PNAS
www.pnas.org

Supplementary Information for

Exploiting regulatory heterogeneity to systematically identify enhancers with high accuracy

Hamutal Arbel[a,b], Sumanta Basu[a,b,e], William W. Fisher[c], Ann S. Hammonds[c], Kenneth H. Wan[c], Soo Park[c], Richard Weiszmann[c], Soile Keranen[c], Clara Henriquez[c], Omid Shams Solari [b], Peter Bickel[1,b,*], Mark D. Biggin[c] ,Susan E. Celniker[1,a,*] and James B. Brown[1,a,b,d,*]

[a]Molecular Ecosystems Biology Department, Division of Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory, Berkeley, CA, USA, 94720
[b]Department of Statistics, University of California, Berkeley, CA, USA, 97420
[c]Biological Systems and Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA, 94720
[d]Centre for Computational Biology, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK
[e]Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY, USA, 14850

[1]To whom correspondence may be addressed. E-mail: JBBrown@lbl.gov, SECelniker@lbl.gov, bickel@stat.berkeley.edu

*Co-senior authors

Short title: Accurate prediction of enhancers

www.pnas.org/cgi/doi/10.1073/pnas.1808833115

This PDF file includes:

Figs. S1 to S6 and associated figure legends
Tables S1 to S3 and associated titles and descriptions

Other supplementary materials for this manuscript include the following:

Dataset S1
Dataset S2

Fig. S1. Roc curves – all active enhancers versus non-enhancers

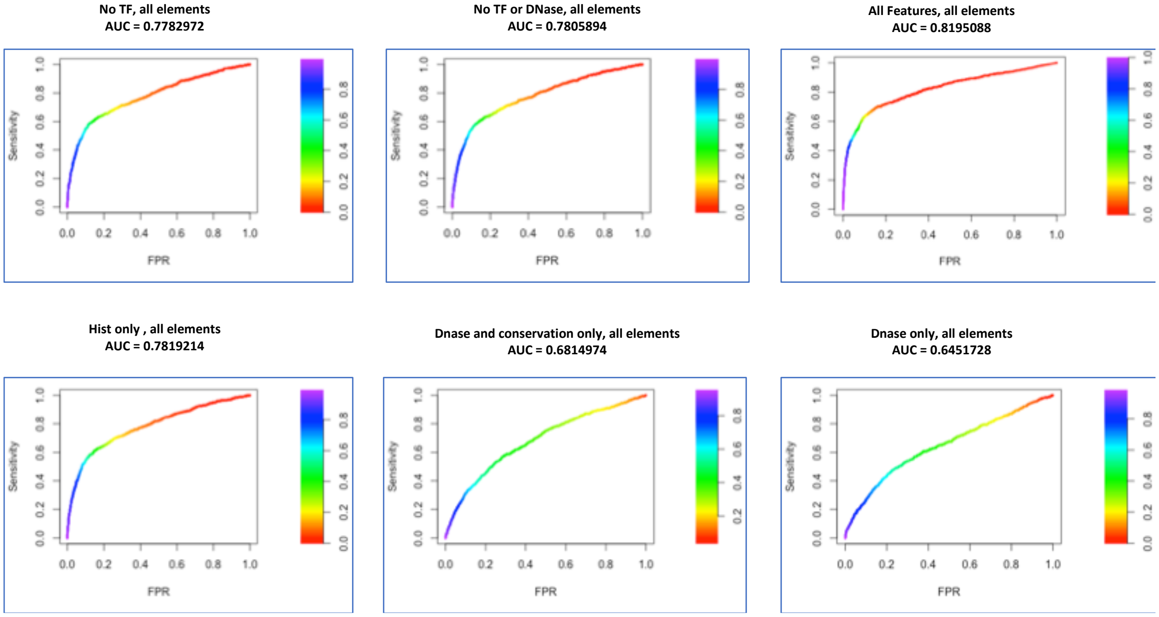


Figure S1: (a) Precision-Recall curves of the three analyses (Random Forest, naïve Bayes, logistic regression) for all elements, class I elements or class II elements. Elements that do not exhibit enhancer activity at stage 5, yet were found to act as enhancers at later stages, are excluded from the analyses. Random Forest PR curve for the data set (dashed blue) shows mediocre performance, with area under the curve (AUC) of 0.72. Predicting class I enhancers, the predictive power rises sharply, AUC = 0.95, while prediction of class II enhancers is close to random guess, with AUC = 0.22. PR curves for logistic regressions almost overlap those of RF, with AUC for all, class I and class II of 0.72, 0.94 and 0.24 respectively. Naïve Bayes performance is poorer, with AUC for all, class I and class II of 0.63, 0.8 and 0.2 respectively.

Fig. S2. Comparison of Machine Learning Methods

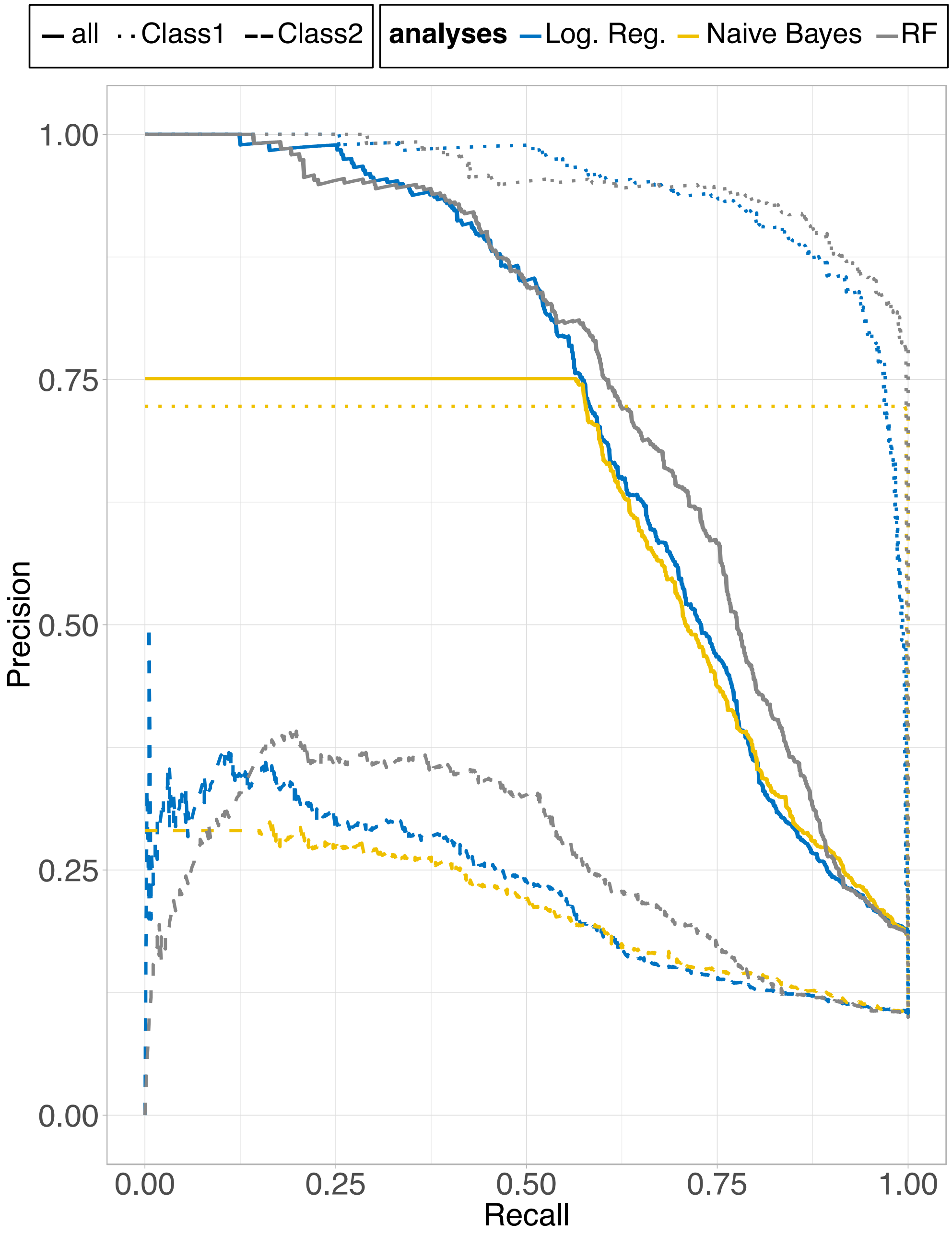
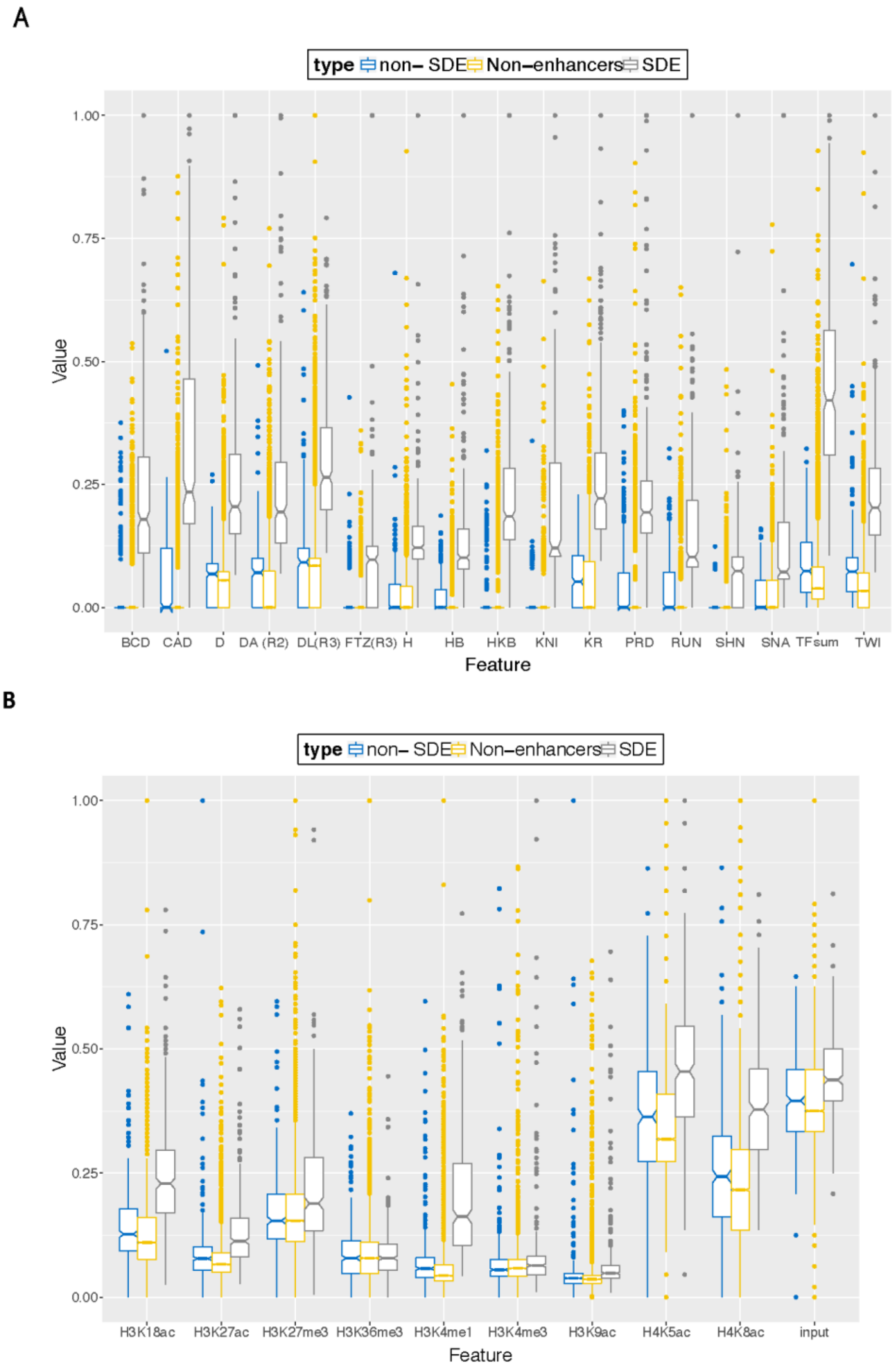


Figure S2: Box plots showing the distribution of features in non-enhancers, class I enhancers and class II enhancers. Sequence specific transcription factors (**a and zld in c**) show the clearest separation of class I from the other enhancers, though a few histone marks (**b**), particularly H3K4me1, also separates class I cleanly. Notably, no feature cleanly separates class II enhancers from non-enhancers. There is no appreciable separation in any of the conservation scores between enhancers and non-enhancers **(c, left)**. DNase accessibility, distance to bidirectional RNA, distance to PolII 2 or distance to transcription start site (**d**) separation between class I and class II enhancers, reinforcing our observation that the data is highly redundant.

Fig. S3. Histograms for SDE and non-SDE enhancers.



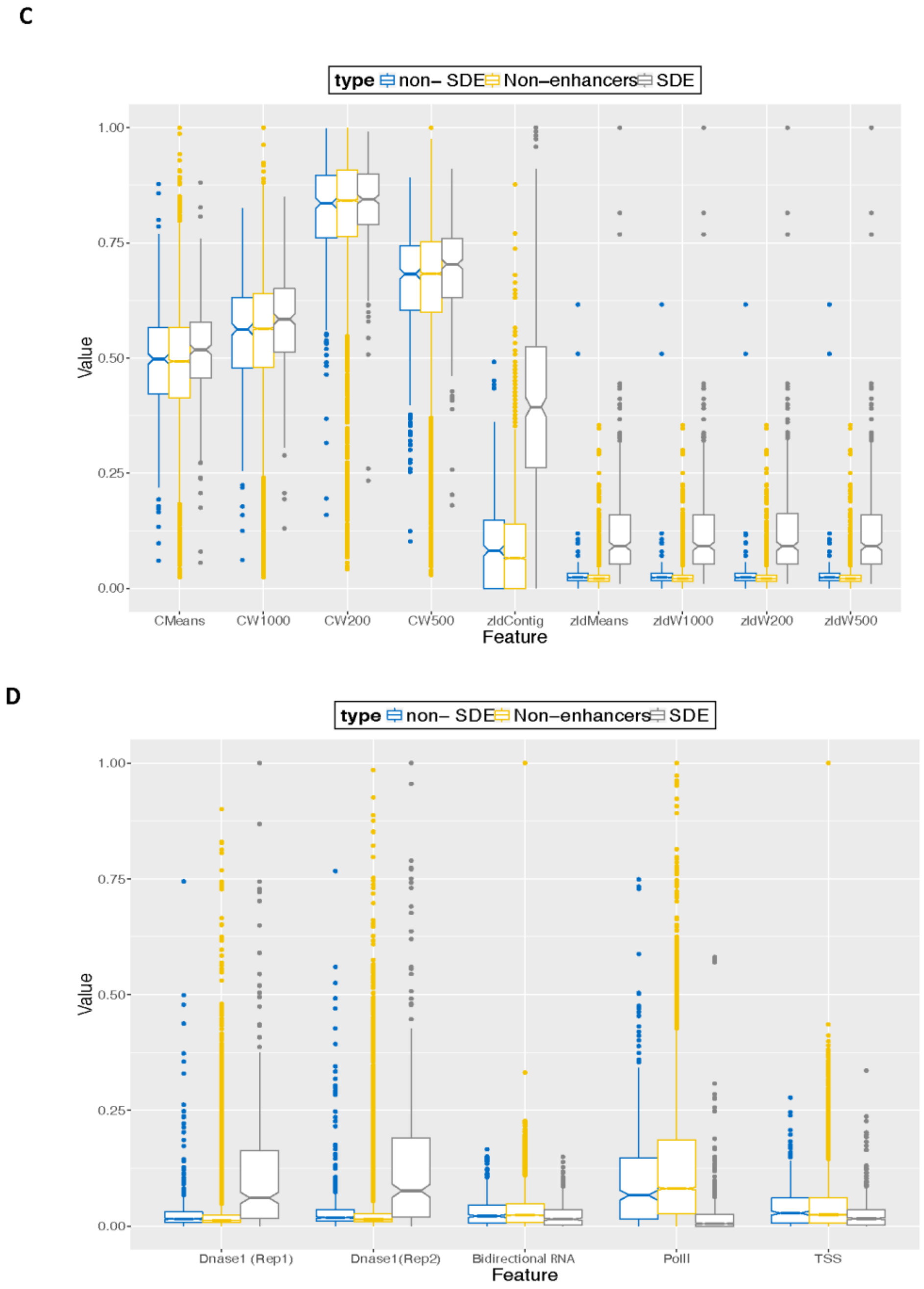
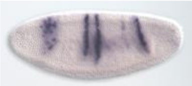
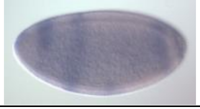
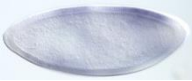
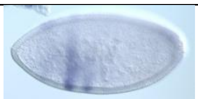
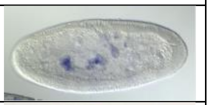
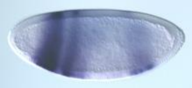
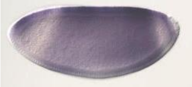
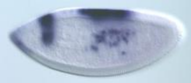
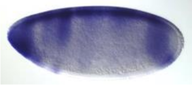
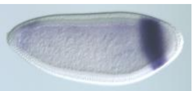
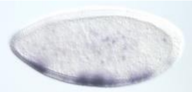
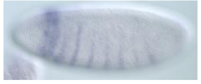
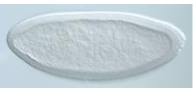
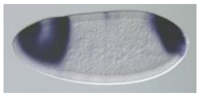
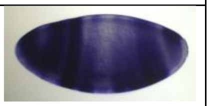
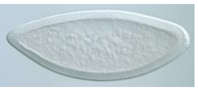
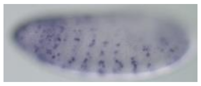
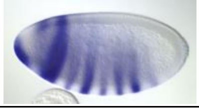
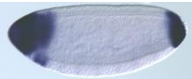
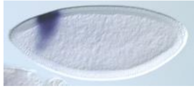
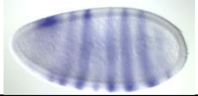
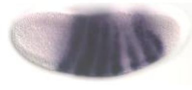
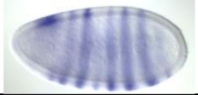
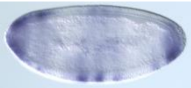


Figure S3: Histogram of expression area for SDE and non-SDE (Class I and Class II enhancers, respectively). (**A**) Transcription factors and related features; (**B**) Chromatin marks and ChIP Input (control); (**C**) Zelda ChIP-seq and controls; (**D**) DNase-seq, CAGE data defining bi-directional promoters, Pol II ChIP-seq data, and CAGE TSS quantifications (including unidirectional transcription start sites).

Fig. S4. Embryonic whole mount *in situ* RNA hybridization validation experiments.

CRM	Lateral	Predicted Regulated Gene	WT gene expression
CEP01101		*salr*	
CEP01102		*N/D*	
CEP01103		*beat-IIIc*	
CEP01104		*dnt*	
CEP01105		*toc*	
CEP01107		*N/D*	
CEP01109		*N/D*	
CEP01110		*N/D*	
CEP01111		*N/A*	
CEP01113		*18w*	
CEP01116		*N/A*	
CEP01117		*trn*	
CEP01118		*N/D*	
CEP01119		*comm2*	
CEP01120		*comm2*	
CEP01121		*N/D*	

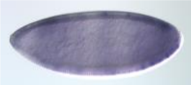
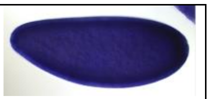
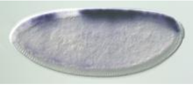
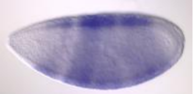
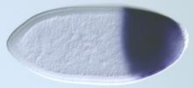
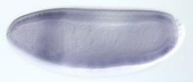
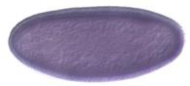
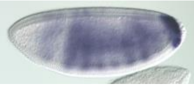
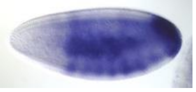
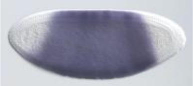
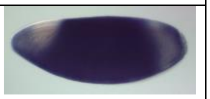
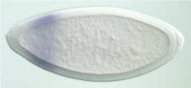
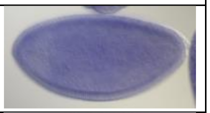
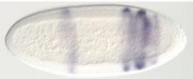
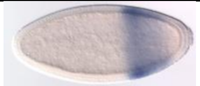
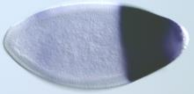
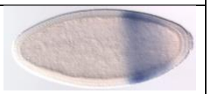
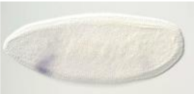
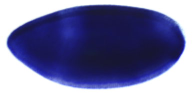
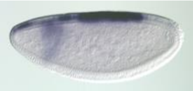
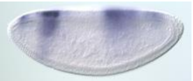
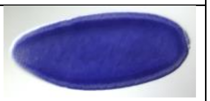
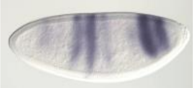
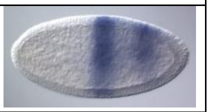
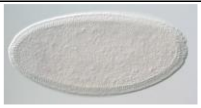
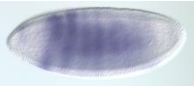
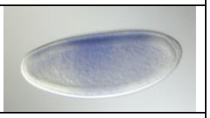
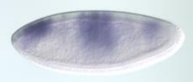
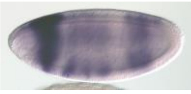
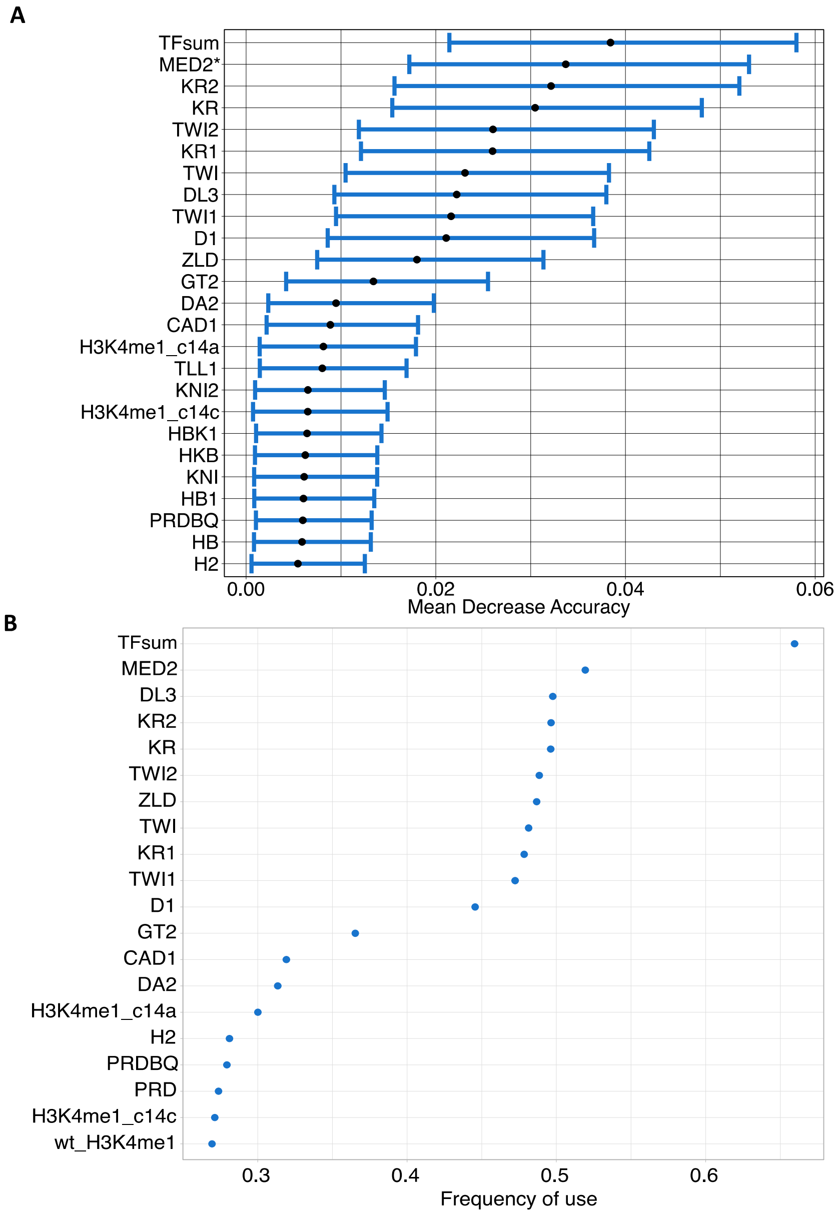
CEP01123		*skd*	
CEP01124		*CG45186*	
CEP01125		*N/D*	
CEP01127		*klar*	
CEP01130		*aay*	
CEP01131		*hth*	
CEP01133		*Akt1*	
CEP01135		*iab-8*	
CEP01136		*iab-8*	
CEP01137		*mira*	
CEP01139		*N/D*	
CEP01140		*heph*	
CEP01143		*Sodh-1*	
CEP01144		*N/A*	
CEP01148		*mew*	
CEP01150		*N/D*	
CEP01151		*N/D*	

Fig. S4: All validation experiments shown along with proximal genes exhibiting similar expression patterns.

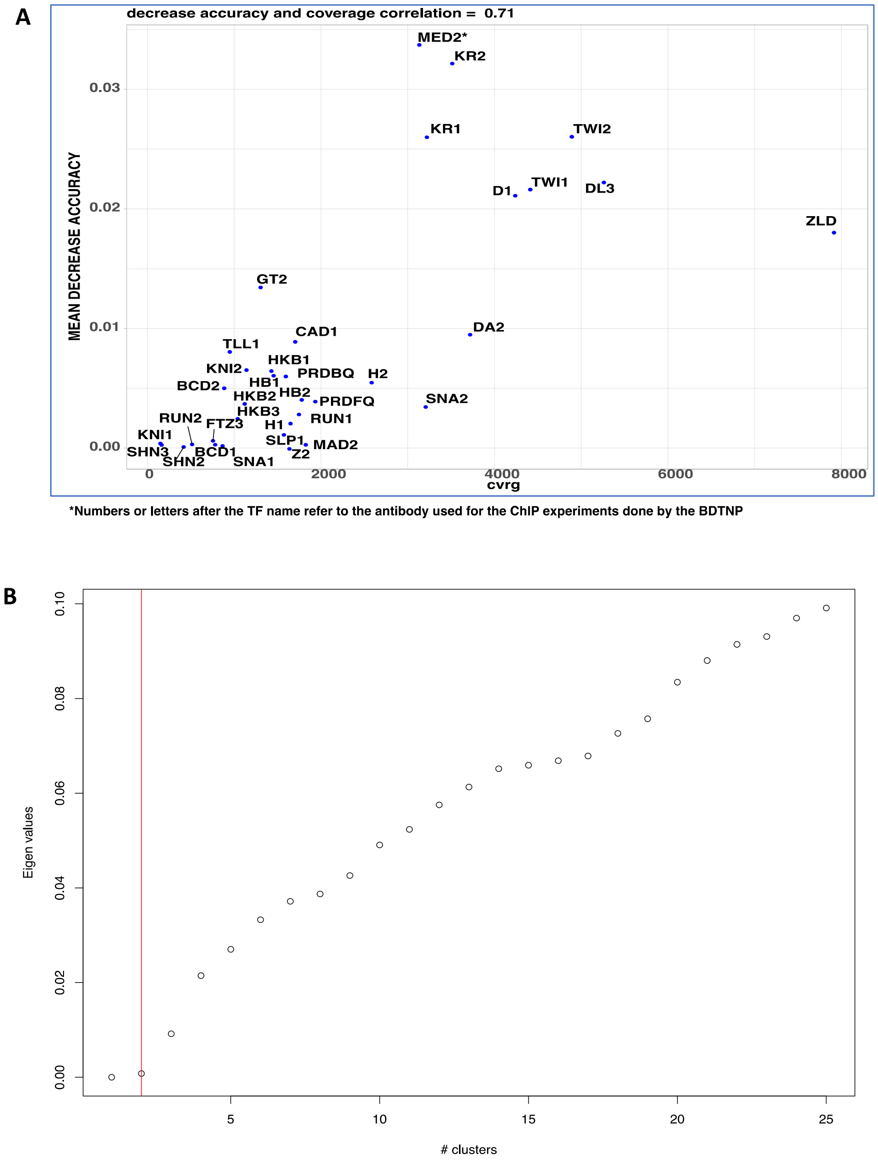
Fig. S5. Feature Importances.



*Numbers or letters following the TF name refer to the antibody used for the ChIP Experiments by the BDTNP

Fig. S5: (A) Feature importance (mean decrease in accuracy upon permutation) is dominated by transcription factors, with the H3K4me1 the only histone mark in the top 25 **(B)** How frequently each of the top 25 features was used by Random Forest in predicting SDE enhancers

Fig. S6. Correlation between feature importance and genome-wide prevalence



*Numbers or letters after the TF name refer to the antibody used for the ChIP experiments done by the BDTNP

Figure S6: (**A**) Correlation between feature importance of transcription factors as measured by mean decrease accuracy and the number of DNA segments in the training set which contain peaks above the 25% FDR of these transcription factors. There are two clusters of transcription factors: low coverage low importance and high coverage high importance. Though there appears to be no correlation inside the clusters, there is an overall correlation of *r* ~ 0.7 between importance and coverage. (**B**) The ordered eigenvalues of the affinity matrix (seven nearest neighbors of Euclidian distance based similarity matrix) of the Random Forests local importance matrix. The jump in value after the second eigenvalue indicates a two-cluster structure

Table S1. All features used in prediction. Transcription factors and DNase number refer to biological replicas

Catergory	Features included
Histone and Histone modifications	H3_c12 H3_c14a H3_c14c H3_c8 H3K18ac_c12 H3K18ac_c14a H3K18ac_c14c H3K18ac_c8 H3K27ac_c12 H3K27ac_c14a H3K27ac_c14c H3K27ac_c8 H3K27me3_c12 H3K27me3_c14a H3K27me3_c14c H3K36me3_c12 H3K36me3_c14a H3K36me3_c14c H3K4me1_c12 H3K4me1_c14a H3K4me1_c14c H3K4me1_c8 H3K4me3_c12 H3K4me3_c14a H3K4me3_c14c H3K4me3_c8 H3K9ac_c12 H3K9ac_c14a H3K9ac_c14c H3K9ac_c8 H4K5ac_c12 H4K5ac_c14a H4K5ac_c14c H4K5ac_c8 H4K8ac_c12 H4K8ac_c14a H4K8ac_c14c H4K8ac_c8 input_c12 input_c14a input_c14c input_c8 wt_H3 wt_H3K18ac wt_H3K4me1
Transcription Factor data	BCD1 BCD2 CAD1 D1 DA2 DL3 FTZ3 GT2 H1 H2 HB1 HB2 HKB1 HKB2 HKB3 KNI1 KNI2 KR1 KR2 MAD2 MED2 PRDBQ PRDFQ RUN1 RUN2 SHN2 SHN3 SLP1 SNA1 SNA2 TLL1 TWI1 TWI2 Z2 ZLD
Transcription factor combinatorics	Sum of all TF, sum of all duplicates for: BCD TWI SNA SHN RUN KR KNI HKB PRD HB H
Conservation scores	Mean, ,Max sliding window of:200, 500 and 1000, longest continuous stretch
ZLD ChiP-seq measurements	Mean, ,Max sliding window of: 200, 500 and 1000, longest continuous stretch
DNA accessibility	DNase1, DNase2
Bi-directional RNA binding	Distance, absolute distance, maximal signal
Exon/intron data	Coding Exons Coverage, All Exons Coverage, Introns Coverage, binary indicators for weather segments contain exons, coding exons or introns
Transcriptional data	Distance to Pol II binding peak, distance to closest transcription start site

Table S2. Results of genomic constructs used as validation set for prediction. St5 expression.1 indicates expression activity observed at stage 5, 0 indicates no expression was observed**.**

Line Name	Prediction Score	FDR range	St5 expression	Arm	Rel 6 Start	Rel 6 End	Rel5 Start	Rel 5 End
CEP01120	0.99282	4%	1	X	13228864	13230982	13122897	13125015
CEP01117	0.99072	4%	1	3R	16864341	16866105	12690063	12691827
CEP01127	0.9894	4%	1	2R	20040447	20042224	15927952	15929729
CEP01103	0.98678	4%	1	3L	2384636	2386548	2384636	2386548
CEP01124	0.97866	25%	1	3L	15700867	15702825	15693967	15695925
CEP01139	0.97716	25%	1	3L	21008165	21010056	21001265	21003156
CEP01143	0.97538	25%	1	3R	31936558	31938611	27762280	27764333
CEP01104	0.9729	25%	1	3L	2879238	2881012	2879238	2881012
CEP01102	0.97246	25%	1	2L	1421000	1422900	1421000	1422900
CEP01107	0.97232	25%	1	2L	3078954	3081088	3078954	3081088
CEP01137	0.96938	25%	1	2L	19326187	19328032	19326187	19328032
CEP01113	0.96856	25%	1	2R	15734639	15736436	11622144	11623941
CEP01125	0.9685	25%	1	3L	15702662	15704639	15695762	15697739
CEP01118	0.96794	25%	0	3L	13076761	13078428	13069861	13071528
CEP01133	0.9651	25%	1	2L	17221300	17223233	17221300	17223233
CEP01130	0.93782	50%	1	3L	16630860	16632285	16623960	16625385
CEP01110	0.93678	50%	1	3L	9421333	9423095	9414433	9416195
CEP01151	0.93162	50%	1	3R	7043578	7044878	2,869,300	2,870,600
CEP01123	0.92818	50%	0	2R	19257755	19259263	15145260	15146768
CEP01136	0.92742	50%	1	X	17745431	17746556	17639464	17640589
CEP01135	0.92714	50%	1	X	17507528	17509231	17401561	17403264
CEP01111	0.9269	50%	1	2L	11343450	11344785	11343450	11344785
CEP01116	0.92332	50%	1	3R	16098827	16100747	11924549	11926469
CEP01144	0.92000	50%	1	2R	10077595	10078595	5,965,100	5,966,100
CEP01150	0.91874	50%	1	3R	19935378	19936578	15,761,100	15,762,300
CEP01101	0.9135	50%	1	3L	489700	491000	489700	491000
CEP01109	0.89746	50%	1	2L	8146372	8147729	8146372	8147729
CEP01119	0.8948	50%	1	3L	13113494	13114703	13106594	13107803
CEP01105	0.89452	50%	0	3R	7162231	7163425	2987953	2989147
CEP01148	0.89386	50%	1	3R	16859478	16860478	12,685,200	12,686,200
CEP01121	0.89286	50%	1	3L	14958837	14959386	14951937	14952486
CEP01140	0.88654	50%	1	3R	29764032	29765191	25589754	25590913

Table S3. ChIP-seq scores for histone marks for each enhancer with low levels H3K27ac. Validated training-set enhancers with a H3K27ac peak binding lower than the median of non-enhancer segments throughout stages 4-6 (cell cycle 12-14), as is the sum of binding in cell cycles 8-14. Maximal binding during cell cycle 8, 12,14a,14c and their sum are shown along with the type of enhancer and the source of enhancer validation data.

	H3K27ac_c12	H3K27ac_c14a	H3K27ac_c14c	H3K27ac_c8	H3K27ac tracks sum	Enhancer type	Validation
Non-enhancer Median	**20**	**17**	**20**	**14**	**78**		
ChIPPCRM5	17	16	13	7	53	SDE	Celniker group
PCE8533	14	15	16	14	59	SDE	Celniker group
ChIPPCRM101	14	16	16	7	53	SDE	Celniker group
VT47178	10	7	11	22	50	SDE	Kvon et al.
PCE8520	15	11	11	0	37	SDE	Celniker group
ChIPPCRM11	15	10	16	14	55	SDE	Celniker group
VT64511	12	13	19	14	58	SDE	Celniker group
VT7078	12	13	17	7	49	SDE	Kvon et al.
VT20119	8	13	18	22	61	SDE	Kvon et al.
PCE8602	10	12	8	14	44	non-SDE	Celniker group
PCE8458	12	17	18	7	54	non-SDE	Celniker group
GMR11C11	17	11	16	14	58	non-SDE	Celniker group
GMR10A07	1	16	10	51	78	non-SDE	Celniker group
VT12768	14	16	16	22	68	non-SDE	Kvon et al.
VT14329	14	16	10	22	62	non-SDE	Kvon et al.
VT14347	14	16	15	7	52	non-SDE	Kvon et al.
VT14726	13	15	15	22	65	non-SDE	Kvon et al.
VT16984	14	12	13	14	53	non-SDE	Kvon et al.
VT19752	14	6	8	22	50	non-SDE	Kvon et al.
VT19895	8	11	6	29	54	non-SDE	Kvon et al.
VT22261	13	13	19	14	59	non-SDE	Kvon et al.
VT23797	19	11	15	22	67	non-SDE	Kvon et al.
VT24637	8	12	14	14	48	non-SDE	Kvon et al.
VT25922	17	12	20	14	63	non-SDE	Kvon et al.
VT26006	20	16	16	14	66	non-SDE	Kvon et al.
VT26012	19	13	15	14	61	non-SDE	Kvon et al.
VT26785	9	15	20	14	58	non-SDE	Kvon et al.
VT27271	12	10	18	29	69	non-SDE	Kvon et al.
VT27272	10	10	13	7	40	non-SDE	Kvon et al.
VT3477	20	16	11	14	61	non-SDE	Kvon et al.
VT35631	0	0	0	0	0	non-SDE	Kvon et al.
VT37817	10	6	6	36	58	non-SDE	Kvon et al.
VT38780	13	13	18	26	70	non-SDE	Kvon et al.
VT39428	4	2	5	0	11	non-SDE	Kvon et al.
VT40566	12	12	15	36	75	non-SDE	Kvon et al.
VT40770	19	13	18	14	64	non-SDE	Kvon et al.
VT41895	14	17	16	14	61	non-SDE	Kvon et al.
VT45119	8	10	18	7	43	non-SDE	Kvon et al.
VT45642	17	8	15	7	47	non-SDE	Kvon et al.
VT45997	3	15	15	14	47	non-SDE	Kvon et al.
VT48569	15	8	20	22	65	non-SDE	Kvon et al.

Table S3, cont.

	H3K27ac_ c12	H3K27ac_ c14a	H3K27ac _c14c	H3K27ac_ c8	H3K27ac tracks sum	Enhancer type	Validation
Non-enhancer Median	**20**	**17**	**20**	**14**	**78**		
VT48827	13	16	15	7	51	non-SDE	Kvon et al.
VT4905	13	10	18	14	55	non-SDE	Kvon et al.
VT4990	8	15	13	22	58	non-SDE	Kvon et al.
VT50230	18	11	15	29	73	non-SDE	Kvon et al.
VT50245	16	17	20	22	75	non-SDE	Kvon et al.
VT56665	4	15	14	7	40	non-SDE	Kvon et al.
VT57294	18	10	6	14	48	non-SDE	Kvon et al.
VT57365	9	8	5	14	36	non-SDE	Kvon et al.
VT57463	6	15	15	14	50	non-SDE	Kvon et al.
VT58480	15	10	12	29	66	non-SDE	Kvon et al.
VT58863	17	13	16	14	60	non-SDE	Kvon et al.
VT60196	14	8	11	14	47	non-SDE	Kvon et al.
VT63194	10	16	16	22	64	non-SDE	Kvon et al.
VT9682	6	11	16	7	40	non-SDE	Kvon et al.
VT28267	20	15	3	36	74	non-SDE	Kvon et al.
VT59438	18	17	18	7	60	non-SDE	Kvon et al.
VT56875	18	15	13	7	53	non-SDE	Kvon et al.
VT56791	17	11	13	22	63	non-SDE	Kvon et al.
VT64886	10	17	18	7	52	non-SDE	Kvon et al.
GMR25F01	8	8	5	14	35	non-SDE	Celniker group
VT57075	9	8	6	7	30	non-SDE	Kvon et al.

Additional Dataset S1 (separate file)

Complete list of genomic segments used in this work, by categories.

Additional Dataset S2 (separate file)

Complete list of genome-wide enhancer predictions.