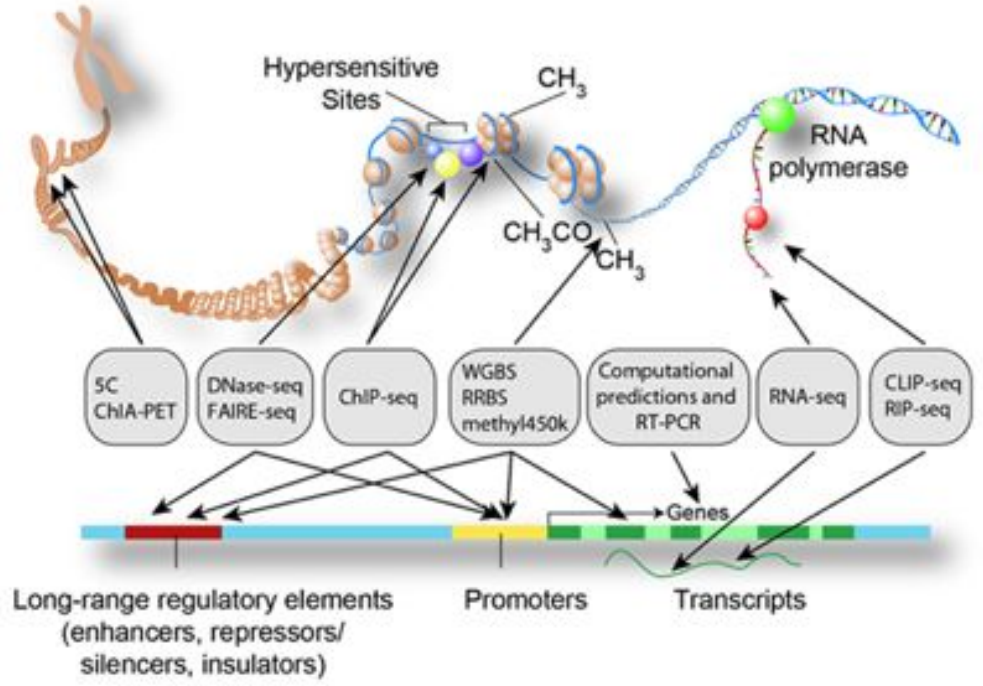
The ENCODE Encyclopedia

Zhiping Weng
University of Massachusetts Medical School
ENCODE 2016: Research Applications and Users Meeting
June 8, 2016

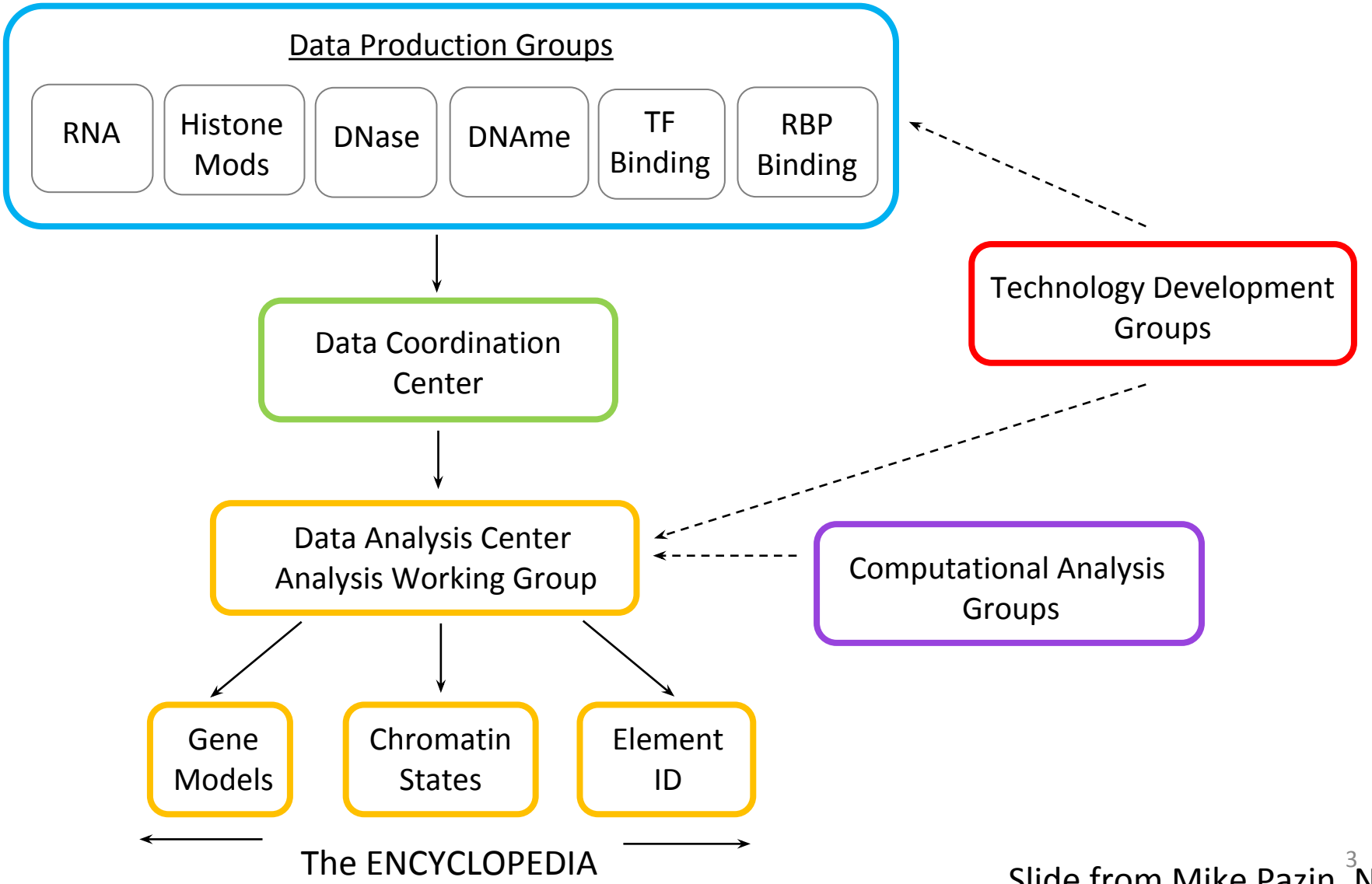
The **Enc**yclopedia **O**f **D**NA **E**lements Consortium



Goals:

- Catalog all functional elements in the genome
- Develop freely available resource for research community
- Study human and mouse

Overview of The ENCODE Consortium



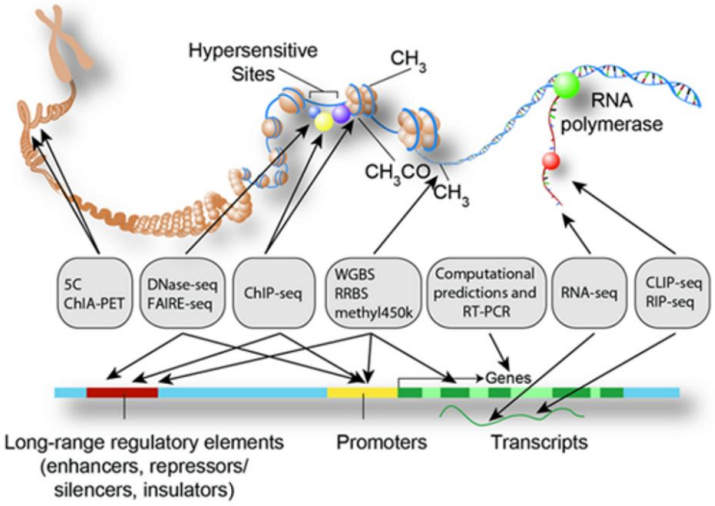
Slide from Mike Pazin, NHGRI

encodeproject.org

ENCODE Data Encyclopedia Materials & Methods Help Search...

About
Matrix
Search

ENCODE …pedia of DNA Elements



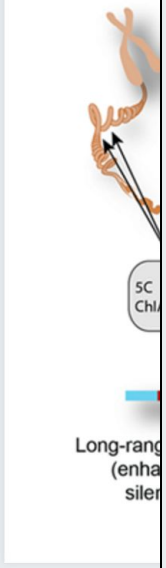
The ENCODE (Encyclopedia of DNA Elements) Consortium is an international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI). The goal of ENCODE is to build a comprehensive parts list of functional elements in the human genome, including elements that act at 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

ENCODE Data Encyclopedia Materials & Methods Help Search...

ENCO

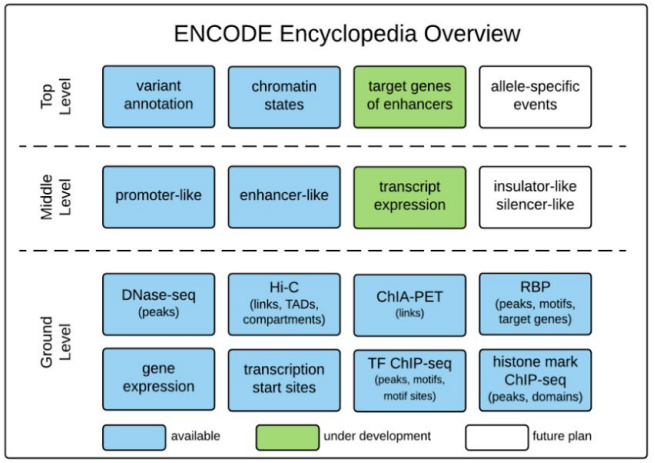


ENCODE Encyclopedia: Genomic annotations

Introduction

The ENCODE Consortium not only produces data, but also analyzes the data in an integrative fashion. The ENCODE Encyclopedia organizes the most salient analysis products into annotations, and provides tools to search and visualize them. The Encyclopedia has three levels of annotations:

- Ground level annotations are typically derived directly from the experimental data.
- Middle level annotations integrate multiple types of experimental data and multiple ground level annotations.
- Top level annotations integrate a broad range of experimental data and ground and middle level annotations.

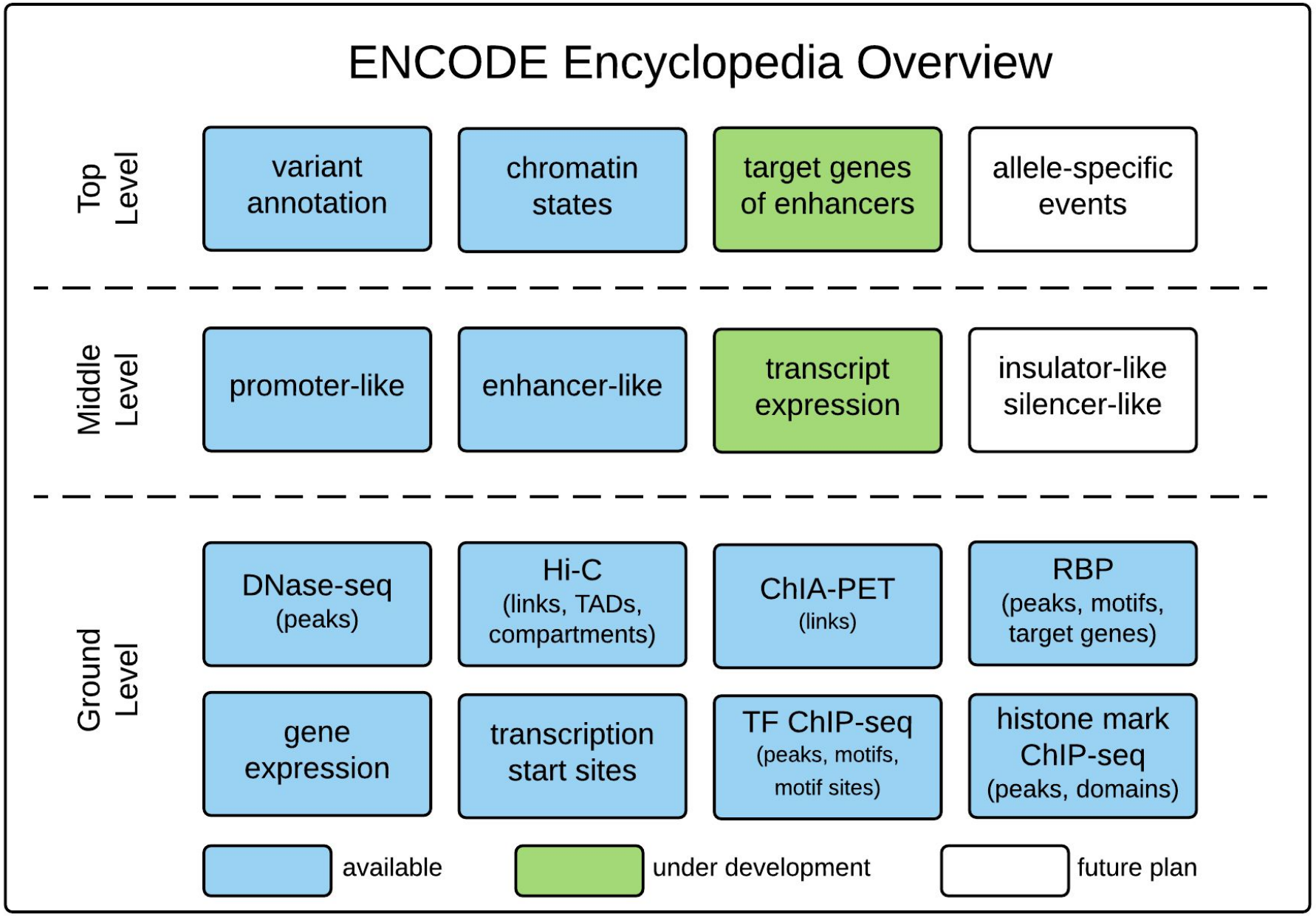


Ground Level Annotations

Gene expression (RNA-seq)

The expression levels of genes annotated by GENCODE 19 in over 100 human cell types and 70 mouse cell types.

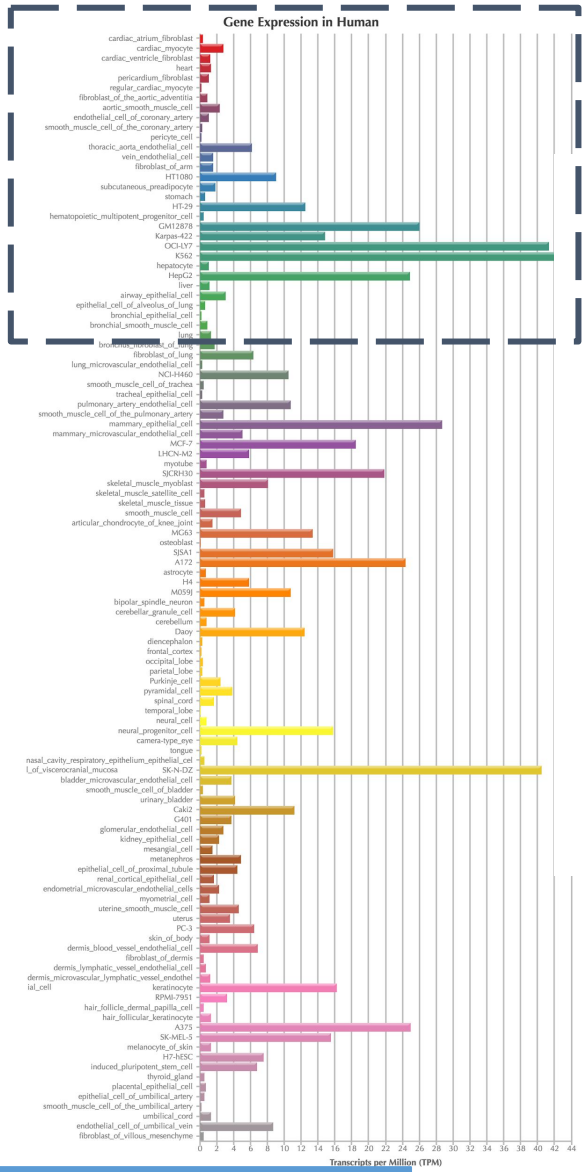
[Long RNA-seq Data | Query | Download | Method]


ENCODE Encyclopedia Overview
Top Level
variant annotation
chromatin states
target genes of enhancers
allele-specific events
Middle Level
promoter-like
enhancer-like
transcript expression
insulator-like silencer-like
Ground Level
DNase-seq (peaks)
Hi-C (links, TADs, compartments)
ChIA-PET (links)
RBP (peaks, motifs, target genes)
gene expression
transcription start sites
TF ChIP-seq (peaks, motifs, motif sites)
histone mark ChIP-seq (peaks, domains)
available
under development
future plan

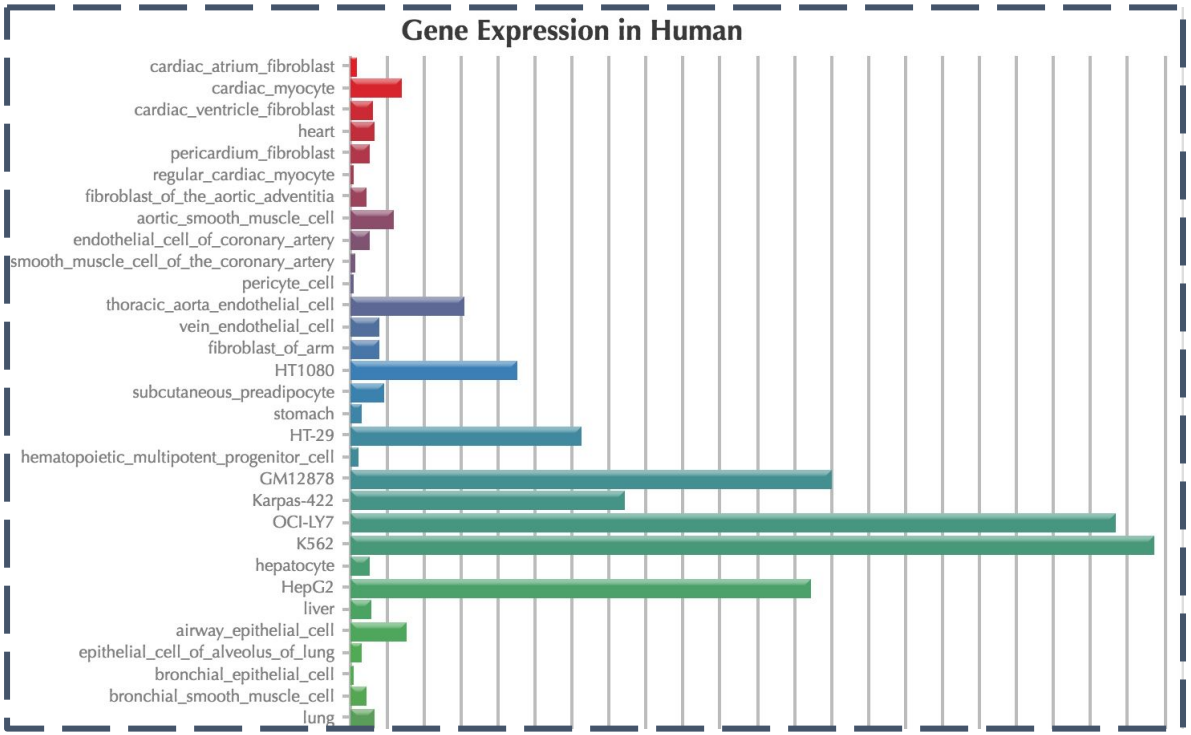
Ground Level Annotations

- Typically derived directly from the experimental data
- Data produced from ENCODE 3 uniform processing pipelines: e.g. peaks and expression quantification

Gene Expression (RNA-seq)



The expression levels of genes annotated by GENCODE 19 in ~60 human cell types.

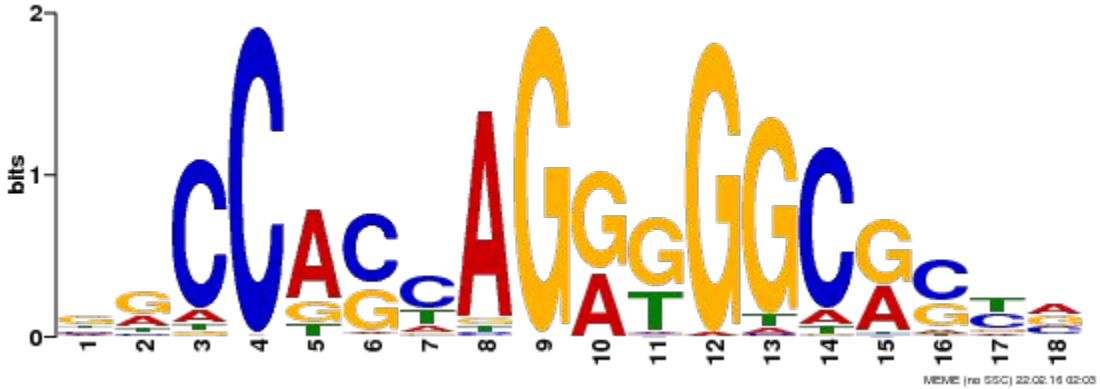


Data produced by: Gingeras, Wold, Lecuyer, Hardison, Graveley

Visualization by: Feng Yue

Transcription Factor Binding (TF ChIP-seq)

Peaks (enriched genomic regions) of TFs computed from ~900 human and mouse ChIP-seq experiments.



Data produced by: Snyder, Myers, Bernstein, Farnham, Stam, Iyer, White, Ren, Struhl, Weissman, Hardison, Wold, Fu

Visualization by: Weng

Ground Level Annotations

Factorbook: Motivation

- Visualizes summarized data centered on TFs
  - not easily shown in a genome browser
  - includes a number of useful analyses and statistical information
    - Average histone profiles
    - Motifs
    - Heat maps
- Transcription Factor (TF)-centric repository of all ENCODE ChIP-seq datasets on TF-binding regions
- Will also visualize ChIP-seq Histone and DNase-seq datasets from ENCODE and ROADMAP soon!

ENCODE ChIP-seq TF Datasets

- Human:
  - 837 ChIP-seq TF datasets
    - 167 TFs
    - 104 cell types
- Mouse:
  - 170 ChIP-seq TF datasets
    - 51 TFs
    - 26 cell types

Last data import: February 29, 2016

Function

Factorbook | human | mouse

BACH1

Function | Histone Profiles | Motif Enrichment | Histone Heatmaps | TF Heatmaps | Nucleosome Profiles

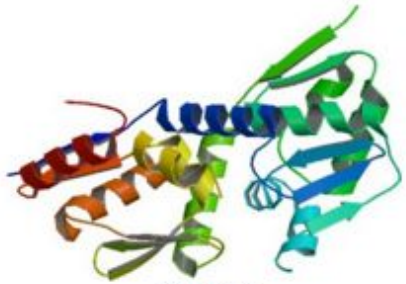
This gene encodes a transcription factor that belongs to the cap'n'collar type of basic region leucine zipper factor family (CNC-bZip). The encoded protein contains broad complex, tramtrack, bric-a-brac/poxvirus and zinc finger (BTB/POZ) domains, which is atypical of CNC-bZip family members. These BTB/POZ domains facilitate protein-protein interactions and formation of homo- and/or hetero-oligomers. When this encoded protein forms a heterodimer with MafK, it functions as a repressor of Maf recognition element (MARE) and transcription is repressed. Multiple alternatively spliced transcript variants have been identified for this gene.

— RefSeq, May 2009

Transcription regulator protein BACH1 is a protein that in humans is encoded by the BACH1 gene.

— wikipedia

BACH1

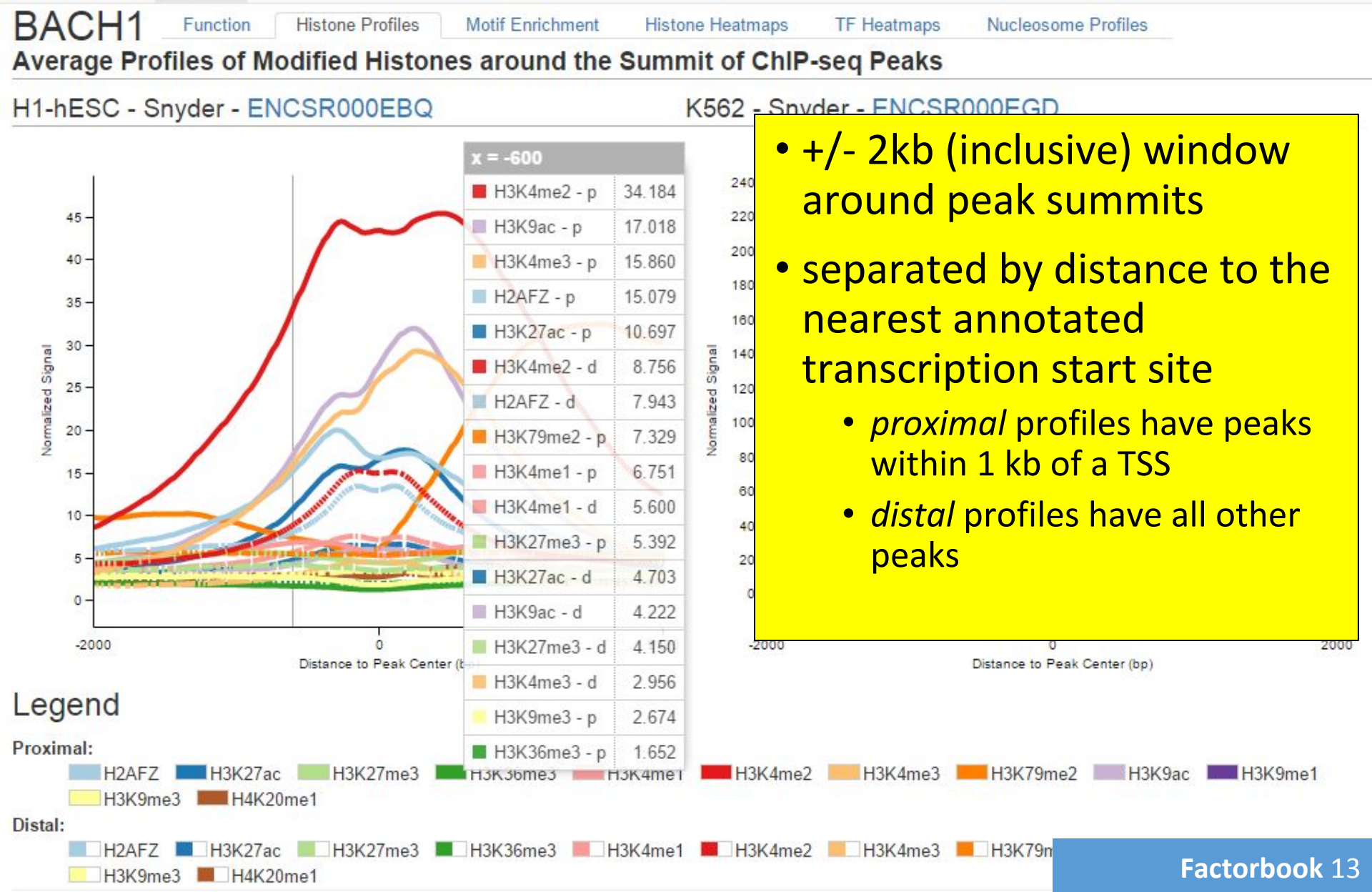


PDB 2IHC

PDB	2IHC
ENCODE	experiments
Ensemble	search
Entrez	571
GO	search
Gene Card	search
HGNC	search
RefSeq	search
UCSC	browse
UniProt	search
Wikipedia	BACH1

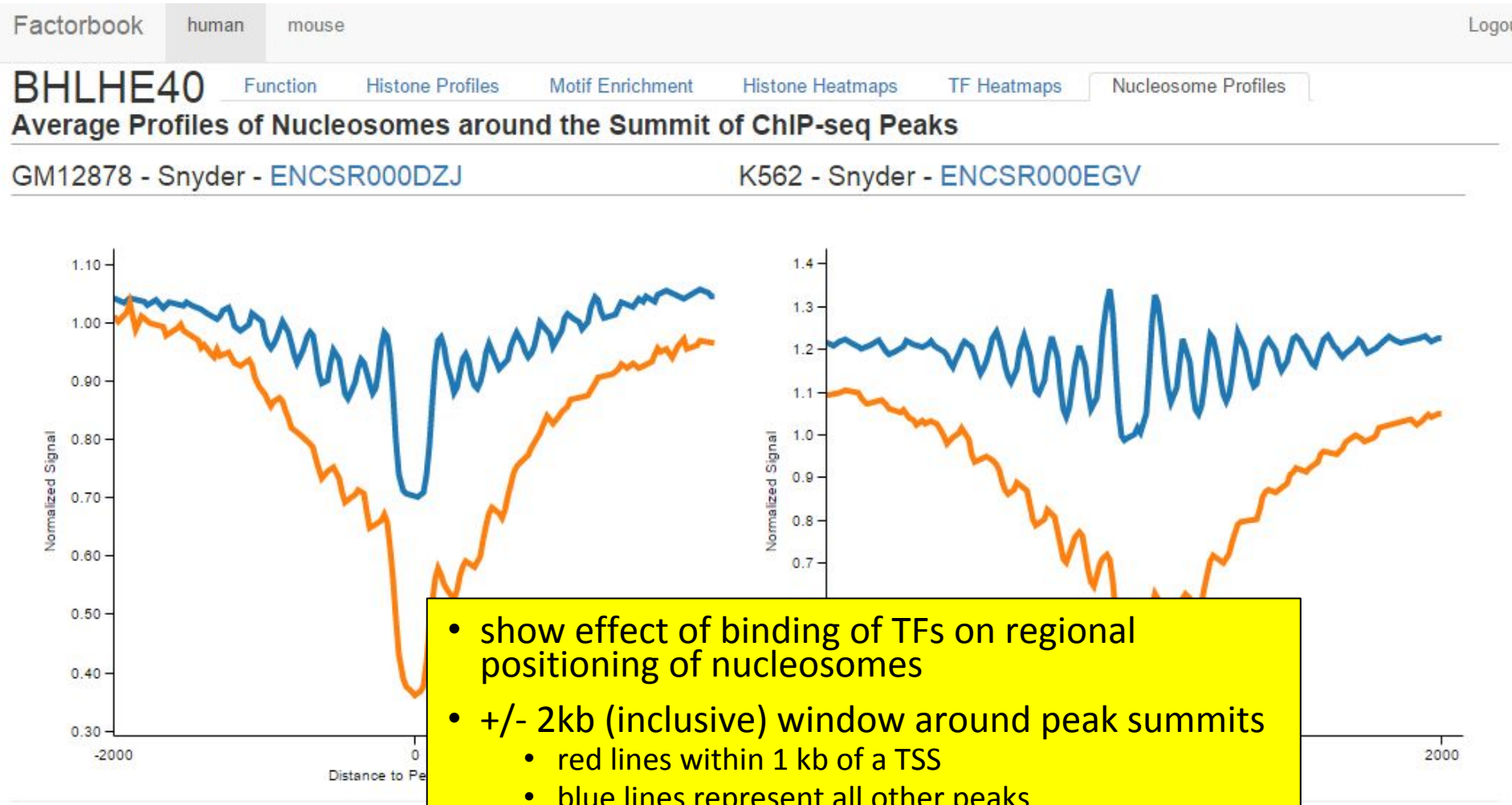
- brief overview of molecular function of TF
  - 3D protein structure of TF (if available)
  - distilled from RefSeq, Gene Card, and wikipedia
  - links to external resources

Average Histone Profiles



- +/- 2kb (inclusive) window around peak summits
- separated by distance to the nearest annotated transcription start site
  - *proximal* profiles have peaks within 1 kb of a TSS
  - *distal* profiles have all other peaks

Average Nucleosome Profiles



- show effect of binding of TFs on regional positioning of nucleosomes
- +/- 2kb (inclusive) window around peak summits
  - red lines within 1 kb of a TSS
  - blue lines represent all other peaks
- data from GM12878 and K562 MNase-seq

Motif Enrichment

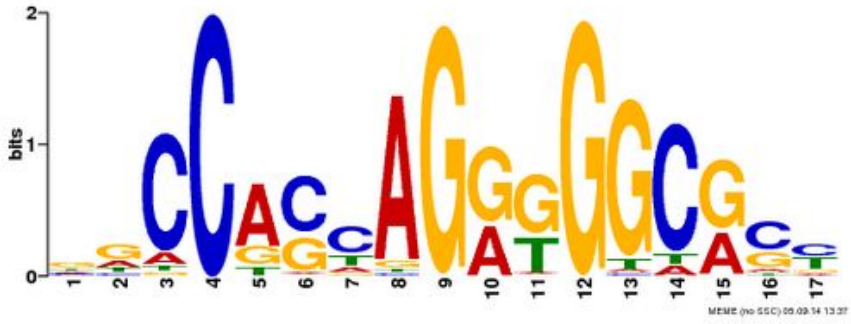
- sequences of the top 500 TF ChIP-seq peaks were used to identify enriched motifs de novo
  - MEME-ChIP
- top 5 motifs shown

M1

484 / 500

3.6e-965

GGCCACCAGGGGGCGCC



Motif Filtering

BACH1

Function | Histone Profiles | Motif Enrichment | Histone Heatmaps | TF Hea

Motifs Enriched in the Top 500 ChIP-seq Peaks

H1-hESC - Snyder | K562 - Snyder

H1-hESC - Snyder - ENCSR000EBQ

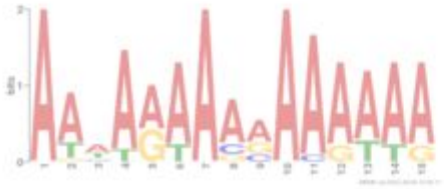
1.
424 / 500
6.2e-1117
AAAGTGCTGAGTCAT

2.
128 / 500
4.4e-106
GCTGAGTCAT

3.
52 / 500
1.6e-27
CCAGCAGGGGGCGCT

4.
67 / 500
3.3e-11
GCGGCCGGGCGGGGGCGAGGGGCGGGAGGG

5.
17 / 500
4.1e-8
AAAAAAAAAAAAAAA

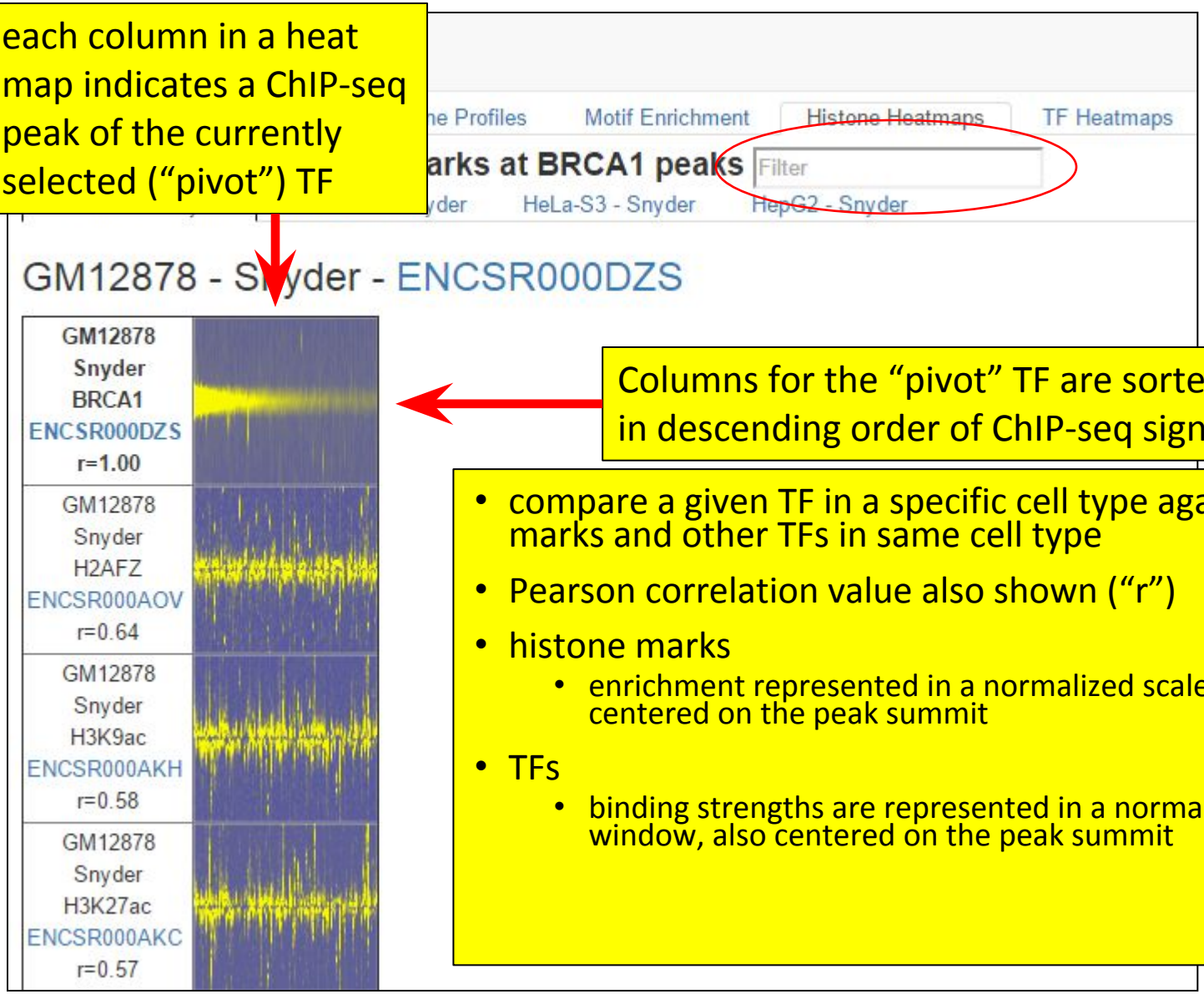


MEME output

Automatically filter out motifs that may not be biologically significant

Histone and TF Heat Maps

each column in a heat map indicates a ChIP-seq peak of the currently selected ("pivot") TF

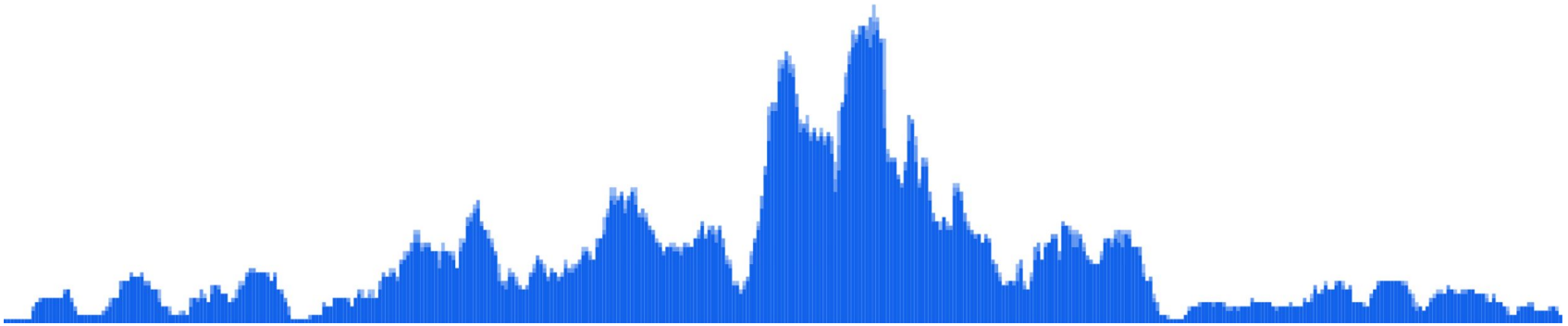


Columns for the "pivot" TF are sorted (left-to-right) in descending order of ChIP-seq signal

- compare a given TF in a specific cell type against the histone marks and other TFs in same cell type
- Pearson correlation value also shown ("r")
- histone marks
  - enrichment represented in a normalized scale over a 10kb window centered on the peak summit
- TFs
  - binding strengths are represented in a normalized scale over a 2kb window, also centered on the peak summit

Histone Mark Enrichment (ChIP-seq)

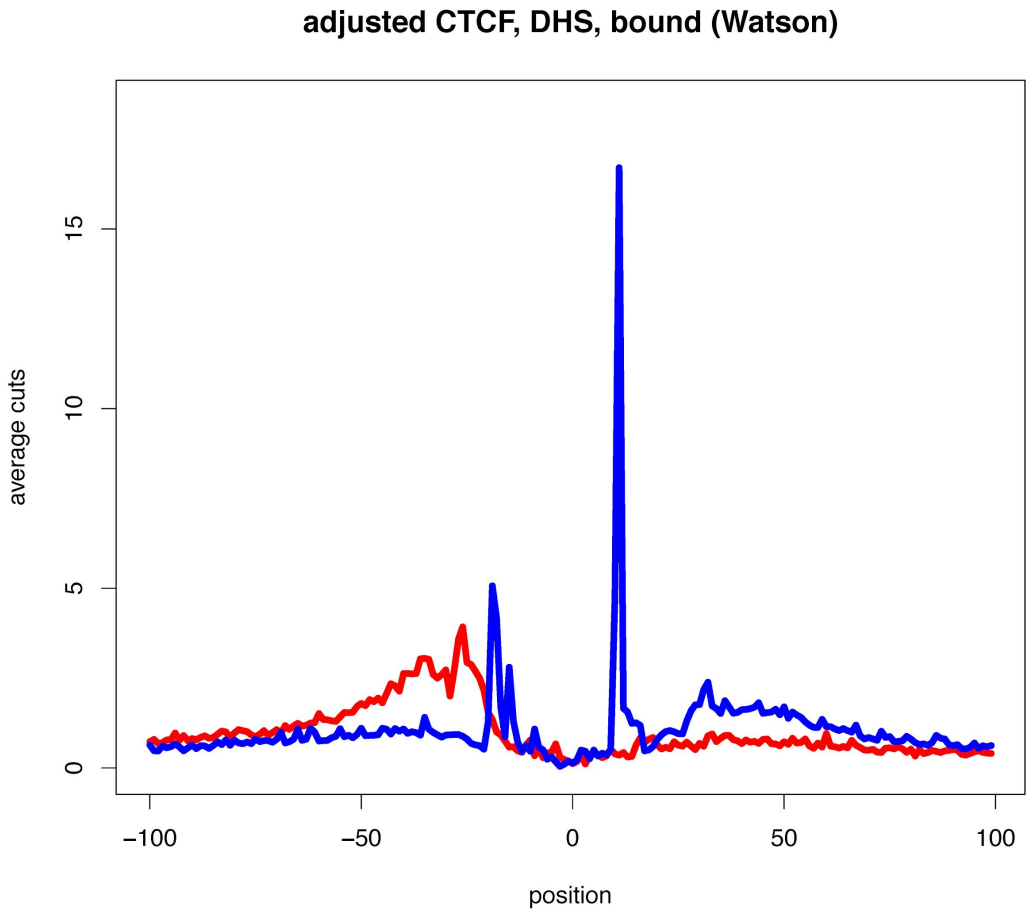
Peaks of a variety of histone marks computed from ~600 ChIP-seq experiments.



Data produced by: Ren, Bernstein, Stam, Farnham, Hardison, Snyder, Wold, Weissman

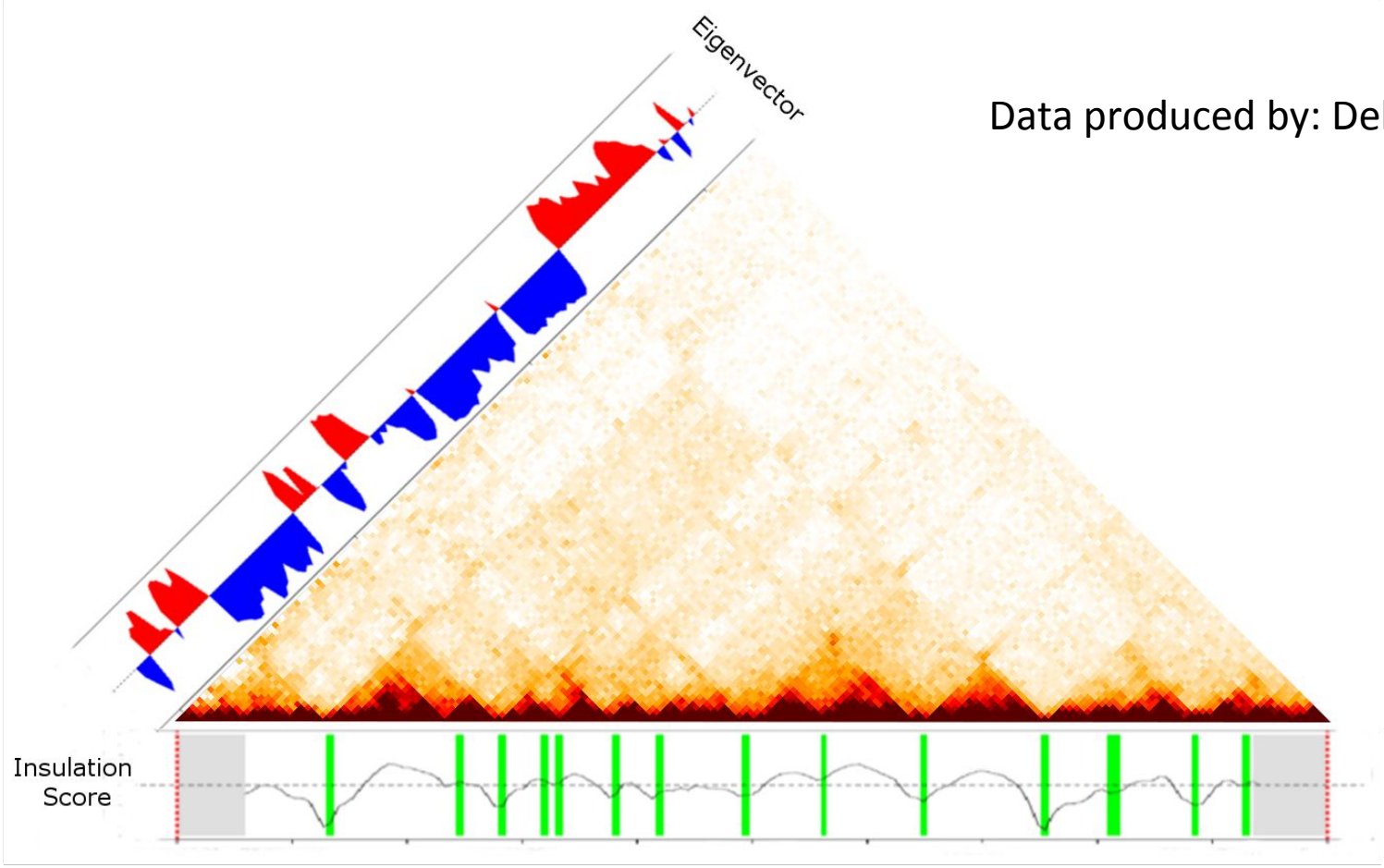
Open chromatin (DNase-seq)

DNase I hypersensitive sites (also known as DNase-seq peaks) computed from ~300 human and mouse experiments.



Data produced by: Stam, Crawford, Hardison

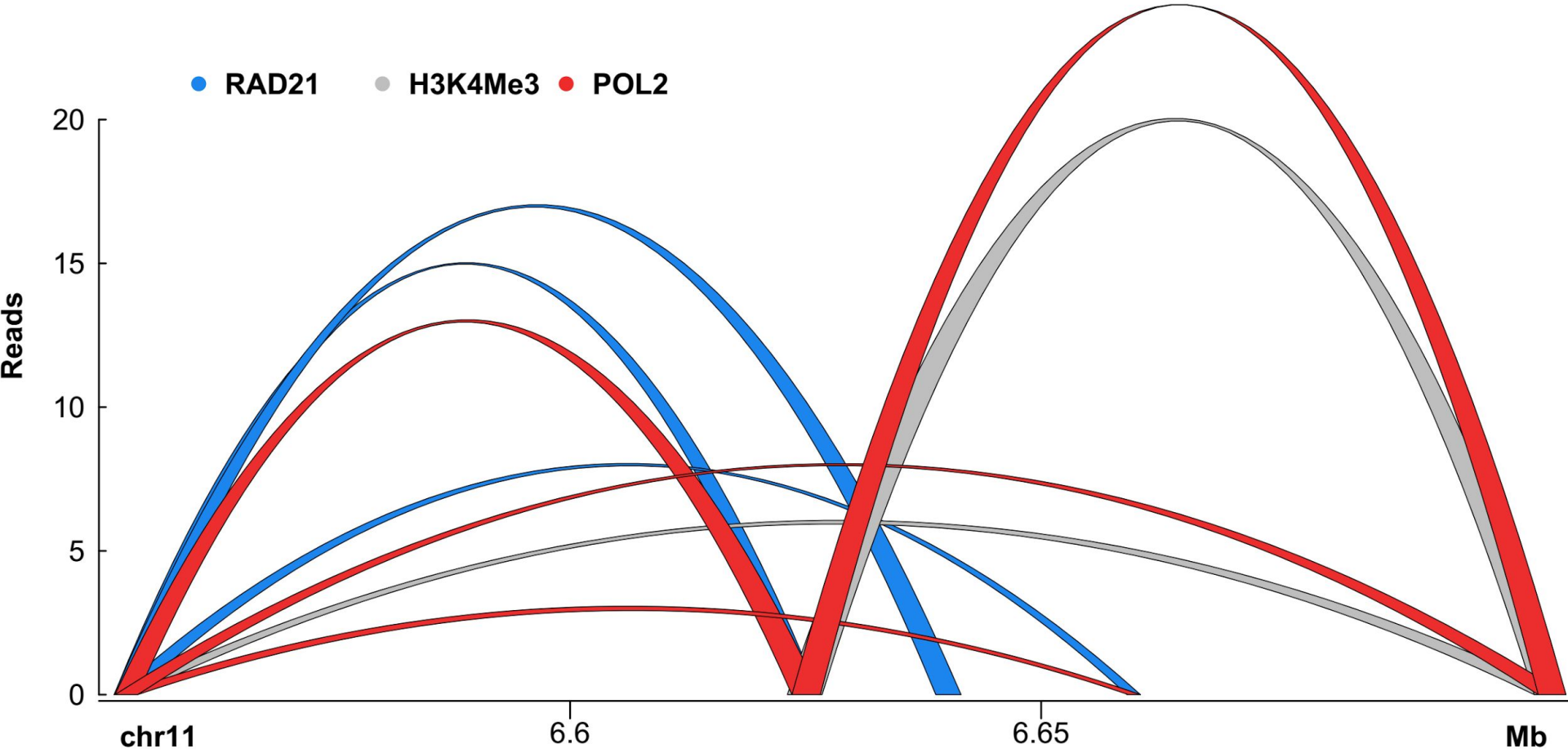
Topologically associating domains (TADs) and Compartments (Hi-C)



Data produced by: Dekker

Promoter-enhancer links (ChIA-PET)

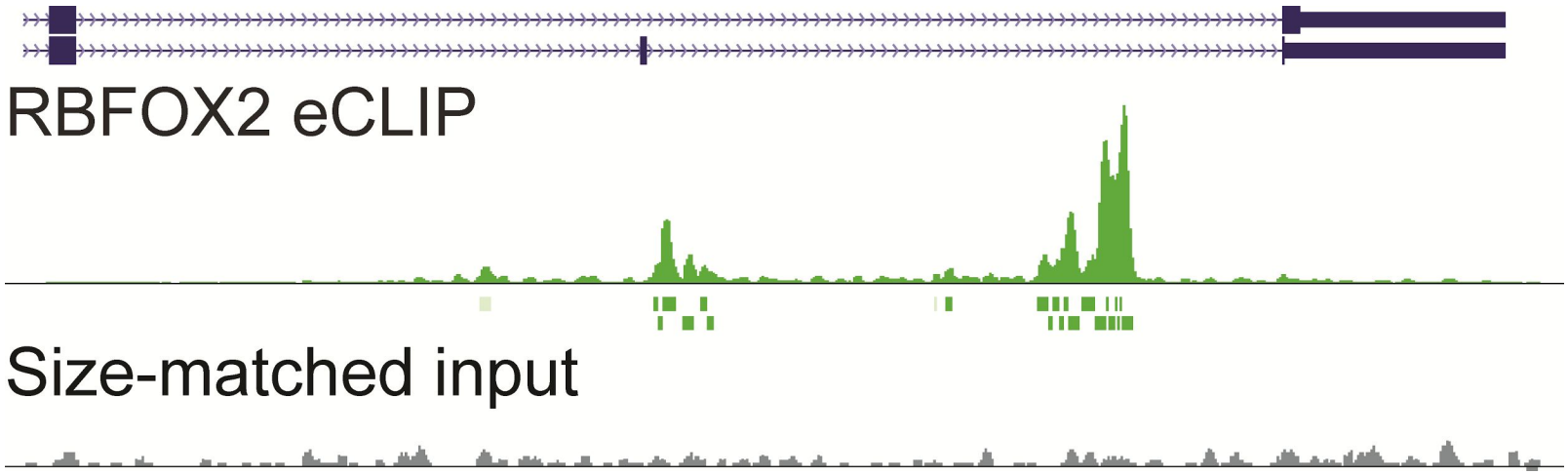
Links between promoters and distal regulatory elements such as enhancers computed from 8 ChIA-PET experiments.



Data produced by: Snyder, Ruan

RNA Binding Protein Occupancy (eCLIP-seq)

Peaks computed from eCLIP-seq data in human cell lines K562 and HepG2 for a large number of RNA Binding Proteins (RBPs).



Data produced by: Yeo

Middle Level Annotations

- Integrate multiple types of experimental data and ground level annotations

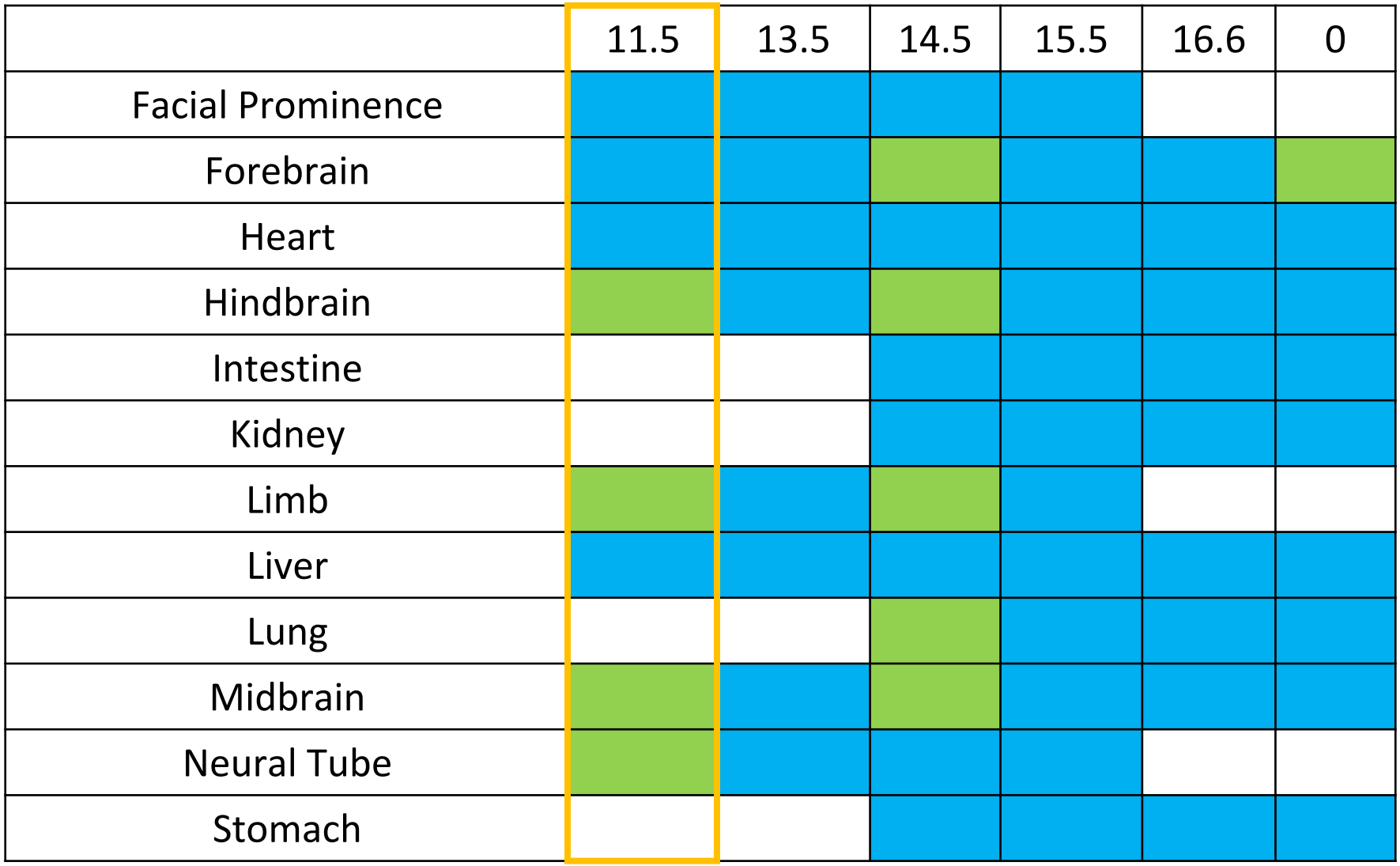
Goals for Predicting Enhancer-like Regions

- Develop an unsupervised method applicable to both human and mouse
- Incorporate different epigenomic datasets such as DNase-seq and H3K27ac
- Apply method to as many cell and tissue types as possible

Rationale for Developing Methods in Mouse

- Rich matrix of data of uniformly processed data:
  - Histone modification ChIP-seq (Bing Ren)
  - RNA-seq (Barbara Wold)
  - DNA methylation (Joe Ecker)
  - DNase-seq (John Stam)

ENCODE Data for Embryonic Mouse



H3K27ac Data - Bing Ren
DNase Data – John Stam

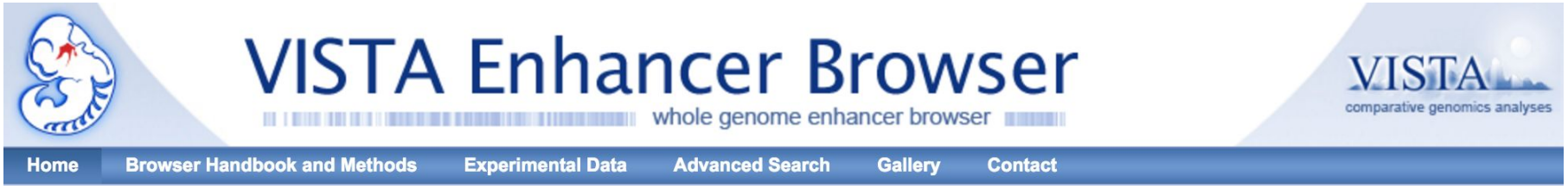
H3K27ac

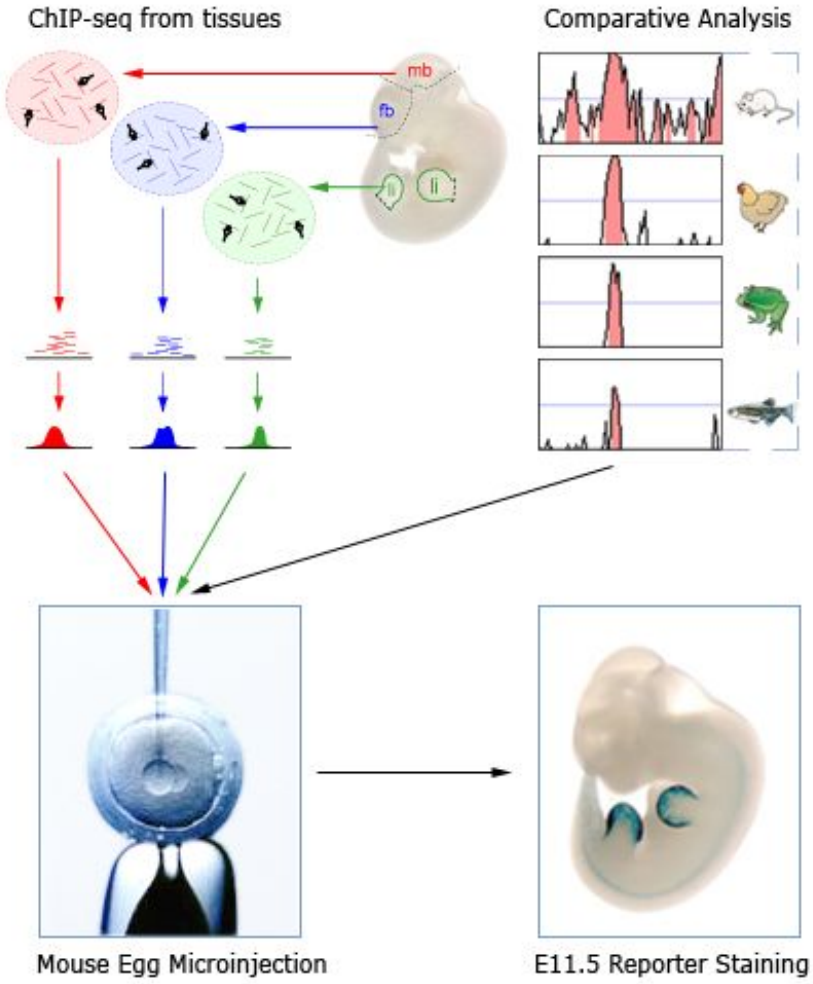
DNase + H3K27ac

Middle Level Annotations

Rationale for Developing Methods in Mouse

- Rich matrix of data of uniformly processed data:
  - Histone Modification ChIP-seq (Bing Ren)
  - RNA-seq (Barbara Wold)
  - DNA Methylation (Joe Ecker)
  - DNase-seq (John Stam)
- Experimental validations of enhancers in embryonic mice:
  - VISTA Database (Len Penacchio & Axel Visel)





- Over 2,000 total tested regions
- Over 200 active enhancers in limb, brain sub regions, and heart

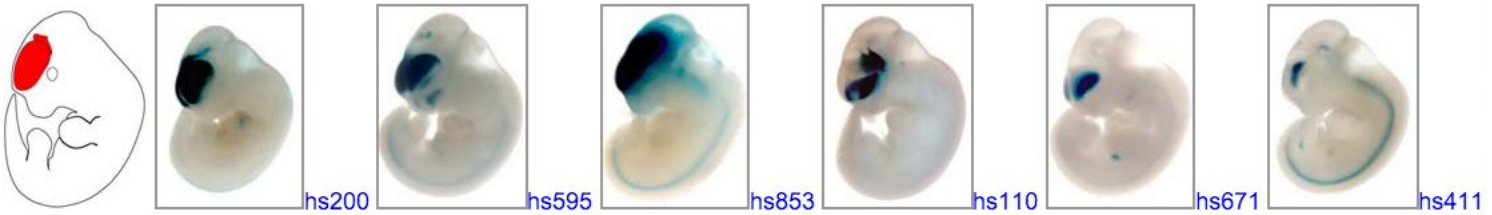
Pennacchio,…, Rubin (2006) *Nature*
Visel, …, Pennacchio (2009) *Nature*

Middle Level Annotations

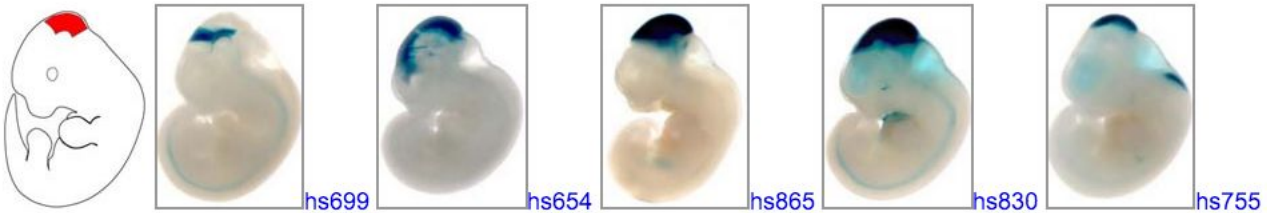
VISTA Database: Examples

Forebrain

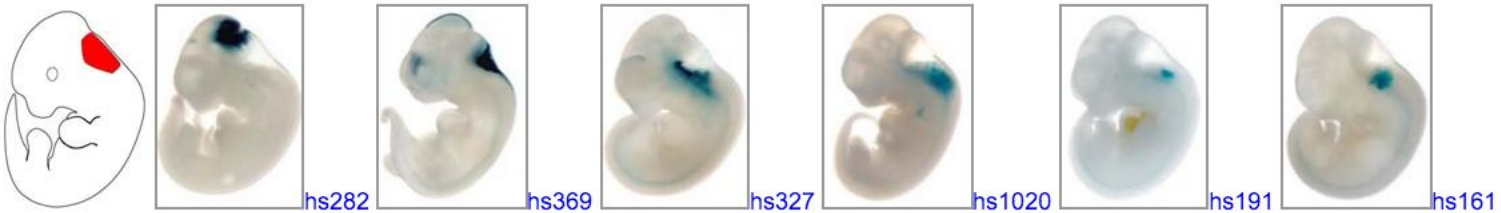
Forebrain includes all parts of the telencephalon and diencephalon.



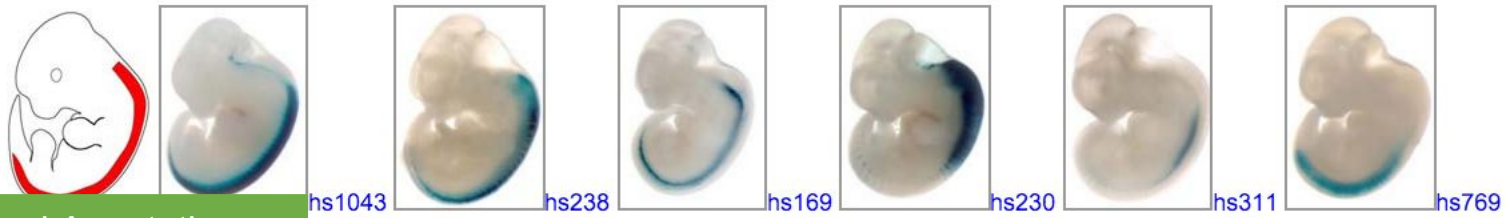
Midbrain



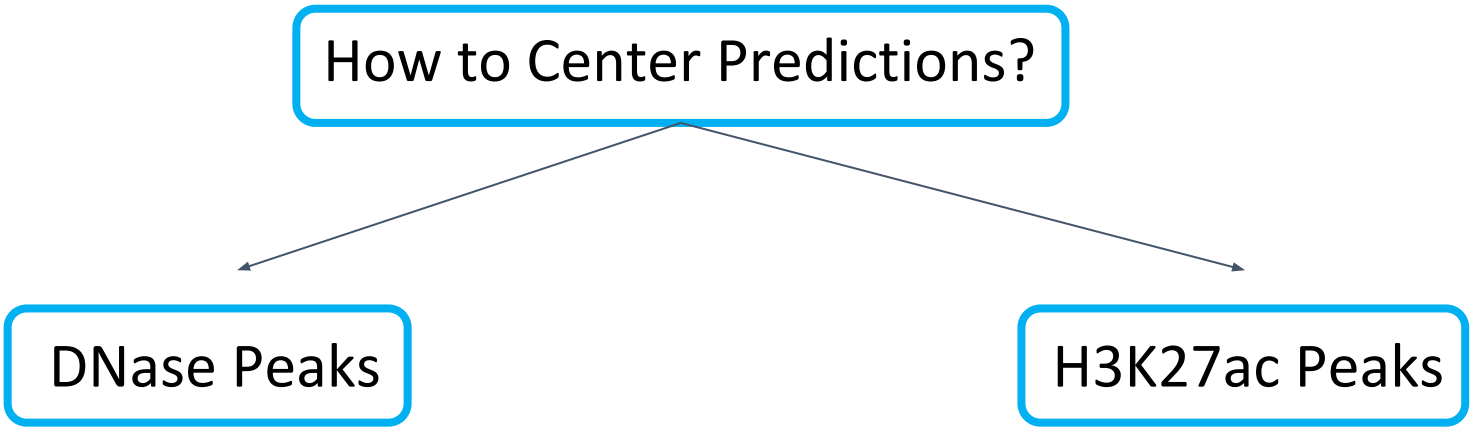
Hindbrain



Neural Tube



Middle Level Annotations



How to Rank Peaks?

p-value

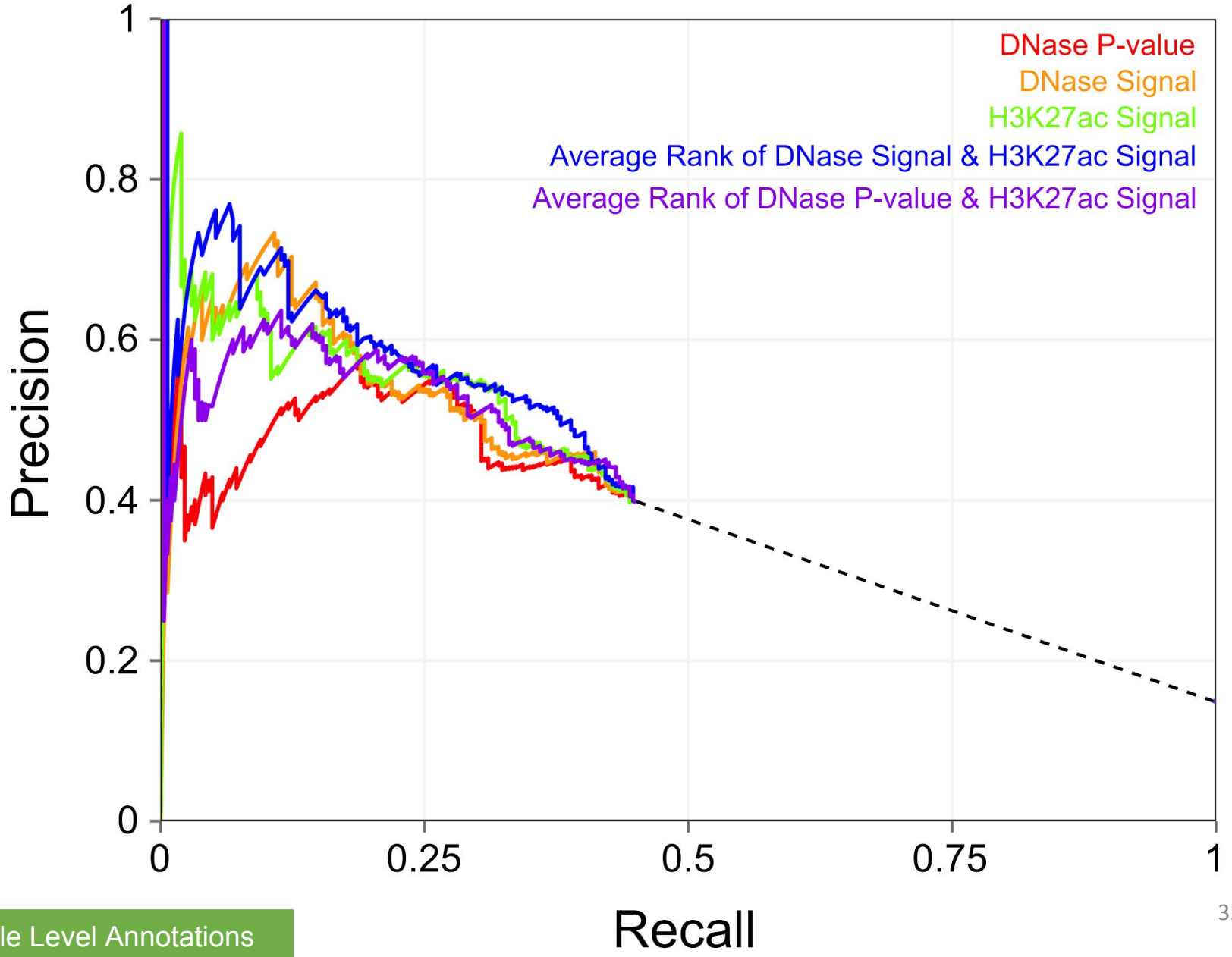
signal

multiple signals:
DNA Methylation
H3K4me1/2/3

Enhancer-like Region Prediction Method

	VISTA Positive	VISTA Negative
Overlaps Peak	True Positive	False Positive
Does Not Overlap Peak	False Negative	True Negative

Midbrain Predictions Centered on DNase Peaks



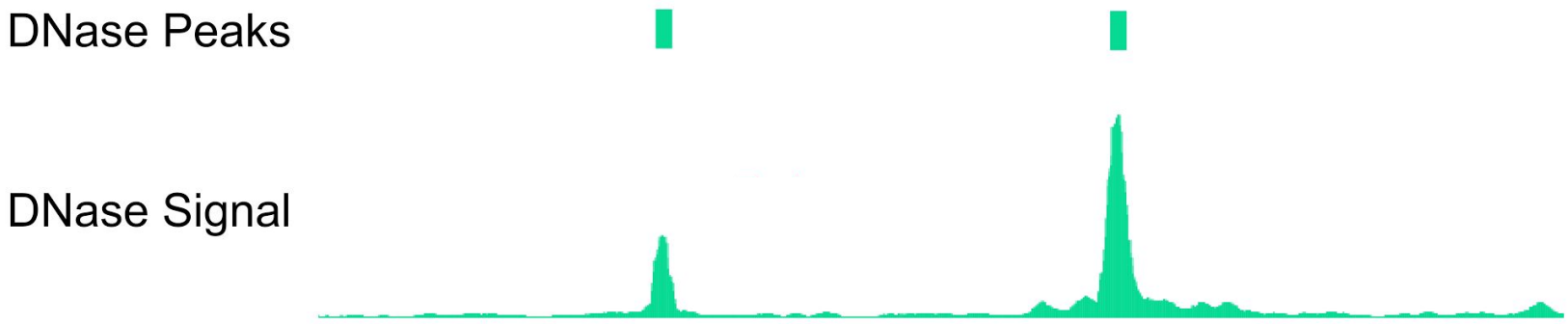
Results

- Centering predictions on DNase peaks results in better performance than centering on H3K27ac peaks

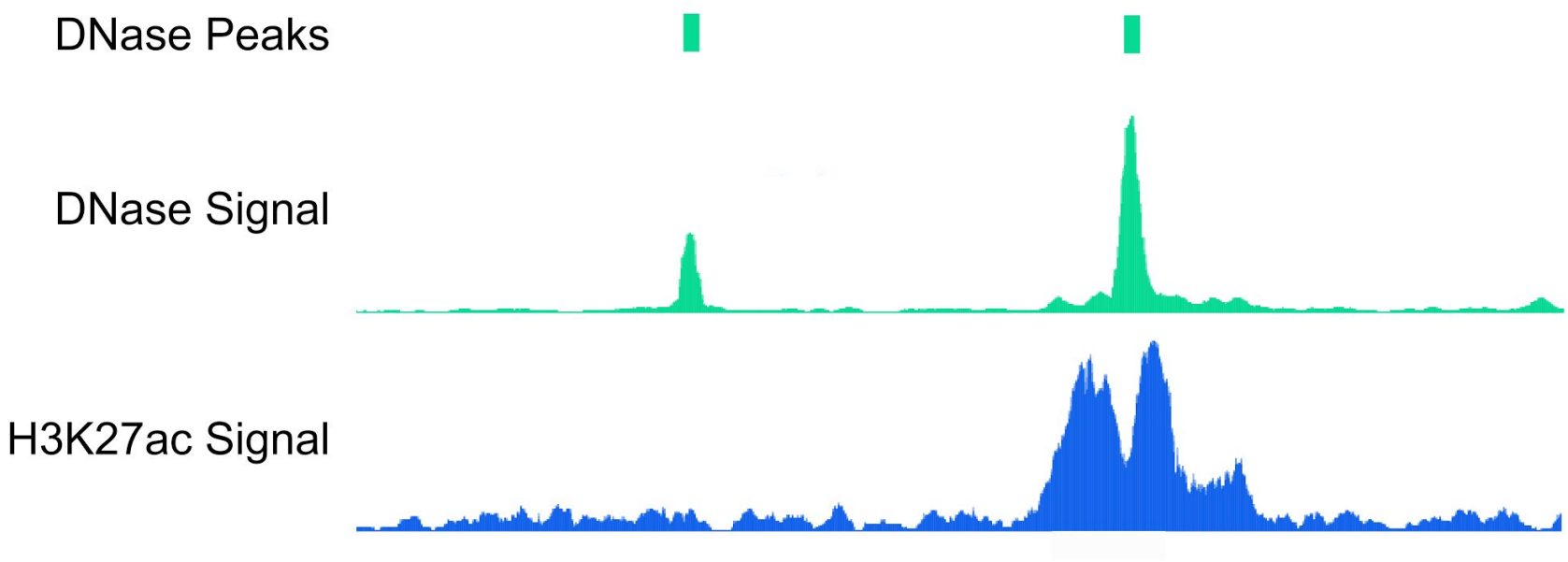
- Incorporating additional data such as DNA methylation and/or H3K4me1/2/3 signal did not improve performance

Enhancer-like Region Prediction Method

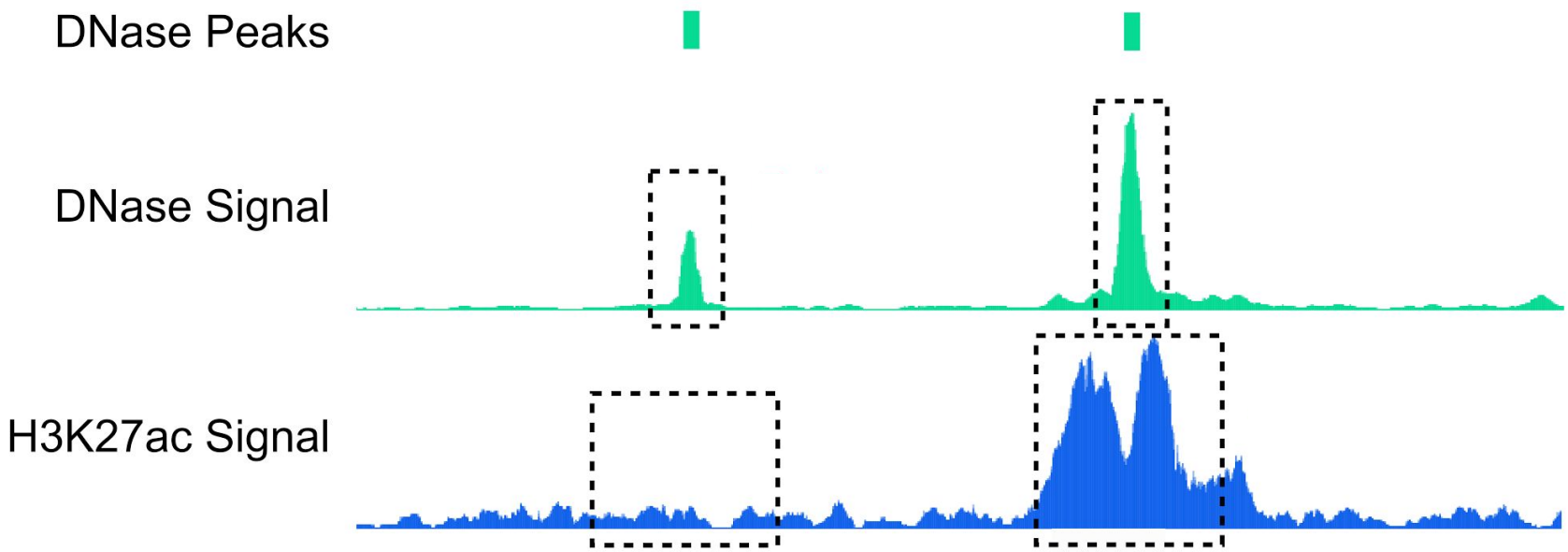
Enhancer-like Region Prediction Method



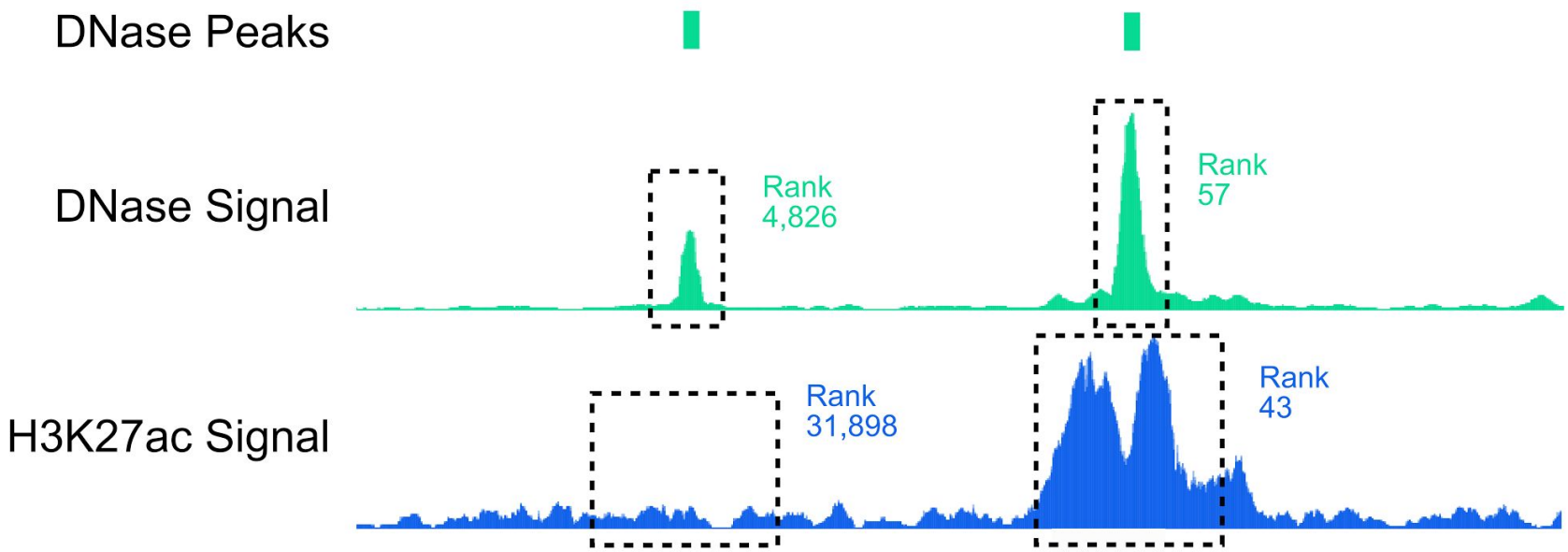
Enhancer-like Region Prediction Method



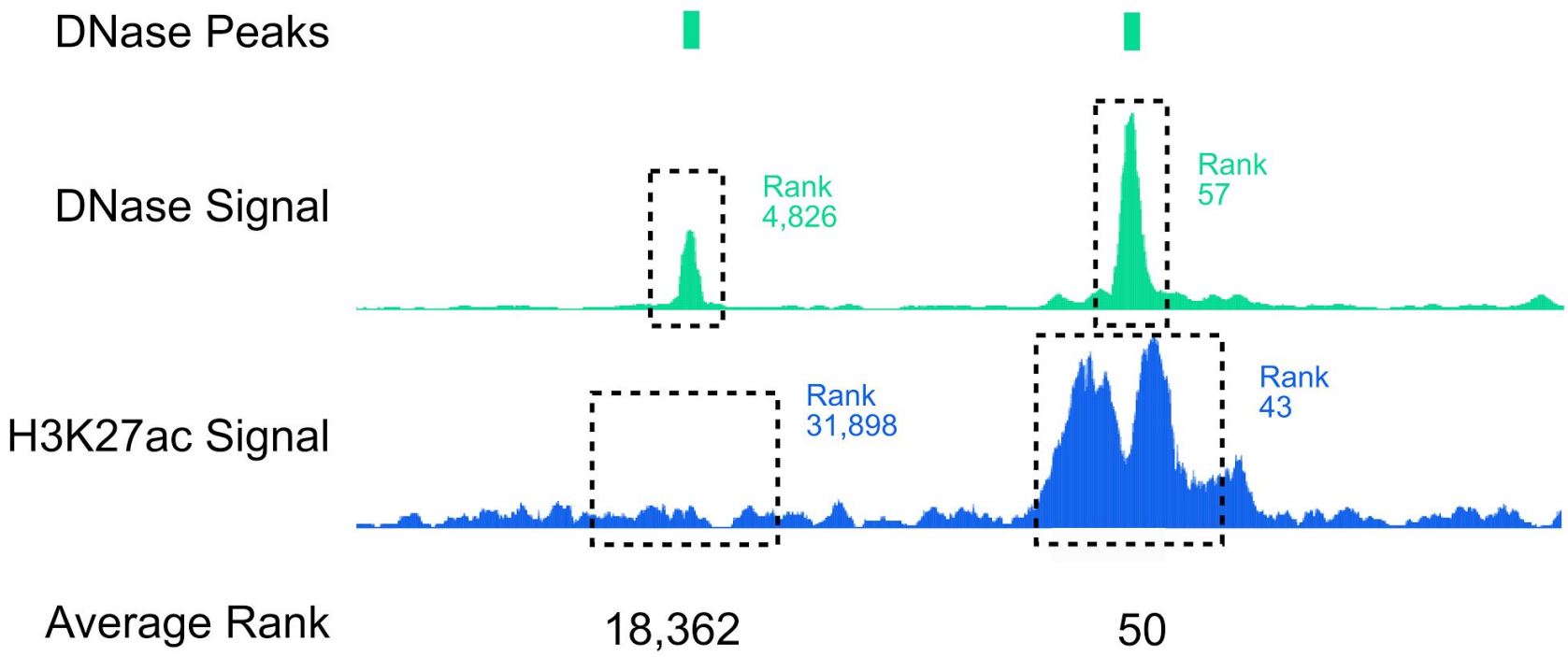
Enhancer-like Region Prediction Method



Enhancer-like Region Prediction Method

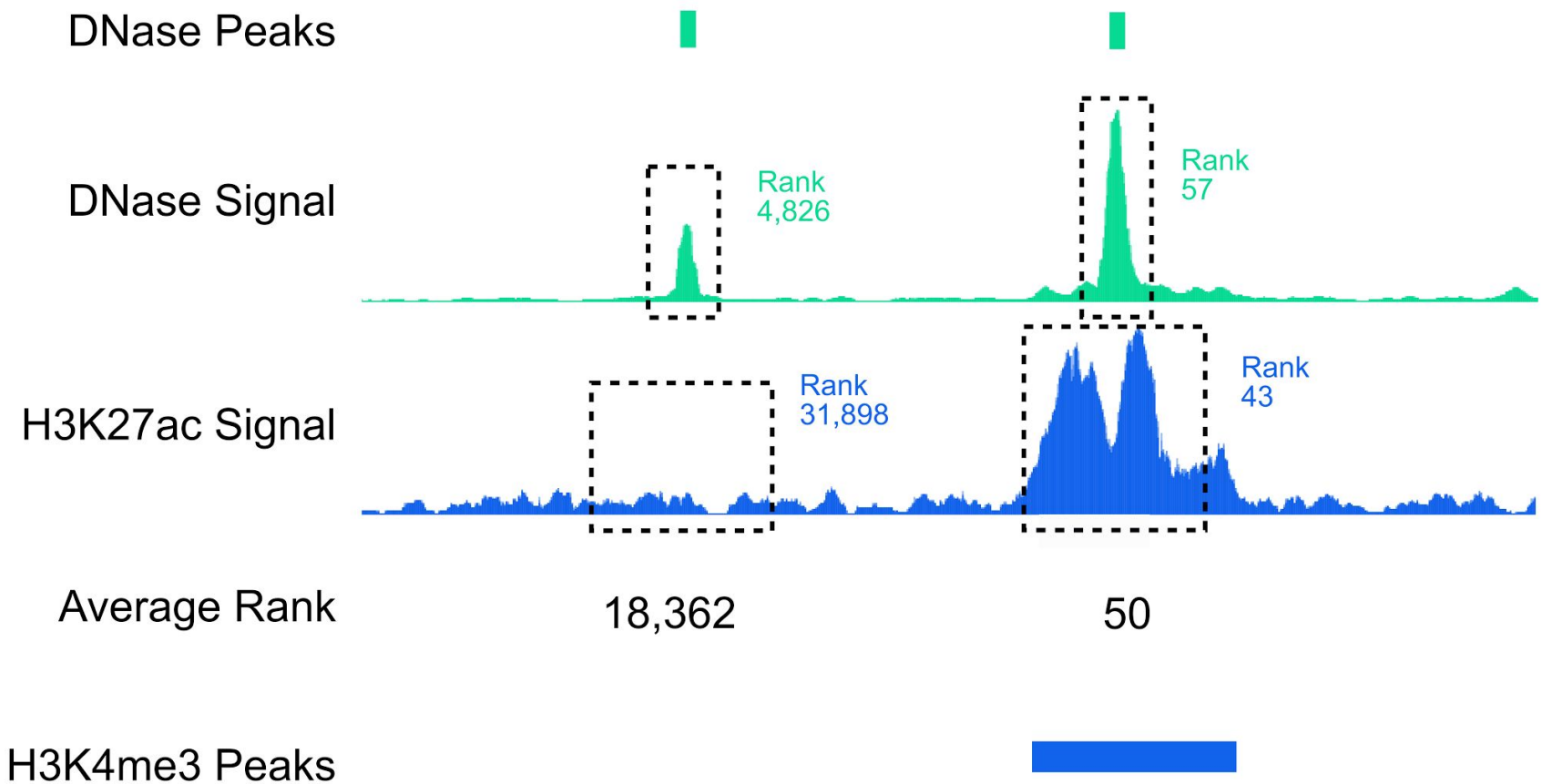


Enhancer-like Region Prediction Method



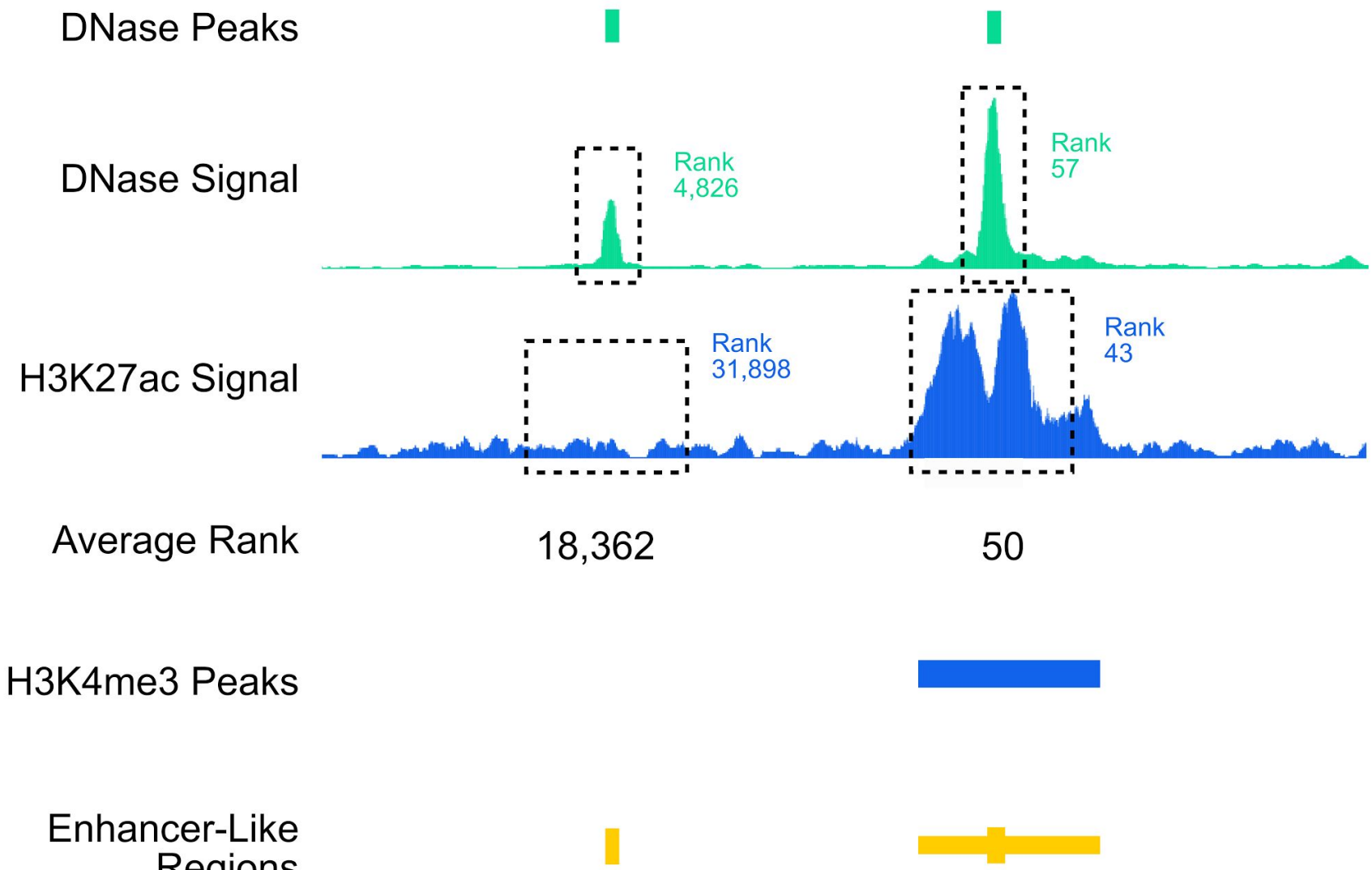
Middle Level Annotations

Enhancer-like Region Prediction Method



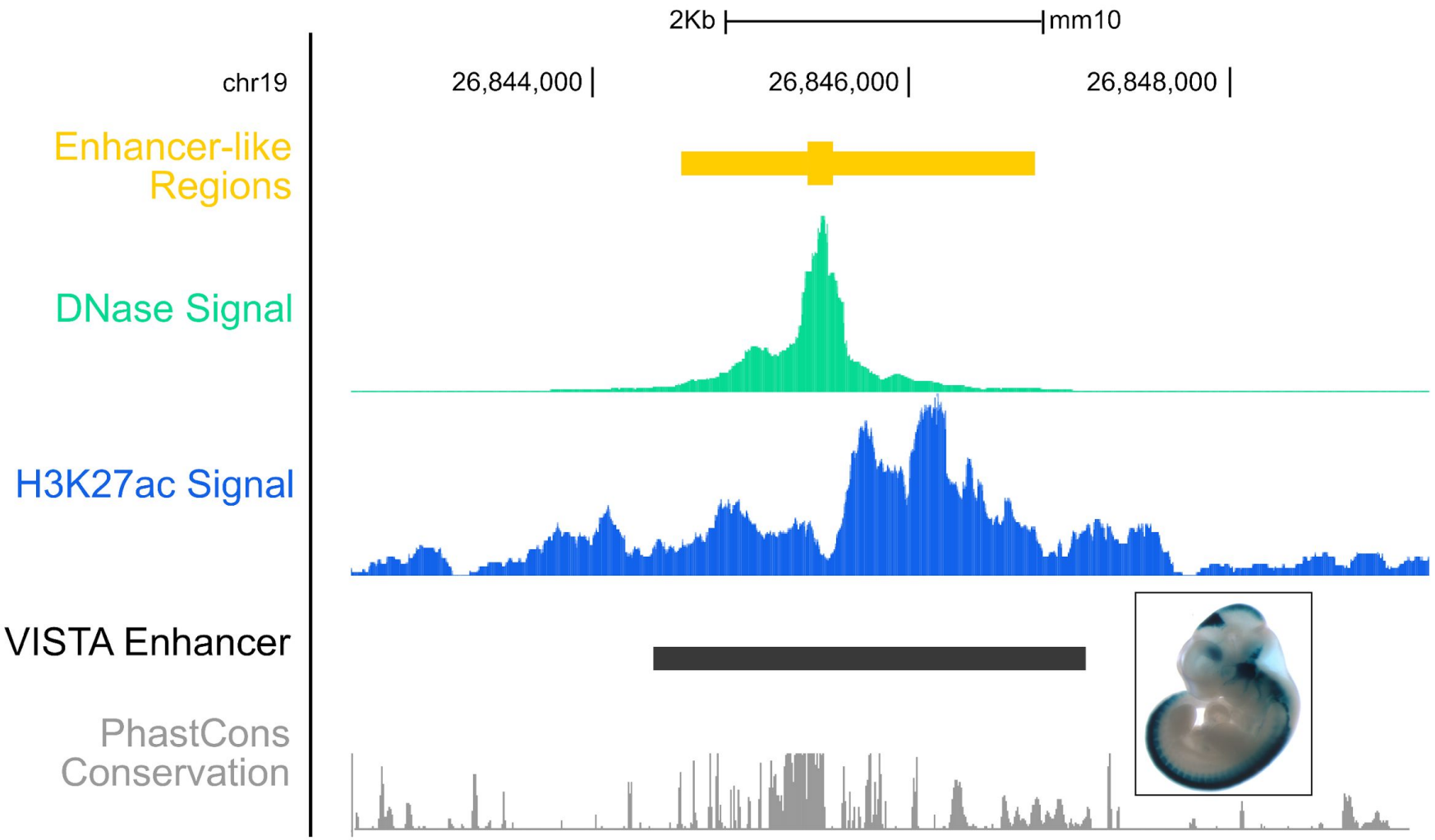
Middle Level Annotations

Enhancer-like Region Prediction Method



Middle Level Annotations

Example - Neural Tube (e11.5) Enhancer



Middle Level Annotations

Goals for Prediction Promoter-like Regions

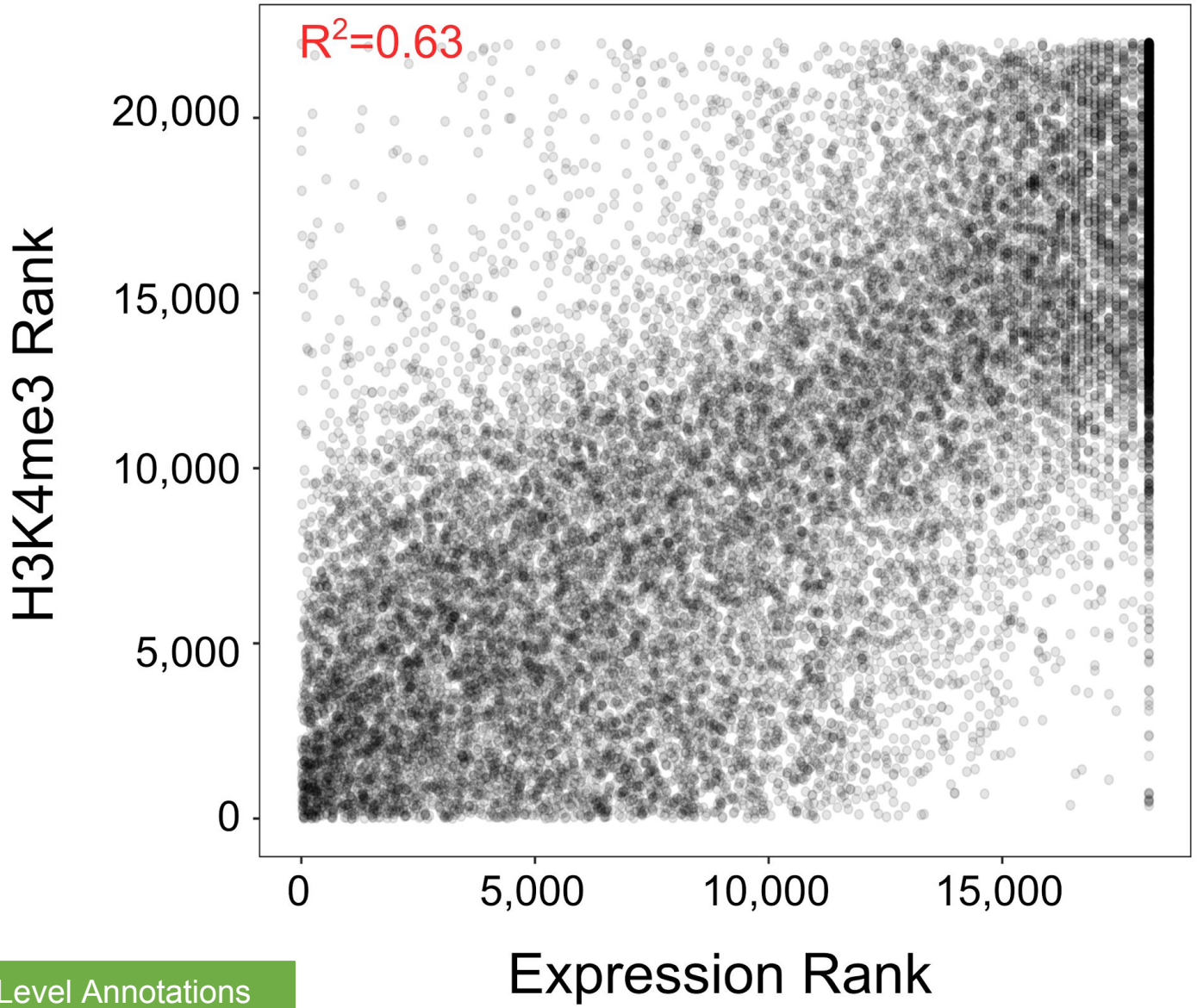
- Develop an unsupervised method applicable to both human and mouse
- Incorporate different epigenomic datasets such as DNase-seq, H3K4me3, and/or H3K27ac
- Apply method to as many cell and tissue types as possible

Promoter Prediction Method

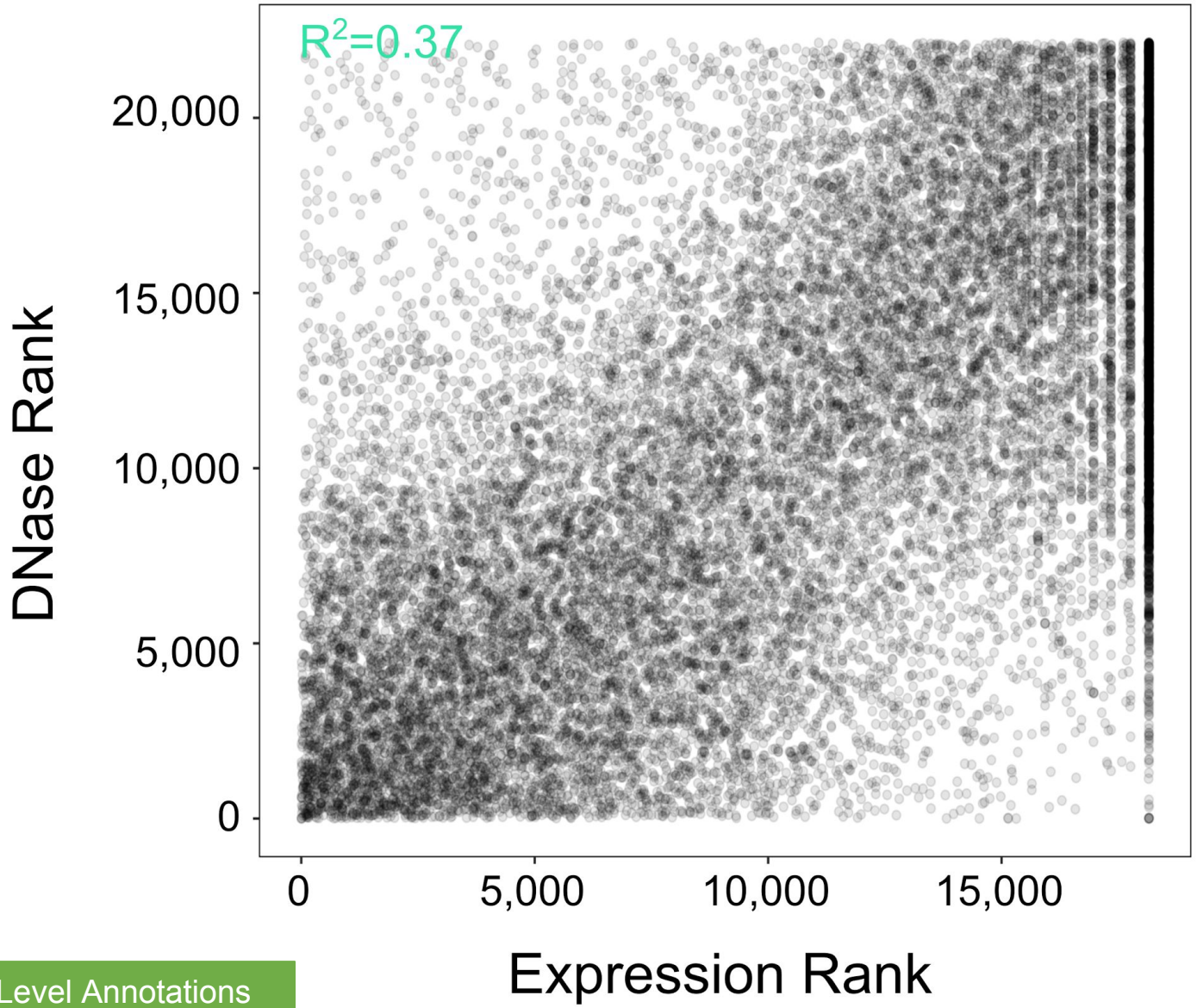
Using a linear model, which features of proximal DNase peaks are most predictive of ranked expression?

Gene	Expression (FPKM)	Ranked Expression		Rank by H3K4me3 Signal	Rank by DNase Signal	Rank by H3K27ac Signal
Gene A	3421	1		1	8	145
Gene B	2329	2		7	345	985
Gene C	432	3		4	2	217
…	…	…		…	…	…

H3K4me3 Signal Only

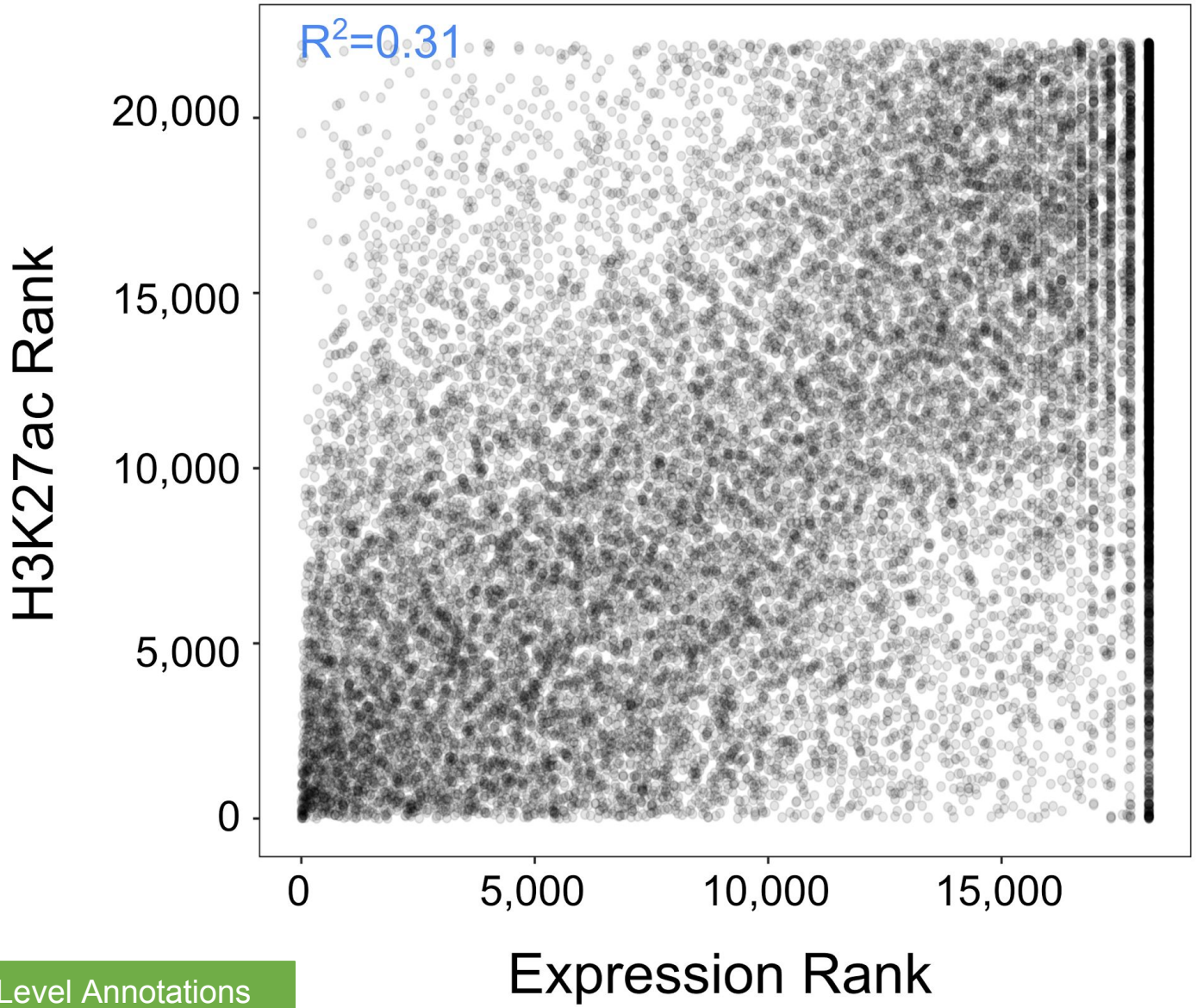


DNase Signal Only



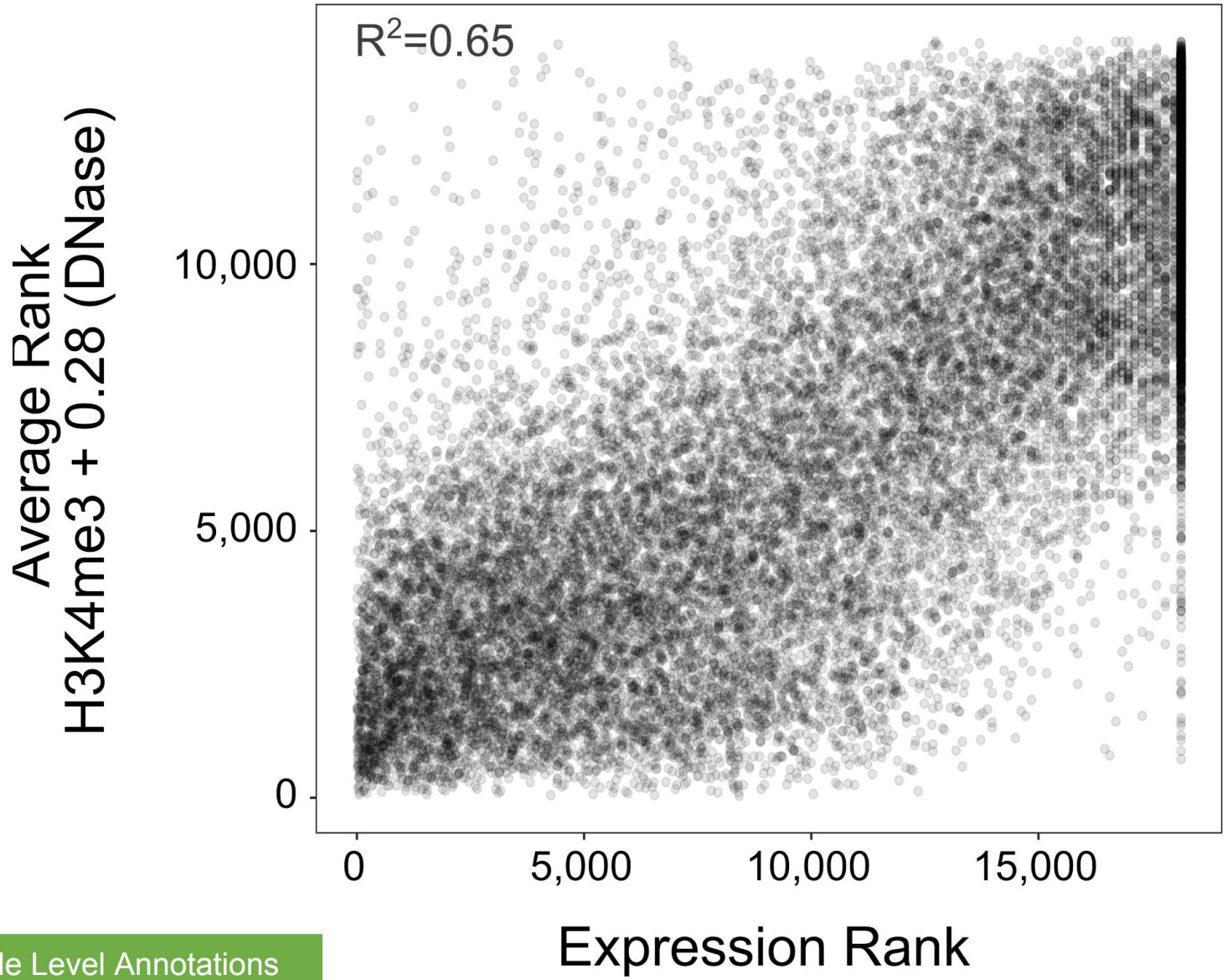
Middle Level Annotations

H3K27ac Signal Only



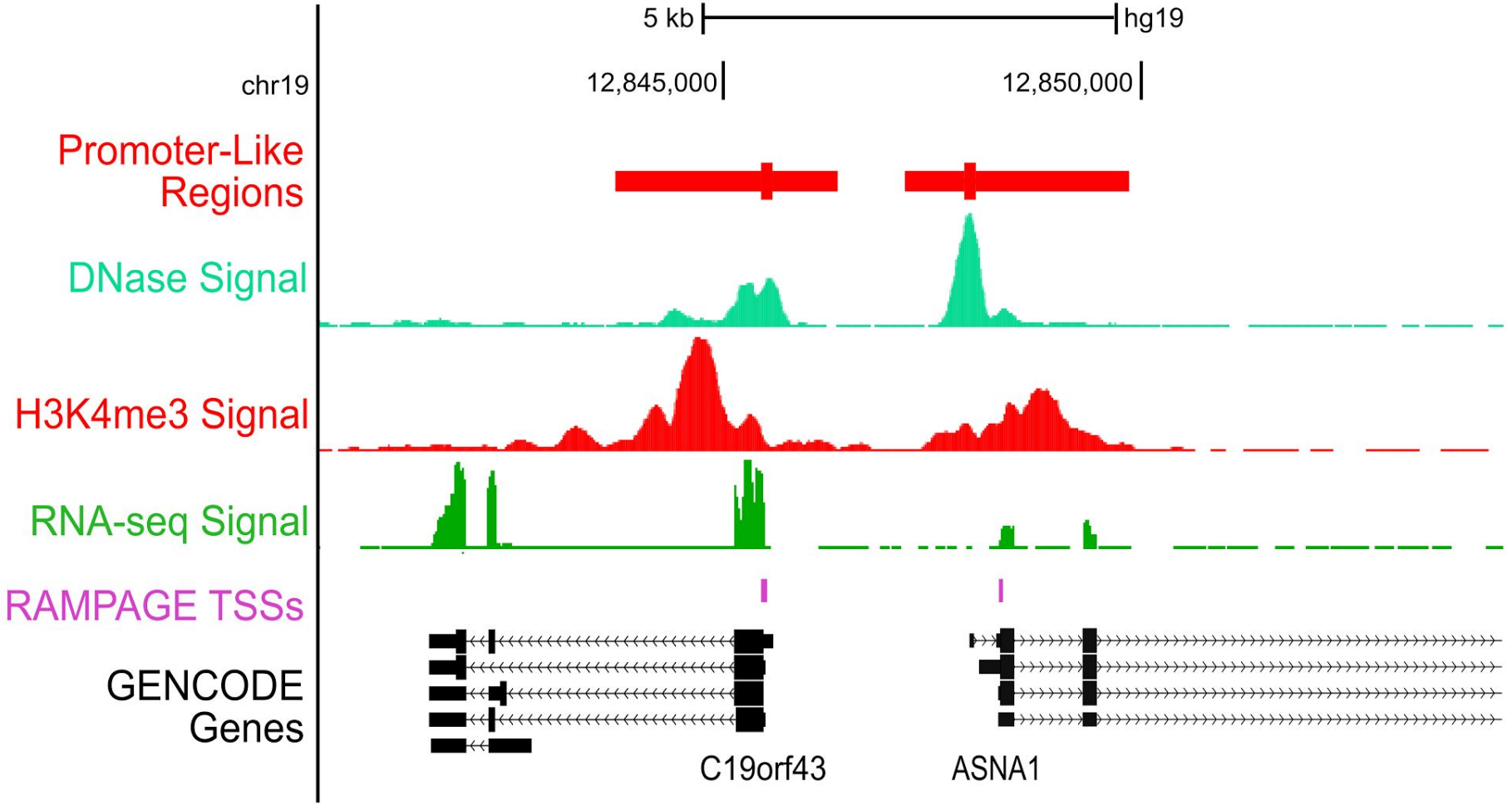
Middle Level Annotations

Best Method: H3K4me3 Signal + 0.28 * DNase Signal



Middle Level Annotations

Example - Promoters in GM12878



Visualization of Enhancer-like and Promoter-like Regions

Demo

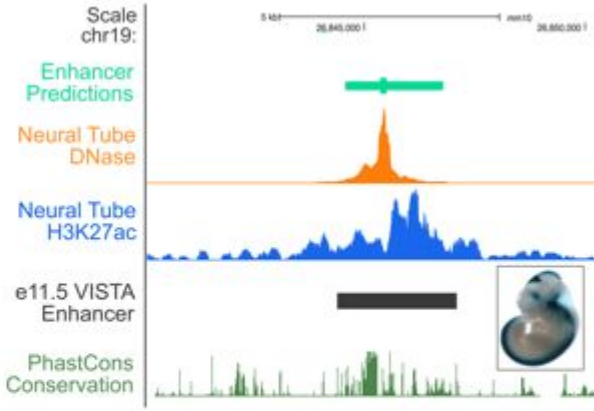
zlab-annotations.umassmed.edu

- Proof of concept for enhancer-like and promoter-like visualization
- Seeking feedback from community
- Provide a sample site for DCC to implement

Annotated genomic regions

Candidate enhancers based on DNase and H3K27ac signals

DNase hypersensitivity and histone modification H3K27ac are well-known indicators of enhancer function. We have developed an unsupervised method that combines DNase and H3K27ac signals in the same cell type to predict enhancers. When tested on mouse transgenic assays, our method shows higher accuracy than DNase and H3K27ac individually. We have applied this method to 45 human cell types and 20 mouse cell types with both DNase and H3K27ac data generated by the ENCODE and Roadmap Epigenomic consortia. You can query these enhancers by genomic locations, nearby genes, or SNPs.



Search Candidate Enhancers

Human (hg19) | Mouse (mm10) | Mouse (mm9)

1) Choose genome

chr1:134054000-134071000 Search

2) Enter gene, SNP, genomic coordinate

Examples: gene, SNP (ex. rs27106747), genomic region, or peak rank (for a single tissue)

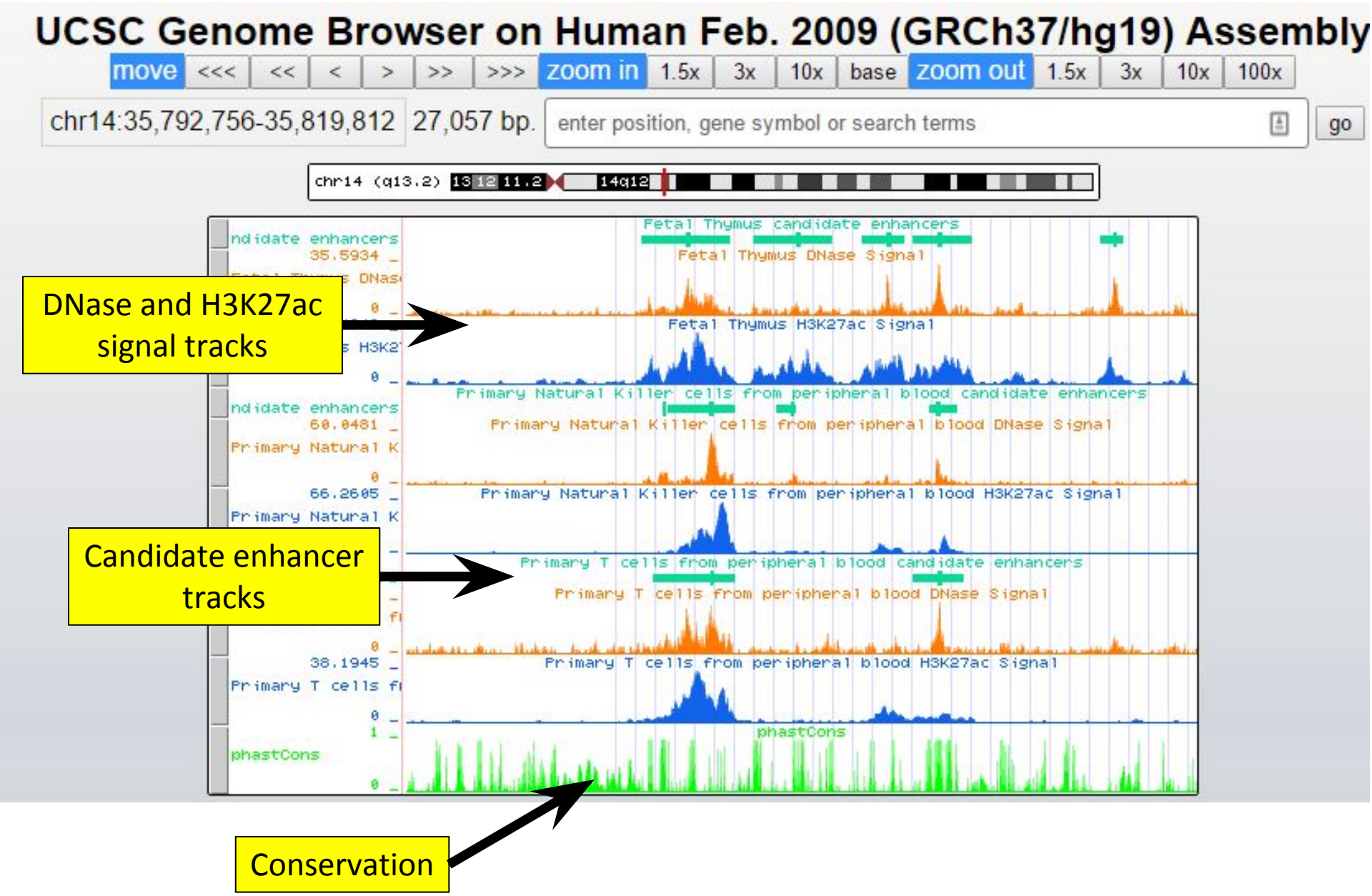
DNase + H3K27ac | DNase | H3K27ac | All | None | Intersect

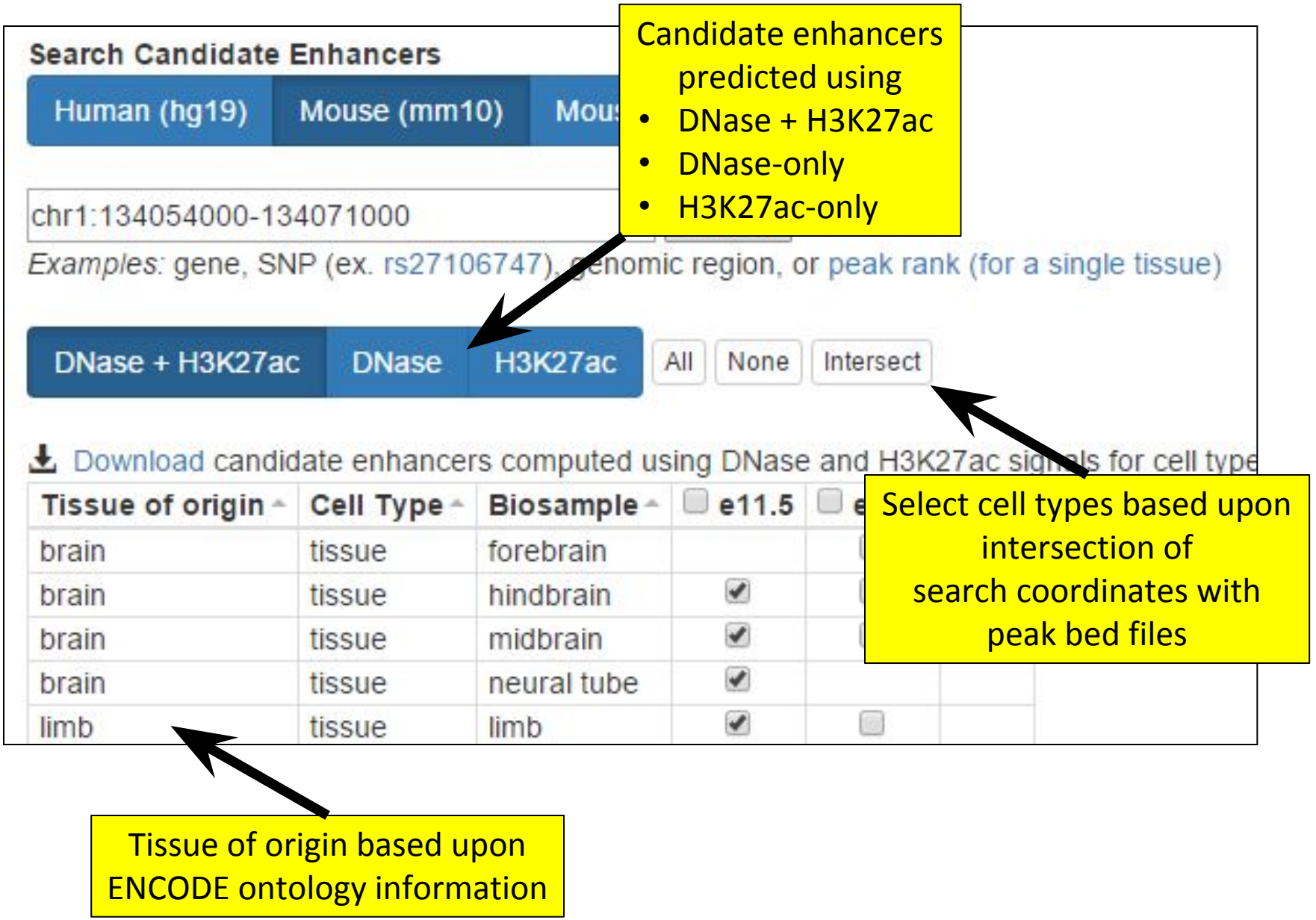
3) Choose cell types

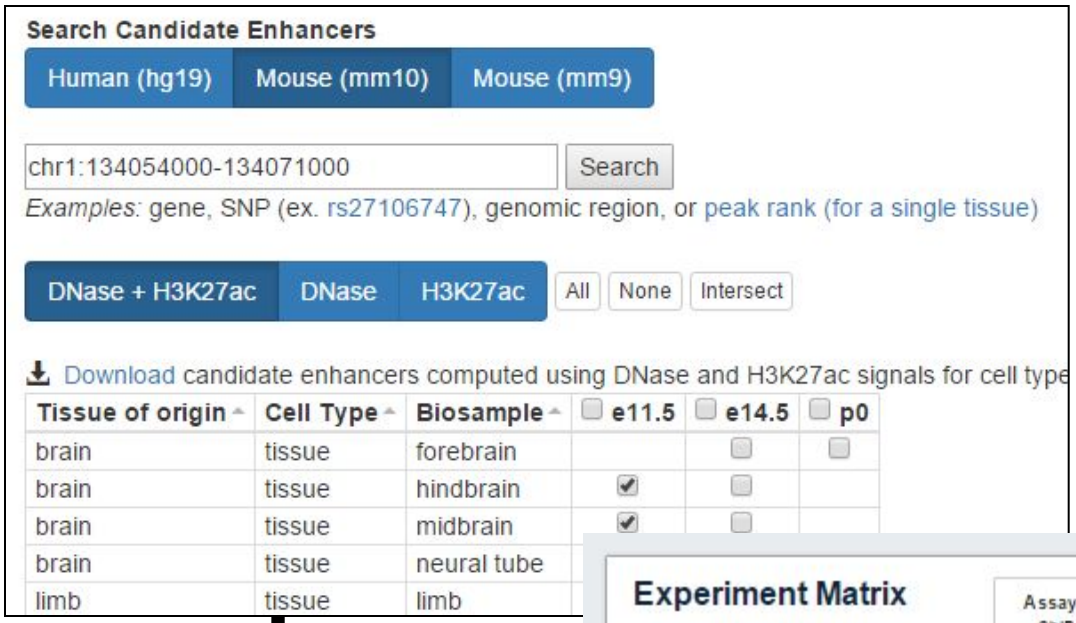
Download candidate enhancers computed using DNase and H3K27ac signals for cell types below (genome wide)

Tissue of origin	Cell Type	Biosample	e11.5	e14.5	p0
brain	tissue	forebrain		☐	☐
brain	tissue	hindbrain	☑	☐	
brain	tissue	midbrain	☑	☐	
brain	tissue	neural tube	☑		
limb	tissue	limb	☑	☐	

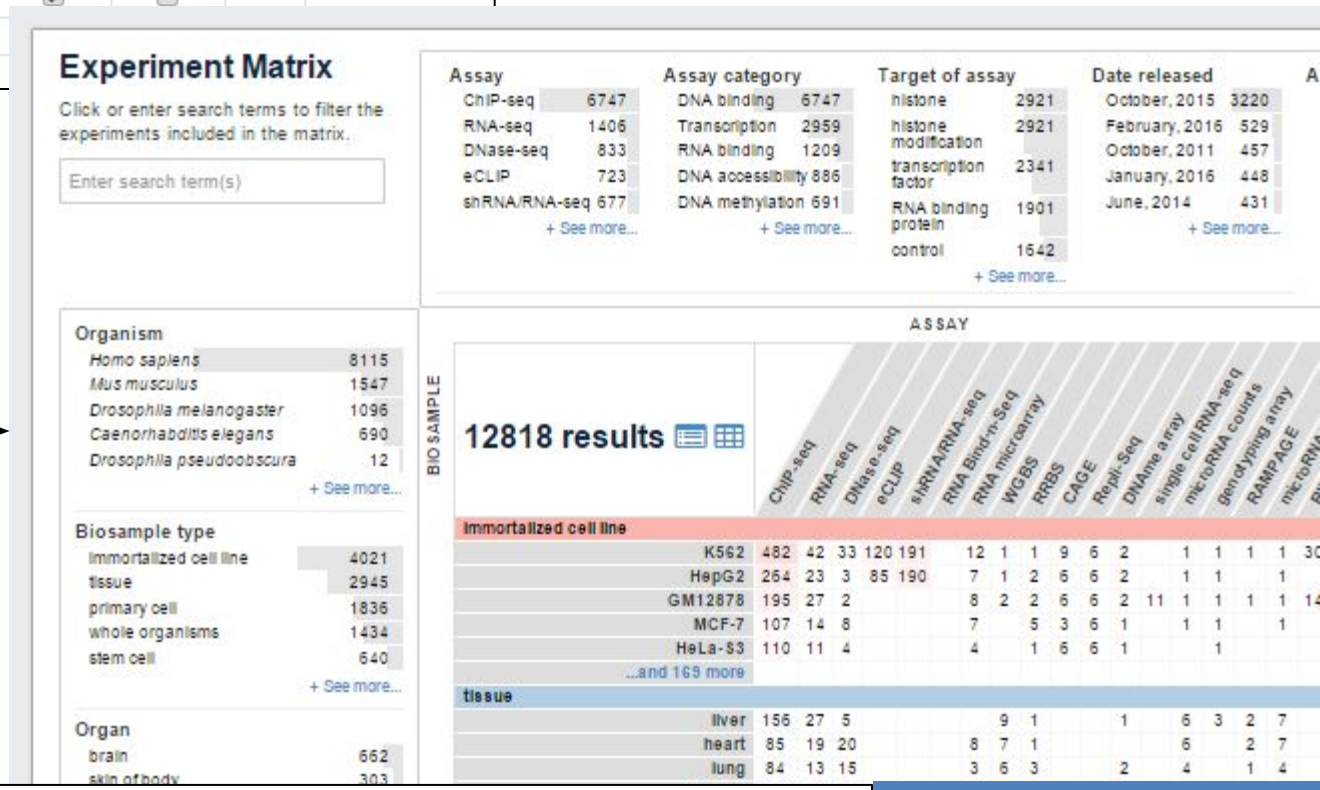
4) View tracks in UCSC Or WashU Genome Browser →


UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x
chr14:35,792,756-35,819,812 27,057 bp.
enter position, gene symbol or search terms
go
chr14 (q13.2) 13 12 11.2 14q12
Fetal Thymus candidate enhancers
Fetal Thymus DNase Signal
Fetal Thymus H3K27ac Signal
Primary Natural Killer cells from peripheral blood candidate enhancers
Primary Natural Killer cells from peripheral blood DNase Signal
Primary Natural Killer cells from peripheral blood H3K27ac Signal
Primary T cells from peripheral blood candidate enhancers
Primary T cells from peripheral blood DNase Signal
Primary T cells from peripheral blood H3K27ac Signal
phastCons
DNase and H3K27ac signal tracks
Candidate enhancer tracks
Conservation


Search Candidate Enhancers
Human (hg19)
Mouse (mm10)
chr1:134054000-134071000
Examples: gene, SNP (ex. rs27106747), genomic region, or peak rank (for a single tissue)
DNase + H3K27ac
DNase
H3K27ac
All
None
Intersect
Download candidate enhancers computed using DNase and H3K27ac signals for cell type
Tissue of origin
Cell Type
Biosample
e11.5
brain tissue forebrain
brain tissue hindbrain
brain tissue midbrain
brain tissue neural tube
limb tissue limb
Candidate enhancers predicted using
• DNase + H3K27ac
• DNase-only
• H3K27ac-only
Select cell types based upon intersection of search coordinates with peak bed files
Tissue of origin based upon ENCODE ontology information



ENCODE DCC Matrix view

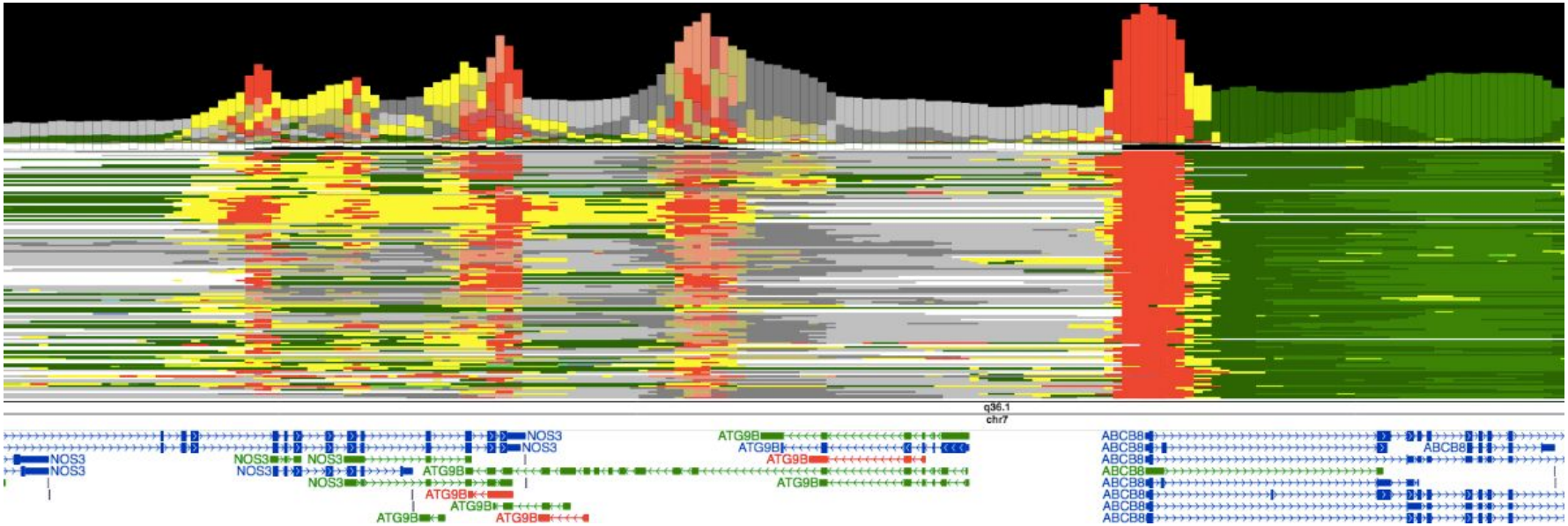


www.encodeproject.org/matrix/?type=Experiment

Top Level Annotations

- Integrate a broad range of experimental data, as well as ground and middle level annotations

Chromatin states

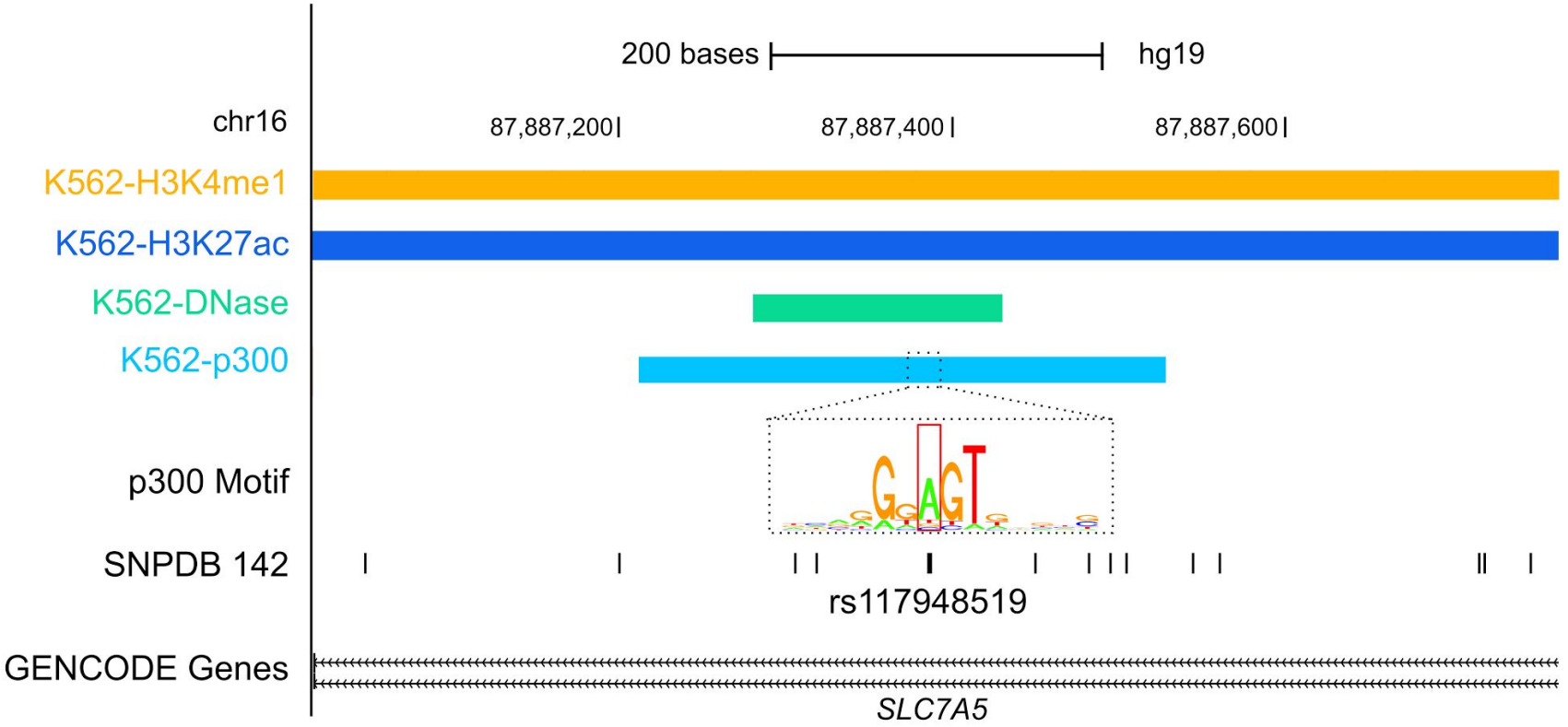


epilogos.broadinstitute.org

Visualization by: Kellis

Top Level Annotations

Variant Annotation



RegulomeDB

HaploReg v4.1

FunSeq2

Visualization by: Snyder, Kellis, Gerstein

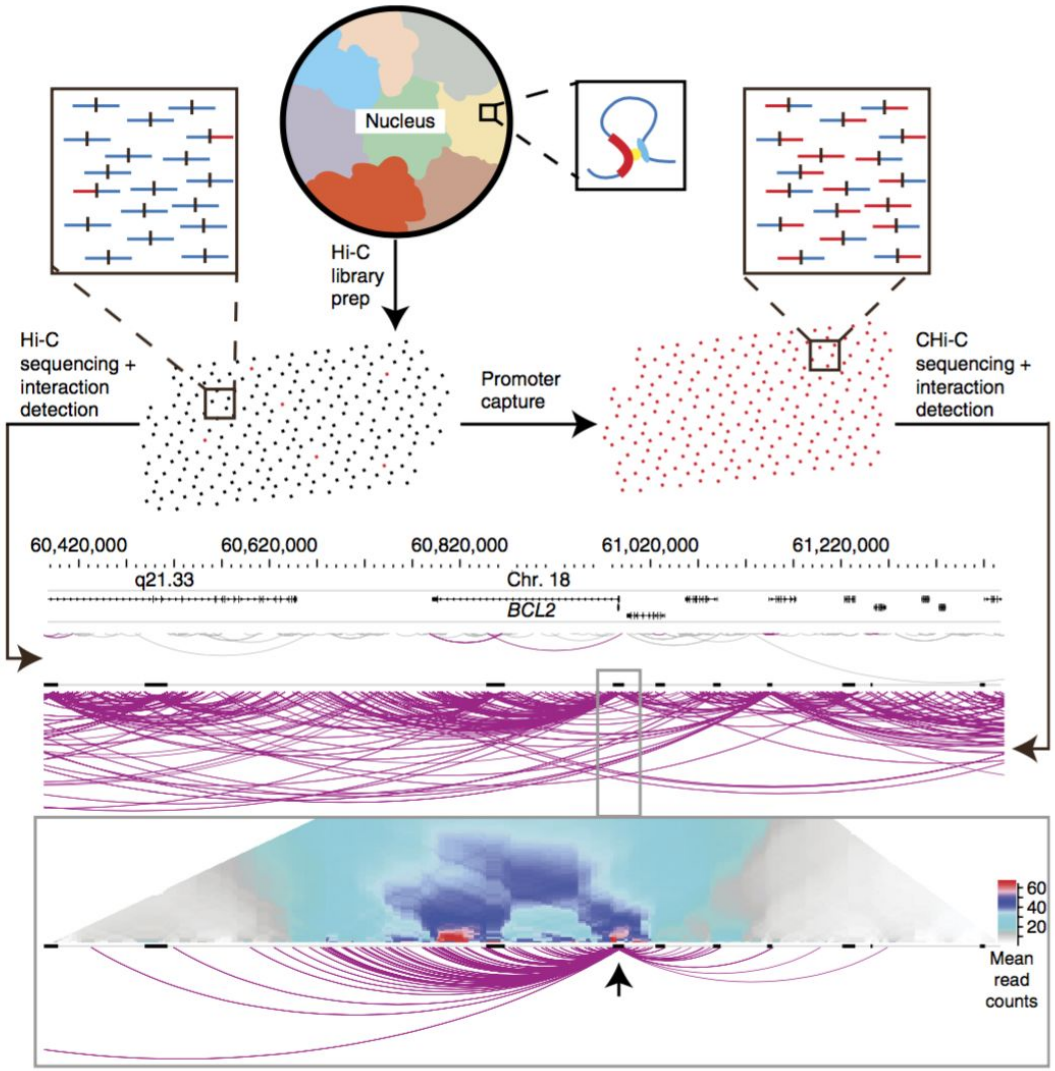
Top Level Annotations

Predicting Target Genes of Enhancers

1. Create benchmark dataset for method comparison
2. Evaluate correlation based methods
3. Integrate additional data to improve performance
4. Input from ENCODE groups & comparison of other methods

Part I: Creating a Benchmark Dataset

Promoter Capture Hi-C



Pros:

- Thousands more high resolution links than previous Hi-C datasets

Cons:

- Links may not represent functional contacts

~50,000 Enhancer-Gene links overlap enhancer-like regions

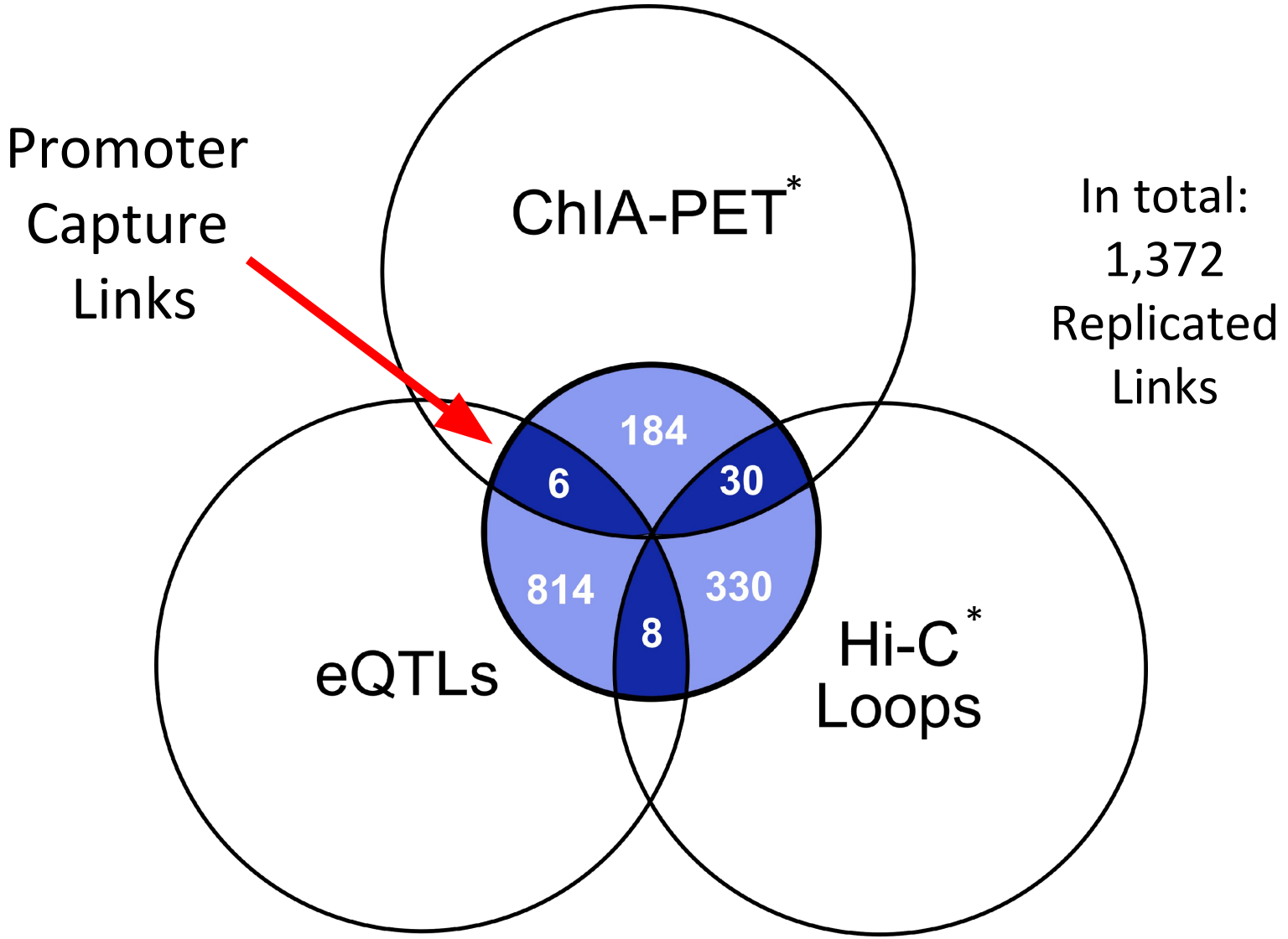
Mifsud, …, Osborne (2015) *Nature Genetics*

Integrating Additional Datasets- GM12878

- ChIA-PET from the Snyder lab targeting RAD21 in GM12878
- eQTLs in lymphoblastoid cells curated by the Kellis Lab in HaploReg (also included LD SNPs $r^2 > 0.8$)
- Hi-C (high resolution) loops in GM12878 from Aiden lab[1]

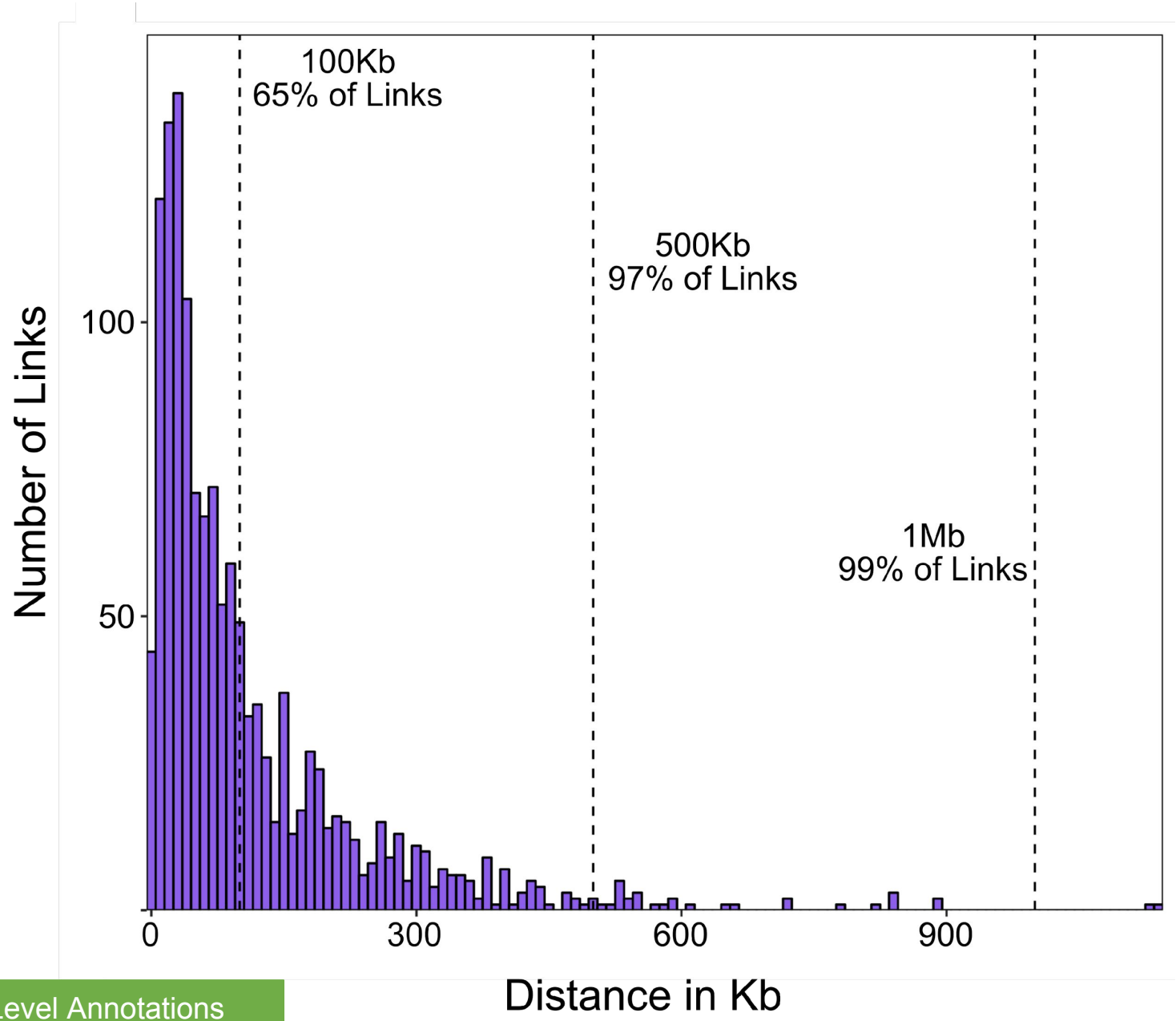
1. Rao, …, Aiden (2014) *Cell*

Overlap of Datasets with Promoter Capture Links



*require one link end to contain only enhancer-like regions and other link end to contain TSSs for only one gene

Distance Between Enhancers and Genes



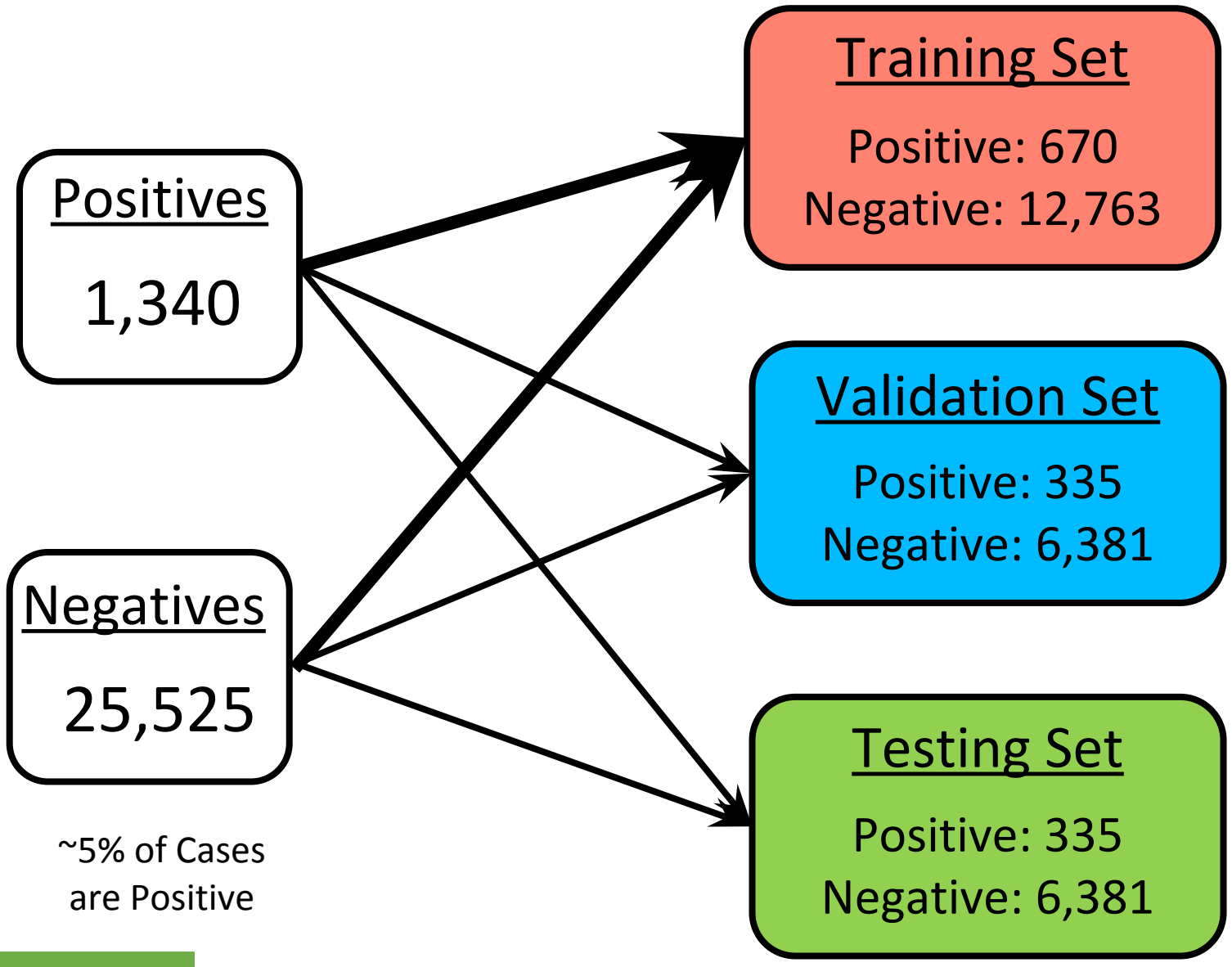
Determining the Negatives

For all enhancer-like regions with at least one positive link, select all genes that meet the following requirements:

#1 – Genes must be within 500Kb

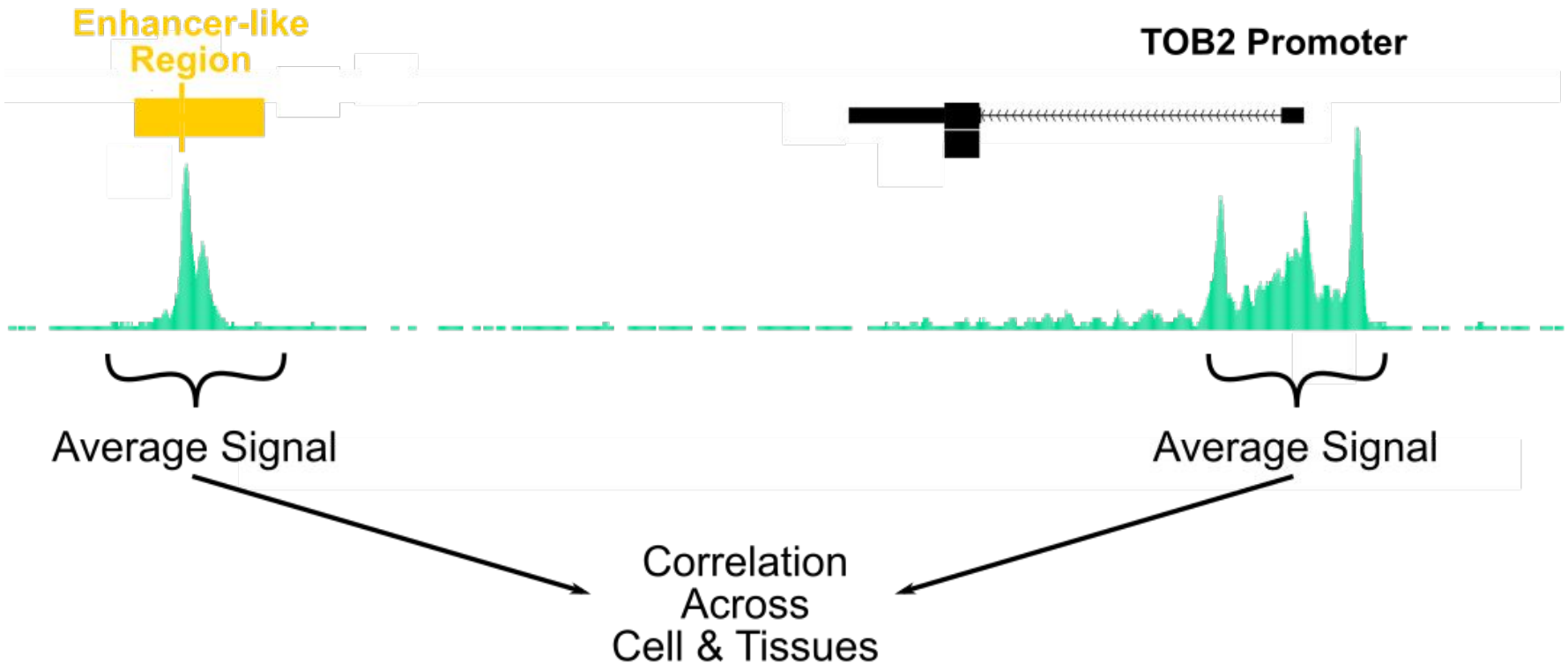
#2 – Genes cannot be linked in any individual dataset (i.e. exclude enhancer-gene pairs with evidence from only one datatype)

Dividing Links into Training, Validation, & Testing Sets



Top Level Annotations

Part II: Evaluation of Correlation Methods



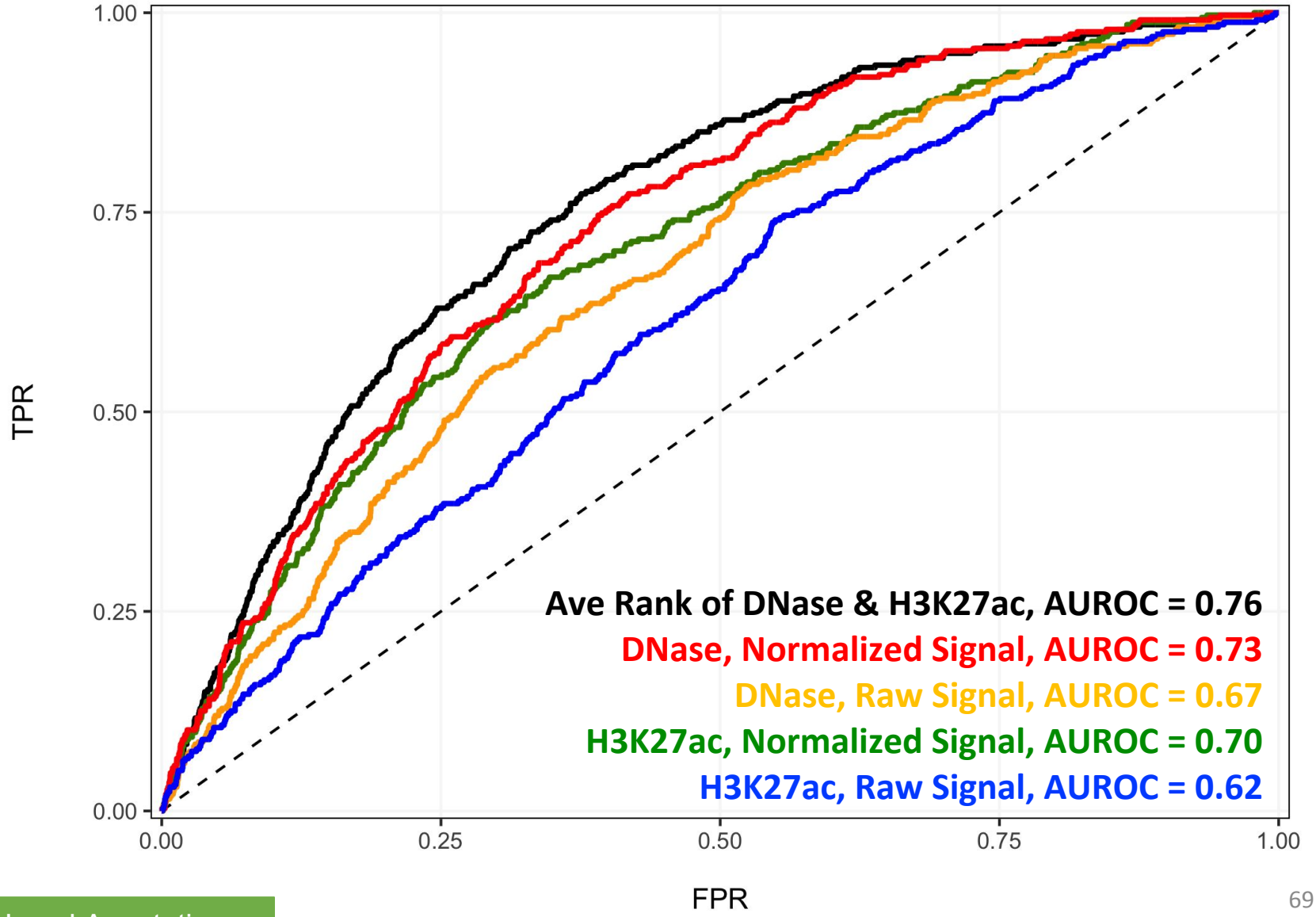
Top Level Annotations

Correlation – Tested Parameters

- Raw signal vs Z-score normalized signal
- DNase signal vs H3K27ac signal
- ENCODE datasets vs. Roadmap datasets
- Pearson vs Spearman correlation
- Rank by correlation coefficient vs permutation p-value[1]

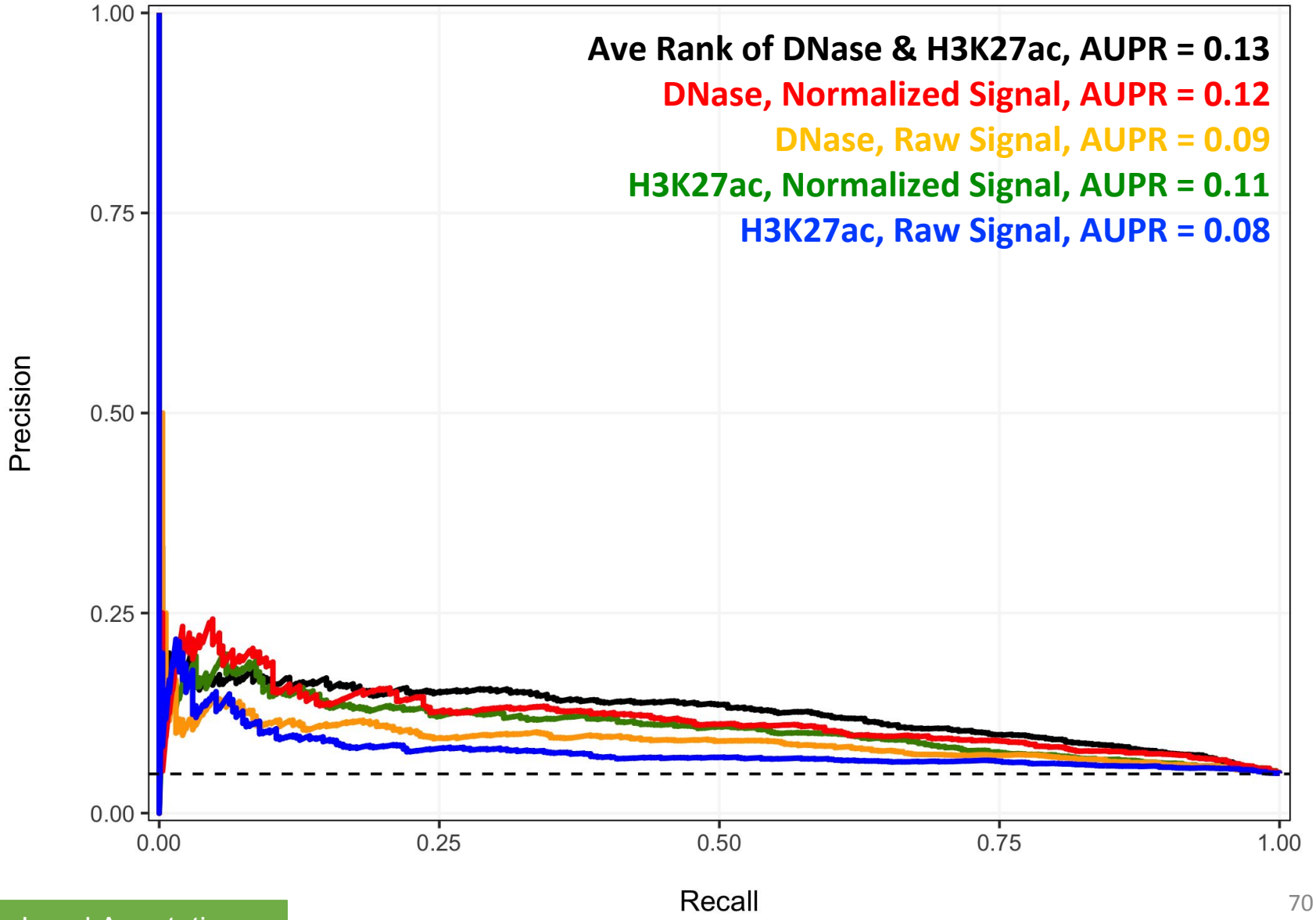
1. Method adapted from Sheffield, …, Furey (2013) *Genome Research*

ROC - Correlation Methods



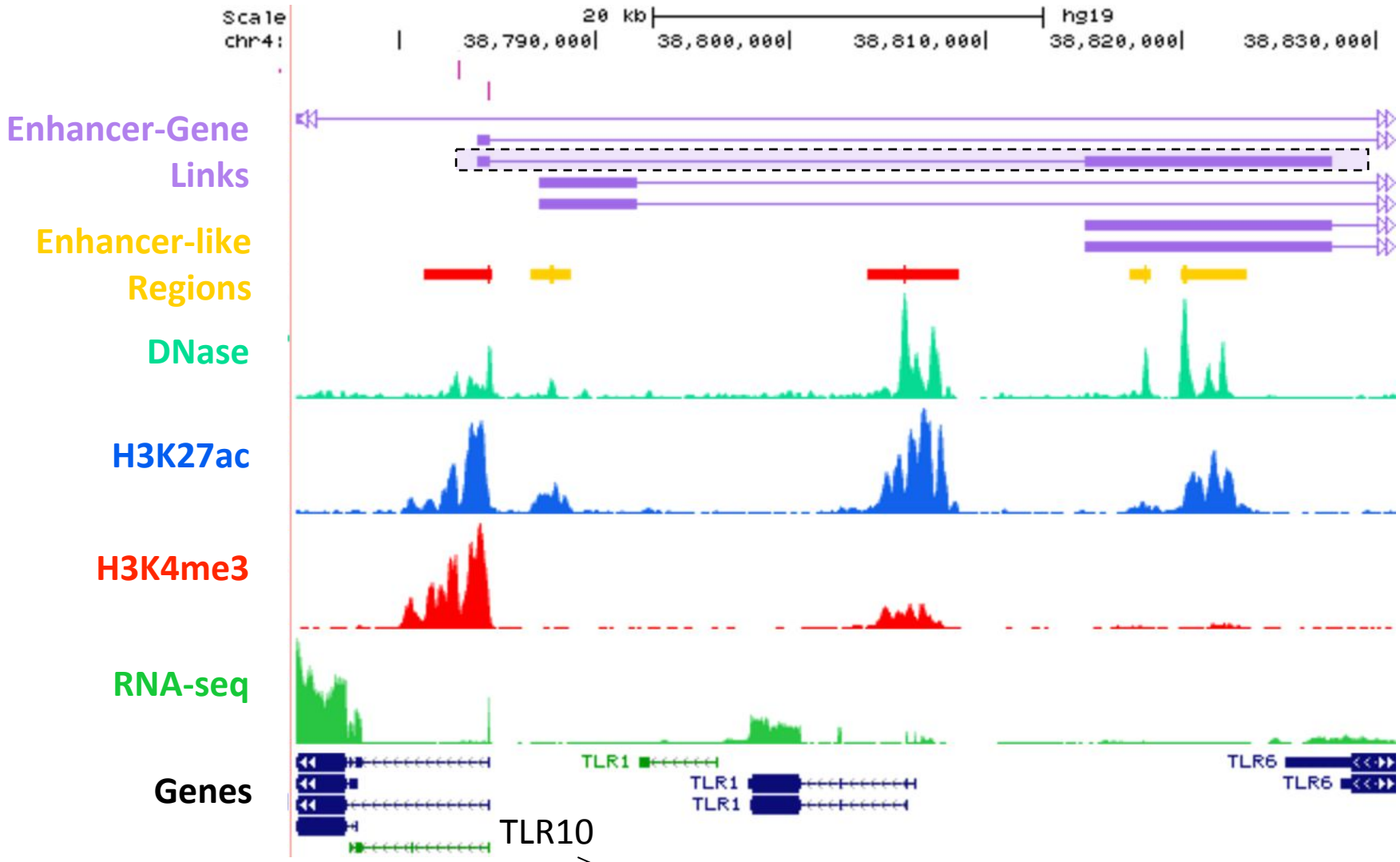
Top Level Annotations

PR - Correlation Methods



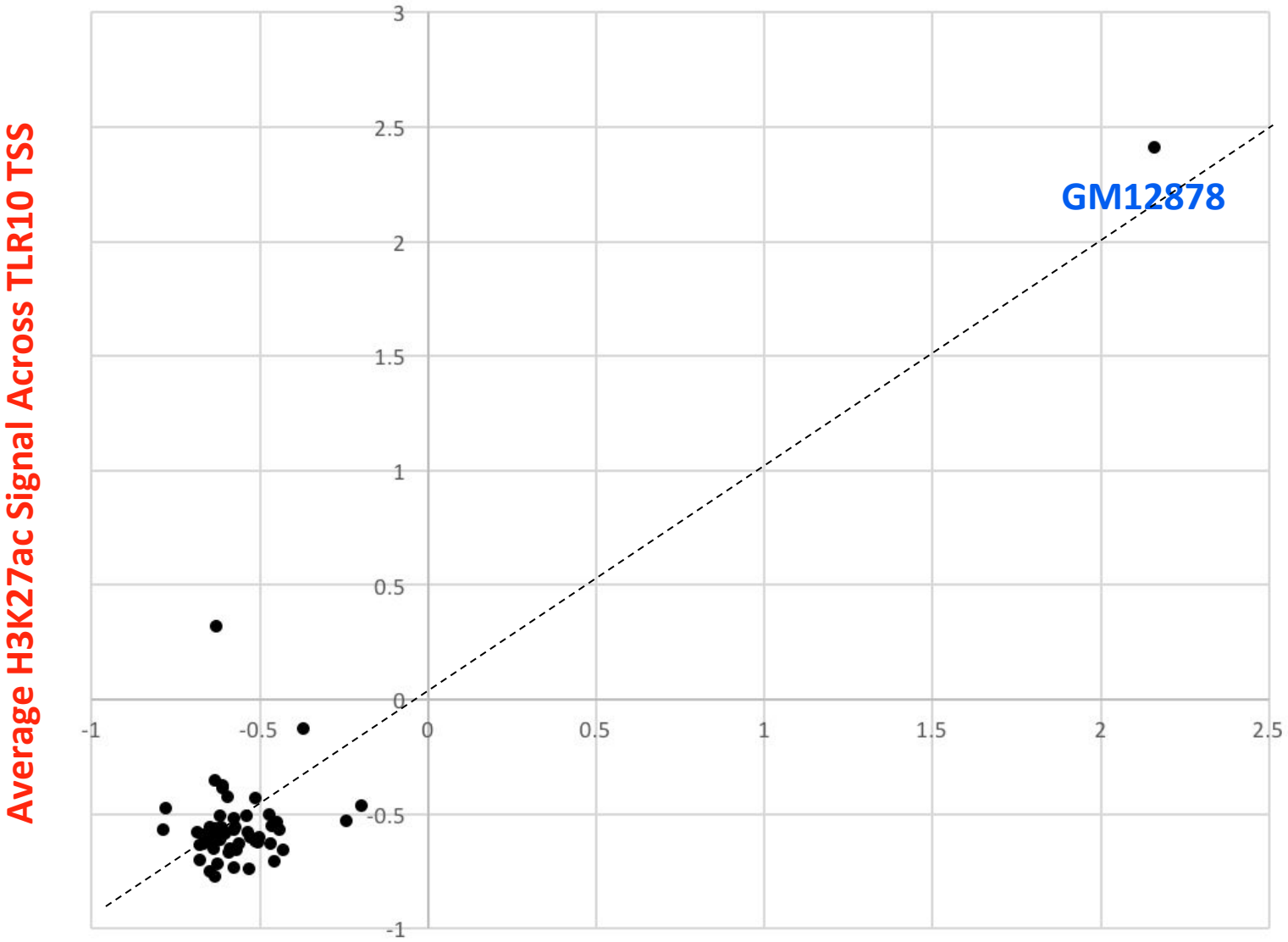
Top Level Annotations

In Some Cases Correlation Accurately Predicts Links

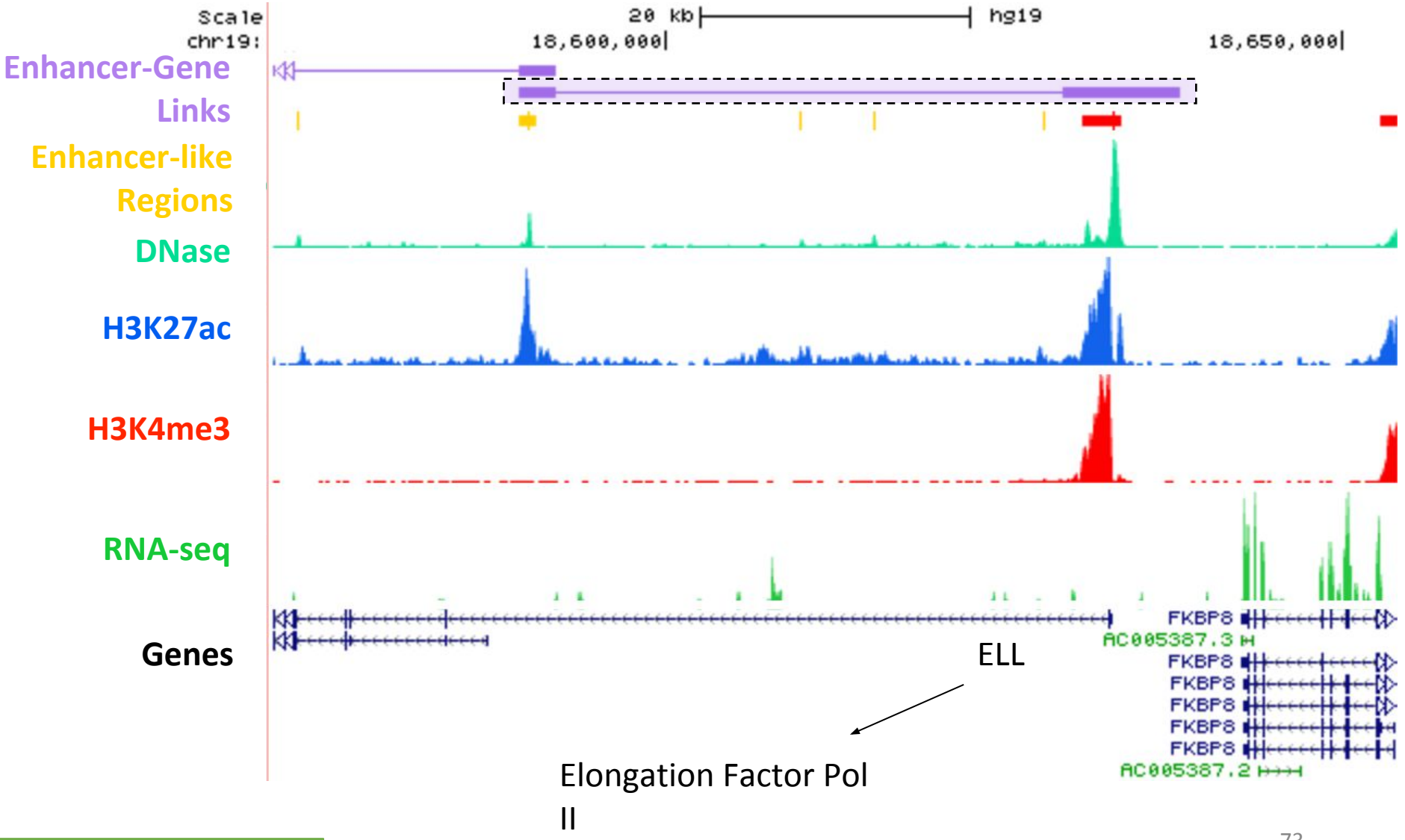


Important to Innate Immune System

In Some Cases Correlation Accurately Predicts Links

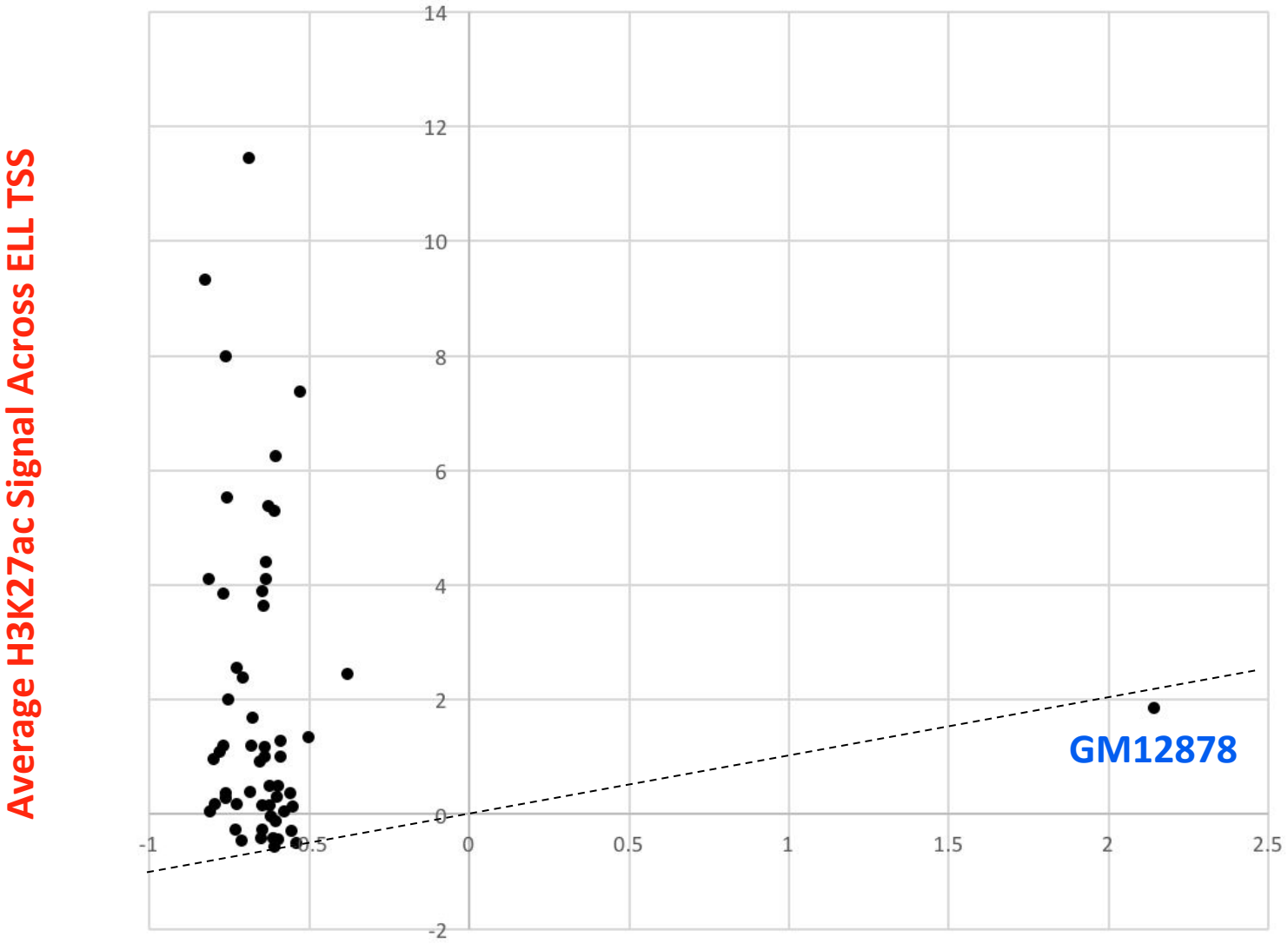


In Many Cases Correlation Does Not Accurately Predict Links



Top Level Annotations

In Many Cases Correlation Does Not Accurately Predict Links



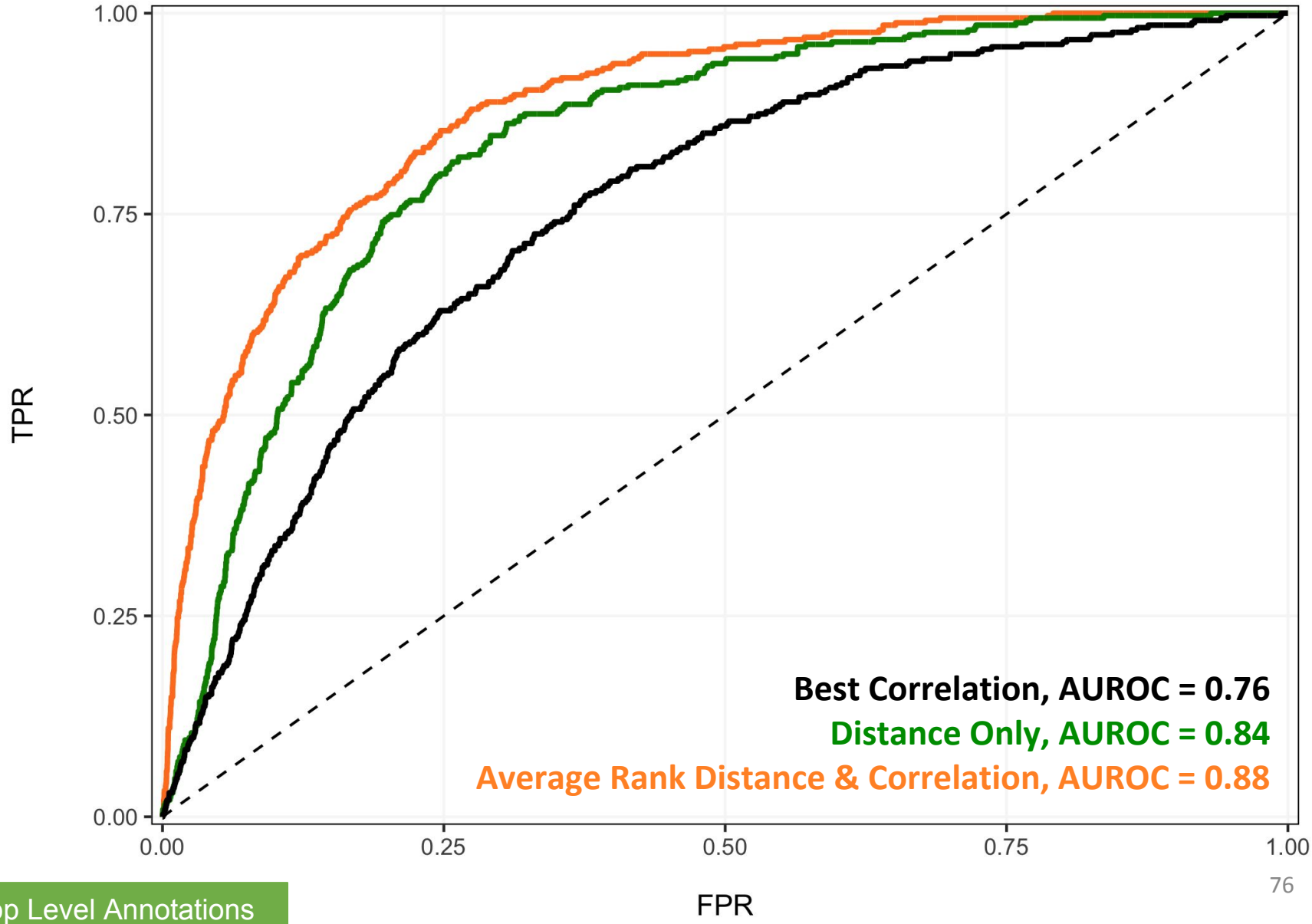
Incorporating Distance Information

Distance is an important feature in predicating enhancer-gene links, but using a hard cutoff (e.g. 100Kb) results in missing 1/3 of links

We instead tested:

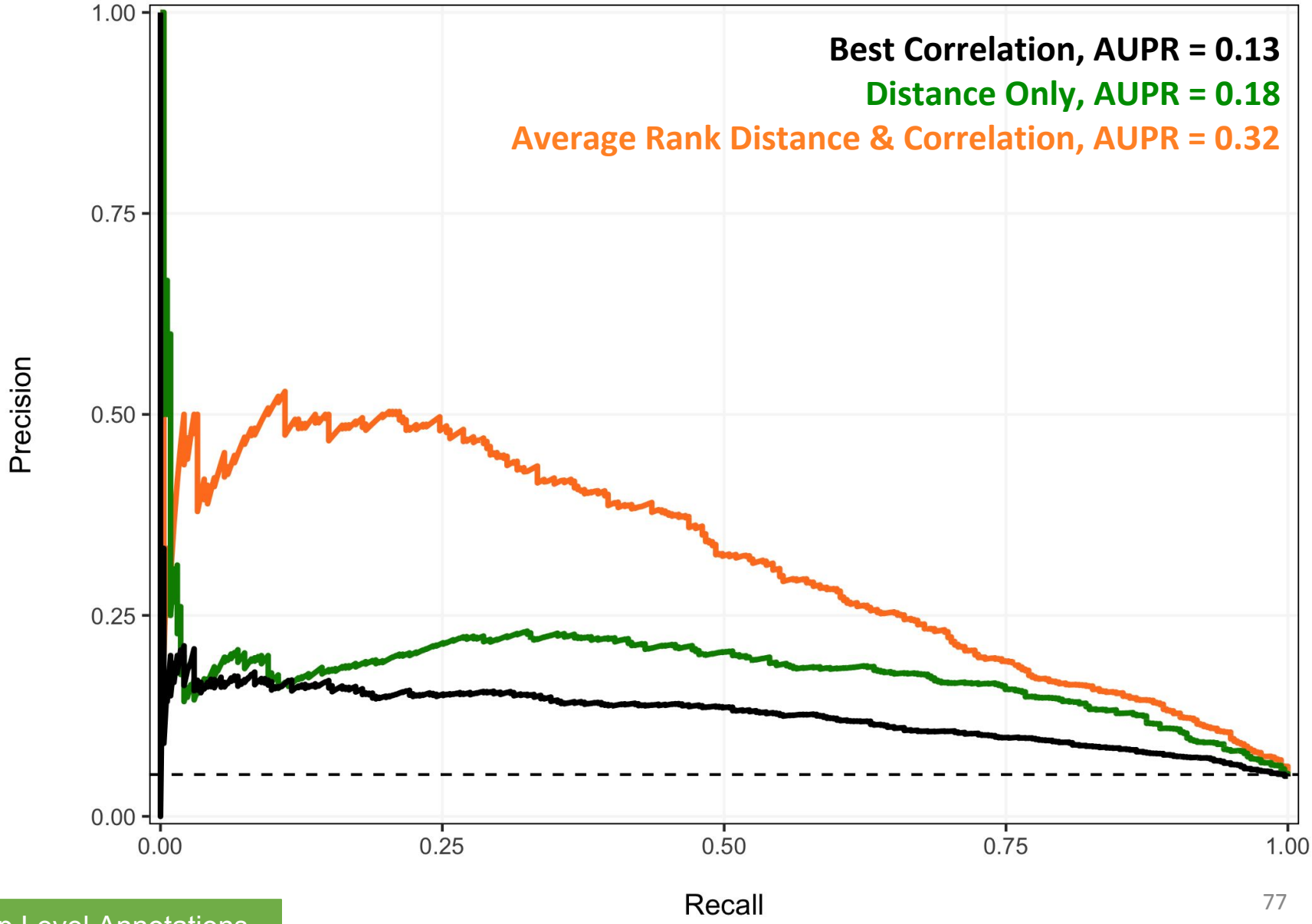
- Ranking by distance
- Average rank of distance and best performing correlation method (average rank of DNase and H3K27ac)

Incorporating Distance Improves Performance



Top Level Annotations

Incorporating Distance Improves Performance



Top Level Annotations

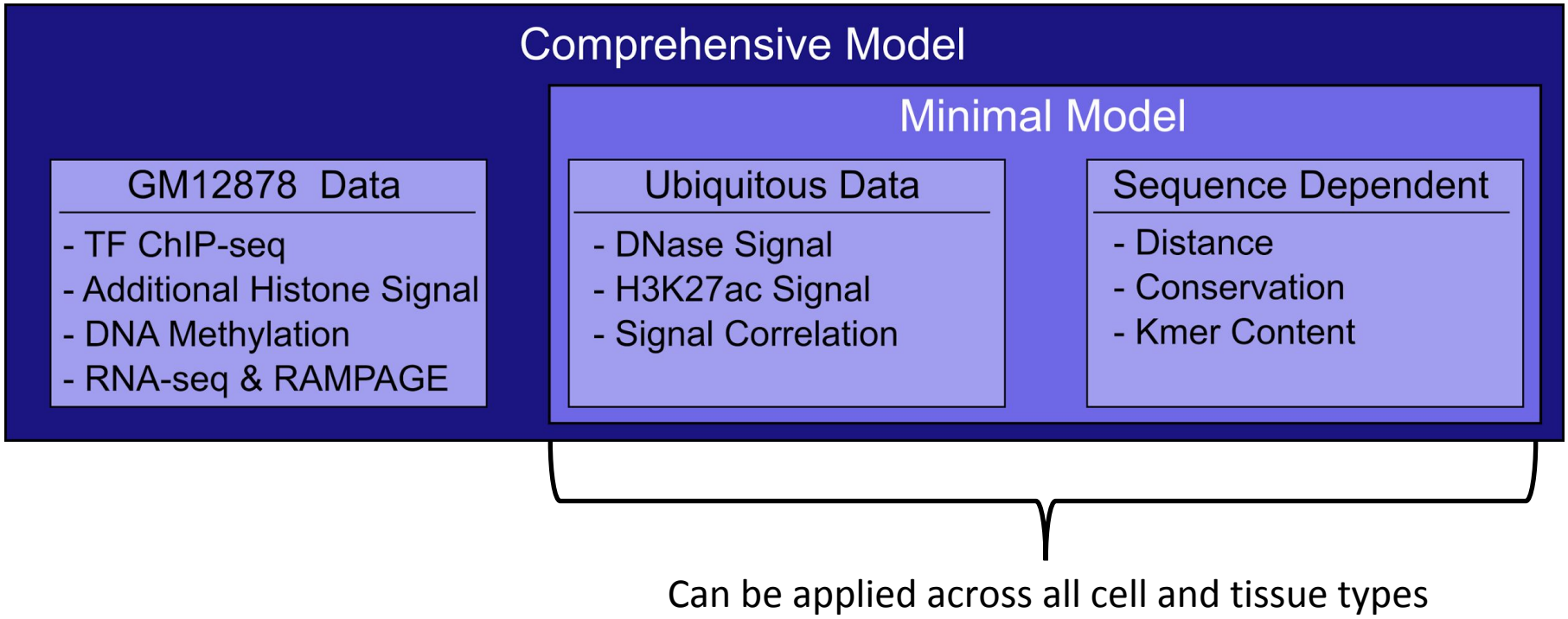
Part II: Conclusions

- For correlation analysis:
  - DNase slightly outperforms H3K27ac
  - It is better to use Z-score normalized signal over raw signal
  - Pearson correlation coefficient out performs Spearman
  - Ranking by correlation coefficient outperforms ranking by p-value (and is much faster!)
- Incorporating distance information dramatically increases performance

Part III: Developing Random Forest Model

Top Level Annotations

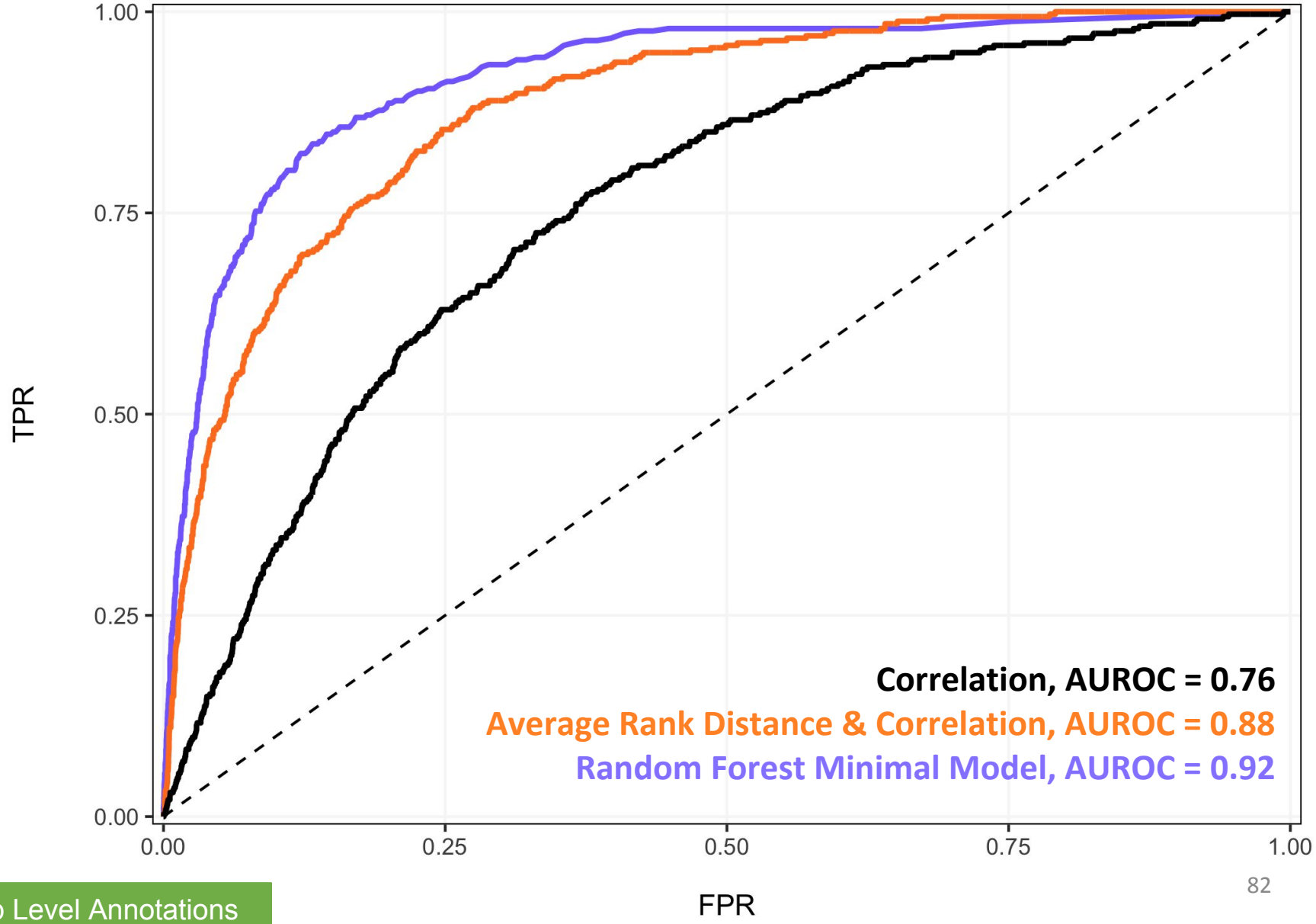
Developing Two Random Forest Models



Minimal Model Features

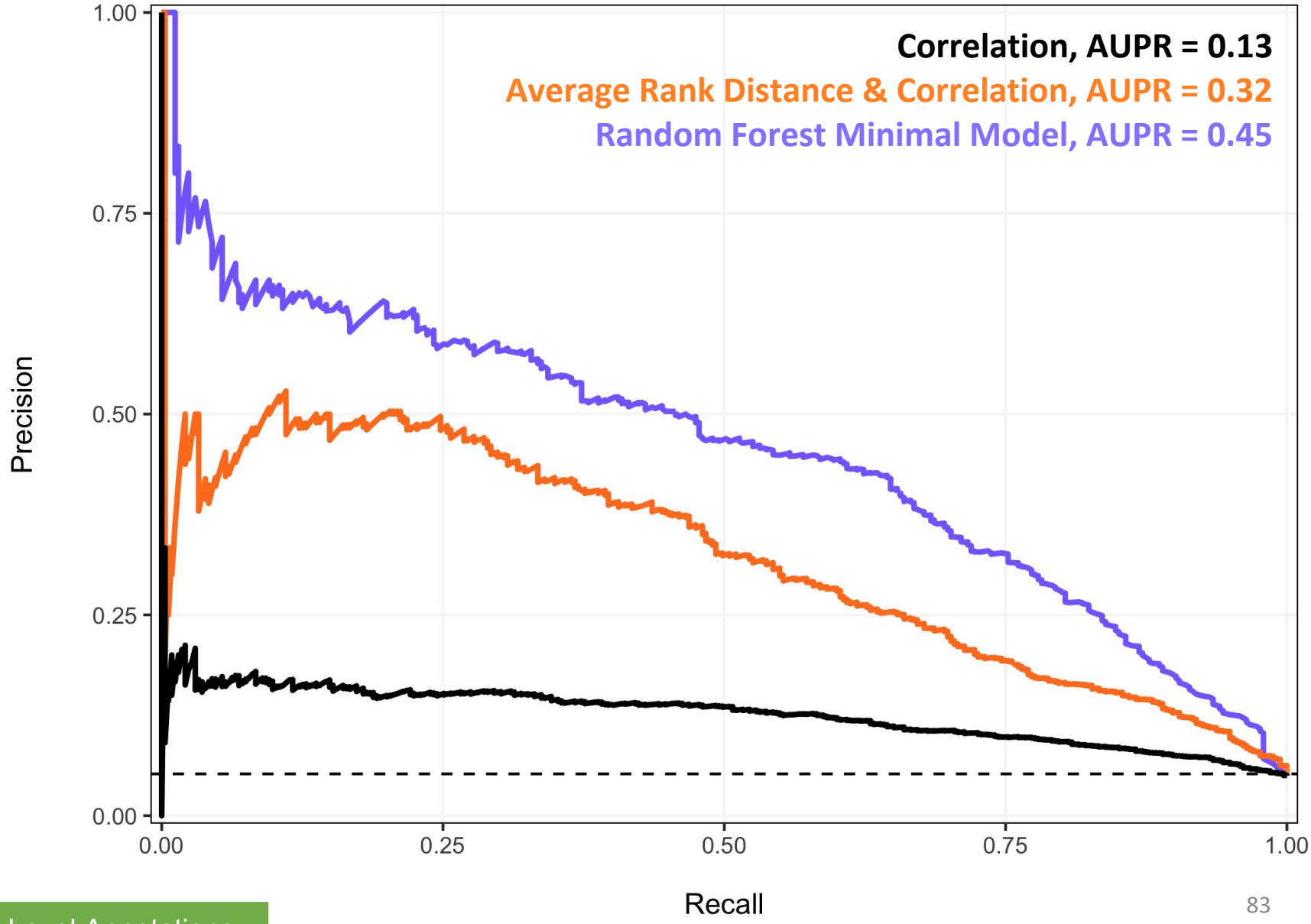
- Minimum distance between enhancer and gene TSS
- Average conservation across enhancer and promoter
- Average DNase Signal across enhancer and promoter
- Average H3K27ac Signal across enhancer and promoter
- Correlation of K-mers (tested 3-6mer)
- Using signals across multiple cell and tissue types:
  - Correlation of DNase signal
  - Mean and standard deviation of DNase signal
  - Correlation of H3K27ac Signal
  - Mean and standard deviation of H3K27ac signal

ROC – Random Forest Minimal Model



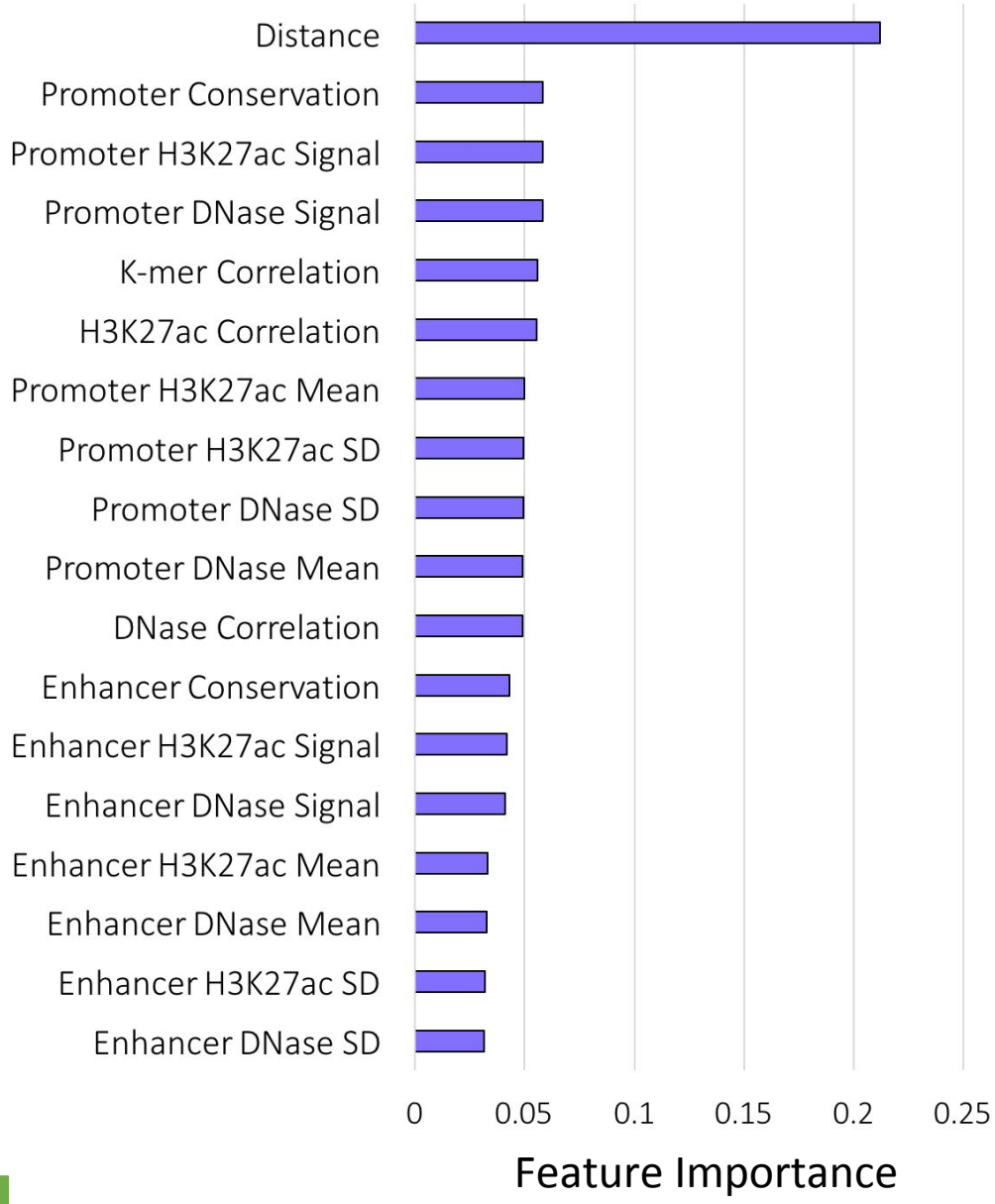
Top Level Annotations

PR – Random Forest Minimal Model



Top Level Annotations

Feature Importance - Minimal Model

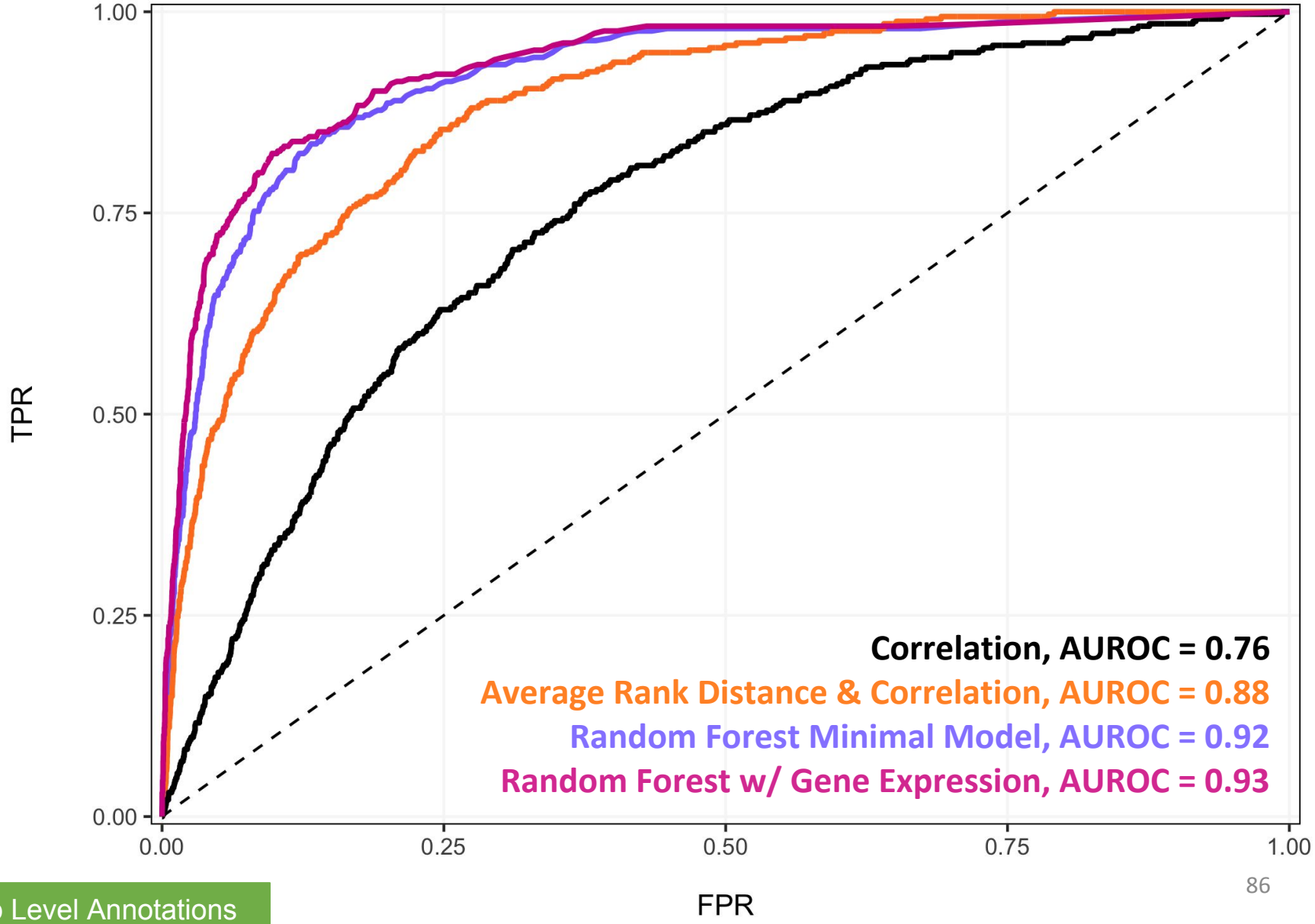


Top Level Annotations

Comprehensive Model Features

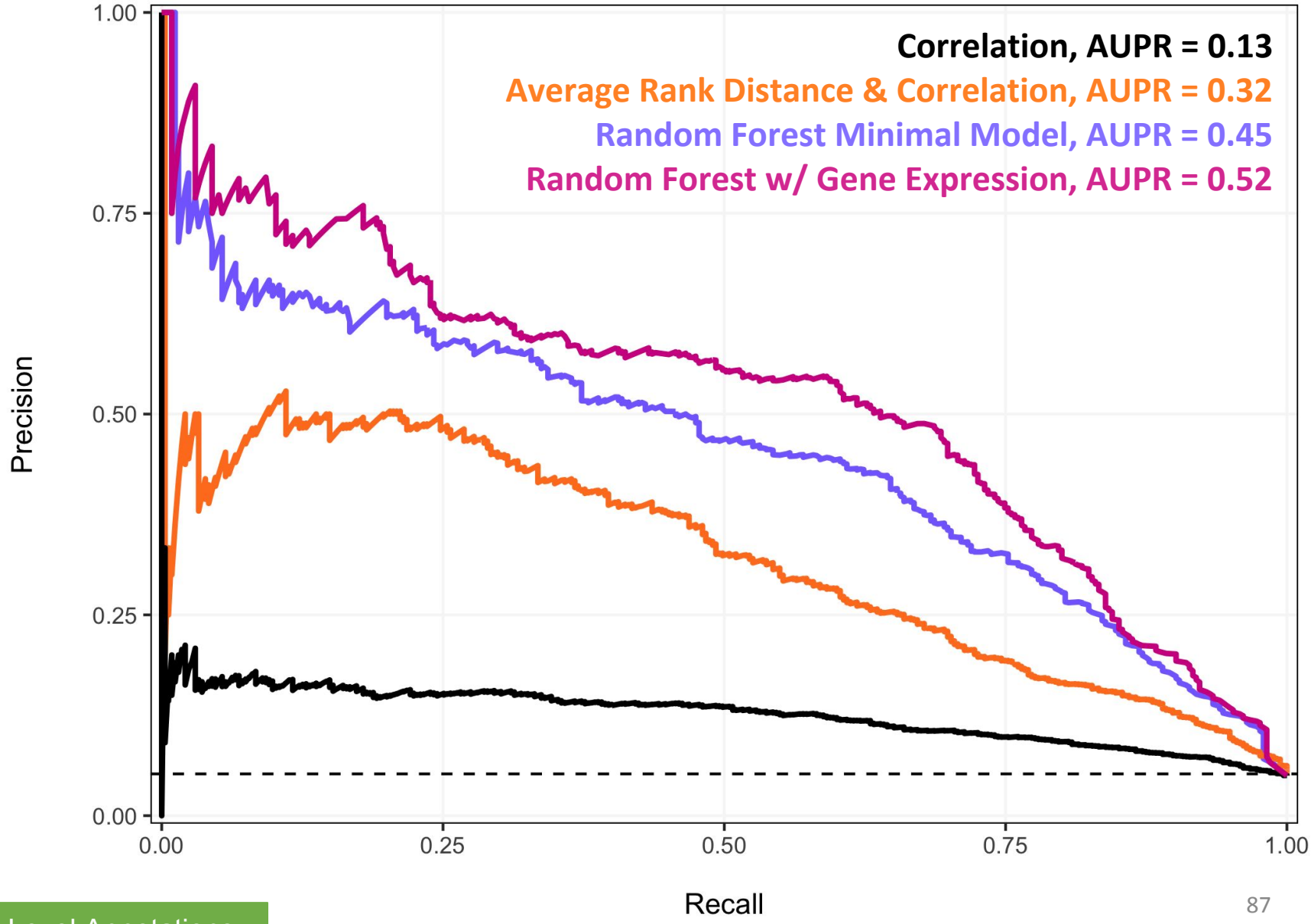
- Minimal model features
- **Gene expression** & RAMPAGE Peaks
- Signal from other Histone Marks (H3K4me1/2/3, H3K27me3, H3K36me3)
- TF peaks signal (Pol2, p300, CTCF)

ROC – Random Forest with Gene Expression



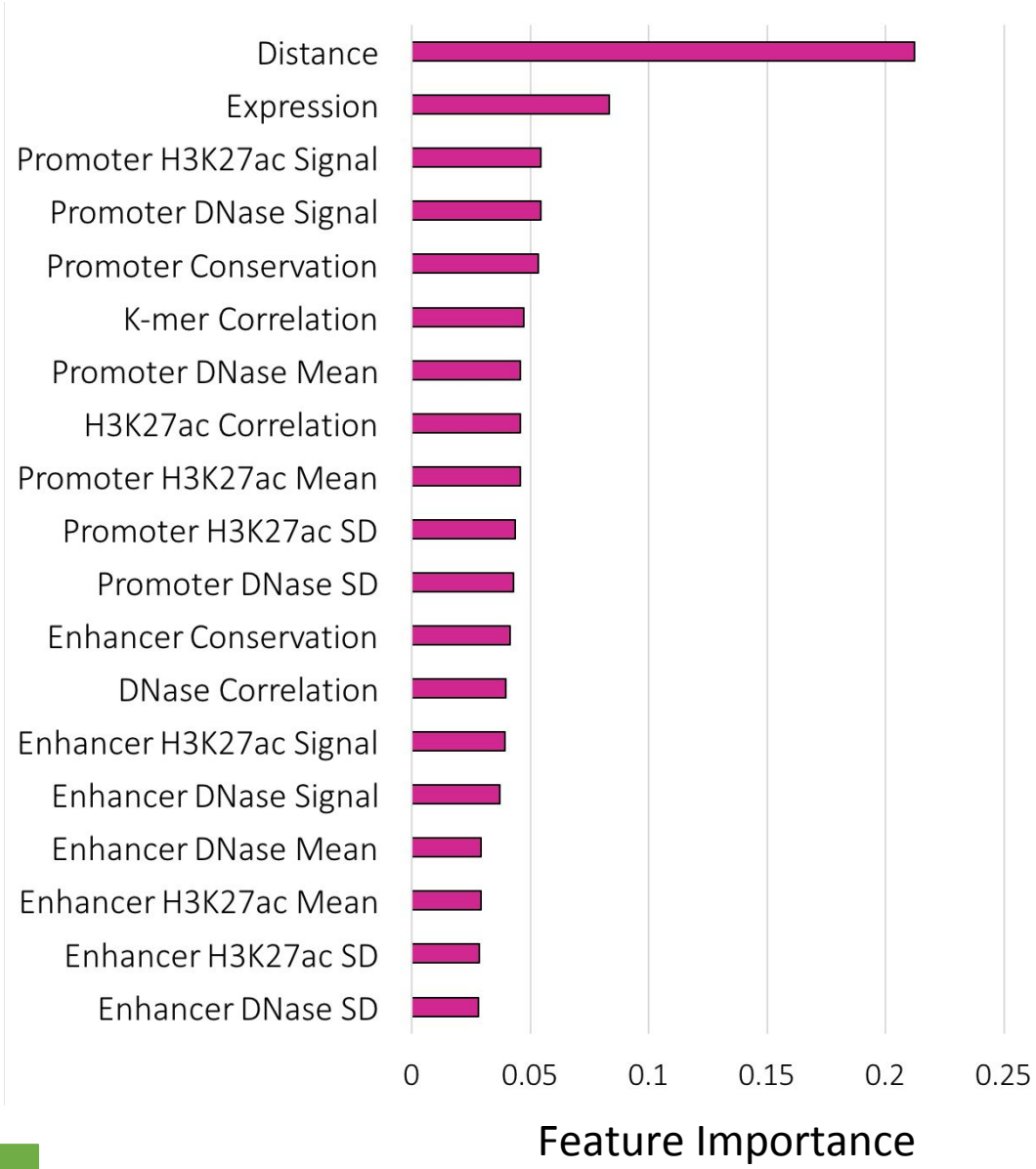
Top Level Annotations

PR – Random Forest with Gene Expression



Top Level Annotations

Feature Importance – RF with Gene Expression



Future Directions

- In corporate additional training and testing data, such as massively parallel reporter assays and STARR-seq
- Retest additional features when training set is large
- Prediction of target genes remains a major challenge.
- We also would like to define other types of regulatory elements.

Top Level Annotations

Acknowledgements

Weng Lab

Jill Moore
Michael Purcaro
Arjan van der Velde
Tyler Borrman
Henry Pratt
Sowmya Iyer
Jie Wang

Stam Lab

John Stamatoyannopoulos
Bob Thurman
Richard Sandstrom

Gerstein Lab

Mark Gerstein
Anurag Sethi

ENCODE Consortium

Brad Bernstein
Ross Hardison
Len Pennacchio
Axel Visel
Bing Ren
Anshul Kundaje
Data Production Groups