

Learning a pairwise epigenomic and transcription factor binding association score across the human genome

Soo Bin Kwon^{1,2, } and Jason Ernst^{1,2,3,4,5,6,7,* }

¹Bioinformatics Interdepartmental Program, University of California, Los Angeles, CA 90095, United States

²Department of Biological Chemistry, University of California, Los Angeles, CA 90095, United States

³Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research, University of California, Los Angeles, CA 90095, United States

⁴Computer Science Department, University of California, Los Angeles, CA 90095, United States

⁵Computational Medicine Department, University of California, Los Angeles, CA 90095, United States

⁶Jonsson Comprehensive Cancer Center, University of California, Los Angeles, CA 90095, United States

⁷Molecular Biology Institute, University of California, Los Angeles, CA 90095, United States

*Corresponding author. Department of Biological Chemistry, University of California, Los Angeles, 615 Charles E Young Drive South, Los Angeles, CA 90095.

Email: jason.ernst@ucla.edu

Associate Editor: Janet Kelso

Abstract

Motivation: Identifying pairwise associations between genomic loci is an important challenge for which large and diverse collections of epigenomic and transcription factor (TF) binding data can potentially be informative.

Results: We developed Learning Evidence of Pairwise Association from Epigenomic and TF binding data (LEPAE). LEPAE uses neural networks to quantify evidence of association for pairs of genomic windows from large-scale epigenomic and TF binding data along with distance information. We applied LEPAE using thousands of human datasets. We show using additional data that LEPAE captures biologically meaningful pairwise relationships between genomic loci, and we expect LEPAE scores to be a resource.

Availability and implementation: The LEPAE scores and the software are available at <https://github.com/ernstlab/LEPAE>.

1 Introduction

Studying genomic loci in the context of pairwise relationships can be informative about the individual loci as well as the larger genomic regions in which they lie (Ernst *et al.* 2011, Thurman *et al.* 2012, Liu *et al.* 2017, Gazal *et al.* 2022). One prominent example of pairwise relationships used to study genomic loci is chromatin interactions identified by assays such as Hi-C (Lieberman-Aiden *et al.* 2009, Rao *et al.* 2014, Mifsud *et al.* 2015). While informative, such assays miss relationships associated with physical interactions beyond the detection thresholds of the assays as well as relationships that are not linked with physical interactions. For example, two loci might be associated in that they are part of the transcribed region of the same gene but do not physically interact.

An alternative strategy to identify pairwise relationships between loci has been based on correlating epigenomic assay signals at distal loci directly or with gene expression across cell types, often focusing on linking enhancers with target genes (Ernst *et al.* 2011, Thurman *et al.* 2012, Corradin *et al.* 2014, He *et al.* 2014). While correlations based on epigenomic mark signals have also been informative, they are limited in the use of available information in large and diverse compendia of

genome-wide maps of histone modifications and variants, chromatin accessibility, chromatin state annotations, and transcription factor (TF) binding and thus are likely limited in the types of associations they capture. Correlation-based measures also cannot effectively reveal associations between two distal loci that have constitutive activity.

To address these issues, we developed a method, Learning Evidence of Pairwise Association from Epigenomic and TF binding data (LEPAE), which systematically identifies pairwise associations between pairs of genomic windows by leveraging a large compendium of epigenomic and TF binding data from diverse assays and cell types along with genomic distance information. LEPAE assumes that pairs of windows at specific distances apart are more likely to have associations reflected in epigenomic and TF binding data than random pairs of windows. Based on this assumption, LEPAE trains a neural network classifier with a Siamese architecture to distinguish pairs of windows at fixed distances apart from randomly mismatched pairs of the same set of windows. LEPAE then applies the trained classifier to score evidence of association for pairs of windows across a genome. LEPAE's approach is an adaptation of the "Learning Evidence of Conservation from Integrated Functional genomics

Received: 1 April 2025. Revised: 6 November 2025. Accepted: 17 December 2025

© The Author(s) 2026. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

data” (LECIF) method, which we previously used to score evidence of conservation at the functional genomics level between human and mouse (Kwon and Ernst 2021). Additionally, LECIF or extensions of the framework have been applied for comparisons between human and pig (Li *et al.* 2023) and in the context of plant genomics (Roces *et al.* 2024). We apply LEPAE to score pairs of windows in the human genome up to 1 Mb apart. Using gene annotations, topological associated domains, chromatin state annotations, and genome-wide association study (GWAS) variants, we validate that the LEPAE score reflects associated or similar properties in pairs of loci. We expect the LEPAE score to be a complementary resource to existing genomic annotations and datasets for understanding relationships among genomic loci.

2 Materials and Methods

2.1 Overview of LEPAE

LEPAE quantifies evidence of association between pairs of binned genomic windows based on epigenomic and TF binding data. LEPAE does this for each pair of windows at a resolution of r bases within D bases. To do this, LEPAE first trains classifiers for each pairwise distance d , where d ranges from r to D bases at increments of r bases. The classifiers are trained to discriminate pairs that are d bases apart as positive pairs from a negative set of pairs. The negative pairs are randomly mismatched positive pairs within each chromosome, which allows for controlling for chromosome-specific characteristics. LEPAE assumes that D is substantially less than the chromosome length and that positive pairs relative to negative pairs more likely represent associated windows given their proximity.

LEPAE trains each classifier using epigenomic and TF binding data as features for the classification. The specific type of classifier LEPAE uses is a Siamese neural network (Bromley *et al.* 1993) (Fig. 1). A similar classifier was previously shown to be effective in the LECIF method for learning patterns from two input feature vectors (Kwon and Ernst 2021). The Siamese neural network architecture forces a classifier to learn the same weights for both windows in each input pair and thus its predictions are invariant to how the two windows are ordered when provided as input to the classifier.

For each pair, LEPAE uses two binary feature vectors corresponding to epigenomic and TF binding data overlapping the pair’s two genomic windows. In our application, the binary features correspond to whether a window overlaps with peak calls from DNase-seq experiments or ChIP-seq experiments of histone modifications, histone variants, or TFs or individual ChromHMM chromatin state annotations (Ernst and Kellis 2012, 2015). These peak calls and chromatin state annotations cover a diverse set of human cell and tissue types generated by the ENCODE (ENCODE Project Consortium *et al.* 2012) or Roadmap Epigenomics (Roadmap Epigenomics Consortium *et al.* 2015) projects (Fig. 1). Because the same set of windows are used in positive and negative pairs, each classifier learns pairwise characteristics of epigenomic and TF binding data from the two windows in a pair as opposed to characteristics of marginal features. During training, chromosomes for which LEPAE will make predictions are held out. Once trained, each classifier

At window size of r bases, for each pairwise distance d bases (up to D bases):

Positive pairs: windows at pairwise distance of d bases

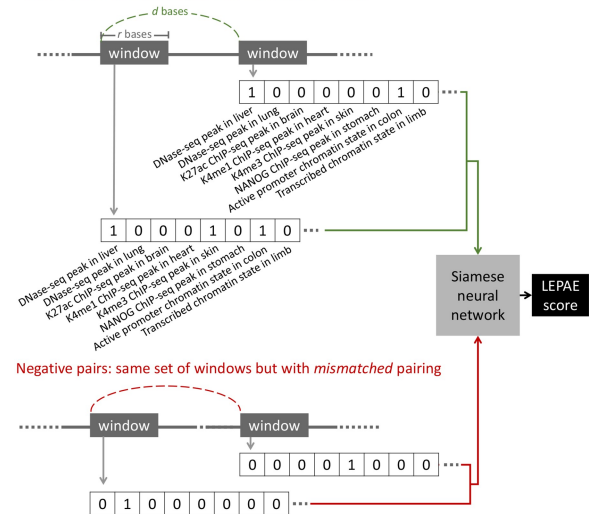


Figure 1 Overview of the LEPAE method. For every genomic window of r bases, LEPAE generates a feature vector from epigenomic and TF binding annotations. Only a subset of the features is shown here. A pair of two such feature vectors are provided as input to a Siamese neural network. At each pairwise distance of d bases, the classifier is trained to distinguish positive pairs, which are genomic windows that are d bases apart, from negative pairs, which are randomly mismatched pairs of the same set of windows. To generate the LEPAE score for a pair of windows that are d bases apart and held out from training, LEPAE ensembles predictions from multiple classifiers trained on random subsets of training data and also for pairwise distances of $d - r$, d , and $d + r$. Feature labels (e.g. DNase-seq peak in liver) are not provided to LEPAE.

makes predictions for pairs of genomic windows at its corresponding distance for chromosomes that were held out from training. LEPAE then ensembles individual classifier predictions to improve performance and consistency across different pairwise distances.

2.2 Defining genomic windows

We binned all autosomal chromosomes and chrX into non-overlapping 1-kb windows and separately into non-overlapping 10-kb windows. We used hg38 as the genome assembly.

2.3 Input features

Each pair of windows was assigned two feature vectors, one corresponding to the upstream window and the other corresponding to the downstream window. For each window, each peak call corresponded to a binary feature. If a genomic window overlapped a peak call in an experiment, the corresponding value in the feature vector was set to 1, otherwise it was set to 0. The chromatin state annotations were one-hot encoded such that each binary feature corresponded to the presence of a chromatin state in a cell or tissue type.

Peak calls for DNase-seq and ChIP-seq experiments of histone modifications, histone variants, and TFs were generated by ENCODE and processed by ENCODE4 pipelines (ENCODE Project

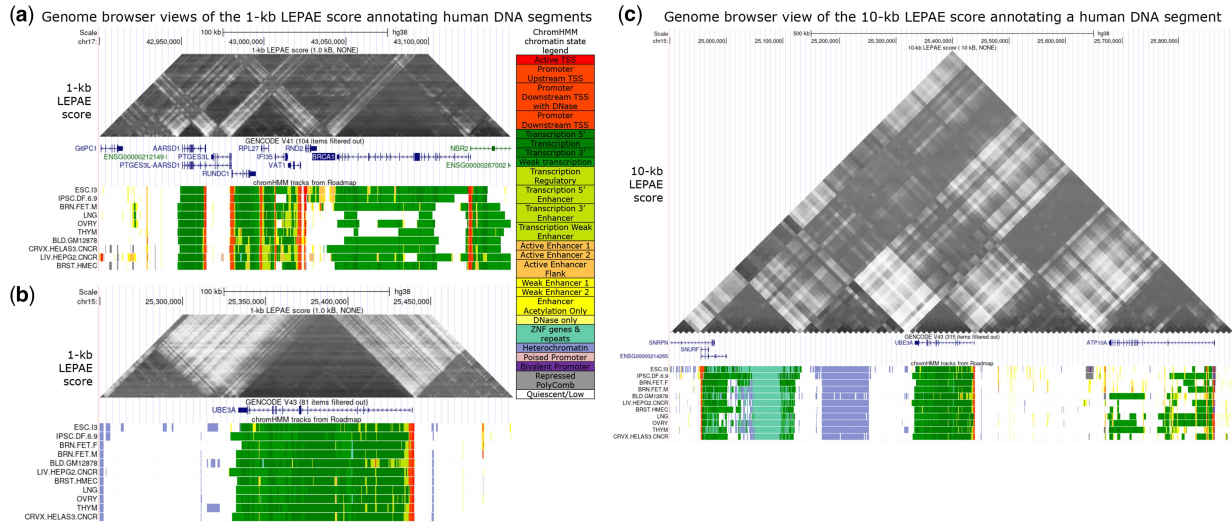


Figure 2 Genome Browser views. (a) UCSC Genome Browser (Haeussler *et al.* 2019) view of the 1-kb LEPAE score annotating gene BRCA1 and its neighboring regions. The score is shown in the top as a heatmap in a format primarily used to visualize Hi-C contact matrices (Durand *et al.* 2016). Darker colors correspond to higher scores. Below the score are GENCODE V41 gene annotations (Frankish *et al.* 2021), followed by ChromHMM chromatin state annotations (Ernst and Kellis, 2012, 2015) for different epigenomes from the Roadmap Epigenomics Project (Roadmap Epigenomics Consortium *et al.* 2015), which were provided as input. While chromatin state annotations from 127 epigenomes were used as input features, only a subset is shown here. State legend is on the right. (b) Same as (a) but for gene UBE3A instead of BRCA1. (c) Same as (b) showing gene UBE3A but for 10-kb LEPAE score instead of 1-kb.

Consortium *et al.* 2012). ChromHMM chromatin state annotations (Ernst and Kellis 2012) were from the 25-state model learned from imputed data for 127 reference epigenomes from the Roadmap Epigenomics Project (Ernst and Kellis 2015; Roadmap Epigenomics Consortium *et al.* 2015) that were previously lifted over from hg19 to hg38 using the liftOver tool (Hinrichs *et al.* 2006). With 3,337 DNase-seq and ChIP-seq datasets and one-hot encoded chromatin state annotations, each feature vector consisted of 6,512 binary values.

2.4 Pairwise distance

For 1-kb windows, we varied the pairwise distance from 1 kb to 100 kb with increments of 1 kb, resulting in 100 distances. For 10-kb windows, we varied the pairwise distance from 10 kb to 1 Mb with increments of 10 kb, resulting in another 100 distances. Distances were measured based on the first base of each window.

2.5 Defining positive and negative pairs

For each pairwise distance d , pairs of windows with d bases between their first bases were defined as positive pairs. To generate negative pairs, we randomly shuffled the pairing of the positive pairs within the same chromosome, resulting in pairs of windows that are not necessarily d bases apart from each other but are from the same chromosome.

2.6 Classifier

Each neural network had a Siamese architecture (Bromley *et al.* 1993) consisting of two identical sub-networks, which share their weights, followed by a final sub-network that combines the

output from the two sub-networks to generate a final prediction. The two input feature vectors were fed into the two sub-networks. All sub-networks had a single hidden layer, resulting in two hidden layers in total.

To set the hyper-parameters of a classifier for a fixed pairwise distance d , window size r , and a particular set of training pairs, LEPAE conducted a random search, where it generated 10 classifiers, each with different randomly selected combination of hyper-parameters. Each classifier was trained on the same set of 50 000 training pairs and evaluated on the same set of 5000 validation pairs. Hyper-parameters were varied as follows during random search: batch size: 16, 32, 64; learning rate: 1e-8, 1e-6, 1e-4; dropout rate: 0, 0.25, 0.5; number of neurons in the initial layer: 32, 64, 128; number of neurons in the final layer: 32, 64, 128.

In training a classifier, along with the original pair, a flipped version of the pair was also provided. In the flipped version of a pair the upstream window is downstream and the downstream window is upstream. This doubled the number of input pairs provided. LEPAE identified the best-performing combination of hyper-parameters by maximizing the AUROC on the validation pairs.

With the best combination of hyper-parameters for a given window size r , pairwise distance d , and set of training pairs, LEPAE then trained 5 new classifiers, each with 50 000 training pairs randomly sampled from the training set but with the same combination of hyper-parameters. To compute the LEPAE score for target pairs, which were not in the set of training pairs, LEPAE computed the average probabilities from an ensemble of classifiers. Specifically, for the target pair with pairwise distance d , LEPAE applied each of the 5 classifiers trained on pairs with pairwise distance of d to obtain the probability of the target pair being classified as a positive pair and also for the flipped pair,

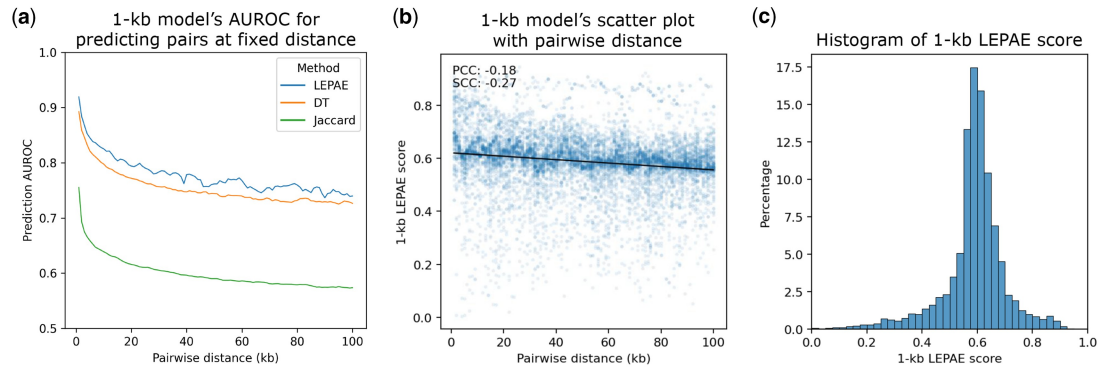


Figure 3 Characteristics of the 1-kb LEPAE score. (a) Relationship between pairwise distance and prediction AUROC for 1-kb model. For each pairwise distance (x -axis), mean prediction AUROC of the 1-kb LEPAE score for distinguishing pairs of windows at the distance from randomly mismatched pairs of the same windows is shown in blue. The mean is computed from two sets of classifiers trained on non-overlapping training sets (Section 2). Mean AUROC values when a decision tree (DT) instead of a neural network was used as the supervised classifier are shown in orange (Supplementary Methods, available as [supplementary data](#) at *Bioinformatics* online). Mean AUROC values computed using Jaccard index instead of the LEPAE score to perform the same classification task are shown in green. Values belonging to the same method are connected by piecewise linear interpolation. (b) Scatter plot showing with a dot for each pair of windows the pairwise distance (x -axis) and the 1-kb LEPAE score (y -axis). Ten thousand random pairs are shown. A linear regression line fitted to the ten thousand random pairs is also shown. PCC and SCC, computed from 1 million randomly selected pairs, are shown in the top left. (c) Distribution of the 1-kb LEPAE score. Forty bins ranging from 0 to 1 with increments of 0.025 were used. A 10-kb score resolution version of this figure is in Fig. 1, available as [supplementary data](#) at *Bioinformatics* online.

resulting in 10 values. In addition, to further reduce variance, for the same target pair of windows with distance of d , LEPAE similarly applied 5 classifiers corresponding to a shorter pairwise distance of $d-r$ to obtain 10 additional values and 5 classifiers corresponding to a longer pairwise distance of $d+r$ to obtain 10 additional values. The assumption here is that classifiers for adjacent pairwise distances will learn similar information. The final LEPAE score of the target pair was the average of these 30 values, except if the pair had the minimum or maximum pairwise distance, the final score was the average of 20 available values since there was no shorter or longer pairwise distance available, respectively. This procedure of hyper-parameter search, training the final classifiers, and ensemble-based predictions was repeated for each pairwise distance d , window size r , and set of training pairs. We used PyTorch (version 0.3.0.post4) (Paszke et al. 2017) for implementation.

2.7 Defining subsets of pairs for training, validation, and test

To generate predictions for all pairs of genomic windows from an odd chromosome or chrX, for a specific pairwise distance d and window size r , we first randomly selected three even chromosomes (chr6, chr16, chr18) and defined a random subset ($n=5000$) of pairs of windows from those three chromosomes as a validation set. We selected pairs of windows from the remaining even chromosomes as a training set. To form a test set, we used a random subset ($n=5000$) of pairs of windows from odd chromosomes.

To generate predictions for all pairs of windows from an even chromosome, we took an analogous approach as above where chr3, chr17, and chr21 were randomly selected for validation. There was no overlap in genomic regions used for training,

validation, and test. chrX was excluded from training, validation, and test, but included for prediction and downstream analyses.

2.8 Filtering pairs

In all analyses, except for reporting predictive performance, all pairs that overlap any assembly gap annotation in either window were excluded since most input features do not map well to these pairs. We downloaded the assembly gap annotation from the UCSC Genome Browser (Haeussler et al. 2019).

3 Results

3.1 Application of LEPAE

We applied LEPAE to over 6000 binary features derived from over 3000 datasets from the ENCODE (ENCODE Project Consortium et al. 2012) or Roadmap Epigenomics (Roadmap Epigenomics Consortium et al. 2015) projects (Section 2). In our application, we applied LEPAE for pairs of 1-kb human genomic windows with distances d ranging from 1 kb to 100 kb in increments of 1 kb (score resolution $r=1$ kb; maximum pairwise distance $D=100$ kb). In addition, we applied LEPAE for pairs of 10-kb genomic windows with pairwise distances d ranging from 10 kb to 1 Mb in increments of 10 kb ($r=10$ kb; $D=1$ Mb). As a result, we annotated more than 302 million and 30 million pairs of 1-kb and 10-kb genomic windows, respectively, with LEPAE scores (Figs 2 and 3c, Fig. 1c, available as [supplementary data](#) at *Bioinformatics* online). We illustrate LEPAE scores at two loci (Fig. 2), which highlights how LEPAE scores can be higher for pairs of loci within the same annotated genes despite making its predictions independent of such annotations.

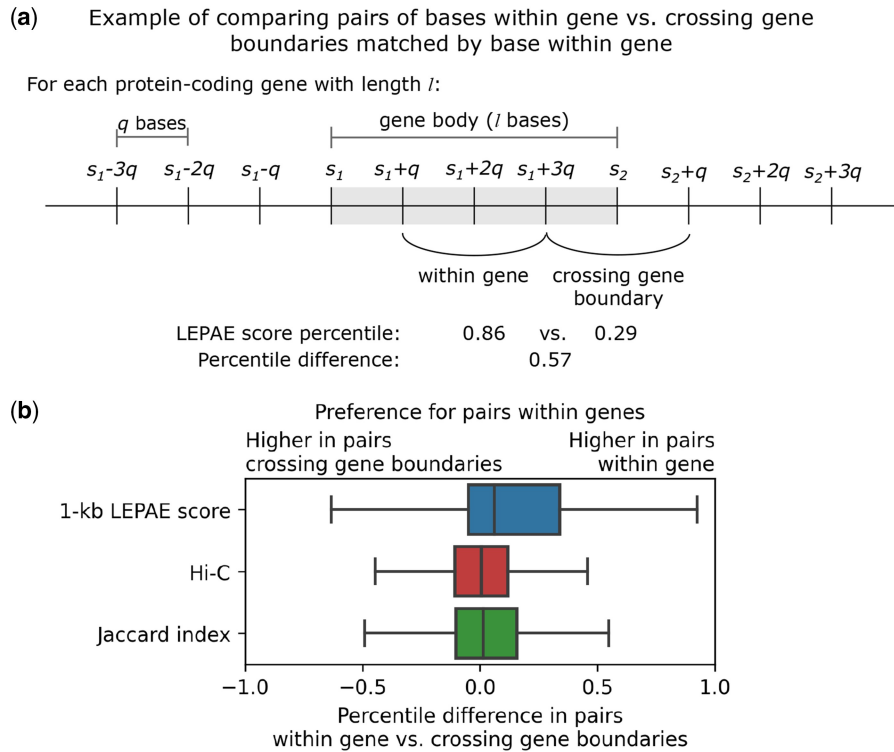


Figure 4 1-kb LEPAE score's relationship to genic annotation. (a) For a protein-coding gene of length l , its quartile length q is used to define bases within the gene body, spanning from the transcription start site (TSS) s_1 to transcription end site (TES) s_2 , and upstream and downstream of the gene body. A pair of bases within a gene (e.g. s_1+q and s_1+3q) is compared to a pair of bases crossing the gene boundary (e.g. s_1+3q and s_2+q), where the two pairs are matched by their pairwise distance (e.g. $2q$) and share one base located within the gene (e.g. s_1+3q). The percentile difference is computed by taking the LEPAE score percentile of the former and subtracting it by the score percentile of the latter. This approach was repeated for normalized Hi-C contact frequency and Jaccard index for comparison (Supplementary Methods, available as [supplementary data](#) at Bioinformatics online). (b) Shown for the 1-kb LEPAE score, normalized Hi-C contact frequency, or Jaccard index is the distribution of percentile differences between pairs with both bases within a gene and pairs with only one base within a gene. A positive percentile difference indicates a preference for pairs within genes over pairs crossing gene boundaries.

3.2 LEPAE's predictive power and relationship to distance

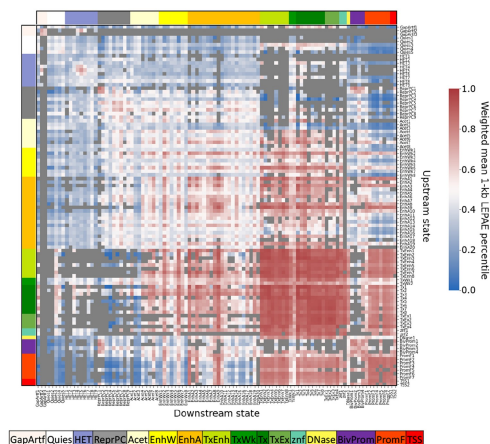
LEPAE has strong predictive power when differentiating positive pairs from negative pairs, all held out from training, especially for pairs with small pairwise distances. For the 1-kb model, we observe the strongest predictive power for pairs with the smallest pairwise distance, 1 kb, with a mean area under the receiver operating characteristic curve (AUROC) of 0.92 (Fig. 3a). For pairs with larger distances, the LEPAE score's predictive power declines, with a minimum mean AUROC of 0.74 at a pairwise distance of 90 kb. Similarly, we observe the greatest predictive power for the 10-kb model with an AUROC of 0.92 at its smallest distance, 10 kb, and with the predictive power declining to the lowest mean AUROC of 0.70 at a pairwise distance of 980 kb (Fig. 1a, available as [supplementary data](#) at Bioinformatics online). Moreover, LEPAE achieves better predictive power than a score learned by decision trees instead of Siamese neural networks at all distances (Fig. 3a, Fig. 1a and Supplementary Methods, available as [supplementary data](#) at Bioinformatics online).

We also confirmed that the LEPAE score is largely distinct from 1D genomic distance (Fig. 3b, Fig. 1b, available as [supplementary data](#) at Bioinformatics online). We observed only

modest negative Pearson correlation coefficient (PCC) values of -0.18 and -0.26 for their correlations with genomic distance, when using the 1-kb and 10-kb LEPAE scores, respectively. Even at the maximum pairwise distances considered (D), 3% and 1% of pairs have a score above the 95th percentile among all scores at 1-kb and 10-kb resolution, respectively. These results indicate that LEPAE can highlight pairs of windows that are distal but exhibit sufficient evidence of association based on their epigenomic and TF binding data.

To confirm that LEPAE learns information beyond similarity in input features, we compared the LEPAE score to Jaccard index, which was computed for every pair of windows to which LEPAE was applied (Supplementary Methods, available as [supplementary data](#) at Bioinformatics online). Specifically, for each pairwise distance we compute the correlation between the LEPAE score and Jaccard index for pairs of windows with that pairwise distance. Consistent with LEPAE learning information beyond similarity in input features, we observed only a moderate correlation with the Jaccard index, with a mean PCC of 0.16 and 0.39 at 1-kb and 10-kb resolutions, respectively (Fig. 2, available as [supplementary data](#) at Bioinformatics online). For the task of differentiating positive pairs from negative pairs, LEPAE achieves consistently better predictive performance than the Jaccard

(a) Percentile of 1-kb LEPAE score for pairs of universal ChromHMM chromatin states with pairwise distance of 5 kb



(b) Pairs of chromatin states at pairwise distance of 5 kb

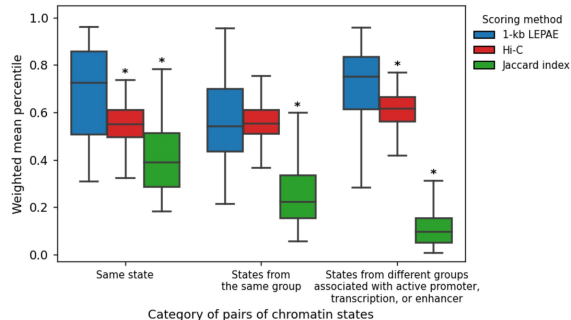


Figure 5 1-kb LEPAE score's relationship to pairs of chromatin states with pairwise distance of 5 kb. (a) Each cell in the heatmap corresponds to a state pair, one annotating the upstream window of a pair of 1-kb windows (row) and the other annotating the downstream 1-kb window (column) with their pairwise distance fixed to 5 kb. The states are from a universal chromatin state annotation (Vu and Ernst 2022) and their ordering in the rows and columns is the same. Color shown next to the topmost row or leftmost column corresponds to the state group of each state along the column or row, respectively, according to the legend on the bottom. Colors shown in the cells correspond to a weighted mean LEPAE score percentile of pairs of windows that are 5 kb apart and are annotated by the states specified in the row and column (Supplementary Methods, available as supplementary data at Bioinformatics online). Color legend for the score is shown on the right. Pairs of states that together cover less than 0.0005% of all the bases annotated by pairs of states were excluded from the analysis with grey shown in their corresponding cells. Similar heatmaps but for Hi-C contact frequency and Jaccard index, for distance of 50 kb, and also for 10-kb score resolution at distances of 50 kb and 500 kb are in Figs 3, 4, 8, and 9, available as supplementary data at Bioinformatics online. (b) Shown for three categories of pairs of ChromHMM chromatin states are the distribution of the weighted mean percentile of the 1-kb LEPAE score, Hi-C contact frequency, or Jaccard index, at a pairwise distance of 5 kb as done in (a). The first category shown on the leftmost position corresponds to pairs of the same state. The second corresponds to pairs of different states belonging to the same state group. The rightmost category corresponds to pairs of states where both states are associated with active promoters, transcription, or enhancers but are from different state groups. These three categories of pairs of states do not overlap with each other. Within each category, the distributions of

index at both resolutions (Fig. 3a, Fig. 1a, available as supplementary data at Bioinformatics online). These results confirm that LEPAE captures information beyond similarity in features.

3.3 High-scoring pairs highlight associated loci based on annotation of genes and chromatin states

We next analyzed the LEPAE score in the context of gene annotations, which were not used in learning the score. Specifically, we asked whether the LEPAE score was higher for pairs of bases that fell within the gene body of the same gene compared to a matched control set. The matched control set consisted of pairs of bases for the same set of genes and the same distance apart but with one base inside and the other outside the gene body. For these control pairs, the region between the two bases crosses the gene boundary (Supplementary Methods, available as supplementary data at Bioinformatics online). We found that pairs of bases within the gene body have higher 1-kb resolution LEPAE score than pairs that crossed gene boundaries (Fig. 4; Wilcoxon signed-rank test $P < .0001$; Supplementary Methods, available as supplementary data at Bioinformatics online). In addition, we found that this difference is stronger than the difference observed with Hi-C data and Jaccard index (Fig. 4; Mann-Whitney U test $P < .0001$; Supplementary Methods, available as supplementary data at Bioinformatics online).

To better understand the nature of pairs of windows that received a high LEPAE score, we also analyzed the score for pairs of genomic elements using a universal chromatin state annotation (Vu and Ernst 2022) (Supplementary Methods, available as supplementary data at Bioinformatics online). This annotation provides a single chromatin state assignment per genomic position based on over 1000 human epigenomic datasets from more than 100 cell or tissue types. In total there were 100 states in this annotation that were manually grouped into 16 broader state groups, which were previously defined in Vu and Ernst (2022) based on model parameters and state enrichments for external genome annotations. This chromatin state annotation was not directly provided to LEPAE as input, though we note that there is overlap in the input datasets provided to LEPAE and the datasets used to learn the chromatin state model. We first computed the percentile of 1-kb LEPAE score for all pairs of windows with pairwise distances of 5 kb and 50 kb, separately, and then the mean percentile for each pair of chromatin states for both pairwise distances (Supplementary Methods, available as supplementary data at Bioinformatics online; Fig. 5, Figs 4

Figure 5. Continued

the weighted mean percentile of the 1-kb LEPAE score, Hi-C contact frequency, and Jaccard index are shown from left to right, colored according to the legend on the right. An asterisk above a distribution denotes significant difference between it and the distribution of weighted 1-kb LEPAE score percentiles for the same category based on a Mann-Whitney U test ($P < .0001$). Similar boxplots but for distance of 50 kb and also for 10-kb score resolution at pairwise distances of 50 kb and 500 kb are in Fig. 5, available as supplementary data at Bioinformatics online.

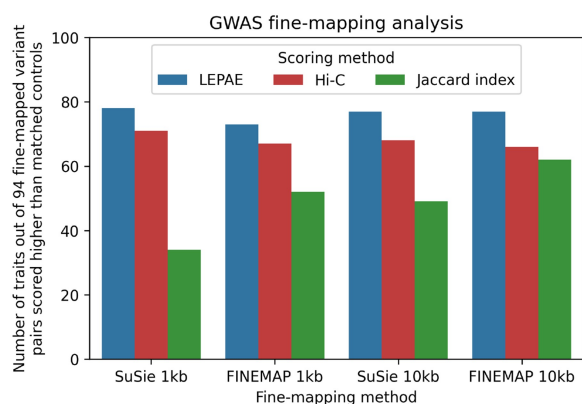


Figure 6 LEPAE score's relationship to fine-mapped GWAS variants. The figure shows out of 94 UK biobank traits (Kanai et al. 2021, 2022), for how many the LEPAE score was higher for pairs of fine-mapped variants than matched control variants. The results are shown for 1-kb resolution, which considers pairs up to 100 kb apart, and 10-kb resolution, which considers pairs up to 1 Mb apart, for both the SuSiE and FINEMAP methods. For comparison, results are also shown for normalized Hi-C contact frequency and the Jaccard index.

and 5a, available as [supplementary data](#) at *Bioinformatics* online). For the 5 kb distance, pairs of the same state had a high mean 1-kb LEPAE score percentile with an average of 0.69, compared to 0.51 for any two states with that pairwise distance (Fig. 5). Notably, pairs of active transcription-associated states (Tx, TxEnh, TxEx) with a pairwise distance of 5 kb had high mean LEPAE score percentiles ranging from 0.84 to 0.96 (Fig. 5a). Similarly, pairs of active promoter-associated states (PromF, TSS) had a mean 1-kb LEPAE score percentile ranging from 0.63 to 0.87 (Fig. 5a). We saw consistent trends using a pairwise distance of 50 kb with pairs of the same state scoring 0.66 on average compared to an overall mean of 0.54 for all state pairs that are 50 kb apart.

We also observed cases of pairs of different types of states having a high mean 1-kb LEPAE score. For instance, at a pairwise distance of 5 kb, pairs of states which are associated with active promoters, transcription, or enhancers, but did not belong to the same state group scored high with a mean LEPAE score percentile of 0.70 on average, close to the average of 0.69 for pairs of the same state (Fig. 5b). Notably, a state associated with enhancer activity in blood and thymus (EnhA9) scored high when paired up with transcription-associated states, such as a state associated with transcription and enhancer activity in blood (TxEnh2) or an active transcription state (Tx1), with the mean 1-kb LEPAE score percentile of 0.92 for both at a pairwise distance of 5 kb (Fig. 5a). This was similar to the mean 1-kb LEPAE score percentile of 0.90 when enhancer state EnhA9 was paired with itself at that pairwise distance. This state also scored high when paired with active promoter chromatin states (PromF1-7, TSS1-2) with the mean LEPAE score percentile ranging from 0.79 to 0.87. At a distance of 50 kb, pairs of states that were previously characterized as being active promoter, transcription, or enhancer associated and belong to different state groups also scored high with a mean 1-kb LEPAE score percentile of 0.72 on average, close to the average of 0.66 for pairs of the same state (Fig. 5a, available as [supplementary data](#) at *Bioinformatics*

online). These results suggest that the LEPAE score captures biologically meaningful relationships between loci that can be associated with different annotations.

We further observed pairs of states that had markedly low mean LEPAE scores. For example, we saw that the quiescent states (Quies1-5) tended to score low when paired up with states associated with transcription or enhancer activity. As a specific example, at a pairwise distance of 5 kb, pairs of a quiescent state (Quies5) and a state characterized by transcription and enhancers (TxEnh2) had a mean 1-kb resolution LEPAE score percentile of 0.04 (Fig. 5a). This suggests that the LEPAE score can also highlight pairs of chromatin states that are less likely to be associated.

We repeated this chromatin state analysis with epigenome-specific state annotations from the Roadmap Epigenomics Project (Roadmap Epigenomics Consortium et al. 2015). We observed that state enrichment patterns were consistent across different epigenomes (Figs 6 and 7, available as [supplementary data](#) at *Bioinformatics* online), suggesting that associations captured by LEPAE are not limited to any single biological context.

We also investigated whether the relationship between the universal chromatin states and the 1-kb LEPAE score are recapitulated with Hi-C contact frequency or Jaccard index (Fig. 5b, Figs 3, 4, and 5a, available as [supplementary data](#) at *Bioinformatics* online; [Supplementary Methods](#), available as [supplementary data](#) at *Bioinformatics* online). We observed that the 1-kb LEPAE score better highlighted pairs of the same state, states from the same state group, and states associated with active promoters, transcription, or enhancers that belong to different state groups (Mann-Whitney U test $P < .001$). For those same set of state pairs, Hi-C contact frequency tended to be close to the median, not differentiating any particular group of state pairs from the rest. On the other hand, Jaccard index at 1-kb resolution scored pairs of the same state high and scored pairs of states that are associated with active promoters, transcription, or enhancers and belong to different groups low, which is expected from the similarity-based nature of the index. We did not observe the patterns captured by the 1-kb resolution scores with 10-kb resolution scores (Figs 5bc, 8, and 9, available as [supplementary data](#) at *Bioinformatics* online). However, we were not necessarily expecting to since the chromatin states are defined at 200-bp resolution, while the much larger 10-kb windows tend to include more states, making it harder to attribute a high score observed in a pair of windows to a specific pair of chromatin states. Overall, these results suggest the LEPAE score can highlight not only similar loci but also loci exhibiting distinct but associated properties.

3.4 Relationship to chromatin contact frequency

We next studied the relationship between LEPAE scores and chromatin interaction data. Specifically, we compared the 1-kb LEPAE score to normalized 1-kb resolution chromatin interaction frequencies collected in a Hi-C experiment in GM12878 ([Supplementary Methods](#), available as [supplementary data](#) at *Bioinformatics* online). We computed the PCC between the LEPAE score and a normalized Hi-C matrix for pairs with the same pairwise distance and observed a low mean PCC (0.06).

We also observed similar results when comparing 10-kb LEPAE score to 10-kb resolution Hi-C data (mean PCC of 0.06). These results suggest that the LEPAE score captures associations largely distinct from Hi-C chromatin interaction frequencies.

We also related the LEPAE score to an annotation of topologically associating domains (TAD) from the same Hi-C experiment ([Supplementary Methods](#), available as [supplementary data](#) at *Bioinformatics* online). Among pairs of windows at a particular pairwise distance, we compared pairs within the same TAD to pairs at the same distance that cross a TAD boundary, with one window inside a TAD and the other outside the TAD. For the 1-kb LEPAE score, at 93 out of 100 pairwise distances, pairs within TADs had a significantly higher score distribution than those crossing TAD boundaries ([Fig. 10](#), available as [supplementary data](#) at *Bioinformatics* online; Mann-Whitney U test $P < .0001$; [Supplementary Methods](#), available as [supplementary data](#) at *Bioinformatics* online). We observed similar results with the 10-kb LEPAE score, with statistically higher score distribution for pairs within TADs at all pairwise distances ([Fig. 10](#), available as [supplementary data](#) at *Bioinformatics* online). While we observed a low overall correlation between LEPAE scores and Hi-C contact frequencies, LEPAE scores do show some level of agreement with Hi-C data as reflected in the TAD analysis. We hypothesize that this stronger agreement with TAD boundaries may be due to TADs representing a more biologically meaningful feature than individual Hi-C interaction signal values and TAD's relative robustness based on being derived from signal values across many genomic positions.

3.5 Relationship to fine-mapped GWAS variants

We next analyzed whether pairs of variants that were fine-mapped to be a causal variant for the same trait at the same loci had a higher LEPAE score than matched control pairs. We did this analysis for a set of 94 UK Biobank traits that were previously fine-mapped using the SuSiE ([Wang et al. 2020](#)) and FINEMAP ([Benner et al. 2016](#)) methods ([Kanai et al. 2021, 2022](#)). For each actual pair of fine-mapped variants, we extracted two matched control pairs of windows selected such that each control pair shared one window with the original pair and the two windows of the control pairs were the same distance apart as the original pair ([Supplementary Methods](#), available as [supplementary data](#) at *Bioinformatics* online). As a result, all control pairs had a fine-mapped variant in one window. We conducted the analysis at 1-kb resolution for all pairs within 100 kb and at 10-kb resolution for all pairs within 1 Mb. Of the 94 traits, 73–78 traits had a higher LEPAE score for the original pairs compared to the control pairs with the specific number of traits varying with the fine-mapping method and score resolution ([Fig. 6](#)). In comparison, only 47 would be expected by chance (binomial test $p < 10^{-7}$). Also in comparison, 34–62 traits had a higher Jaccard index for the original pairs than control pairs, and 66–71 traits for Hi-C contact frequency ([Fig. 6](#)). These results suggest that pairs of loci with higher LEPAE scores may be more likely to share phenotypic associations.

4 Discussion

Here we presented LEPAE, a method that scores evidence for pairwise association between genomic windows based on a large collection of functional genomic datasets. LEPAE is a variant of LECIF ([Kwon and Ernst 2021](#)), a method to score evidence of conservation between two species based on functional genomics data. In contrast, LEPAE scores evidence of association based on functional genomics data for pairs of loci within the same species.

We applied LEPAE with more than 3000 functional genomic datasets including maps of chromatin accessibility, histone modifications and variants, TF binding, and chromatin state annotations from various tissue and cell types. The resulting LEPAE scores had greater predictive power for differentiating pairs of windows at a fixed distance from randomly mismatched pairs of the same windows than a baseline approach of computing the Jaccard index of input features. Using external annotations not provided as input features, including gene annotations, fine-mapped variants from GWAS, and TADs, we presented analyses suggesting LEPAE scores reflect biologically meaningful associations.

While the results we presented suggest LEPAE will be informative for studying pairwise relationships between genomic loci, it does have limitations. As the score characterizes relationships at fixed distances, it is not applicable to studying interchromosomal associations. Also, the score does not directly provide information about relationships involving more than two pairs of loci.

While here we applied LEPAE at 1-kb and 10-kb resolutions with maximum distances of 100 kb and 1 Mb, respectively, we could in the future aim for a finer resolution and larger pairwise distances at a higher computational cost. When training LEPAE, here we assumed the maximum pairwise distance D was substantially less than the chromosome length and restricted negative pairs to consist of windows from the same chromosome to control for chromosome-specific characteristics during classification. An avenue for future investigation is to extend LEPAE for predictions at substantially larger distances by removing the restriction that the windows in negative pairs come from the same chromosome. In addition, we have defined positive and negative examples for training based on genomic distance though our method is not dependent on all positive pairs actually representing true associations and similarly negative pairs not. However, future work could consider incorporating additional information such as chromatin contact frequency to better reflect the likelihood of association.

Moreover, although here we applied LEPAE to epigenomic and TF binding data from the human genome, future work could investigate applying LEPAE with additional types of functional genomics data such as RNA-seq data and also to other species with less extensive data than human, exploring its applicability to varying sizes and types of input data. Another direction of future work is to integrate the intra-species LEPAE scores with the inter-species LECIF score to identify pairs of human loci and their homologous mouse loci with high LEPAE and LECIF scores, highlighting associations conserved across multiple dimensions. We expect our existing LEPAE scores as well as the approach

applied to other data will be a resource for studying relationships between genomic loci.

Acknowledgements

We thank the ENCODE and Roadmap Epigenomics consortia for generating the data and making it publicly available. We thank members of the Ernst lab for useful discussions.

Author contributions

Soo Bin Kwon (Conceptualization [lead], Data curation [lead], Formal analysis [lead], Investigation [lead], Methodology [lead], Software [lead], Validation [lead], Visualization [lead], Writing—original draft [lead], Writing—review & editing [lead]) and Jason Ernst (Conceptualization [lead], Data curation [supporting], Formal analysis [supporting], Funding acquisition [lead], Investigation [supporting], Methodology [lead], Supervision [lead], Writing—original draft [supporting], Writing—review & editing [lead])

Supplementary material

[Supplementary material](#) is available at *Bioinformatics* online.

Conflict of interests

No competing interest is declared.

Funding

We acknowledge funding from US National Institutes of Health (DP1DA044371, U01MH130995, and U01HG012079 to J.E.), a Rose Hills Innovator Award (J.E.), and a UCLA Dissertation Year Fellowship (S.K.).

Data availability

The LEPAE scores and the software are available at the LEPAE repository, <https://github.com/ernstlab/LEPAE>, and Zenodo, <https://doi.org/10.5281/zenodo.17403040>. Links to data files from the ENCODE (ENCODE Project Consortium *et al.* 2012) or Roadmap Epigenomics Projects (Roadmap Epigenomics Consortium *et al.* 2015) that were used to generate input features to LEPAE are listed in the same repository as well as in [Table 1](#), available as [supplementary data](#) at *Bioinformatics* online. For analysis using gene boundaries, we used GENCODE annotations (V38) (Frankish *et al.* 2021). The in situ Hi-C data for GM12878 (Rao *et al.* 2014) used for the Hi-C analysis is experiment 4DNES3JX38V5 (Rao *et al.* 2014) on the 4DN Nucleome Data Portal (Reiff *et al.* 2022) and available from <https://4dn-open-data-public.s3.amazonaws.com/fourfront-webprod/wfoutput/a98ca64a-861a-4a8c-92e9-586af457b1fb/4DNFI1UEG1HD.hic>. TAD coordinates for the experiments are from the 3D Genome Browser (Wang *et al.* 2020) and available from <http://3dgenome.fsm.northwestern.edu/downloads/hg38.TADs.zip>.

The variants used in the fine-mapping analysis are available from <https://www.finucanelab.org/data> (Kanai *et al.* 2021, 2022). Universal chromatin state annotations were downloaded from https://github.com/ernstlab/full_stack_ChromHMM_annotations (Vu and Ernst 2022).

References

- Benner C, Spencer CCA, Havulinna AS *et al.* FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* 2016;**32**:1493–501. <https://doi.org/10.1093/bioinformatics/btw018>
- Bromley J, Guyon I, LeCun Y *et al.* Signature verification using a “Siamese” time delay neural network. In: *Proceedings of the 7th International Conference on Neural Information Processing Systems*. San Francisco, CA, USA: Morgan Kaufmann, Publishers Inc. 1993, 737–44.
- Corradin O, Saiakhova A, Akhtar-Zaidi B *et al.* Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. *Genome Res* 2014;**24**:1–13. <https://doi.org/10.1101/gr.164079.113>
- Durand NC, Robinson JT, Shamim MS *et al.* Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst* 2016;**3**:99–101. <https://doi.org/10.1016/j.cels.2015.07.012>
- ENCODE Project Consortium *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;**489**:57–74. <https://doi.org/10.1038/nature11247>
- Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* 2012;**9**:215–6. <https://doi.org/10.1038/nmeth.1906>
- Ernst J, Kellis M. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nat Biotechnol* 2015;**33**:364–76. <https://doi.org/10.1038/nbt.3157>
- Ernst J, Kheradpour P, Mikkelsen TS *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 2011;**473**:43–9. <https://doi.org/10.1038/nature09906>
- Frankish A, Diekhans M, Jungreis I *et al.* GENCODE 2021. *Nucl Acids Res* 2021;**49**:D916–23. <https://doi.org/10.1093/nar/gkaa1087>
- Gazal S, Weissbrod O, Hormozdiari F *et al.* Combining SNP-to-gene linking strategies to identify disease genes and assess disease omnigenicity. *Nat Genet* 2022;**54**:827–36. <https://doi.org/10.1038/s41588-022-01087-y>
- Haeussler M, Zweig AS, Tyner C *et al.* The UCSC genome browser database: 2019 update. *Nucleic Acids Res* 2019;**47**:D853–8. <https://doi.org/10.1093/nar/gky1095>
- He B, Chen C, Teng L, Tan K. Global view of enhancer-promoter interactome in human cells. *Proc Natl Acad Sci USA* 2014;**111**:E2191–99. <https://doi.org/10.1073/pnas.1320308111>
- Hinrichs AS, Karolchik D, Baertsch R *et al.* The UCSC genome browser database: update 2006. *Nucleic Acids Res* 2006;**34**:D590–8. <https://doi.org/10.1093/nar/gkj144>
- Kanai M, Ulirsch JC, Karjalainen J *et al.* Insights from complex trait fine-mapping across diverse populations. *medRxiv*. 2021. <https://doi.org/10.1101/2021.09.03.21262975>

- Kanai M, Elzur R, Zhou W *et al.* Meta-analysis fine-mapping is often miscalibrated at single-variant resolution. *Cell Genom* 2022;**2**:100210. <https://doi.org/10.1016/j.xgen.2022.100210>
- Kwon S, Ernst J. Learning a genome-wide score of human-mouse conservation at the functional genomics level. *Nat Commun* 2021;**12**:2495. <https://doi.org/10.1038/s41467-021-22653-8>
- Li J, Zhao T, Guan D *et al.* Learning functional conservation between human and pig to decipher evolutionary mechanisms underlying gene expression and complex traits. *Cell Genom* 2023;**3**:100390. <https://doi.org/10.1016/j.xgen.2023.100390>
- Lieberman-Aiden E, van Berkum NL, Williams L *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 2009;**326**:289–93. <https://doi.org/10.1126/science.1181369>
- Liu Y, Sarkar A, Kheradpour P *et al.* Evidence of reduced recombination rate in human regulatory domains. *Genome Biol* 2017;**18**:193. <https://doi.org/10.1186/s13059-017-1308-x>
- Mifsud B, Tavares-Cadete F, Young AN *et al.* Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat Genet* 2015;**47**:598–606. <https://doi.org/10.1038/ng.3286>
- Paszke A, Gross S, Chintala S *et al.* Automatic differentiation in PyTorch. NIPS 2017 Workshop Autodiff, Long Beach, CA, USA, 2017..
- Rao SSP, Huntley MH, Durand NC *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 2014;**159**:1665–80. <https://doi.org/10.1016/j.cell.2014.11.021>
- Reiff SB, Schroeder AJ, Kirli K *et al.* The 4D nucleome data portal as a resource for searching and visualizing curated nucleomics data. *Nat Commun* 2022;**13**:2365. <https://doi.org/10.1038/s41467-022-29697-4>
- Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* 2015;**518**:317–30. <https://doi.org/10.1038/nature14248>
- Roces V, Guerrero S, Alvarez A *et al.* Plantfunco: integrative functional genomics database reveals clues into duplicates divergence evolution. *Mol Biol Evol* 2024;**41**:msae042. <https://doi.org/10.1093/molbev/msae042>
- Thurman RE, Rynes E, Humbert R *et al.* The accessible chromatin landscape of the human genome. *Nature* 2012;**489**:75–82. <https://doi.org/10.1038/nature11232>
- Vu H, Ernst J. Universal annotation of the human genome through integration of over a thousand epigenomic datasets. *Genome Biol* 2022;**23**:9. <https://doi.org/10.1186/s13059-021-02572-z>
- Wang G, Sarkar A, Carbonetto P *et al.* A simple new approach to variable selection in regression, with application to genetic fine mapping. *J R Stat Soc Series B Stat Methodol* 2020;**82**:1273–300. <https://doi.org/10.1111/rssb.12388>