Supplementary Materials

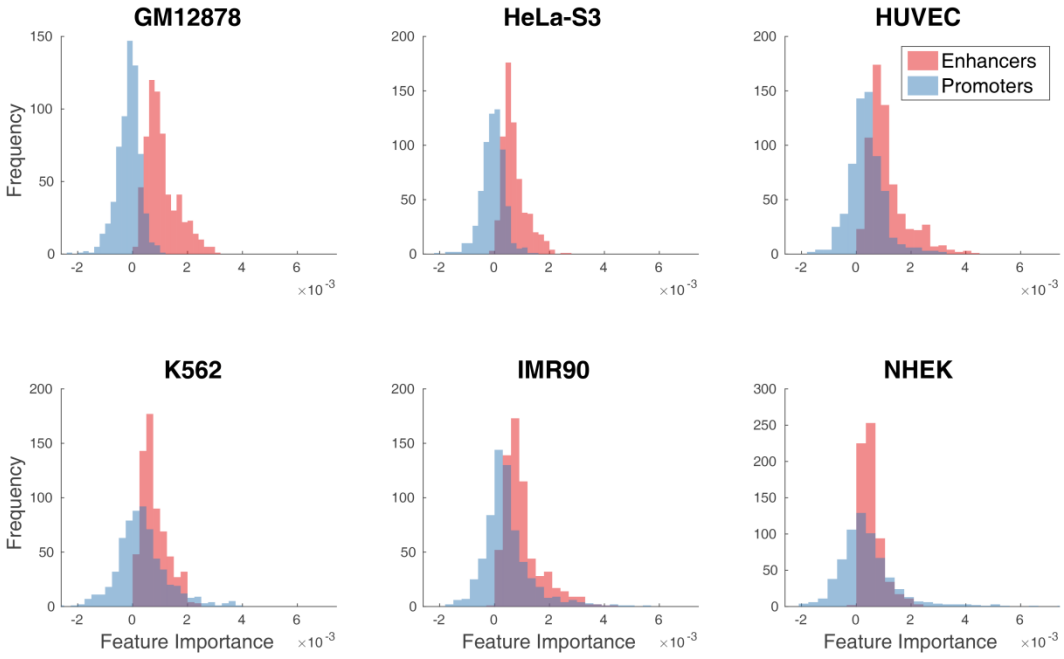

Figure S1: Distribution of feature importance scores for enhancers and promoters in each cell line.

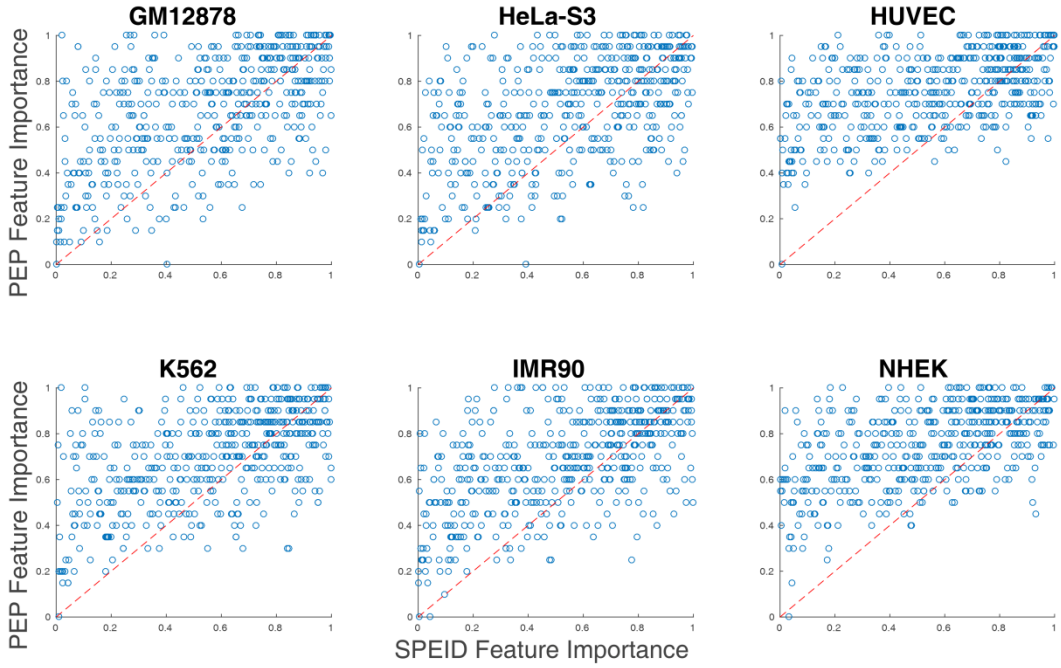

Figure S2: Comparison of feature importance scores between PEP and SPEID. The dotted line indicates the identity.

	Cell Type					
Model	GM12878	HeLa-S3	HUVEC	IMR90	K562	NHEK
SPEID	0.85	0.81	0.75	0.78	0.85	0.94
TargetFinder (E/P)	0.59	0.61	0.48	0.48	0.61	0.83
TargetFinder (EE/P)	0.84	0.83	0.71	0.83	0.81	0.83
TargetFinder (E/P/W)	0.81	0.87	0.77	0.78	0.85	0.90
PEP-Motif	0.83	0.85	0.73	0.80	0.81	0.84
PEP-Word	0.82	0.83	0.74	0.82	0.81	0.86
PEP-Integrate	0.84	0.85	0.75	0.84	0.82	0.88

Table S1: F_1 scores of different EPI prediction methods for each cell line. SPEID and PEP results are reported on E/P/W dataset.

	GM12878	HeLa-S3	HUVEC	K562	IMR90	NHEK
Enhancers						
GM12878	1.00	0.38	0.38	0.26	0.34	0.28
HeLa-S3	0.38	1.00	0.37	0.24	0.35	0.33
HUVEC	0.38	0.37	1.00	0.33	0.37	0.34
K562	0.26	0.24	0.33	1.00	0.31	0.30
IMR90	0.34	0.35	0.37	0.31	1.00	0.29
NHEK	0.28	0.33	0.34	0.30	0.30	1.00
Promoters						
GM12878	1.00	0.19	0.32	0.19	0.21	0.04
HeLa-S3	0.19	1.00	0.18	0.14	0.17	0.03
HUVEC	0.32	0.18	1.00	0.24	0.25	0.10
K562	0.19	0.14	0.24	1.00	0.25	0.16
IMR90	0.21	0.17	0.25	0.25	1.00	0.13
NHEK	0.04	0.03	0.10	0.16	0.13	1.00

Table S2: Rank-correlations between importances of features cross cell lines.

Chromosome	Enhancer Start	Enhancer End	Promoter Start	Promoter End	Number of Patients
chr1	26221800	26222200	26324400	26324800	3
chr1	26222200	26222400	26324400	26324800	3
chr1	26223200	26223400	26324400	26324800	3
chr9	107914000	107914200	107690400	107690600	3
chr9	107917000	107917200	107690400	107690600	3
chr9	107917200	107917600	107690400	107690600	3
chr9	107917600	107917800	107690400	107690600	3
chr16	58055200	58055600	58163200	58163600	6
chr16	58055800	58057200	58163200	58163600	4
chr16	58057200	58057400	58163200	58163600	4
chr16	67701800	67702000	67840400	67841000	4
chr16	67702000	67702400	67840400	67841000	3
chr16	67706800	67707000	67840400	67841000	4
chr16	67707400	67707800	67840400	67841000	3
chr16	67710400	67710600	67840400	67841000	4
chr16	67710600	67710800	67840400	67841000	5
chr19	4575600	4576000	4639200	4639600	3
chr19	6516000	6516200	6737600	6737800	5
chr19	6517200	6517400	6737600	6737800	4
chr19	6517400	6518400	6737600	6737800	3
chr19	6519200	6519600	6737600	6737800	4
chr19	6519600	6519800	6737600	6737800	3
chr20	9958800	9959200	10414800	10415800	3
chr22	42095200	42095800	42342800	42343200	3
chrX	150922000	150922600	151999000	151999600	3
chrX	150922600	150923400	151999000	151999600	3
chrX	151808200	151808400	151999000	151999600	3

Table S3: Positions of the EPIs which are possibly disrupted in at least 3 patients. The position of an EPI includes the chromosome of the enhancer/promoter, the start and end positions of the enhancer, and the start and end positions of the paired promoter. The number of patients where the corresponding EPI is possibly disrupted in also shown.

Cell Line ID	Sample name	Sample group	Epigenome mnemonic
29	E029	HSC & B-cell	BLD.CD14.PC
30	E030	HSC & B-cell	BLD.CD15.PC
32	E032	HSC & B-cell	BLD.CD19.PPC
34	E034	Blood & T-cell	BLD.CD3.PPC
36	E036	HSC & B-cell	BLD.CD34.CC
46	E046	HSC & B-cell	BLD.CD56.PC
50	E050	HSC & B-cell	BLD.MOB.CD34.PC.F
51	E051	HSC & B-cell	BLD.MOB.CD34.PC.M
114	E116	ENCODE2012	BLD.GM12878
121	E123	ENCODE2012	BLD.K562.CNCR

Table S4: Additional JEME [55] cell lines used to evaluate robustness of SPEID to dataset construction. To construct these additional cell lines, we used data from JEME, which collected 127 human cell types, tissue types and cell lines from ENCODE and Roadmap Epigenomics. An elastic net method, was used to predict enhancer-TSS (transcription start site) interactions. We performed filtering based on the confidence score of each connection pair, and removing all pairs with confidence scores less than 0.5. Then, we chose 10 cell lines that have top 10 largest number of enhancer-TSS pairs remained and use them as our EPI positive dataset. Finally, we constructed the negative dataset to satisfy 4 constraints:

1. Promoter variety control (i.e., all promoters that appear in the negative dataset should also appear in the positive dataset)
2. Enhancer variety control (i.e., all enhancers that appear in the negative dataset should appear in the positive dataset)
3. Promoter frequency control (i.e., the occurrence frequency of each promoter in the negative dataset should be the same as that in the positive dataset)
4. Enhancer frequency control (i.e., the occurrence frequency of each enhancer in the negative dataset should be the same as that in the positive dataset)

Cell Line ID	Sample Size	Prediction Accuracy
29	49379	0.557 (+/- 0.0097)
30	43577	0.592 (+/- 0.0101)
32	46905	0.612 (+/- 0.0096)
34	43221	0.596 (+/- 0.0101)
36	38618	0.612 (+/- 0.0106)
46	38775	0.621 (+/- 0.0105)
50	48602	0.556 (+/- 0.0098)
51	47307	0.539 (+/- 0.0100)
114	48055	0.600 (+/- 0.0096)
121	67185	0.597 (+/- 0.0081)

Table S5: Prediction performance of SPEID on additional cell lines constructed from JEME.