Genome Biology

METHOD

Open Access

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

Tevfik Umut Dincer[1,2] and Jason Ernst[1,2,3,4,5,6,7*]

*Correspondence:
jason.ernst@ucla.edu

[1] Bioinformatics Interdepartmental Program, University of California, Los Angeles, Los Angeles, CA 90095, USA
[2] Department of Biological Chemistry, University of California, Los Angeles, Los Angeles, CA 90095, USA
[3] Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research at University of California, Los Angeles, Los Angeles, CA 90095, USA
[4] Computer Science Department, University of California, Los Angeles, Los Angeles, CA 90095, USA
[5] Jonsson Comprehensive Cancer Center, University of California, Los Angeles, Los Angeles, CA 90095, USA
[6] Molecular Biology Institute, University of California, Los Angeles, Los Angeles, CA 90095, USA
[7] Department of Computational Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA

Abstract

We introduce ChromActivity, a computational framework for predicting and annotating regulatory activity across the genome through integration of multiple epigenomic maps and various functional characterization datasets. ChromActivity generates genomewide predictions of regulatory activity associated with each functional characterization dataset across many cell types based on available epigenomic data. It then for each cell type produces ChromScoreHMM genome annotations based on the combinatorial and spatial patterns within these predictions and ChromScore tracks of overall predicted regulatory activity. ChromActivity provides a resource for analyzing and interpreting the human regulatory genome across diverse cell types.

Keywords: Epigenome, Gene regulation, CRISPR screens, Massively parallel reporter assays, Hidden Markov model, Genome annotation, Machine learning

Background

Transcriptional regulation of gene expression is controlled by a large set of regulatory elements distributed across the genome [1–3]. Identifying and predicting regulatory elements is important to advancing our understanding of cellular processes and gaining insight into the genetic basis of common diseases [1, 4, 5].

Epigenomic data, such as maps of histone modifications, histone variants, and chromatin accessibility, have been powerful resources for the identification of candidate regulatory elements within the genome [4, 6–9]. Such data are now available across hundreds of different cell or tissue types based on the efforts of large consortium projects [7, 9, 10] as well as contributions from individual labs [6, 11]. Maps of chromatin marks have enabled the prediction of regulatory elements in hundreds of cell types, often through unsupervised approaches such as calling peaks on single marks [12] or the

identification of combinatorial and spatial patterns of multiple marks using chromatin state models [13–16].

However, despite its extensive utility, unsupervised integration of chromatin marks does not take advantage of information from functional characterization assays to potentially better predict regulatory regions. Functional characterization assays complement chromatin marks by enabling direct testing of genomic regions for regulatory activity in high-throughput [17–20] by either incorporating sequences of candidate regulatory elements into cells via plasmids or by manipulating or interfering with the genome itself using lentiviral integrases or CRISPR-based technologies [3]. Plasmid-based assays [21], such as barcoded Massively Parallel Reporter Assays (MPRAs) [22, 23] or Self-Transcribing Active Regulatory Region Sequencing (STARR-seq) assays [24], typically measure the expression of a reporter gene on a plasmid containing the candidate regulatory element, serving as an indicator of the expression level that it is likely to induce in the cell. In contrast, genomically integrated assays target the genome directly in its native environment, for example by altering the epigenetic landscape near a candidate regulatory element (e.g., CRISPR interference screens that use dCas9 with an attached KRAB repressor domain [25]). Notably, only a subset of regulatory element predictions based on epigenomic data typically validate in functional characterization assays [3].

While functional characterization assays provide a more direct assessment of regulatory activity, they can also have some limitations. They are less widely available across cell types compared to chromatin mark datasets, partly due to cost and resource constraints associated with these specialized assays, and also because of technical challenges such as achieving sufficient transfection efficiency in certain cell types [26]. Another drawback for some assays is limited genomic coverage: functional characterization assays often provide readouts for a limited subset of genomic regions, whereas chromatin mark data can be mapped genomewide. Integrative approaches that combine the broad availability of chromatin marks with direct testing of functional assays have the potential to computationally extend the cell type coverage of functional testing assays.

Several existing methods have used data from high-throughput functional characterization assays as training data for supervised methods that predict regulatory activity [27, 28] or for predicting effects of individual sequence mutations based on features including sequence [29–32]. However, these methods generally focus on scoring sites or bases within the same cell type for which training data is available. As many sequence and transcription factor binding features are cell type specific, a method optimized to make predictions within a cell type it is trained in might be less effective at making predictions that generalize well across cell types. Additionally, the reliance on a single functional characterization assay or dataset, as commonly seen in existing methods, could introduce biases to the predictions, given that technical differences even within the same assay type have been shown to impact the readouts [19].

To address the challenges of predicting regulatory activity genomewide across a range of cell and tissue types, we propose ChromActivity, a computational framework that integrates chromatin marks with a variety of functional characterization datasets. ChromActivity employs a supervised learning approach to generate genomewide regulatory activity predictions and annotations across multiple cell types. ChromActivity is designed to effectively generalize across both cell types and genomic loci and to

produce annotations that reflect differences between functional characterization assays. We apply ChromActivity in over one hundred human cell and tissue types to generate a set of genomewide regulatory activity prediction tracks, where each track is based on a model that is specifically trained on one of 11 functional characterization datasets. ChromActivity generates ChromScoreHMM genome annotations, which correspond to combinatorial and spatial patterns in the prediction tracks. ChromActivity also generates ChromScore tracks, composite genomewide regulatory activity prediction scores on a per-cell or tissue type basis that reflects the mean predicted regulatory activity based on the different functional characterization datasets. The ChromActivity framework and associated annotations provide a resource for analyzing gene regulatory activity across a broad range of human cell and tissue types.

Results

Overview of the ChromActivity framework

We developed ChromActivity to provide annotations and scores of predicted regulatory activity across human cell and tissue types by leveraging information from both epigenomic data and a variety of functional characterization datasets. As a first step in the ChromActivity framework, a separate model is trained for each functional characterization dataset to predict the relative likelihood of each 25-bp genomic interval showing activity based on chromatin mark features across the entire human genome. These individual prediction scores are then integrated to produce ChromScoreHMM annotations, which are unsupervised genome annotations built on top of ChromHMM [13, 14]. The scores are also integrated into a single combined score, ChromScore (Fig. 1, Additional file 1: Fig. S1).

ChromActivity makes predictions for any cell type with chromatin mark data available (For ease of presentation, we use the term "cell type" to refer to cell types, tissue types, and reference epigenomes collectively). Notably, ChromActivity operates without assuming any functional characterization data is available in the cell types for which it predicts. This is important as currently many cell types have extensive chromatin mark data but lack corresponding functional characterization assay data. ChromActivity's approach relies on the observation that the same chromatin mark patterns generally mark regulatory regions in different cell types, though their specific genomic locations can vary [4, 9]. This contrasts with specific DNA sequence or transcription factor binding patterns which can mark regulatory regions only in specific cell types and are therefore not used as features in ChromActivity.

Initially, ChromActivity trains a separate bagging ensemble of regularized logistic regression models for each input functional characterization dataset. These models are trained with labels derived from the readouts for the genomic regions tested by the functional characterization datasets, which include both plasmid-based (MPRAs, STARR-seq screens) [33–37] and genome-integrated assays (CRISPR-dCas9 screens) [25, 38] from multiple different conditions and cell types (Methods). ChromActivity uses features derived from signal tracks and peak calls of individual chromatin marks, as well as chromatin state annotations [13, 14]. In addition to using the signal directly at the tested loci, ChromActivity incorporates spatial information from the signal track by extracting the signal at 25 bp resolution within 2-kb windows centered around the tested loci

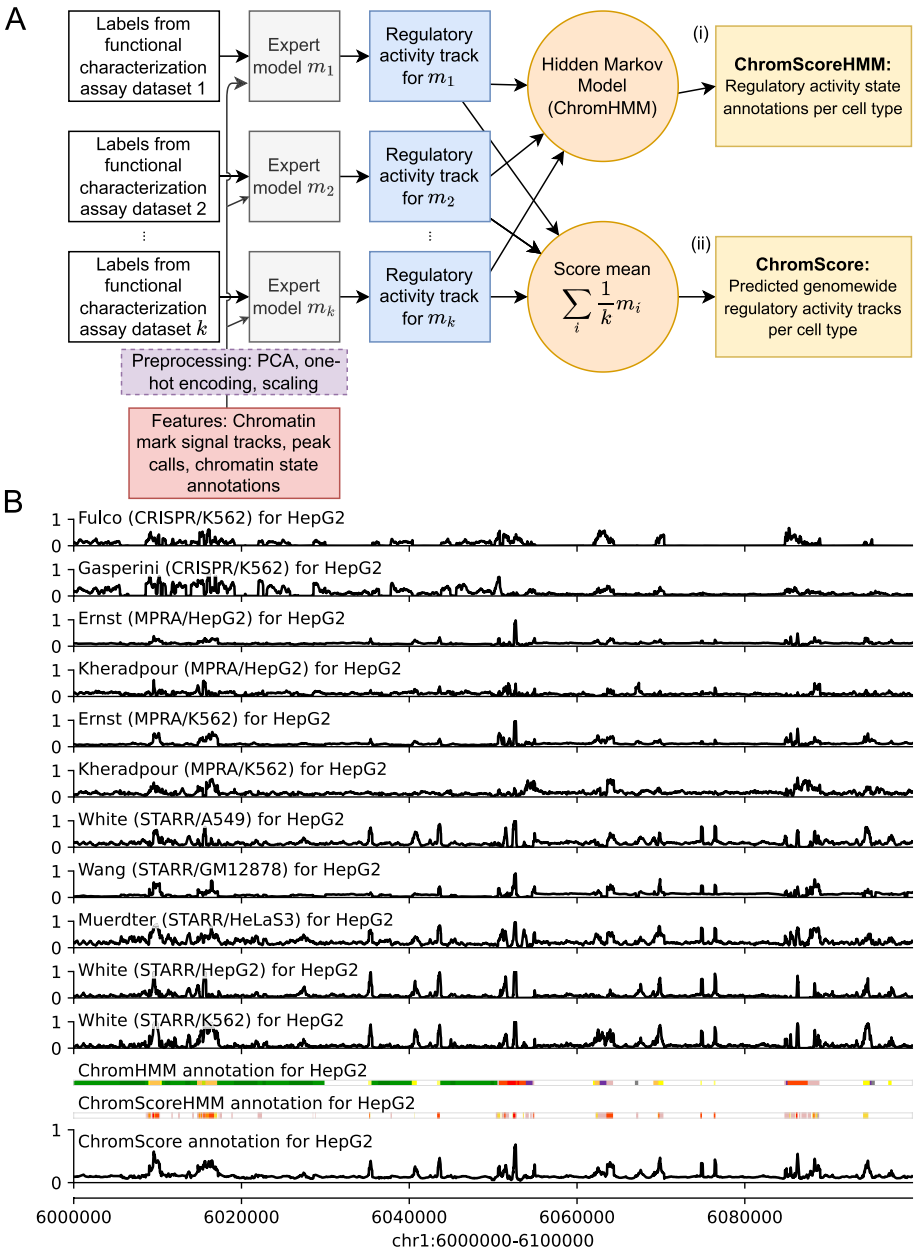

Fig. 1 Overview of the ChromActivity framework. **A** Flowchart of the ChromActivity framework. ChromActivity takes as input regulatory activity labels from targeted genomic regions from *k* different functional characterization datasets (stacked white blocks, upper left). Using features based on chromatin mark signals, peak calls, and chromatin state annotations for the targeted regions (red block, lower left) which it preprocesses (purple block, lower left), it trains a separate classifier ("expert") for each functional characterization dataset. Each expert provides a predicted genomewide regulatory activity score track specific to a functional characterization dataset (stacked blue blocks). ChromActivity then uses the score tracks to generate two complementary outputs reflecting predictions of regulatory activity for each cell type (yellow blocks, right): (*i*) ChromScoreHMM annotations, which are annotations of the genome into states generated by integrating combinatorial and spatial patterns in the expert prediction score tracks using ChromHMM and (*ii*) ChromScore tracks, which are continuous genomewide regulatory activity score tracks based on the mean individual expert scores at each 25-bp interval. **B** Visualization of regulatory activity score tracks for each expert, ChromHMM chromatin state annotations (25-state imputed model), the ChromScore track, and ChromScoreHMM annotations in HepG2 for genomic interval chr1:6,000,000–6,100,000 (hg19). ChromScoreHMM and ChromHMM color legends are shown in Additional file 1: Fig. S1

and then uses principal component analysis (PCA) to reduce the number of these additional signal features per mark from 81 to 3 (Methods). ChromActivity trains each logistic regression ensemble on a single functional characterization dataset, which we term ChromActivity experts. ChromActivity then applies these experts to make predictions across the entire genome in a large number of cell types, only one of which each expert would have seen training data from.

As individual expert predictions can in some cases disagree on predictions of regulatory activity, ChromActivity uses the individual expert predictions to generate genome annotations corresponding to combinatorial and spatial patterns of top predicted positions of regulatory activity from the different experts within a cell type. To do this, ChromActivity applies ChromHMM [13, 14] with input based on the different expert predictions to generate what we term ChromScoreHMM genome annotations (Methods). Relative to ChromHMM annotations, which are defined directly based on chromatin marks, these states are intended to more directly correspond to regions which have chromatin mark annotations predictive of regulatory activity in all or specific subsets of functional characterization datasets.

In addition to generating the ChromScoreHMM annotations, the ChromActivity framework includes a final step where predictions from different experts are averaged to generate ChromScore, a single genomewide regulatory activity potential score track for each cell type. ChromScore provides a numerical score between 0 and 1 of predicted regulatory activity potential for any 25-bp interval of the genome.

Training and evaluation of ChromActivity experts

We applied ChromActivity to imputed signal tracks and peak calls for ten histone modifications: H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me2, H3K4me3, H3K79me2, H3K9ac, H3K9me3, H4K20me1, histone variant H2 A.Z, and the DNase-I hypersensitivity (DNase) signal for 127 cell types from the Roadmap Epigenomics compendium [9, 39]. We used imputed data as it enabled us to apply our method with more marks across more cell types in a uniform manner than would be possible with observed data. We also included features based on a one-hot encoding of the 25-state ChromHMM chromatin state annotation that was previously trained on the same 12 imputed marks as used for our mark features.

We generated binary activating and neutral labels (Methods) for each genomic locus in 11 functional characterization datasets (Table S1). Five different cell types (A549 lung carcinoma, GM12878 lymphoblastoid, HeLa-S3 cervical carcinoma, HepG2 liver carcinoma, and K562 myelogenous leukemia cell types) were represented in the functional characterization datasets. Of the 11 datasets, two were CRISPR-dCas9-based assays (Fulco/K562 [38], Gasperini/K562 [25]). Additionally, there were nine plasmid-based assays (Methods), which we further classified into four MPRAs (Ernst/HepG2, Ernst/K562 [34], Kheradpour/HepG2, Kheradpour/K562 [33]) and five STARR-seq-derived assays (Muerdter/HeLaS3 [35], Wang/GM12878 [36], White/A549 [7], White/HepG2, and White/K562 [37]).

The total number of genomic loci used in training each individual expert ranged from 816 to 38,452 (Additional file 1: Fig. S2C). On average, 8.98% of genomic loci in a given dataset were within 100 bp of any locus in any other dataset. Across dataset pairs, this

overlap varied from 0.01% to complete overlap in the cases of Ernst/HepG2 with Ernst/K562 and Kheradpour/HepG2 with Kheradpour/K562. The fraction of DNase-I hypersensitive sites (Additional file 1: Fig. S2B) and the chromatin state distributions of the loci (Additional file 1: Fig. S2C) also varied across the datasets, which was expected as the datasets employed diverse strategies for selecting loci for testing.

While our main focus is cell type generalization, to establish reference points for the prediction difficulty for each dataset, we first evaluated predictive performance of the experts at distinguishing activating vs. neutral labeled loci in unseen partitions of the same dataset in which they were trained (Additional file 1: Fig. S3A). We found median out-of-sample prediction AUROCs ranged from 0.65 (Kheradpour/K562) to 0.93 (White/HepG2), with mean AUROC across all experts of 0.80. Expert predictive performance generally increased with the number of loci used in training (Spearman correlation: 0.75, Additional file 1: Fig. S3B). The expert models trained on the STARR-seq datasets Muerdter/HeLaS3, Wang/GM12878, White/A549, White/K562, and White/HepG2 all had relatively high median AUROCs (0.80, 0.83, 0.89, 0.91 and 0.93 respectively) compared to experts trained on other assay types (average AUROCs of 0.75 for MPRAs, 0.72 for CRISPR-based screens). In addition to the larger size of their training data, another possible contribution to the higher predictive performance of STARR-seq-based experts could be differences in the distribution of loci tested, which in the STARR-seq data include a broader and more diverse set of loci (Methods).

Genomewide expert predictions

For each of the 127 cell types, ChromActivity computed a score track for each expert predictor reflecting its genomewide regulatory activity predictions (Fig. 1B). We quantified the agreement among the individual expert regulatory activity scores based on the mean of pairwise Pearson correlations computed across the genome (Fig. 2A, Methods). The different expert predictions exhibited moderate agreement with an average pairwise Pearson correlation of 0.37 across all pairs of 11 score tracks and cell types. The pairwise correlations of experts ranged from -0.14 to 0.90, with the extremes corresponding to experts trained on the pair Gasperini/K562 (CRISPR-based) and Wang/GM12878 (STARR-seq) and the pair White/A549 (STARR-seq) and White/HepG2 (STARR-seq), respectively. We observed higher correlations within predictions from experts trained on plasmid-based (mean correlation 0.51) and CRISPR-based (correlation 0.52) functional characterization datasets than the correlations between plasmid-based and CRISPR-based experts (mean correlation 0.09).

Correspondingly, clustering the experts based on pairwise correlations of genomewide predictions revealed two main clusters (Fig. 2A). The first cluster included predictions from the two CRISPR-based experts, Fulco/K562 and Gasperini/K562. The second cluster included predictions from all but two plasmid experts (average pairwise Pearson correlation 0.67) and itself contained two subclusters. One subcluster included predictions from the three White lab experts and Muerdter/HeLaS3 (average correlation 0.86) and the other contained the Ernst/K562, Ernst/HepG2, and Wang/GM12878 experts (average correlation 0.67). Outside of the two main clusters, there were two experts, Kheradpour/K562 and Kheradpour/HepG2, which had low correlations with each other (0.22) and with predictions from other experts (average correlations 0.28 and 0.19).

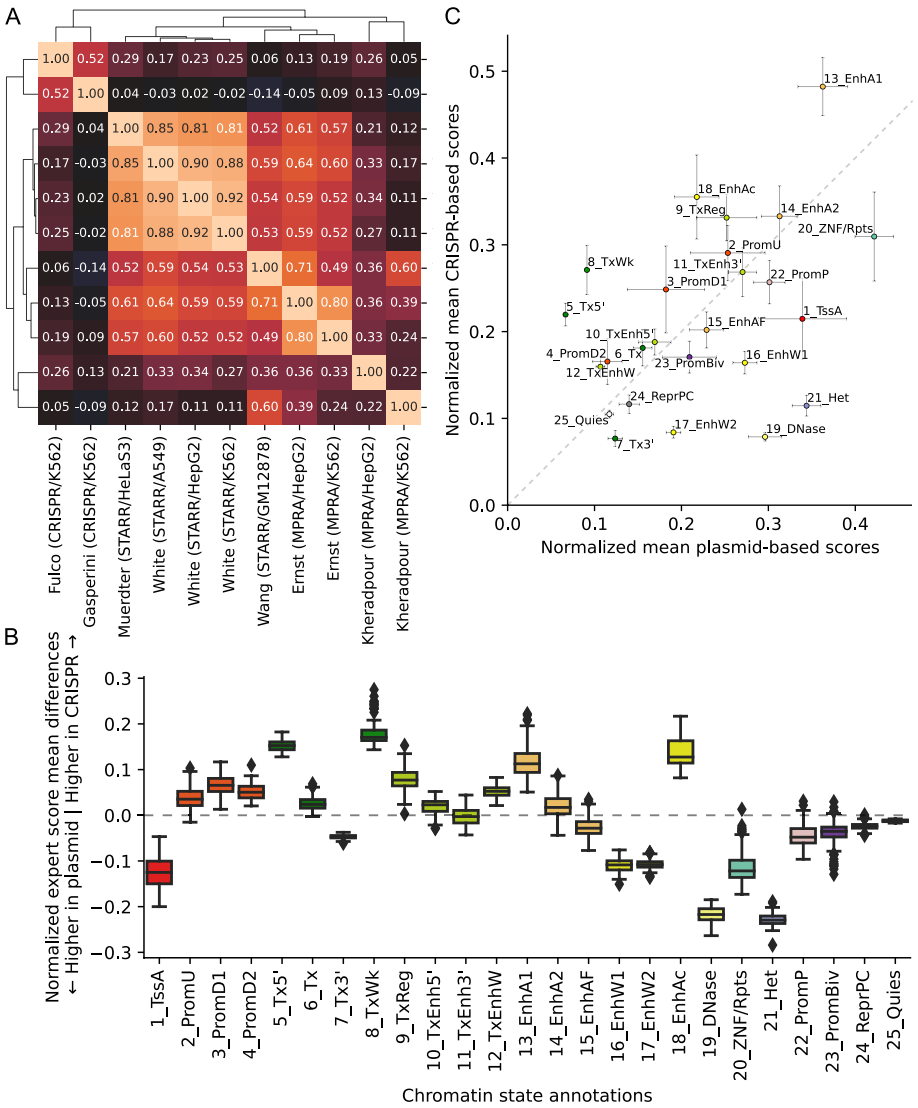

Fig. 2 Correlation of individual expert scores and comparison of plasmid-based and CRISPR-based experts. **A** Heatmap of mean genomewide Pearson correlations between expert model tracks clustered with hierarchical clustering, averaged over cell types. **B** Box plots of mean normalized score differences across cell types between experts trained on nine plasmid-based and two CRISPR-based functional characterization datasets in different ChromHMM chromatin states [13, 39]. The boxes represent quartiles and whiskers indicate maximum and minimum score differences between plasmid-based and CRISPR-based experts. Individual mean scores averaged across cell types, for each expert separately, is shown in Additional file 1: Fig. S5. The corresponding box plot distributions of means across cell types for each expert and each state is shown in Additional file 1: Fig. S6. Box colors correspond to the predefined ChromHMM imputed 25-state model colors. **C** Scatter plot of mean normalized expert scores for plasmid-based vs. CRISPR-based functional characterization datasets per chromatin state, averaged over cell types. Error bars indicate standard deviation of score means across cell types. Chromatin state abbreviations: active promoters (1_TssA, 2_PromU, 3_PromD1, 4_PromD2), transcribed regions (5_Tx5′, 6_Tx, 7_Tx3′, 8_TxWk), transcribed and regulatory regions (9_TxReg, 10_TxEnh5′, 11_TxEnh3′, 12_TxEnhW), active enhancers (13_EnhA1, 14_EnhA2, 15_EnhAF), weak enhancers (16_EnhW1, 17_EnhW2, 18_EnhAc), primary DNase (19_DNase), ZNF genes and repeats (20_ZNF/Rpts), heterochromatin (21_Het), poised/bivalent promoters (22_PromP, 23_PromBiv), repressed polycomb (24_ReprPC), and quiescent/low (25_Quies)

We investigated potential reasons for low or negative score correlations among some pairs of experts. We hypothesized that genomic regions corresponding to chromatin states not well represented in the training data could be associated with reduced correlations. To test this, we analyzed the genomewide score correlations of selected pairs of expert predictions after excluding genomic positions corresponding to individual or pairs of chromatin states (Additional file 1: Fig. S4). We focused on two expert pairs with the lowest pairwise correlations (Gasperini/CRISPR/K562 and Wang/STARR-seq/GM12878; Gasperini/CRISPR/K562 and Kheradpour/MPRA/K562) and two expert pairs that were of the same cell or assay type (Ernst/MPRA/K562 and Gasperini/CRISPR/K562; Gasperini/CRISPR/K562 and Fulco/CRISPR/K562) that had low correlations. We observed that removing particular combinations of transcription-associated chromatin states resulted in positive and higher correlations (Additional file 1: Fig. S4). Notably, chromatin states 8_TxWk and 5_Tx5' had few or no loci tested within the Gasperini/CRISPR/K562 dataset (Additional file 1: Fig. S2) and removing loci assigned to these states when computing pairwise correlations resulted in the largest correlation increases for these pairs that included Gasperini/CRISPR/K562.

We observed considerable variability in the chromatin states prioritized by different experts (Additional file 1: Fig. S5), notably between plasmid-based and CRISPR-based experts. For instance, regions overlapping the heterochromatin-associated 21_Het state had substantially greater normalized predicted regulatory activity based on the plasmid-based experts compared to CRISPR-based experts (Fig. 2B, Additional file 1: Fig. S6). This is consistent with DNA sequences that are active in the plasmid context but are repressed by H3K9me3 marked heterochromatin in the native chromatin context.

ChromScoreHMM genome annotations

To better understand the relationships between ChromActivity's expert model predictions and to generate an integrated genome annotation based on them, we developed ChromScoreHMM. ChromScoreHMM identifies combinatorial and spatial patterns within the expert predictions and uses these patterns to annotate the genome at 25-bp resolution. ChromScoreHMM starts by binarizing the expert model predictions based on a top percentage threshold computed separately for each expert in each cell type, which we set to 2% (Methods). It then uses the binarized predictions across the cell types as input to ChromHMM [13, 14] to learn a multivariate hidden Markov model (Fig. 3A, Additional file 1: Fig. S7). ChromScoreHMM learns a model across cell types using the concatenated approach of ChromHMM, leading to a shared set of states across cell types but cell type-specific assignments. The states capture distinct combinatorial and spatial patterns of expert predictions, and the resulting genome annotation is referred to as ChromScoreHMM annotations.

We focused our analysis on a ChromScoreHMM model with 15 states (Methods). We numbered the states in decreasing order of mean emission parameter values (Fig. 3A) and divided the states into three subgroups consisting of what we characterized as multi-expert states (states 1–10), single expert states (states 11–14), and the no expert state (state 15) (Additional file 1: Fig. S1). The multi-expert states all had at least two experts with emission probabilities ≥ 0.20. The single expert states all had a single expert with an emission probability of ≥ 0.90 and no other experts with

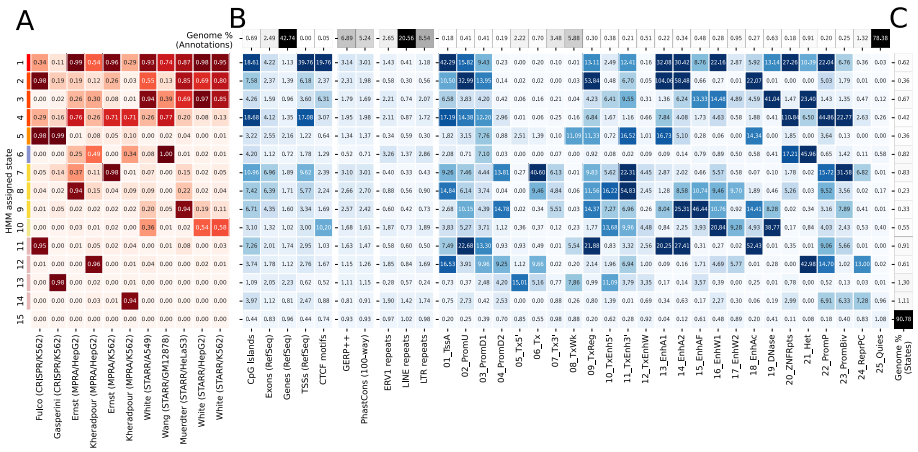

Fig. 3 ChromScoreHMM emission parameters and enrichments. **A** Emission parameters of a ChromScoreHMM model learned based on combinatorial and spatial patterns of top scoring predictions of each expert (top 2% of predictions, Methods). Each row of the heatmap corresponds to a ChromScoreHMM state (states 1–15, color legend on left margin) and each column a different input expert model. Emission parameter values correspond to the probability in that state of observing a top scoring prediction for that expert model. **B** Overlap fold enrichments for (1) sequence and gene based annotations: CpG islands [40], exons, gene bodies, and transcription start sites from RefSeq [41], CTCF motifs from HOMER [42], (2) evolutionary conservation related annotations: GERP++ [43] and PhastCons 100 vertebrates conserved elements [44], (3) ERV1, LINE and LTR repeat elements from RepeatMasker [45], (4) ChromHMM annotations, 25-state model [13, 39]. *Top row* ("Genome % (Annotations)"): Percentage of the genome covered by each annotation type, calculated by dividing the number of genomic base pairs covered by a given annotation by the total size of the genome. As some annotations can overlap each other, the sum of these percentages exceeds 100%. **C** Percentage of the genome assigned to each ChromScoreHMM state. See Additional file 1: Fig. S8 and Additional file 1: Fig. S9 for additional enrichments. Red shading: emission parameters, blue shading: fold enrichments, black shading: genome percentages. Enrichments and percentages are medians across cell types

emission probability ≥ 0.10. In the no expert state, all experts had < 0.001 emission probability. The multi-expert and single expert states covered in total 5.2 and 3.9% of the genome respectively, while the no expert state was by far the most common state covering 90.8% of the genome.

Among the multi-expert states, state 1 was the only state that had emission probabilities > 0.10 for all 11 experts. It had relatively high emission probabilities (> 0.50) for eight of the experts based on MPRA and STARR-seq datasets, and moderate emission probability (0.10–0.50) for the two experts based on CRISPR-dCas9 datasets and one MPRA dataset. State 2 had high emission probabilities for the Fulco/K562 (CRISPR-based) expert and all but one STARR-seq-based expert. State 3 had moderate or high emission probabilities for all the STARR-seq-based experts and two of the MPRA experts. State 4 had moderate or high emission probabilities for all experts except for White/HepG2. In contrast, state 5 was dominated by two CRISPR-based experts, with emission probabilities ≥ 0.98 for both, while the highest non-CRISPR expert emission probability was 0.10. States 6–9 had one expert that had a very high emission probability (≥ 0.95) and moderate emissions for one or two other experts. For instance, state 7 had very high (0.98) and moderate (0.37) emission probabilities for the Ernst/K562 and Ernst/HepG2 experts, respectively, while state 8 had high (0.94) and moderate (0.22) emission probabilities to the Ernst/HepG2 and Muerdter/

HeLaS3 experts, respectively. State 10 was uniquely associated with the three White lab STARR-seq datasets, with each associated expert having a moderate emission probability and other experts having low emission probability (< 0.05), suggesting that this state may be capturing aspects of this particular STARR-seq protocol or other types of batch effects.

The four single expert states (states 11–14) were each associated with one expert. States 11 and 13 were associated with the CRISPR-based experts Fulco/K562 and Gasperini/K562 respectively, while states 12 and 14 were associated with the experts for Kheradpour HepG2 and K562 datasets, respectively. The experts associated with the single expert states had below average pairwise score correlations with other experts (0.21 and 0.04 for Fulco/K562 and Gasperini/K562, respectively, and 0.19 and 0.28 for Kheradpour/K562 and Kheradpour/HepG2, respectively, compared to average pairwise correlation over all pairs of 0.37, Fig. 2A). These results suggest that these single expert states might be capturing dataset-specific signals or biases.

Enrichment analysis of ChromScoreHMM states

To better understand genomic properties of individual ChromScoreHMM states, we computed state enrichments for various genomic annotations (Fig. 3, Additional file 1: Figs. S8, S9). Some of the annotations used for the enrichments were also inputs to ChromActivity, specifically ChromHMM chromatin states (Fig. 3) and chromatin mark peak calls (Additional file 1: Fig. S8C). Other annotations were independent of ChromActivity's predictions, including CpG islands, CCCTC-binding factor (CTCF) motifs, transcription start sites, exons, gene bodies, various repeat elements, and evolutionarily conserved elements (Fig. 3). We also computed the proportion and fold enrichments of the ChromScoreHMM states in the genomic neighborhood of TSSs (Additional file 1: Fig. S10). Additionally, we computed the normalized average prediction score for individual experts in each state (Additional file 1: Fig. S9C).

Seven of the ChromScoreHMM states showed strong (> tenfold) enrichments for at least one of the active enhancer or flanking chromatin states 13_EnhA1, 14_EnhA2, or 15_EnhAF (Fig. 3), including 6 multi-expert (states 1, 2, 3, 5, 8, 9) and one single expert state (state 11, corresponding to the Fulco/K562 expert). State 2 (most associated with Fulco/K562, Muerdter/HeLaS3, and the White STARR-seq experts) had the strongest enrichments for the active enhancer states 13_EnhA1 and 14_EnhA2 among all the ChromScoreHMM states, with median fold enrichments of 104.0 and 58.5 fold respectively, while state 9 had the highest enrichment for the active enhancer flanking state, 15_EnhAF (46.4 fold).

Among all the states, state 1 (associated with broad expert activity) was most strongly enriched for both the TSS associated chromatin state 1_TssA (42.3 fold) and annotated TSSs themselves (39.8 fold), with a sharp peak in fold enrichment just around the TSS that decreases to approximately 1.7 fold 2 kb upstream and downstream of the TSS (Additional file 1: Fig. S10C). State 1 was also highly enriched for CpG islands (18.6 fold) and CTCF motifs (19.8 fold).

Eight states did not show strong enrichment for any of the active enhancer or flanking states. Of these, three of them still showed moderate enrichment for conserved bases (1.3–3.0 fold) including two multi-expert states (states 7, 10) and one single expert state

(state 12). State 4 (associated with moderate to high emissions for many experts) was also notable in that it was strongly enriched for CpG islands (18.7 fold) and for the chromatin states associated with poised promoters (22_PromP, 44.9 fold) and ZNF genes and repeats (20_ZNF/Rpts, 110.8 fold). State 7 (most associated with Ernst/K562 and Ernst/HepG2) was notable in that it showed the strongest enrichment for the bivalent promoter state (23_PromBiv, 31.6 fold) and also showed strong enrichment (> tenfold) for two transcribed states (6_Tx and 11_TxEnh3'). State 10, which was associated with a subset of the STARR-seq-based experts (White/A549, White/HepG2, White/K562), showed strong enrichment for a DNase-specific chromatin state (19_DNase, enrichment 38.8 fold). The presence of DNase without histone modifications is often associated with CTCF binding and candidate insulator regions [46]. Consistent with that, state 10 had a 10.2 fold enrichment for CTCF motifs.

Interestingly, some ChromScoreHMM states showed enrichment for both chromatin states associated with repression and activation. For instance, state 12 (the single expert Kheradpour/HepG2 state) was strongly enriched for the polycomb repressed chromatin state (24_ReprPC, 13.0 fold), the repressive heterochromatin associated chromatin state (21_Het, 43.0 fold) and the poised promoter state (22_PromP, 14.7 fold), but also the active TSS state (1_TssA, 16.5 fold). Similarly, state 3 was enriched for the repressive 21_Het state (23.4 fold) while also being enriched for moderately active states like 15_EnhAF (13.3 fold) and 16_EnhW1 (14.5 fold).

Three states (states 6, 14, and 15) were predominantly associated with repressive or quiescent genomic chromatin states. State 6 (most associated with Wang/GM12878, Kheradpour/HepG2, and Kheradpour/K562) had the strongest enrichment of any state for the 21_Het chromatin state (46.0 fold) and also for LTRs (2.9 fold) while having the strongest depletion of conserved bases (0.71 fold). State 14 (the single expert state for Kheradpour/K562) was enriched for the repressive poised promoter (22_PromP), bivalent promoter (23_PromBiv), and repressed polycomb (24_ReprPC) states (6.9, 6.3, and 7.3 fold respectively). Additionally, state 14 showed the weakest depletion for the Quiescent chromatin state (25_Quies) among single or multi-expert states (0.96 fold) with the Quiescent chromatin state comprising 75% of the state. State 15 (the no expert state) was the only ChromScoreHMM state to show enrichment for the Quiescent chromatin state (1.1 fold).

Notably, all ChromScoreHMM states with high emission parameter values for CRISPR-based experts (states 2, 5, 11, 13) were depleted for the 21_Het heterochromatin chromatin state (Fig. 3), which was not the case in general for states with high emission parameter values for plasmid-based experts. This is consistent with our analysis of individual CRISPR-based and plasmid-based experts (Fig. 2B), which showed 21_Het was being assigned higher scores by the plasmid-based experts compared to the CRISPR-based experts, likely marking regulatory sequences that are repressed in their native chromatin context.

ChromScoreHMM annotations displayed substantial variation in mean gene expression between states and specific positions relative to the TSS (Additional file 1: Fig. S11). Most ChromScoreHMM states were more enriched at TSSs of high expression genes than low expression genes (Additional file 1: Fig. S12). States 14 and 15, which were among the states associated with repressive or quiescent genomic regions, were

exceptions to this. State 6, which was also predominantly associated with repressive or quiescent chromatin states, was more enriched for low expression genes upstream and downstream of the TSS but was more enriched for high expression genes at the TSS (Additional file 1: Fig. S12B). Meanwhile, states 5 and 13, which are mainly associated with CRISPR-based experts, were more enriched downstream of the TSSs of high expression genes compared to low expression genes.

We also evaluated the enrichment of ChromScoreHMM states for experimentally determined transcription factor binding locations matched to the cell type of the annotations, which we would expect to enrich in bases of regulatory activity, and compared the enrichments to that of states of a ChromHMM model. For this, we focused on seven cell types with a large number of ChIP-seq experiments from ENCODE (A549, GM12878, H1 hESC, HeLaS3, HepG2, Huvec, K562). For each cell type, we computed enrichments for the center base of peaks after combining peak calls for all experiments in the cell type considered. We analyzed the cumulative fold enrichments of peak centers relative to the cumulative genome coverage of states when the states are ordered based on their enrichment. The ChromHMM model we used was the same 25-state model that was input to ChromActivity. Despite having fewer states, for six of the seven cell lines the most enriched ChromScoreHMM annotations were collectively able to have greater enrichment for transcription factor binding peak centers while covering more of the genome than a set of most enriched ChromHMM states (Additional file 1: Fig. S13). The one exception was H1 hESC, which may be partly because of increased TF binding in these cells without being associated with regulatory activity. Consistent with that, the chromatin state associated with DNase-I hypersensitivity without any active histone modifications (19_DNase) had a 40-fold enrichment for transcription factor binding peak centers compared to between 11- and 21-fold in other cell types considered. However, some of the reduced performance in H1 hESC might also be reflective of epigenome differences in this cell type relative to other cell types considered. Overall, these analyses suggests that ChromScoreHMM offers advantages in many cell types for identifying a limited set of high-confidence regulatory sites while ChromHMM provides complementary benefits in other cases.

ChromScore regulatory activity predictions

ChromActivity also averages the outputs of its individual expert predictions to generate a cell type specific regulatory activity score, termed ChromScore (Fig. 4A, Methods). ChromScore provides a single continuous score track for each cell type, where higher scores correspond to higher average predicted regulatory activity potential (Fig. 4A).

We investigated if ChromScore, which was trained based on functional characterization assay data in a limited number of cell types, would generalize to new cell types without functional characterization data. To evaluate the cell type generalization performance of ChromScore to predict regulatory activity in unseen cell types, first we generated modified versions of our ensemble models in which functional characterization datasets of each cell type were removed from the training data. Next, we generated and evaluated ChromScore tracks for the held-out cell types using the modified models at loci not seen in training (Fig. 4B, Additional file 1: Fig. S14).

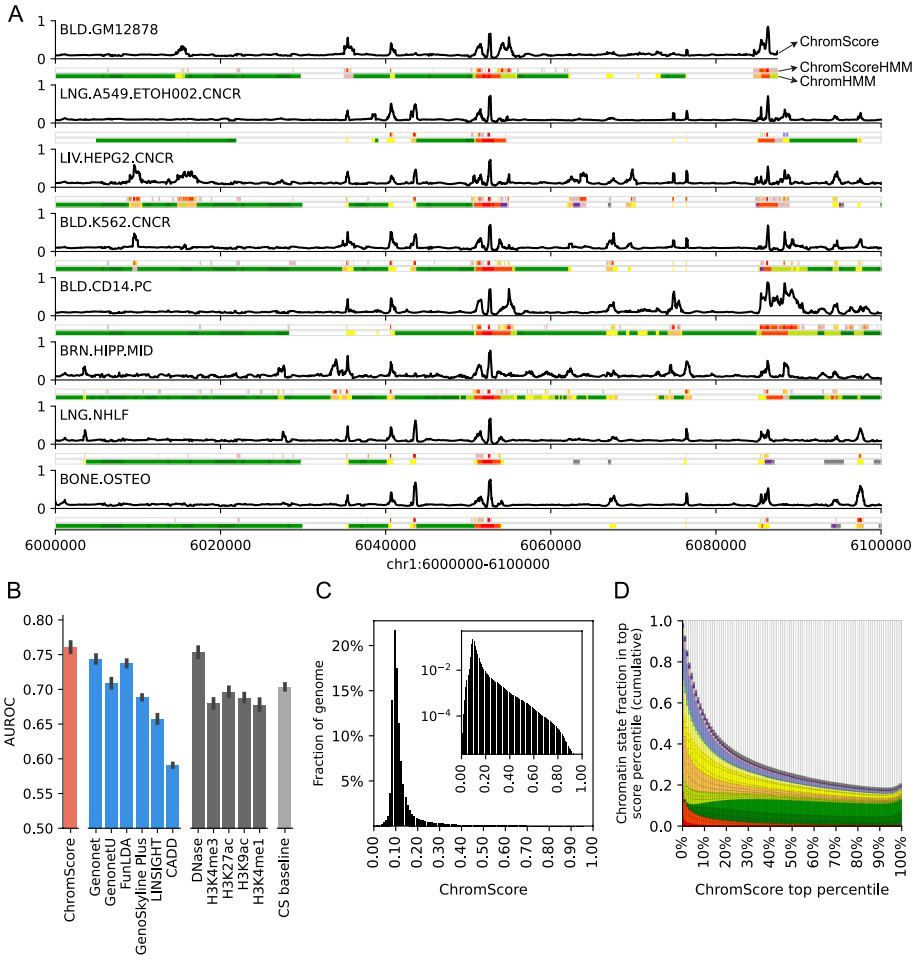

Fig. 4 ChromScore tracks, cell type generalization performance evaluations and score distributions. **A** Visualization of ChromScore tracks in eight cell types shown above ChromScoreHMM and ChromHMM annotations in the same cell types for genomic interval chr1:6,000,000–6,100,000 (hg19). The cell types shown represent examples of both those with and without functional characterization training data (cell types with training data: GM12878, A549, HepG2, and K562; cell types without training data: CD14 primary monocytes, brain hippocampus cells, NHLF lung fibroblast primary cells, and osteoblast primary cells). **B** A comparison of cell type generalization performance of ChromScore to existing scores, single marks, and a chromatin state baseline. The bars correspond to the mean area under receiver operator characteristic (AUROC) across 11 functional characterization datasets. The first bar shows the performance of ChromScore. For ChromScore evaluations, expert models trained on the same cell type as the evaluation dataset were not used. The next six bars show the performance of existing scores [27, 47–51], which are followed by bars for the imputed signal tracks for DNase I hypersensitivity, H3K4me3, H3K27ac, H3K9ac, and H3K4me1. The last bar shows the mean ensemble of the chromatin state baseline models for all datasets (CS baseline, Methods). Error bars indicate standard error across evaluations. **C** Genomewide distribution of ChromScore values, averaged over cell types. Inset: log scaled. **D** Cumulative chromatin state fraction for top ChromScore percentiles. Each bin corresponds to an additional top 1% of scores. See Additional file 1: Fig. S1 for chromatin state color legend

We compared the ChromScore predictions to a set of baselines and existing score tracks from other methods for predicting activating vs. neutral labels of loci tested with functional characterization assays. The baselines included those based on individual chromatin marks and one based on chromatin state assignments (Methods). The existing score tracks included several scores that provided cell type-specific

regulatory activity estimates integrating multiple epigenomic datasets (GenoNet, GenoNet-U [27], FunLDA [47], Genoskyline Plus [48]). In addition, we compared to two scores that also integrate epigenomic annotations but do so in a non-cell type-specific manner and consider a diverse set of other annotations, CADD [49] and LINSIGHT [50] (Methods).

ChromScore predictions had a substantially higher mean AUROC score (0.76) relative to all the baselines (AUROC range 0.67–0.70) except to DNase signal for which it was marginally better (0.75) (Fig. 4B). Among the existing scores we evaluated, AUROCs ranged from 0.59 (CADD) to 0.74 (Genonet and FunLDA). Comparing ChromScore's predictive performance to its underlying experts indicated that ChromScore performed similar to or better than the highest scoring experts in the majority of evaluations (Additional file 1: Fig. S15). ChromScore performed better in plasmid-based dataset evaluations (mean AUROC 0.79) compared to CRISPR-based dataset evaluations (mean AUROC 0.59, Additional file 1: Fig. S16), possibly because it was trained on more plasmid-based datasets. Notably, while ChromScore and DNase signal showed similar cell type generalization performances, they were only moderately correlated across cell types (median Spearman correlation 0.26, Additional file 1: Fig. S17). Furthermore, the chromatin state distributions of top ChromScore regions differed considerably from top DNase regions (Additional file 1: Fig. S18) particularly for the 20_ZNF/Rpts (18.01 vs. 0.11 fold), 2_PromU (12.19 vs 29.20 fold), and 3_PromD1 (7.08 vs. 34.05 fold) states.

We compared ChromScore predictions when training and using features based on observed data rather than imputed data, which yielded equivalent cell type generalization performance (Additional file 1: Fig. S19). We also investigated the relative importance of subsets of features on cell type generalization performance with a feature ablation study (Additional file 1: Fig. S20). Removing individual features linked to specific marks, removing chromatin state features or removing peak features all had small effects on AUROCs (0.73–0.74, compared to 0.76). However, removing all signal features led to a more substantial decrease in prediction performance (0.70). These analyses suggest that it would be possible to use a smaller set of features than we used here and obtain comparable predictive performance, but also value of at least having multiple signal features. Removing DNase signal tracks, DNase peaks calls, and chromatin state annotations, which were in part-based on DNase signal tracks, from the model leads to only a modest decrease in cell type generalization performance, indicating that the model's overall performance has limited dependence on DNase-based features.

The median ChromScore across the genome and all 127 cell types was 0.10 (Fig. 4C), with top-scoring genomic regions (highest 2% genomewide) having a ChromScore >0.35 on average. A number of the chromatin states with high mean ChromScores (Fig. 5C) and high fold enrichments within top-scoring genomic regions (Additional file 1: Fig. S18A) across cell types included states typically associated with regulatory activity, such as 13_EnhA1 (mean score 0.41, fold enrichment 29.40 fold), 14_EnhA2 (mean score 0.35, fold enrichment 23.43), and 1_TssA (mean score 0.32, fold enrichment 17.46). Interestingly, other chromatin states such as 20_ZNF/Rpts (mean score 0.41, 18.01 fold) and 21_Het (mean score 0.32, 14.34 fold) also displayed high mean

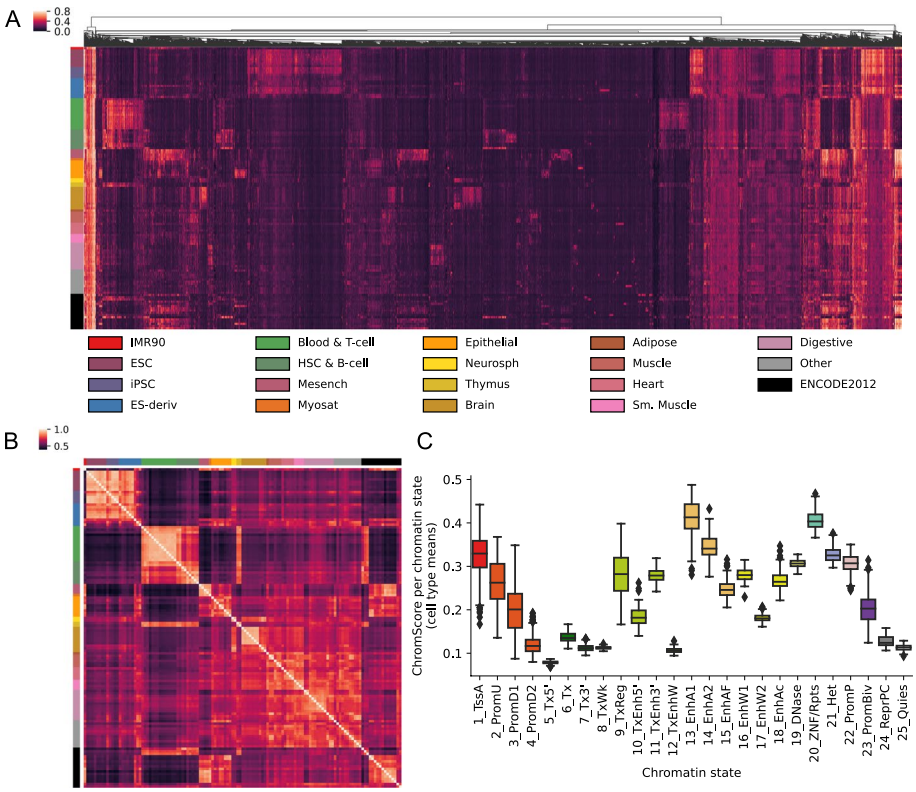

Fig. 5 ChromScore across cell types. **A** Heatmap showing ChromScores at 20,000 randomly selected bases across the genome (columns) that had a score difference of > 0.25 between at least two cell types for 127 Roadmap Epigenomics cell types (rows) (Methods). Columns are hierarchically clustered and rows are sorted based on Roadmap Epigenomics tissue groups [9]. The tissue groups of the rows are indicated on the left and their color legend is displayed at the bottom. **B** Heatmap of ChromScore Pearson correlations across all pairs of 127 cell types, which are ordered and colored as in A. **C** Distribution of mean ChromScores per chromatin state per cell type

scores and top-scoring region enrichments. The high mean scores and top-scoring region enrichments of 20_ZNF/Rpts and 21_Het appeared to be mainly driven by plasmid-based experts which, as previously shown, were more likely to assign higher scores on average to 20_ZNF/Rpts and 21_Het-annotated genomic regions (Additional file 1: Figs. S5, S6).

Analyzing ChromScore across many cell types enabled us to identify some genomic loci that were predicted to show near-universal activity across a diverse range of cell types. Approximately 0.19% of loci across the genome were predicted to be highly active (top 2% ChromScore) in over 90% of all cell types. We also observed that cell types that were more biologically similar had greater correlation in their ChromScore (Fig. 5 A, B). In particular, cell types within the same Roadmap Epigenomics tissue group [9] had an average Pearson correlation of 0.80 compared to a correlation of 0.62 for predictions crossing different tissue groups, reflecting ChromScore's ability to capture cell and tissue-specific behavior.

We analyzed fold enrichments for genomic repeat elements in top-scoring ChromScore regions (top 2%, Additional file 1: Fig. S21A, E), and observed enrichments for long terminal repeats (LTRs, fold enrichment 1.71), particularly the endogenous

retroviral sequence 1 (ERV1) subclass (fold enrichment 2.04), and depletions for long interspersed nuclear elements (LINEs, fold enrichment 0.56) and short interspersed nuclear elements (SINEs, fold enrichment 0.41). The enrichment for LTRs is consistent with previous reports showing LTRs association with activating gene expression [34, 52–54]. Top DNase regions by signal, in comparison, were depleted for LTRs, LINEs, and SINEs (Additional file 1: Fig. S21D, H). Plasmid-based experts and CRISPR-based experts prioritized different repeat classes, with LTRs being enriched in bases prioritized by plasmid-based experts but depleted in CRISPR-based experts and SINEs including the subclass of Alu elements showing the opposite trend (Additional file 1: Fig. S21 B, C, F, G). This could suggest LTRs being repressed in the genome but drive expression in a plasmid context. The enrichment of Alus in bases prioritized by CRISPR-based experts is consistent with the enrichment of both for transcribed regions [55] (Additional file 1: Fig. S5).

ChromScore moderately correlated with in vivo gene expression at the TSS (Pearson correlation 0.41, Additional file 1: Fig. S22A, Methods). However, some chromatin marks showed stronger correlations around the TSS, such as H3K9ac (Pearson correlation 0.59 at TSS + 500 bp). We note that correlations for ChromScore were not necessarily expected to surpass that of all chromatin marks with expression, since ChromScore heavily relies on plasmid-based experts, which while providing an assessment of the inherent regulatory activity of a DNA sequence, do not reflect the full in vivo chromatin context. Correlation patterns varied across functional characterization assay types and individual experts (Additional file 1: Fig. S22B), with CRISPR-based experts showing higher correlations upstream and downstream of the TSS (Additional file 1: Fig. S22C) but lower correlations at the TSS compared to plasmid-based experts (Mean Pearson correlations 0.32 for plasmid-based experts, 0.23 for CRISPR-based experts, Additional file 1: Fig. S22D). This observation is consistent with the lower CRISPR expert scores observed in the 1_TssA chromatin state, which is primarily associated with active TSSs, compared to those of plasmid-based experts, as well as the higher scores observed in upstream promoter (2_PromU) and downstream promoter (3_PromD1, 4_PromD2) chromatin states (Additional file 1: Fig. S5). These findings highlight the distinct patterns among ChromScore expert tracks in predicting gene expression around the TSS.

Discussion

We introduced ChromActivity, a computational framework that predicts gene regulatory element activity across diverse cell types by integrating information from chromatin marks and multiple functional characterization datasets. ChromActivity first trains a set of experts with each expert trained on a different individual functional characterization dataset. It then applies these trained predictors to make predictions for each cell type. Using these predictions, ChromActivity produces two complementary integrative outputs for each cell type. One of them is ChromScoreHMM, which annotates the genome into states representing combinatorial and spatial patterns in the expert's regulatory activity track predictions. The other is ChromScore, which is a cell type-specific continuous numerical score of predicted regulatory activity potential across the genome based on combining the individual expert predictions. We applied ChromActivity using

chromatin mark data from 127 cell types in the Roadmap Epigenomics compendium and data from 11 functional characterization datasets.

We observed that different experts prioritized different subsets of the genome, in some cases corresponding to the assay or experimental protocol of the functional characterization dataset it was trained on. For example, plasmid-based experts on average assigned higher regulatory activity prediction scores to H3K9me3 heterochromatin-associated genomic regions compared to CRISPR-based experts, which was expected as the plasmid-based experts were trained based on loci tested outside of their native chromatin context (Fig. 2B). These differences enabled us to distinguish genomic regions with likely H3K9me3-associated repressive activity from inactive regions. We also observed differences between CRISPR-based and plasmid-based experts in terms of their correlations with gene expression at and around the TSS and their predictions of regulatory activity for different classes of repeat elements. Given these differences, specific applications may benefit from utilizing either plasmid-based or CRISPR-based expert predictions, or different ChromScoreHMM states. For example, plasmid-based expert tracks and associated ChromScoreHMM states could be preferred for applications focused on predicted regulatory activity inherent in genomic sequences, independent of any regulatory effect of chromatin marks, while the CRISPR-based expert tracks and associated ChromScoreHMM states could be preferred for applications focused on predicted regulatory activity in the native genomic context.

Some of the ChromScoreHMM states corresponded to genomic regions with predicted regulatory activity in different types of functional characterization assays, while others were more specific to a specific assay or likely associated with dataset-specific signals or biases. Notably, ChromScoreHMM states explicitly capture genomic regions that are more active in CRISPR-dCas9 or STARR-seq assays. We showed that ChromScoreHMM states corresponded to substantial enrichment differences for various annotations, including gene annotations, repeat elements, chromatin states, and chromatin mark peaks. Further, the spatial distribution of ChromScoreHMM states relative to the TSSs of nearby genes varied depending on the expression of the genes. As expected, most states were more enriched at or around the TSSs of high expression genes compared to low expression genes, except for the few states associated with repressive or quiescent genomic regions.

ChromScoreHMM, while building on the ChromHMM method, provides a distinct genome annotation that complements ChromHMM annotations. In particular, ChromScoreHMM annotations are defined based on combinatorial and spatial patterns in supervised predictions of regulatory activity corresponding to different functional characterization datasets, while ChromHMM annotations are defined directly based on the combinatorial and spatial patterns of chromatin marks. ChromScoreHMM annotations thus more directly correspond to different classes of predicted regulatory activity, while ChromHMM annotations can capture chromatin mark patterns not expected to correspond to differences in regulatory activity reflected in functional characterization assays. In particular, high emission parameters for a state in a ChromScoreHMM model can be directly interpreted to be associated with high predicted regulatory activity based on one or more functional characterization datasets, which is not the case for ChromHMM models. In addition, the ChromScoreHMM

annotations were generated at 25-bp resolution, higher than the 200-bp resolution typically used for ChromHMM annotations. A challenge with higher resolution annotations with ChromHMM is the positions with locally highest signal for histone modifications are more likely to be the positions of nucleosomes and less likely the actual regulatory bases. By using supervised information, ChromScoreHMM can predict a position has greater regulatory activity even if it does not have a greater value for any input feature. This is likely reflected in the greater cumulative enrichments and genome coverage among the most enriched states for the center of transcription factor binding peaks in all but one of the cell types evaluated. The exception of H1 hESC also highlighted potential complementary advantages of ChromHMM.

ChromScore is based on an ensemble of predictors trained on a variety of functional characterization datasets thus avoiding an overreliance on the biases associated with any one dataset. We demonstrated the generalizability of ChromScore predictions across cell types through evaluations of predictive performance in unseen cell types. Top ChromScore regions were highly enriched for enhancer chromatin states as well as classes of repeat elements previously shown to be associated with regulatory activation [34, 54]. We also showed that the predictions across 127 cell types exhibited cell type-restricted activity corresponding to known biological groupings of cell types.

There are several potential avenues for future work building on the current ChromActivity framework. One avenue would be to expand the set of functional characterization datasets used as input to ChromActivity, including adding additional recent CRISPR-based ones and datasets from additional cell types. A challenge to incorporating many additional functional characterization datasets in addition to availability has been the lack of uniform processing. However, this is changing with additional uniformly processed datasets beginning to accumulate in repositories [56], facilitating their inclusion in future models. A second avenue for future work would be to develop an improved way to combine expert predictions into a score other than the current strategy of averaging of predictions. This could potentially involve an approach that assigns different weights to different experts globally, for instance based on an estimated level of the noise for the labels on which they were trained, or in a locus-specific manner based on how similar the locus is to those for which the experts were trained. Future extensions of ChromActivity could potentially improve performance in CRISPR-based assays by assigning larger weights to CRISPR-based experts or including additional CRISPR-based. Another avenue of future work would be to extend ChromActivity to directly predict repressive loci. We designed ChromActivity primarily to predict activation as the information in the functional characterization datasets that we considered for repression was more limited and inconsistent. However, some functional characterization datasets are informative of repression [34] and could be incorporated into an extended framework that explicitly models repression. ChromActivity focuses on predicting regulatory activity independent of target genes, a potential direction for future work is to evaluate the use of its predictions as input to complementary frameworks to map distal regulatory elements to their targets [4, 38, 57]. Future work could also investigate applying ChromActivity to additional cell types in human as well as to non-human species. However, we expect ChromActivity to already be a resource for analyzing and interpreting the human regulatory genome across diverse cell types.

Conclusions

The ChromActivity framework provides integrative annotation of regulatory activity across diverse human cell types by combining data from multiple chromatin marks using supervised information from functional characterization assays. Specifically, the framework generates ChromScoreHMM states that can capture both assay-specific and broadly shared predictions of regulatory activity in addition to the summary ChromScore regulatory activity prediction tracks. The ChromScoreHMM state annotations and ChromScore tracks constitute what we expect to be a valuable resource for analyzing gene regulation across many human cell types, and we expect future applications of ChromActivity to further extend the coverage of cell types and conditions with such annotations.

Methods

Dataset selection and label extraction

We derived labeled training data from 11 functional characterization datasets for ChromActivity (Table S1). All datasets were of experiments in cell types for which there was matched uniformly processed chromatin mark data available from the Roadmap Epigenomics consortium. The chromatin mark data for these cell types were all originally generated by the ENCODE project consortium [7].

The individual datasets we used in abbreviated notation are as follows: Ernst/HepG2, Ernst/K562 [34], Kheradpour/HepG2, Kheradpour/K562 [33], Muerdter/HeLaS3 [35], Wang/GM12878 [36], White/A549 [7, 58], White/HepG2, White/K562 [37], Fulco/K562 [38], and Gasperini/K562 [25]. The cell types covered by the individual datasets are as follows: A549 lung carcinoma (epigenome identifier E114), GM12878 lymphoblastoid (epigenome identifier E116), HeLa-S3 cervical carcinoma (epigenome identifier E117), HepG2 liver carcinoma (epigenome identifier E118), and K562 myelogenous leukemia (epigenome identifier E123).

ChromActivity treats predicting regulatory activity as captured by functional characterization assays as a binary classification task and specifically focuses on differentiating activating regions from assumed neutral regions. For input into ChromActivity, we defined binary activating vs. neutral labels for each genomic region in each functional characterization dataset using dataset specific procedures described below. A subset of the datasets reported repressive sequences in addition to neutral and activating (Ernst/HepG2, Ernst/K562, Kheradpour/HepG2, Kheradpour/K562). For these datasets, for consistency with other datasets that did not measure repression, we still treated it a binary classification task but excluded reported repressive regions while training the corresponding expert. We also provided ChromActivity a "reference nucleotide" within each region used for training, which we selected as the base we considered most likely representative of the regulatory activity. The specific procedure for selecting the base (e.g., center of construct, nucleotide with the highest signal) depended on the functional characterization dataset and is described below.

The Ernst/HepG2 and Ernst/K562 datasets [34] used a dense tiling of MPRA constructs combined with the SHARPR computational method to assign continuous regulatory activity scores to 5-bp intervals within the tiled regions. For each individual tiled

region, we identified the position with absolute maximum value and defined S_{absmax} of the region to be the value at that position if non-negative and otherwise the negative of it. We assigned activating labels to tiled regions with S_{absmax} values exceeding 1, and neutral labels to regions with S_{absmax} values between -1 and 1. We filtered any regions with scores under -1 to exclude likely repressive regulatory regions from the training dataset. This procedure yielded 2405 activating and 10,894 neutral regions for Ernst/HepG2 and 2519 activating and 10,162 neutral regions for Ernst/K562. The reference nucleotide for each region was the center base of the 5-bp interval with the highest absolute maximum SHARPR score.

For the MPRA datasets Kheradpour/HepG2 and Kheradpour/K562 [33], we used the precomputed *p*-values associated with regulatory activity for each construct against scrambled controls. The constructs with regulatory activity under the expressed *p*-value threshold of 0.05 were labeled activating and the rest were labeled neutral. This yielded 541 activating and 1548 neutral regions for Kheradpour/HepG2 and 347 activating and 1742 neutral regions for Kheradpour/K562. The reference nucleotide for each region was the center nucleotide of the sequence motif originally used in the experimental design, which also was the center nucleotide of the construct. We excluded any synthetic sequences not represented in the genome from the dataset.

For the STARR-seq-based datasets White/A549, White/HepG2, and White/K562, we obtained STARRPeaker 1.0 [37] peak calls with ENCODE accessions ENCFF646OQS, ENCFF047LDJ, and ENCFF045TVA respectively. We assigned activating labels to the top 10% of the peak calls by the normalized signal output/input fold change value. For the neutral regions, we randomly selected bases from the genome, excluding any that overlapped the ENCODE list of excluded regions [59]. For each activating region, we picked three neutral regions from the genome. This procedure yielded 6929 activating and 20,787 neutral regions for White/A549, 5199 activating and 15,597 neutral regions for White/HepG2, and 3571 activating and 10,713 neutral regions for White/K562. The reference nucleotide for each region was the center nucleotide of the peak, which corresponded to the base with the highest normalized signal output/input fold change value.

For Muerdter/HeLaS3 [35], which was also a STARR-seq dataset, we obtained peak calls from https://data.starklab.org/publications/muerdter_boryn_2017/peaks_inhibitor_correctedEnrichment4_supp.table3.tsv, which corresponded to STARR-seq peaks with corrected fold-enrichment values above 4, the threshold used for peak calling in [35]. We applied the same random regions selection procedure as above to generate three neutral regions for each activating region. This yielded 9613 activating and 28,839 neutral regions for Muerdter/HeLaS3. The reference nucleotide was the center of the peak for each region.

The Wang/GM12878 dataset [36] is based on a combined experimental and computational functional characterization method called High-resolution Dissection of Regulatory Activity (HiDRA). The experimental part of the method is based on a variant of STARR-seq called ATAC-STARR-seq, which first applies a selection step based on ATAC-seq to identify regions of open chromatin and then applies STARR-seq to these selected regions instead of the whole genome. For Wang/GM12878, we first obtained peak calls for "HiDRA driver elements" identified by the HiDRA-SHARPR2 pipeline and the HiDRA RNA/DNA ratio score track (GEO accession GSE104001). We then assigned

activating labels to the driver elements with RNA/DNA ratios above 1. To generate the neutral regions, we randomly selected nucleotides from "HiDRA tiled regions" that were not also in "HiDRA active regions," maintaining a label ratio of three neutral regions for each activating region. This yielded 2409 activating and 7227 neutral regions for Wang/GM12878. The reference nucleotide was the center of the peak for each region.

For Gasperini/K562, a CRISPR-dCas9 dataset [25], we obtained the data for the scaled-up experiment from GEO (accession GSE120861), which included genomic regions targeting DNase-I hypersensitive sites with various combinations of H3K27ac, p300, GATA1, and RNA Pol II binding. We filtered gRNA readouts to only include predefined target regions that resulted in a decrease in a candidate gene's expression (regression coefficient "beta" column < 0) and excluded loci that were flagged in the "outlier_gene" column. We followed the methodology provided in the paper [25] to aggregate gRNA readouts to gRNA groups targeting the same locus. Target loci with adjusted empirical *p*-values below 0.05 for any of the measured target genes were labeled activating. Regions that failed to reach that threshold for any genes were labeled neutral. This procedure yielded 432 activating and 5122 neutral regions for Gasperini/K562. The reference nucleotide was the midpoint of a target region.

For Fulco/K562, also a CRISPR-dCas9 dataset, we obtained the published adjusted *p*-values associated for tested candidate regulatory element-gene pairs (E–G pairs) in K562 [38]. This dataset contains aggregated data from 10 CRISPR-based functional characterization studies [60–69], with the vast majority of data points (> 99%) generated by perturbation with the CRISPRi-FlowFISH screen, which makes use of CRISPR-dCas9 with an attached KRAB domain, targeting DNase-I hypersensitive sites within 450 kb of 30 selected genes. We used the same procedure as Ref. [38] to exclude any E–G pairs that (i) had less than 80% power to a detect 25% effect on gene expression or (ii) had a fraction change in gene expression that was positive after CRISPR interference, since it suggests repression. We used a *p*-value threshold of 0.05 to assign the activating and neutral labels E–G pairs. Candidate regulatory elements that were in at least one E–G pair were assigned to the activating label, while all other elements were assigned the neutral label. This yielded 69 activating and 747 neutral regions for Fulco/K562. The reference nucleotide was the center nucleotide of the element.

Genomic coordinates not in hg19 were converted to hg19 using the liftOver utility from the UCSC genome browser [70], specifically from hg18 for Kheradpour HepG2/K562 datasets and from hg38 for White A549/HepG2/K562 datasets. All training, testing, and analysis was done in the hg19 human genome assembly, except we have also provided as a resource hg38 liftOver ChromScoreHMM and ChromScore annotations. For all datasets, loci not in chromosomes 1 through 22 or chromosome X were filtered out.

Feature extraction and preprocessing

ChromActivity uses three classes of features in the models: chromatin mark signals, chromatin mark peak calls, and ChromHMM chromatin states. For the chromatin signal and peak call features, we used imputed signal tracks and narrow peak calls on imputed signal tracks, respectively, for the following 12 chromatin marks: DNase I hypersensitivity (DNase), H2A.Z, H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me2,

H3K4me3, H3K79me2, H3K9ac, H3K9me3, H4K20me1 [9]. For the chromatin state features, we used chromatin state annotations based on the 25-state ChromHMM model based on imputed data [13, 39]. All these features were available across 127 cell types.

For each region considered, ChromActivity extracts the signal features within a 2-kb window at 25 base intervals centered around a reference nucleotide associated with the region, yielding 81 features per mark. In our application here, this resulted in 972 intermediate signal features for the 12 marks. For each mark, ChromActivity then applies principal component analysis (PCA) to its 81 signal features and selects the top three principal components. In our application here, the first three principal components explained on average 97% of the variance across training regions. ChromActivity retains the original signal value at the reference nucleotide thus reducing the number of signal features from 81 to 4 per mark. In evaluations and analyses that involved dividing the dataset into training and test partitions, PC component weights were learned from training partitions in each dataset and then applied to the test partitions.

For each chromatin mark peak, ChromActivity includes a binary indicator variable for the presence of the peak at the reference nucleotide. It also includes features corresponding to a one-hot encoding of the 25-chromatin state annotation at the reference nucleotides. Altogether, this procedure yields 85 features used for classification: 36 PCA signal features, 12 original signal value features, 12 chromatin mark peak features, and 25 one-hot encoded chromatin state annotations features. All features are standardized (based on the training partition for evaluations involving train and test sets) to have mean zero and a variance of one before training.

Training, evaluation and genomewide prediction track generation of the expert models

In the supervised learning component of ChromActivity, ChromActivity uses a bagging ensemble of regularized logistic regression classifiers to generate the individual experts, which has the advantages of being robust and providing well-calibrated probability estimates that reflect the class membership of the training data. For each functional characterization dataset, ChromActivity trained an ensemble of classifiers based on the extracted labels and features as described above. Each ensemble contained 100 binary logistic regression classifiers with a L2-norm penalty trained on a random drawing of training data points. The data points were drawn with replacement to obtain the same number of data points as the initial training set, i.e., a bagging ensemble. The regularization strength C of the logistic regression classifiers was set to the default value of 1 and assigned label weights of $\mathrm{w}(\mathrm{y} = \textit{activating}) = \mathrm{n}_{\mathrm{neutral}}$ and $\mathrm{w}(\mathrm{y} = \textit{neutral}) = 3\mathrm{n}_{\mathrm{activating}}$ to the label classes, where $w(\mathrm{y} = y')$ indicates label class weight for y', and $n_{y'}$ indicates number of data points of label y'. The label weights correspond to an effective label ratio of 0.25 (activating/(activating+neutral)) across different datasets (Additional file 1: Fig. S2A) instead of a balanced ratio so the resulting score better highlights genomic regions with high regulatory activity potential.

To evaluate the predictive performance of ChromActivity's expert models on the functional characterization datasets they were trained on, we randomly generated 20 train/test partitions per dataset with a 4:1 train:test ratio, stratified by label to ensure

consistent label ratios across the partitions. The models were trained as described above and applied to each test set to obtain AUROC metrics in Additional file 1: Fig. S3.

To produce the genomewide expert score tracks, ChromActivity applied the experts (trained on the entire training data) at 25-bp intervals across the genome to predict the activating class label probability at the center nucleotide in the interval (i.e., 13th nucleotide). ChromActivity produced ChromScore tracks by taking the mean value of the individual expert predictions for each 25-bp interval.

We computed a normalized version of the expert scores for analyses in which the distributions of the expert model scores are directly compared on the same sets of genomic loci (Fig. 2B, C, Additional file 1: Figs. S5, S6, S9). The normalization procedure we implemented was based on quantile normalization. Specifically, to establish the reference distribution, we first computed expert model scores for 10 million randomly selected genomic locations, removing regions in the ENCODE excluded list [59]. We sorted the expert scores and computed the median expert score for each ranked entry. We then computed 1000 quantile bins of each expert score distribution and generated mappings from the quantile bins to the corresponding median expert scores. Score values from experts are mapped to normalized score values using these mappings. We computed the mean normalized expert score values over all experts to generate the normalized ChromScore track used in Additional file 1: Fig. S9.

ChromScoreHMM annotations

To generate the ChromScoreHMM annotations, ChromActivity first converts the continuous score tracks associated with expert models into binarized input for ChromHMM (version 1.23). These annotations are generated at 25 bp resolution, corresponding to the resolution of the predictions, instead of the default ChromHMM resolution of 200 bp. For the main analysis, the binarization threshold per score track was set such that the 25-bp bins within the top 2% of model scores were assigned to 1 and the rest were assigned to 0.

We used ChromHMM's LearnModel subcommand with the following command line flags: -b 25 -n 128 -p 4 -d -1 -lowmem. This configuration corresponds to a score bin size of 25 bases, using 128 randomly selected cell type and chromosome combinations per Baum-Welch training iteration, 4 threads running the standard Baum-Welch algorithm, with the change in estimated log-likelihood stopping criterion disabled and reduced memory usage mode. The number of chromatin states for the main analysis was set to 15. Emission and transmission parameters of the model are shown in Additional file 1: Fig. S7.

To determine the number of states and the binarization threshold we ran models with the number of chromatin states set to 10, 15, and 25 and binarization thresholds of top 1, 2, 5, and 10%. We focused on a 15-state model as it provided a good balance between model expressivity and interpretability for multiple values of the binarization threshold. The binarization threshold presented a tradeoff: a higher binarization threshold risks missing a larger number of true regulatory sites or evidence that a regulatory site is supported by multiple expert's top predictions, while a lower binarization threshold could over-assign the genome into regulatory states (Additional file 1: Fig. S23). We opted to use a binarization threshold of 2%, which provided a reasonable tradeoff with approximately

9.3% of the genome in a cell type on average in ChromScoreHMM states associated with at least one expert (Fig. 3) and 5.0% of the 25-bp intervals in the genome were above the binarization threshold in two or more experts (Additional file 1: Fig. S23).

ChromScoreHMM overlap fold enrichments

We computed overlap fold enrichments using ChromHMM's OverlapEnrichment command with command line flag -b 25. CpG island coordinates and RefSeq gene coordinates [41] were the ones included with ChromHMM (version 1.23), originally downloaded from the UCSC genome browser. RefSeq annotations were the version available on July 26, 2015. RepeatMasker [45] repeat element coordinates and PhastCons 100-way conserved element annotations [44] were obtained from the UCSC Genome Browser [40, 71]. CTCF motif instances were obtained from HOMER known motifs (track version 191020) [42]. GERP++ conserved element annotations were obtained from http://mendel.stanford.edu/SidowLab/downloads/gerp [43].

Transcription factor binding enrichment analysis

We conducted a comparative analysis of enrichment for the center base of transcription factor binding peaks between ChromScoreHMM and ChromHMM state annotations. For this comparison, we used the annotations from the 15-state ChromScoreHMM model and the annotations from the 25-state imputation-based ChromHMM model, the latter of which was input features to ChromActivity. We focused on annotations from seven reference epigenomes corresponding to seven cell lines (A549 (E114), GM12878 (E116), H1 hESC (E003), HeLa-S3 (E117), HepG2 (E118), Huvec (E122), and K562 (E123)) based on the availability extensive uniformly processed ChIP-seq data from a prior ENCODE consortium compendium [7]. Specifically, we collected ChIP-seq from this directory https://hgdownload.soe.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgTfbsUniform/ which had 35 experiments for A549, 90 for GM12878, 58 for H1 hESC, 64 for HeLa-S3, 77 for HepG2, 14 for Huvec, and 150 for K562.

For each cell type considered, we separately concatenated peaks for all the ChIP-seq experiments into a single file. We then used the OverlapEnrichment command of ChromHMM to compute enrichments for the concatenated peaks of each cell type both for the ChromHMM and ChromScoreHMM annotations. We used the "-center" flag to have the enrichment based on the center of the base of the peak. We used the "-multicount" flag to count a base multiple times in the overlap if it was the center base of multiple peaks in the concatenated file. For ChromScoreHMM enrichment, we used the "-b 25" to specify the annotation was at 25 bp resolution while for ChromHMM we used the default "-b 200" option.

Separately for each cell type, we sorted the 15-ChromScoreHMM states based on their enrichment for the peak centers. We then plotted on the x-axis the cumulative genome coverage of the states based on this ordering. On the y-axis, we plotted the cumulative fold enrichment based on the same ordering. The cumulative fold enrichment was computed by first multiplying for each state the percentage of the genome the state covers by its fold enrichment for peak centers to obtain the percentage of peak centers the state contains. The cumulative sum of those percentages was then computed. Finally, the

cumulative sum of percentage of peak centers was divided by the cumulative percentage of genome coverage to obtain cumulative fold enrichments.

Analysis of expert score and ChromScore distributions

Pairwise expert score correlations

We computed pairwise Pearson correlations between pairs of expert scores at 500,000 randomly selected bases of the genome, excluding any region on the ENCODE excluded regions list v2 for hg19 [59].

Chromatin state score distributions

To determine chromatin state score distributions for the 25-state ChromHMM annotations [13, 39], we sampled 2.5 million loci from the genome excluding those in ENCODE excluded regions v2 as above and extracted their chromatin states and associated scores.

Cluster heatmap of ChromScores across cell types and tissue group correlations

To generate a cluster heatmap of scores across cell types, we randomly selected 20,000 bases from the genome among those for which ChromScore showed a difference of at least 0.25 between at least one of the 127 cell types. We filtered for score differences to highlight genomic loci with different regulatory activity potential across cell types. Roadmap Epigenomics tissue groupings were obtained from the metadata section of the Roadmap Epigenomics data portal [9]. Loci were clustered using the euclidean average linkage metric implemented in scipy.cluster.hierarchy.linkage in the SciPy package [72]. We excluded the "ENCODE2012" and "Other" tissue groupings when computing the mean ChromScore correlations within and across tissue groups.

Evaluating ChromScore cell type generalization performance

To estimate the generalization performance of ChromScore in unseen cell types, we trained five modified versions of the model, one for each cell type with characterization data available. Each version was constructed such that it did not have access to training data in one particular cell type (i.e., one of A549, GM12878, HeLa-S3, HepG2, or K562). To evaluate predictive performance for a dataset of a particular cell type, we used the version of the model with that cell type removed. In addition to holding out cell types, we also spatially partitioned the genome into 5-kb chunks and assigned each chunk to the training or testing partition with probability 0.75 (i.e., 3:1 train:test ratio). We repeated this process 20 times per dataset.

We compared the performance of ChromScore to a set of baselines and existing scores. The baselines included the individual imputed chromatin mark signals (DNase, H2A.Z, H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me2, H3K4me3, H3K79me2, H3K9ac, H3K9me3, and H4K20me1) in the matched cell types obtained from the Roadmap Epigenomics compendium. In addition, they included a simple chromatin state baseline model, which generated a single score track for each cell type by mapping a chromatin state annotation at a specific position to the average fraction of positive labels within the training partition for each dataset.

We also compared ChromScore to various cell type-specific and non-cell type-specific external scores that integrate different epigenomic datasets and in some cases with other

annotations, specifically FunLDA [47], GenoSkyline Plus [48], LINSIGHT [50], CADD [49, 51], and GenoNet/GenoNet-U [27]. Precomputed FunLDA scores were downloaded from http://www.funlda.com/download. GenoSkyline Plus annotations were obtained from http://zhaocenter.org/GenoSkyline. LINSIGHT annotation was downloaded from http://compgen.cshl.edu/LINSIGHT. CADD v1.4 scores were obtained from https://cadd.gs.washington.edu. We used the browser track (hg19) version of the CADD scores, which are based on the highest scoring single-nucleotide variant for each genomic position as described in https://github.com/kircherlab/CADD-browserTracks.

GenoNet used two distinct models, a supervised version ("GenoNet") which was trained on MPRA data [33, 73] and was only applied to the three cell types K562, HepG2, GM12878, and an unsupervised version ("GenoNet-U") which did not use any functional characterization data and was applied to the remaining 124 Roadmap Epigenome cell types. Precomputed GenoNet scores for K562, HepG2, and GM12878 and precomputed Genonet-U scores for the remaining Roadmap Epigenomics cell type were obtained from https://zenodo.org/record/3336208 [27, 74]. Precomputed GenoNet-U scores in K562, HepG2, and GM12878 were not available, and instead we computed using a custom script based on the description of the method. The output of our implementation was confirmed to produce nearly identical predictions (Pearson correlation >0.99) to the GenoNet-U scores in all 124 of Roadmap epigenome cell types for which it was available. For cell types for which a supervised GenoNet score was not available, we used the GenoNet-U score as the GenoNet score.

To evaluate cell type generalization performance of the ChromScore model trained on observed rather than imputed data, we used observed histone modification signals and peak annotations processed by the Roadmap Epigenomics consortium [9] for data for the cell types K562, HepG2, A549, GM12878, and HeLa-S3 originally generated by the ENCODE Consortium. All these cell types had observed data for all marks in addition to a functional testing dataset we considered. We excluded chromatin state features during training so the models would be more directly comparable between observed and imputed data. Consistent with our previous approach, we removed experts trained on cell types matching the evaluation cell type.

Expression analyses around TSSs

We downloaded the RPKM expression matrix for protein coding genes for 56 Roadmap Epigenomes from the Roadmap Epigenomics data portal (https://egg2.wustl.edu/roadmap/data/byDataType/rna/expression/57epigenomes.RPKM.pc.gz), along with the corresponding Ensembl gene annotations (Ensembl v65, hg19) [75]. For each cell type, we categorized the genes into high expression (defined as $\log_2(\text{RPKM}+1)>1$) and low expression (defined as $\log_2(\text{RPKM}+1)<0.01$) genes. Across the 56 cell types, 62% of genes were categorized as high expression and 15% of genes were categorized as low expression on average.

To investigate expression of nearby genes for ChromScoreHMM states, we identified the ChromScoreHMM states within 24-kb windows centered around the TSSs of the genes, sampled at 200 base intervals. For genes on the negative strand, we flipped the position indices so that positive offset values always corresponded to the direction of the gene body. We computed $\log_2(\text{RPKM}+1)$ values for each gene based on the Roadmap

Epigenomics RPKM expression matrix for protein coding genes. For each ChromScore-HMM state and position offset, we then computed the mean $\log_2(\text{RPKM} + 1)$ value across genes and cell types.

For the ChromScore expression correlations analysis, we first computed ChromScore and individual expert model scores within a 24-kb window centered around TSSs in all cell types with expression data available. We elected to center our intervals around the TSS, as distance to the TSS in general can be a major confounder for analyses reporting association between distal loci and gene expression. We also extracted chromatin mark signal values for the same windows for comparison. We mirrored the score windows for the genes on the negative strand around the TSS to align the upstream segments and the gene bodies. We then computed Pearson correlations between scores or signal values and log expression with a pseudocount ($\log_2(\text{RPKM} + 1)$) for each 25-bp interval centered around the TSS and averaged them over the cell types.

ChromScore repeat element enrichments

We downloaded RepeatMasker [45] repeat elements from the UCSC Genome Browser [40, 71], using the repClass column to identify the LINE, SINE, and LTR elements and the repFamily column to identify the ERV and Alu elements. We randomly selected 1 million nucleotides from hg19 on chromosomes 1 through 22 and chromosome X, determined if they overlapped a repeat element with the bedtools intersect command, and computed ChromScore and expert prediction scores for each plasmid-based and CRISPR-based expert for the nucleotides. Mean plasmid and CRISPR scores were obtained by taking the mean score of the respective experts at each nucleotide. We grouped the scores within 200 quantiles (i.e., each quantile representing 0.5% of the nucleotides) and computed fold enrichments for each quantile for each repeat type compared to the genome background.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-025-03579-6.

Additional file 1: Figures S1–S23, Table S1.

Additional file 2. Review history.

Acknowledgements

We thank Heather Han for conducting some related preliminary work and members of the Ernst lab for insightful discussions. We acknowledge the ENCODE and Roadmap Epigenomics Consortia and the individual labs that generated the data that we used.

Peer review information

Andrew Cosgrove was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions

TD implemented ChromActivity, compiled and curated input datasets, performed analyses, and wrote the paper. JE conceived and conceptualized the study, suggested analyses, and wrote the paper. All authors read and approved the final manuscript.

Funding

US National Institutes of Health (DP1DA044371, U01MH130995, U01MH105578, UH3 NS104095, U01HG012079); US National Science Foundation (1254200, 2125664); Rose Hills Innovator Award, and the UCLA Jonsson Comprehensive Cancer Center and Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research Ablon Scholars Program.

Data availability
The ChromActivity software, model weights and links to the ChromScoreHMM and ChromScore annotations (hg19 and hg38 liftOver) for the 127 Roadmap cell types are available on https://github.com/ernstlab/ChromActivity and archived on Zenodo under https://doi.org/10.5281/zenodo.15009259 [76, 77]. The software is available under an MIT license. Signal and score tracks were processed using pyBigWig v0.3.18 [78] and UCSC utilities v369 [70, 79]. Genomic coordinates were mapped across genome assemblies using liftOver v369 [70, 79]. Analyses involving genome intervals used BedTools v2.30.0 [80], pyBedTools v0.9.0 [81], and bedops v2.4.41 [82] packages. Hierarchical clustering used in heatmap visualizations is implemented in SciPy [72]. scikit-learn v1.1.2 [83] was used for data preprocessing, model training, inference, and evaluations. ChromScoreHMM annotations were generated using ChromHMM v.1.23 [13, 14]. The ChromHMM software is available from https://ernstlab.github.io/ChromHMM/. We used matplotlib [84] and Seaborn [85] for plotting and visualization. All other packages were obtained from the conda-forge and bioconda [86] repositories.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 13 July 2023 Accepted: 15 April 2025
Published online: 09 May 2025

References

1. Chatterjee S, Ahituv N. Gene regulatory elements, major drivers of human disease. Annu Rev Genomics Hum Genet. 2017;18:45–63.
2. Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, et al. The human transcription factors. Cell. 2018;172:650–65.
3. Gasperini M, Tome JM, Shendure J. Towards a comprehensive catalogue of validated and target-linked human enhancers. Nat Rev Genet. 2020;21:292–310.
4. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. 2011;473:43–9.
5. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012;337:1190–5.
6. Barski A, Cuddapah S, Cui K, Roh T-Y, Schones DE, Wang Z, et al. High-resolution profiling of histone methylations in the human genome. Cell. 2007;129:823–37.
7. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74.
8. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, et al. The accessible chromatin landscape of the human genome. Nature. 2012;489:75–82.
9. 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.
10. Stunnenberg HG, Abrignani S, Adams D, de Almeida M, Altucci L, Amin V, et al. The International Human Epigenome Consortium: a blueprint for scientific collaboration and discovery. Cell. 2016;167:1145–9.
11. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods. 2013;10:1213–8.
12. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based Analysis of ChIP-Seq (MACS). Genome Biol. 2008;9: R137.
13. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nat Methods. 2012;9:215–6.
14. Ernst J, Kellis M. Chromatin-state discovery and genome annotation with ChromHMM. Nat Protoc. 2017;12:2478–92.
15. Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nat Methods. 2012;9:473–6.
16. Libbrecht MW, Chan RCW, Hoffman MM. Segmentation and genome annotation algorithms for identifying chromatin state and other genomic patterns. PLOS Comput Biol. 2021;17: e1009423.
17. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, et al. An atlas of active enhancers across human cell types and tissues. Nature. 2014;507:455–61.
18. Andersson R, Sandelin A. Determinants of enhancer and promoter activities of regulatory elements. Nat Rev Genet. 2020;21:71–87.
19. Klein JC, Agarwal V, Inoue F, Keith A, Martin B, Kircher M, et al. A systematic evaluation of the design and context dependencies of massively parallel reporter assays. Nat Methods. 2020;17:1083–91.
20. Chen PB, Fiaux PC, Zhang K, Li B, Kubo N, Jiang S, et al. Systematic discovery and functional dissection of enhancers needed for cancer cell fitness and proliferation. Cell Rep. 2022;41: 111630.
21. Inoue F, Ahituv N. Decoding enhancers using massively parallel reporter assays. Genomics. 2015;106:159–64.

22. Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, et al. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nat Biotechnol. 2012;30:271–7.
23. Gallego Romero I, Lea AJ. Leveraging massively parallel reporter assays for evolutionary questions. Genome Biol. 2023;24:26.
24. Arnold CD, Gerlach D, Stelzer C, Boryń ŁM, Rath M, Stark A. Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science. 2013;339:1074–7.
25. 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.
26. Chong ZX, Yeap SK, Ho WY. Transfection types, methods and strategies: a technical review. PeerJ. 2021;9: e11165.
27. He Z, Liu L, Wang K, Ionita-Laza I. A semi-supervised approach for predicting cell-type specific functional consequences of non-coding variation using MPRAs. Nat Commun. 2018;9:5199.
28. Sethi A, Gu M, Gumusgoz E, Chan L, Yan K-K, Rozowsky J, et al. Supervised enhancer prediction with epigenetic pattern recognition and targeted validation. Nat Methods. 2020;17:807–14.
29. Kreimer A, Zeng H, Edwards MD, Guo Y, Tian K, Shin S, et al. Predicting gene expression in massively parallel reporter assays: a comparative study. Hum Mutat. 2017;38:1240–50.
30. Kreimer A, Yan Z, Ahituv N, Yosef N. Meta-analysis of massively parallel reporter assays enables prediction of regulatory function across cell types. Hum Mutat. 2019;40:1299–313.
31. Zhou J, Theesfeld CL, Yao K, Chen KM, Wong AK, Troyanskaya OG. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. Nat Genet. 2018;50:1171–9.
32. Movva R, Greenside P, Marinov GK, Nair S, Shrikumar A, Kundaje A. Deciphering regulatory DNA sequences and noncoding genetic variants using neural network models of massively parallel reporter assays. PLoS ONE. 2019;14: e0218073.
33. 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.
34. 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.
35. 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.
36. 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.
37. 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.
38. 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.
39. Ernst J, Kellis M. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. Nat Biotechnol. 2015;33:364–76.
40. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, et al. The UCSC Table Browser data retrieval tool. Nucleic Acids Res. 2004;32(Database issue):D493-496.
41. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44:D733-745.
42. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010;38:576–89.
43. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLOS Comput Biol. 2010;6: e1001025.
44. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005;15:1034–50.
45. Smit A, Hubley R, Green P. RepeatMasker Open-4.0 [Internet]. 2013. Available from: http://www.repeatmasker.org.
46. Vu H, Ernst J. Universal annotation of the human genome through integration of over a thousand epigenomic datasets. Genome Biol. 2022;23:9.
47. Backenroth D, He Z, Kiryluk K, Boeva V, Petukhova L, Khurana E, et al. FUN-LDA: a latent dirichlet allocation model for predicting tissue-specific functional effects of noncoding variation: methods and applications. Am J Hum Genet. 2018;102:920–42.
48. Lu Q, Powles RL, Abdallah S, Ou D, Wang Q, Hu Y, et al. Systematic tissue-specific functional annotation of the human genome highlights immune-related DNA elements for late-onset Alzheimer's disease. PLOS Genet. 2017;13: e1006933.
49. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019;47:D886–94.
50. Huang Y-F, Gulko B, Siepel A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. Nat Genet. 2017;49:618–24.
51. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46:310–5.
52. Bannert N, Kurth R. Retroelements and the human genome: new perspectives on an old relation. Proc Natl Acad Sci. 2004;101(suppl_2):14572–9.
53. Criscione SW, Zhang Y, Thompson W, Sedivy JM, Neretti N. Transcriptional landscape of repetitive elements in normal and cancer human cells. BMC Genomics. 2014;15: 583.
54. Ali A, Han K, Liang P. Role of transposable elements in gene regulation in the human genome. Life. 2021;11: 118.
55. Zhang X-O, Pratt H, Weng Z. Investigating the potential roles of SINEs in the human genome. Annu Rev Genomics Hum Genet. 2021;22:199–218.
56. Luo Y, Hitz BC, Gabdank I, Hilton JA, Kagda MS, Lam B, et al. New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. Nucleic Acids Res. 2020;48:D882–9.

57. Liu Y, Sarkar A, Kheradpour P, Ernst J, Kellis M. Evidence of reduced recombination rate in human regulatory domains. Genome Biol. 2017;18:193.
58. White K. ENCSR895FDL. ENCODE data portal. 2020. https://doi.org/10.17989/ENCSR895FDL.
59. Amemiya HM, Kundaje A, Boyle AP. The ENCODE Blacklist: identification of problematic regions of the genome. Sci Rep. 2019;9:9354.
60. Thakore PI, D'Ippolito AM, Song L, Safi A, Shivakumar NK, Kabadi AM, et al. Highly specific epigenome editing by CRISPR-Cas9 repressors for silencing of distal regulatory elements. Nat Methods. 2015;12:1143–9.
61. Xu J, Shao Z, Li D, Xie H, Kim W, Huang J, et al. Developmental control of polycomb subunit composition by GATA factors mediates a switch to non-canonical functions. Mol Cell. 2015;57:304–16.
62. Fulco CP, Munschauer M, Anyoha R, Munson G, Grossman SR, Perez EM, et al. Systematic mapping of functional enhancer–promoter connections with CRISPR interference. Science. 2016;354:769–73.
63. Ulirsch JC, Nandakumar SK, Wang L, Giani FC, Zhang X, Rogov P, et al. Systematic functional dissection of common genetic variation affecting red blood cell traits. Cell. 2016;165:1530–45.
64. Wakabayashi A, Ulirsch JC, Ludwig LS, Fiorini C, Yasuda M, Choudhuri A, et al. Insight into GATA1 transcriptional activity through interrogation of cis elements disrupted in human erythroid disorders. Proc Natl Acad Sci U S A. 2016;113:4434–9.
65. Klann TS, Black JB, Chellappan M, Safi A, Song L, Hilton IB, et al. CRISPR–Cas9 epigenome editing enables high-throughput screening for functional regulatory elements in the human genome. Nat Biotechnol. 2017;35:561–8.
66. Liu SJ, Horlbeck MA, Cho SW, Birk HS, Malatesta M, He D, et al. CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. Science. 2017;355: eaah7111.
67. Xie S, Duan J, Li B, Zhou P, Hon GC. Multiplexed engineering and analysis of combinatorial enhancer activity in single cells. Mol Cell. 2017;66:285-299.e5.
68. Huang J, Li K, Cai W, Liu X, Zhang Y, Orkin SH, et al. Dissecting super-enhancer hierarchy based on chromatin interactions. Nat Commun. 2018;9:943.
69. Qi Z, Xie S, Chen R, Aisa HA, Hon GC, Guan Y. Tissue-specific gene expression prediction associates vitiligo with SUOX through an active enhancer. bioRxiv. 2018;2:337196.
70. Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. BigWig and BigBed: enabling browsing of large distributed datasets. Bioinformatics. 2010;26:2204–7.
71. Navarro Gonzalez J, Zweig AS, Speir ML, Schmelter D, Rosenbloom KR, Raney BJ, et al. The UCSC Genome Browser database: 2021 update. Nucleic Acids Res. 2021;49:D1046–57.
72. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17:261–72.
73. Tewhey R, Kotliar D, Park DS, Liu B, Winnicki S, Reilly SK, et al. Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. Cell. 2016;165:1519–29.
74. Ionita-Laza I. GenoNet scores for human genome assembly GRCh37, Zenodo. 2019. https://doi.org/10.5281/ZENODO.3336208.
75. Cunningham F, Allen JE, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, et al. Ensembl 2022. Nucleic Acids Res. 2022;50:D988–95.
76. Dincer TU, Ernst J. ChromActivity. Zenodo. 2025. https://doi.org/10.5281/zenodo.15009259.
77. Dincer TU, Ernst J. ChromActivity. Github. 2025. https://github.com/ernstlab/chromactivity.
78. Ryan D, Gökçen Eraslan, Grüning B, Betts E, Ramirez F, Nezar Abdennur, et al. deeptools/pyBigWig: 0.3.18 [Internet]. Zenodo; 2021 [cited 2023 Apr 3]. Available from: https://zenodo.org/record/4515486.
79. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome Res. 2002;12:996–1006.
80. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.
81. Dale RK, Pedersen BS, Quinlan AR. Pybedtools: a flexible python library for manipulating genomic datasets and annotations. Bioinformatics. 2011;27:3423–4.
82. Neph S, Kuehn MS, Reynolds AP, Haugen E, Thurman RE, Johnson AK, et al. BEDOPS: high-performance genomic feature operations. Bioinformatics. 2012;28:1919–20.
83. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in python. J Mach Learn Res. 2011;12:2825–30.
84. Hunter JD. Matplotlib: a 2D graphics environment. Comput Sci Eng. 2007;9:90–5.
85. Waskom M. seaborn: statistical data visualization. J Open Source Softw. 2021;6:3021.
86. Grüning B, Dale R, Sjödin A, Chapman BA, Rowe J, Tomkins-Tinch CH, et al. Bioconda: sustainable and comprehensive software distribution for the life sciences. Nat Methods. 2018;15:475–6.

Publisher's Note