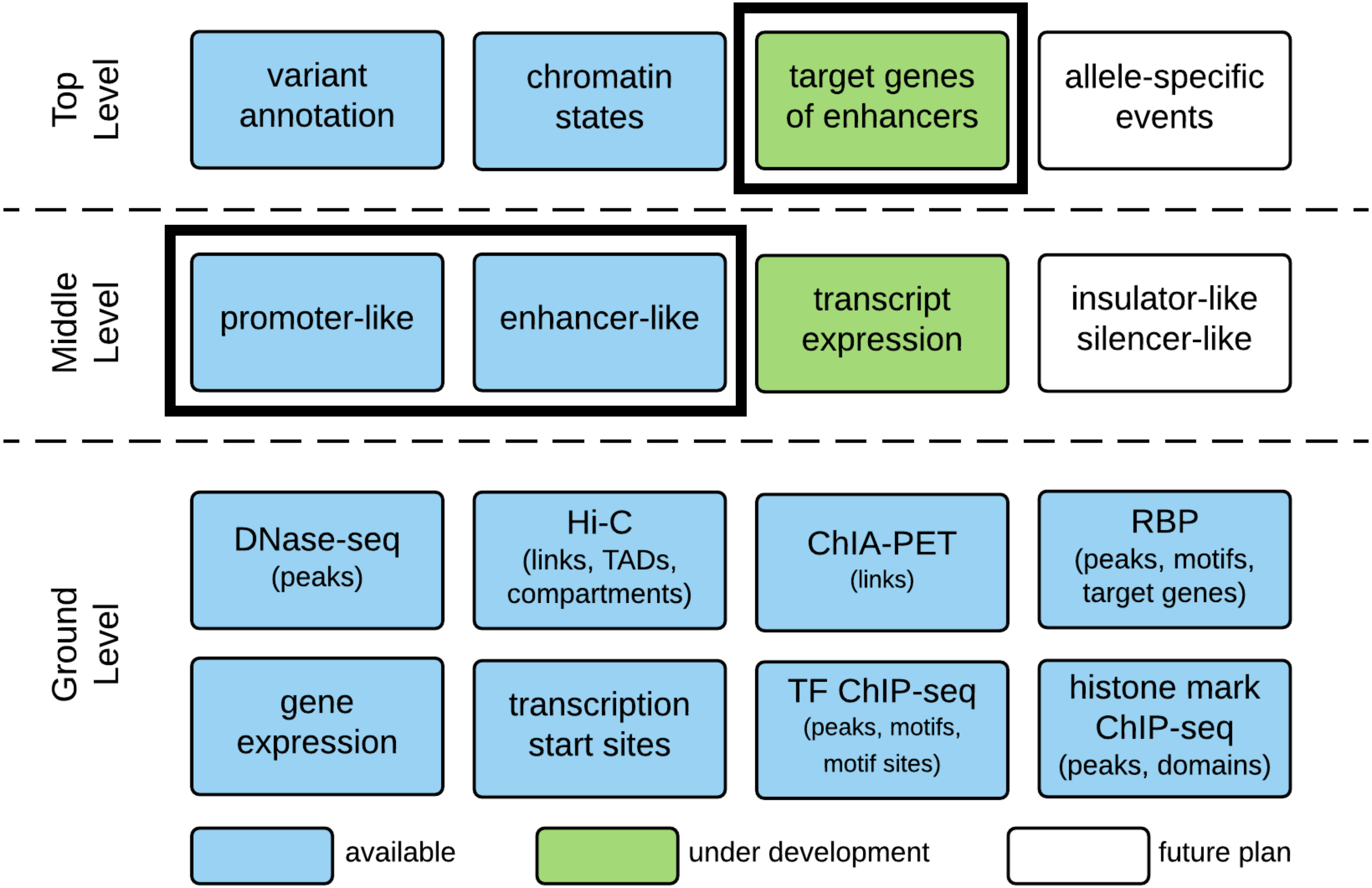
UMASS
MEDICAL
SCHOOL

Expanding the Encyclopedia: Connecting Regulatory Elements with Target Genes

Jill E. Moore
Advisor: Zhiping Weng
University of Massachusetts Medical School
Program in Bioinformatics and Integrative Biology

AWG Call – May 2016

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

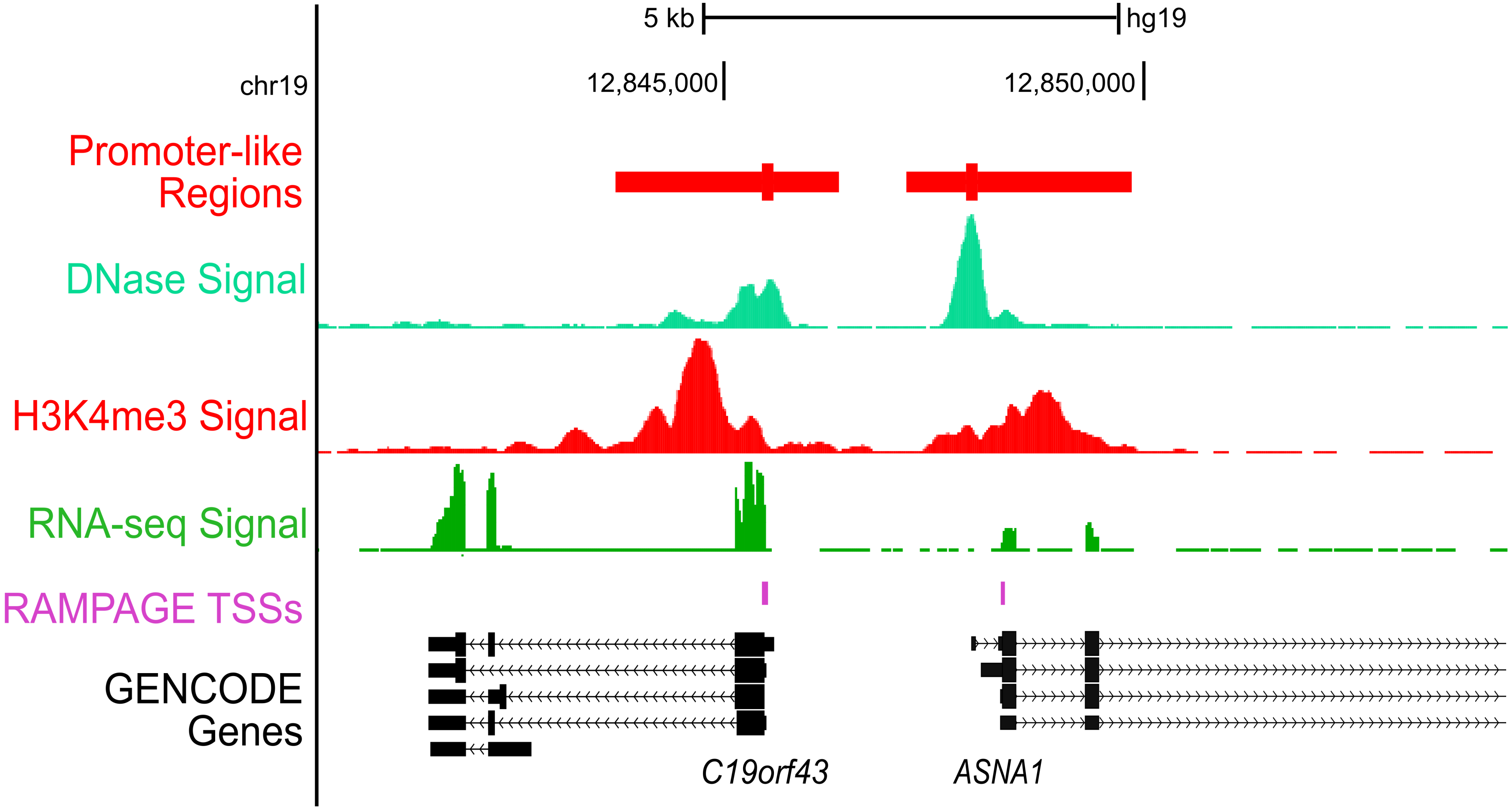
Ground Level

- 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

https://www.encodeproject.org/data/annotations/

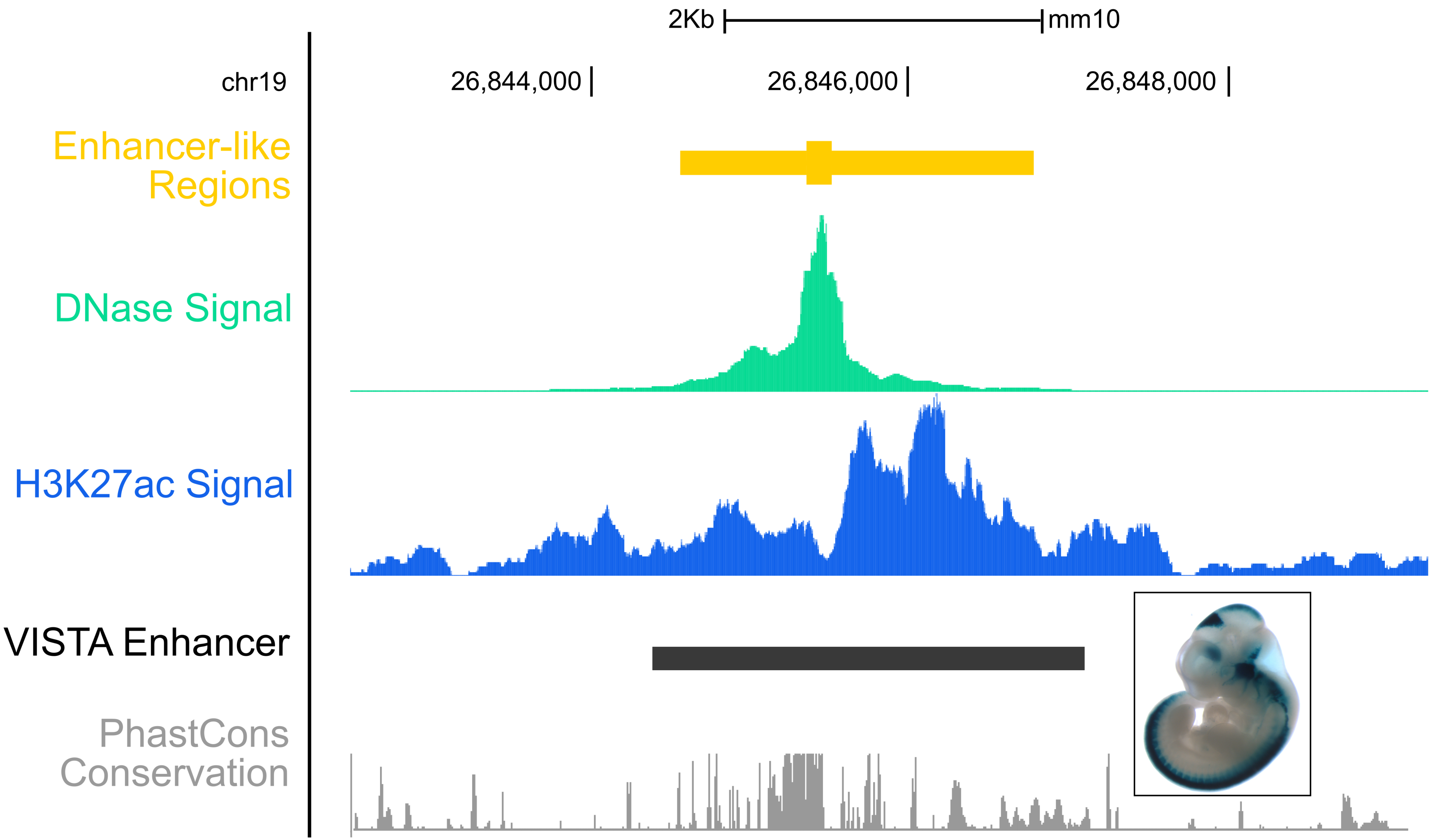
Promoter-like Regions



We predict Promoter-like regions by ranking DNase peaks by the average rank of H3K4me3 and DNase signals

http://zlab-annotations.umassmed.edu/promoters/methods

Enhancer-like Regions



We predict Enhancer-like regions by ranking DNase peaks by the average rank of H3K27ac and DNase signals

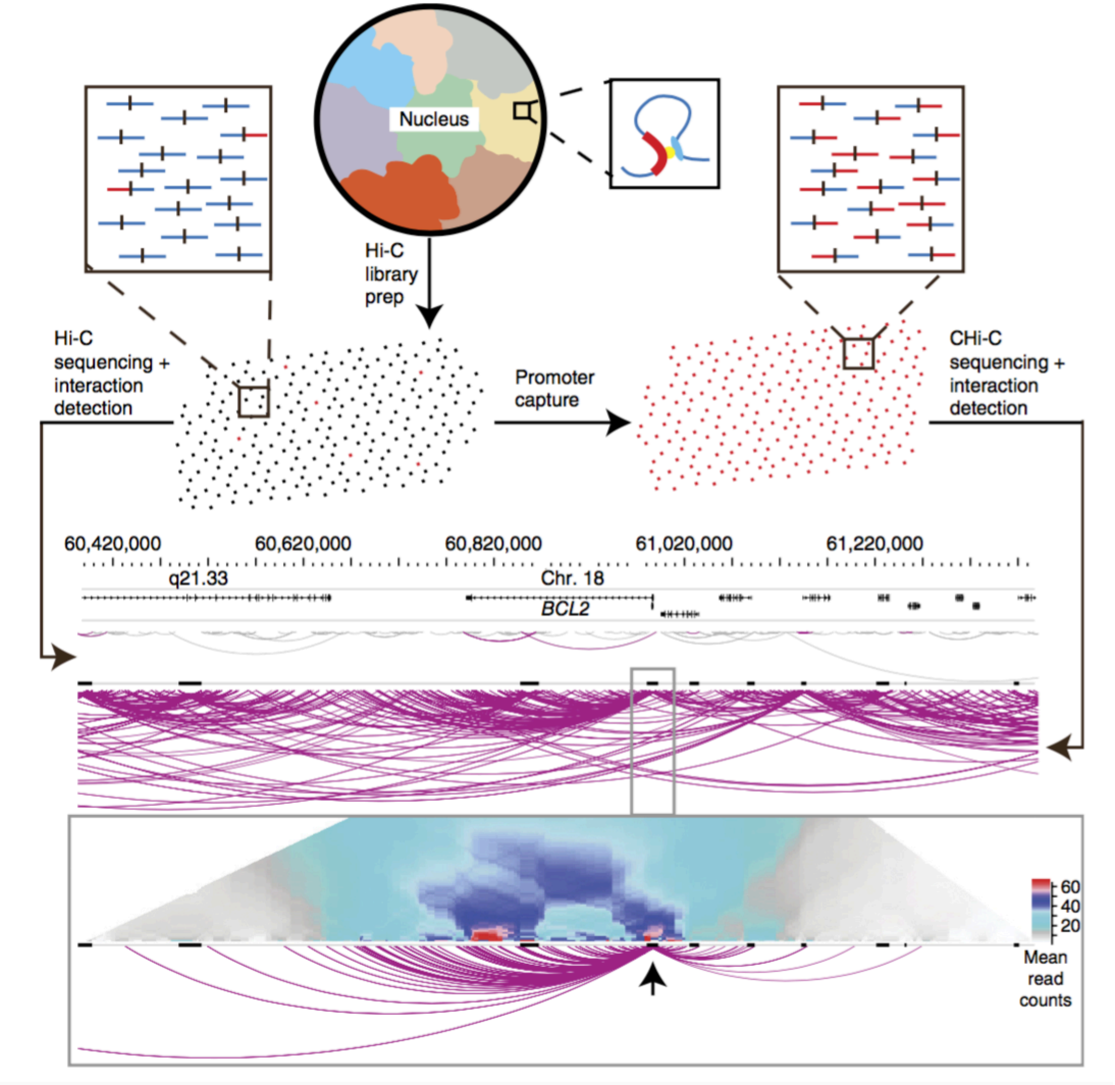
http://zlab-annotations.umassmed.edu/enhancers/methods

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

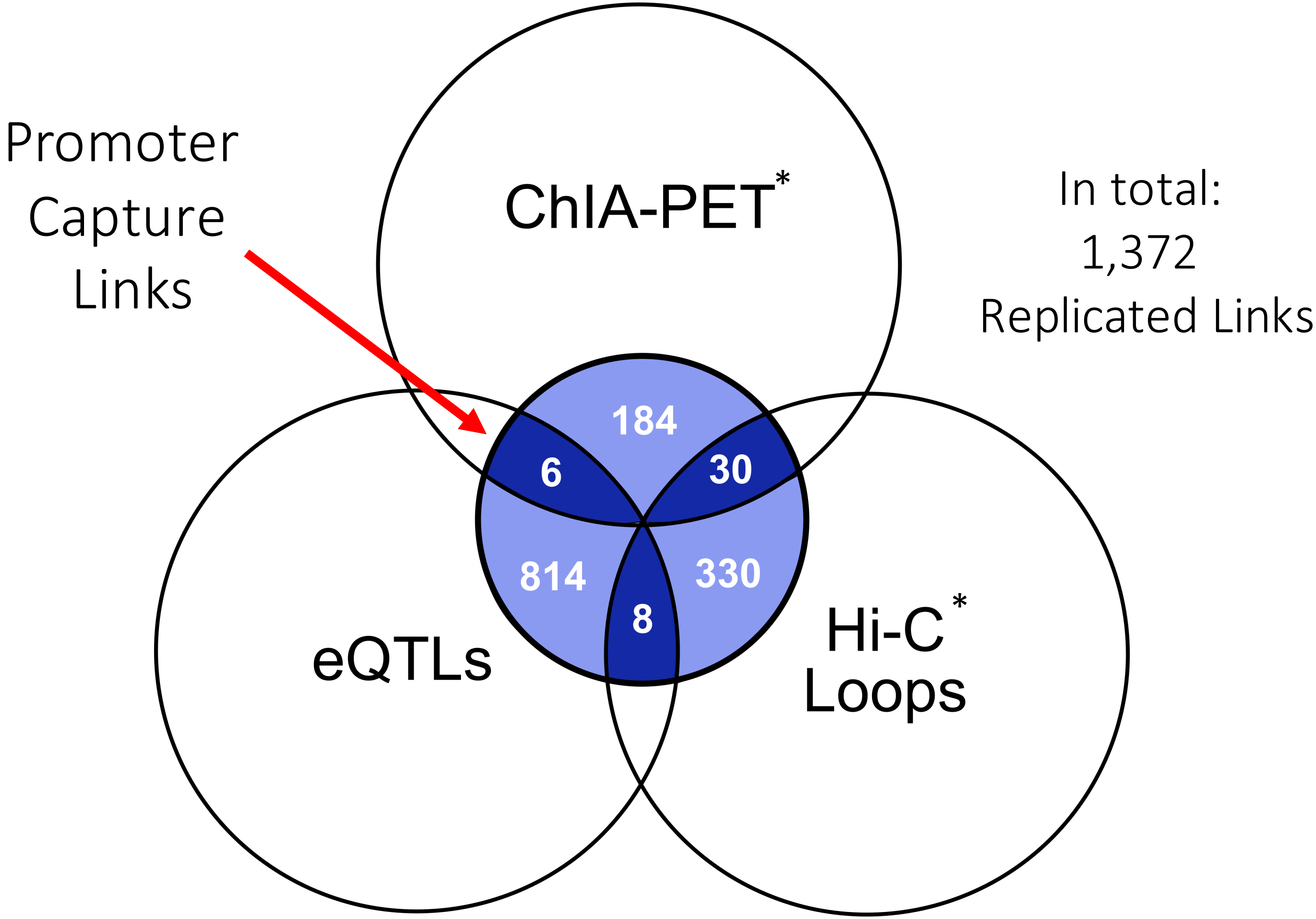
- ChIA-PET from the Snyder lab targeting RAD21 in GM12878

- 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*

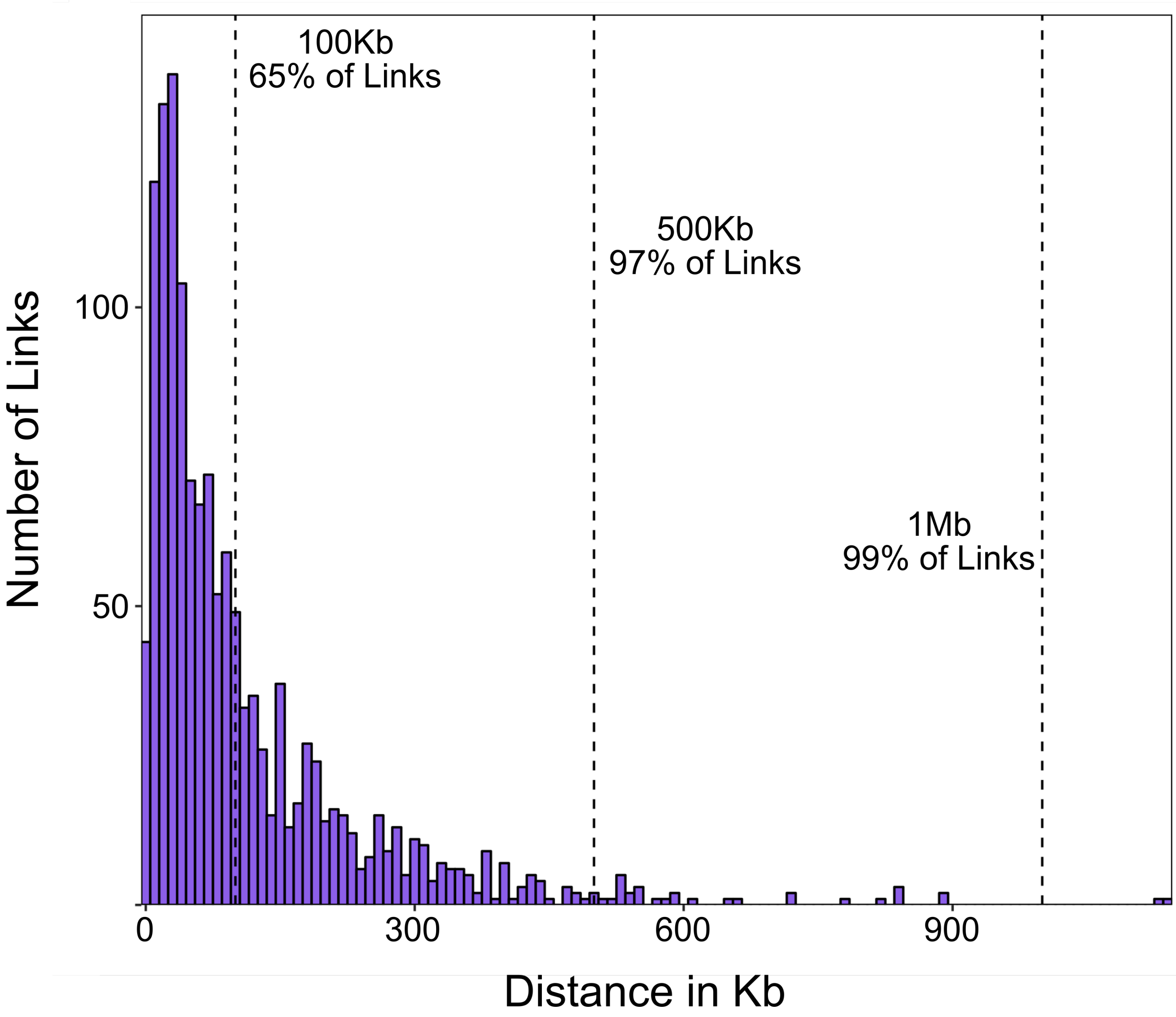
Overlap of Datasets with Promoter Capture Links



In total:
1,372
Replicated 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



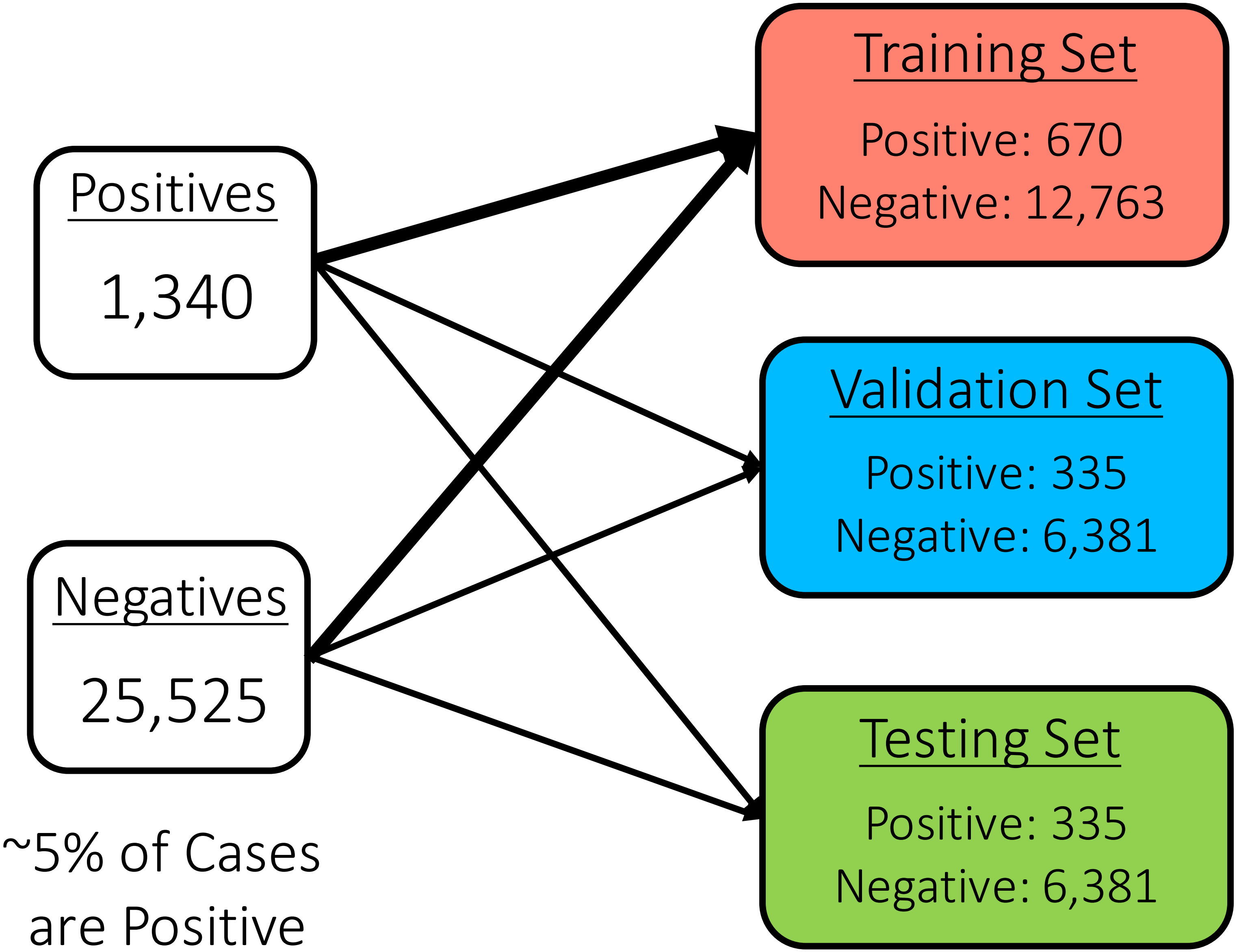
Determining the Negatives

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



Positives

1,340

Negatives

25,525

Training Set

Positive: 670

Negative: 12,763

Validation Set

Positive: 335

Negative: 6,381

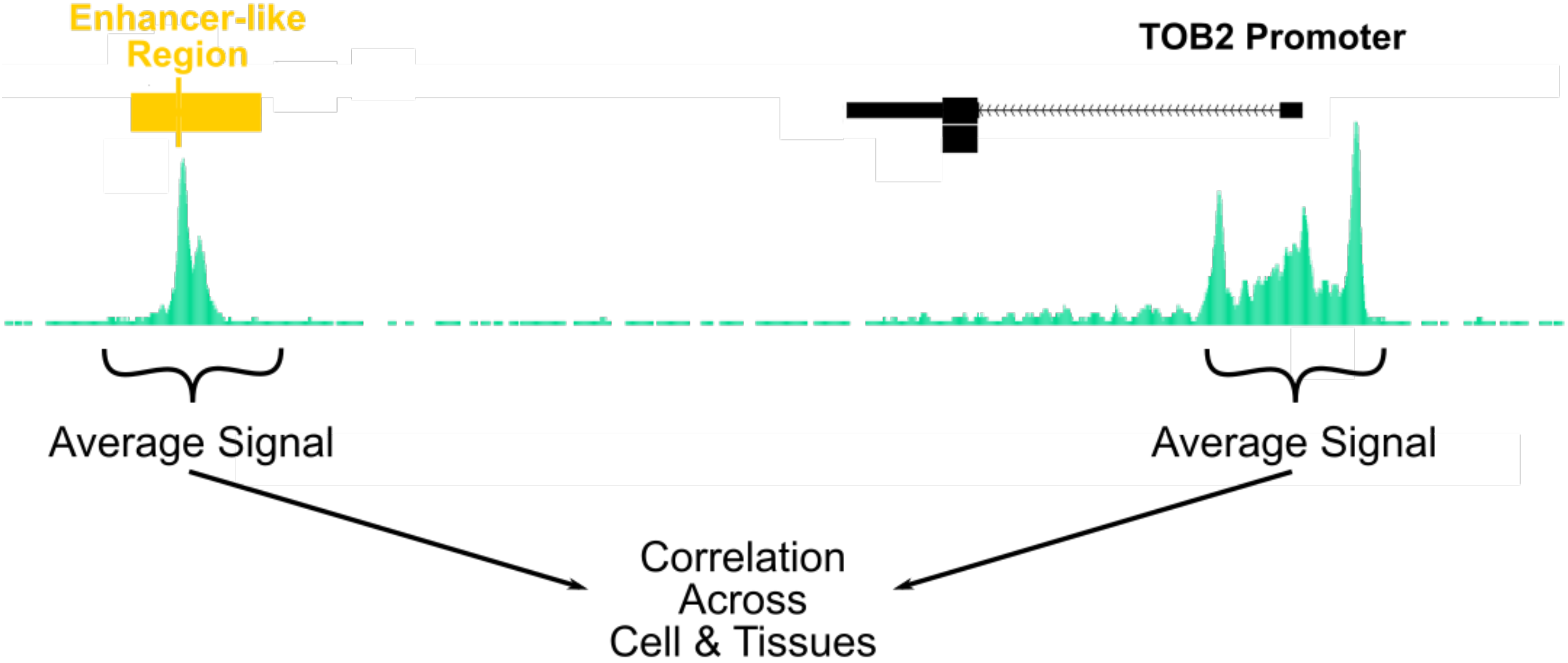
Testing Set

Positive: 335

Negative: 6,381

~5% of Cases are Positive

Part II: Evaluation of Correlation Methods



Enhancer-like Region

TOB2 Promoter

Average Signal

Average Signal

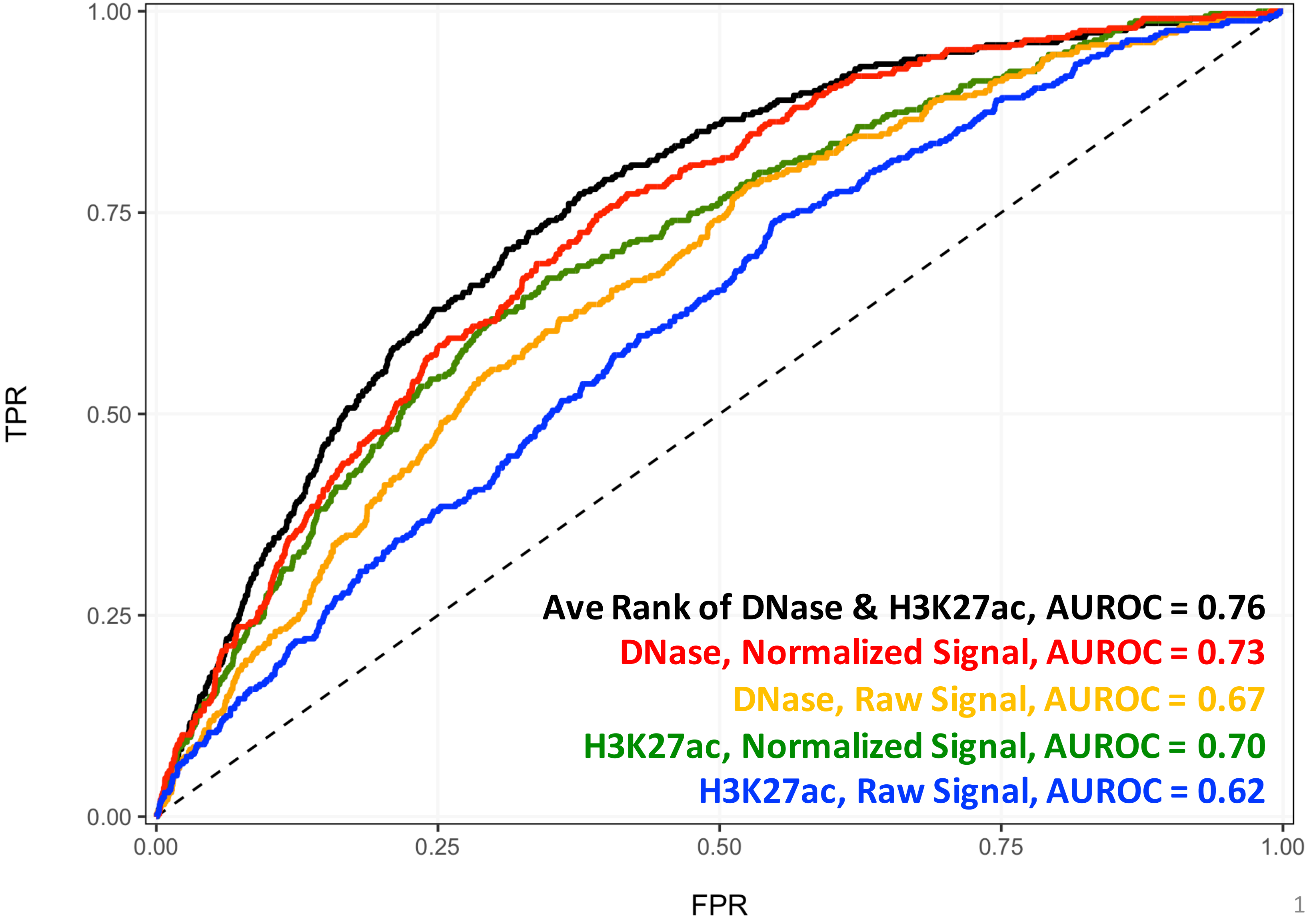
Correlation Across Cell & Tissues

Correlation – Tested Parameters

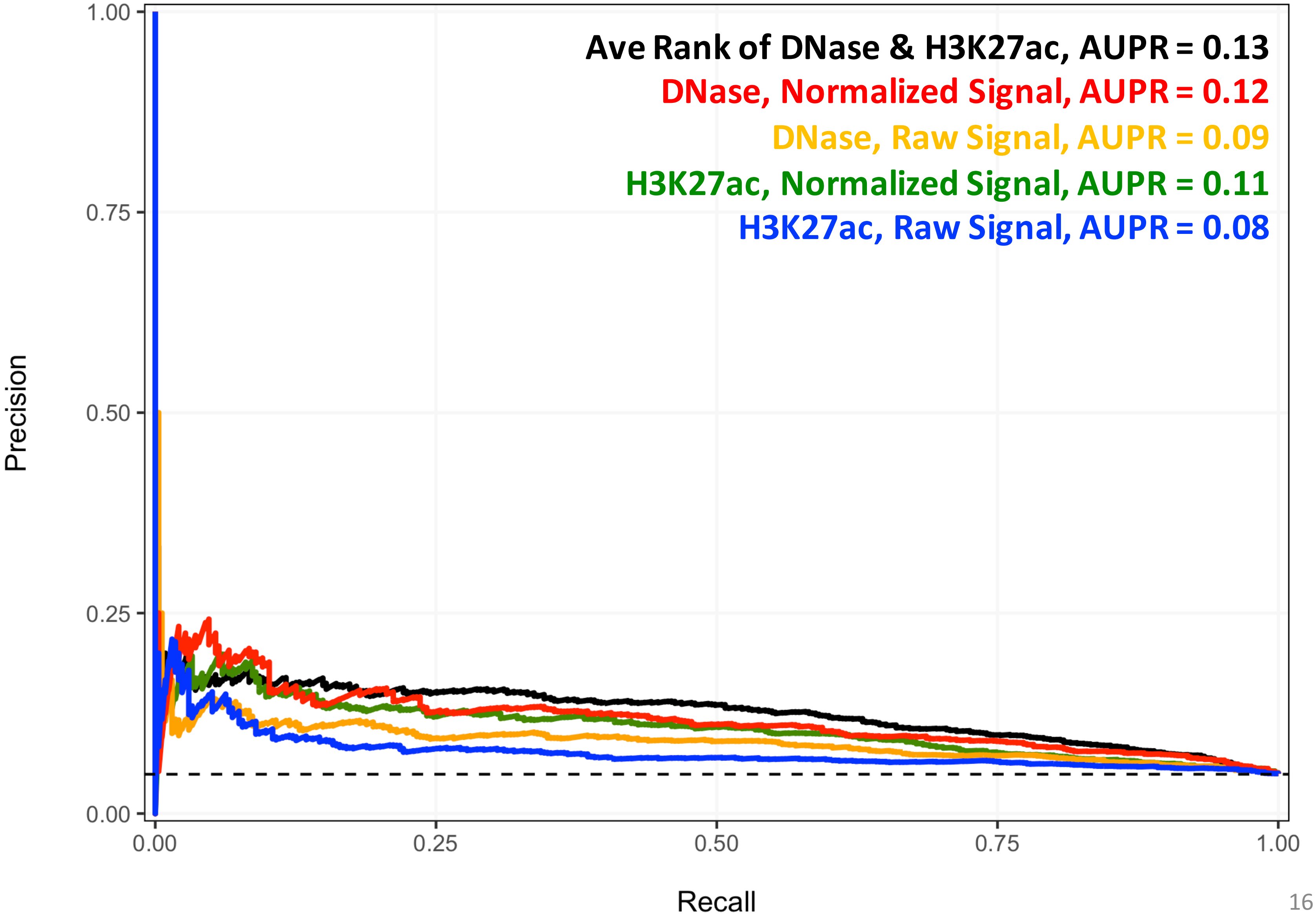
- Raw signal vs Z-score normalized signal
- DNase signal vs H3K27ac signal
- ENCODE datasets vs. Roadmap datasets
- Pearson vs Spearman correlation
- Rank by correlation coefficient vs permutation p-value[1]

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

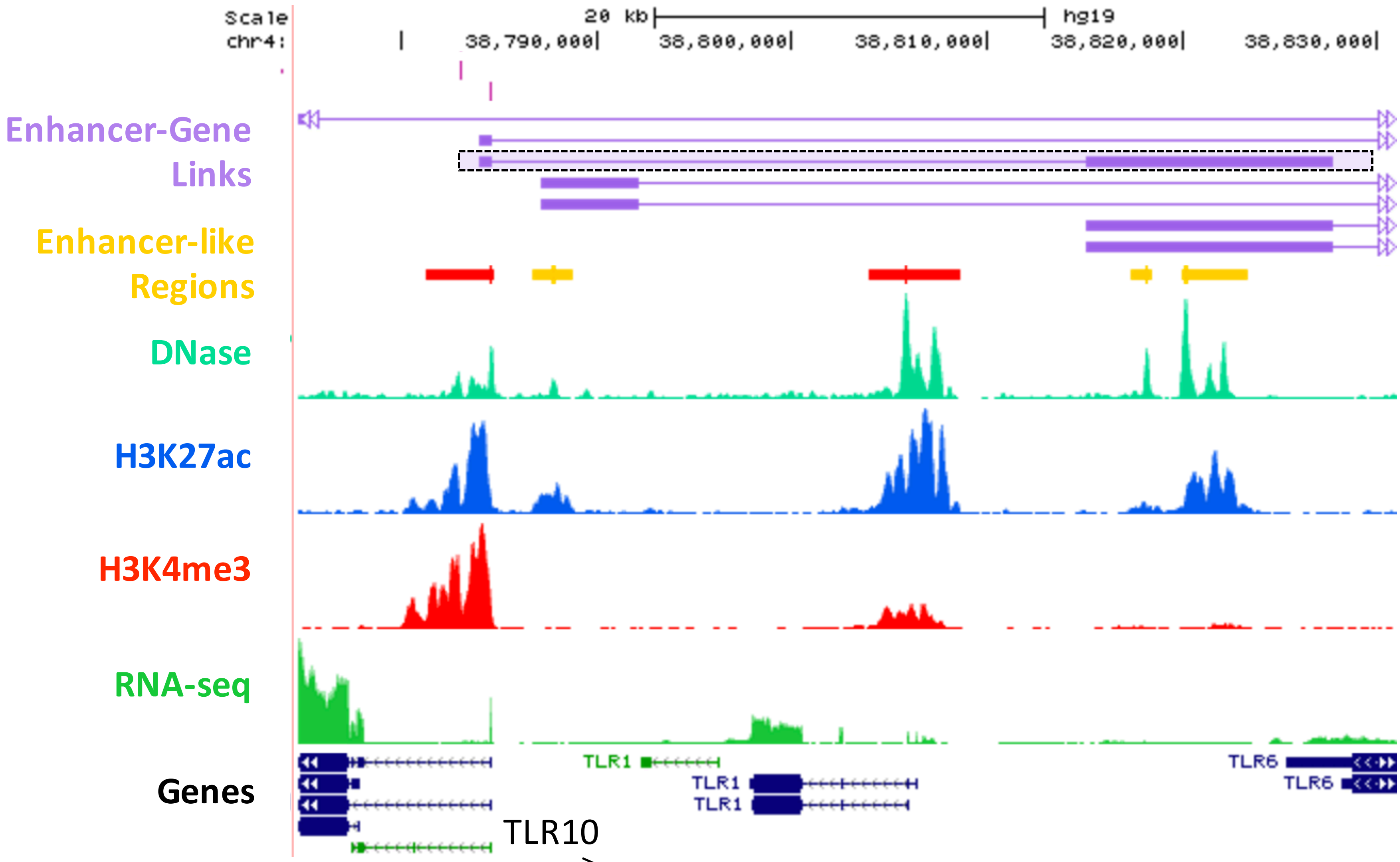
ROC - Correlation Methods



PR - Correlation Methods



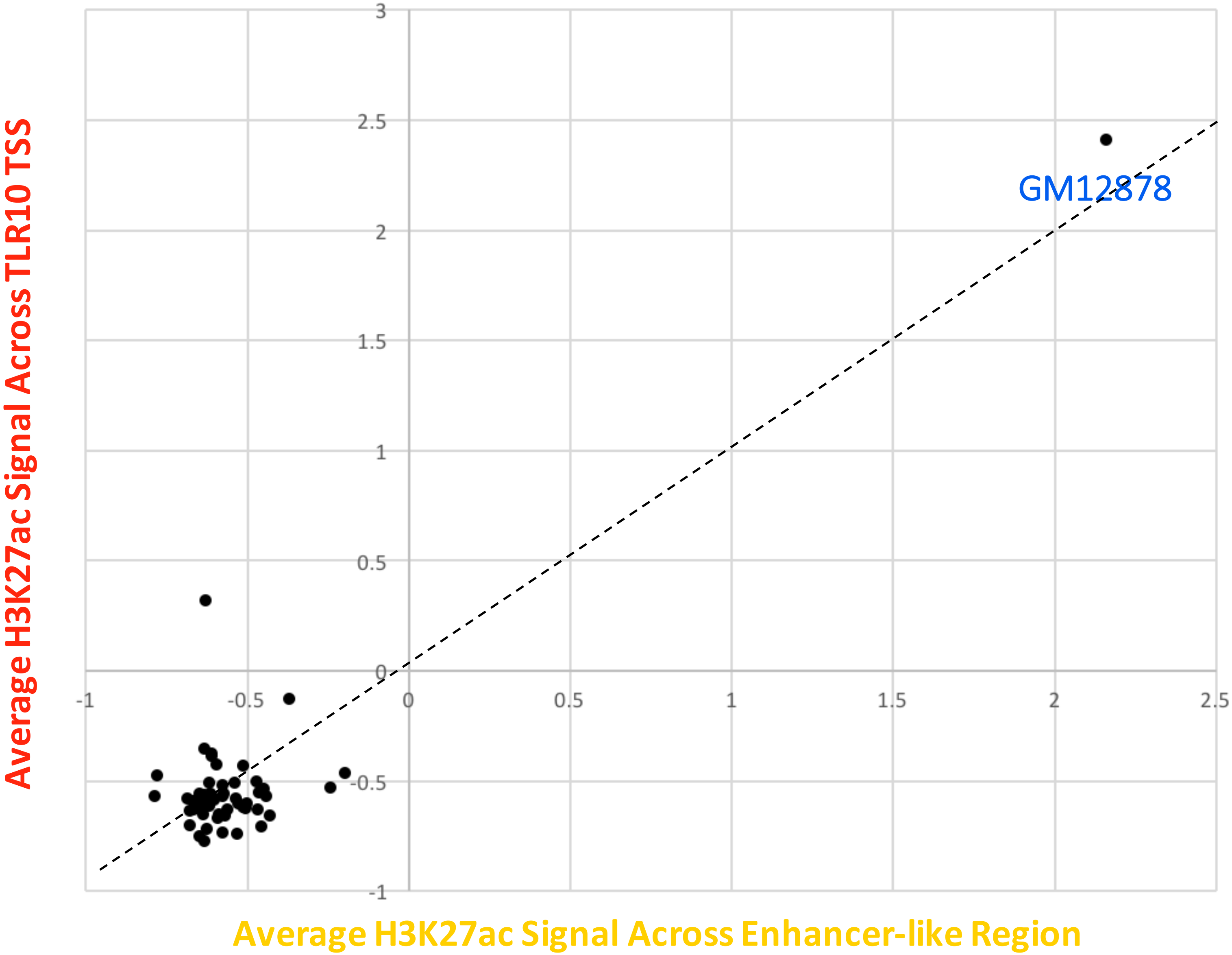
In Some Cases Correlation Accurately Predicts Links



TLR10

Important to Innate Immune System

In Some Cases Correlation Accurately Predicts Links



In Many Cases Correlation Does Not Accurately Predict Links

Enhancer-Gene Links

Enhancer-like Regions

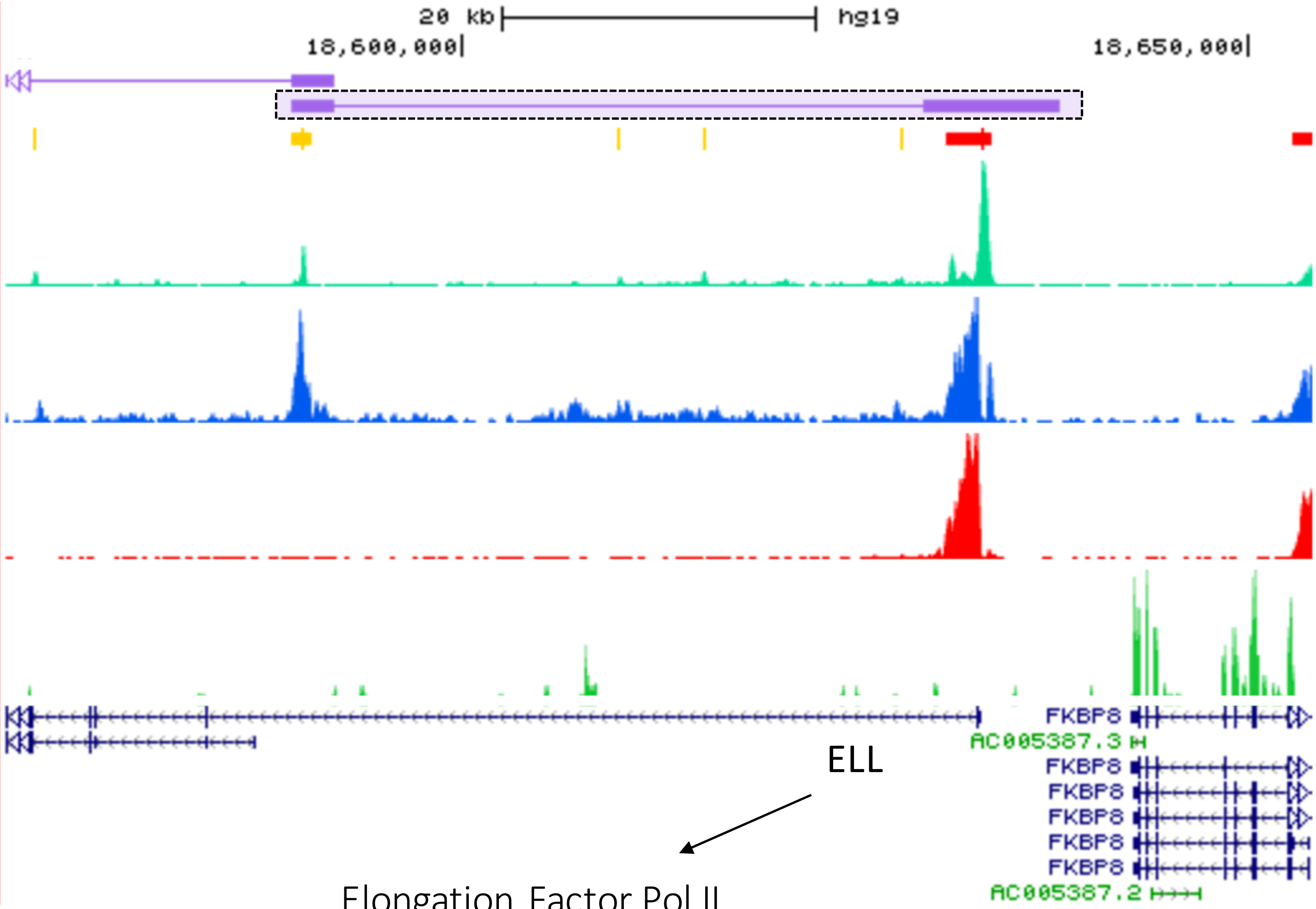
DNase

H3K27ac

H3K4me3

RNA-seq

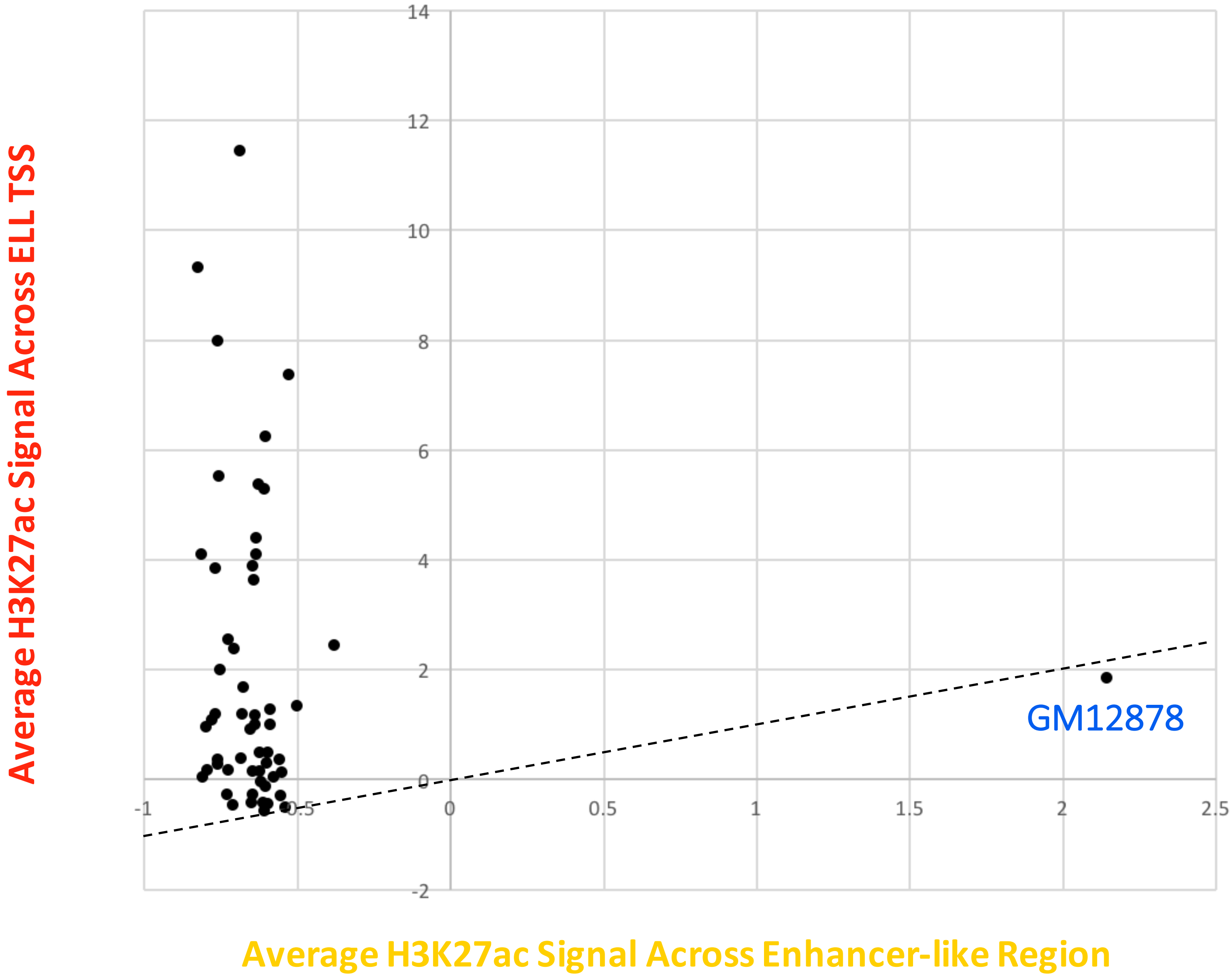
Genes



ELL

Elongation Factor Pol II

In Many Cases Correlation Does Not Accurately Predict Links



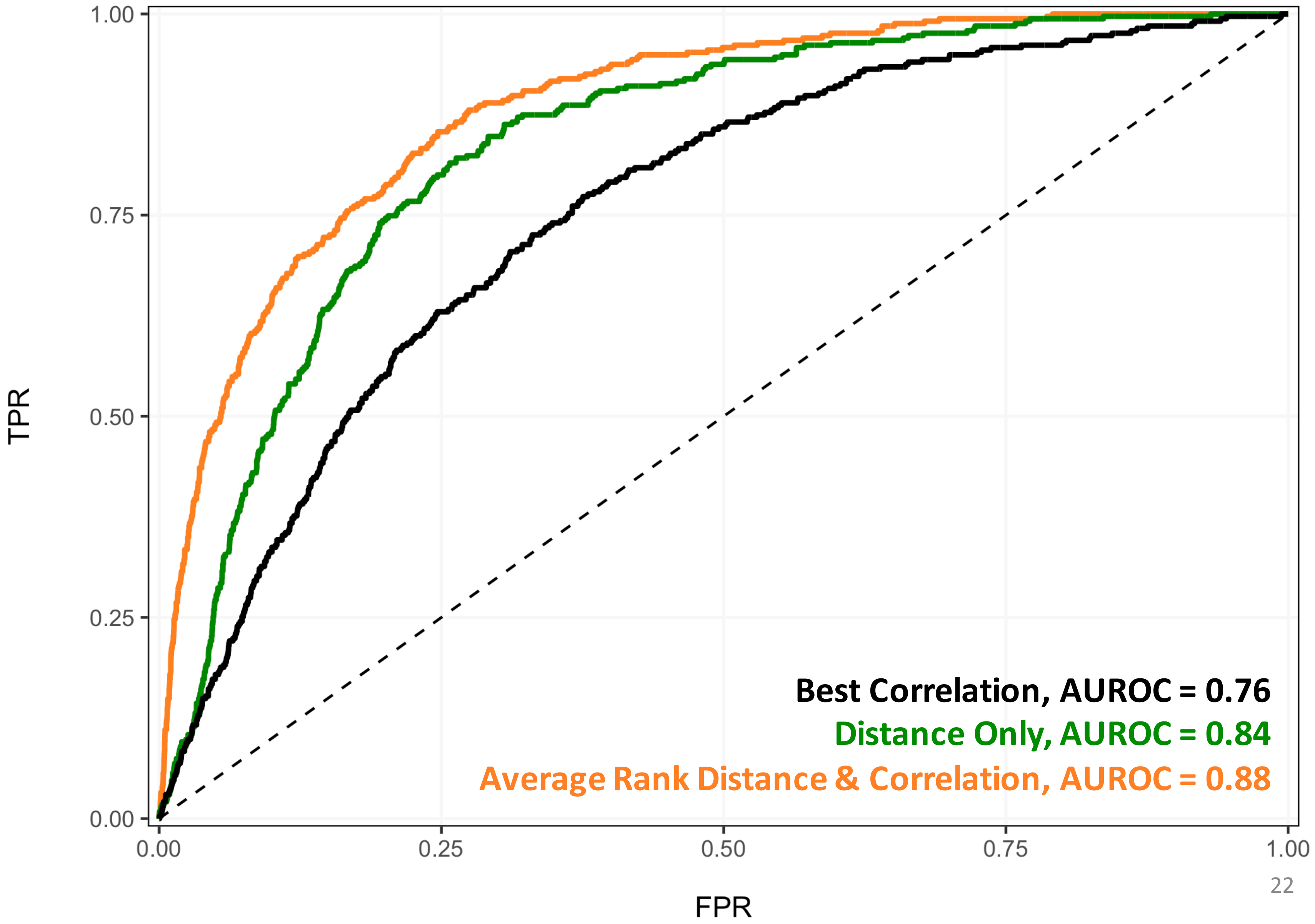
Incorporating Distance Information

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

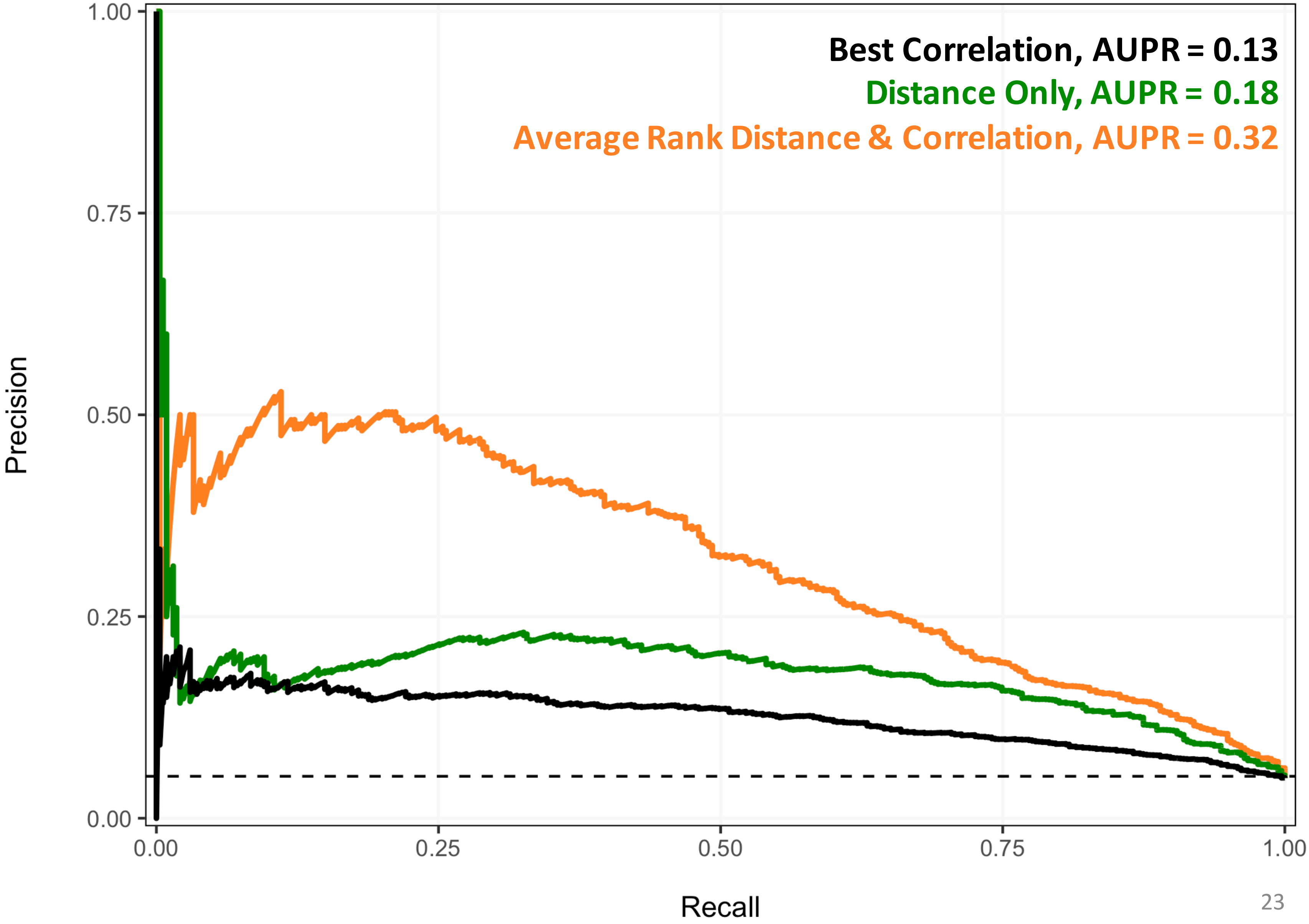
We instead tested:

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

Incorporating Distance Improves Performance



Incorporating Distance Improves Performance

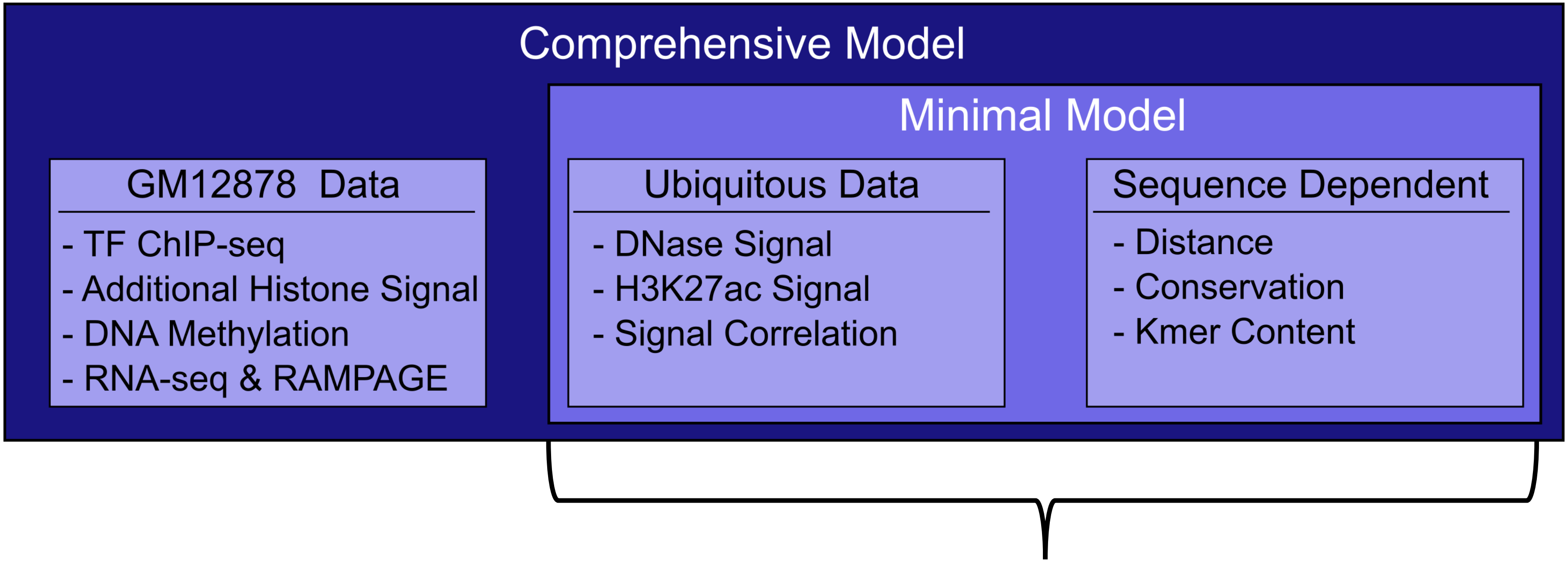


Part II: Conclusions

- 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

Developing Two Random Forest Models



Comprehensive Model

Minimal Model

GM12878 Data
- TF ChIP-seq
- Additional Histone Signal
- DNA Methylation
- RNA-seq & RAMPAGE

Ubiquitous Data
- DNase Signal
- H3K27ac Signal
- Signal Correlation

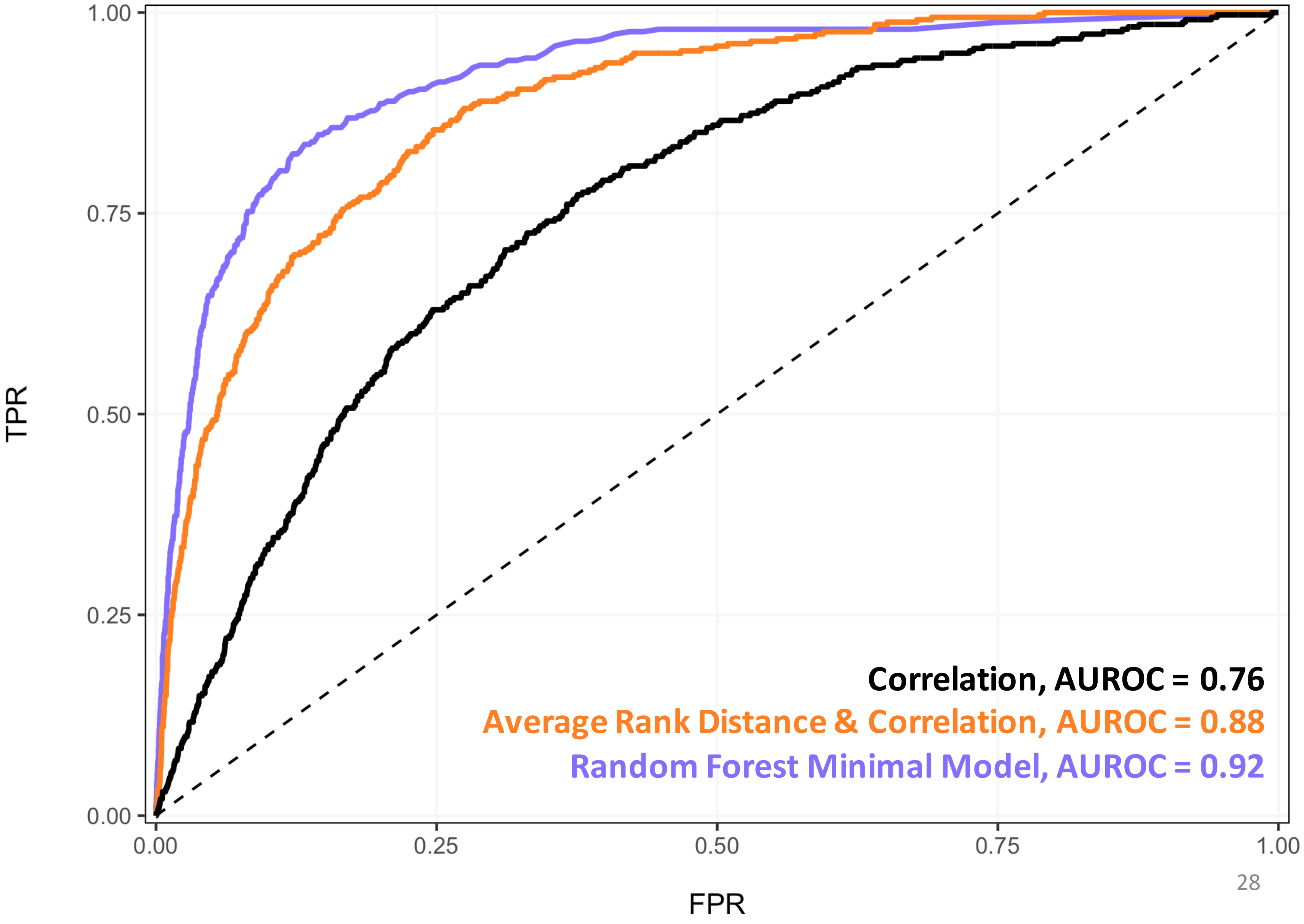
Sequence Dependent
- Distance
- Conservation
- Kmer Content

Can be applied across all cell and tissue types

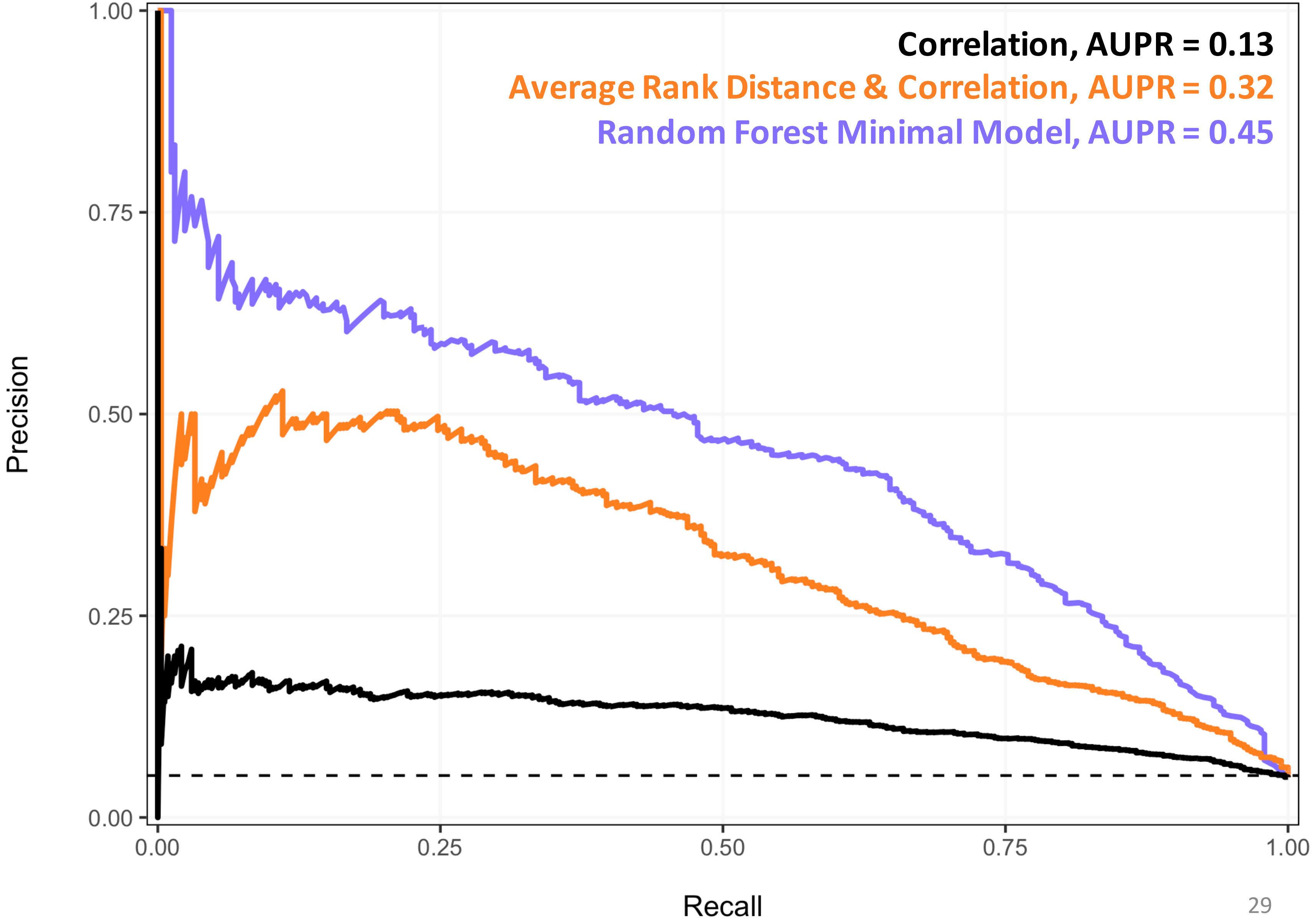
Minimal Model Features

- 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

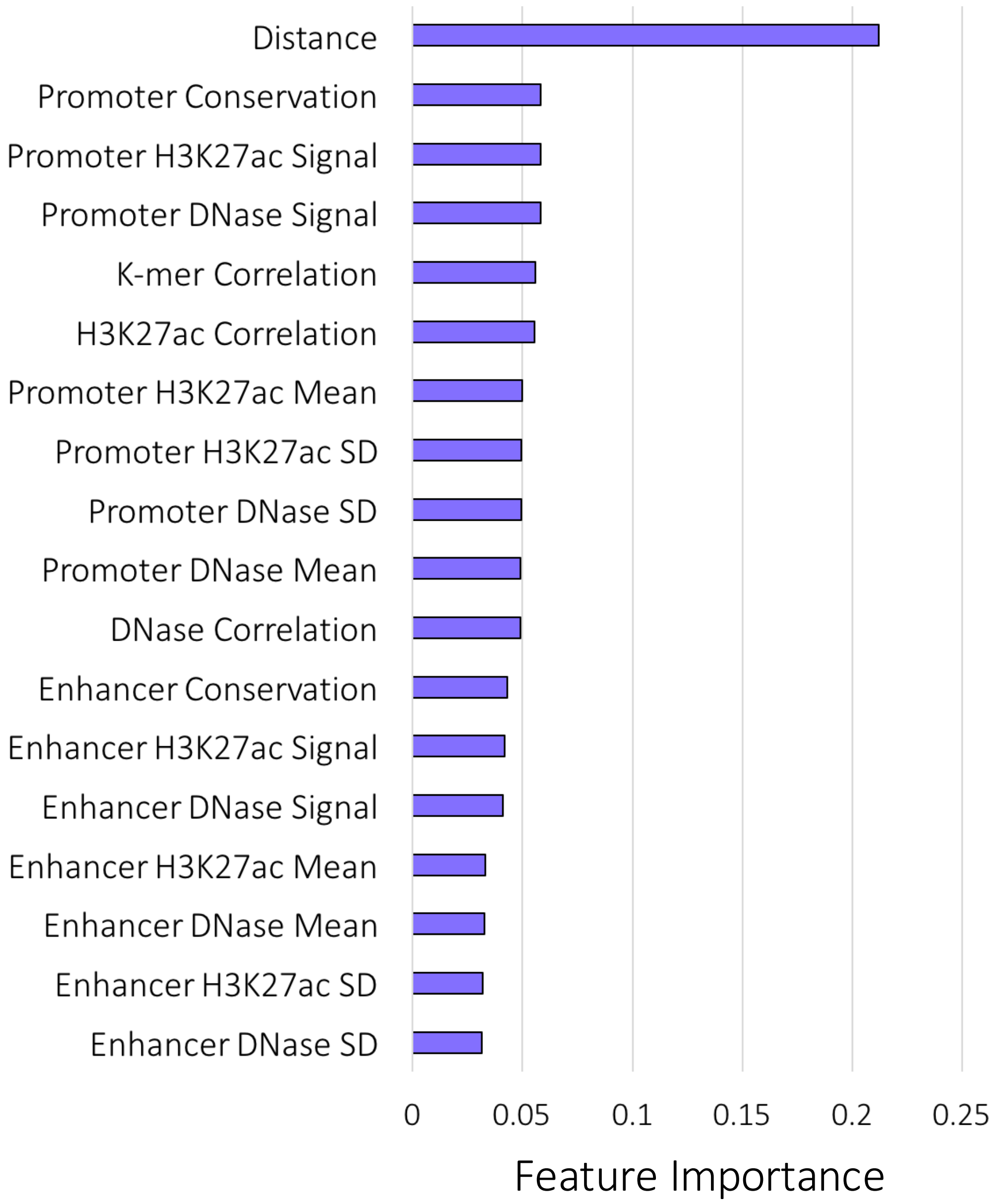
ROC – Random Forest Minimal Model



PR – Random Forest Minimal Model



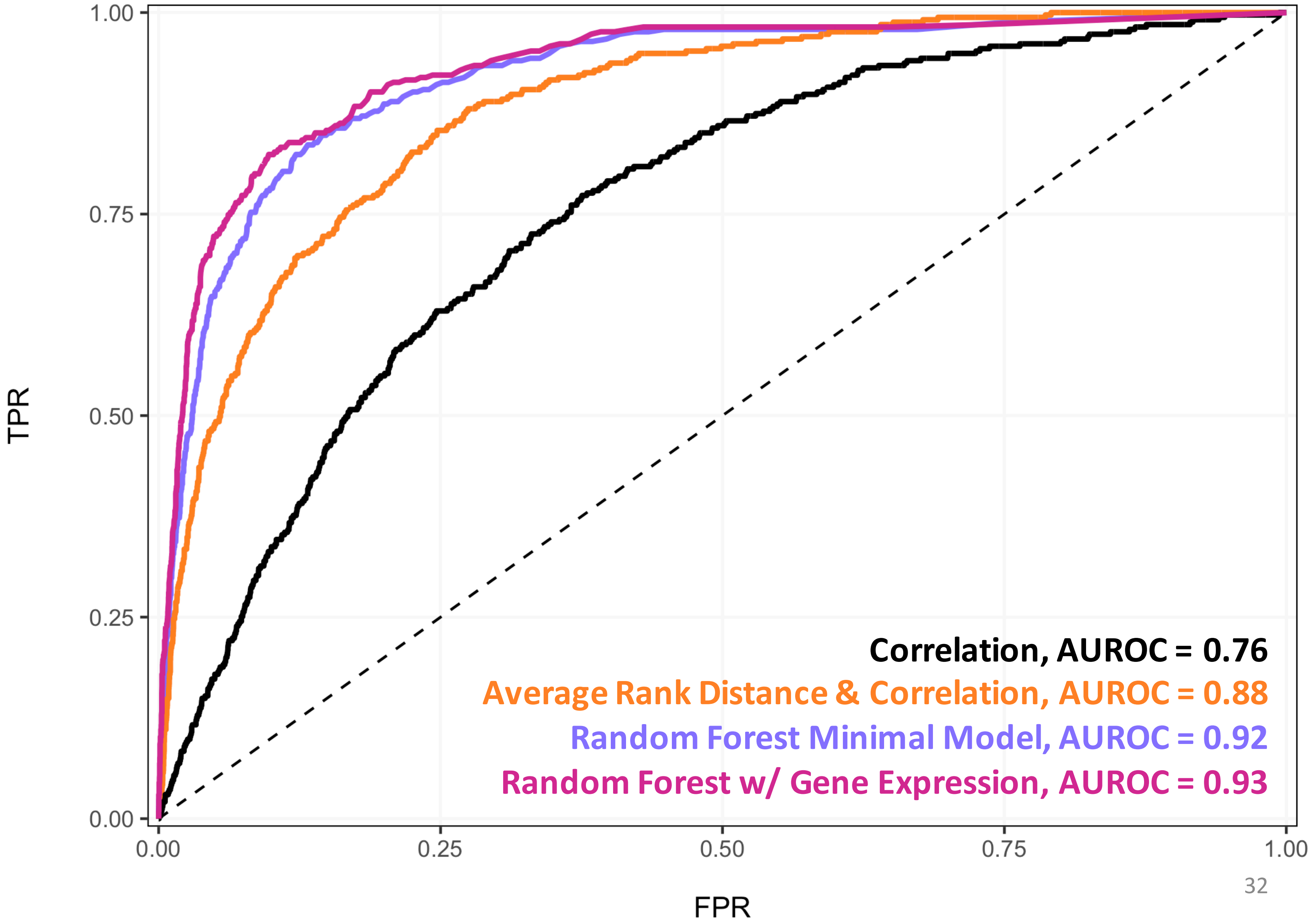
Feature Importance - Minimal Model



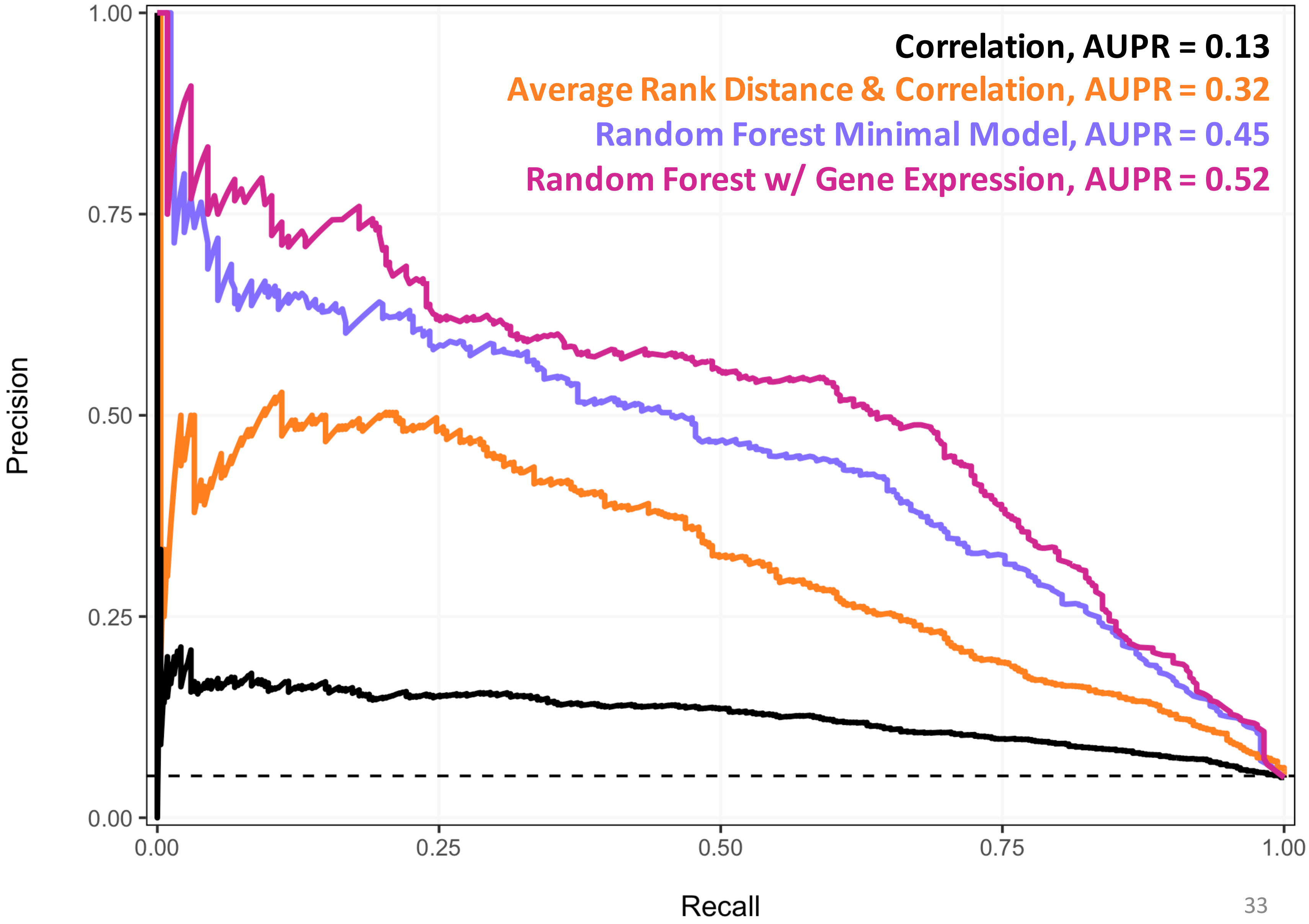
Comprehensive Model Features

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

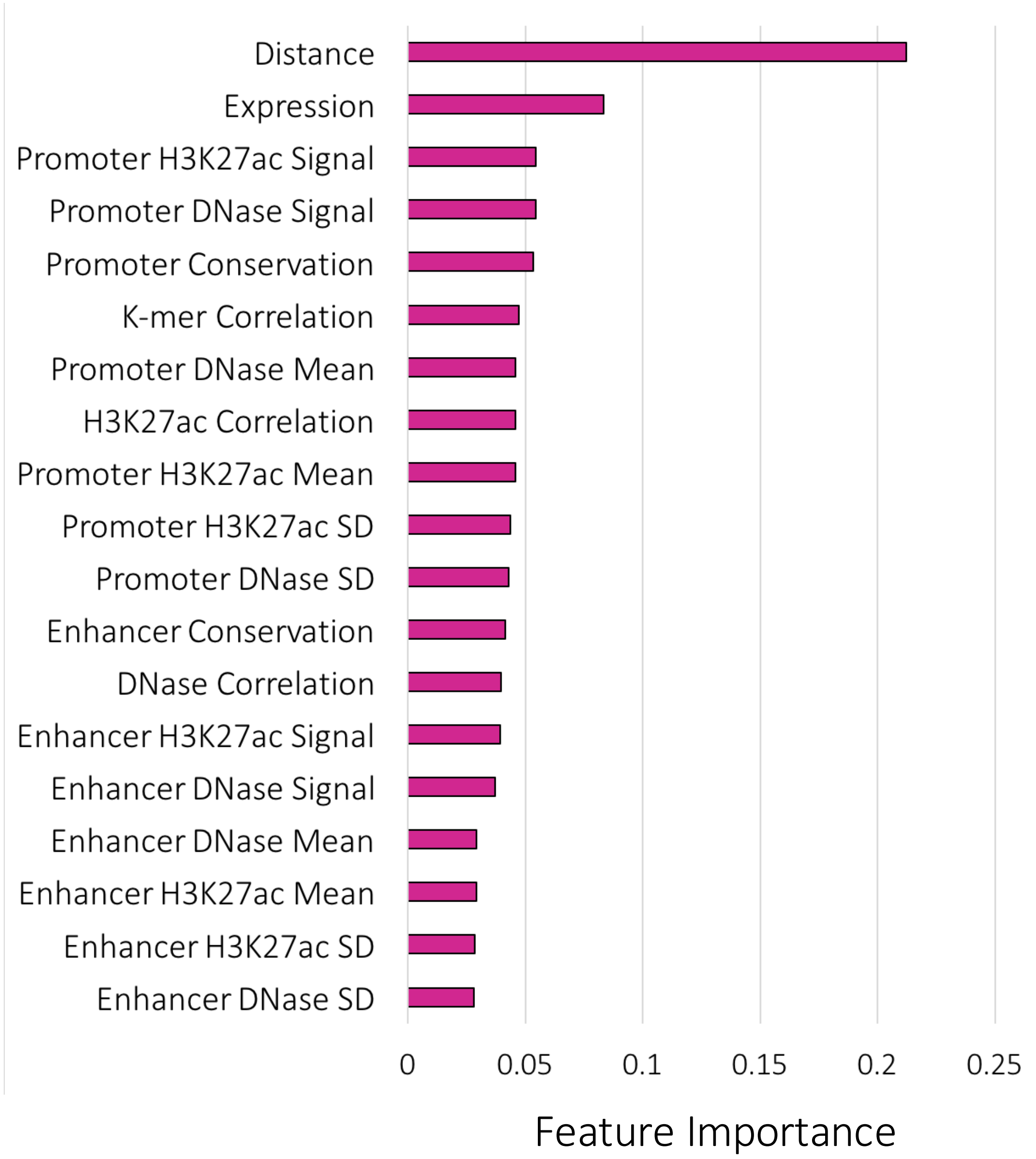
ROC – Random Forest with Gene Expression



PR – Random Forest with Gene Expression



Feature Importance – RF with Gene Expression

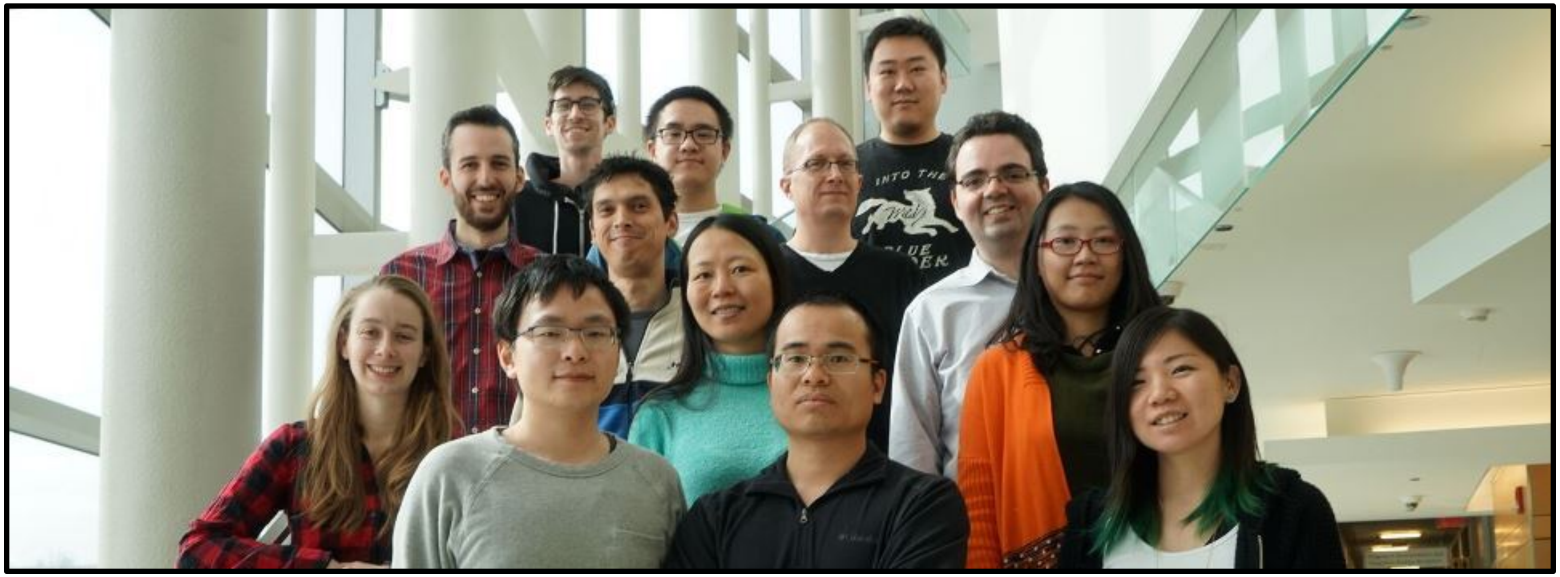


Future Directions

- Apply minimal model to all cell & tissue types in Encyclopedia
- Continue to develop comprehensive model by incorporating more data
- Input from other ENCODE groups – compare other methods

Part IV: Discussion

Acknowledgements

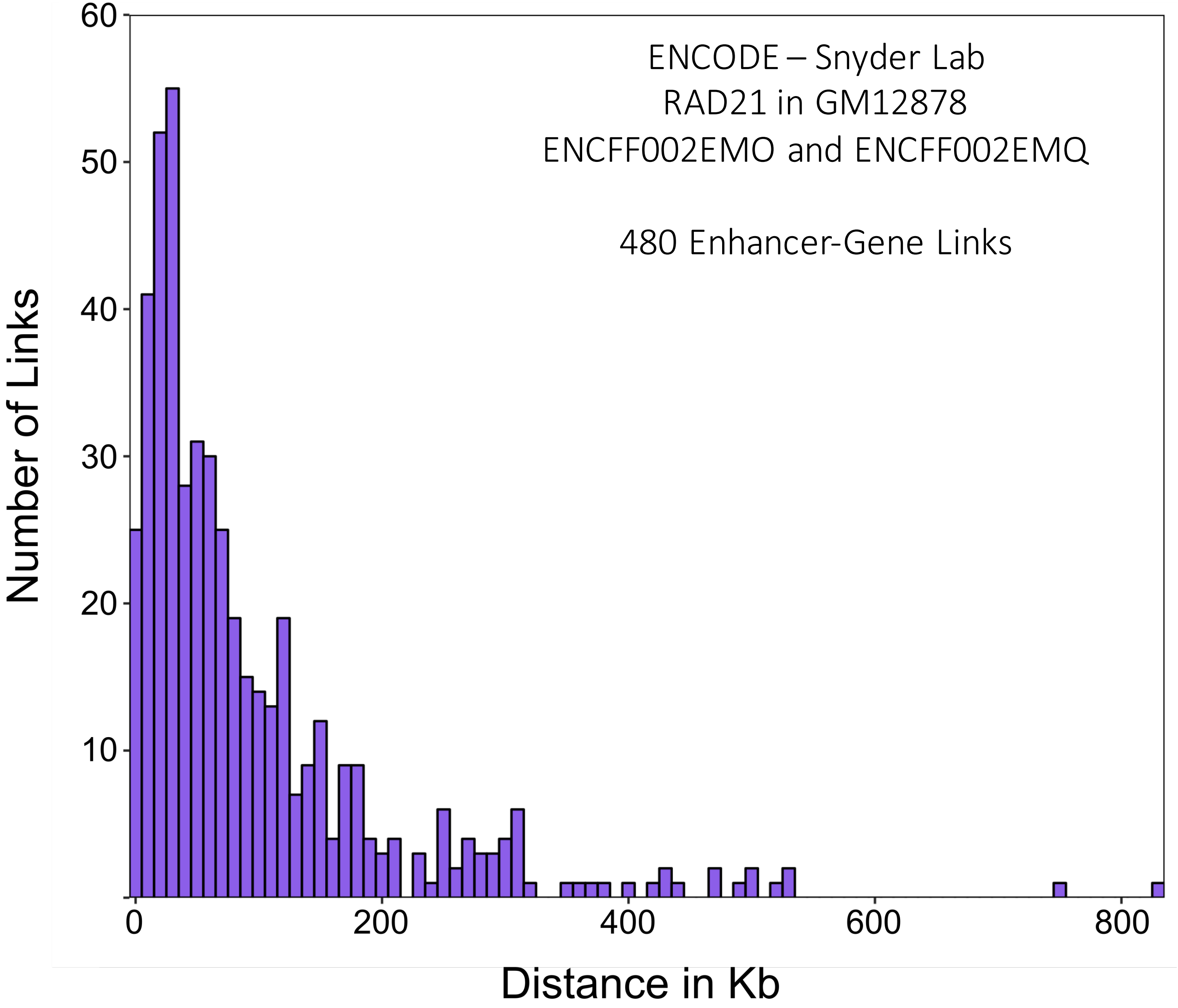


Zhiping Weng, PI
Michael Purcaro
Arjan van der Velde
Tyler Borrman
Henry Pratt
Sowmya Iyer
Jie Wang

UMASS MEDICAL SCHOOL

Supplementary Slides

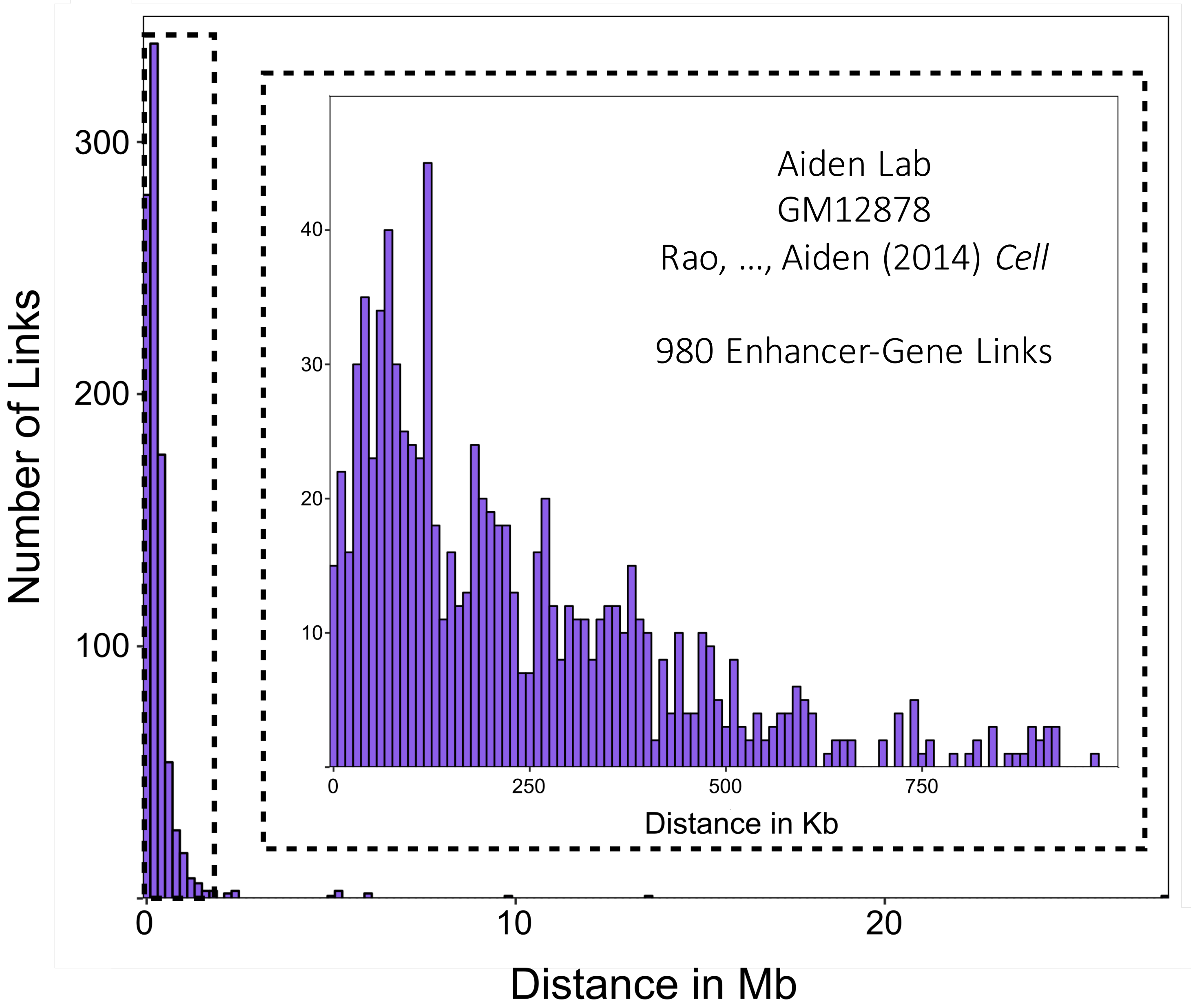
ChIA-PET Datasets Distance Distribution



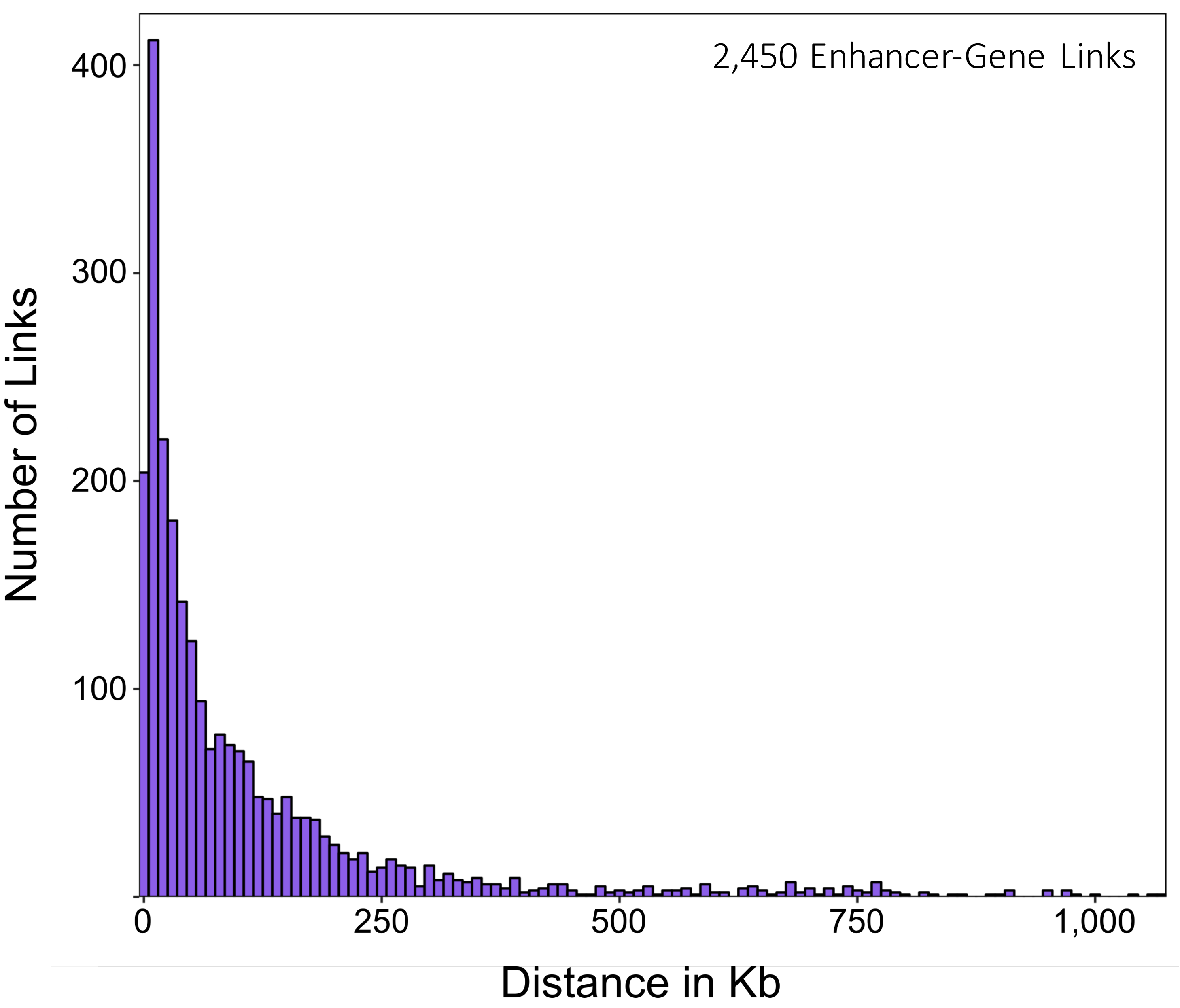
ENCODE – Snyder Lab
RAD21 in GM12878
ENCFF002EMO and ENCFF002EMQ

480 Enhancer-Gene Links

Aiden Lab Hi-C Distance Distribution



Lymphoblastoid eQTLs Distance Distribution



Normalizing Raw Signal Using Z Scores

	Cell Type 1	Cell Type2	…	Cell Type N
Peak 1	100.5	3.2	…	0
Peak 2	12.3	80.4	…	64.9
Peak 3	2.1	0	…	21.9
…	…	…	…	…
Peak M	45.3	3.1		5.4

$$z = \frac{x - \text{colMean}}{\text{colSD}}$$

	Cell Type 1	Cell Type2	…	Cell Type N
Peak 1	2.0	-0.6	…	-2.0
Peak 2	-2.3	7.0	…	0.6
Peak 3	-2.8	-1.0	…	-1.1
…	…	…	…	…
Peak M	-0.7	-0.7		-1.7

Correlation Results

AUROC	ENCODE Pearson	Roadmap Pearson	ENCODE Spearman	Roadmap Spearman
DNase-Norm	0.7320	0.7148	0.7192	0.7095
DNase-Raw	0.6700	0.6877	0.6534	0.6847
H3K27ac-Norm	0.7015	0.7187	0.6940	0.7008
H3K27ac-Raw	0.6176	0.6971	0.6145	0.6739
Average Rank-Norm	0.7556	0.7459	0.7441	0.7310
Average Rank-Raw	0.6750	0.7188	0.6602	0.7014

AURPR	ENCODE Pearson	Roadmap Pearson	ENCODE Spearman	Roadmap Spearman
DNase-Norm	0.1158	0.1047	0.1051	0.1043
DNase-Raw	0.0890	0.1002	0.0926	0.0947
H3K27ac-Norm	0.1059	0.1164	0.1009	0.1021
H3K27ac-Raw	0.0763	0.1018	0.0696	0.0938
Average Rank-Norm	0.1252	0.1219	0.1168	0.1137
Average Rank-Raw	0.0937	0.1111	0.0909	0.1020