 1 2 1 6 1

Data Types



- Raw signals
- Raw signal peak calling outputs (e.g. PeakSeq results)
- Relatively course-grain peak data

The Game Plan

Can we reduce transcription factor binding landscapes into categories?

- Scan across genome, looking for promoters
- Bin promoters appropriately
- Score binding at each promoter
- Clustering analysis

RefSeq

Overview

- Curated database of genes
- New versions released as frequently as Firefox
- Includes pseudogenes, haplotype variations, and predicted genes

Defining Promoter?

- Only upstream from TSS?
- Incredibly far regulatory regions?
- Intronic regulation?
- Post termination regulatory elements?
- **1000 bp upstream**

Defining Promoter?

- Only upstream from TSS?
- Incredibly far regulatory regions?
- Intronic regulation?
- Post termination regulatory elements?
- **1000 bp upstream**

Defining Promoter?

- Only upstream from TSS?
- Incredibly far regulatory regions?
- Intronic regulation?
- Post termination regulatory elements?
- **1000 bp upstream**

Defining Promoter?

- Only upstream from TSS?
- Incredibly far regulatory regions?
- Intronic regulation?
- Post termination regulatory elements?
- 1000 bp upstream

Defining Promoter?

- Only upstream from TSS?
- Incredibly far regulatory regions?
- Intronic regulation?
- Post termination regulatory elements?
- **1000 bp upstream**

Binning

- How do we quantitatively analyze promoter presence?
- Break promoter regions into bins for a finer metric?
- Do we give weights to bins as a function of their position?
- **Single, unweighted 1000 bp bin of counts**

Binning

- How do we quantitatively analyze promoter presence?
- Break promoter regions into bins for a finer metric?
- Do we give weights to bins as a function of their position?
- **Single, unweighted 1000 bp bin of counts**

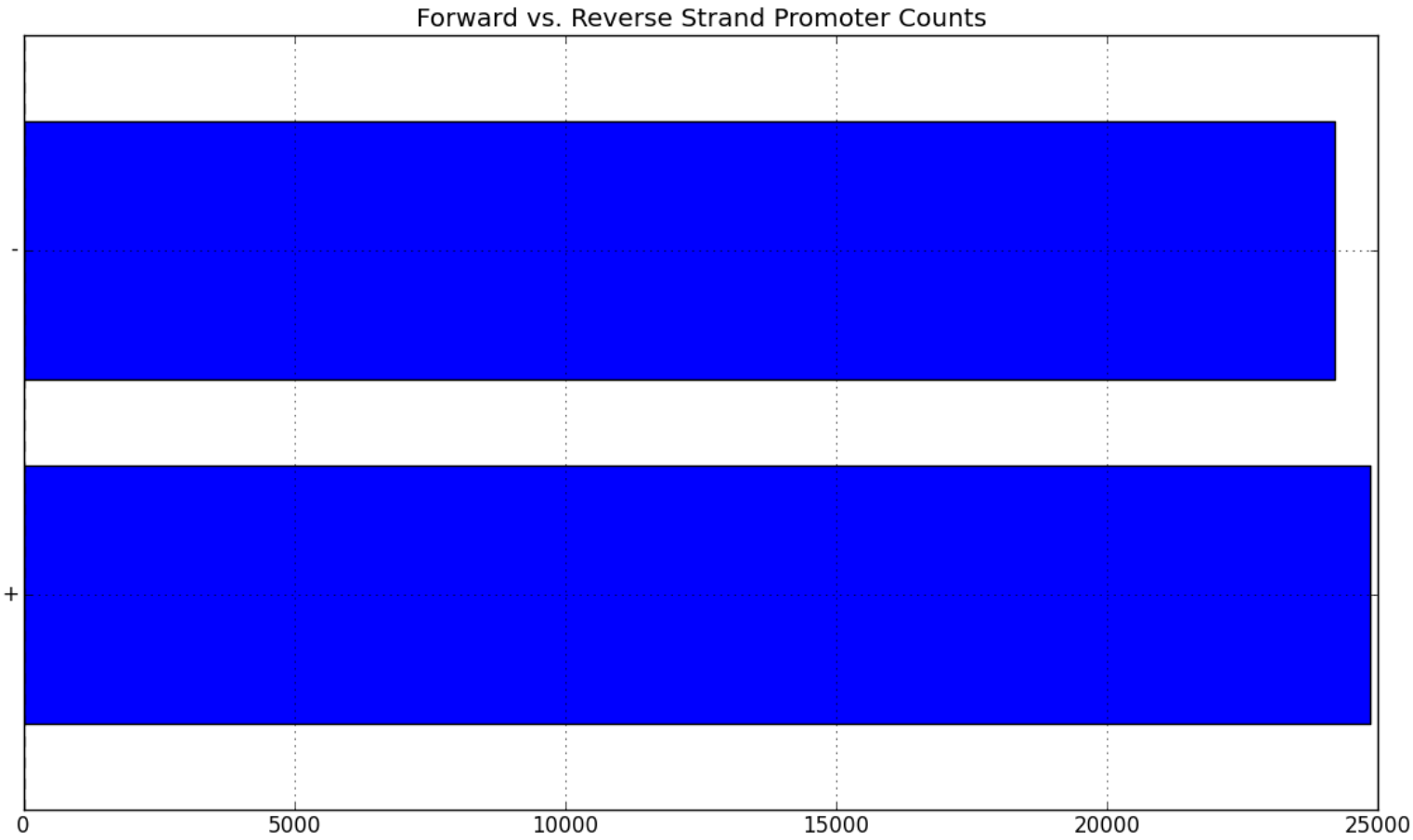
Binning

- How do we quantitatively analyze promoter presence?
- Break promoter regions into bins for a finer metric?
- Do we give weights to bins as a function of their position?
- **Single, unweighted 1000 bp bin of counts**

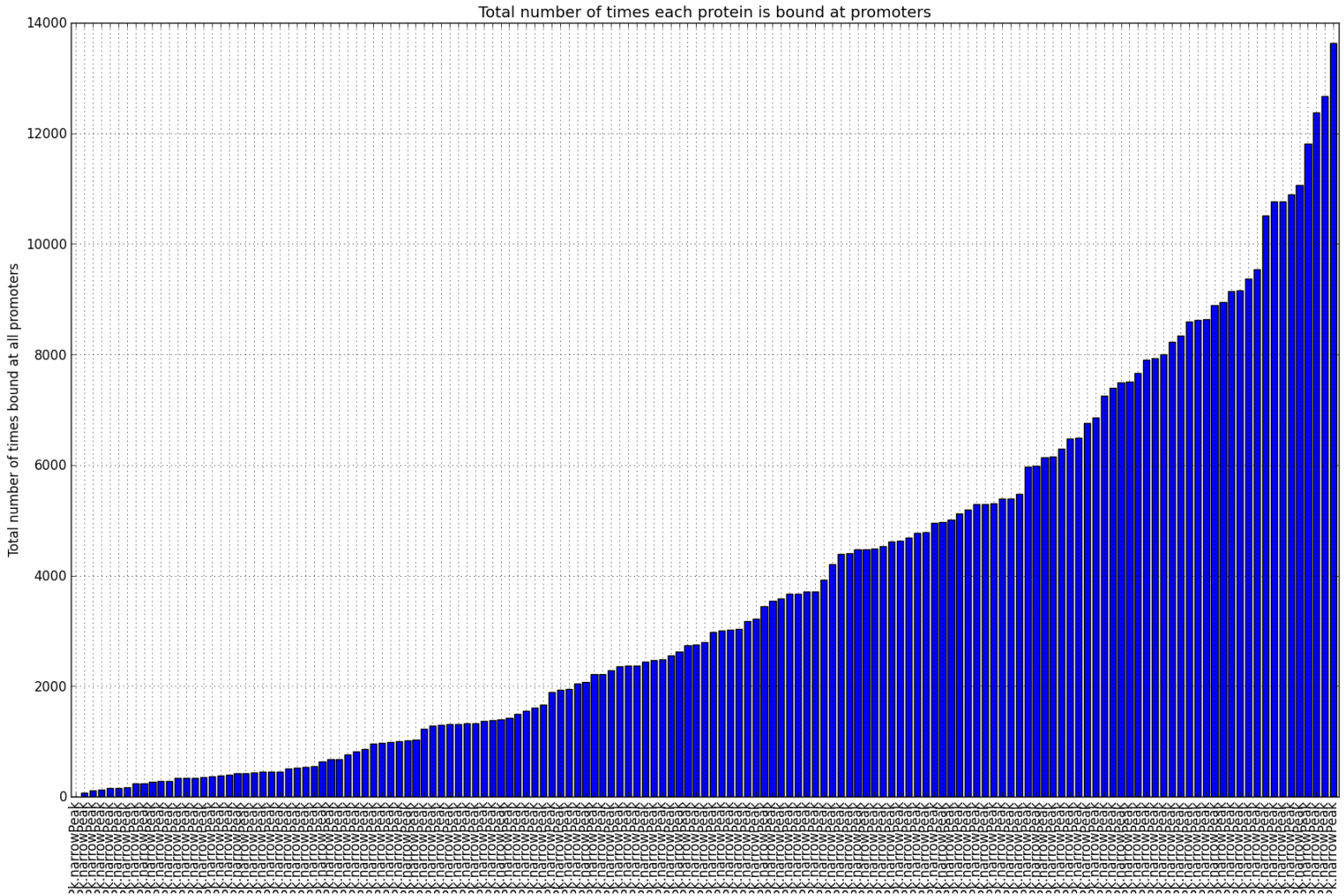
Binning

- How do we quantitatively analyze promoter presence?
- Break promoter regions into bins for a finer metric?
- Do we give weights to bins as a function of their position?
- **Single, unweighted 1000 bp bin of counts**

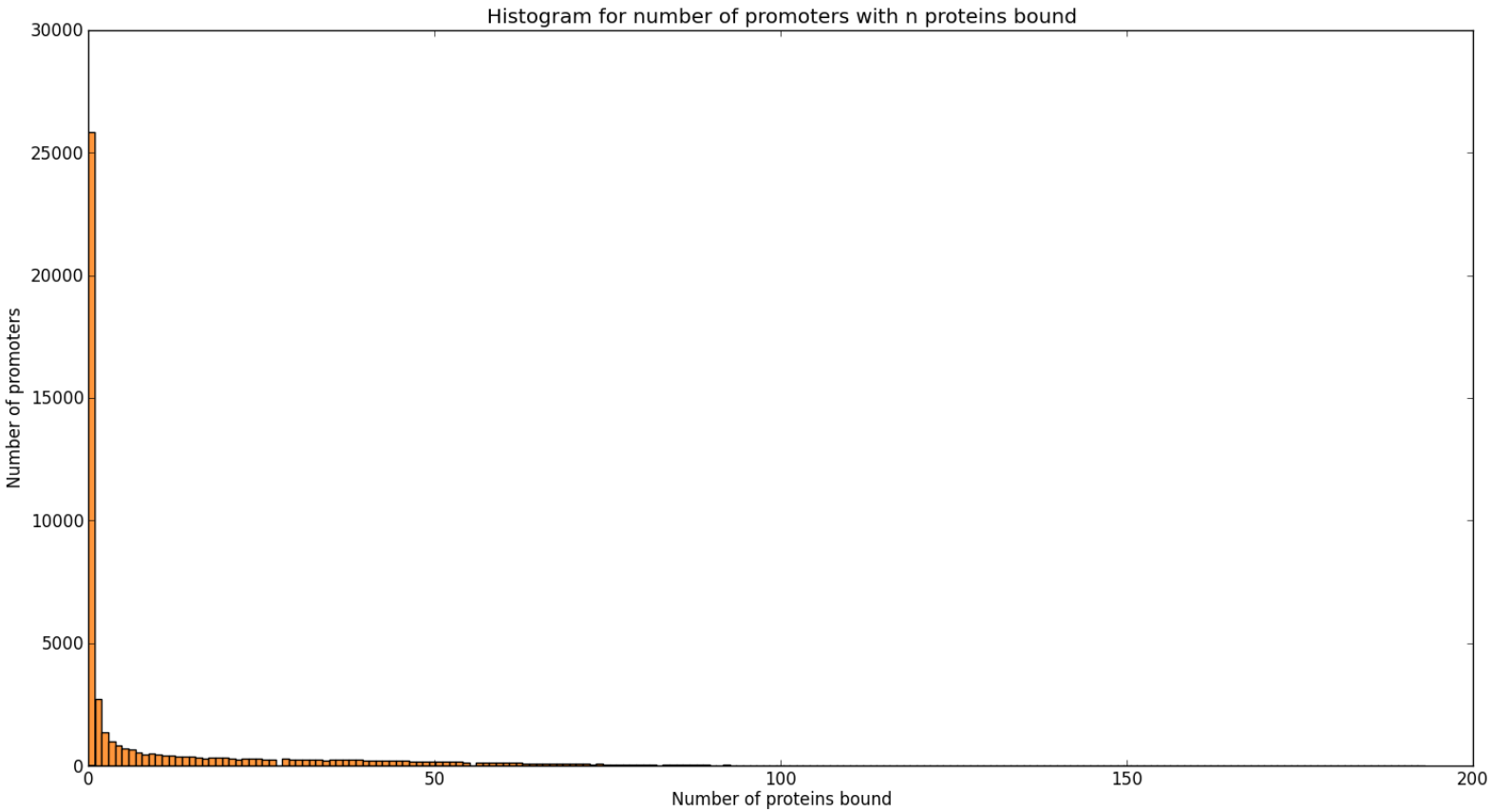
Forward vs. Backward Strand



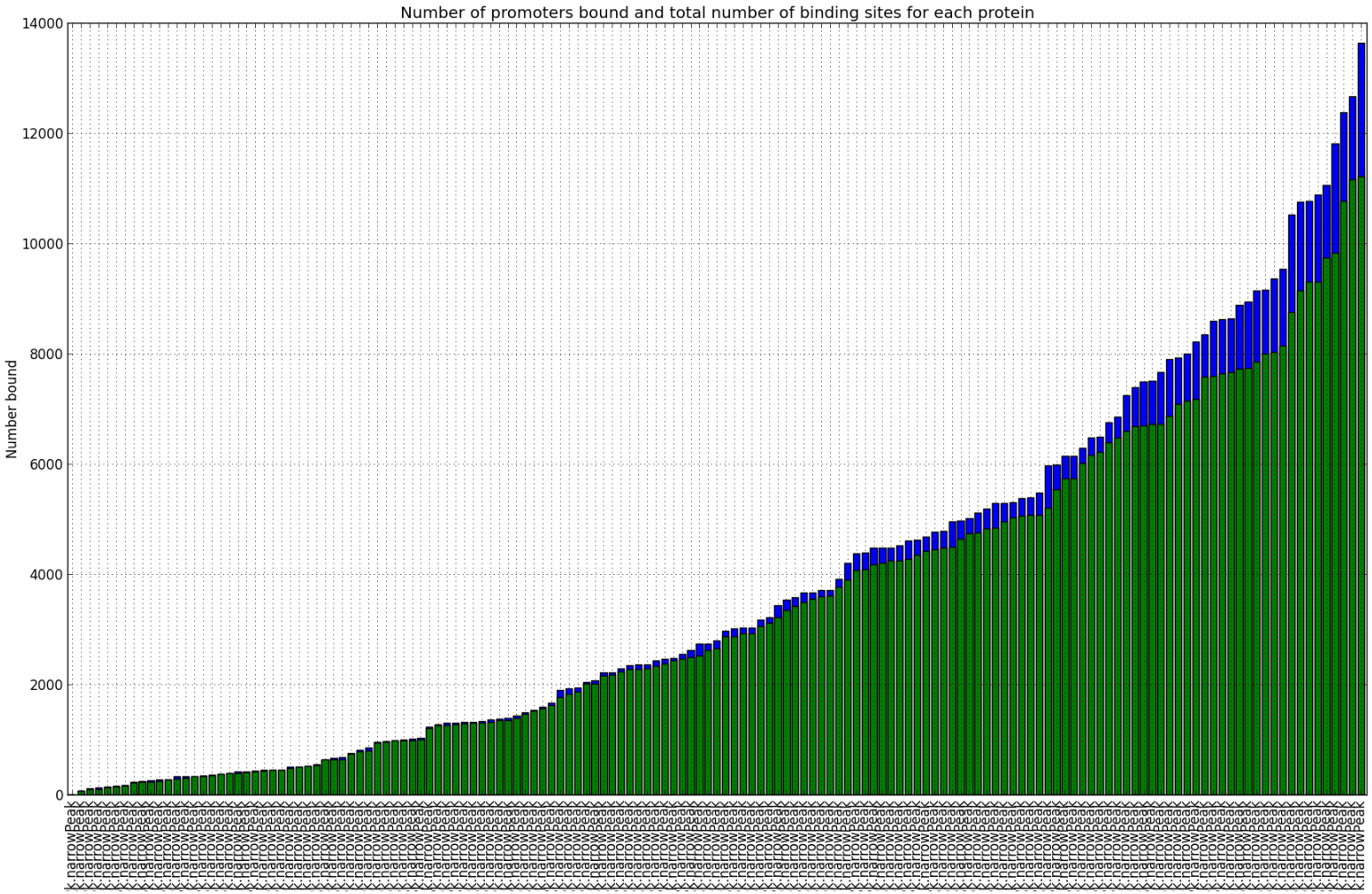
TF Binding Frequency



Histogram of promoter binding

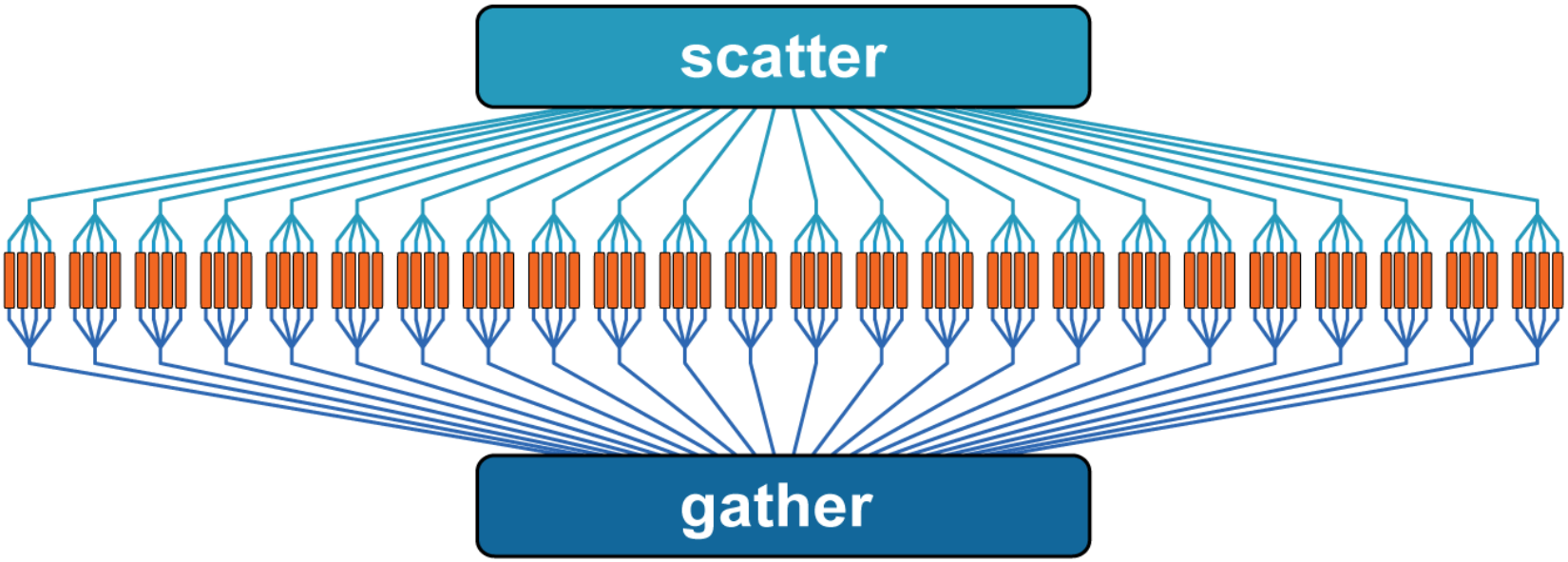


Promoter/TFBS intersections



Computational Efficiency

- This was an exercise in program optimization
- Original algorithm took about 5 days, optimized/parallelized algorithm took just a few hours



Clustering

The unsupervised grouping of information such that groups have similar elements that are dissimilar from elements in other groups

Clustering

Minkowski Distance ($p = 2 \rightarrow$ Euclidean Distance)

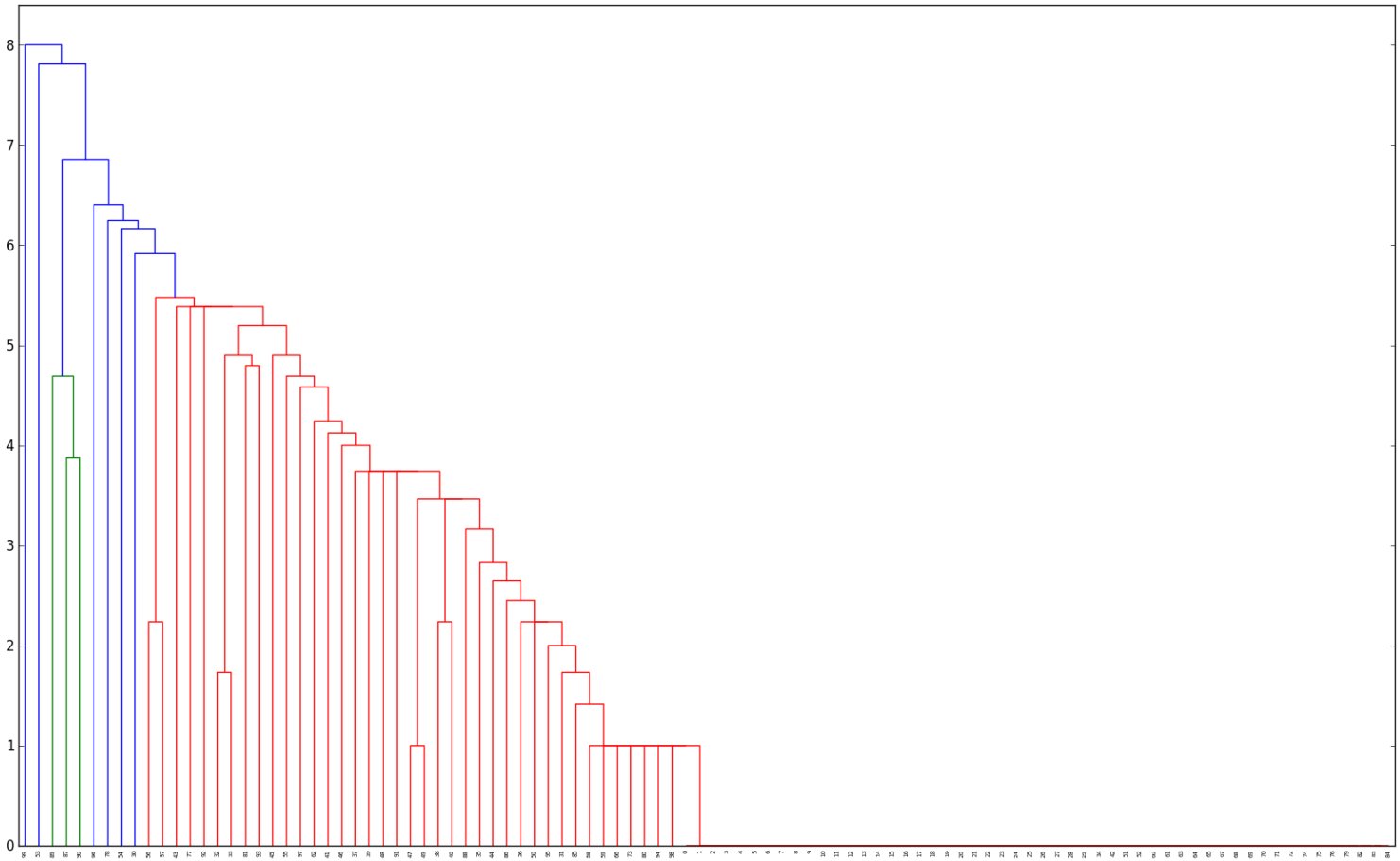
$$\left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}}$$

Clustering

Simple Linkage Clustering

$$D(X, Y) = min(d(x, y)) \; \forall \; (x, y) \epsilon (X, Y)$$

Clustering (simple linkage exhibiting chaining)

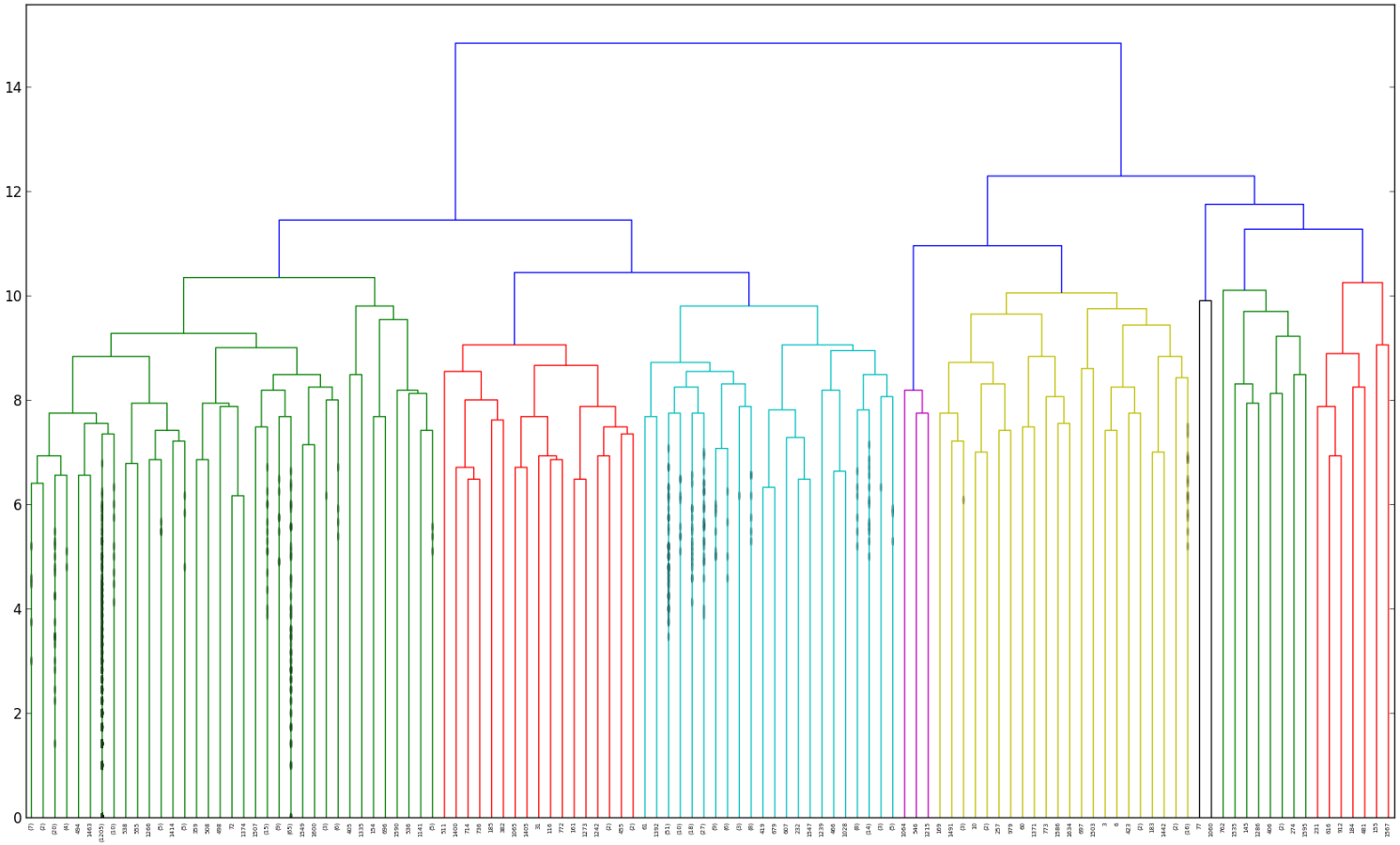


Clustering

Complete Linkage Clustering

$$D(X, Y) = max(d(x, y)) \; \forall \; (x, y) \epsilon (X, Y)$$

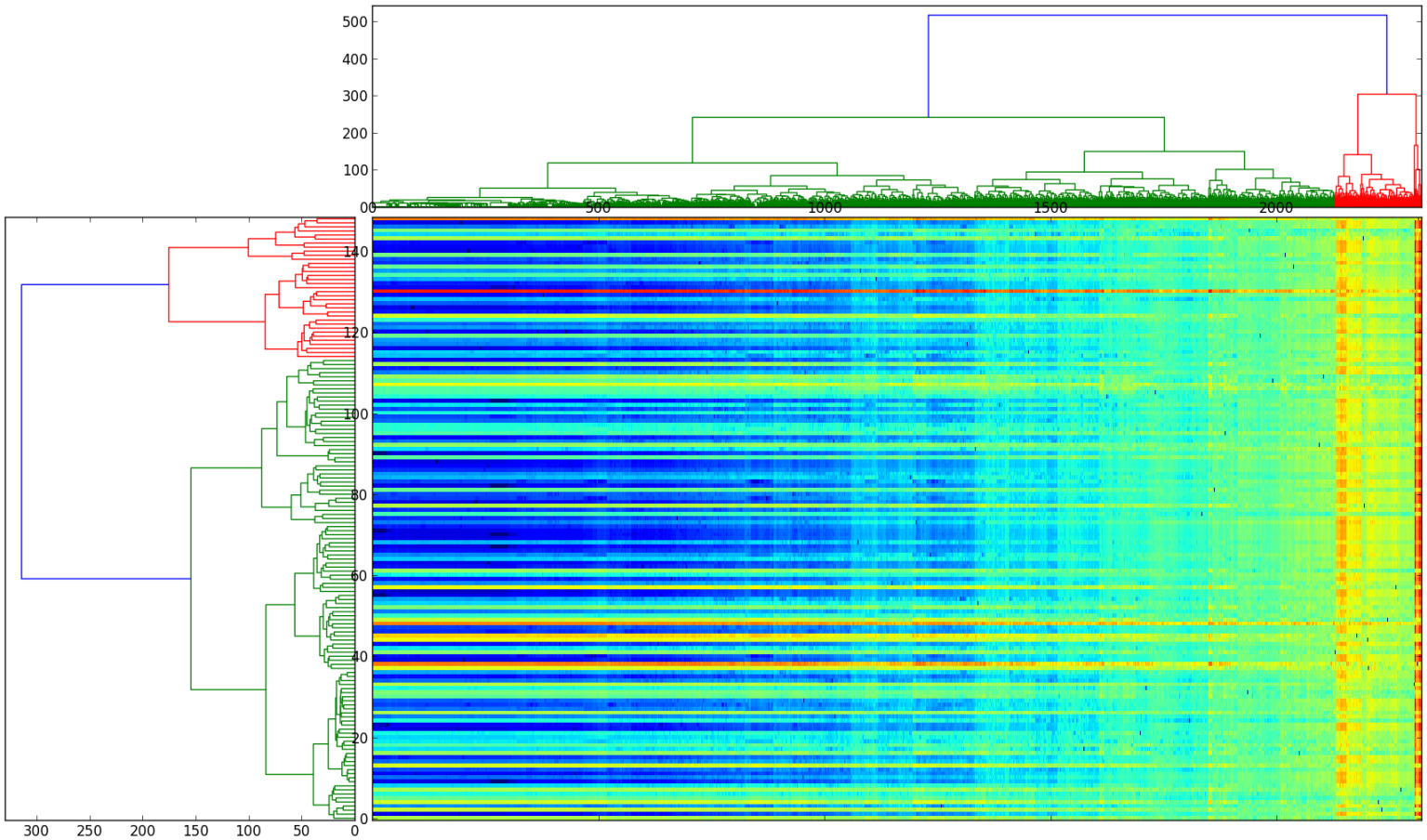
Clustering (complete linkage)



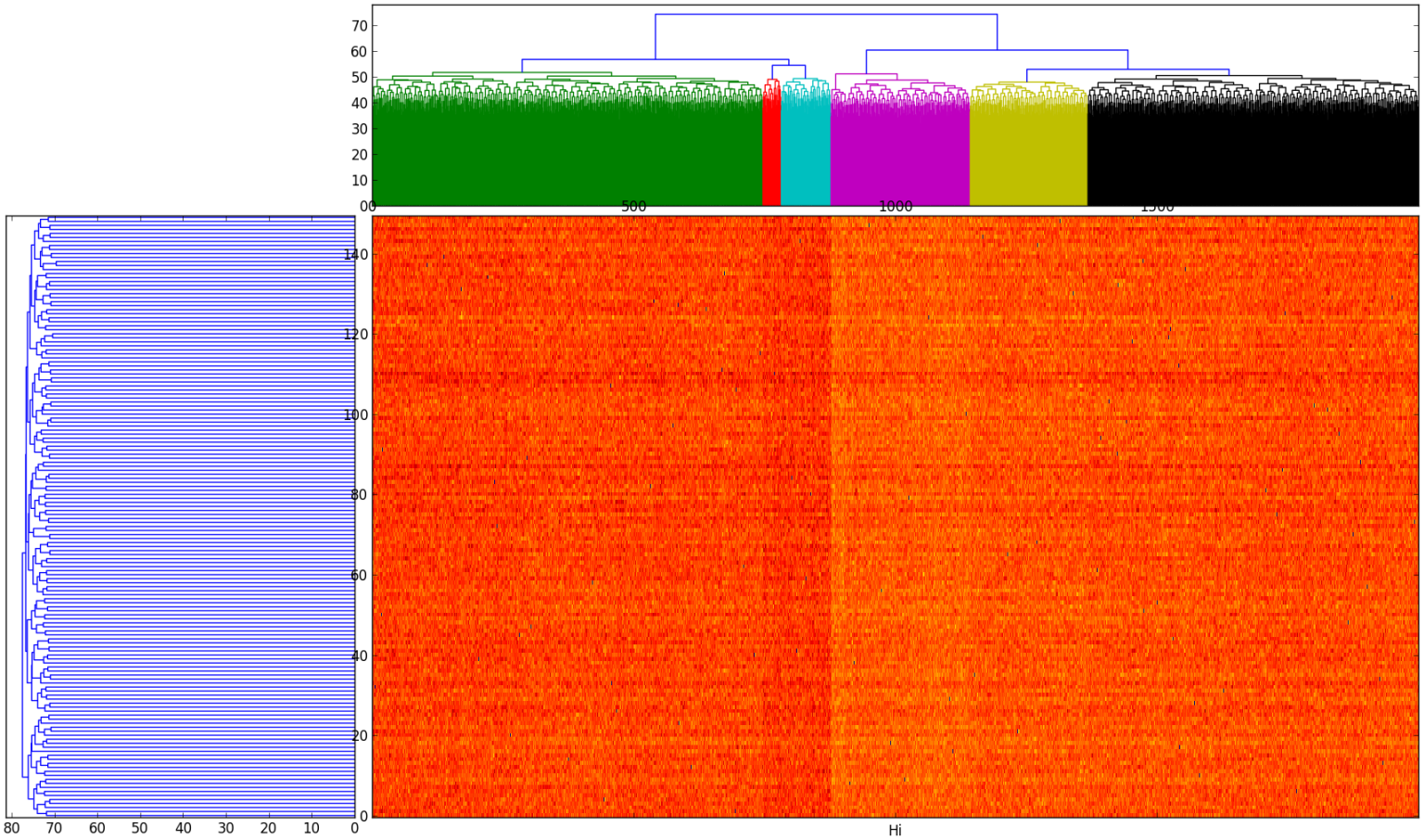
Computational Efficiency

- Pairwise distance calculations require a LOT of RAM

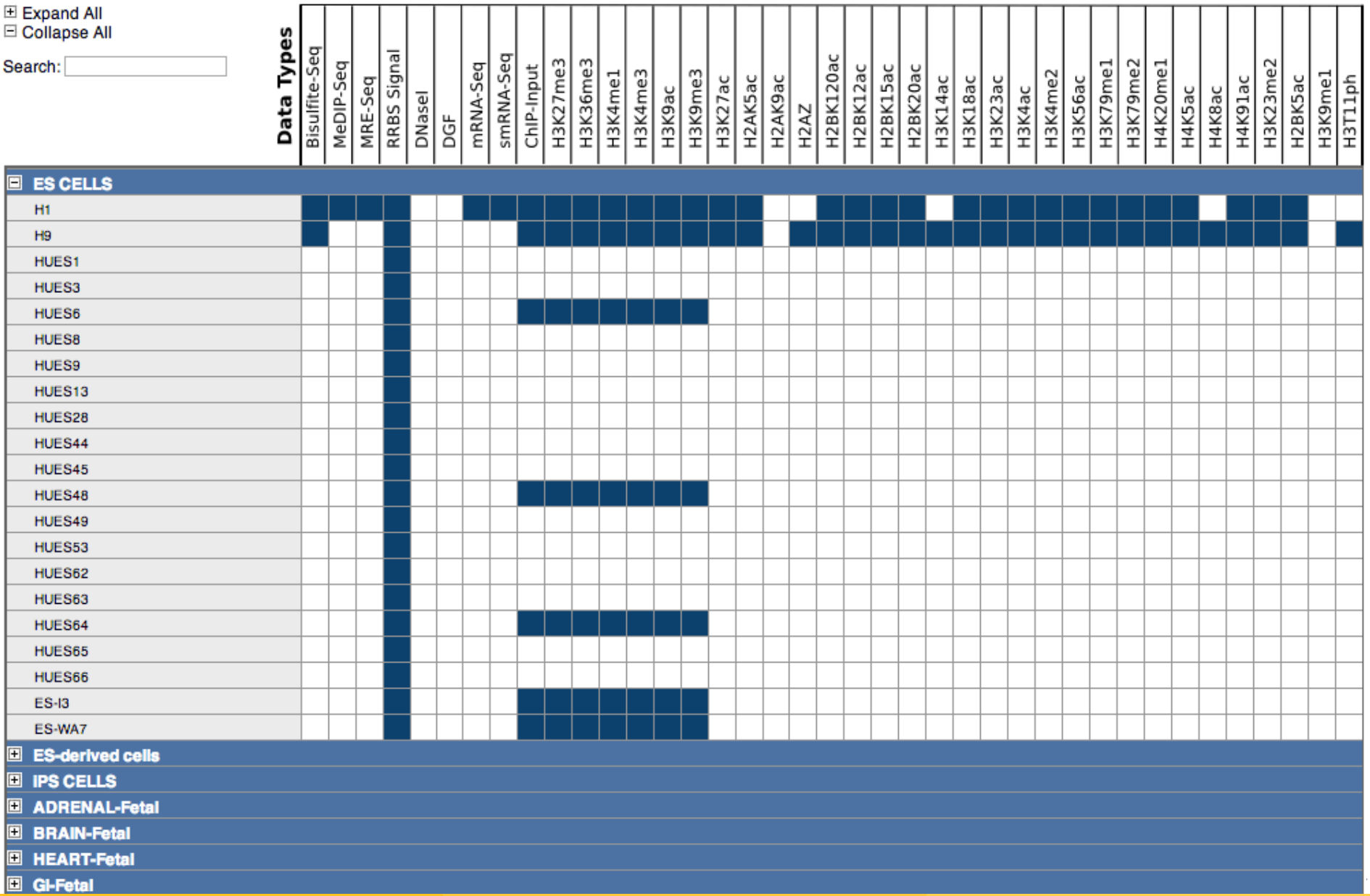
2-way clustering heatmap



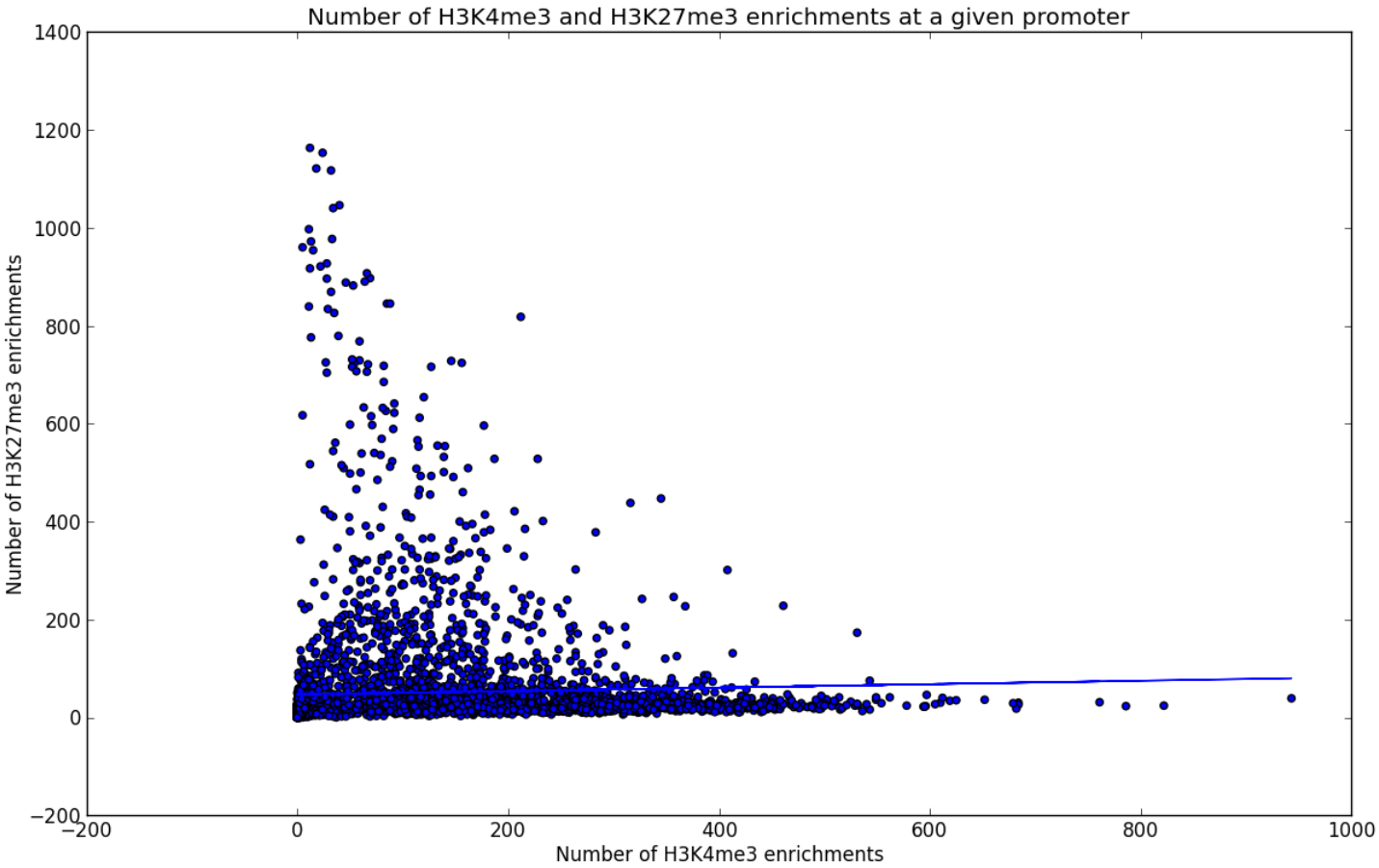
2-way clustering heatmap (with random data arrays)



Histone Modifications (Roadmap Dataset)



A quick correlation test



$R = 0.042$

Conclusions

- There are numerous methods of promoter binning that prove useful for complexity reduction
- Unsupervised clustering of preprocessed promoter data yield results with biological significance

Next Steps

- Incorporation of nucleosome enrichment, methylation, and histone modification data in a more meaningful way
- Further refining of cluster analysis pipeline to uncover more unknown biology

Thanks!

- Jeremy Gunawardena and the HMS Department of Systems Biology!
- My collaborator Tobias Ahsendorf!
- PRISE and Greg Llacer!
- The lovely PRISE staff!
- This audience!