










In the format provided by the authors and unedited.

Supervised enhancer prediction with epigenetic pattern recognition and targeted validation

Anurag Sethi ^{1,8}, Mengting Gu ^{2,3,8}, Emrah Gumusgoz⁴, Landon Chan⁵, Koon-Kiu Yan ¹, Joel Rozowsky ¹, Iros Barozzi⁶, Veena Afzal⁶, Jennifer A. Akiyama⁶, Ingrid Plajzer-Frick⁶, Chengfei Yan¹, Catherine S. Novak⁶, Momoe Kato⁶, Tyler H. Garvin⁶, Quan Pham⁶, Anne Harrington⁶, Brandon J. Mannion ⁶, Elizabeth A. Lee⁶, Yoko Fukuda-Yuzawa⁶, Axel Visel ⁶, Diane E. Dickel ⁶, Kevin Y. Yip ⁷, Richard Sutton⁴, Len A. Pennacchio⁶ and Mark Gerstein ^{1,2,3} ✉

¹Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, USA. ²Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA. ³Department of Computer Science, Yale University, New Haven, CT, USA. ⁴Department of Internal Medicine, Section of Infectious Diseases, Yale University School of Medicine, New Haven, CT, USA. ⁵School of Medicine, The Chinese University of Hong Kong, Hong Kong, China. ⁶Functional Genomics Department, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. ⁷Department of Computer Science, The Chinese University of Hong Kong, Hong Kong, China. ⁸These authors contributed equally: Anurag Sethi, Mengting Gu. ✉e-mail: mark@gersteinlab.org

Supplementary Information
A framework for supervised enhancer prediction with epigenetic pattern
recognition and targeted validation across organisms

Methods

Whole-genome prediction of regulatory regions in H1-hESC

To predict enhancers and promoters on the whole genome, we utilized the six-parameter machine learning model shown in Figure 2. The histone and DHS signals from the ENCODE consortium [1] were used to predict enhancers and promoters in H1-hESCs. The histone signals were converted to log-fold enrichment (with respect to control signal) before we scanned it with the matched filter. There were 43,463 active regulatory regions predicted in the human genome (< 2% of genome). All regions within 2kb of any TSS were annotated as promoters, and active regulatory regions that were more than 2kb from any TSS were annotated as enhancers. The distribution of the expression of the closest gene (GENCODE v19 TSS [2]) from the ENCODE RNA-seq dataset for H1-hESCs was compared to the expression of all genes from H1-hESCs. The Wilcoxon test was used to measure the significance of changes in gene expression.

Transgenic mouse enhancer assay

The representative transgenic embryo images for enhancers that displayed reproducible activity are summarized in Extended Data Fig 8. Experimental results are summarized in Supplementary Tables 4-9, where element numbers are the unique identifiers from the VISTA Enhancer Browser (<https://enhancer.lbl.gov>). Details and raw embryo images for each tested element can be found on the browser using the genome coordinate.

All animal work was reviewed and approved by the Lawrence Berkeley National Laboratory Animal Welfare Committee. All mice used in this study were housed at the Animal Care Facility (the ACF) at LBNL. Mice were monitored daily for food and water intake, and animals were inspected weekly by the Chair of the Animal Welfare and Research Committee and the head of the animal facility in consultation with the veterinary staff. The LBNL ACF is accredited by the American Association for the Accreditation of Laboratory Animal Care International.

Assessment with mESC FIREWACH assay peaks

In Extended Data Fig 9, the promoters are defined as FIREWACH peaks within 2kb of any TSS (GENCODE release vM4 [3]); enhancers were defined as FIREWACH peaks more than 2kb from any TSS. The larger distance (2kb) for defining promoters was used because of the larger size of the mouse genome. The FIREWACH assay was performed in a transduction assay and was based on ChIP-seq peaks of a few key TFs. Hence, we did not split the FIREWACH peaks into active and poised enhancers and promoters. The ENCODE histone and DHS datasets for mESCs were used to predict enhancers and promoters.

Enhancer validation experiment

Cell lines

WA01 or H1-hESCs were obtained from WiCell and maintained feeder free on matrigel-coated plates in mTESR1 medium (StemCell Technologies) supplemented with penicillin and streptomycin. All human cell line experiments were conducted in compliance with relevant ethical regulations. Roughly once weekly cell colonies were dissociated using dispase, and absence of differentiation was confirmed by visual inspection and periodically staining cells using anti-SSEA4 conjugated to FITC and by performing flow cytometry. Other cell types (HOS and A549 obtained from ATCC and TZM-bl from the AIDS Reagent Repository) were maintained in DMEM supplemented with 10% fetal calf serum and passaged twice weekly using trypsin-EDTA.

Preparation of HIV vector, cellular transduction, and analysis

Self-inactivating (SIN) HIV vector pFG12 was modified in that the UBC promoter driving eGFP along with the WPRE was removed and replaced with a 1.4 kb IRES-eGFP cassette. Upstream of the IRES, a 142 bp basal Oct 4 promoter (5'-CCTCCCTCTCCTCCACCCATCCAGGGGGCGGGGCCAGAGGTCAAGGCTAGTGGG TGGGACTGGGGAGGGAGAGAGGGGTTGAGTAGTCCCTTCGCAAGCCCTCATTTCAC CAGGCCCGCGCTTGGGGCGCCTTCCTTCCCC-3'; coordinates on chromosome 6, negative strand: 31138398-31138539) was inserted, which overlaps with the TSS of *Oct4* but not with the coding sequence. A unique Xba 1 site was present just upstream of the basal Oct4 promoter, used for cloning of test insert DNA fragments. Each test DNA fragment was amplified from genomic DNA using nested PCR and Takara LA enzyme. Typical initial PCR amplification conditions were 98°C for 10 s, 55°C for 15 s, and 68°C for 3 min for 30 cycles using 100-200 ng of genomic DNA, with the annealing temperature being variable, depending upon the T_m of the primer pair. For the second (internal) round of PCR, only 1-2% of the original product was used under similar PCR conditions, but for 15 cycles.

PCR products were individually cloned into TOPO pCRII-blunt vector (Invitrogen) and insert identity was confirmed by both restriction digests and dideoxy sequencing. All DNA inserts were cloned into the unique Xba 1 site of the HIV vector described above using compatible cohesive ends, and in each case both orientations of the insert within the vector were confirmed by appropriate restriction digests.

HIV vector supernatants were prepared by co-transfecting 35 mm tissue culture wells of 293T cells (~75-80% confluence), each with 5 µg of HIV transfer vector (HIV-TV) with the DNA element of interest, HIV packaging vector, and pME VSV G (encoding Indiana strain VSV G). After 48-72 hours, vector supernatant was harvested, centrifuged at 3000 x g for 10 min, and stored at -80°C until use.

In order to transduce the WA01 hESCs, cells were first lifted using dispase, washed extensively, and plated in the presence of ROCK Inhibitor Y-27632 (StemCell Technologies) on matrigel-coated plates. After a few hours, cells were transduced for 4-6 h with lentiviral vector supernatant. After 48-72 h single cell suspensions were again prepared using dispase and Y-27632 and cells were analyzed for eGFP expression by

flow cytometry, collecting 10,000 events. For all other cell lines, cells were plated the day before in a 12-well format, transduced using the indicated amounts of vector supernatant, refed the following day, and analyzed for eGFP expression 48-72 h later, as described above.

The activity of each element was assessed by flow cytometric readout of eGFP expression 48-72h after transduction. All readouts were recorded as fold changes normalized to the empty SIN HIV vector transduction control. The mean fold change of each tested element in each direction is detailed in Supplementary data 1, where 0 occurred when the number of positive cells was less than that of the negative control according to flow cytometric gating. The experimental results for H1hESC are plotted in Extended Data Fig 12.

We calculated the log fold changes of each replicate. Then, for each element in each direction, we calculated the mean log fold changes as well as the standard deviations. We used t-distribution to test if insertion of one element incur significantly positive log fold changes ($p < 0.05$). The t-distribution is used as a more stringent test for data from normal distribution but of small sample sizes. This was done for both forward and reverse orientations separately and elements that were positive in either orientation were considered to be active. Based on this criteria, we did the calculations in each human cell line.

Code availability

Source code with detailed description of our model is distributed through our website <http://matchedfilter.gersteinlab.org>. We also provided a dockerized image at this site that can be directly downloaded to work in different platforms. In addition, the dataset that we used to train our model is available at the website under the *Training* directory. The metaprofiles for the model is under the *Metaprofile* directory. To test the tool, use the datasets provided in the *sample* directory along with the instructions in the README file. We have also put the genome-wide predictions of regulatory elements in different mouse tissues and human cell lines under the directory *Predicted-Regulatory-Elements*.

Supplementary Table 1 – Performance of matched filter models with single epigenetic feature. The performance of matched filter models from different epigenetic marks for predicting regulatory regions are compared using AUROC and AUPR. The data STARR-seq experiments using multiple core promoters were used to train each matched filter. Due to imbalance in real datasets (more negatives than positives), the performance was compared on data with negatives being 10 times more than positives.

Feature	AUROC	AUPR
H3K27ac	0.95	0.80
H3K4me1	0.70	0.56
H3K4me2	0.90	0.73
H3K4me3	0.82	0.71
H3K9ac	0.92	0.82
H4K12ac	0.86	0.69
H3	0.78	0.64
H1	0.84	0.71
H2BK5ac	0.89	0.76
H4K8ac	0.81	0.61
H4K5ac	0.81	0.63
H4K16ac	0.87	0.74
H3K18ac	0.84	0.68
H3K9me1	0.69	0.55
H3K79me2	0.73	0.49
H3K27me2	0.76	0.58
H2Av	0.60	0.45
H3K27me3	0.77	0.56
H3K23ac	0.65	0.44
H3K79me3	0.66	0.46
H3K27me1	0.58	0.36
H4	0.69	0.49
H3K36me1	0.50	0.35
H3K9me3	0.60	0.43
H3K9me2	0.61	0.41
H3K36me3	0.54	0.34
H4K20me1	0.44	0.29

H3K79me1	0.47	0.30
DHS	0.88	0.79

Supplementary Table 2 – Performance of matched filter models with single epigenetic feature for predicting promoters and enhancers. The performance of matched filter models from different epigenetic marks for predicting enhancers and promoters are compared using AUROC and AUPR. The data STARR-seq experiments using multiple core promoters were used to train each matched filter. Due to imbalance in real datasets (more negatives than positives), the performance was compared on data with negatives being 10 times more than positives. The numbers within (outside) parentheses refer to the accuracy of models for predicting promoters (enhancers).

Feature	AUROC	AUPR
H3K27ac	0.91 (0.96)	0.60 (0.73)
H3K4me1	0.88 (0.60)	0.42 (0.16)
H3K4me2	0.84 (0.92)	0.21 (0.48)
H3K4me3	0.62 (0.92)	0.09 (0.65)
H3K9ac	0.85 (0.94)	0.24 (0.70)
H4K12ac	0.90 (0.93)	0.33 (0.58)
H3	0.78 (0.83)	0.26 (0.48)
H1	0.83 (0.92)	0.36 (0.61)
H2BK5ac	0.91 (0.96)	0.59 (0.70)
H4K8ac	0.90 (0.86)	0.55 (0.37)
H4K5ac	0.89 (0.86)	0.52 (0.41)
H4K16ac	0.90 (0.90)	0.52 (0.40)
H3K18ac	0.90 (0.88)	0.60 (0.47)
H3K9me1	0.53 (0.81)	0.09 (0.44)
H3K79me2	0.70 (0.83)	0.10 (0.27)
H4K27me2	0.68 (0.85)	0.19 (0.44)
H2Av	0.63 (0.78)	0.15 (0.36)
H3K27me3	0.81 (0.86)	0.20 (0.36)
H3K23ac	0.55 (0.71)	0.07 (0.20)
H3K79me3	0.61 (0.74)	0.08 (0.23)
H3K27me1	0.72 (0.57)	0.12 (0.12)
H4	0.69 (0.68)	0.13 (0.21)
H3K36me1	0.75 (0.58)	0.19 (0.18)
H3K9me3	0.59 (0.64)	0.11 (0.15)
H3K9me2	0.62 (0.63)	0.09 (0.15)
H3K36me3	0.60 (0.62)	0.09 (0.14)

H4K20me1	0.55 (0.50)	0.07 (0.10)
H3K79me1	0.54 (0.58)	0.06 (0.12)

Supplementary Table 3 - Summary of predicted mouse regulatory regions in six different tissues. The SVM model is used to predict enhancers in a genome wide fashion using the relevant histone marks and DNAase-Seq datasets measured in each tissue. Regions within 2 kb of a transcript (GENCODE) were labeled as proximal regulatory regions while the rest are predicted to be distal regulatory regions. The numbers within parentheses refer to the percentage of all predictions within that category.

Tissue	Regulatory regions	Distal regulatory regions	Proximal regulatory regions
Forebrain	35,509	24,423 (68.8%)	11,086 (31.2%)
Hindbrain	32,855	22,659 (69.0%)	10,196 (31.0%)
Limb	38,232	26,761 (70.0%)	11,471 (30.0%)
Midbrain	33,451	22,947 (68.6%)	10,504 (31.4%)
Heart	30,739	20,282 (66.0%)	10,457 (34.0%)
Neural Tube	38,933	27,033 (69.4%)	11,900 (30.6%)

Supplementary Table 4 – Transgenic mouse reporter assay results for 103 top human homolog regions. The results from the transgenic mouse assays and the regions tested are displayed. The element number refers to the label of the element in VISTA while the name refers to the label in the ENCODE database. The regions tested were predicted based on predictions in human cell lines.

Element #	Name	hg19 coordinates	Result summary
hs2346	EN202	chr4:23932061-23933692	8/11 eye, 5/11 facial mesenchyme
hs2349	EN205	chr22:47048605-47050100	Negative
hs2353	EN209	chr10:97267716-97269342	4/6 heart
hs2357	EN214	chr1:214280595-214282080	8/11 heart
hs2359	EN216	chr3:42113230-42114717	Negative
hs2371	EN228	chr17:55618678-55620173	Negative
hs2372	EN229	chr2:109252387-109254056	Negative
hs2373	EN230	chr20:43201171-43202669	Negative
hs2374	EN231	chr1:225954390-225955885	4/5 branchial arch
hs2375	EN232	chr17:71287045-71288497	Negative
hs2377	EN234	chr6:163630391-163631925	Negative
hs2378	EN235	chr11:12203825-12205249	Negative
hs2380	EN237	chr20:46012576-46013656	Negative
hs2382	EN240	chr3:186123841-186125332	Negative
hs2384	EN242	chr2:20778294-20779806	10/10 heart, 7/10 ear, 5/10 other
hs2387	EN245	chr7:130012949-130014460	Negative
hs2393	EN251	chr20:17839843-17841338	Negative
hs2394	EN252	chr6:108909808-108911282	Negative
hs2397	EN255	chr6:46020500-46022001	Negative
hs2399	EN257	chr6:43760764-43762277	Negative
hs2400	EN258	chr21:29655315-29656764	Negative
hs2403	EN261	chr11:8753701-8755208	Negative
hs2404	EN262	chr1:203660971-203662806	Negative
hs2405	EN263	chr6:17931980-17933492	Negative
hs2412	EN270	chr4:129278773-129280245	Negative
hs2414	EN272	chr4:47826466-47828052	5/5 heart
hs2415	EN273	chr22:28028233-28029715	Negative
hs2417	EN275	chr4:128406285-128407745	Negative

hs2418	EN276	chr1:92310736-92312231	Negative
hs2419	EN277	chr7:82039621-82041108	12/12 somites; 11/12 limb, 10/12 eye, 9/12 brachial arch
hs2420	EN278	chr10:5627988-5629809	Negative

Supplementary Table 5 - Transgenic reporter assay results for 20 top tier elements of e11.5 mouse forebrain. The results from the transgenic mouse assays and the regions tested are displayed. The element number refers to the label of the element in VISTA while the name refers to the label in the ENCODE database. The regions tested were predicted based on predictions in mouse e11.5 forebrain tissue.

Element #	Name	mm9 coordinates	Result summary
mm1303	mEN351	chr10:60995425-61000401	Negative
mm1304	mEN352	chr15:75477139-75480138	4/7 forebrain
mm1305	mEN353	chr9:121210706-121215001	Negative
mm1332	mEN354	chr4:134812990-134815987	Negative
mm1306	mEN356	chr1:38253589-38258706	Negative
mm1333	mEN357	chr1:40002378-40007534	7/9 forebrain, 7/9 cranial nerve, 7/9 dorsal root ganglion
mm1307	mEN358	chr13:34377263-34382362	3/5 forebrain, 3/5 midbrain, 3/5 hindbrain
mm1308	mEN359	chr4:97313903-97317906	Negative
mm1309	mEN360	chr11:117204339-117209430	8/8 forebrain
mm1310	mEN362	chr12:12714218-12718924	5/6 forebrain, 5/6 midbrain
mm1311	mEN363	chr4:62272177-62276366	Negative
mm1328	mEN366	chr2:101430145-101434498	8/9 forebrain, 8/9 midbrain, 6/9 limb, 6/9 shoulder
mm1312	mEN367	chr2:103464143-103467689	3/5 forebrain, 4/5 hindbrain
mm1334	mEN368	chr13:84921377-84926070	5/10 forebrain
mm1329	mEN369	chr18:34290952-34294024	8/10 nose, 7/10 neck
mm1313	mEN373	chr2:130315050-130319592	3/6 forebrain
mm1316	mEN381	chr6:93768350-93773377	4/9 forebrain, 4/9 midbrain, 4/9 hindbrain
mm1314	mEN382	chr6:91094557-91099332	7/7 forebrain, 7/7 midbrain, 7/7 hindbrain, 4/7 trigeminal V (ganglion, cranial)
mm1315	mEN383	chr16:23502881-23507429	7/8 forebrain, 7/8 hindbrain, 4/8 neural tube
mm1317	mEN388	chr1:99435074-99439318	3/4 forebrain, 3/4 midbrain, 3/4 hindbrain, 3/4 neural tube

Supplementary Table 6 - Transgenic reporter assay results for 22 middle tier elements of e11.5 mouse forebrain. The results from the transgenic mouse assays and the regions tested are displayed. The element number refers to the label of the element in VISTA while the name refers to the label in the ENCODE database. The regions tested were predicted based on predictions in e11.5 mouse forebrain tissue.

Element #	Name	mm9 coordinates	Result summary
mm1336	mEN391	chr8:89675106-89678195	3/4 forebrain
mm1338	mEN395	chr12:5273244-5276374	8/10 ear
mm1339	mEN396	chr16:37812733-37815651	Negative
mm1340	mEN397	chr5:77915965-77918950	4/9 forebrain, 5/9 hindbrain, 7/9 limb
mm1364	mEN400	chr6:112763556-112766918	Negative
mm1365	mEN401	chr3:63673741-63676349	Negative
mm1341	mEN402	chr14:73633763-73636133	6/6 forebrain, 6/6 midbrain, 6/6 hindbrain, 6/6 limb, 3/6 blood vessel
mm1366	mEN403	chr5:119115486-119118887	Negative
mm1348	mEN405	chr11:107623487-107625498	11/11 abdomen
mm1367	mEN406	chr9:95713136-95716028	3/8 midbrain, 5/8 hindbrain, 7/8 ear
mm1368	mEN409	chr2:117252816-117256342	3/6 forebrain
mm1349	mEN410	chr11:77738264-77741018	Negative
mm1369	mEN411	chr1:160195598-160198177	3/4 midbrain, 3/4 hindbrain, 3/4 neck
mm1370	mEN412	chr3:76269644-76273343	7/7 forebrain, 4/7 midbrain, 4/7 hindbrain
mm1371	mEN413	chr9:13502414-13505204	6/6 Hindbrain, 3/6 neural tube
mm1372	mEN414	chr1:75284862-75287747	5/12 forebrain
mm1342	mEN415	chr1:12993828-12996637	Negative
mm1345	mEN420	chr4:24144061-24147950	Negative
mm1346	mEN421	chr2:165845157-165848962	4/5 midbrain
mm1375	mEN424	chr2:168518619-168521392	4/5 hindbrain, 3/5 neural tube
mm1376	mEN425	chr13:12594345-12597146	4/5 forebrain, 4/5 midbrain, 4/5 hindbrain, 4/5 eye, 4/5 neural tube
mm1347	mEN429	chrX:96761917-96764647	5/8 midbrain

Supplementary Table 7 - Transgenic reporter assay results for 20 bottom tier elements of e11.5 mouse forebrain. The results from the transgenic mouse assays and the regions tested are displayed. The element number refers to the label of the element in VISTA while the name refers to the label in the ENCODE database. The regions tested were predicted based on predictions in e11.5 mouse forebrain tissue.

Element #	Name	mm9 coordinates	Result summary
mm1389	mEN432	chr17:4038923-4041381	Negative
mm1406	mEN439	chr9:120511026-120513650	5/8 midbrain
mm1391	mEN440	chr2:132252190-132254838	5/5 forebrain, 4/5 nose, 3/5 heart
mm1398	mEN442	chr5:99701432-99704258	Negative
mm1392	mEN444	chr3:97896495-97899340	Negative
mm1401	mEN445	chr7:142329604-142332301	3/4 forebrain, 3/4 midbrain
mm1393	mEN448	chr12:80896781-80899359	5/7 blood vessels
mm1386	mEN451	chr6:114752658-114755344	Negative
mm1394	mEN453	chr8:118687658-118690168	4/7 facial mesenchyme, 6/7 hindbrain, 7/7 neural tube
mm1402	mEN454	chr2:170661658-170664941	6/12 heart
mm1403	mEN456	chr13:39972062-39974802	6/6 forebrain, 5/6 facial mesenchyme, 6/6 neural tube, 5/6 midbrain, 6/6 hindbrain
mm1390	mEN458	chr18:69706161-69709018	Negative
mm1395	mEN462	chrX:22474415-22477133	10/11 forebrain, 10/11 neural tube
mm1388	mEN463	chr3:51711493-51714567	7/8 forebrain, 7/8 hindbrain, 7/8 eye, 7/8 midbrain, 6/8 heart, 5/8 ear, 5/8 nose, 5/8 brancial arch
mm1396	mEN464	chr7:6803218-6806107	Negative
mm1397	mEN465	chr7:83582528-83585249	3/5 forebrain
mm1399	mEN466	chrX:99180954-99183481	Negative
mm1387	mEN467	chr4:131588028-131592067	4/5 forebrain
mm1405	mEN468	chr7:141869653-141872185	Negative
mm1400	mEN469	chr15:30441205-30443717	5/7 Trigeminal V (ganglion, cranial), 5/7 tail

Supplementary Table 8 - Transgenic reporter assay results for 20 top tier elements of e11.5 mouse heart. The results from the transgenic mouse assays and the regions tested are displayed. The element number refers to the label of the element in VISTA while the name refers to the label in the ENCODE database. The regions tested were predicted based on predictions e11.5 mouse heart tissue.

Element #	Name	mm9 coordinates	Result summary
mm1318	mEN472	chr6:50304038-50307302	Negative
mm1319	mEN473	chr18:5185220-5188223	Negative
mm1330	mEN474	chr13:68710011-68714227	5/7 heart, 3/7 abdomen
mm1320	mEN475	chr7:87263494-87267152	Negative
mm1321	mEN476	chr6:39491754-39496348	3/4 heart, 3/4 nose, 3/4 shoulder
mm1352	mEN478	chr16:32852130-32856370	9/9 heart
mm1353	mEN480	chr6:145403783-145408604	Negative
mm1322	mEN481	chr18:65673857-65677447	5/6 midbrain
mm1323	mEN484	chr11:98762967-98767955	5/7 abdomen
mm1324	mEN485	chr2:84358122-84360960	6/10 abdomen
mm1331	mEN487	chr18:61508433-61511882	Negative
mm1335	mEN488	chr8:81264247-81267464	4/8 heart, 4/8 branchial arch
mm1337	mEN489	chr9:40879715-40882950	3/6 liver
mm1354	mEN492	chr2:33696983-33701358	Negative
mm1355	mEN495	chr1:75401691-75406385	Negative
mm1343	mEN499	chr11:54692427-54697431	4/4 heart
mm1344	mEN500	chr4:57549003-57553035	5/6 heart
mm1325	mEN502	chr2:30860459-30863597	Negative
mm1326	mEN509	chr17:30685485-30689495	4/5 heart
mm1327	mEN510	chr3:121438015-121440547	Negative

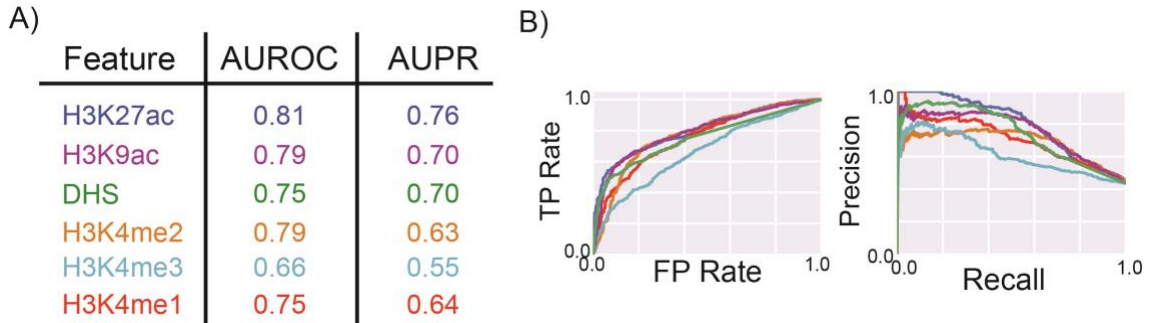
Supplementary Table 9 - Transgenic reporter assay results for 20 middle tier elements of e11.5 mouse heart. The results from the transgenic mouse assays and the regions tested are displayed. The element number refers to the label of the element in VISTA while the name refers to the label in the ENCODE database. The regions tested were predicted based on predictions in e11.5 mouse heart tissue.

element #	Name	mm9 coordinates	Result summary
mm1350	mEN514	chr18:39388954-39391193	Negative
mm1407	mEN515	chr7:116850326-116855192	5/6 heart, 6/6 somite
mm1351	mEN518	chr19:53485525-53487959	7/9 facial mesenchyme, 5/6 ear
mm1377	mEN521	chr2:156639496-156642147	3/5 somite
mm1356	mEN524	chr5:102395744-102399405	Negative
mm1378	mEN526	chr9:21360965-21364026	Negative
mm1357	mEN527	chr1:31158444-31161289	Negative
mm1381	mEN528	chr19:10734265-10738378	Negative
mm1358	mEN530	chr1:68825903-68828605	Negative
mm1379	mEN531	chr7:35050573-35054815	8/9 heart, 8/9 limb, 4/9 eye
mm1380	mEN532	chr2:44909457-44913512	Negative
mm1382	mEN534	chr10:69105930-69109995	Negative
mm1359	mEN535	chr12:81069617-81072879	4/9 heart, 8/9 branchial arch, 5/9 abdomen
mm1360	mEN536	chr3:121735128-121737942	Negative
mm1361	mEN539	chr16:37892230-37895304	7/9 heart, 8/9 forebrain, 9/9 limb, 5/9 blood vessels
mm1384	mEN543	chr11:102911136-102914616	Negative
mm1383	mEN545	chr6:50286189-50288925	Negative
mm1362	mEN546	chr8:11356668-11359383	Negative
mm1363	mEN548	chr1:129651379-129655643	Negative
mm1385	mEN549	chr8:92516582-92519349	3/5 heart

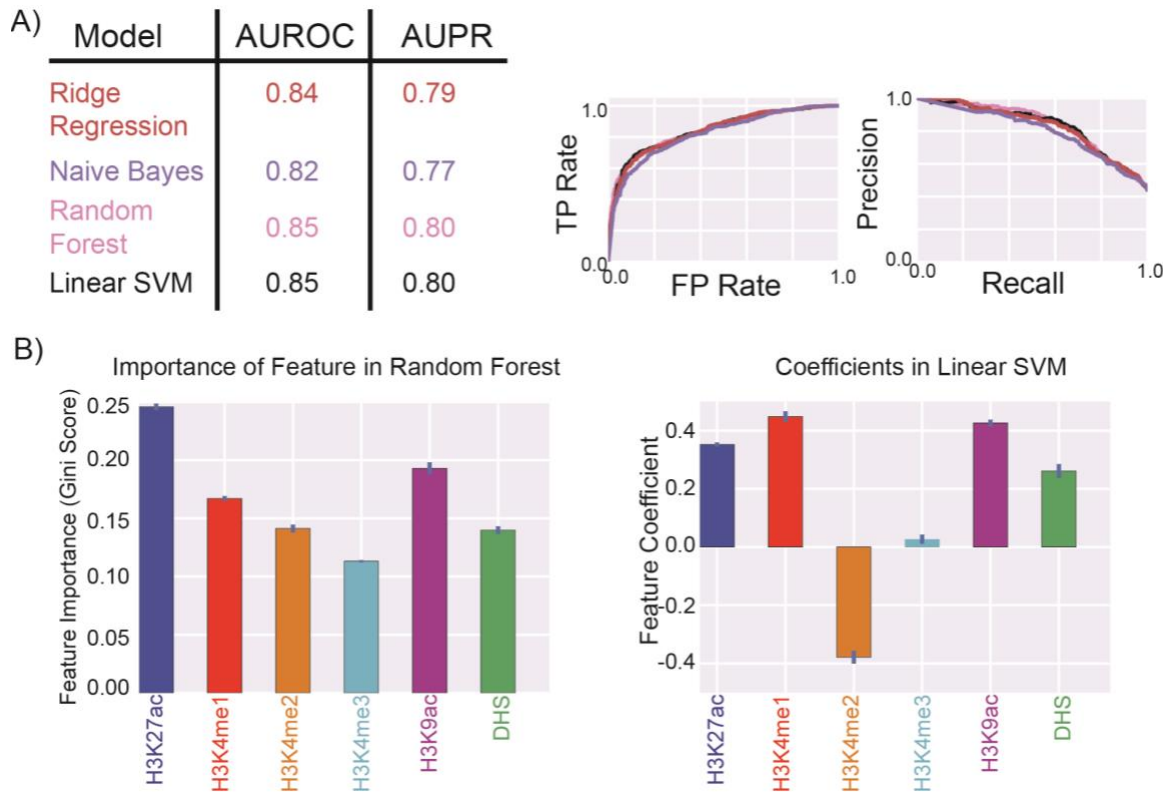
Supplementary Table 10 - Validation results for 20 putative enhancers in four different cell lines. The results from the enhancer validation assay for 20 putative enhancers in four different cell lines are displayed. The activity in different cell lines is also labeled.

Region	H1-hESC	HOS	A549	TZM-bl
chr1:19253310-19254069	Positive	Positive	Positive	Positive
chr2:231809337-231809988	Non-positive	Positive	Non-positive	Positive
chr11:65679112-65679919	Positive	Positive	Non-positive	Positive
chr12:125039037-125040700	Positive	Positive	Non-positive	Positive
chr13:113921562-113922944	Positive	Positive	Non-positive	Positive
chr14:77422602-77423265	Positive	Positive	Non-positive	Non-positive
chr17:2929462-2930394	Non-positive	Non-positive	Non-positive	Non-positive
chr22:31662162-31663116	Non-positive	Positive	Positive	Non-positive
chr1:54839458-54841157	Positive	Non-positive	Non-positive	Non-positive
chr3:128150669-128152511	Positive	Non-positive	Non-positive	Non-positive
chr4:6246837-6247511	Positive	Non-positive	Non-positive	Non-positive
chr7:1956626-1958036	Positive	Non-positive	Non-positive	Non-positive
chr7:73448387-73448811	Positive	Non-positive	Positive	Non-positive
chr9:132976212-132977003	Positive	Non-positive	Non-positive	Non-positive
chr9:138892812-1338893419	Non-positive	Non-positive	Non-positive	Non-positive
chr11:44307337-44308437	Non-positive	Non-positive	Non-positive	Non-positive
chr12:52536500-52539000	Non-positive	Non-positive	Non-positive	Non-positive
chr13:24121112-24121886	Non-positive	Positive	Positive	Non-positive
chr14:75905362-75907344	Non-positive	Positive	Positive	Non-positive
chr18:12271615-12272169	Non-positive	Positive	Positive	Positive
Overall	11/20	10/20	6/20	6/20

Figures and Captions:



Supplementary Figure 1: Training matched filter with all STARR-seq peaks. All STARR-seq peaks without filtering were used in the tenfold cross validation. Following the same procedure, metaprofiles were created on the training data and the matched filter scores are evaluated on the remaining test dataset. A) The performance of each epigenetic mark on all STARR-seq peaks assessed by AUROC and AUPR is shown in the table. B) The ROC and PR curves of each feature are plotted in their corresponding colors.

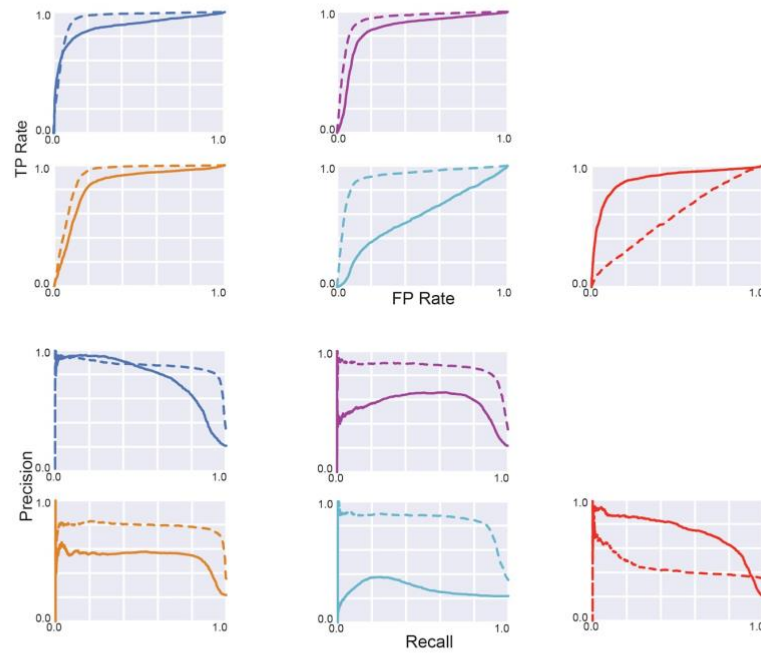


Supplementary Figure 2: Training matched filter with all STARR-seq peaks. The integrated models were trained on all STARR-seq peaks and evaluated by ten-fold cross-validations. A) The performances of the integrated models trained on all STARR-seq peaks are shown, with the ROC and PR curves of each model plotted in their corresponding colors. B) The importance of each feature (measured by the Gini score and the feature coefficient) in the integrated models is shown. Error bars show the standard deviations in the ten-fold cross-validations.

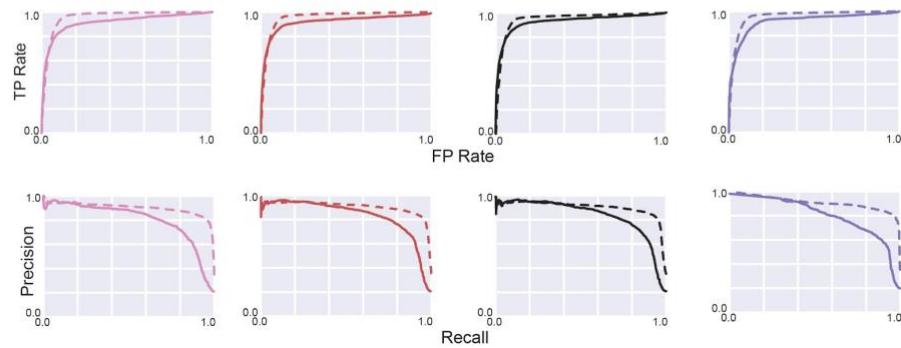
A)

Feature	AUROC	AUPR
H3K27ac	0.88 (0.94)	0.78 (0.87)
H3K9ac	0.86 (0.94)	0.56 (0.86)
H3K4me2	0.84 (0.92)	0.53 (0.79)
H3K4me3	0.58 (0.91)	0.28 (0.84)
H3K4me1	0.89 (0.58)	0.74 (0.44)
Random Forest	0.91 (0.94)	0.81 (0.90)
Ridge Regression	0.93 (0.96)	0.84 (0.90)
Linear SVM	0.92 (0.95)	0.84 (0.90)
Naive Bayes	0.92 (0.96)	0.82 (0.91)

B)



C)

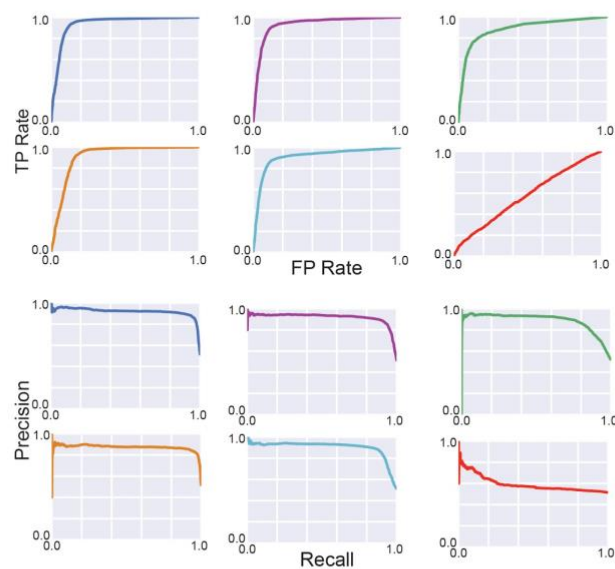


Supplementary Figure 3: Transferability of models across cell lines. The performance of the BG3-trained matched filters of different epigenetic marks and statistical models for predicting S2 cell line active promoters (dashed lines) and enhancers (solid lines) were compared. A) The AUROC and AUPR for each matched filter and statistical model are tabulated. The individual ROC and PR curves for B) each matched filter and C) each statistical model are shown.

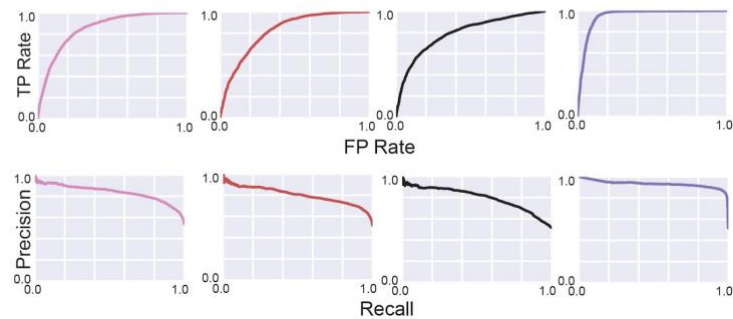
A)

Feature	AUROC	AUPR
H3K27ac	0.94	0.92
H3K9ac	0.93	0.92
DHS	0.89	0.89
H3K4me2	0.91	0.87
H3K4me3	0.91	0.90
H3K4me1	0.57	0.59
Random Forest	0.85	0.84
Ridge Regression	0.82	0.80
Linear SVM	0.79	0.80
Naive Bayes	0.95	0.93

B)



C)

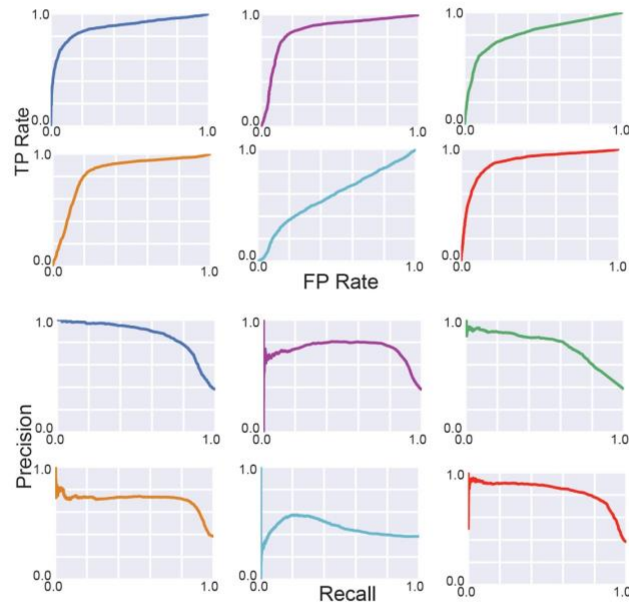


Supplementary Figure 4: Accuracy of enhancer-trained matched filter and statistical models for promoter prediction. The performance of the enhancer-trained matched filters of different epigenetic marks and statistical models for predicting active promoters was compared. A) The AUROC and AUPR for each matched filter and statistical model were tabulated. The individual ROC and PR curves for each matched filter (B) and each statistical model (C) are shown.

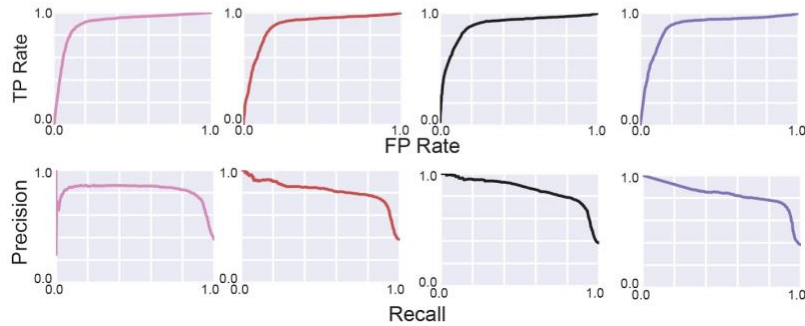
A)

Feature	AUROC	AUPR
H3K27ac	0.88	0.86
H3K9ac	0.86	0.73
DHS	0.82	0.77
H3K4me2	0.83	0.70
H3K4me3	0.58	0.46
H3K4me1	0.89	0.83
Random Forest	0.91	0.82
Ridge Regression	0.89	0.80
Linear SVM	0.90	0.86
Naive Bayes	0.88	0.83

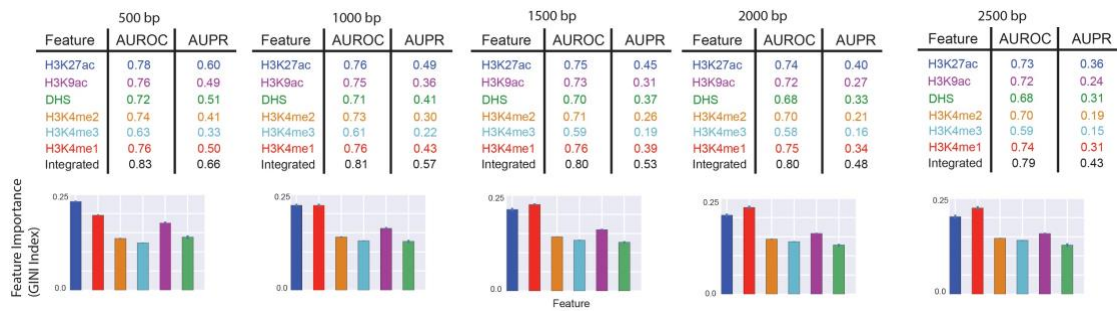
B)



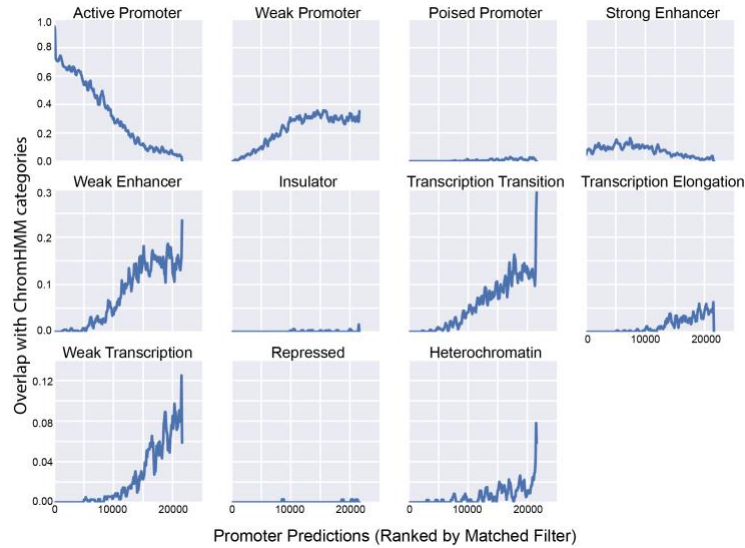
C)



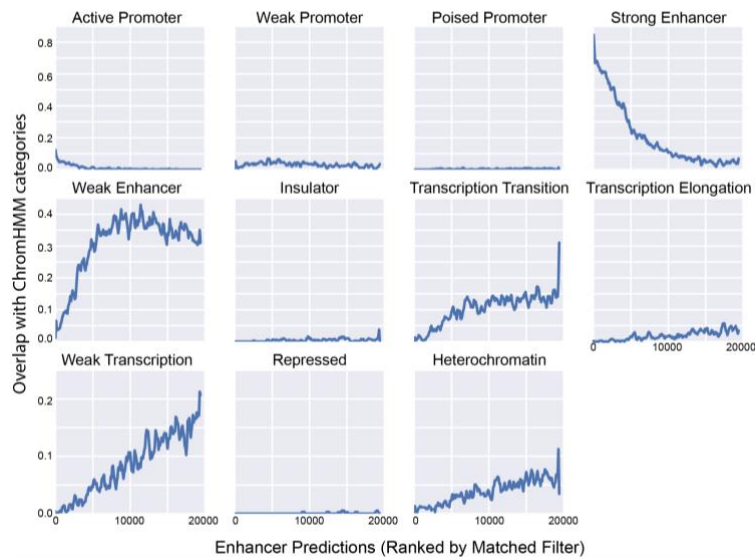
Supplementary Figure 5: Accuracy of promoter-trained matched filter and statistical models for enhancer prediction. The performance of the promoter-trained matched filters of different epigenetic marks and statistical models for predicting active enhancers was compared. A) The AUROC and AUPR for each matched filter and statistical model are tabulated. The individual ROC and PR curves for each matched filter (B) and each statistical model (C) are shown.



Supplementary Figure 6. Accuracy of matched filter predicting enhancers using different distance cutoffs. The STARR-seq peaks are divided into promoters and enhancer by their distance to the nearest TSS. Using varying cutoff of the distance, from 500bp to 2,500bp, performance of each epigenetic mark and the integrated model is shown in the table. The bar plot below each table shows the feature coefficient of each epigenetic mark in the integrated model, with error bars showing the standard deviations of the weights as measured from ten-fold cross-validations.



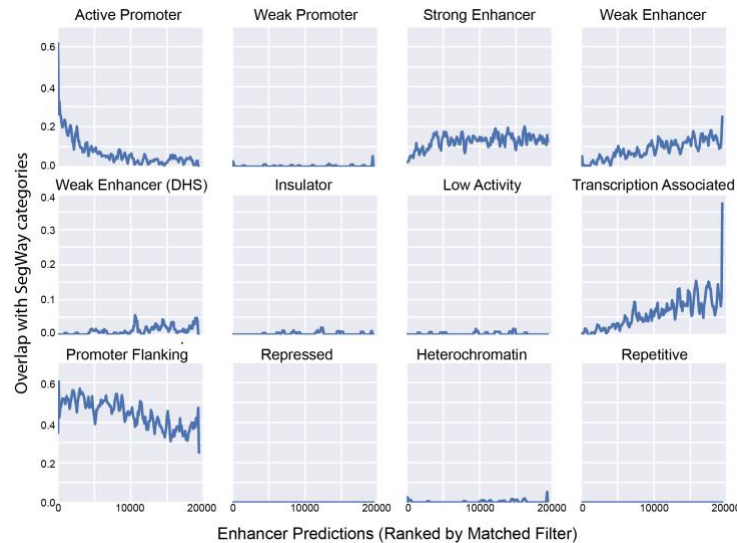
Supplementary Figure 7: Overlap of predicted promoters with chromatin states predicted by ChromHMM. The promoters predicted to be active by matched filter in the H1-hESC cell line were compared with the chromatin states predicted using chromHMM. Most of the matched filter promoters were also predicted to be either strong or weak promoters by chromHMM, whereas some of the other matched filter promoters were labeled as weak enhancers or transcription-related elements in chromHMM. Very few inactive regions and insulators were predicted to be promoters by matched filter. Note that the boundaries of the elements can be very different as chromHMM promoters can be tens of kilobases in length.



Supplementary Figure 8: Overlap of predicted enhancers with chromatin states predicted by ChromHMM. The enhancers predicted to be active by matched filter in the H1-hESC cell line were compared with the chromatin states predicted using chromHMM. Most of the matched filter enhancers were also predicted to be either strong or weak enhancers by chromHMM, whereas some of the other matched filter enhancers were labeled as transcription-related elements in chromHMM. Very few inactive regions and insulators were predicted to be enhancers by matched filter.



Supplementary Figure 9: Overlap of predicted promoters with chromatin states predicted by Segway. The promoters predicted to be active by matched filter in the H1-hESC cell line were compared with the chromatin states predicted using Segway. Most of the matched filter promoters were also predicted to be either active promoters by Segway, whereas some of the other matched filter promoters were labeled as promoter-flanking or transcription-related elements in Segway. Very few inactive regions and insulators were predicted to be promoters by matched filter. Note that the boundaries of the elements can be very different.



Supplementary Figure 10: Overlap of predicted enhancers with chromatin states predicted by Segway. The enhancers predicted to be active by matched filter in the H1-hESC cell line were compared with the chromatin states predicted using Segway. Most of the matched filter enhancers were also predicted to be promoters or enhancers by Segway, whereas some of the other matched filter enhancers were labeled as either promoter-flanking or transcription-related elements in Segway. Very few inactive regions and insulators were predicted to be enhancers by matched filter.

References:

1. Dunham, I., et al., *An integrated encyclopedia of DNA elements in the human genome*. Nature, 2012. **489**(7414): p. 57-74.
2. Harrow, J., et al., *GENCODE: The reference human genome annotation for The ENCODE Project*. Genome Research, 2012. **22**(9): p. 1760-1774.
3. Mudge, J.M., et al., *Creating reference gene annotation for the mouse C57BL6/J genome assembly*. Mammalian Genome, 2015. **26**(9-10): p. 366-378.