Supplementary Information for

ChromActivity: Integrative epigenomic and functional characterization assay based annotation of regulatory activity across diverse human cell types

Tevfik Umut Dincer and Jason Ernst

Supplementary Figures

Fig. S1: Color legends for ChromScoreHMM states and the ChromHMM imputed 25-state model.

Fig. S2: Genomic coverage of functional characterization datasets.

Fig. S3: Predictive performance of ChromScore expert models in held-out loci of the same functional characterization dataset.

Fig. S4: Genomewide score correlations for selected pairs of expert predictions after removing loci assigned to individual chromatin states or combinations of chromatin states.

Fig. S5: Mean normalized expert scores per chromatin state.

Fig. S6: Boxplots of mean normalized expert scores for each chromatin state, distributions across cell types.

Fig. S7: ChromScoreHMM emission and transition parameters for the 15-state model.

Fig. S8: ChromScoreHMM emission parameters and chromatin mark peak enrichments for the 15-state model.

Fig. S9: ChromScoreHMM emission parameters and mean model scores for the 15-state model.

Fig. S10: ChromScoreHMM state proportions around transcription start sites.

Fig. S11: ChromScoreHMM gene expression patterns around transcription start sites.

Fig. S12: ChromScoreHMM overlap enrichments around TSSs of high expression and low expression genes.

Fig. S13: Cumulative transcription factor binding peak enrichments and genome coverage for ChromScoreHMM and ChromHMM.

Fig. S14: Predictive performance of ChromScore, external models and baselines.

Fig. S15: Predictive performance of ChromScore and individual experts.

Fig. S16: Predictive performance of ChromScore, external models and baselines in CRISPR and plasmid-based functional characterization datasets.

Fig. S17: Spearman correlations between ChromScore and chromatin mark signal tracks.

Fig. S18: Chromatin state fold enrichments of top ChromScore and DNase-I hypersensitive regions.

Fig. S19: Cell type generalization performance of the ChromScore model trained on imputed vs. observed Roadmap Epigenomics dataset for each evaluation dataset.

Fig. S20: Feature ablation study on cell type generalization performance.

Fig. S21: Repeat element fold enrichments for ChromScore, mean plasmid-based expert score, mean CRISPR-based expert score and DNase signal quantiles.

Fig. S22: Cell type averages of Pearson correlations between gene expression and ChromScore, individual chromatin mark signals, expert scores.

Fig. S23: Percentage of genome above the binarization threshold by one or more experts.

Supplementary Tables

Table S1: Overview of the functional characterization datasets.

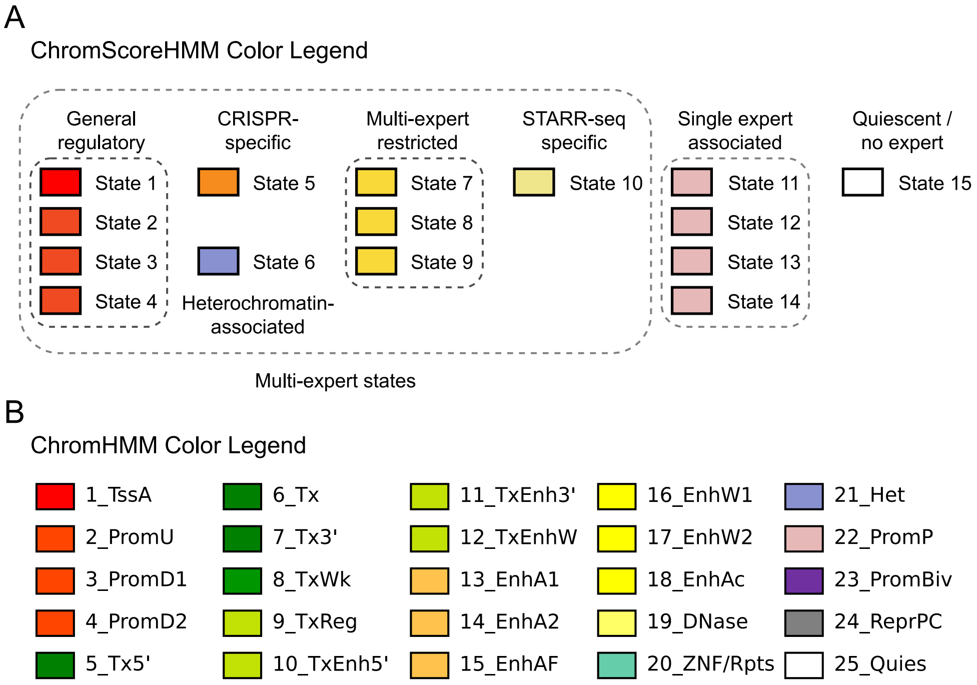


Fig. S1: Color legends for ChromScoreHMM states and the ChromHMM imputed 25-state model. (A) The color legend shows the ChromScoreHMM states grouped into seven color groups for visualization purposes. The groups and their associated states are as follows: General regulatory (States 1, 2, 3 and 4), which are broadly associated with regulatory activity across the majority of experts; CRISPR-specific (State 5), which is associated with CRISPR-based experts; Heterochromatin-associated (State 6), which is highly enriched for a heterochromatin chromatin state; Multi-expert restricted (States 7, 8 and 9), which are associated with two or three experts; STARR-seq specific (State 10), which is associated with a subset of STARR-seq-based experts; Single expert associated (States 11, 12, 13 and 14), which are emission parameters dominated by a single expert; Quiescent/no expert (State 15), which is associated with minimal predicted regulatory activity in all experts. **(B)** Color legend for the ChromHMM imputed 25-state model [1, 2].

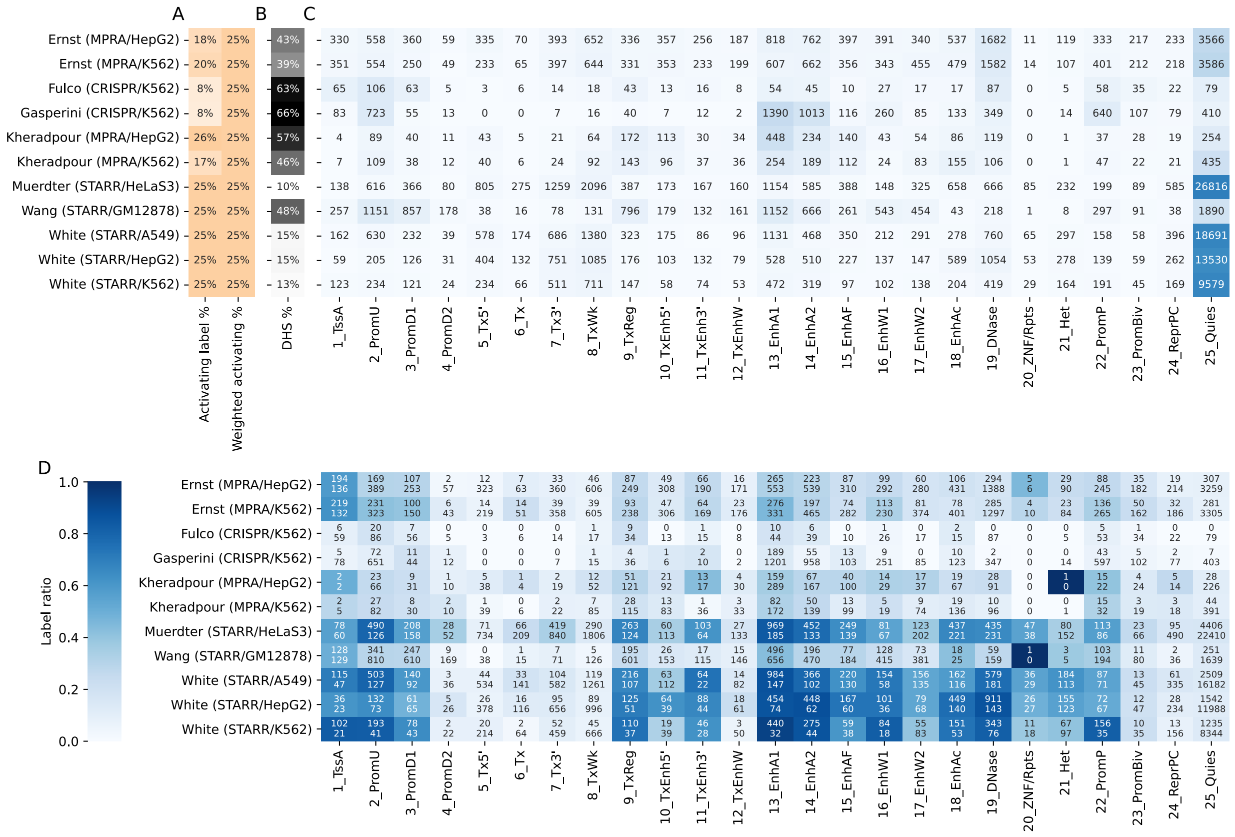


Fig. S2: Genomic coverage of functional characterization datasets. (A) Activating label %: Fraction of dataset loci labeled activating. Weighted activating %: Effective activating label fraction after label class weighting (Methods). Weighting was not necessary for STARR-seq datasets as neutral loci were selected with a 3-to-1 neutral to activating label ratio (Methods). **(B)** DHS %: Fraction of dataset loci overlapping peaks of DNase in corresponding cell type. Peak calls were based on imputed data obtained from Roadmap Epigenomics [3]. **(C)** Total number of dataset loci in each chromatin state. **(D)** Number of dataset loci labeled activating (top) or neutral (bottom) in each chromatin state. Heatmap color indicates activating label fraction as indicated by the color legend on the left.

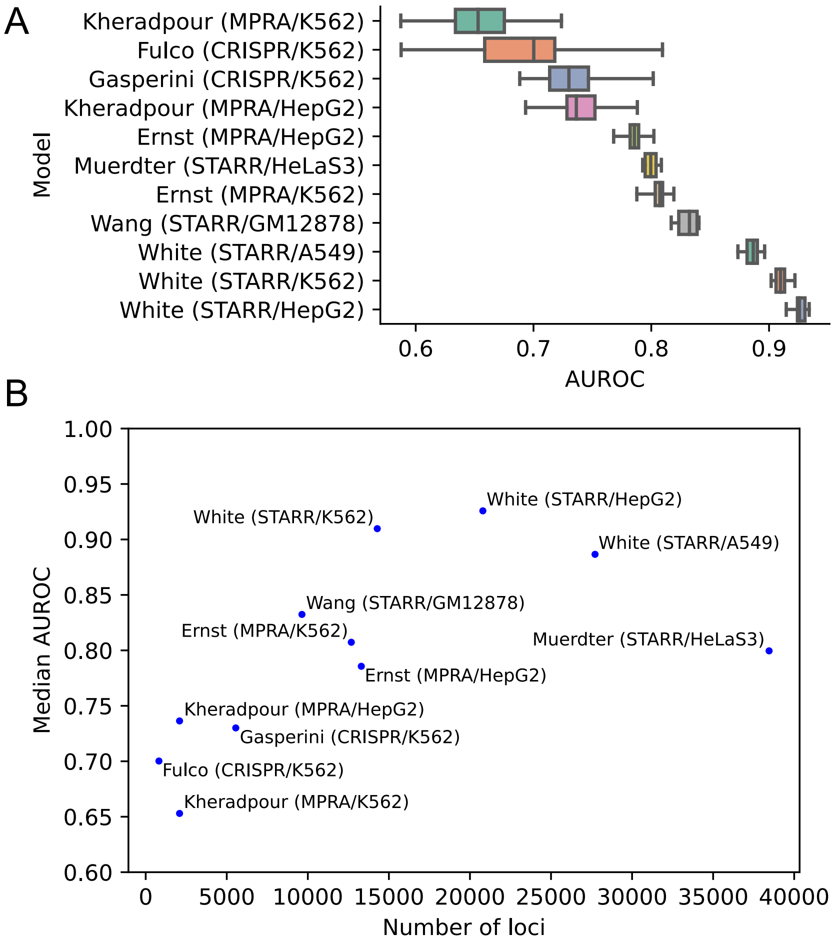


Fig. S3: Predictive performance of ChromScore expert models in held-out loci of the same functional characterization dataset. (A) Area under receiver operator characteristic curve (AUROC) distribution across 20 random permutation cross-validations for each expert model at predicting held-out loci in the same functional characterization dataset (Methods). The box represents quartiles and whiskers indicate maximum and minimum AUROCs across the train/test shuffles. **(B)** Median AUROCs for each expert across the train/test shuffles vs. number of genomic loci used in training (Spearman correlation 0.75).

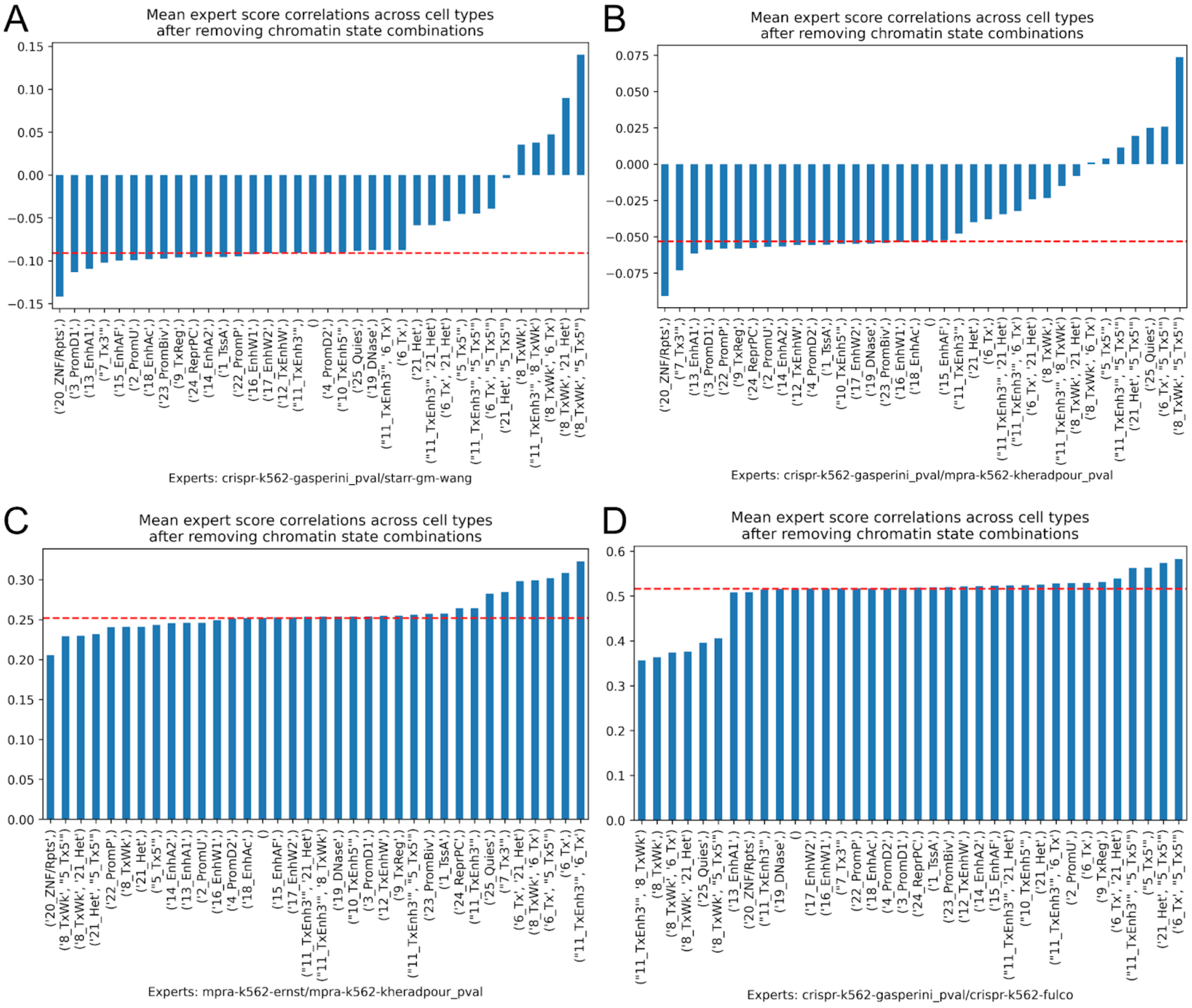
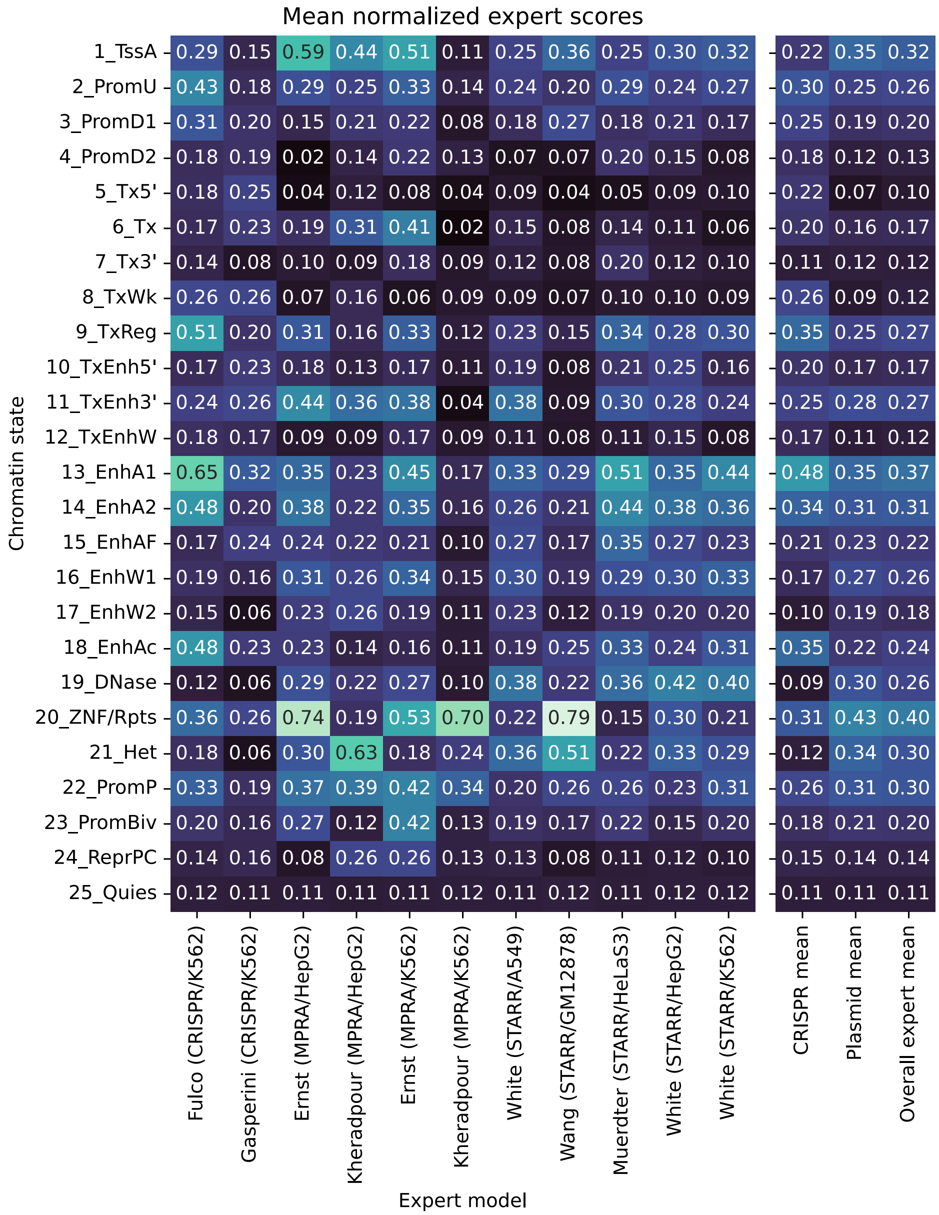


Fig. S4: Genomewide score correlations for selected pairs of expert predictions after removing loci assigned to individual chromatin states or combinations of chromatin states. The selected pairs all had low or negative correlations before removing the loci, for which results are shown in bar-graphs, are **(A)** Gasperini/CRISPR/K562 and Wang/STARR-seq/GM12878, **(B)** Gasperini/CRISPR/K562 and Kheradpour/MPRA/K562, **(C)** Ernst/MPRA/K562 and Gasperini/CRISPR/K562, **(D)** Gasperini/CRISPR/K562 and Fulco/CRISPR/K562. Score correlations are averaged over 30 randomly selected cell types. The bar with the label "()" and the dashed red line indicate correlations with no chromatin states removed. Chromatin states in pairs of states for which removal of loci that led to the largest increase in correlation include the transcription associated chromatin states 5_Tx5', 6_Tx, 8_TxWk, 11_TxEnh3', and Heterochromatin associated state 21_Het.



	Fulco (CRISPR/K562)	Gasperini (CRISPR/K562)	Ernst (MPRA/HepG2)	Kheradpour (MPRA/HepG2)	Ernst (MPRA/K562)	Kheradpour (MPRA/K562)	White (STARR/A549)	Wang (STARR/GM12878)	Muerdter (STARR/HeLaS3)	White (STARR/HepG2)	White (STARR/K562)	CRISPR mean	Plasmid mean	Overall expert mean
1_TssA	0.29	0.15	0.59	0.44	0.51	0.11	0.25	0.36	0.25	0.30	0.32	0.22	0.35	0.32
2_PromU	0.43	0.18	0.29	0.25	0.33	0.14	0.24	0.20	0.29	0.24	0.27	0.30	0.25	0.26
3_PromD1	0.31	0.20	0.15	0.21	0.22	0.08	0.18	0.27	0.18	0.21	0.17	0.25	0.19	0.20
4_PromD2	0.18	0.19	0.02	0.14	0.22	0.13	0.07	0.07	0.20	0.15	0.08	0.18	0.12	0.13
5_Tx5'	0.18	0.25	0.04	0.12	0.08	0.04	0.09	0.04	0.05	0.09	0.10	0.22	0.07	0.10
6_Tx	0.17	0.23	0.19	0.31	0.41	0.02	0.15	0.08	0.14	0.11	0.06	0.20	0.16	0.17
7_Tx3'	0.14	0.08	0.10	0.09	0.18	0.09	0.12	0.08	0.20	0.12	0.10	0.11	0.12	0.12
8_TxWk	0.26	0.26	0.07	0.16	0.06	0.09	0.09	0.07	0.10	0.10	0.09	0.26	0.09	0.12
9_TxReg	0.51	0.20	0.31	0.16	0.33	0.12	0.23	0.15	0.34	0.28	0.30	0.35	0.25	0.27
10_TxEnh5'	0.17	0.23	0.18	0.13	0.17	0.11	0.19	0.08	0.21	0.25	0.16	0.20	0.17	0.17
11_TxEnh3'	0.24	0.26	0.44	0.36	0.38	0.04	0.38	0.09	0.30	0.28	0.24	0.25	0.28	0.27
12_TxEnhW	0.18	0.17	0.09	0.09	0.17	0.09	0.11	0.08	0.11	0.15	0.08	0.17	0.11	0.12
13_EnhA1	0.65	0.32	0.35	0.23	0.45	0.17	0.33	0.29	0.51	0.35	0.44	0.48	0.35	0.37
14_EnhA2	0.48	0.20	0.38	0.22	0.35	0.16	0.26	0.21	0.44	0.38	0.36	0.34	0.31	0.31
15_EnhAF	0.17	0.24	0.24	0.22	0.21	0.10	0.27	0.17	0.35	0.27	0.23	0.21	0.23	0.22
16_EnhW1	0.19	0.16	0.31	0.26	0.34	0.15	0.30	0.19	0.29	0.30	0.33	0.17	0.27	0.26
17_EnhW2	0.15	0.06	0.23	0.26	0.19	0.11	0.23	0.12	0.19	0.20	0.20	0.10	0.19	0.18
18_EnhAc	0.48	0.23	0.23	0.14	0.16	0.11	0.19	0.25	0.33	0.24	0.31	0.35	0.22	0.24
19_DNase	0.12	0.06	0.29	0.22	0.27	0.10	0.38	0.22	0.36	0.42	0.40	0.09	0.30	0.26
20_ZNF/Rpts	0.36	0.26	0.74	0.19	0.53	0.70	0.22	0.79	0.15	0.30	0.21	0.31	0.43	0.40
21_Het	0.18	0.06	0.30	0.63	0.18	0.24	0.36	0.51	0.22	0.33	0.29	0.12	0.34	0.30
22_PromP	0.33	0.19	0.37	0.39	0.42	0.34	0.20	0.26	0.26	0.23	0.31	0.26	0.31	0.30
23_PromBiv	0.20	0.16	0.27	0.12	0.42	0.13	0.19	0.17	0.22	0.15	0.20	0.18	0.21	0.20
24_ReprPC	0.14	0.16	0.08	0.26	0.26	0.13	0.13	0.08	0.11	0.12	0.10	0.15	0.14	0.14
25_Quies	0.12	0.11	0.11	0.11	0.11	0.12	0.11	0.12	0.11	0.12	0.12	0.11	0.11	0.11

Fig. S5: Mean normalized expert scores per chromatin state. Rows correspond to chromatin states in the 25-state ChromHMM model based on imputed marks [1, 2, 4]. First 11 columns correspond to individual expert models. Values correspond to the mean normalized expert score averaged over the 127 cell types. The rightmost three columns correspond to the mean of the normalized scores across CRISPR-based experts, plasmid-based experts and all experts. See Methods for a description of the normalization procedure.

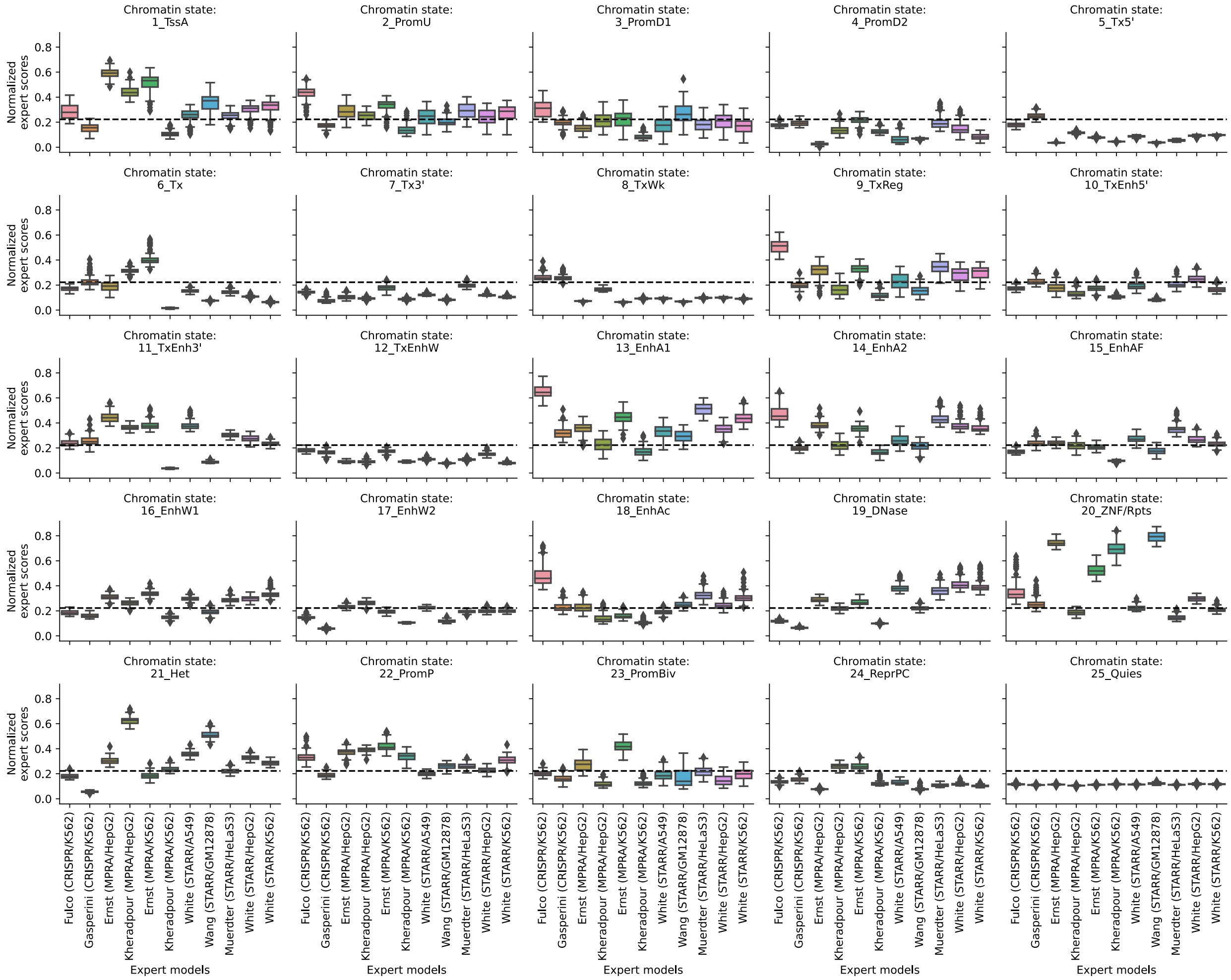


Fig. S6: Boxplots of mean normalized expert scores for each chromatin state, distributions across cell types. Each subplot corresponds to a chromatin state in a 25-state ChromHMM model. Each boxplot within a subplot indicates the distribution of mean normalized expert scores across cell types for the indicated expert and chromatin state (Methods). Horizontal line in the middle of the box indicates median, the box indicates upper and lower quartiles and whiskers indicate maximum and minimum scores. Dashed line indicates average normalized score across chromatin states.

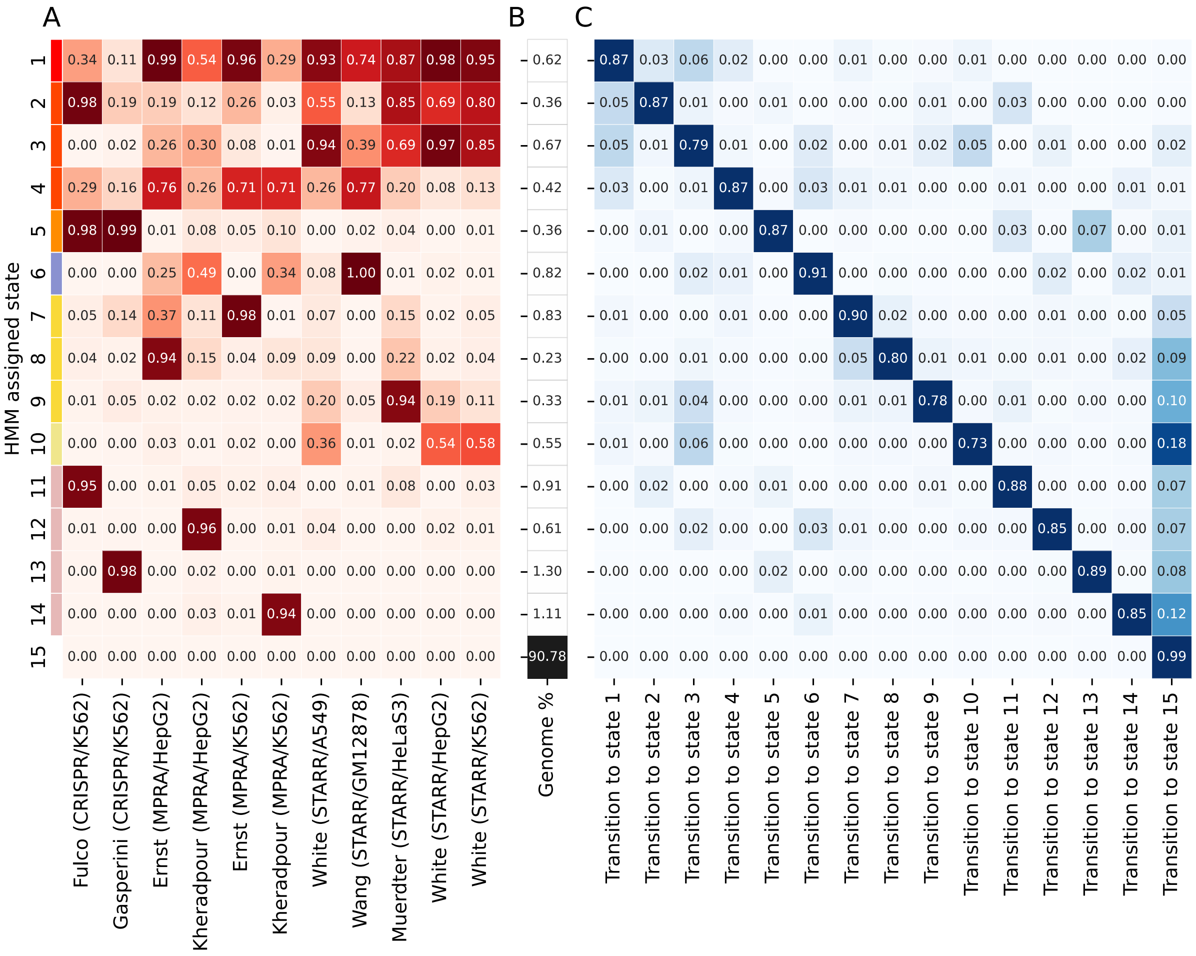


Fig. S7: ChromScoreHMM emission and transition parameters for the 15-state model. (A) Emission parameters of the model (see Figure 3 for details). **(B)** Percentage of the genome assigned to each ChromScoreHMM state. **(C)** Transition parameters of the model. Each value corresponds to the probability when in the state of the row of transitioning to the state of the column.

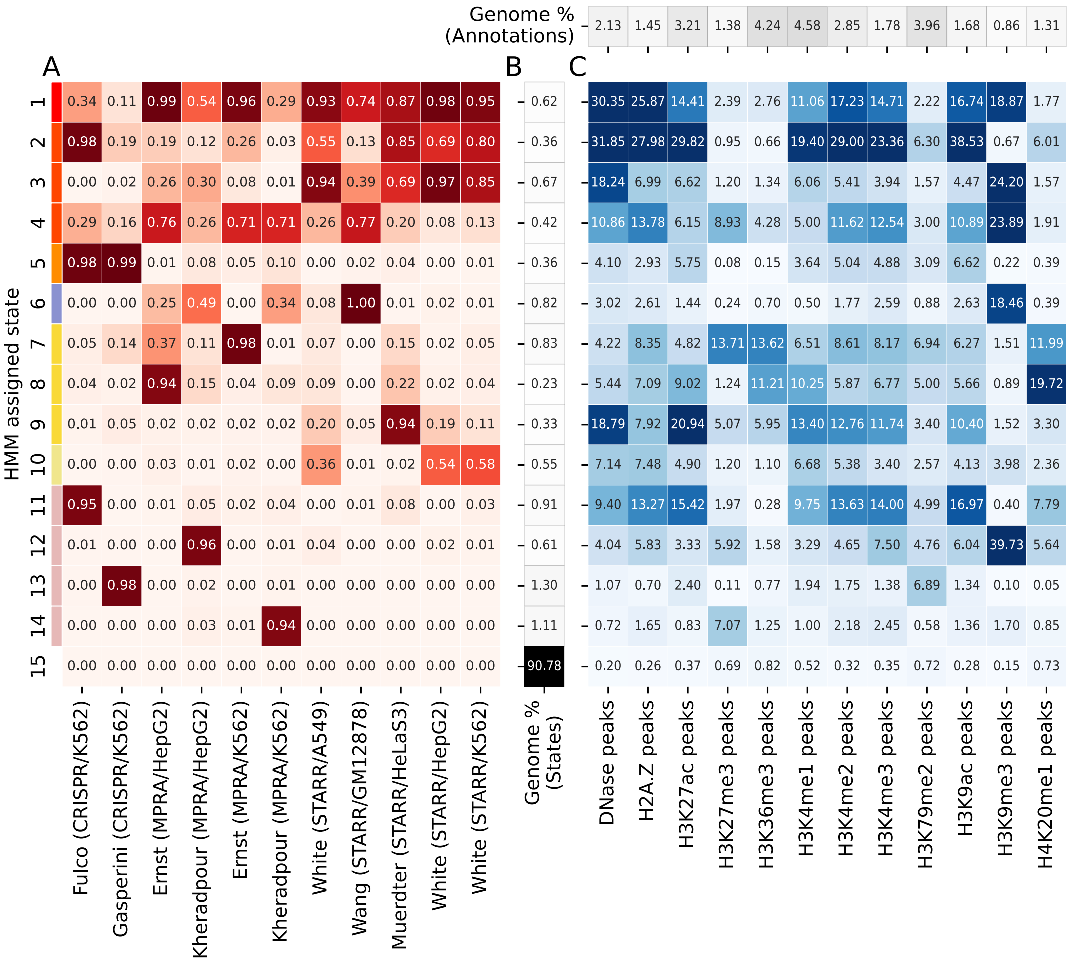


Fig. S8: ChromScoreHMM emission parameters and chromatin mark peak enrichments for the 15-state model. (A) Emission parameters of the ChromScoreHMM model (see Figure 3 for details). **(B)** Percentage of the genome assigned to each ChromScoreHMM state. **(C)** Fold enrichments for peak calls for individual chromatin marks based on imputed data from Roadmap Epigenomics. Top row indicates the percentage of the genome occupied by the peak call annotation. Enrichments and percentages are medians across cell types.

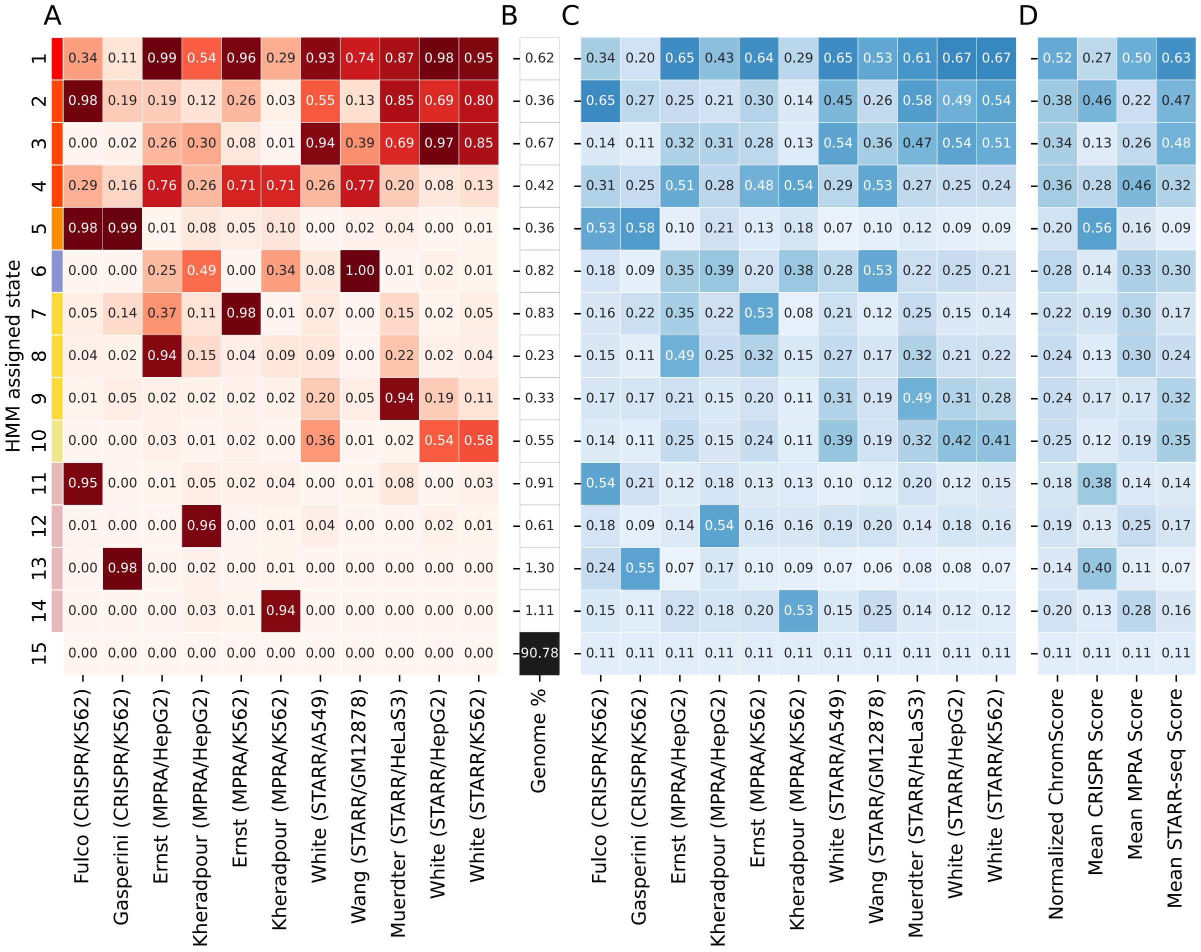


Fig. S9: ChromScoreHMM emission parameters and mean model scores for the 15-state model. (A): Emission parameters of ChromScoreHMM model (see Figure 3 for details). **(B)** Percentage of the genome assigned to each ChromScoreHMM state. **(C)** Mean normalized scores (Methods) for each expert model in each ChromScoreHMM state. **(D)** Mean normalized scores across ChromScoreHMM states based on combining predictions from multiple expert models. Normalized ChromScore refers to the mean normalized expert score track (Methods). Mean CRISPR, MPRA and STARR-seq scores refer to scores generated by taking the ensemble averages of normalized score tracks for expert models trained on CRISPR, MPRA and STARR-seq datasets respectively (Methods). Scores and percentages shown are medians across cell types.

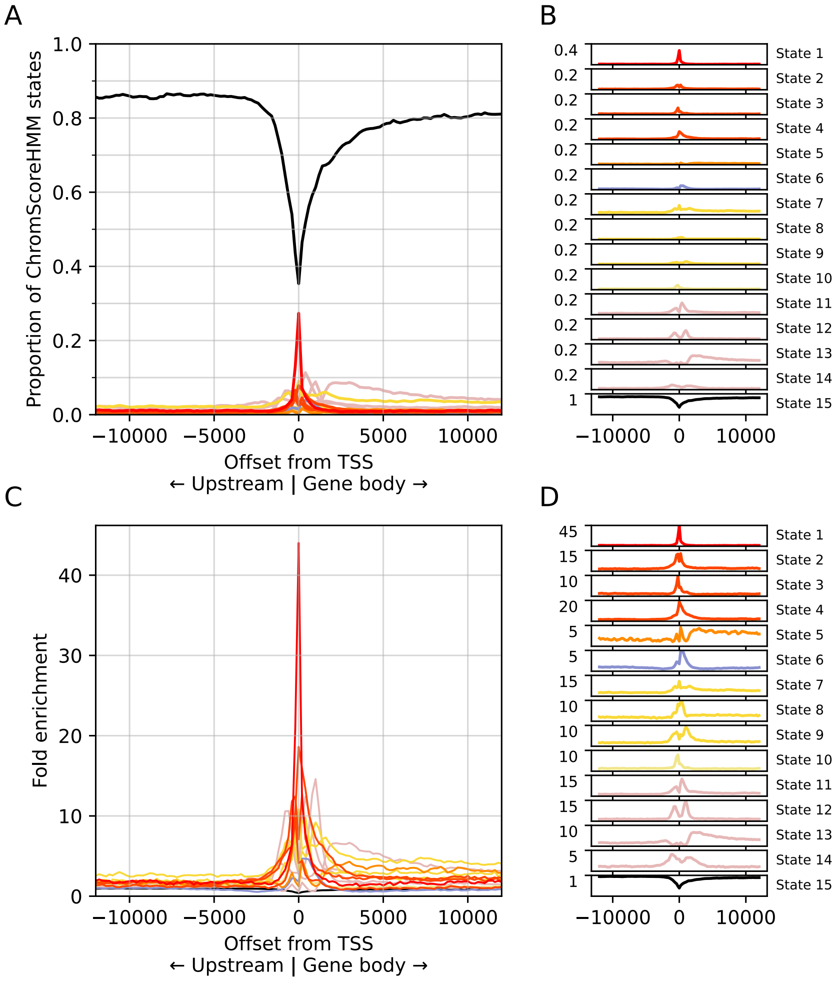


Fig. S10: ChromScoreHMM state proportions around transcription start sites. (A) Proportion of ChromScoreHMM state assignments represented by each ChromScoreHMM state as a function of genomic position centered around TSSs, averaged over cell types. See B for color legend. Negative offset values are upstream of the transcription start site and positive offset values are downstream. Offset zero refers to the TSS. **(B)** Individual plots of ChromScoreHMM state proportions and color legend. State 15 color changed from white to black for legibility. Bottom limits for the y-axes are 0 (omitted for clarity), upper limits as indicated on the plot. **(C)** Fold enrichments of ChromScoreHMM states as a function of genomic position, similar to (A). **(D)** Individual plots of ChromScoreHMM fold enrichments, similar to (B).

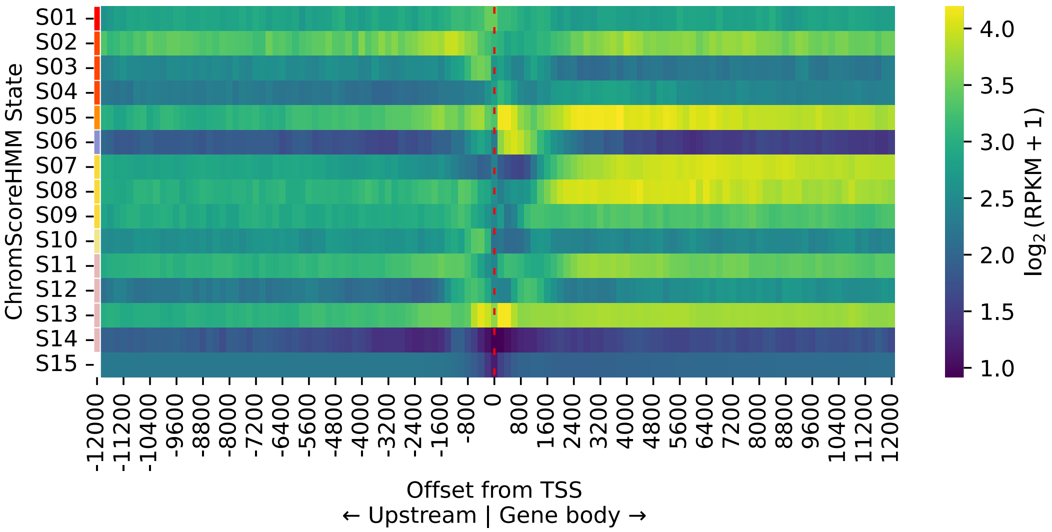


Fig. S11: ChromScoreHMM gene expression patterns around transcription start sites. Mean expression of nearby genes for positions within a +/-12 kb window around the TSSs (dashed red line) occupied by each ChromScoreHMM state. Each row corresponds to a ChromScoreHMM state and each column a position relative to a TSS. Expression values for each offset position and ChromScoreHMM state are the average $\log_2$(RPKM+1) values across both genes and cell types. RPKM: reads per kilobase of exon per million mapped reads.

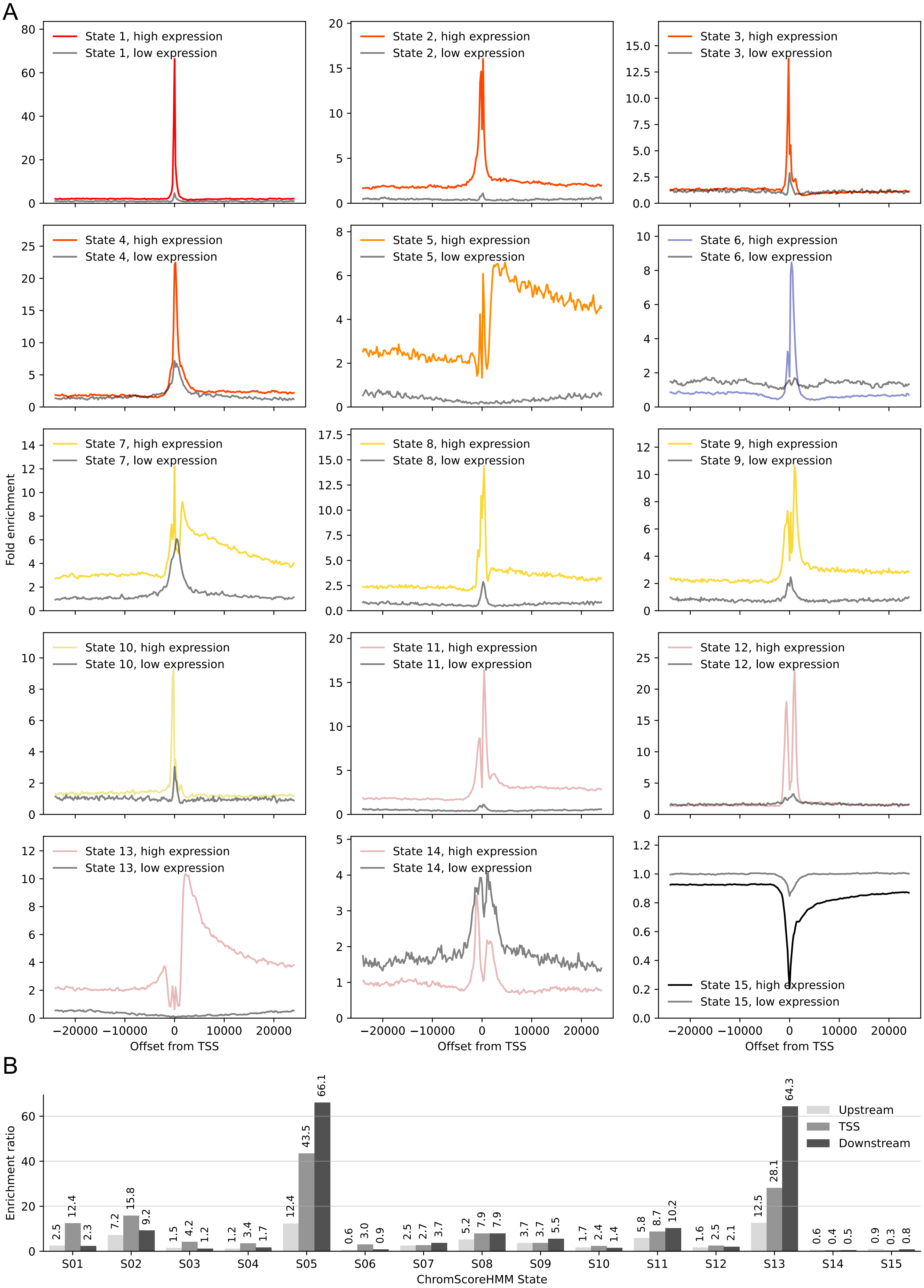

A
State 1, high expression
State 1, low expression
State 2, high expression
State 2, low expression
State 3, high expression
State 3, low expression
State 4, high expression
State 4, low expression
State 5, high expression
State 5, low expression
State 6, high expression
State 6, low expression
State 7, high expression
State 7, low expression
State 8, high expression
State 8, low expression
State 9, high expression
State 9, low expression
State 10, high expression
State 10, low expression
State 11, high expression
State 11, low expression
State 12, high expression
State 12, low expression
State 13, high expression
State 13, low expression
State 14, high expression
State 14, low expression
State 15, high expression
State 15, low expression
Fold enrichment
−20000 −10000 0 10000 20000
Offset from TSS
B
Enrichment ratio
0
20
40
60
Upstream
TSS
Downstream
S01 2.5 12.4 2.3
S02 7.2 15.8 9.2
S03 1.5 4.2 1.2
S04 1.2 3.4 1.7
S05 12.4 43.5 66.1
S06 0.6 3.0 0.9
S07 2.5 2.7 3.7
S08 5.2 7.9 7.9
S09 3.7 3.7 5.5
S10 1.7 2.4 1.4
S11 5.8 8.7 10.2
S12 1.6 2.5 2.1
S13 12.5 28.1 64.3
S14 0.6 0.4 0.5
S15 0.9 0.3 0.8
ChromScoreHMM State

Fig. S12: ChromScoreHMM overlap enrichments around TSSs of high expression and low expression genes. (A) Each subplot corresponds to a ChromScoreHMM state and shows the fold enrichments relative to the TSS for both a set of high expression and low expression genes as indicated by the color legend in each plot. High expression color corresponds to ChromScoreHMM state colors (Fig. S1). Negative offset values are upstream of the transcription start site and positive offset values are downstream. Offset zero refers to the TSS. Gene expression thresholds for high and low expression genes are $\log_2$(RPKM+1) > 1 and $\log_2$(RPKM+1) < 0.01, respectively. **(B)** Ratio of enrichments for high expression relative to low expression genes for each ChromScoreHMM state restricted to positions upstream of the TSS, at the TSS, or downstream of the TSS. The enrichment ratios correspond to the number of bases covered by a ChromScoreHMM state within an interval around high expression genes over the number of bases covered within the equivalent interval around low expression genes, each normalized by the total number of bases in the corresponding high expression and low expression gene intervals before computing the ratios. Upstream and downstream ratios computed over the 0.5-10 kb interval centered around the TSS. TSS ratio computed over the interval -200 to 200 bp.

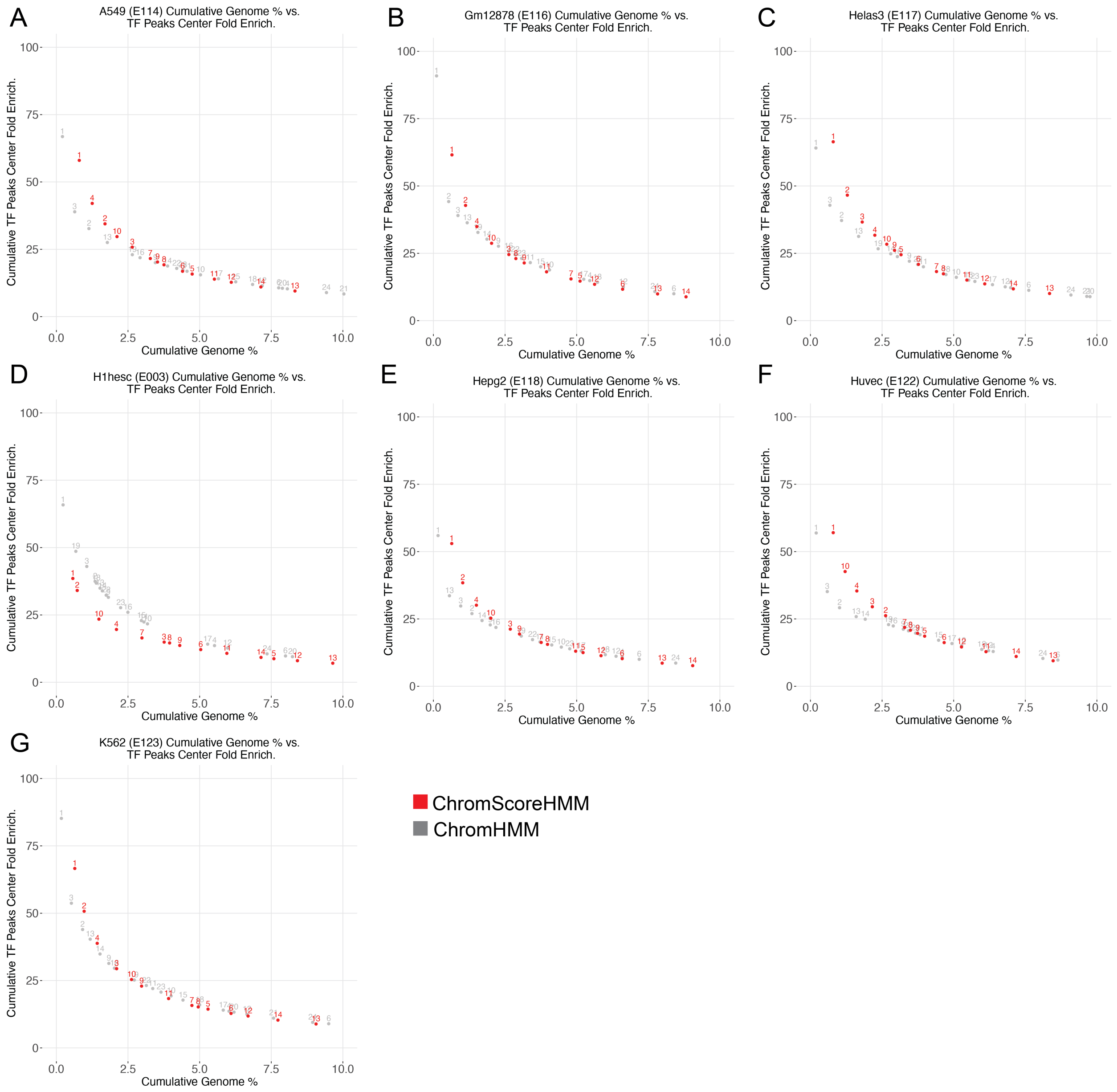


Fig. S13: Cumulative transcription factor binding peak enrichments and genome coverage for ChromScoreHMM and ChromHMM annotations. **(A)** The plot shows for A549 (E114) on the x-axis the cumulative % of the genome covered by the states and on the y-axis the cumulative fold enrichment of ChromScoreHMM states (red) and ChromHMM (gray) for the center base of transcription factor binding peaks when states are ordered based on enrichment. Points are labeled by the corresponding state. **(B-G)** The same as **A** except for (**B**) GM12878 (E116), **(C)** HeLa-S3 (E117), **(D)** H1hesc (E003), **(E)** Hepg2 (E118), **(F)** Huvec (E122), **(G)** K562 (E123).

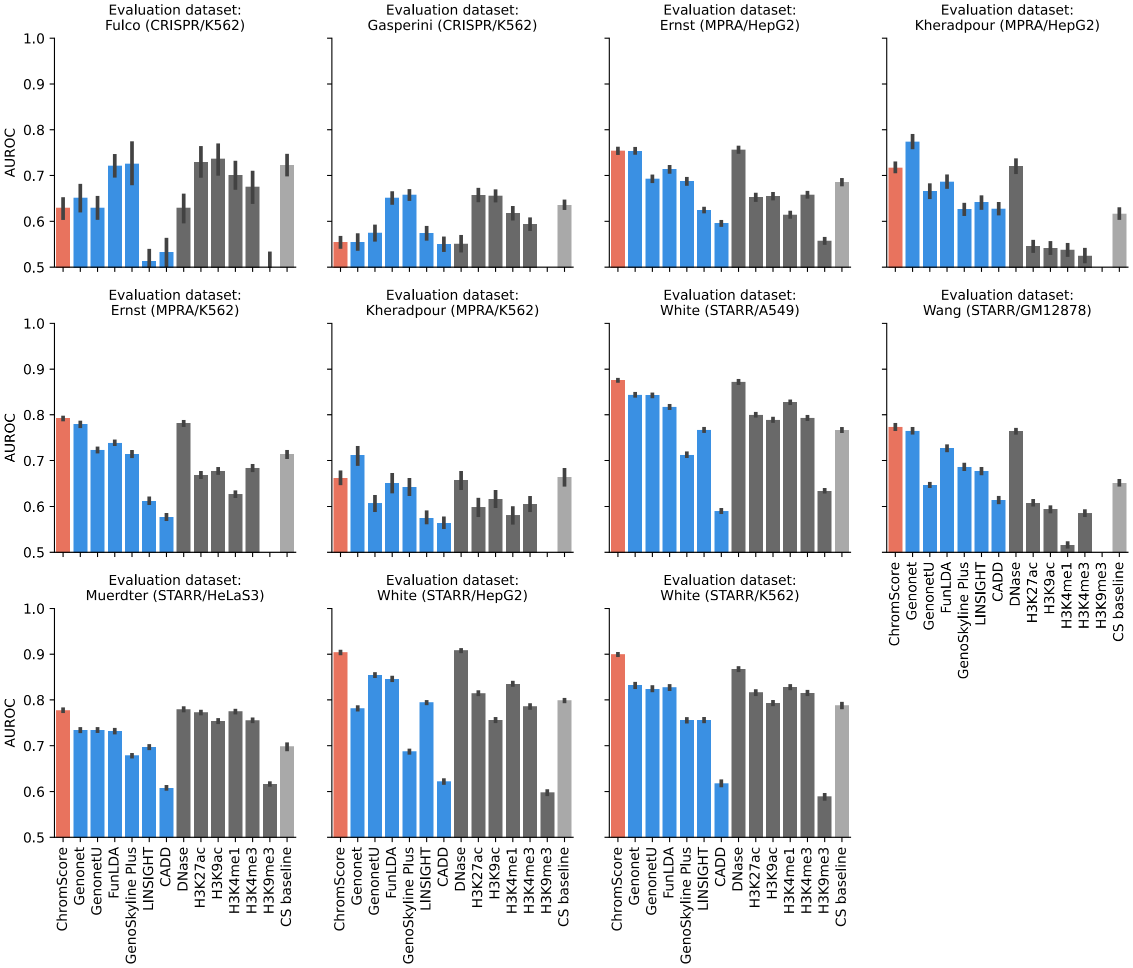


Fig. S14: Predictive performance of ChromScore, external models and baselines. Each panel corresponds to a functional characterization dataset used for evaluating regulatory activity predictions. ChromScore excludes all training data from evaluation cell types or from the same genomic position. Error bars indicate standard error across test partitions. AUROC: Area under receiver operator characteristic curve. Red bar: ChromScore, blue bars: external scores (Genonet, GenonetU, FunLDA, GenoSkylinePlus, LINSIGHT, CADD), dark gray bars: individual chromatin mark signals, light gray bar: chromatin state baseline model.

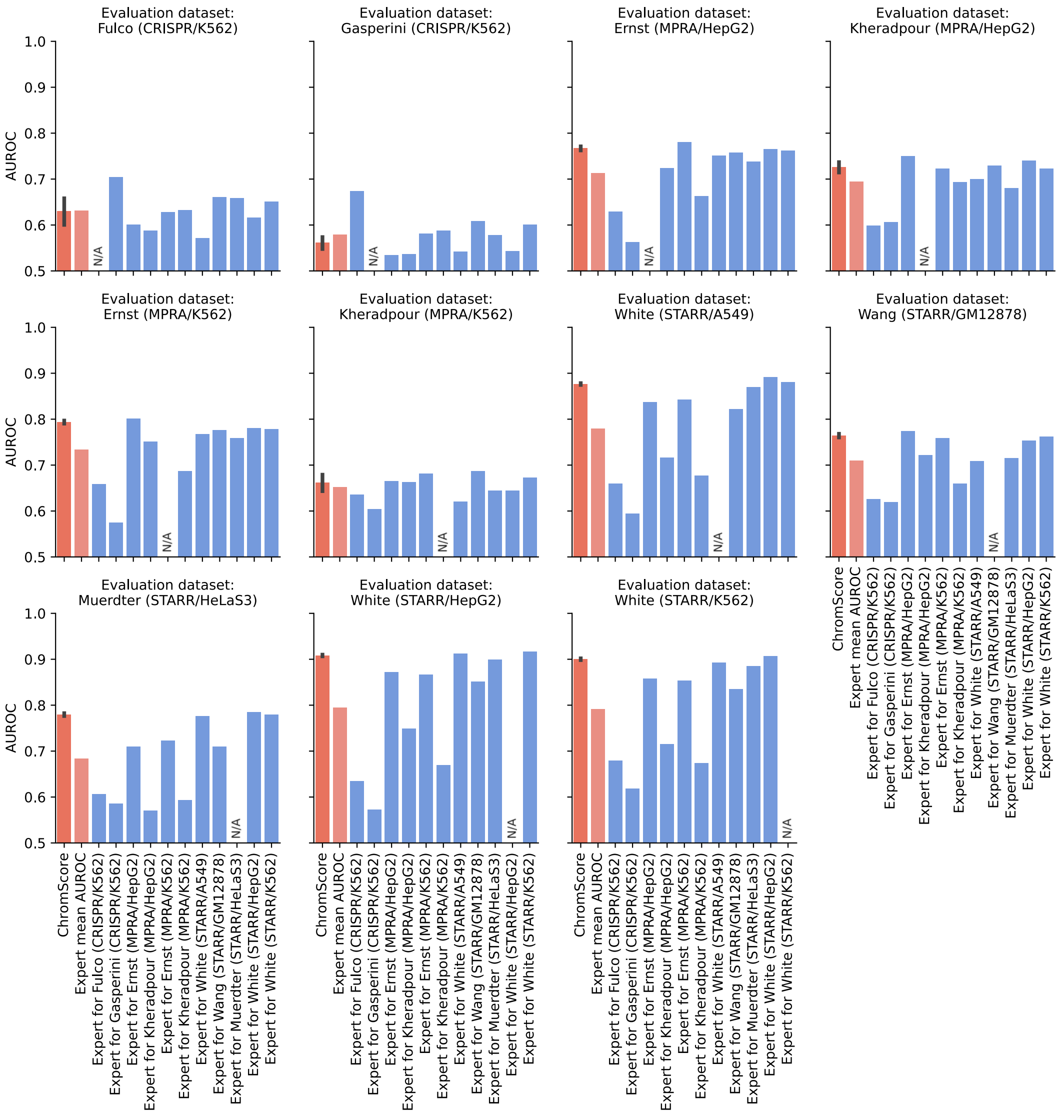


Fig. S15: Predictive performance of ChromScore and individual experts. Each panel corresponds to a functional characterization dataset used for evaluating regulatory activity predictions. Expert models trained on a given dataset were not evaluated on the same dataset and were marked "N/A". See Fig. S3 for performance of expert models on the same dataset under cross-validation. In evaluating ChromScore, we excluded all training data from the evaluation cell types or from the same genomic position (Figure 4B, Methods). Error bar indicates standard error across ChromScore test partitions. AUROC: Area under receiver operator characteristic curve. Red bar: ChromScore, light red bar: mean AUROC of individual experts for the given evaluation, blue bars: individual experts.

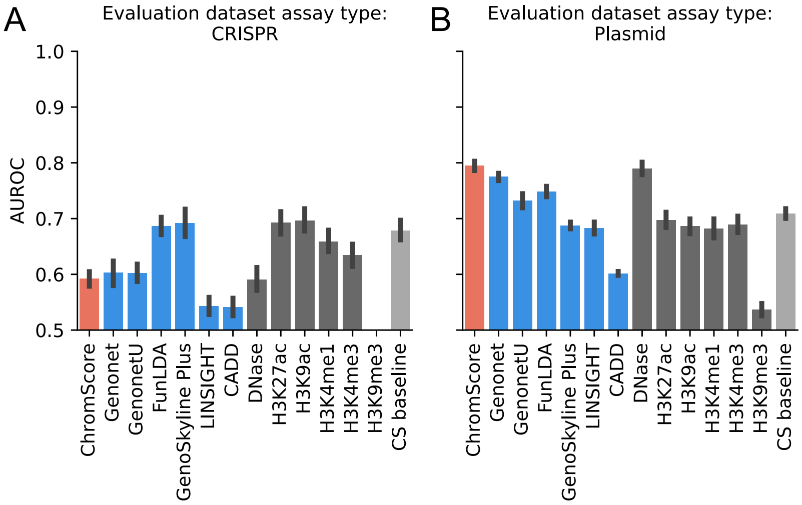


Fig. S16: Predictive performance of ChromScore, external models and baselines in CRISPR and plasmid-based functional characterization datasets. (A) Mean score performance evaluated in CRISPR-based datasets. **(B)** Mean score performance evaluated in plasmid-based datasets. Error bars indicate standard error across evaluations. AUROC: Area under receiver operator characteristic curve. Red bar: ChromScore, blue bars: external scores (Genonet, GenonetU, FunLDA, GenoSkylinePlus, LINSIGHT, CADD), dark gray bars: individual chromatin mark signals, light gray bar: chromatin state baseline model.

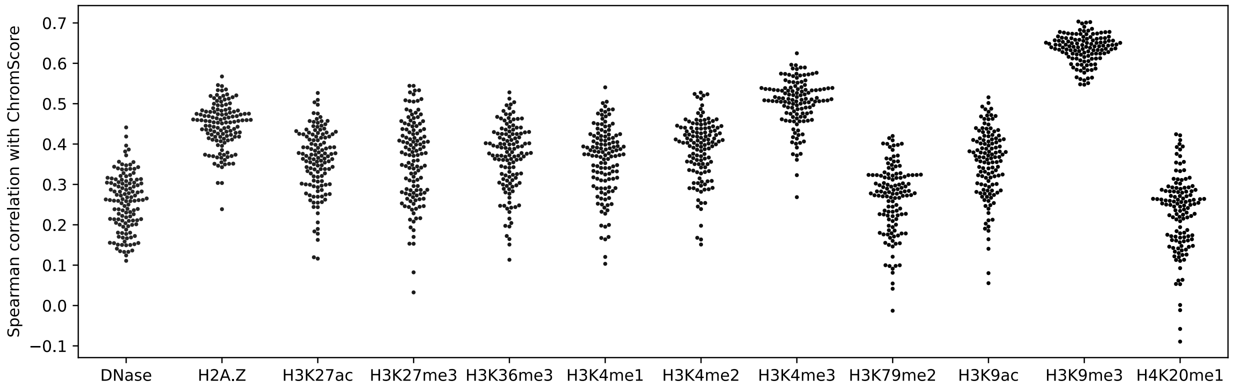


Fig. S17: Spearman correlations between ChromScore and chromatin mark signal tracks. Each point indicates correlation between the indicated mark and ChromScore in one cell type.

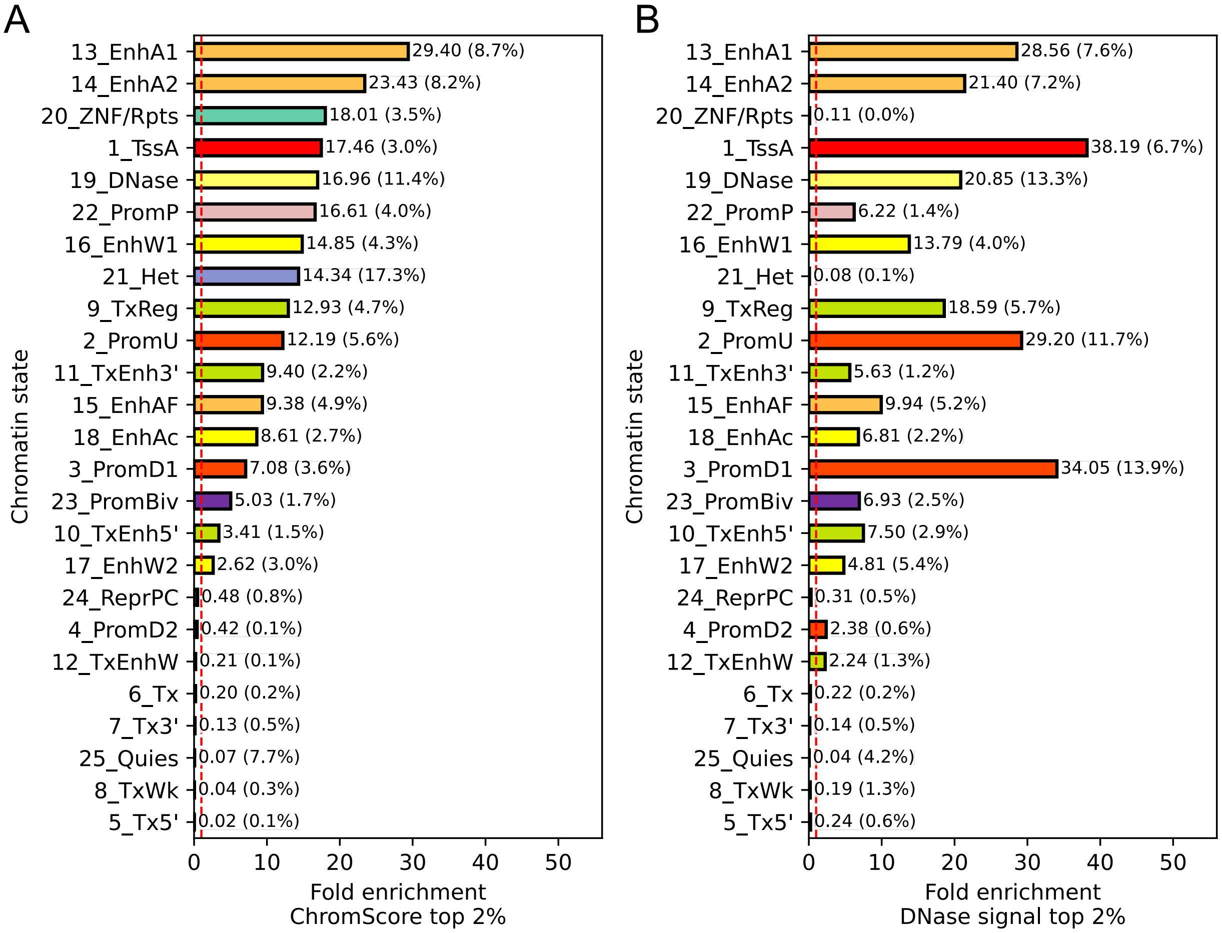


Fig. S18: Chromatin state fold enrichments of top ChromScore and DNase-I hypersensitive regions. Chromatin state fold enrichments for genomic regions assigned the top 2% of **(A)** ChromScore, **(B)** DNase-I hypersensitivity signal, geometric mean over cell types. Dashed red line indicates fold enrichment 1. Chromatin state composition average percentages within the top 2% of regions shown in parentheses. States are ordered as in **A** by ChromScore enrichment.

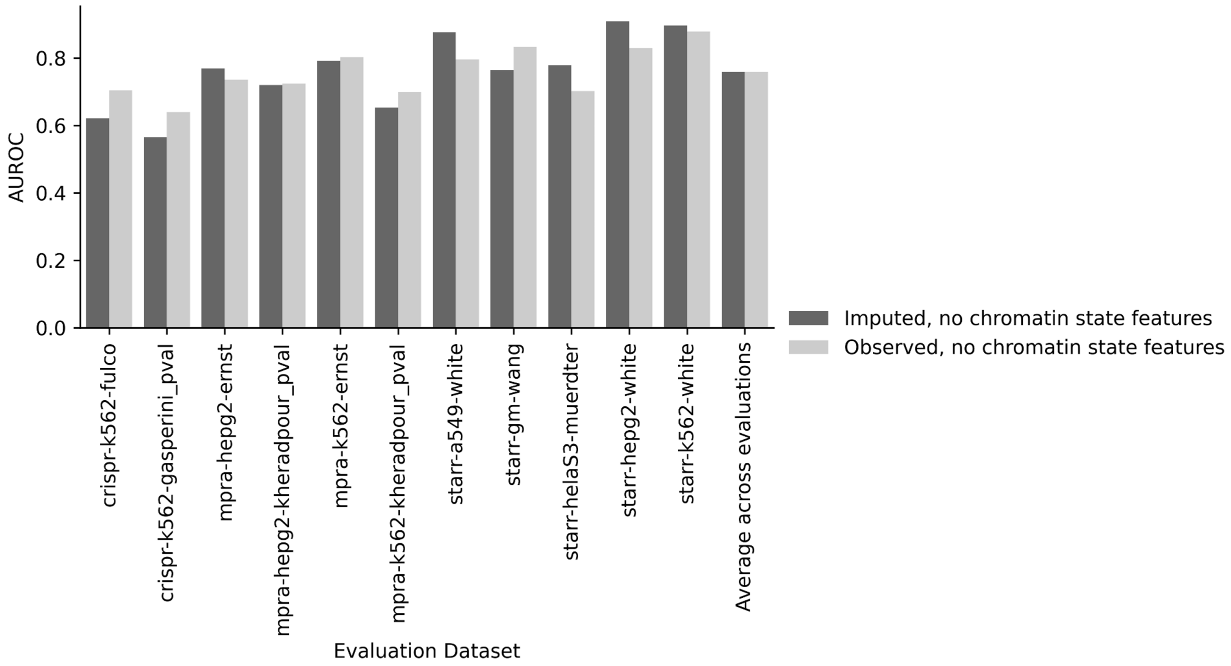


Fig. S19: Cell type generalization performance of the ChromScore model trained on imputed vs. observed Roadmap Epigenomics dataset for each evaluation dataset. Each pair of bars shows the AUROC for imputed and observed data for one functional testing dataset with the last pair of bars showing the average AUROC over all the datasets. Both models achieved equivalent average AUROCs across the evaluation datasets: 0.758 for the model trained on imputed data and 0.758 for the model trained on observed data, with small differences across individual evaluation datasets. To enable direct comparison between models, chromatin state features were removed during training of both models.

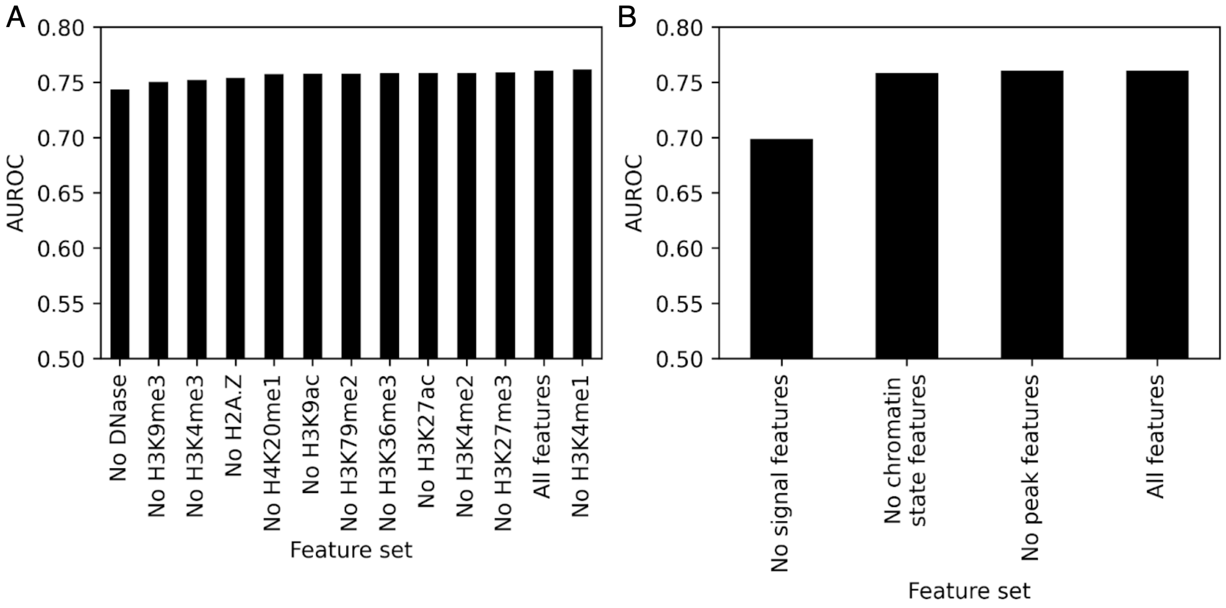


Fig. S20: Feature ablation study on cell type generalization performance. (A) Cell type generalization AUROCs averaged across all evaluation datasets after removing signal tracks and peaks for the mark as well as chromatin state annotations, which include marks as features. **(B)** Cell type generalization AUROCs averaged across all evaluation datasets after removing all peak, chromatin state or signal features.

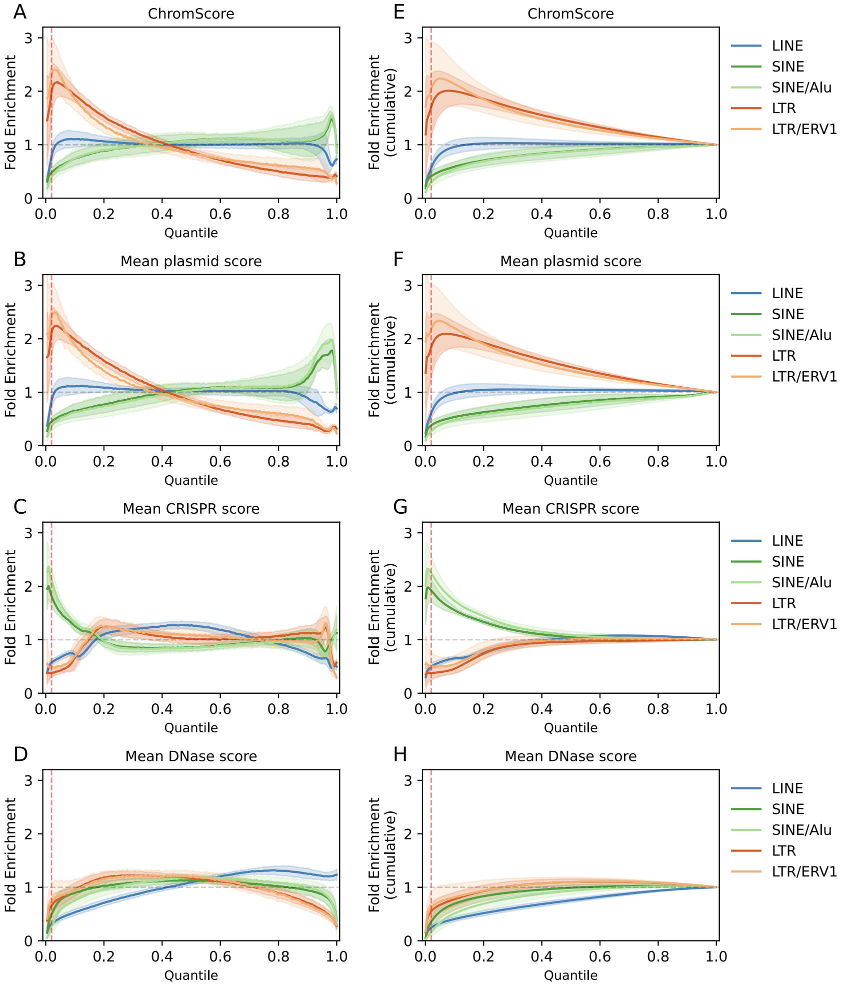


Fig. S21: Repeat element fold enrichments for ChromScore, mean plasmid-based expert score, mean CRISPR-based expert score and DNase signal quantiles. Fold enrichments for long interspersed nuclear elements (LINEs), all short interspersed nuclear elements (SINEs), the Alu subclass of SINEs, all long terminal repeats (LTRs) and the endogenous retroviral sequence 1 (ERV1) subclass of LTRs, computed for **(A)** ChromScore, **(B)** mean plasmid-based expert score, **(C)** mean CRISPR-based expert score, **(D)** DNase signal shown as a function of quantile when values were sorted in decreasing order **(E-H)** same as (A-D) except for cumulative fold enrichments. Lines indicate mean enrichments across 127 cell types, shaded areas indicate standard deviations across cell types. Fold enrichments are computed relative to the genome background. Red vertical dashed line denotes the top 2% of scores. Gray horizontal line indicates fold enrichment of 1.

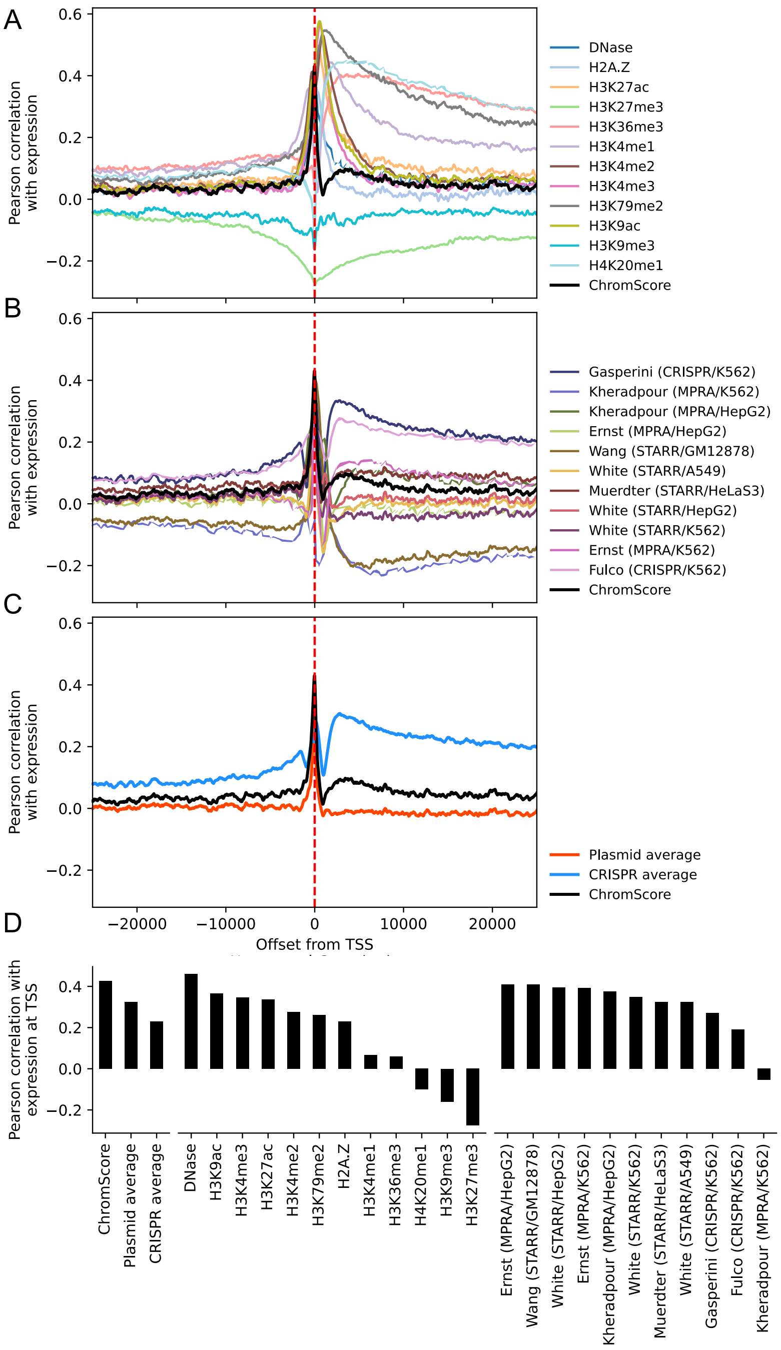
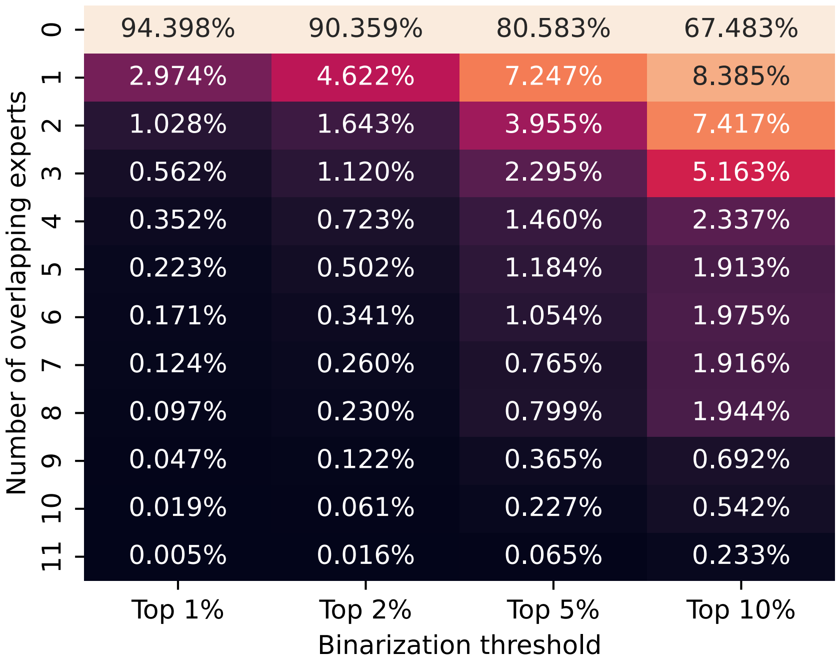


Fig. S22: Cell type averages of Pearson correlations between gene expression and ChromScore, individual chromatin mark signals, expert scores. Correlation of gene expression ($\log_2$(RPKM+1)) with ChromScore and (**A**) individual chromatin mark signals, **(B)** individual expert tracks, **(C)** and average of plasmid-based expert tracks and CRISPR-based expert tracks as a function of position relative to the TSS within a 24 kb window. **(D)** Bar graph reporting correlation with gene expression of A-C specifically at the TSS.



Number of overlapping experts	Top 1%	Top 2%	Top 5%	Top 10%
0	94.398%	90.359%	80.583%	67.483%
1	2.974%	4.622%	7.247%	8.385%
2	1.028%	1.643%	3.955%	7.417%
3	0.562%	1.120%	2.295%	5.163%
4	0.352%	0.723%	1.460%	2.337%
5	0.223%	0.502%	1.184%	1.913%
6	0.171%	0.341%	1.054%	1.975%
7	0.124%	0.260%	0.765%	1.916%
8	0.097%	0.230%	0.799%	1.944%
9	0.047%	0.122%	0.365%	0.692%
10	0.019%	0.061%	0.227%	0.542%
11	0.005%	0.016%	0.065%	0.233%

Fig. S23: Percentage of genome above the binarization threshold by one or more experts. The columns correspond to the binarization thresholds of top 1%, 2%, 5%, and 10% of the expert prediction scores. Each row corresponds to a number of experts. Values correspond to the percentage of 25 bp intervals in the genome for which exactly the number of experts of the row had scored it above the binarization threshold of the column.

Table S1: Overview of the functional characterization datasets.

Name	Cell type	Assay technology	Unit of observation	Citation	Number of activating loci	Number of neutral loci	Total number of data points
Fulco/K562	K562	CRISPR-dCas9	Center nucleotide of target site	[5]	69	747	816
Gasperini/K562	K562	CRISPR-dCas9	Center nucleotide of target site	[6]	432	5122	5554
Kheradpour/HepG2	HepG2	MPRA	Individual construct	[7]	541	1548	2089
Kheradpour/K562	K562	MPRA	Individual construct	[7]	347	1742	2089
Ernst/HepG2	HepG2	MPRA	Nucleotide with absolute maximum score in tiled region	[8]	2405	10,894	13,299
Ernst/K562	K562	MPRA	Nucleotide with absolute maximum score in tiled region	[8]	2519	10,162	12,681
Wang/GM12878	GM12878	STARR-seq	Center of driver elements	[9]	2409	7227	9636
Muerdter/HeLaS3	HeLa-S3	STARR-seq	Center of peak call	[10]	9613	28,839	38,452
White/A549	A549	STARR-seq	Center of peak call	[11]	6929	20,787	27,716
White/HepG2	HepG2	STARR-seq	Center of peak call	[12]	5199	15,597	20,796
White/K562	K562	STARR-seq	Center of peak call	[12]	3571	10,713	14,284

References

1. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nat Methods. 2012;9:215–6.

2. Ernst J, Kellis M. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. Nat Biotechnol. 2015;33:364–76.

3. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518:317–30.

4. Ernst J, Kellis M. Chromatin-state discovery and genome annotation with ChromHMM. Nat Protoc. 2017;12:2478–92.

5. Fulco CP, Nasser J, Jones TR, Munson G, Bergman DT, Subramanian V, et al. Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. Nat Genet. 2019;51:1664–9.

6. Gasperini M, Hill AJ, McFaline-Figueroa JL, Martin B, Kim S, Zhang MD, et al. A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. Cell. 2019;176:377-390.e19.

7. Kheradpour P, Ernst J, Melnikov A, Rogov P, Wang L, Zhang X, et al. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. Genome Res. 2013;23:800–11.

8. Ernst J, Melnikov A, Zhang X, Wang L, Rogov P, Mikkelsen TS, et al. Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. Nat Biotechnol. 2016;34:1180–90.

9. Wang X, He L, Goggin SM, Saadat A, Wang L, Sinnott-Armstrong N, et al. High-resolution genome-wide functional dissection of transcriptional regulatory regions and nucleotides in human. Nat Commun. 2018;9:5380.

10. Muerdter F, Boryń ŁM, Woodfin AR, Neumayr C, Rath M, Zabidi MA, et al. Resolving systematic errors in widely used enhancer activity assays in human cells. Nat Methods. 2018;15:141–9.

11. White K. ENCSR895FDL. Datasets. Roadmap Epigenomics. 10.17989/ENCSR895FDL. 2020.

12. Lee D, Shi M, Moran J, Wall M, Zhang J, Liu J, et al. STARRPeaker: uniform processing and accurate identification of STARR-seq active regions. Genome Biol. 2020;21:298.