Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin

Sean Whalen Rebecca M. Truty Katherine S. Pollard

March 4, 2016

Nature Genetics: doi:10.1038/ng.3539

Supplemental Material

Supplementary Note

This study demonstrates that complex genomic signatures strongly distinguish the true targets of active enhancers from other active but non-interacting promoters in the same loci. These signatures are primarily based on patterns of protein binding and epigenetic modifications on the looping chromatin. A unique feature of our approach is the combination of high resolution genome-wide Hi-C interaction data [1] with the vast functional genomics datasets provided by the ENCODE and Roadmap Epigenomics projects partitioned by enhancer, promoter, and window regions. By integrating these diverse datasets and examining their relevance to enhancer-promoter interactions, we computed the most predictive datasets and highlighted the complex interplay between regulatory proteins and DNA in the three-dimensional genome.

Careful examination of many enhancer-promoter pairs across cell lines suggests several broad rules influence TargetFinders score of an enhancer-promoter interaction: 1) whether the enhancer looks truly active in a given cell line (reducing false positive annotations), 2) whether the target promoter has lineage-specific marks, 3) whether the target is actively elongating, 4) whether alternate promoters near the target have marks of repression or paused polymerase, 5) whether other pairs are interacting in the window, 6) whether the interaction crosses a contact domain or TAD, and 7) whether known looping factors are bound near the enhancer and promoter or within the window.

Our predictive accuracy and biological insights depended critically on the decision to include genomic data from the window between each enhancer and promoter in the analyses. We discovered these window features dominated those encoding chromatin states at the promoter and enhancer themselves. Because all enhancers and promoters we studied, including non-interacting pairs, had sufficient activation marks to be called by ChromHMM/Segway, our analysis revealed more subtle and complex genomic signatures that distinguish regulatory targets from poised, paused, or insulated promoters. The genomic signature of looping DNA has several components. First, interacting pairs are depleted for insulator and contact domain crossings in the window (Supplementary Figures 8 and 10), particularly for more distal interactions. Second, interacting pairs are depleted for cohesin complex bound to the window (Supplementary Figure 11), although it is prevalent near the enhancer and promoter. Third, DNA between interacting enhancers and promoters tends to lack activating TFs and epigenetic marks of elongation (Supplementary Figure 12) which could indicate the presence of an alternative promoter target, and indeed is depleted for active promoters (Supplementary Figure 13). On the other hand, windows do contain epigenetic marks associated with heterochromatin (Supplementary Figure 7), polycomb-associated proteins, and co-factors of CTCF associated with its insulator function. Looping interactions in the window are highly enriched (Supplementary Figure 14), strongly supporting existing evidence for TADs or contact domains and suggesting window features may be a proxy for domain membership.

Alternate software implementation

To confirm robustness to underlying software, an alternative version of the *TargetFinder* pipeline was implemented in R using the `caret` [2], `randomForest` [3], `gbm` [4], and `glmnet` [5] packages. Results were consistent with the Python version.

Alternate promoter and enhancer definitions

Before transitioning to annotation-based enhancers, we defined candidate enhancers as transcription factor binding sites (TFBS) identified by CENTIPEDE [6], lifted these over from the hg18 to hg19 assembly, and clustered them [7] using the DBSCAN algorithm [8] with `eps = 300` and `min_samples = 1`. Finally, we intersected the resulting TFBS clusters with EP300, H3K27ac, and H3K4me1 ChIP-seq peaks from the same cell line and retained all clusters that overlapped at least one of these ChIP-seq marks. Clusters closer than 10Kb to the nearest promoter were discarded. Promoters were defined using a fixed 1 Kb window centered at the transcription start site of each expressed gene. This approach had comparable performance for 5C-assayed interactions but was outperformed by segmentation based enhancer and promoter definitions for Hi-C-assayed interactions.

Nature Genetics: doi:10.1038/ng.3539

Alternate interaction data

To assess the robustness of our approach to interaction data, we used chromosome conformation capture carbon copy (5C) data from ENCODE that also identifies physically interacting segments of the genome [9]. Enhancers were intersected with forward 5C fragments and promoters with reverse 5C fragments. Following the ENCODE standard for interaction significance, enhancer-promoter pairs with fragments found to interact across both 5C biological replicates were given positive labels in our training data. To select negatives matching the distribution of interaction distances, positives were first assigned a bin number using quantile discretization of the distance between enhancers and promoters. For each positive distance bin, 200 negatives were generated by randomly selecting non-interacting enhancer-promoter pairs within the ENCODE pilot regions. The number of negatives per bin was limited by the number of active promoters covered by reverse 5C fragments. Performance and variable importance results were generally consistent between our 5C and Hi-C analyses, but performance was better using the larger Hi-C dataset.

Data quality affects predictive accuracy

Read length, read quality, alignment quality, sequencing depth, antibodies, cell lines, and lab protocols all influence the quality of ChIP-seq data and in turn affect downstream predictive models. For example, a protein expected to be important could be redundant with another more discriminative protein, or may merely lack a good antibody. Supporting this notion, we observe large variance in CTCF binding across data sets from different labs (Supplementary Figure 6), which likely contributes to quantitative differences in the predictive importance of a feature. We added ENCODE phase 3 experiments to *TargetFinder* and found they generally receive a better predictive rank than older ones.

Nature Genetics: doi:10.1038/ng.3539

Supplemental Tables

Cell Line	Promoters	Enhancers	Interacting Pairs	Non-Interacting Pairs
K562	8196	82806	1977	39500
GM12878	8453	100036	2113	42200
HeLa-S3	7794	103460	1740	34800
HUVEC	8180	65358	1524	30400
IMR90	5253	108996	1254	25000
NHEK	5254	144302	1291	25600

Supplementary Table 1: Counts of promoters, enhancers, interacting pairs, and non-interacting pairs per cell line.

Supplementary Table 2: Datasets and accession ids for each cell line, with blanks indicating unavailability.

	Cell Line					
	K562	GM12878	HeLa-S3	HUVEC	IMR90	NHEK
CAGE	GSE34448	GSE34448	GSE34448	GSE34448		GSE34448
ChIP-seq (ARID3A)	GSE31477					
ChIP-seq (ATF1)	GSE31477					
ChIP-seq (ATF2)		GSE32465				
ChIP-seq (ATF3)	GSE32465	GSE32465				
ChIP-seq (BACH1)	GSE31477					
ChIP-seq (BATF)		GSE32465				
ChIP-seq (BCL11A)		GSE32465				
ChIP-seq (BCL3)	GSE32465	GSE32465				
ChIP-seq (BCLAF1)	GSE32465	GSE32465				
ChIP-seq (BDP1)	GSE31477		GSE31477		GSE47849	
ChIP-seq (BHLHE40)	GSE31477	GSE31477				
ChIP-seq (BRCA1)		GSE31477	GSE31477			
ChIP-seq (BRF1)	GSE31477		GSE31477			
ChIP-seq (BRF2)	GSE31477		GSE31477			
ChIP-seq (CBX2)	GSE29611					
ChIP-seq (CBX3)	GSE32465					
ChIP-seq (CBX8)	GSE29611					
ChIP-seq (CCNT2)	GSE31477					
ChIP-seq (CEBPB)	GSE32465	GSE32465	GSE31477		GSE31477	
ChIP-seq (CEBPD)	GSE32465					
ChIP-seq (CHD1)	GSE29611	GSE31477			GSE31477	
ChIP-seq (CHD2)	GSE31477	GSE31477	GSE31477			
ChIP-seq (CHD4)	GSE29611					
ChIP-seq (CHD7)	GSE29611					
ChIP-seq (CREB1)	GSE32465	GSE32465				
ChIP-seq (CREBBP)	GSE29611					
ChIP-seq (CTCF)	GSE32465	GSE29611	GSE29611	GSE29611	GSE31477	GSE29611
ChIP-seq (CTCFL)	GSE32465					
ChIP-seq (CUX1)	GSE31477	GSE31477				
ChIP-seq (E2F1)			GSE31477			
ChIP-seq (E2F4)	GSE31477	GSE31477	GSE31477			
ChIP-seq (E2F6)	GSE32465		GSE31477			
ChIP-seq (EBF1)		GSE32465				
ChIP-seq (EGR1)	GSE32465	GSE32465				

Nature Genetics: doi:10.1038/ng.3539

Supplementary Table 2: Datasets and accession ids for each cell line, with blanks indicating unavailability.

	Cell Line					
	K562	GM12878	HeLa-S3	HUVEC	IMR90	NHEK
ChIP-seq (ELF1)	GSE32465	GSE32465				
ChIP-seq (ELK1)	GSE31477	GSE31477	GSE31477			
ChIP-seq (ELK4)			GSE31477			
ChIP-seq (EP300)	GSE29611	GSE31477	GSE31477		GSE43070	
ChIP-seq (ETS1)	GSE32465	GSE32465				
ChIP-seq (EZH2)	GSE29611	GSE29611	GSE29611	GSE29611		GSE29611
ChIP-seq (FOS)	GSE31363	GSE31477	GSE31477	GSE31477		
ChIP-seq (FOSL1)	GSE32465					
ChIP-seq (FOXM1)		GSE32465				
ChIP-seq (GABPA)	GSE32465	GSE32465	GSE32465			
ChIP-seq (GATA1)	GSE31477					
ChIP-seq (GATA2)	GSE32465			GSE31477		
ChIP-seq (GDOWN1)					GSE33128	
ChIP-seq (GRHL3)						GSE42180
ChIP-seq (GTF2B)	GSE31477					
ChIP-seq (GTF2F1)	GSE31477		GSE31477			
ChIP-seq (GTF3C2)	GSE31477		GSE31477			
ChIP-seq (H1B)					GSE26979	
ChIP-seq (H2AK5ac)					GSE16256	
ChIP-seq (H2AK9ac)					GSE16256	
ChIP-seq (H2AY)					GSE54847	
ChIP-seq (H2AZ)	GSE29611	GSE29611	GSE29611	GSE29611	GSE16256	GSE29611
ChIP-seq (H2BK120ac)					GSE16256	
ChIP-seq (H2BK12ac)					GSE16256	
ChIP-seq (H2BK15ac)					GSE16256	
ChIP-seq (H2BK20ac)					GSE16256	
ChIP-seq (H2BK5ac)					GSE16256	
ChIP-seq (H3K14ac)					GSE16256	
ChIP-seq (H3K18ac)					GSE16256	
ChIP-seq (H3K23ac)					GSE16256	
ChIP-seq (H3K27ac)	GSE29611	GSE29611	GSE29611	GSE29611	GSE16256	GSE29611
ChIP-seq (H3K27me3)	GSE29611	GSE29611	GSE29611	GSE29611	GSE16256	GSE29611
ChIP-seq (H3K36me3)	GSE29611	GSE29611	GSE29611	GSE29611	GSE16256	GSE29611
ChIP-seq (H3K4ac)					GSE16256	
ChIP-seq (H3K4me1)	GSE29611	GSE29611	GSE29611	GSE29611	GSE16256	GSE29611
ChIP-seq (H3K4me2)	GSE29611	GSE29611	GSE29611	GSE29611	GSE16256	GSE29611
ChIP-seq (H3K4me3)	GSE29611	GSE29611	GSE29611	GSE29611	GSE16256	GSE29611
ChIP-seq (H3K56ac)					GSE16256	
ChIP-seq (H3K79me1)					GSE16256	
ChIP-seq (H3K79me2)	GSE29611	GSE29611	GSE29611	GSE29611	GSE16256	GSE29611
ChIP-seq (H3K9ac)	GSE29611	GSE29611	GSE29611	GSE29611	GSE16256	GSE29611
ChIP-seq (H3K9me1)	GSE29611			GSE29611		GSE29611
ChIP-seq (H3K9me3)	GSE29611	GSE29611	GSE29611	GSE29611	GSE16256	GSE29611
ChIP-seq (H3k9me1)					GSE16256	
ChIP-seq (H4K16ac)					GSE58740	
ChIP-seq (H4K20me1)	GSE29611	GSE29611	GSE29611	GSE29611	GSE16256	GSE29611
ChIP-seq (H4K5ac)					GSE16256	
ChIP-seq (H4K8ac)					GSE16256	
ChIP-seq (H4K91ac)					GSE16256	

Nature Genetics: doi:10.1038/ng.3539

Supplementary Table 2: Datasets and accession ids for each cell line, with blanks indicating unavailability.

	Cell Line					
	K562	GM12878	HeLa-S3	HUVEC	IMR90	NHEK
ChIP-seq (HA-E2F1)			GSE31477			
ChIP-seq (HCFC1)	GSE31477		GSE31477			
ChIP-seq (HDAC1)	GSE29611					
ChIP-seq (HDAC2)	GSE32465					
ChIP-seq (HDAC6)	GSE29611					
ChIP-seq (HDAC8)	GSE31363					
ChIP-seq (HMGN3)	GSE31477					
ChIP-seq (IKZF1)		GSE31477				
ChIP-seq (IRF1)	GSE31477					
ChIP-seq (IRF3)		GSE31477	GSE31477			
ChIP-seq (IRF4)		GSE32465				
ChIP-seq (JUN)	GSE31477		GSE31477	GSE31477		
ChIP-seq (JUNB)	GSE31363					
ChIP-seq (JUND)	GSE31477	GSE31477	GSE31477			
ChIP-seq (KAT2A)		GSE31477	GSE31477			
ChIP-seq (KAT2B)	GSE29611					
ChIP-seq (KDM1A)	GSE29611					
ChIP-seq (KDM5B)	GSE29611					
ChIP-seq (LMNB1)					GSE53332	
ChIP-seq (MAFF)	GSE31477					
ChIP-seq (MAFK)	GSE31477	GSE31477	GSE31477		GSE31477	
ChIP-seq (MAX)	GSE32465	GSE31477	GSE31477	GSE31477		
ChIP-seq (MAZ)	GSE31477	GSE31477	GSE31477		GSE31477	
ChIP-seq (MECP2)					GSE47678	
ChIP-seq (MEF2A)	GSE32465	GSE32465				
ChIP-seq (MEF2C)		GSE32465				
ChIP-seq (MTA3)		GSE32465				
ChIP-seq (MXI1)	GSE31477	GSE31477	GSE31477		GSE31477	
ChIP-seq (MYC)	GSE31477	GSE33213	GSE31477	GSE33213		
ChIP-seq (NCOR1)	GSE29611					
ChIP-seq (NELFE)	GSE31477					
ChIP-seq (NFATC1)		GSE32465				
ChIP-seq (NFE2)	GSE31477	GSE31477				
ChIP-seq (NFIC)		GSE32465				
ChIP-seq (NFKB1)		GSE31477				
ChIP-seq (NFYA)	GSE31477	GSE31477	GSE31477			
ChIP-seq (NFYB)	GSE31477	GSE31477	GSE31477			
ChIP-seq (NR2C2)	GSE31477	GSE31477	GSE31477			
ChIP-seq (NR2F2)	GSE32465					
ChIP-seq (NR4A1)	GSE31363					
ChIP-seq (NRF1)	GSE31477	GSE31477	GSE31477			
ChIP-seq (PAX5)		GSE32465				
ChIP-seq (PBX3)		GSE32465				
ChIP-seq (PHF8)	GSE29611					
ChIP-seq (PML)	GSE32465	GSE32465				
ChIP-seq (POLR2A)	GSE32465	GSE32465	GSE32465	GSE31477	GSE31477	GSE29611
ChIP-seq (POLR3A)			GSE31477			
ChIP-seq (POLR3D)					GSE47849	
ChIP-seq (POLR3G)	GSE31477	GSE31477			GSE47849	

Nature Genetics: doi:10.1038/ng.3539

Supplementary Table 2: Datasets and accession ids for each cell line, with blanks indicating unavailability.

	Cell Line					
	K562	GM12878	HeLa-S3	HUVEC	IMR90	NHEK
ChIP-seq (POLR3GL)					GSE47849	
ChIP-seq (POU2F2)		GSE32465				
ChIP-seq (PRDM1)			GSE31477			
ChIP-seq (RAD21)	GSE32465	GSE32465	GSE31477		GSE31477	
ChIP-seq (RB1)					GSE19899	
ChIP-seq (RBBP5)	GSE29611					
ChIP-seq (RBL2)					GSE19899	
ChIP-seq (RCOR1)	GSE31477	GSE31477	GSE31477		GSE31477	
ChIP-seq (RELA)					GSE43070	
ChIP-seq (REST)	GSE32465	GSE32465	GSE32465			
ChIP-seq (RFX5)	GSE31477	GSE31477	GSE31477		GSE31477	
ChIP-seq (RNF2)	GSE29611					
ChIP-seq (RUNX3)		GSE32465				
ChIP-seq (RXRA)		GSE32465				
ChIP-seq (SAP30)	GSE29611					
ChIP-seq (SETDB1)	GSE31477					
ChIP-seq (SIN3A)		GSE31477				
ChIP-seq (SIN3AK20)	GSE32465					
ChIP-seq (SIRT6)	GSE31477					
ChIP-seq (SIX5)	GSE32465	GSE32465				
ChIP-seq (SMARCA4)	GSE31477		GSE31477			
ChIP-seq (SMARCB1)	GSE31477		GSE31477			
ChIP-seq (SMARCC1)			GSE31477			
ChIP-seq (SMARCC2)			GSE31477			
ChIP-seq (SMC3)	GSE31477	GSE31477	GSE31477			
ChIP-seq (SP1)	GSE32465	GSE32465				
ChIP-seq (SP2)	GSE32465					
ChIP-seq (SPI1)	GSE32465	GSE32465				
ChIP-seq (SREBF1)		GSE31477				
ChIP-seq (SREBF2)		GSE31477				
ChIP-seq (SRF)	GSE32465	GSE32465				
ChIP-seq (STAT1)	GSE31477	GSE31477	GSE31477			
ChIP-seq (STAT2)	GSE31477					
ChIP-seq (STAT3)		GSE31477	GSE31477			
ChIP-seq (STAT5A)	GSE32465	GSE32465				
ChIP-seq (SUMO2-C)	GSE66448					
ChIP-seq (SUMO2-HS)	GSE66448					
ChIP-seq (SUPT20H)		GSE31477	GSE31477			
ChIP-seq (SUZ12)	GSE29611					
ChIP-seq (TAF1)	GSE32465	GSE32465	GSE32465			
ChIP-seq (TAF7)	GSE32465					
ChIP-seq (TAL1)	GSE31477					
ChIP-seq (TBL1XR1)	GSE31477	GSE31477				
ChIP-seq (TBP)	GSE31477	GSE31477	GSE31477			
ChIP-seq (TCF12)		GSE32465				
ChIP-seq (TCF3)		GSE32465				
ChIP-seq (TCF7L2)			GSE31477			
ChIP-seq (TEAD4)	GSE32465					
ChIP-seq (TFAP2A)			GSE31477			

Nature Genetics: doi:10.1038/ng.3539

Supplementary Table 2: Datasets and accession ids for each cell line, with blanks indicating unavailability.

	Cell Line					
	K562	GM12878	HeLa-S3	HUVEC	IMR90	NHEK
ChIP-seq (TFAP2C)			GSE31477			
ChIP-seq (THAP1)	GSE32465					
ChIP-seq (TP53)					GSE58740	
ChIP-seq (TRIM28)	GSE32465					
ChIP-seq (UBTF)	GSE31477					
ChIP-seq (USF1)	GSE32465	GSE32465				
ChIP-seq (USF2)	GSE31477	GSE31477	GSE31477			
ChIP-seq (WDR5)						GSE42180
ChIP-seq (WHSC1)	GSE29611					
ChIP-seq (WRNIP1)		GSE31477				
ChIP-seq (XRCC4)	GSE31477					
ChIP-seq (YAP1)					GSE61852	
ChIP-seq (YY1)	GSE32465	GSE31477				
ChIP-seq (ZBTB33)	GSE32465	GSE32465				
ChIP-seq (ZBTB7A)	GSE32465					
ChIP-seq (ZC3H11A)	GSE31477					
ChIP-seq (ZEB1)		GSE32465				
ChIP-seq (ZKSCAN1)			GSE31477			
ChIP-seq (ZMIZ1)	GSE31477					
ChIP-seq (ZNF143)	GSE31477	GSE31477	GSE31477			
ChIP-seq (ZNF263)	GSE31477					
ChIP-seq (ZNF274)	GSE31477	GSE31477	GSE31477			
ChIP-seq (ZNF384)	GSE31477	GSE31477				
ChIP-seq (ZZZ3)		GSE31477	GSE31477			
DNase-seq	GSE29692	GSE29692	GSE29692	GSE29692	GSE18927	GSE29692
FAIRE-seq	GSE35239	GSE35239	GSE35239	GSE35239		GSE35239
Methyl-RRBS	GSE27584	GSE27584	GSE27584		GSE27584	
Total	136	100	73	23	56	20

Nature Genetics: doi:10.1038/ng.3539

Cell Line	Region	Model	TP	TN	FP	FN	AUC	MCC	FDR	F_1
GM12878	EE/P	Baseline	1788	35951	35149	1771	0.50	0.00	0.95	0.09
GM12878	EE/P	Boosted Trees	2861	70742	358	698	0.90	0.84	0.11	0.84
GM12878	EE/P	Linear SVM	1697	64449	6651	1862	0.69	0.26	0.80	0.29
GM12878	EE/P	Decision Tree	2262	69250	1850	1297	0.80	0.57	0.45	0.59
GM12878	E/P	Baseline	1033	20945	21255	1080	0.49	-0.01	0.95	0.08
GM12878	E/P	Boosted Trees	1134	41600	600	979	0.76	0.57	0.35	0.59
GM12878	E/P	Linear SVM	1137	34159	8041	976	0.67	0.18	0.88	0.20
GM12878	E/P	Decision Tree	893	40867	1333	1220	0.70	0.38	0.60	0.41
GM12878	E/P/W	Baseline	1033	20945	21255	1080	0.49	-0.01	0.95	0.08
GM12878	E/P/W	Boosted Trees	1595	41978	222	518	0.87	0.81	0.12	0.81
GM12878	E/P/W	Linear SVM	1016	37311	4889	1097	0.68	0.23	0.83	0.25
GM12878	E/P/W	Decision Tree	1170	40966	1234	943	0.76	0.49	0.51	0.52
HUVEC	EE/P	Baseline	922	19060	19540	1010	0.49	-0.01	0.95	0.08
HUVEC	EE/P	Boosted Trees	1227	38298	302	705	0.81	0.70	0.20	0.71
HUVEC	EE/P	Linear SVM	1105	29145	9455	827	0.66	0.16	0.90	0.18
HUVEC	EE/P	Decision Tree	949	37460	1140	983	0.73	0.44	0.55	0.47
HUVEC	E/P	Baseline	746	15180	15220	778	0.49	-0.00	0.95	0.09
HUVEC	E/P	Boosted Trees	604	30035	365	920	0.69	0.48	0.38	0.48
HUVEC	E/P	Linear SVM	916	19848	10552	608	0.63	0.11	0.92	0.14
HUVEC	E/P	Decision Tree	604	29432	968	920	0.68	0.36	0.62	0.39
HUVEC	E/P/W	Baseline	746	15180	15220	778	0.49	-0.00	0.95	0.09
HUVEC	E/P/W	Boosted Trees	1053	30239	161	471	0.84	0.76	0.13	0.77
HUVEC	E/P/W	Linear SVM	952	23857	6543	572	0.70	0.21	0.87	0.21
HUVEC	E/P/W	Decision Tree	836	29595	805	688	0.76	0.50	0.49	0.53
HeLa-S3	EE/P	Baseline	1080	21506	21394	1066	0.50	0.00	0.95	0.09
HeLa-S3	EE/P	Boosted Trees	1685	42667	233	461	0.89	0.82	0.12	0.83
HeLa-S3	EE/P	Linear SVM	901	38089	4811	1245	0.65	0.20	0.84	0.23
HeLa-S3	EE/P	Decision Tree	1379	41850	1050	767	0.81	0.58	0.43	0.60
HeLa-S3	E/P	Baseline	800	17260	17540	940	0.48	-0.02	0.96	0.08
HeLa-S3	E/P	Boosted Trees	1020	34223	577	720	0.78	0.59	0.36	0.61
HeLa-S3	E/P	Linear SVM	799	28340	6460	941	0.64	0.15	0.89	0.18
HeLa-S3	E/P	Decision Tree	841	33874	926	899	0.73	0.45	0.52	0.48
HeLa-S3	E/P/W	Baseline	800	17260	17540	940	0.48	-0.02	0.96	0.08
HeLa-S3	E/P/W	Boosted Trees	1473	34631	169	267	0.92	0.87	0.10	0.87
HeLa-S3	E/P/W	Linear SVM	974	30116	4684	766	0.71	0.25	0.83	0.26
HeLa-S3	E/P/W	Decision Tree	1240	34087	713	500	0.85	0.66	0.37	0.67
IMR90	EE/P	Baseline	983	19033	18767	914	0.51	0.01	0.95	0.09

Nature Genetics: doi:10.1038/ng.3539

IMR90	EE/P	Boosted Trees	1468	37647	153	429	0.88	0.83	0.09	0.83
IMR90	EE/P	Linear SVM	912	33224	4576	985	0.68	0.22	0.83	0.25
IMR90	EE/P	Decision Tree	1164	36821	979	733	0.79	0.55	0.46	0.58
IMR90	E/P	Baseline	620	12512	12488	634	0.50	-0.00	0.95	0.09
IMR90	E/P	Boosted Trees	500	24670	330	754	0.69	0.47	0.40	0.48
IMR90	E/P	Linear SVM	537	21300	3700	717	0.64	0.16	0.87	0.20
IMR90	E/P	Decision Tree	491	24122	878	763	0.68	0.34	0.64	0.37
IMR90	E/P/W	Baseline	620	12512	12488	634	0.50	-0.00	0.95	0.09
IMR90	E/P/W	Boosted Trees	913	24839	161	341	0.86	0.78	0.15	0.78
IMR90	E/P/W	Linear SVM	673	22090	2910	581	0.71	0.26	0.81	0.28
IMR90	E/P/W	Decision Tree	721	24359	641	533	0.77	0.53	0.47	0.55
K562	EE/P	Baseline	1420	27970	27030	1330	0.51	0.01	0.95	0.09
K562	EE/P	Boosted Trees	2074	54716	284	676	0.87	0.81	0.12	0.81
K562	EE/P	Linear SVM	914	50718	4282	1836	0.63	0.19	0.82	0.23
K562	EE/P	Decision Tree	1696	53542	1458	1054	0.80	0.55	0.46	0.57
K562	E/P	Baseline	1015	19891	19609	962	0.51	0.01	0.95	0.09
K562	E/P	Boosted Trees	1093	38996	504	884	0.77	0.60	0.32	0.61
K562	E/P	Linear SVM	1150	27381	12119	827	0.64	0.13	0.91	0.15
K562	E/P	Decision Tree	925	38316	1184	1052	0.72	0.42	0.56	0.45
K562	E/P/W	Baseline	1015	19891	19609	962	0.51	0.01	0.95	0.09
K562	E/P/W	Boosted Trees	1604	39296	204	373	0.90	0.84	0.11	0.85
K562	E/P/W	Linear SVM	754	35541	3959	1223	0.64	0.19	0.84	0.23
K562	E/P/W	Decision Tree	1359	38631	869	618	0.83	0.63	0.39	0.65
NHEK	EE/P	Baseline	836	15549	15451	723	0.52	0.02	0.95	0.09
NHEK	EE/P	Boosted Trees	1226	30823	177	333	0.89	0.82	0.13	0.83
NHEK	EE/P	Linear SVM	777	23466	7534	782	0.63	0.13	0.91	0.16
NHEK	EE/P	Decision Tree	979	30164	836	580	0.80	0.56	0.46	0.58
NHEK	E/P	Baseline	609	12904	12696	682	0.49	-0.01	0.95	0.08
NHEK	E/P	Boosted Trees	702	25223	377	589	0.76	0.58	0.35	0.59
NHEK	E/P	Linear SVM	642	16769	8831	649	0.58	0.07	0.93	0.12
NHEK	E/P	Decision Tree	574	24844	756	717	0.71	0.41	0.57	0.44
NHEK	E/P/W	Baseline	609	12904	12696	682	0.49	-0.01	0.95	0.08
NHEK	E/P/W	Boosted Trees	1121	25507	93	170	0.93	0.89	0.08	0.90
NHEK	E/P/W	Linear SVM	867	19067	6533	424	0.71	0.20	0.88	0.20
NHEK	E/P/W	Decision Tree	962	25122	478	329	0.86	0.69	0.33	0.70
4 Lines	EE/P	Baseline	5149	102855	103945	5203	0.50	-0.00	0.95	0.09
4 Lines	EE/P	Boosted Trees	8299	203595	3205	2053	0.89	0.75	0.28	0.76
4 Lines	EE/P	Linear SVM	5422	168799	38001	4930	0.67	0.18	0.88	0.20
4 Lines	EE/P	Decision Tree	5993	201092	5708	4359	0.78	0.52	0.49	0.54

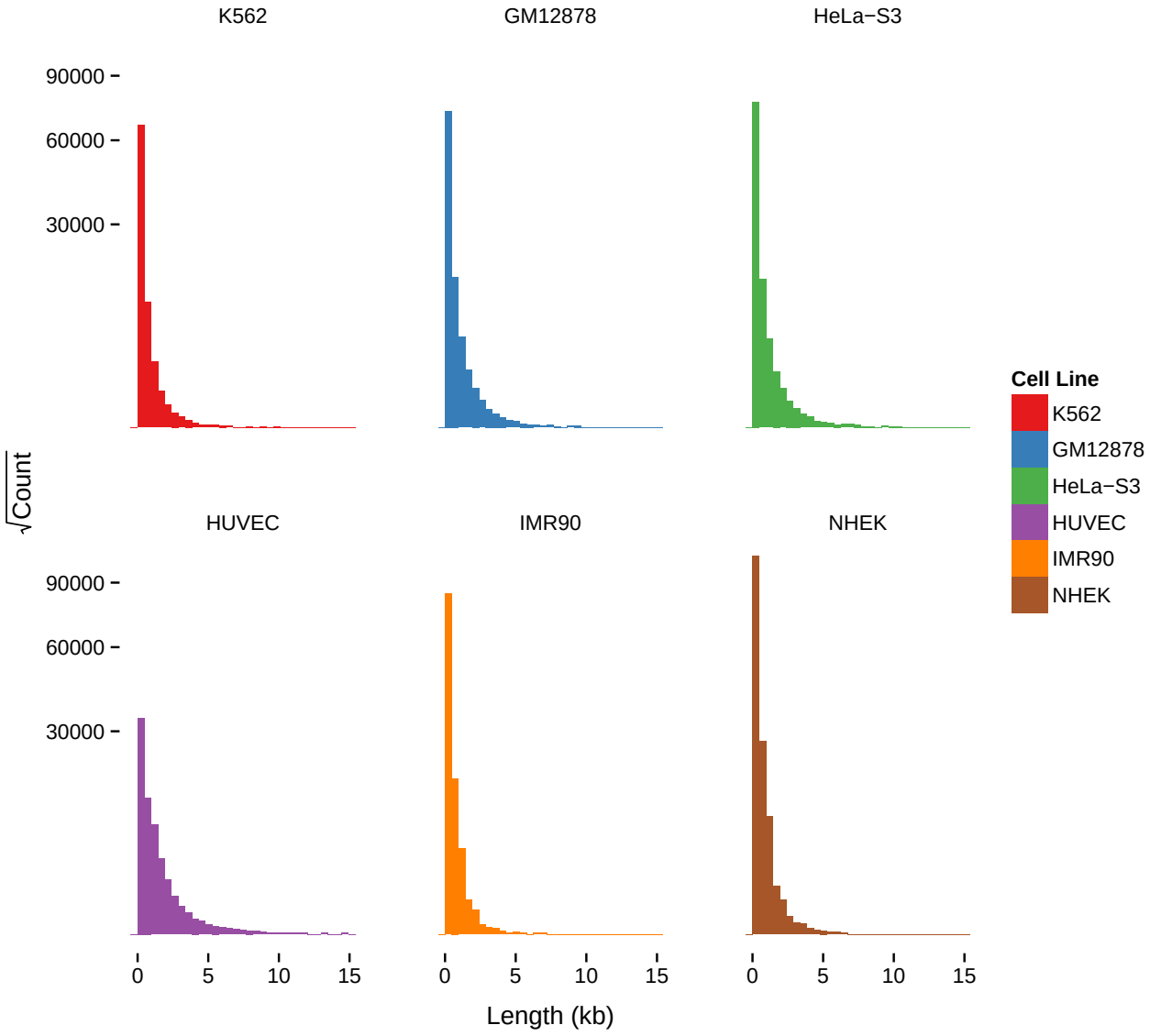
Nature Genetics: doi:10.1038/ng.3539

4 Lines	E/P	Baseline	3536	70482	71018	3548	0.50	-0.00	0.95	0.09
4 Lines	E/P	Boosted Trees	4340	137646	3854	2744	0.79	0.55	0.47	0.57
4 Lines	E/P	Linear SVM	3642	99030	42470	3442	0.61	0.10	0.92	0.14
4 Lines	E/P	Decision Tree	2733	136578	4922	4351	0.68	0.34	0.64	0.37
4 Lines	E/P/W	Baseline	3536	70482	71018	3548	0.50	-0.00	0.95	0.09
4 Lines	E/P/W	Boosted Trees	5408	140413	1087	1676	0.88	0.79	0.17	0.80
4 Lines	E/P/W	Linear SVM	4303	105061	36439	2781	0.67	0.17	0.89	0.18
4 Lines	E/P/W	Decision Tree	3831	137256	4244	3253	0.76	0.48	0.53	0.51

Supplementary Table 3: **TargetFinder performance on held out data.** Counts of true and false positives and negatives are given for each classifier and cell line, as well as several summary metrics including area under the ROC curve, Matthews correlation coefficient (ϕ), and a balance of precision and recall (F_1). Performance was measured separately for classifiers using features at enhancer and promoter loci (E/P), features at extended enhancer (+/- 3 Kb) and promoter loci (EE/P), and features using the full window between enhancer and promoter in addition to the enhancer and promoter loci themselves (E/P/W). This table demonstrates the importance of binding activity outside the enhancer and promoter for accurate prediction of interactions, and that much of this activity occurs in a relatively small region around the enhancer – though the complete window typically maximizes performance. Baseline performance was estimated by generating random predictions from the empirical distribution of class labels.

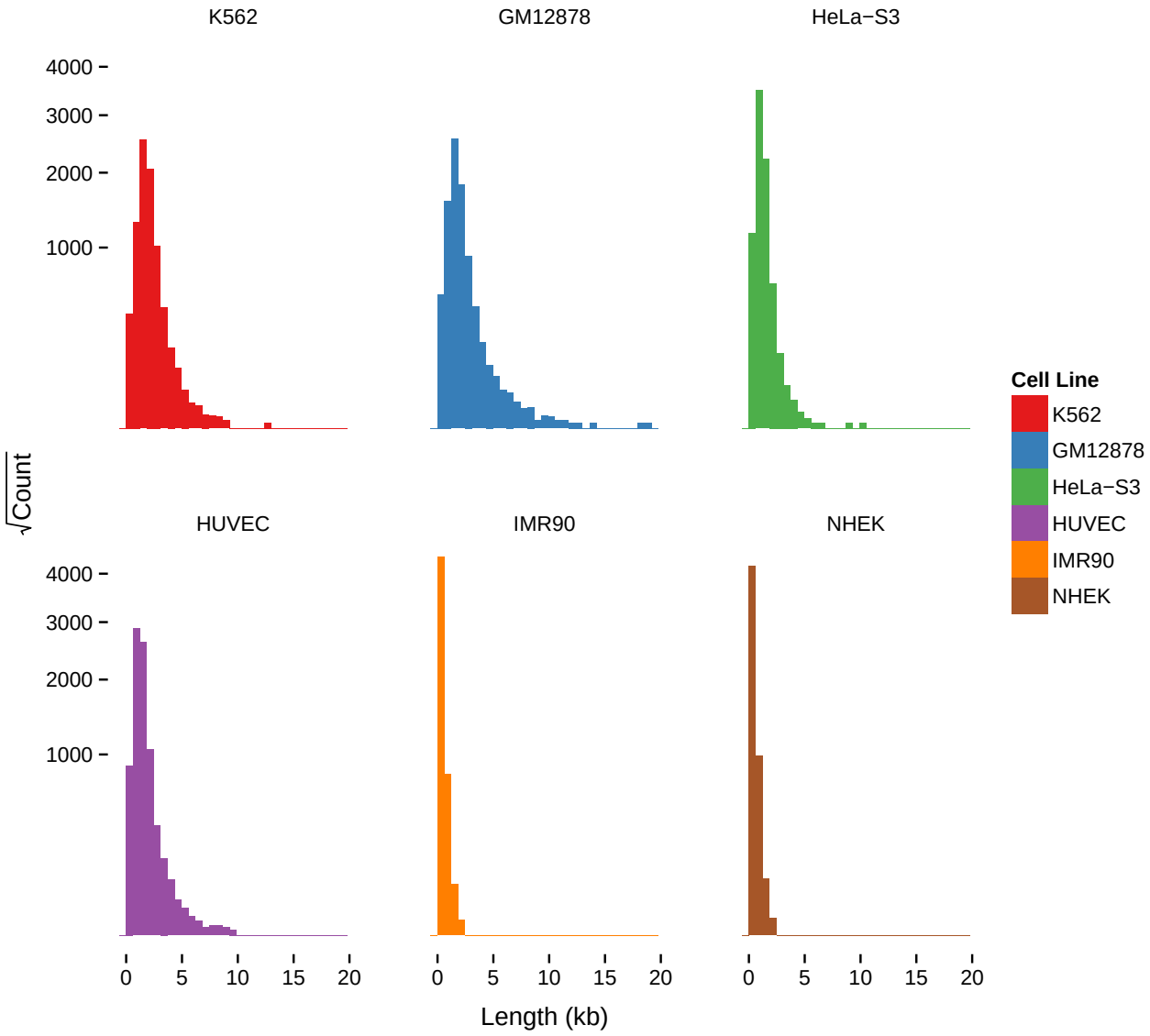
Nature Genetics: doi:10.1038/ng.3539

Supplemental Figures



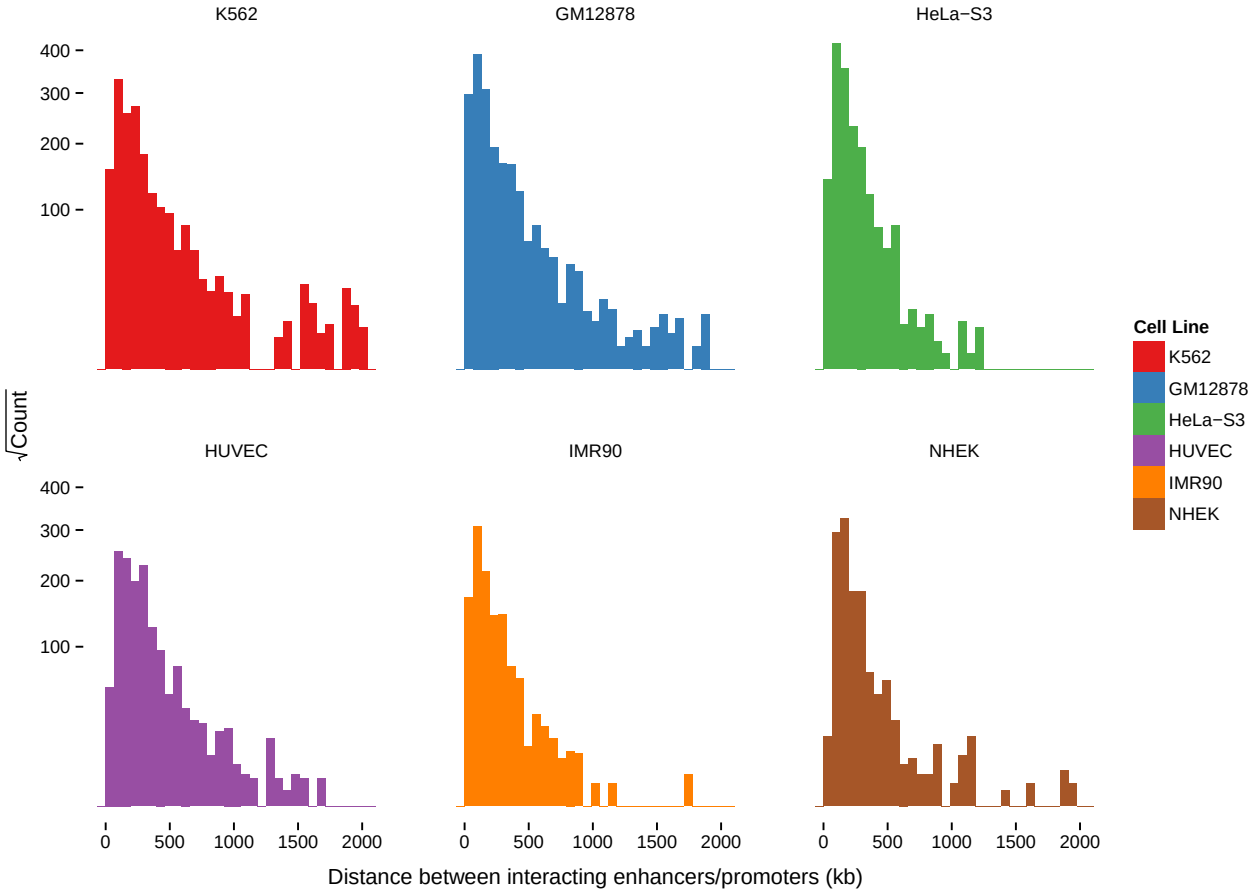
Supplementary Figure 1: Distribution of active enhancer lengths for each cell line.

Nature Genetics: doi:10.1038/ng.3539



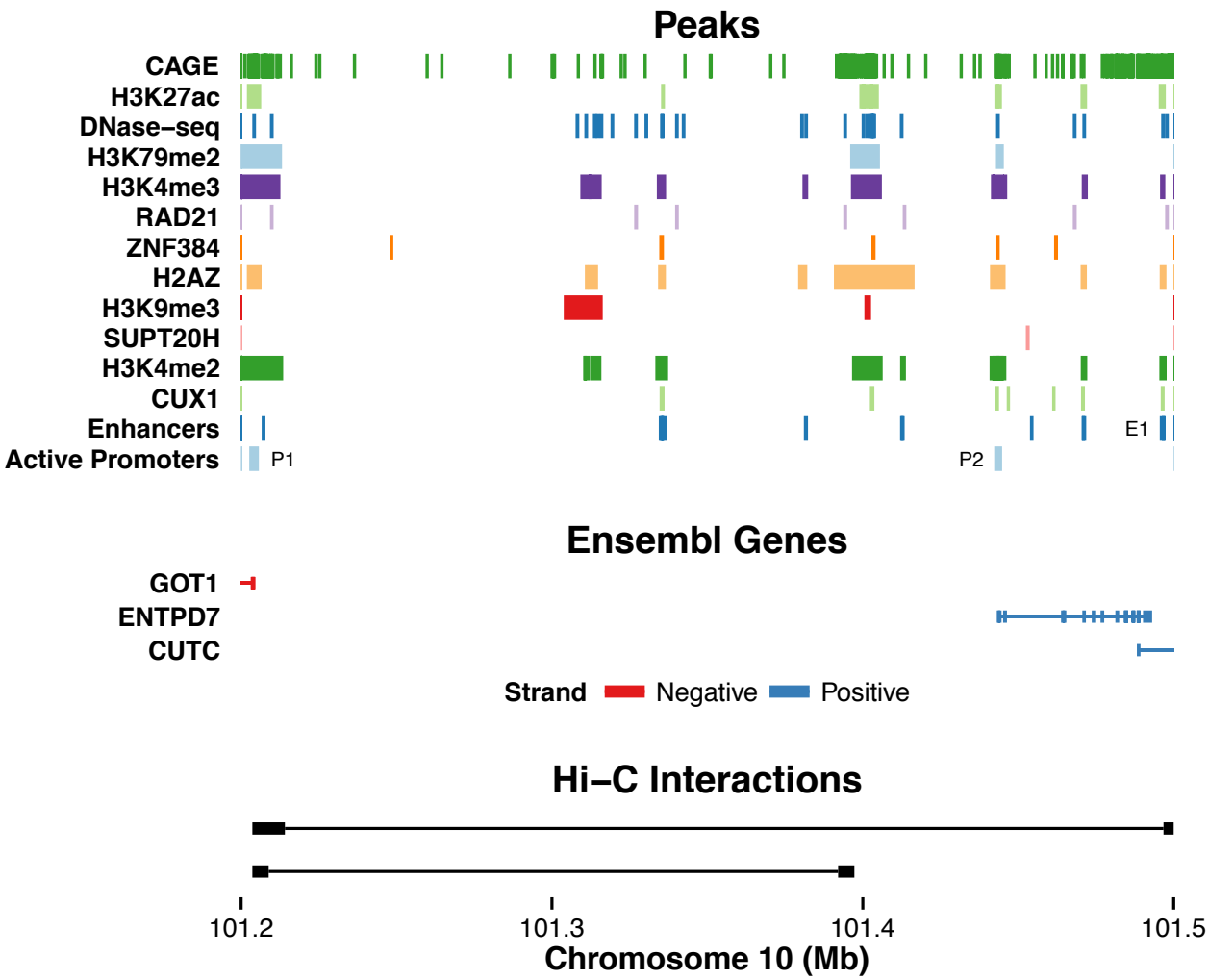
Supplementary Figure 2: Distribution of active promoter lengths for each cell line.

Nature Genetics: doi:10.1038/ng.3539



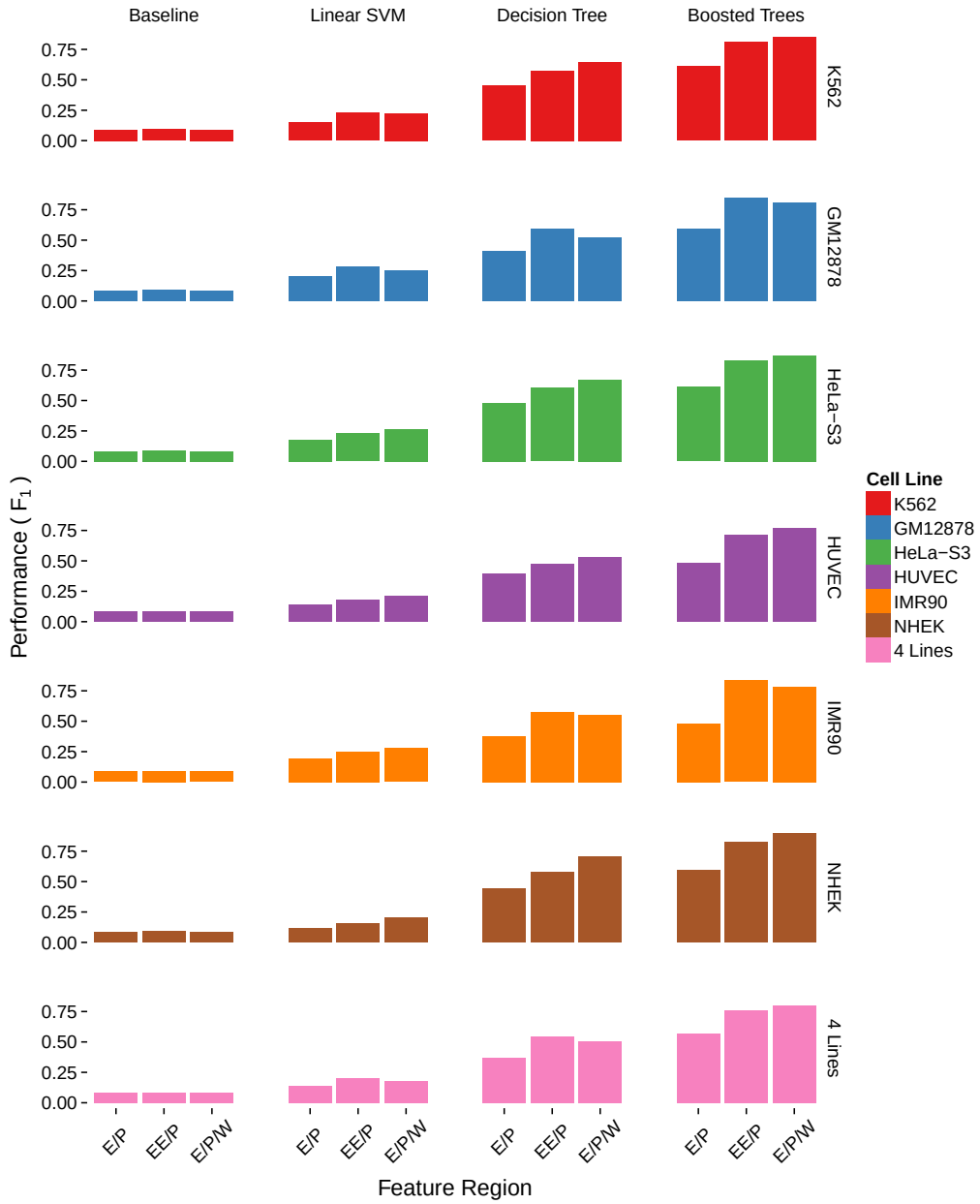
Supplementary Figure 3: Distribution of genomic distance between interacting enhancers and promoters for each cell line.

Nature Genetics: doi:10.1038/ng.3539



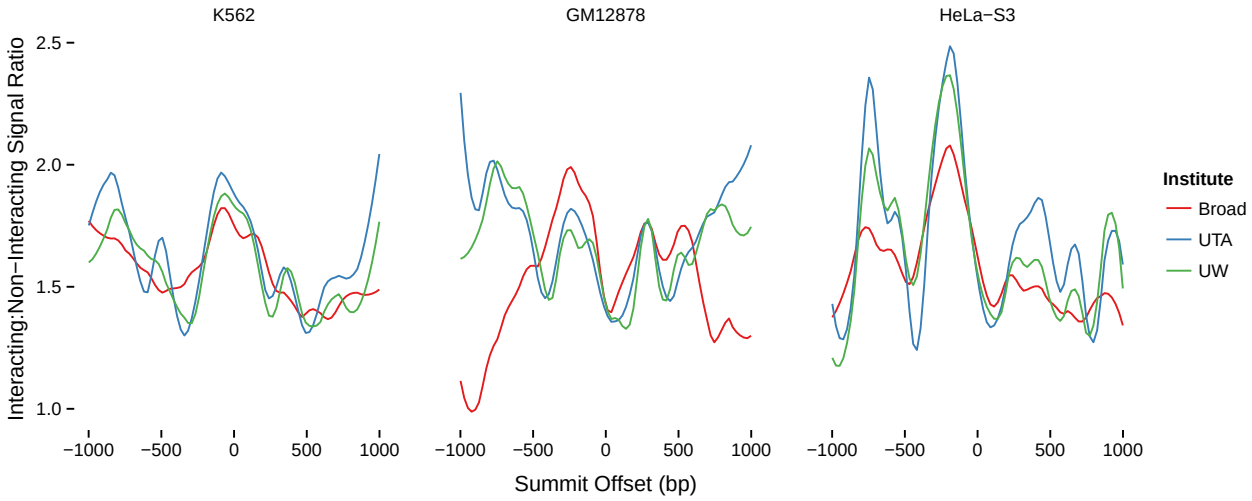
Supplementary Figure 4: Significant peaks for the top 12 predictive datasets of an interacting promoter (P1) and enhancer (E1) in GM12878, separated by other active promoters and enhancers. Active enhancers are segments marked "E" by combined ChromHMM/Segway annotations, and active promoters are segments marked "TSS" and expressed in GM12878 with RPKM > 0.3. Ensembl genes are also displayed, with introns denoted as thin lines and exons as squares. Left and right fragments of the Hi-C assay are also shown to visually confirm E1 interacts with P1. This figure demonstrates the potential complexities of interpreting interactions. First, note that P1 is the target of multiple regulatory elements, one of which was not classified by ChromHMM/Segway as an active enhancer. This regulatory element lacks a RAD21 peak commonly associated with looping and also has a heterochromatin-associated H3K9me3 mark. However, RAD21 is not always present at looping interactions, and H3K9me3 is also associated with activation in cancer cells [10]. Second, note that E1 is an intronic enhancer, existing within the CUTC gene but interacting with P1 via long-range looping. Finally, note that non-target P2 lacks RAD21 but otherwise has many activation-associated marks. Thus, the looping interaction of an enhancer and promoter is defined by the complex interplay of protein binding and other features that often cannot be explained by the presence or absence of a single feature.

Nature Genetics: doi:10.1038/ng.3539

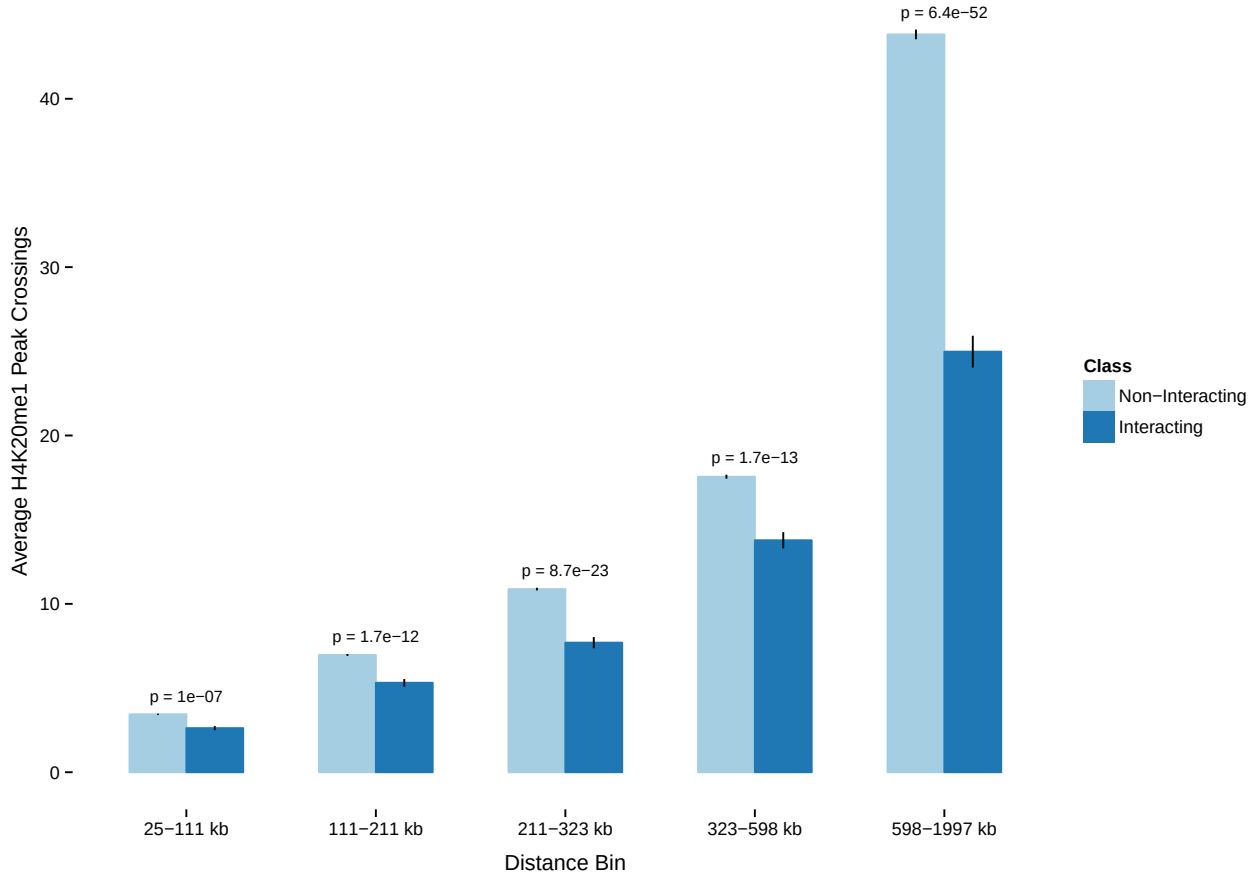


Supplementary Figure 5: **Held-out performance by cell line, model type, and region.** Performance was measured separately for classifiers using features at enhancer and promoter loci (E/P), features at extended enhancer (+/- 3 Kb) and promoter loci (EE/P), and features using the full "window" between enhancer and promoter in addition to the enhancer and promoter loci themselves (E/P/W). This figure demonstrates the importance of capturing feature interactions, as well as binding activity outside the enhancer and promoter. Much of the relevant activity flanks the enhancer, though use of the complete window typically maximizes performance. Baseline performance was estimated by generating random predictions from the empirical distribution of class labels. Reported performance is a balance of precision and recall (F_1) averaged over 10 cross-validation folds.

Nature Genetics: doi:10.1038/ng.3539

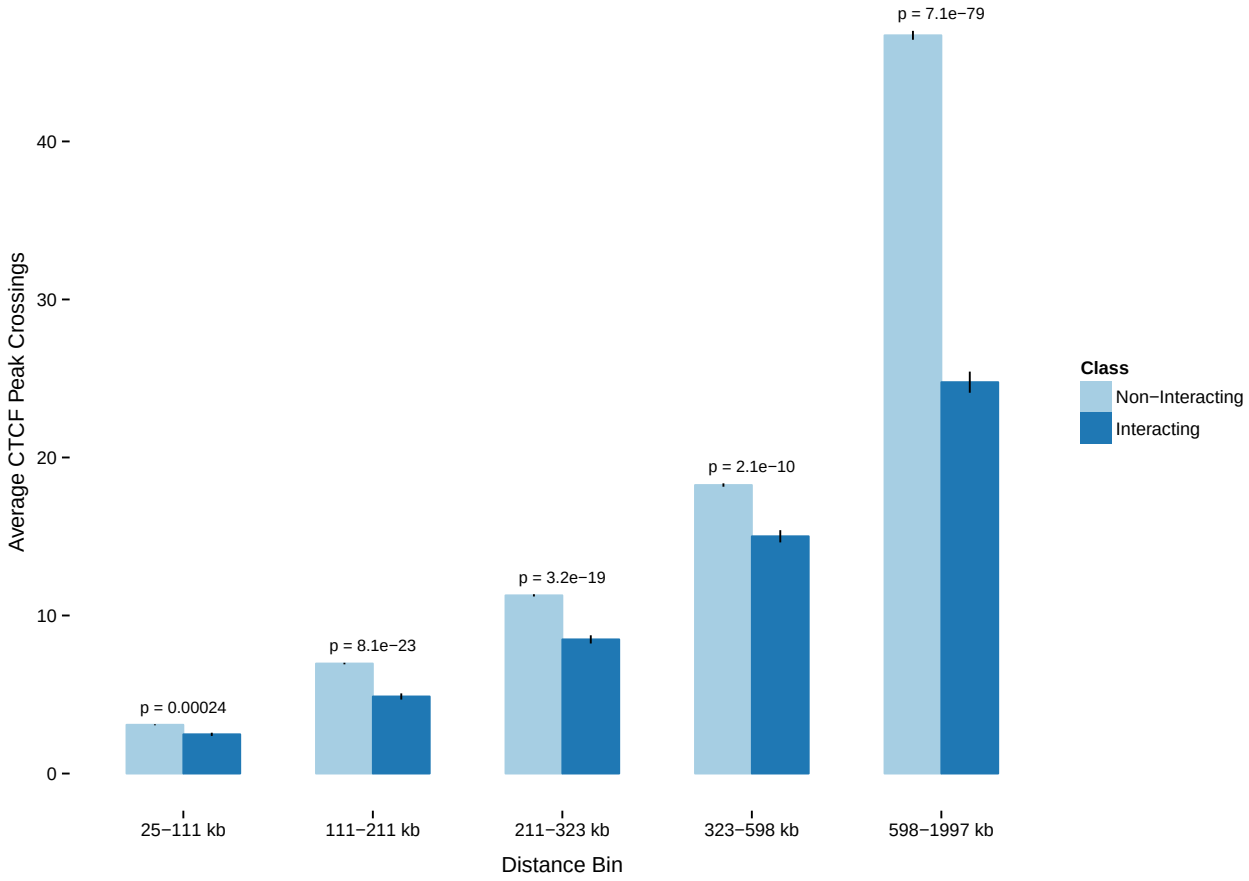


Supplementary Figure 6: Consistency of CTCF ChIP-seq signals performed at 3 different institutions across 3 ENCODE cell lines. Signals are centered at transcription start sites and displayed as the ratio of signal at interacting vs non-interacting sites. This figure shows how variability between, and even within, cell lines contributes to the difficulty of using a model trained on one cell line to predict interactions in another.



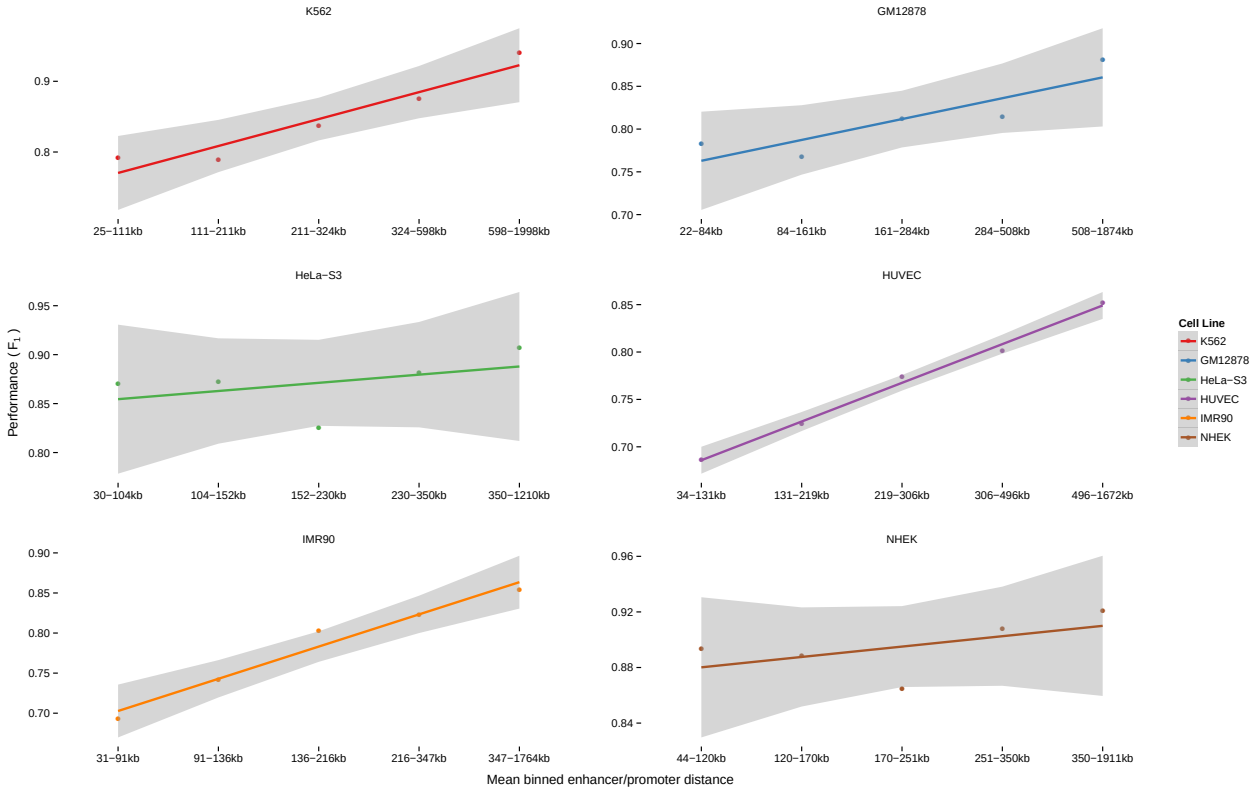
Supplementary Figure 7: Average number of H4K20me1 peaks crossed by interacting and non-interacting enhancer-promoter pairs, grouped by distance bin. The difference in mean number of crossings is significant for each distance bin (two-sample Wilcoxon test) and increases for more distal pairs. This shows repression-associated heterochromatin is depleted in the window of interacting enhancer-promoter pairs.

Nature Genetics: doi:10.1038/ng.3539



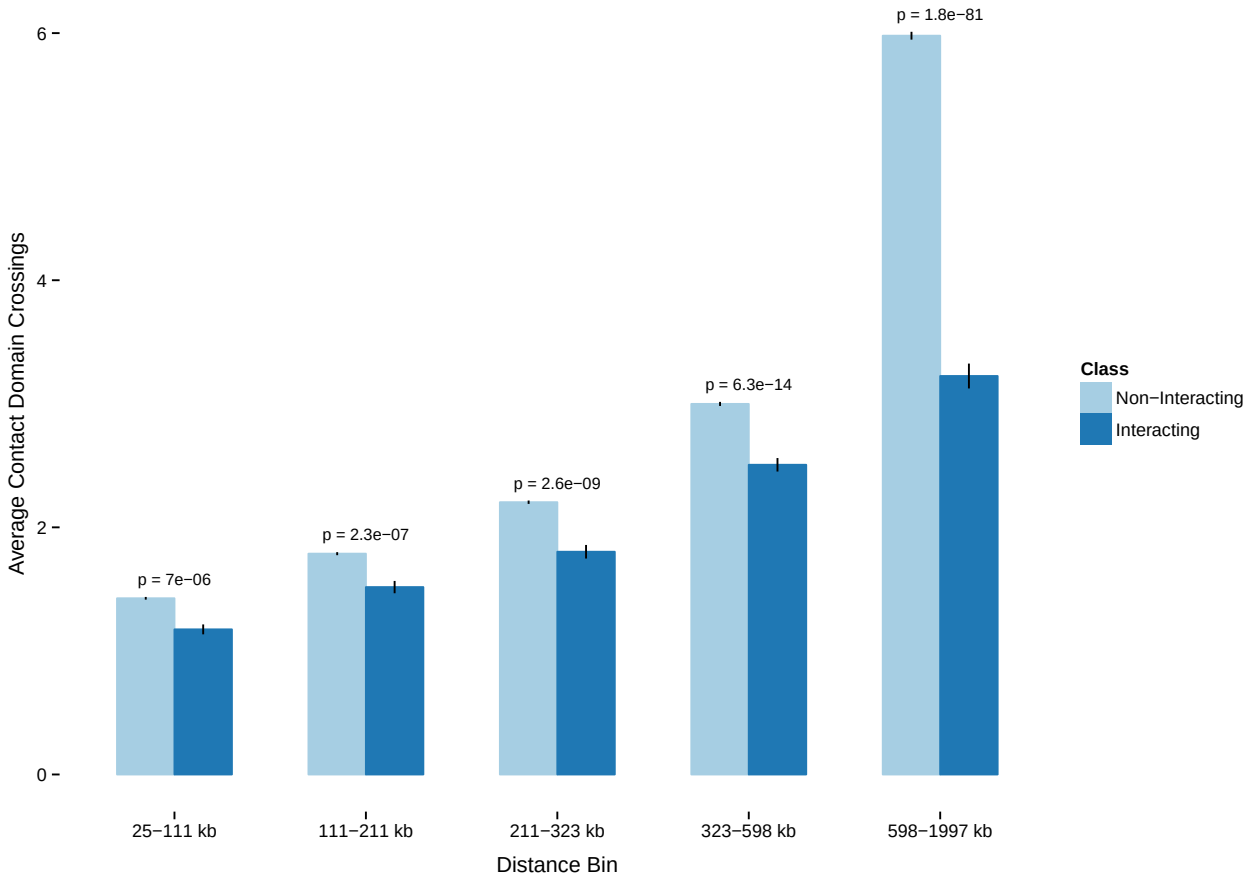
Supplementary Figure 8: Average number of CTCF peak summits crossed by interacting and non-interacting enhancer-promoter pairs, grouped by distance bin. The difference in mean number of crossings is significant for each distance bin (two-sample Wilcoxon test) and increases for more distal pairs. This reinforces current evidence that interactions preferentially occur within TADs or contact domains demarcated by CTCF acting as an insulator to prevent spurious enhancer-promoter interactions.

Nature Genetics: doi:10.1038/ng.3539



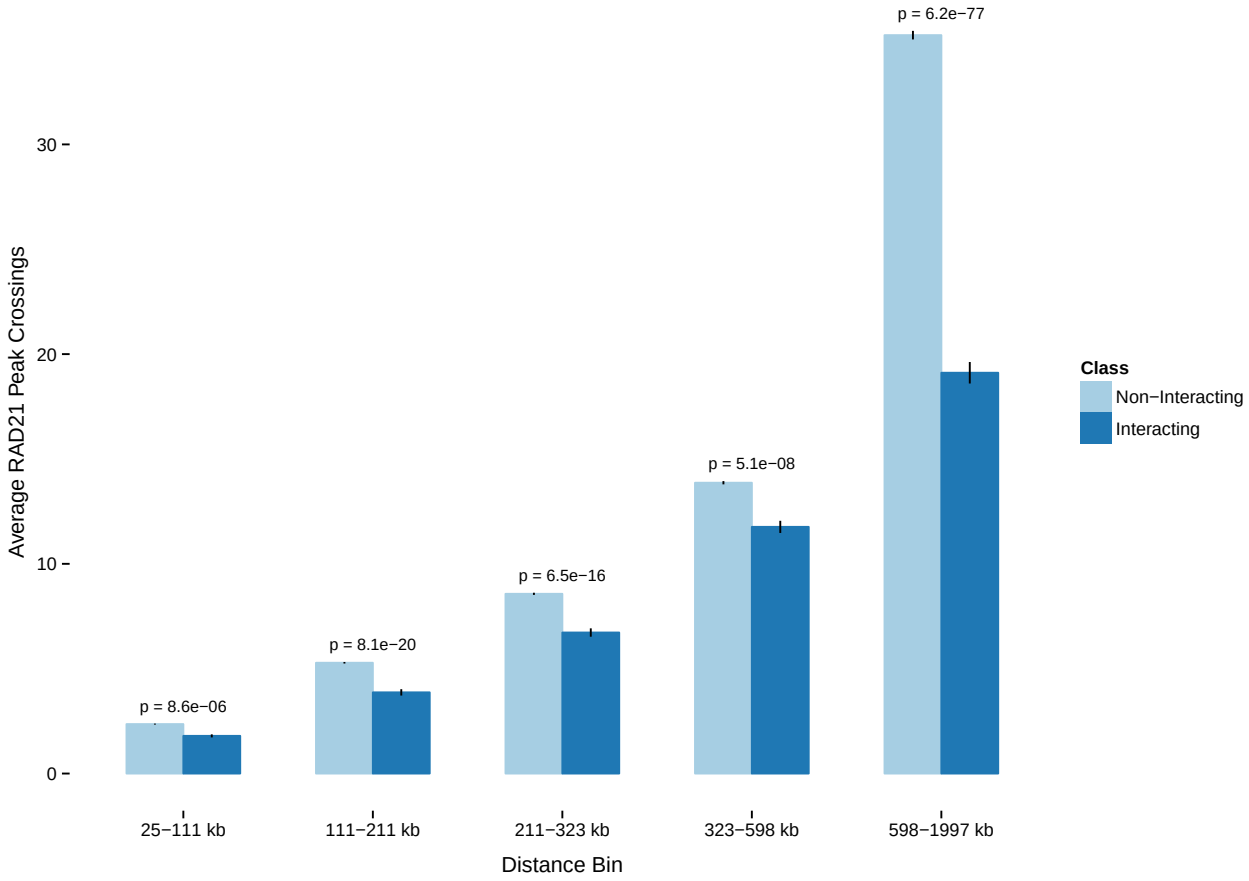
Supplementary Figure 9: Performance as a function of distance between enhancer and promoter. Samples are assigned to distance bins such that each bin has an equal number of interacting pairs. Performance is often higher for extremely distal interaction bins since TAD boundaries reduce the probability of interaction, and such boundaries are more likely to occur within the window between distal enhancer-promoter pairs. (Error bars = 95% c.i.)

Nature Genetics: doi:10.1038/ng.3539



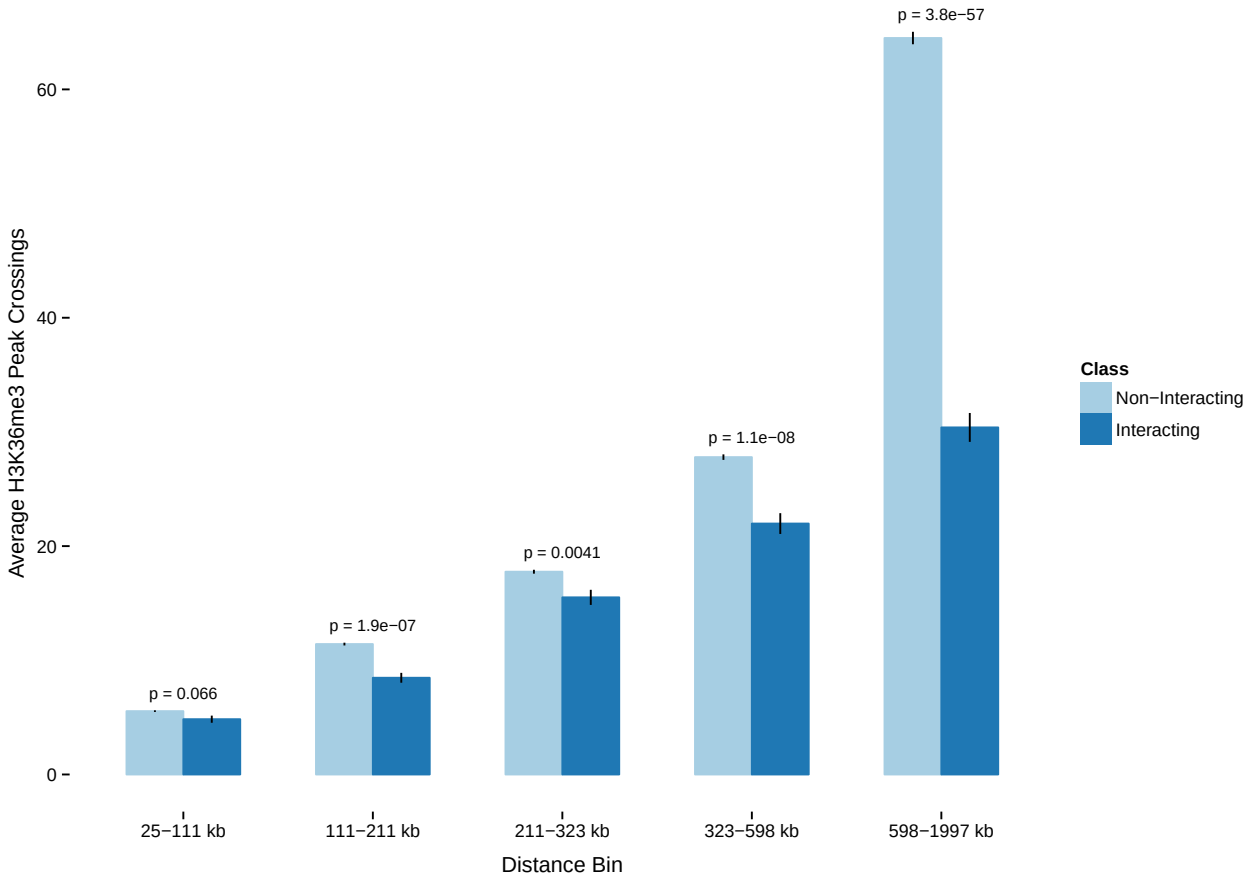
Supplementary Figure 10: Average number of contact domains (estimated by the arrowhead algorithm of Rao et al. [1]) crossed by interacting and non-interacting enhancer-promoter pairs, grouped by distance bin. The difference in mean number of crossings is significant for each distance bin (two-sample Wilcoxon test) and increases for more distal pairs.

Nature Genetics: doi:10.1038/ng.3539



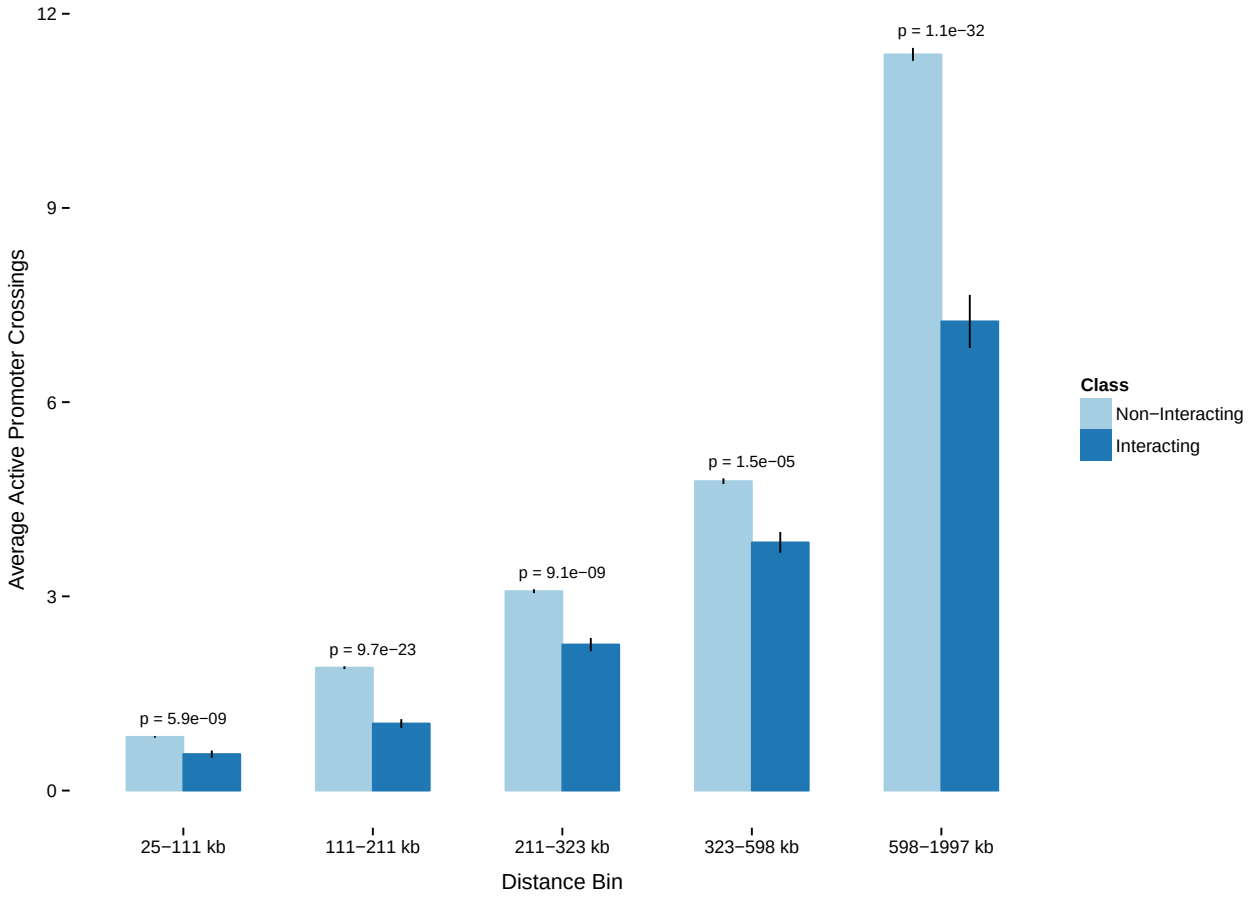
Supplementary Figure 11: Average number of RAD21 peak summits crossed by interacting and non-interacting enhancer-promoter pairs, grouped by distance bin. The difference in mean number of crossings is significant for each distance bin (two-sample Wilcoxon test) and increases for more distal pairs. This suggests looping interactions within the window of a candidate enhancer-promoter pair act as a kind of insulator that decreases the likelihood of interaction.

Nature Genetics: doi:10.1038/ng.3539



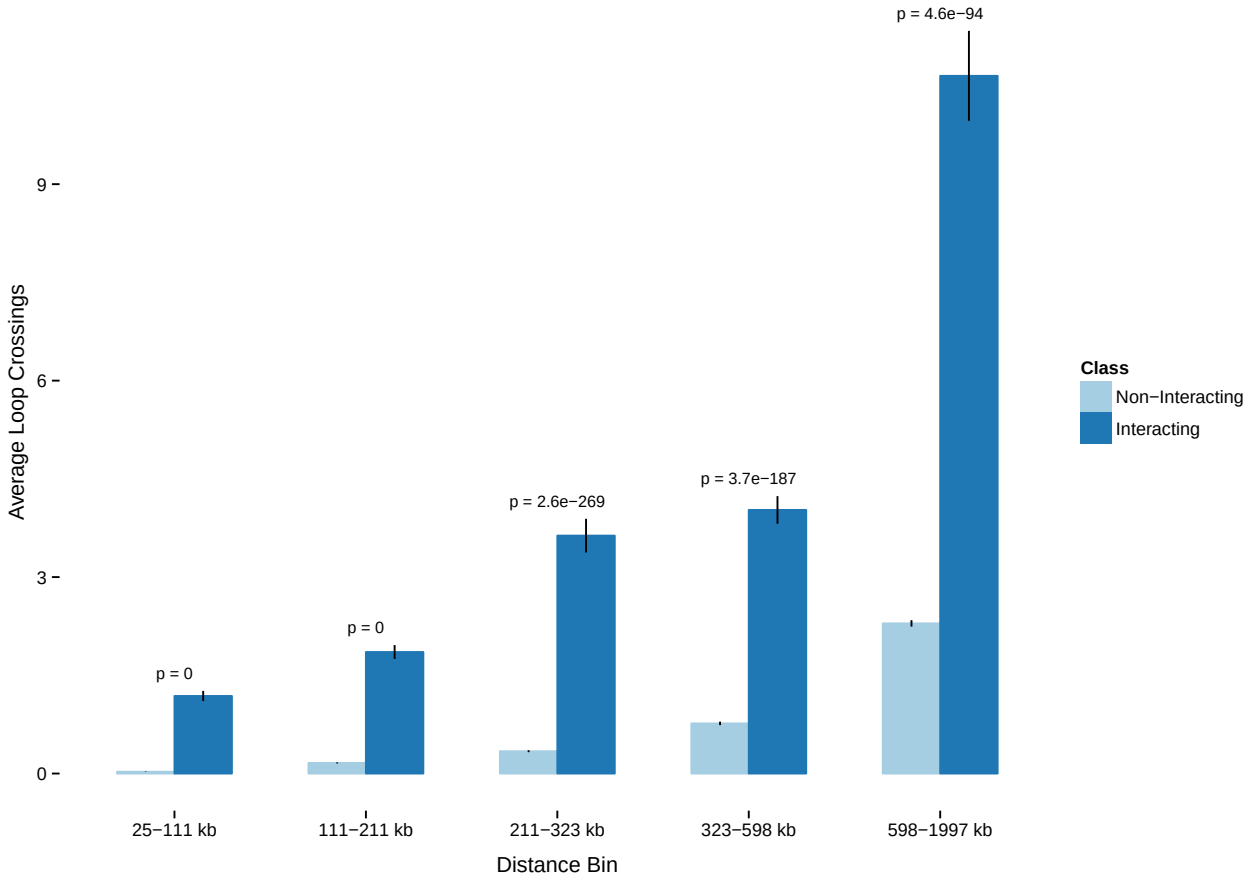
Supplementary Figure 12: Average number of H3K36me3 peaks crossed by interacting and non-interacting enhancer-promoter pairs, grouped by distance bin. The difference in mean number of crossings is significant for all but the first distance bin (two-sample Wilcoxon test) and increases for more distal pairs. This suggests elongation within the window of a candidate enhancer-promoter pair decreases the likelihood of interaction.

Nature Genetics: doi:10.1038/ng.3539



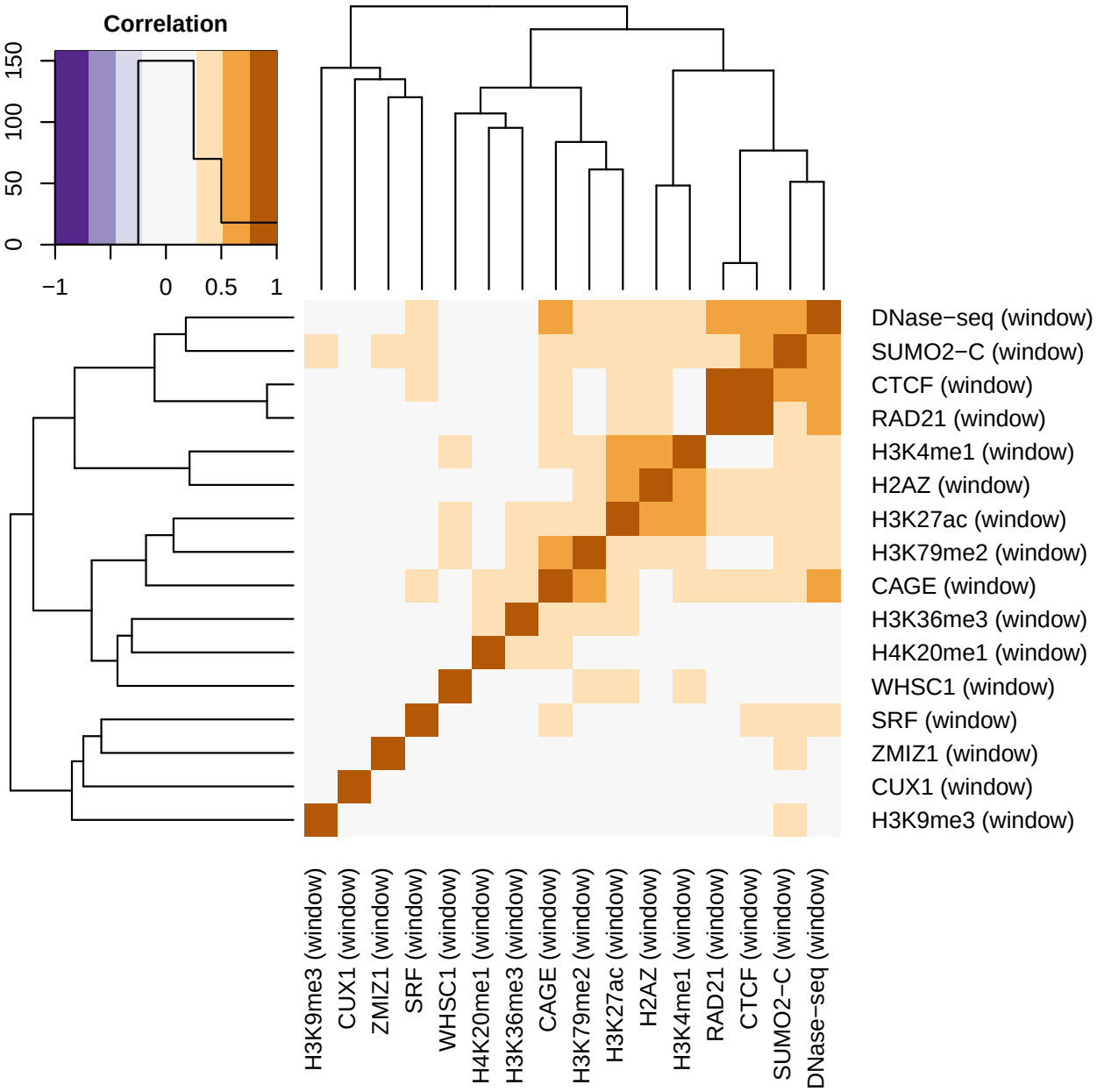
Supplementary Figure 13: Average number of active promoters crossedby interacting and non-interacting enhancer-promoter pairs, grouped by distance bin. The difference in mean number of crossings is significant for each distance bin (two-sample Wilcoxon test) and increases for more distal pairs. Active promoters within the window of a candidate interaction offer alternate targets for the enhancer and thus decrease the likelihood of interaction.

Nature Genetics: doi:10.1038/ng.3539



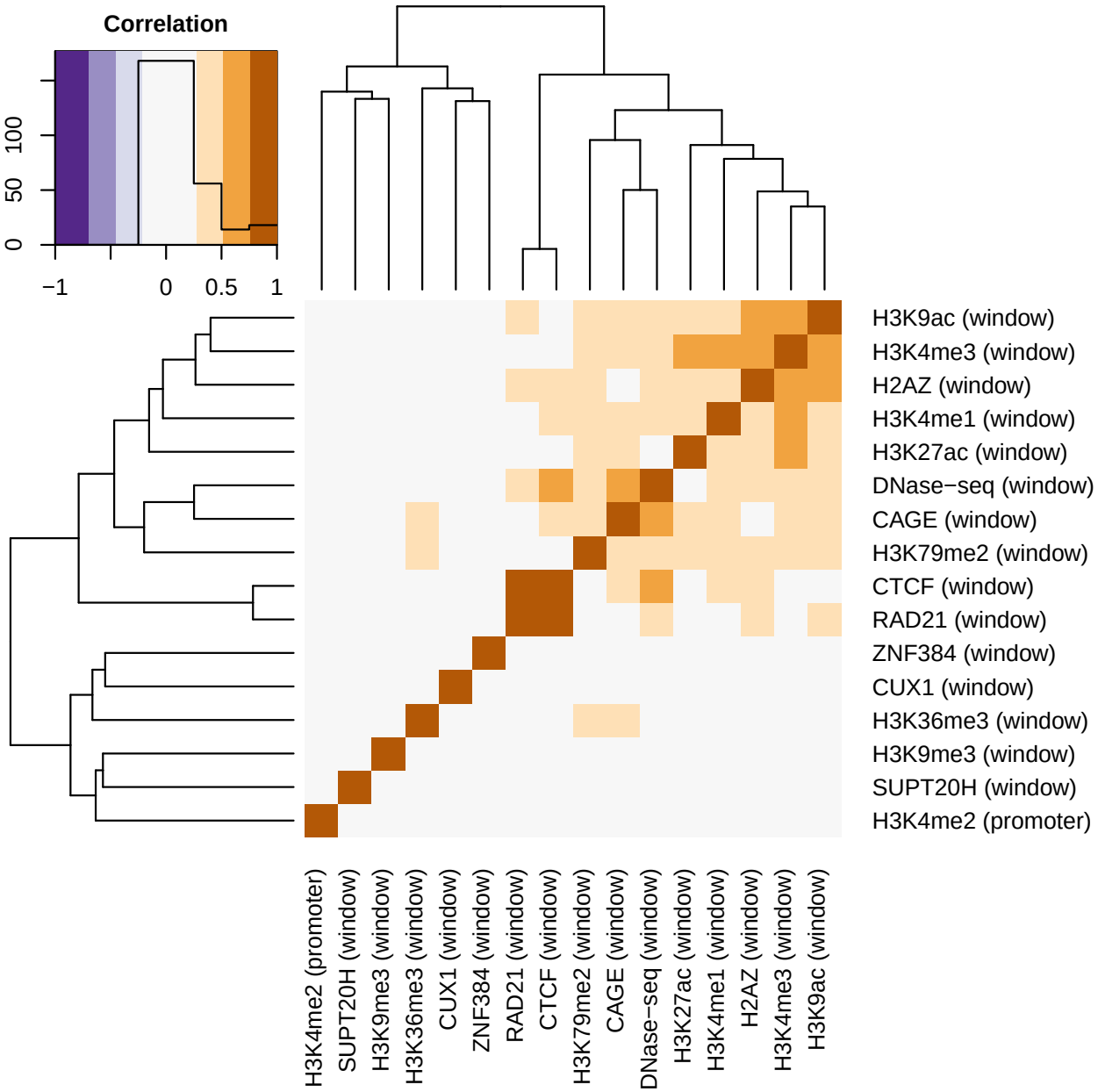
Supplementary Figure 14: Average number of looping interactions crossed by interacting and non-interacting enhancer-promoter pairs, grouped by distance bin. The difference in mean number of crossings is significant for each distance bin (two-sample Wilcoxon test) and increases for more distal pairs. This supports the idea that interactions are enriched within TADs or contact domains and that window features may be a proxy for such domains.

Nature Genetics: doi:10.1038/ng.3539



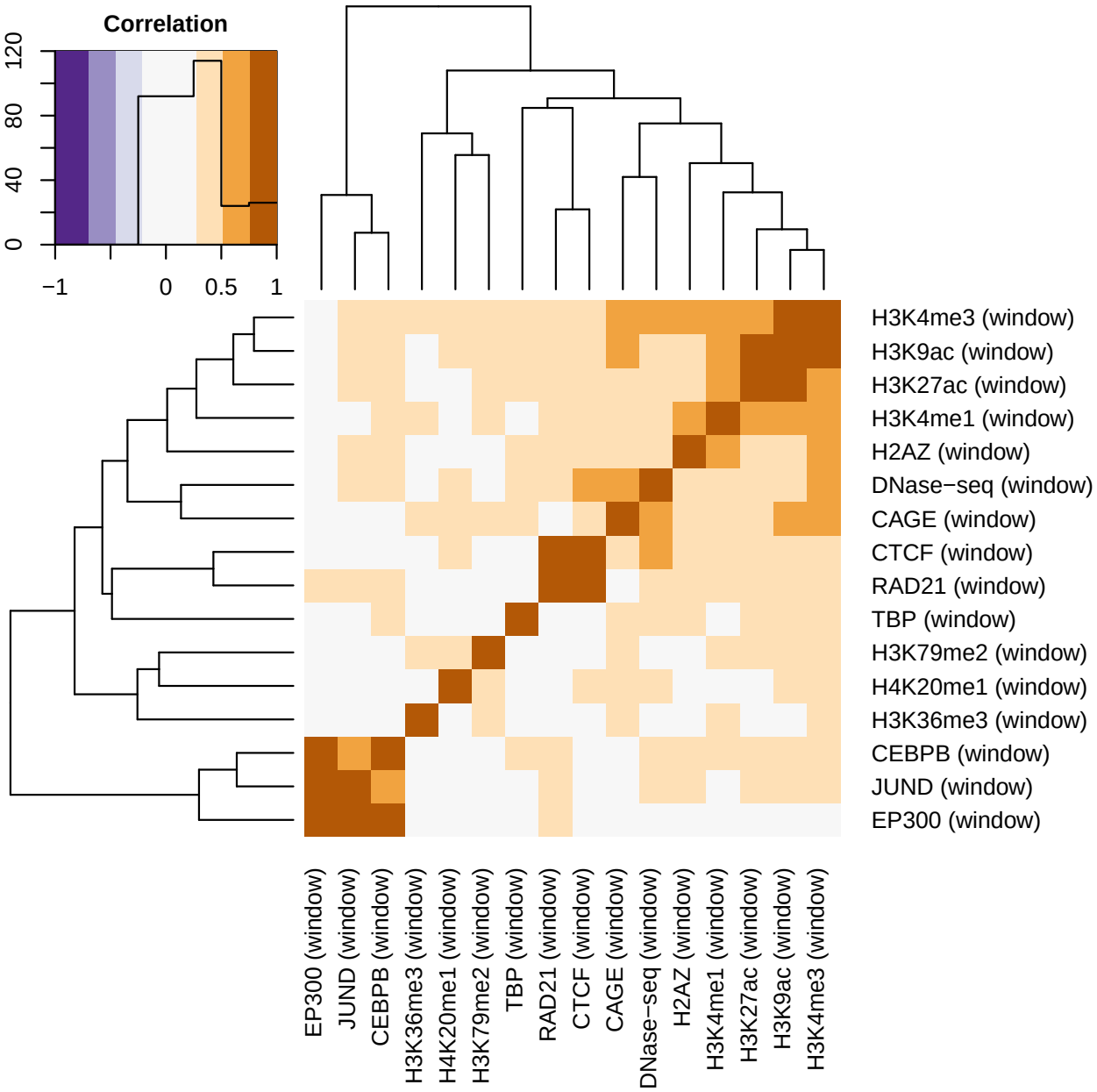
Supplementary Figure 15: Cluster heatmap of the correlation between the top 16 predictive features for K562.

Nature Genetics: doi:10.1038/ng.3539



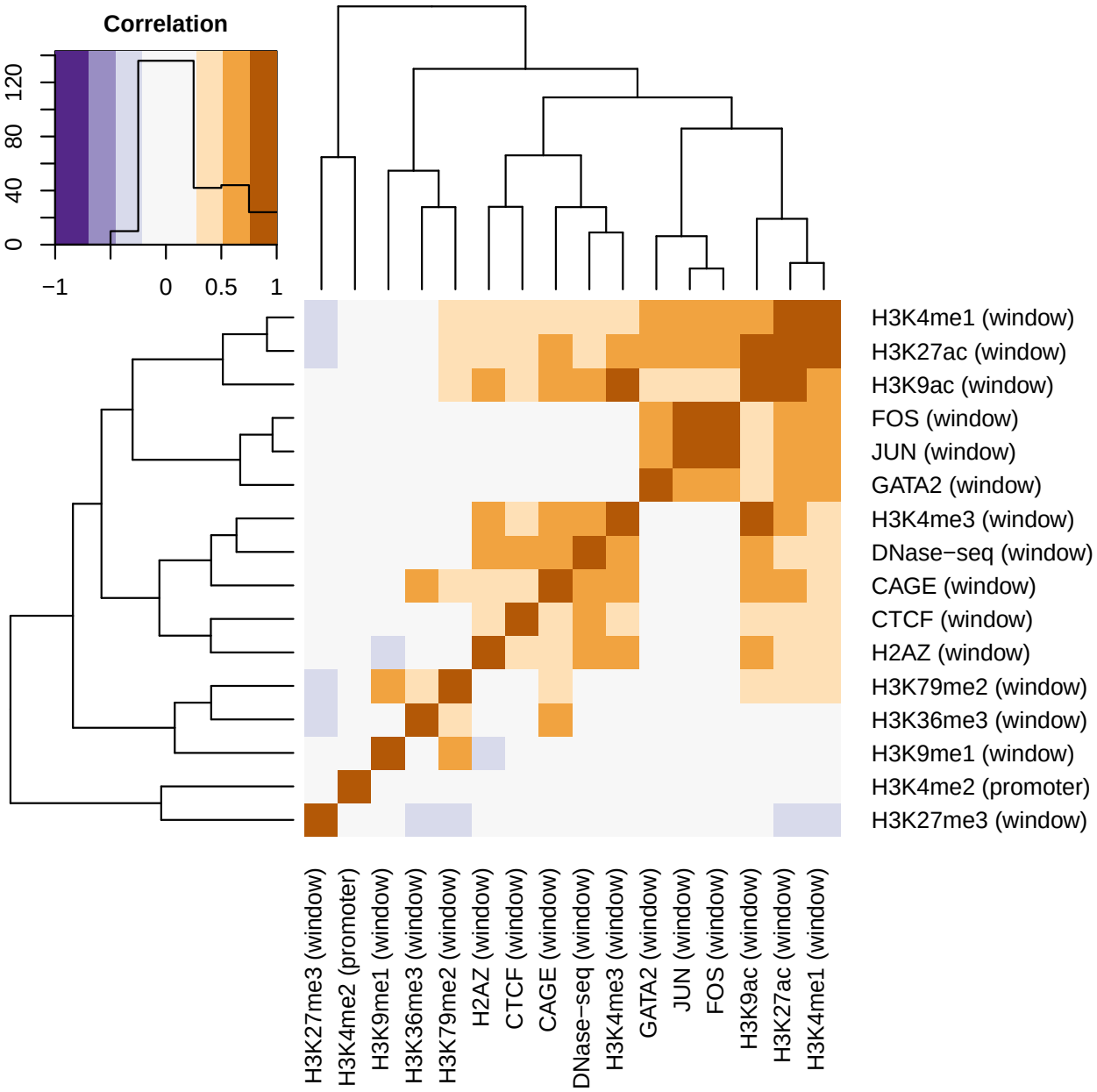
Supplementary Figure 16: Cluster heatmap of the correlation between the top 16 predictive features for GM12878.

Nature Genetics: doi:10.1038/ng.3539



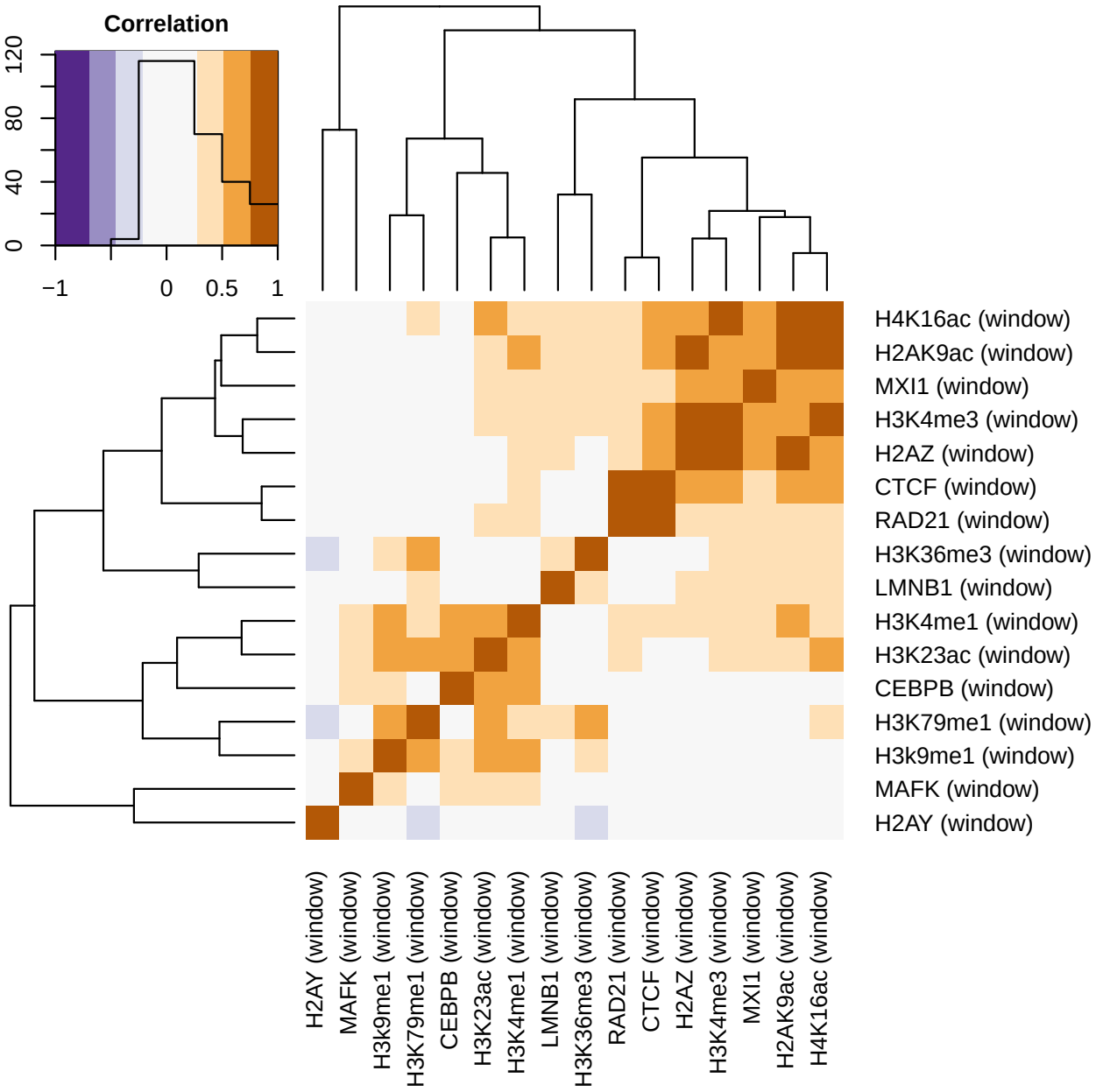
Supplementary Figure 17: Cluster heatmap of the correlation between the top 16 predictive features for HeLa-S3.

Nature Genetics: doi:10.1038/ng.3539



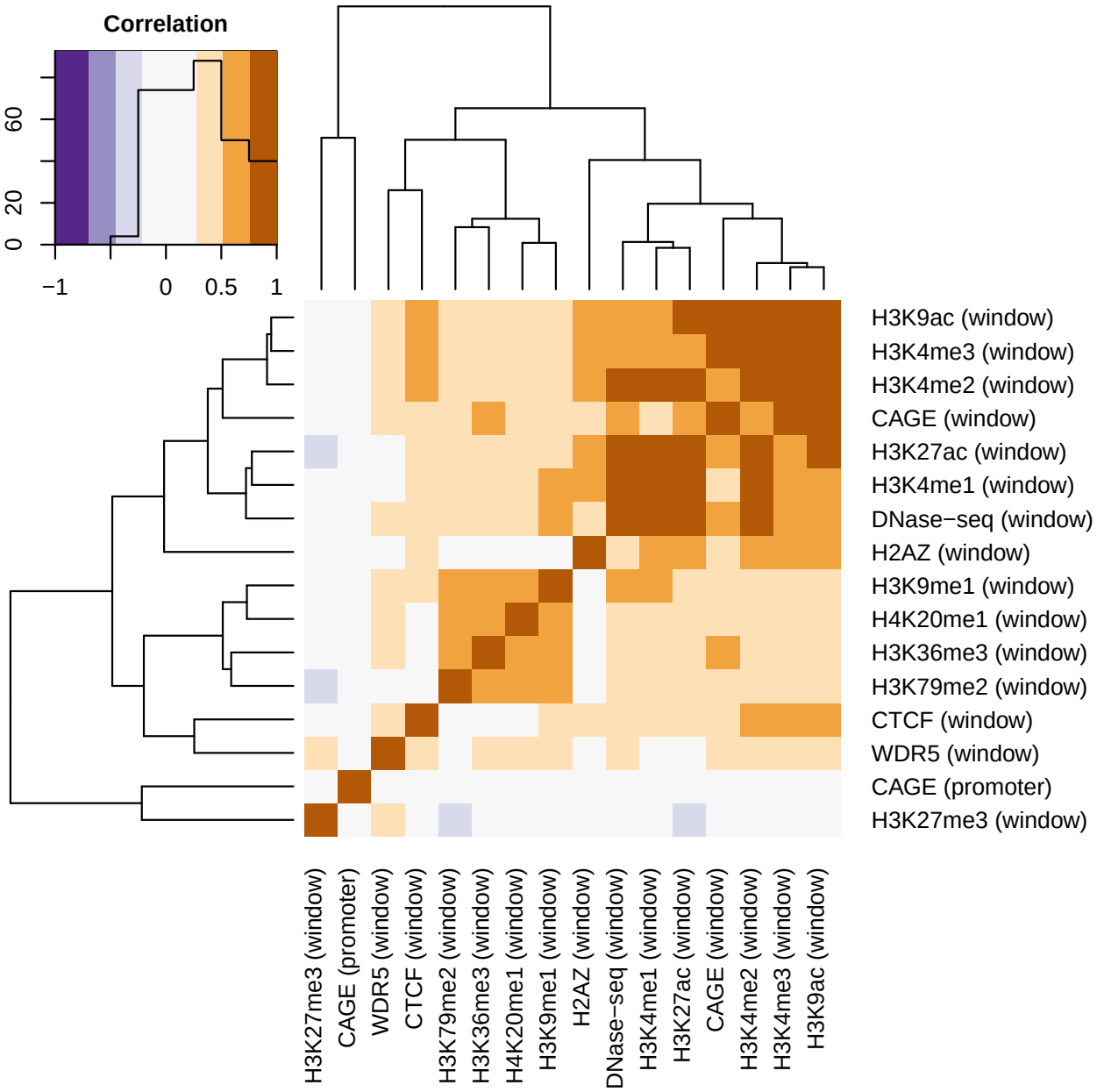
Supplementary Figure 18: Cluster heatmap of the correlation between the top 16 predictive features for HUVEC.

Nature Genetics: doi:10.1038/ng.3539



Supplementary Figure 19: Cluster heatmap of the correlation between the top 16 predictive features for IMR90.

Nature Genetics: doi:10.1038/ng.3539



Supplementary Figure 20: Cluster heatmap of the correlation between the top 16 predictive features for NHEK.

Nature Genetics: doi:10.1038/ng.3539

References

1. Rao, S. S. *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **159,** 1665–80 (2014).
2. Kuhn, M. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software* **28,** 1–26 (2008).
3. Law, A. & Wiener, M. Classification and Regression by randomForest. *R News* **2,** 18–22 (2002).
4. Ridgeway, G. *Generalized boosted models: A guide to the gbm package* 2005.
5. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* **33,** 1–22 (2010).
6. Pique-Regi, R. *et al.* Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Research* **21,** 447–55 (2011).
7. Yan, J. *et al.* Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites. *Cell* **154,** 801–13 (2013).
8. Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. *A density-based algorithm for discovering clusters in large spatial databases with noise* in *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining* (1996), 226–231.
9. Dostie, J. *et al.* Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements. *Genome Research* **16,** 1299–1309 (2006).
10. Wiencke, J. K., Zheng, S, Morrison, Z & Yeh, R.-F. Differentially expressed genes are marked by histone 3 lysine 9 trimethylation in human cancer cells. *Oncogene* **27,** 2412–21 (2008).

Nature Genetics: doi:10.1038/ng.3539