



Predicting Target Genes: Analysis of Baseline Methods

Jill E. Moore

Weng Lab

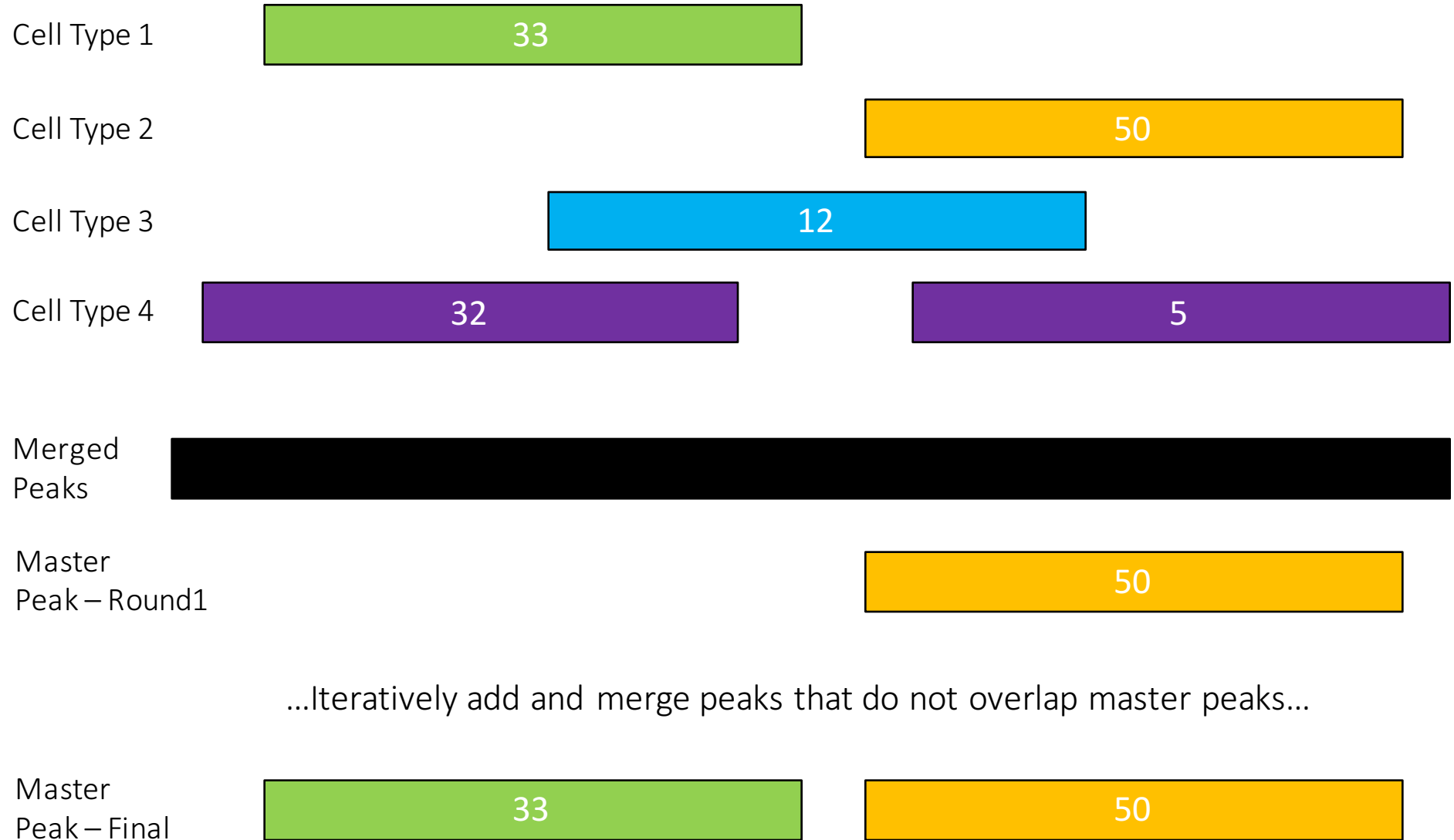
University of Massachusetts Medical School

December 2015

Current State of ENCODE Encyclopedia

- Candidate enhancers and promoters for DNase hypersensitivity, annotated with histone marks H3K27ac and H3K4me1 which are enriched at enhancers, H3K4me3 which is enriched at promoters, H3K9ac which is enriched at both enhancers and promoters, as well as ChIP peaks of transcription factors. Out of 177 cell types with DNase-seq data, we annotated 45 cell types with H3K27ac, 48 cell types with H3K4me1, 94 cell types with H3K4me3, and 27 cell types with H3K9ac in a cell type specific manner. [[Download methods](#)]
 - Distal DNase peaks [[Download](#)]
 - Proximal DNase peaks [[Download](#)]
 - Distal H3K27ac annotations (cell type specific) [[Download](#)]
 - Distal H3K4me1 annotations (cell type specific) [[Download](#)]
 - Distal H3K9 annotations (cell type specific) [[Download](#)]
 - Proximal H3K4me3 annotations (cell type specific) [[Download](#)]
 - Proximal H3K9ac annotations (cell type specific) [[Download](#)]
 - Distal TF binding sites [[Download](#)]
 - Proximal TF binding sites [[Download](#)]
- Gene expression over ~60 cell types with genes annotated by GENCODE 19 [[Query tool at Penn State](#) | [Visualize data](#) | [Download data](#) | [Download methods](#)]
- Transcription start site (TSS) lists [[View README](#)]
 - GENCODE v19 TSS [[Download](#)]
 - GENCODE v19 TSS stratified by strict Fantom5 CAGE clusters [[Download](#)]
 - GENCODE v19 TSS stratified by robust Fantom5 CAGE clusters [[Download](#)]
 - GENCODE v19 TSS stratified by permissive Fantom5 CAGE clusters [[Download](#)]

Creating DNase Master Peaks – Stam Lab Pipeline



Current State of ENCODE Encyclopedia

- Candidate enhancers and promoters for DNase hypersensitivity, annotated with histone marks H3K27ac and H3K4me1 which are enriched at enhancers, H3K4me3 which is enriched at promoters, H3K9ac which is enriched at both enhancers and promoters, as well as ChIP peaks of transcription factors. Out of 177 cell types with DNase-seq data, we annotated 45 cell types with H3K27ac, 48 cell types with H3K4me1, 94 cell types with H3K4me3, and 27 cell types with H3K9ac in a cell type specific manner. [[Download methods](#)]

- Distal DNase peaks [[Download](#)]
- Proximal DNase peaks [[Download](#)]
- Distal H3K27ac annotations (cell type specific) [[Download](#)]
- Distal H3K4me1 annotations (cell type specific) [[Download](#)]
- Distal H3K9 annotations (cell type specific) [[Download](#)]
- Proximal H3K4me3 annotations (cell type specific) [[Download](#)]
- Proximal H3K9ac annotations (cell type specific) [[Download](#)]
- Distal TF binding sites [[Download](#)]
- Proximal TF binding sites [[Download](#)]

How do we link regulatory elements to genes?

- Gene expression over ~60 cell types with genes annotated by GENCODE 19 [[Query tool at Penn State](#) | [Visualize data](#) | [Download data](#) | [Download methods](#)]
- Transcription start site (TSS) lists [[View README](#)]
 - GENCODE v19 TSS [[Download](#)]
 - GENCODE v19 TSS stratified by strict Fantom5 CAGE clusters [[Download](#)]
 - GENCODE v19 TSS stratified by robust Fantom5 CAGE clusters [[Download](#)]
 - GENCODE v19 TSS stratified by permissive Fantom5 CAGE clusters [[Download](#)]

Methods for Predicting Target Genes of Distal Regulatory Elements

- Correlation of genomic data

Systematic Localization of Common Disease-Associated Variation in Regulatory DNA

Matthew T. Maurano,^{1*} Richard Humbert,^{1*} Eric Rynes,^{1*} Robert E. Thurman,¹ Eric Haugen,¹ Hao Wang,¹ Alex P. Reynolds,¹ Richard Sandstrom,¹ Hongzhu Qu,^{1,2} Jennifer Brody,³ Anthony Shafer,¹ Fidencio Neri,¹ Kristen Lee,¹ Tanya Kutyaivin,¹ Sandra Stehling-Sun,¹ Audra K. Johnson,¹ Theresa K. Canfield,¹ Erika Giste,¹ Morgan Diegel,¹ Daniel Bates,¹ R. Scott Hansen,⁴ Shane Neph,¹ Peter J. Sabo,¹ Shelly Heimfeld,⁵ Antony Raubitschek,⁶ Steven Ziegler,⁶ Chris Cotsapas,^{7,8} Nona Sotoodehnia,^{3,9} Ian Glass,¹⁰ Shamil R. Sunyaev,¹¹ Rajinder Kaul,⁴ John A. Stamatoyannopoulos^{1,12†}

ARTICLE

doi:10.1038/nature11232

The accessible chromatin landscape of the human genome

Robert E. Thurman^{1*}, Eric Rynes^{1*}, Richard Humbert^{1*}, Jeff Vierstra¹, Matthew T. Maurano¹, Eric Haugen¹, Nathan C. Sheffield², Andrew B. Stergachis¹, Hao Wang¹, Benjamin Vernot¹, Kavita Garg³, Sam John¹, Richard Sandstrom¹, Daniel Bates¹, Lisa Boatman⁴, Theresa K. Canfield¹, Morgan Diegel¹, Douglas Dunn¹, Abigail K. Ebersol⁴, Tristan Frum⁴, Erika Giste¹, Audra K. Johnson¹, Ericka M. Johnson⁴, Tanya Kutyaivin¹, Bryan Lajoie⁵, Bum-Kyu Lee⁶, Kristen Lee¹, Darin London², Dimitra Lotakis⁴, Shane Neph¹, Fidencio Neri¹, Eric D. Nguyen⁴, Hongzhu Qu^{1,7}, Alex P. Reynolds¹, Vaughn Roach¹, Alexias Safi¹, Minerva E. Sanchez⁴, Amartya Sanyal², Anthony Shafer¹, Jeremy M. Simon⁸, Lingyun Song², Shinny Vong¹, Molly Weaver¹, Yongqi Yan⁴, Zhancheng Zhang⁸, Zhuzhu Zhang⁸, Boris Lenhard^{9†}, Muneesh Tewari³, Michael O. Dorschner¹⁰, R. Scott Hansen⁴, Patrick A. Navas⁴, George Stamatoyannopoulos⁴, Vishwanath R. Iyer⁶, Jason D. Lieb⁸, Shamil R. Sunyaev¹¹, Joshua M. Akey¹, Peter J. Sabo¹, Rajinder Kaul⁴, Terrence S. Furey⁸, Job Dekker⁵, Gregory E. Crawford⁵ & John A. Stamatoyannopoulos^{1,12}

Methods for Predicting Target Genes of Distal Regulatory Elements

- 3D Chromatin Assays: ChIA-PET and 3C/4C/5C/Hi-C

LETTER

doi:10.1038/nature12716

Chromatin connectivity maps reveal dynamic promoter–enhancer long–range associations

Yubo Zhang^{1*†}, Chee-Hong Wong^{1*}, Ramon Y. Birnbaum^{2**†}, Guoliang Li^{3,4}, Rebecca Favaro⁵, Chew Yee Ngan¹, Joanne Lim⁴, Eunice Tai⁴, Huay Mei Poh⁴, Eleanor Wong⁴, Fabianus Hendriyan Mulawadi⁴, Wing-Kin Sung⁴, Silvia Nicolis⁵, Nadav Ahituv², Yijun Ruan³ & Chia-Lin Wei^{1,4}

Published online 3 September 2013

Nucleic Acids Research, 2013, Vol. 41, No. 22 **10391–10402**
doi:10.1093/nar/gkt785

Combining Hi-C data with phylogenetic correlation to predict the target genes of distal regulatory elements in human genome

Yulan Lu, Yuanpeng Zhou and Weidong Tian*

State Key Laboratory of Genetic Engineering, Department of Biostatistics and Computational Biology, School of Life Science, Fudan University, Shanghai 200433, China

Received May 14, 2013; Revised July 25, 2013; Accepted August 11, 2013

Methods for Predicting Target Genes of Distal Regulatory Elements

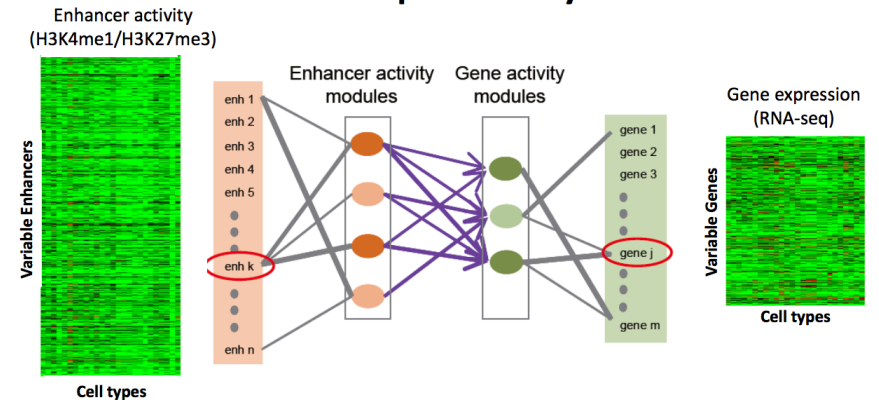
- Machine learning methods

Learning three-dimensional regulation of gene expression

Jianrong Wang
Anshul Kundaje
Manolis Kellis

AWG Presentation: 4/10/15

A novel prob. model for enhancer-gene linking using chromatin-expression dynamics



Joint learning of mixed-membership probabilistic model

- Mixed membership gene modules (which genes active in which cell types)
- Mixed membership enhancer modules (which enhancers active in which cell types)
- Prob. non-linear linking of Gene module to enhancer module
- Cell-type specific enhancer to gene linking

Methods for Predicting Target Genes of Distal Regulatory Elements

- Machine learning methods

Massive data integration enables discovery of gene regulatory enhancers and their targets

Katie Pollard

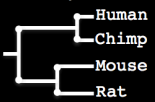
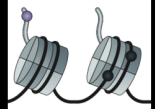
**Gladstone Institutes, Institute for Human Genetics,
Division of Biostatistics - UCSF**

AWG Presentation: 9/12/14

TargetFinder: Training

Teach a machine learning algorithm to discriminate true versus false enhancer-promoter interactions based on their features.

<p>Training Data</p> <p>Active enhancer expressed gene Positives = 5C + Negatives = 5C -</p>	<p style="text-align: center;">Computational Algorithm</p> <p>Decision trees: good for interacting features</p> <p>Ensemble learning: build many imperfect classifiers and combine them to improve prediction accuracy</p>
---	---

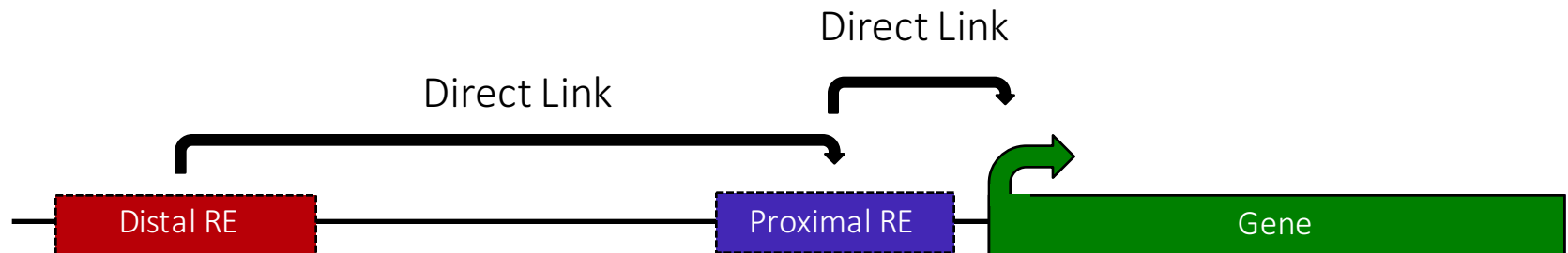
<p>Features</p> <p style="text-align: center;">Evolutionary Conservation</p>  <p style="text-align: center;">Conserved synteny of enhancer and promoter</p>	<p style="text-align: center;">Functional Genomics</p>  <p>ChIA-PET RNA-seq ChIP-seq (TFs, histones) enhancer/promoter/window ChromHMM & Segway</p>	<p style="text-align: center;">Sequence Annotations</p> <p>AAAA, AAAC, AAAG, AAAT, AACA, AACG, AAGC, AACT, AAGA, AAGC, AAGG, AAGT, AATA, AATC, AATG, AATT, ACAA, ACAC, ACAG, ACAT, ...</p> <p>K-mer correlation Annotated functions and pathways of gene and enhancer bound TFs</p>
---	--	---

Goal: Predict Target Genes for Distal Regulatory Elements

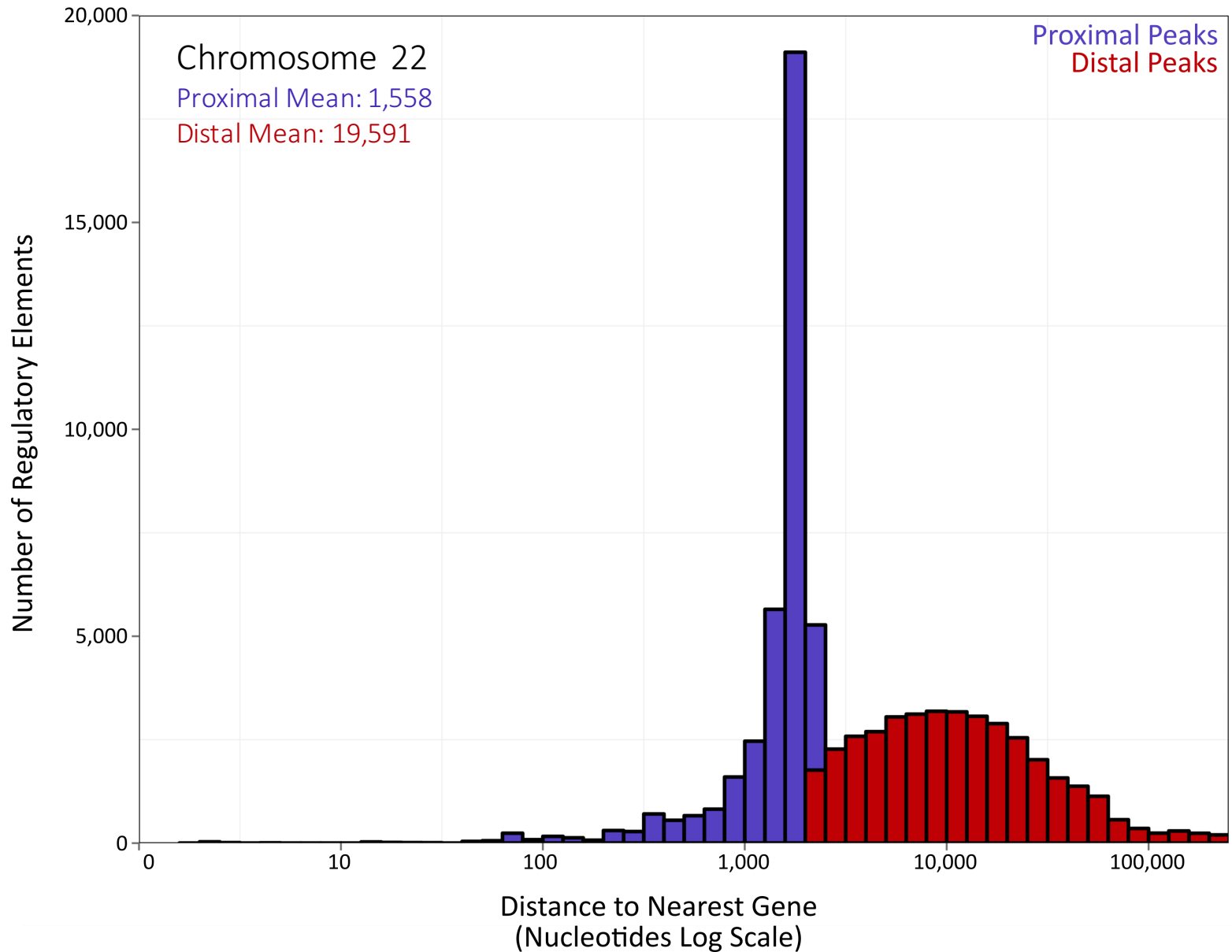
- Approach #1 – Nearest Gene
- Approach #2 – Correlation of genomic & epigenomic data
- Approach #3 – Machine learning methods

Approach #1 – Nearest Gene

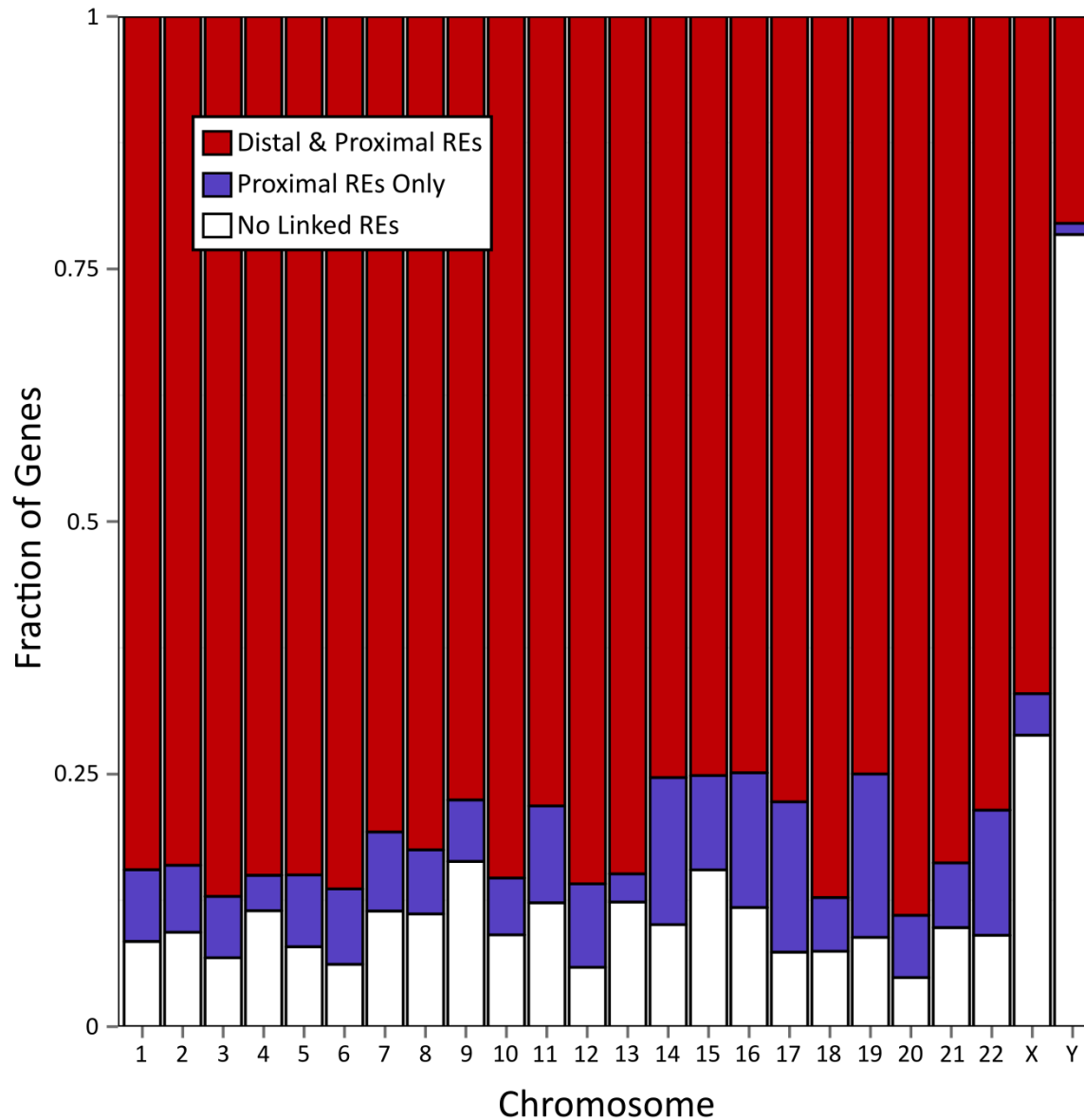
1. Assigned proximal REs to genes using Gencode V19 TSSs
2. Linked distal REs to nearest proximal RE
3. Links were calculated using center point of REs



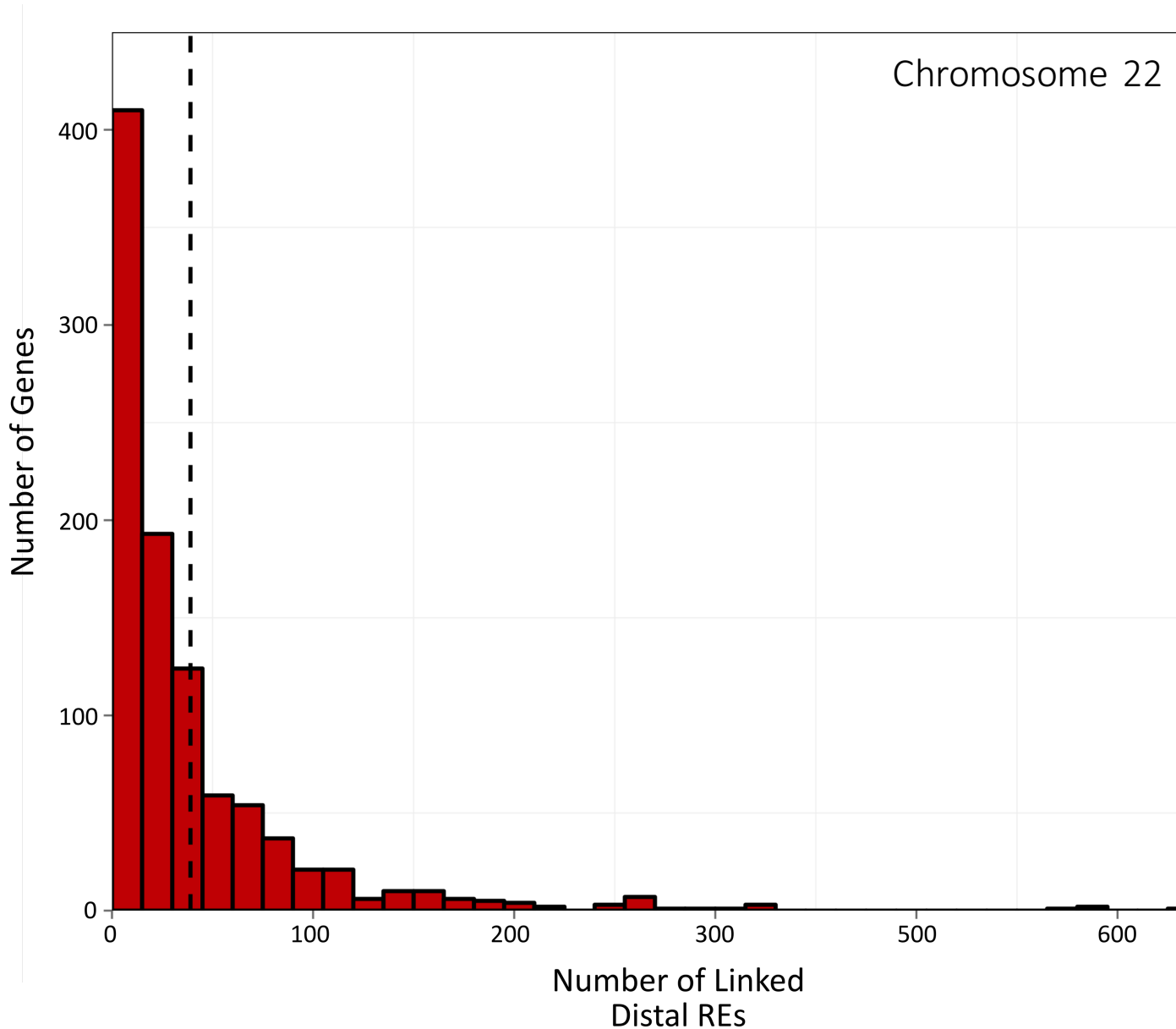
Distance to Nearest Gene



Most Genes are Linked with a Distal RE

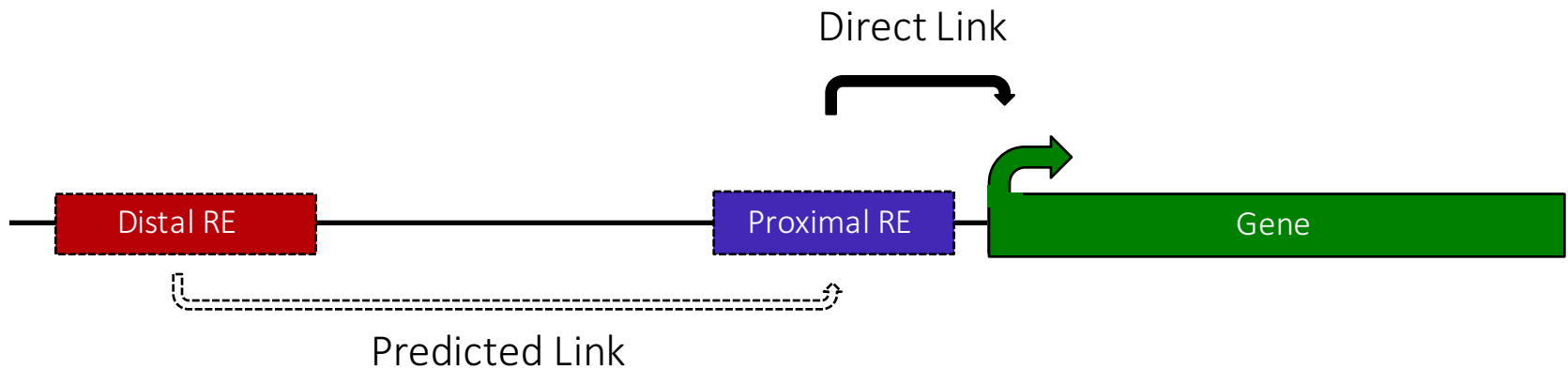


Average of 39 Linked Distal REs Per Gene



Approach #2 – DNase Signal Correlation

1. Assigned proximal REs to genes using Gencode V19 TSSs
2. Linked distal REs to proximal REs using DNase I signal correlation



Cell Type	Distal Signal	Proximal Signal
GM12878	120.1	99.4
K562	3.4	2.6
HepG2	50.8	60.3
...

Datasets

	ENCODE2	ENCODE3- October Freeze	Roadmap	Total
DNase (Raw Signal)	261	35	yes	296+
DNase (Fold Change)	0	0	53	53
H3K27ac (Raw Signal)	23	0	yes	23+
H3K27ac (Fold Change)	0	15	98	113

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

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

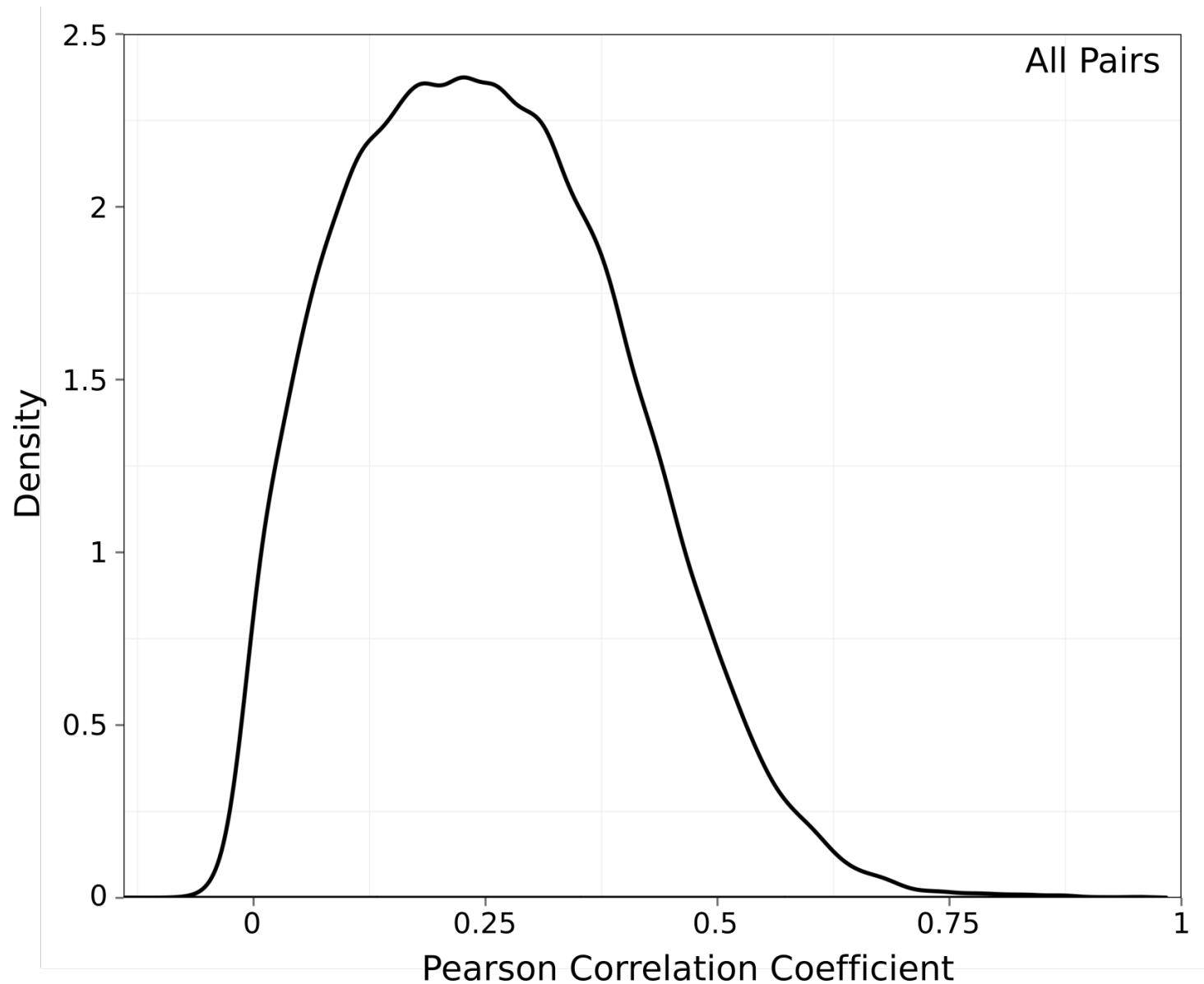
	Cell Type 1	Cell Type2	...	Cell Type N
Peak 1	2.0	-0.6	...	-2.0
Peak 2	-2.3	7.0	...	0.6
Peak 3	-2.8	-1.0	...	-1.1
...
Peak M	-0.7	-0.7		-1.7

DNase Results

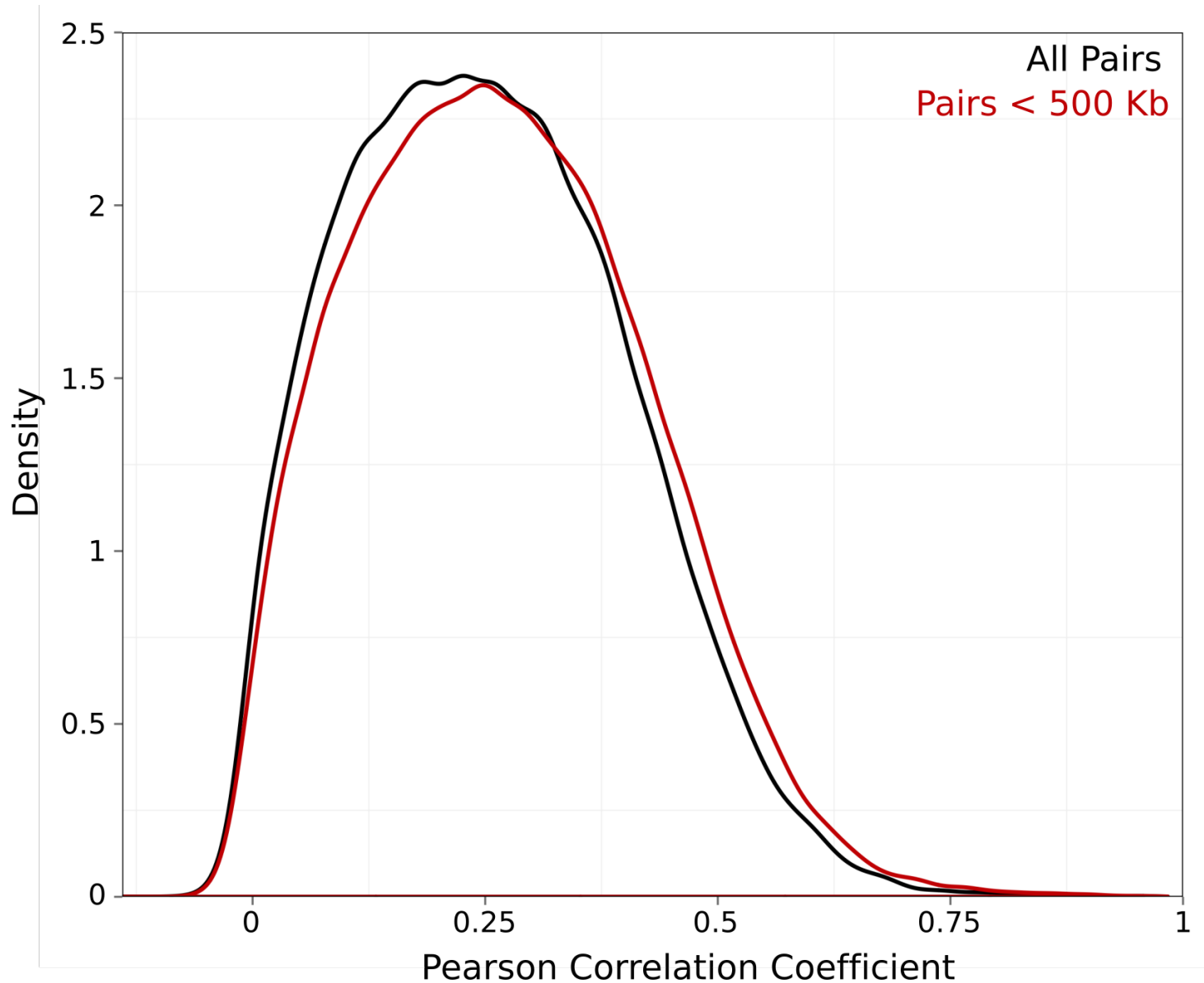
Correlation With Raw Signal Data

All initial analyses were conducted on chromosome 22

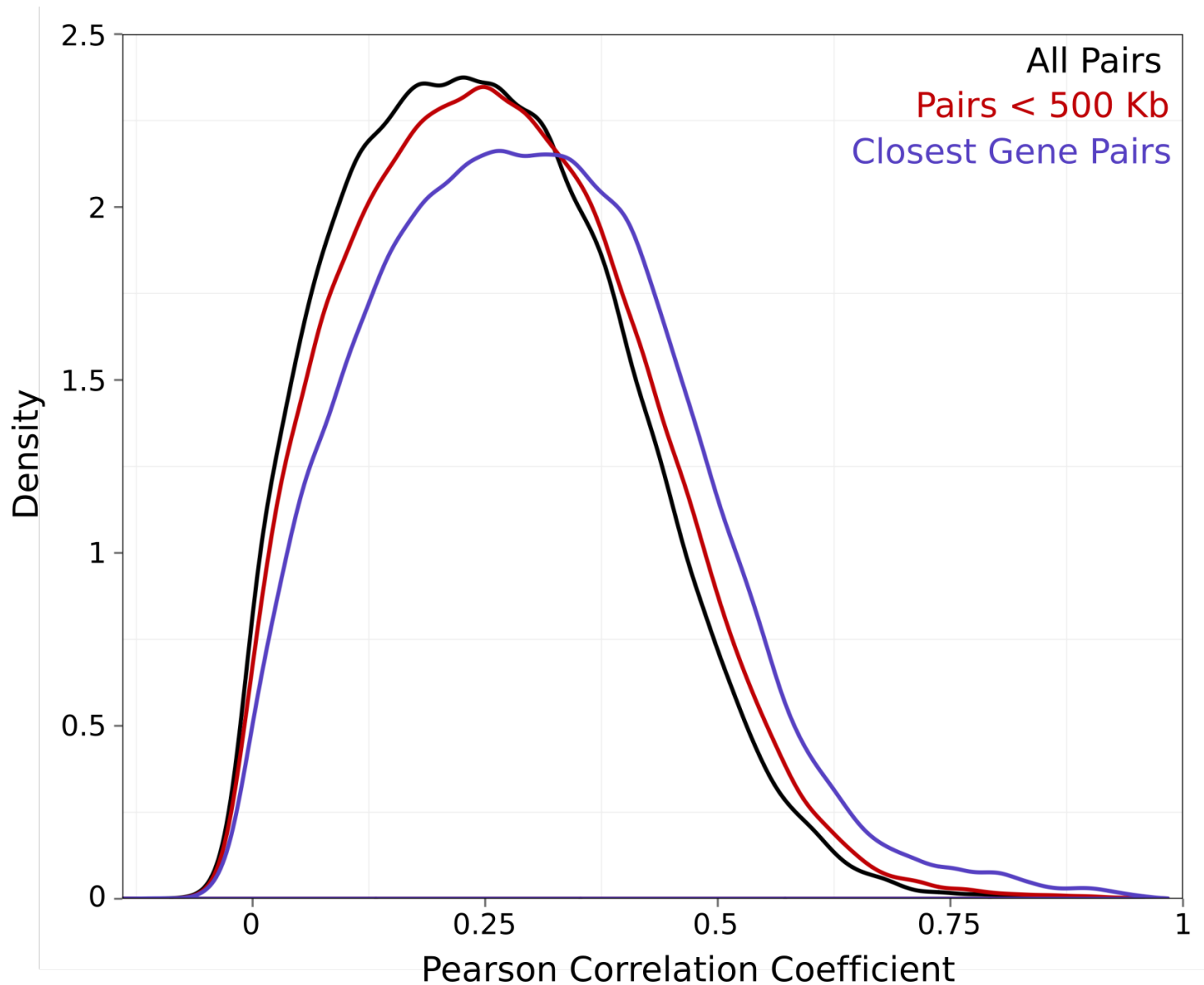
Distribution of PCC for All Distal-Proximal Pairs



Distribution of PCC for Pairs < 500 Kb



Distribution of PCC for Closest Gene Pairs

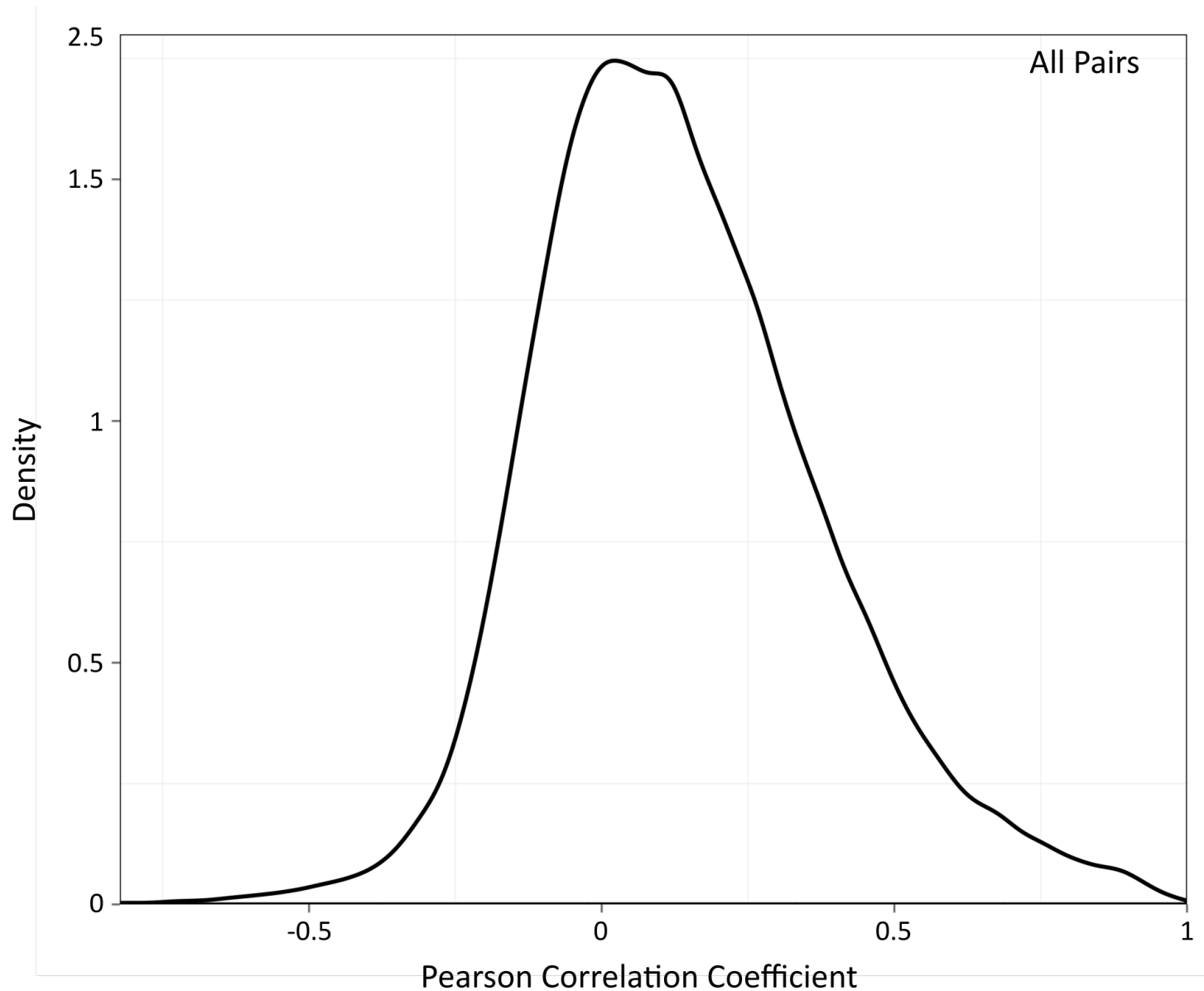


H3K27ac Results

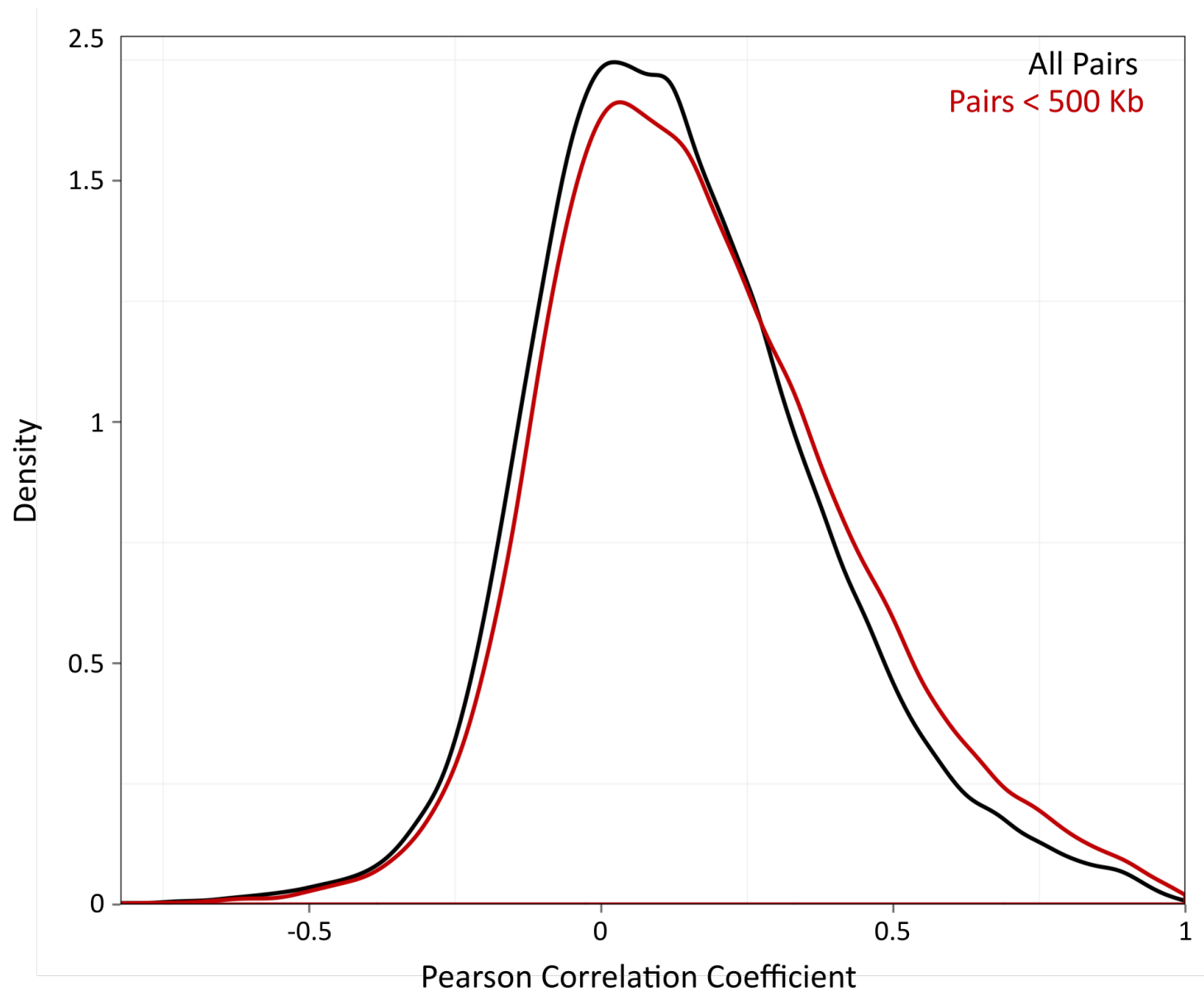
Correlation With Raw Signal Data

All initial analysis was conducted on chromosome 22

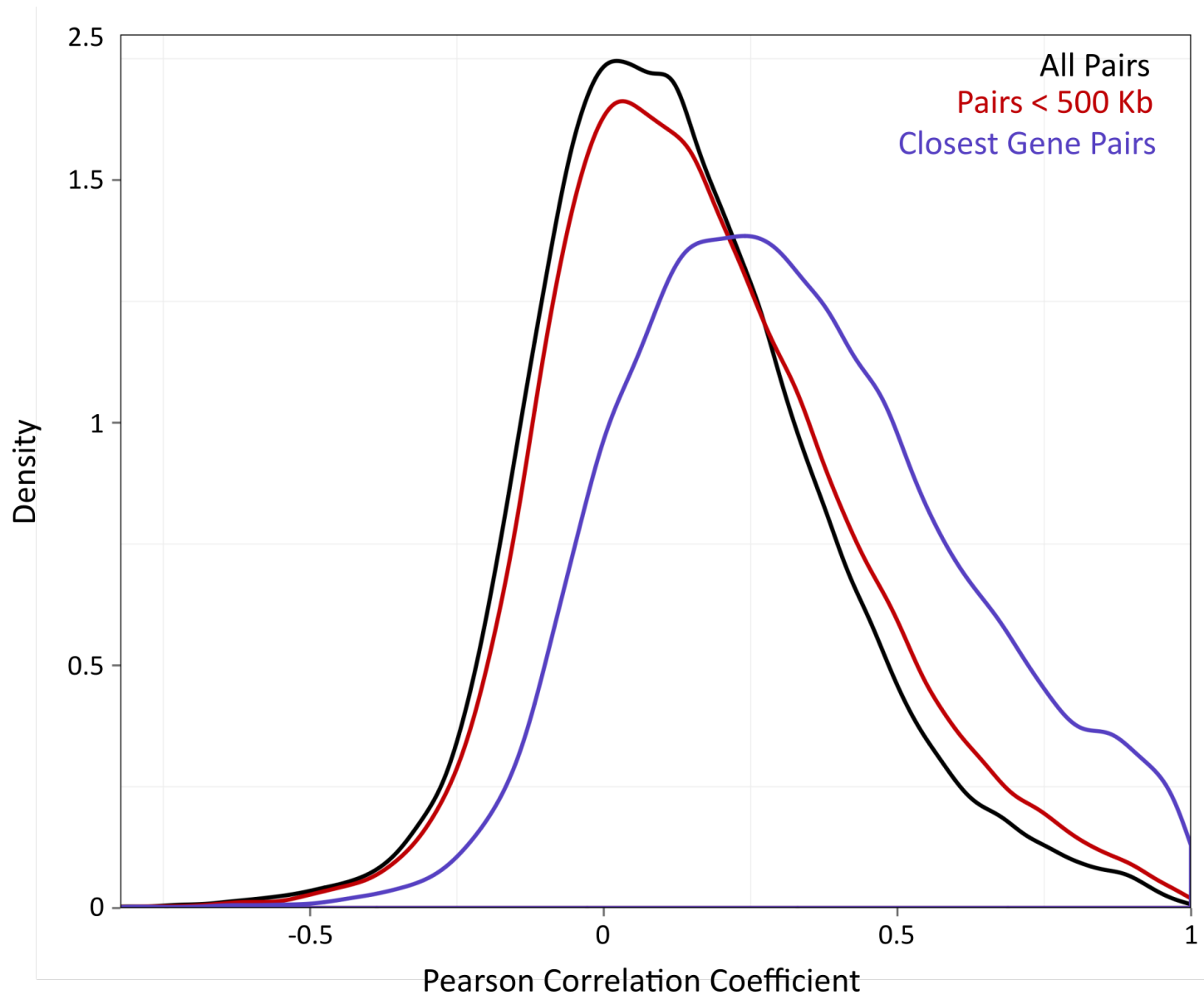
Distribution of PCC for All Distal-Proximal Pairs



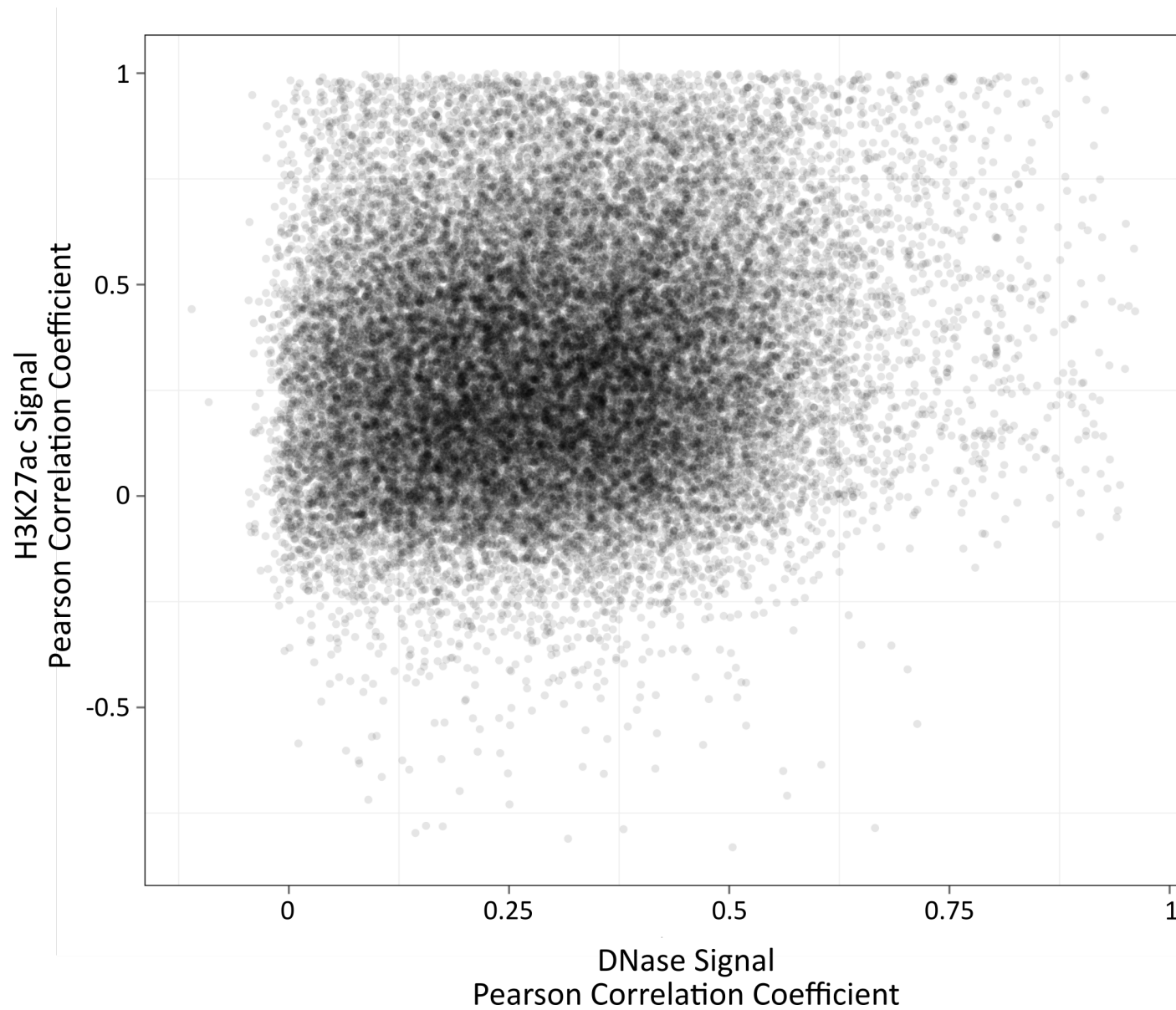
Distribution of PCC for Pairs < 500 Kb



Distribution of PCC for Closest Gene Pairs



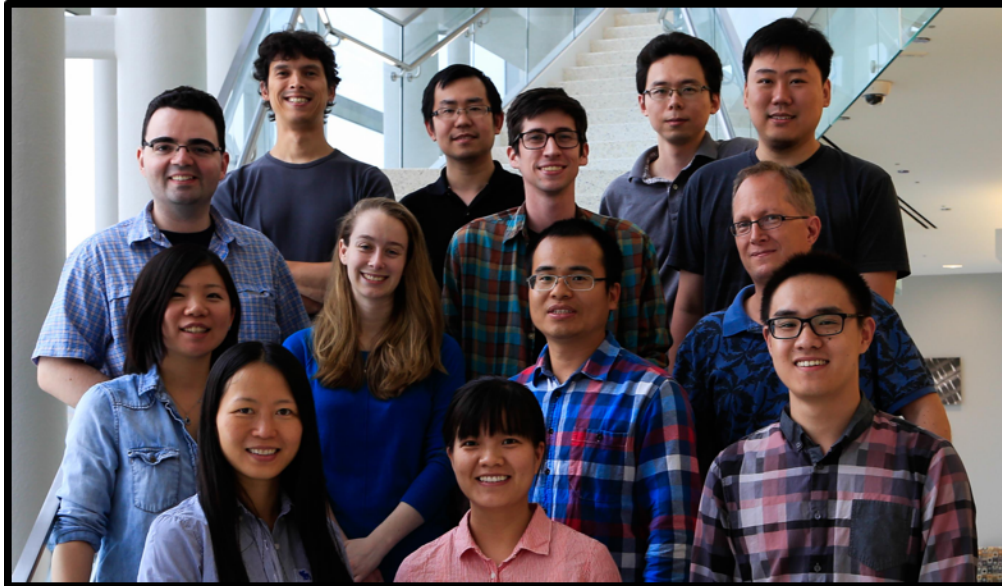
Closest Gene Pairs: DNase vs H3K27ac PCC



Future Directions

- Compare raw signal vs. fold change for DNase and H3K27ac correlation
- Incorporate Hi-C and ChIA-PET data
- How does correlation look across known gene-enhancer pairs?
- Develop benchmark as a metric for testing other methods
- Other suggestions?

Acknowledgements



Advisor: Zhiping Weng

Weng Lab

- Michael Purcaro
- Arjan van der Velde
- Tyler Borrman
- Henry Pratt

Stam Lab

- Richard Sandstrom