Bioinformatics, 38(19), 2022, 4457–4465
https://doi.org/10.1093/bioinformatics/btac535
Advance Access Publication Date: 5 August 2022
Original Paper

OXFORD

Genome analysis

Prediction of gene co-expression from chromatin contacts with graph attention network

Ke Zhang[1,2], Chenxi Wang[3,4], Liping Sun[3] and Jie Zheng [1,5,*]

[1]School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China, [2]Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai 200050, China, [3]iHuman Institute, ShanghaiTech University, Shanghai 201210, China, [4]School of Life Science and Technology, ShanghaiTech University, Shanghai 201210, China and [5]Shanghai Engineering Research Center of Intelligent Vision and Imaging, ShanghaiTech University, Shanghai 201210, China

*To whom correspondence should be addressed.
Associate Editor: Pier Luigi Martelli

Received on December 7, 2021; revised on July 12, 2022; editorial decision on July 15, 2022

Abstract

Motivation: The technology of high-throughput chromatin conformation capture (Hi-C) allows genome-wide measurement of chromatin interactions. Several studies have shown statistically significant relationships between gene–gene spatial contacts and their co-expression. It is desirable to uncover epigenetic mechanisms of transcriptional regulation behind such relationships using computational modeling. Existing methods for predicting gene co-expression from Hi-C data use manual feature engineering or unsupervised learning, which either limits the prediction accuracy or lacks interpretability.

Results: To address these issues, we propose HiCoEx (Hi-C predicts gene co-expression), a novel end-to-end framework for explainable prediction of gene co-expression from Hi-C data based on graph neural network. We apply graph attention mechanism to a gene contact network inferred from Hi-C data to distinguish the importance among different neighboring genes of each gene, and learn the gene representation to predict co-expression in a supervised and task-specific manner. Then, from the trained model, we extract the learned gene embeddings as a model interpretation to distill biological insights. Experimental results show that HiCoEx can learn gene representation from 3D genomics signals automatically to improve prediction accuracy, and make the black box model explainable by capturing some biologically meaningful patterns, e.g., in a gene contact network, the common neighbors of two central genes might contribute to the co-expression of the two central genes through sharing enhancers.

Availability and implementation: The source code is freely available at https://github.com/JieZheng-ShanghaiTech/HiCoEx.

Contact: zhengjie@shanghaitech.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The chromatin organization in 3D space plays essential role in transcriptional regulation (Ahn *et al.*, 2021; Bhat *et al.*, 2021). Advances in the chromosome conformation capture (3C) technique and 3C-based derivative techniques, such as 4C, 5C and Hi-C, enable the 3D characterization of chromatin interactions (Yu and Ren, 2017). In particular, with a genome-wide measurement, Hi-C allows to comprehensively explore the relationships between 3D genomic organization and gene regulation and reveal the underlying epigenetic mechanisms behind such relationships. For instance, Dong *et al.* (2010) observed that the Hi-C interactions between many gene pairs are consistent with their co-expression levels, suggesting that co-expression between two genes is strongly associated with their chromatin interactions. Ibn-Salem *et al.* (2017) analyzed the relationships among genes' spatial distribution, gene expression and evolution of paralogous gene pairs. They found that a large proportion of paralogous gene pairs tend to be both co-expressed and co-localized within the same topological associated domains, and these pairs of genes usually share common enhancers. These works considered the information of local spatial contacts between a pair of genes when explaining their co-expression but neglected the indirect contacts connecting two genes through intermediate genes (Dekker and Misteli, 2015; Sandhu *et al.*, 2012).

To gain a better understanding of whole-genome chromatin contacts, many studies have performed network analysis on these

contacts by constructing the so-called *gene contact networks* (Thibodeau *et al.*, 2017; Zhang and Ma, 2020), where a node represents a genomic locus and an edge connecting two nodes corresponds to their spatial contacts. In such a network, two genes without direct spatial interactions may still have connections in terms of 3D genomics, e.g. they could have contacts with a common set of genes. Therefore, the network-based methods are capable of representing high-level relationships among multiple genes.

Recently, there has been increasing interest in jointly studying the genome conformation and transcriptional regulation by computational methods (Cao *et al.*, 2020; Tian *et al.*, 2020). Several studies have attempted to establish predictive models, to quantify the interplay between 3D genomic structure and gene co-expression (Babaei *et al.*, 2015; Varrone *et al.*, 2020). Here, the key is to extract representative features of the gene contact network for the co-expression prediction. A previous method calculates five topological properties of a gene contact network as the node representations to predict co-expression (Babaei *et al.*, 2015). To capture the correlation between chromatin topology and gene co-expression, two embedding-based methods in Varrone *et al.* (2020) encode gene features by latent vectors. One method uses matrix factorization to decompose a gene contact network into factors as the gene embeddings. The other method uses the random walk-based node2vec algorithm (Grover and Leskovec, 2016) to learn the gene embedding. These unsupervised methods in gene representation learning cannot extract gene features specific to the task of co-expression prediction. Moreover, these approaches do not adequately explore the biological explanations behind gene embeddings.

Graph neural network (GNN) has emerged as a powerful tool for graph representation learning and link prediction tasks in *Bioinformatics*, such as the predictions of protein–protein interactions (Fout *et al.*, 2017) and chromatin interactions (Lanchantin and Qi, 2020; Zhang and Ma, 2020). In particular, graph convolution network (GCN) generalizes the traditional convolution to graphs by propagating information of neighboring nodes for each central node (Kipf and Welling, 2017; Niepert *et al.*, 2016). The propagation process can be regarded as finding connectivity patterns and integrating these patterns into the latent embeddings. Furthermore, attention mechanisms have been integrated into many models as they assign weights to different parts of input data, which can improve prediction performance.

Based on these observations, we present HiCoEx (Hi-C predicts gene co-expression), a GNN-based method for the prediction of gene co-expression using spatial chromatin contacts. With the gene contact network as the input, HiCoEx employs the graph attention mechanism to capture the characteristics of 3D chromatin structures in an end-to-end manner. We collected Hi-C data and RNA-seq data from different tissues and cell lines of human. Experimental results on these real datasets show that our method can effectively learn the latent graph representations and significantly outperform existing methods. We also explored biological mechanisms behind the predictive model, explaining the structural information encoded in gene embeddings. Our analyses suggest that common neighbors might contribute to the co-expression of the two central genes through enhancer sharing. Our framework of supervised learning followed by explanation technique can facilitate the discovery of transcriptional regulation by 3D chromatin structures.

2 Materials and methods

2.1 Overview of HiCoEx

By pre-processing Hi-C and RNA-seq data, we obtain the input of HiCoEx, i.e. the adjacency matrix of a gene contact network **P** and the adjacency matrix of a gene co-expression network **Q** for genes in the same chromosome. The size of both matrices is $N \times N$, where N is the number of genes in the chromosome. In matrix **P**, an element $p_{ij} = 1$ if genes v_i and v_j have significant spatial contact, and $p_{ij} = 0$ otherwise. Similarly, in matrix **Q**, an element $q_{ij} = 1$ if v_i and v_j are significantly co-expressed, and $q_{ij} = 0$ otherwise. Our goal is to learn the embedding of each gene by graph representation learning and then predict whether each pair of genes have co-expression relationship.

Existing methods for predicting co-expression from Hi-C data either design a graph kernel or learn a random walk from the gene contact network to encode the spatial features for each gene, which are unsupervised learning and as such their gene representations do not consider gene co-expression (Babaei *et al.*, 2015; Varrone *et al.*, 2020). Here, we introduce a GNN-based model to learn the gene representations in a supervised manner, i.e. incorporating signals from both gene contact network and gene co-expression network into the node embeddings. Figure 1A illustrates our overall framework. The proposed architecture of HiCoEx (Fig. 1B) consists of a graph attention layer (Velickovic *et al.*, 2018) and a feed-forward layer.

2.2 Data pre-processing

We collected published Hi-C data from 12 types of tissues and cell lines, and downloaded the corresponding RNA-seq datasets of the same tissues and cell lines from GTEx for normal samples and from The Cancer Genome Atlas (TCGA) for tumor samples, following the procedure used in Varrone *et al.* (2020). We also used a published Hi-C dataset of human pancreatic islets from Greenwald *et al.* (2019) and a corresponding RNA-seq dataset from Fadista *et al.* (2014) (see Table 1 for the summary of all datasets). According to the HbA1c index of each donor (American Diabetes Association, 2010) which measures the diabetic degree, we divided the RNA-seq dataset into two subsets, which are named islet healthy (HbA1c $<$ 6.5) and islet diabetic (HbA1c $\geq$ 6.5). Here, we considered autosomes and intra-chromosomal relationships only, although our framework can be easily adapted to sex chromosomes and inter-chromosomal relationships.

For the Hi-C datasets used in Varrone *et al.* (2020), we chose the resolution of 40 kb. For the pancreatic islet Hi-C dataset, since the processed contact matrix was not binned at 40 kb resolution, we selected the Hi-C data of resolution 50 kb which is the closest to 40 kb among the available resolutions. To normalize each Hi-C contact matrix, we used the method of iterative correction (Imakaev *et al.*, 2012), to preserve the local connectivity within the gene contact network during gene embedding. Next, to map contacts from the bin level to the gene level, we queried the transcription start sites (TSS) of all genes from the Ensembl database with GRCh37/hg19 genome assembly. Then, we map each gene to a bin that contains the TSS of that gene. As such, we can get the contact between two genes using the contact between their corresponding bins, following the TSS-mapping-based procedure in Babaei *et al.* (2015). Such gene-level contact information will be used to construct a gene contact network for each autosome. For the RNA-seq data, we used transcripts per kilobase million to measure gene expression levels from GTEx, RNA-Seq by expectation maximization for the data from TCGA and trimmed mean of M-values for the pancreatic islet data, as these measures were used in the original databases. We filtered out lowly expressed genes, i.e. those genes that have zero expression values in 80% of samples.

2.3 Constructing gene contact network and gene co-expression network

To estimate the co-expression between each pair of genes within the same chromosome, we calculated the Pearson correlation coefficient (PCC) between the two genes' expression profiles over the samples. As such, for each of the 22 autosomes, we obtained a co-expression matrix, the element of which is the PCC value. If the PCC value is above the 90th percentile across all the 22 co-expression matrices, then the gene pair is counted as significantly co-expressed (i.e. positive label); otherwise, it is counted as not co-expressed (i.e. a negative label). A gene co-expression network is constructed for each autosome by keeping only those significantly co-expressed gene pairs (COPs) as edges.

Similarly, we constructed a gene contact network for each autosome. To select reliable contacts for our prediction, we only took significant Hi-C contacts as the edges, i.e. above the 80th percentile of all Hi-C contacts in a chromosome. In this way, each autosome is associated with a gene contact matrix and a gene co-expression matrix, both containing elements of 0 or 1.

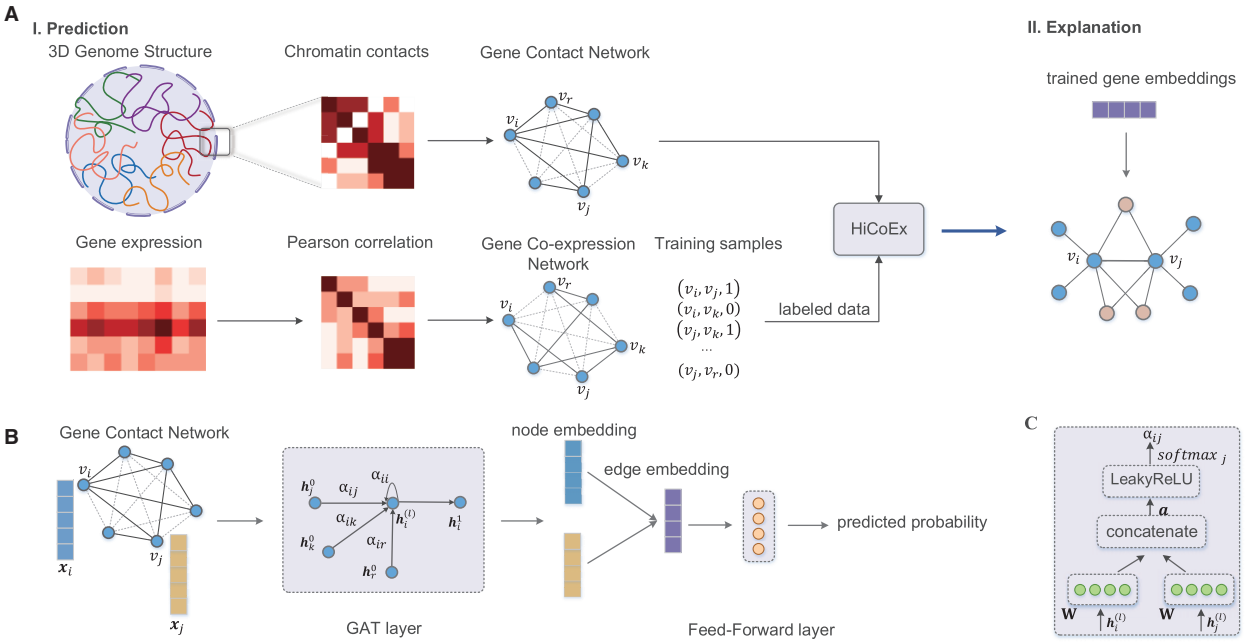

Fig. 1. Overview of HiCoEx. (**A**) The basic workflow. In the prediction module, Hi-C data are used to infer contacts within each chromosome which lead to a gene contact network, and gene expression data are used to construct a gene co-expression network. Dark solid lines in both networks represent that two genes have contacts or co-expression, while light-dashed lines represent the lack of either relationship. Two genes with co-expression relationship are defined as a positive instance and otherwise are taken as a negative instance. The topology of the gene contact network and the labeled data of co-expression are fed to HiCoEx for training. In the explanation module, the trained gene embeddings are first explained by topological properties and then the biological meaning is explored in the union subgraph of a pair of genes. (**B**) The architecture of HiCoEx. $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are the input feature vectors of gene v_i and gene v_j which are initialized randomly. Node embeddings are generated by a graph attention layer, and the node embeddings of two genes are merged through element-wise product to calculate the edge embedding. After passing through a feed-forward layer, the final score is obtained, which is the predicted probability of having the co-expression relationship between v_i and v_j. (**C**) Attention mechanism within a graph attention layer, which computes the attention coefficient α_{ij} between v_i and v_j

Table 1. Summary of datasets used in our experiments

Dataset	No. of gene nodes (RNA-seq)	No. of samples (RNA-seq)	No. of edges (gene contact network, intra-chrom)	No. of edges (gene co-expression network, intra-chrom)	Hi-C sample
Pancreatic islet healthy (Greenwald *et al.*, 2019)	13 747	66	1 039 257	533 849	T
Pancreatic islet diabetic (Greenwald *et al.*, 2019)	13 853	11	1 052 201	544 004	T
Adrenal gland (Schmitt *et al.*, 2016)	20 705	258	279 731	1 190 506	T
Aorta (Schmitt *et al.*, 2016)	20 528	432	694 523	1 133 207	T
Hippocampus (Schmitt *et al.*, 2016)	20 930	197	407 151	1 160 528	T
Left ventricle (Schmitt *et al.*, 2016)	19 011	432	434 768	994 113	T
Pancreas rep.1 (Schmitt *et al.*, 2016)	20 235	328	346 286	1 118 191	T
Pancreas rep.2 (Schmitt *et al.*, 2016)	20 235	328	135 865	1 118 191	T
Lung rep.1 (Schmitt *et al.*, 2016)	21 903	578	188 239	1 302 634	T
Lung rep.2 (Schmitt *et al.*, 2016)	21 903	578	293 378	1 302 634	T
Lung cell (Rao *et al.*, 2014)	21 903	578	1 618 428	1 302 634	CL
Breast cancer (Barutcu *et al.*, 2015)	14 519	1218	66 522	609 747	CL
Breast normal (Le Dily *et al.*, 2019)	21 353	459	166 293	1 286 132	CL
Prostate cancer (Rhie *et al.*, 2019)	14 643	550	657 946	583 446	CL

Note: Note that the islet dataset is split into healthy and diabetic subsets according to the HbA1c index of samples in RNA-seq data. For the column of Hi-C sample, T means Hi-C data sequenced from a tissue and CL means Hi-C data from a cell line.

2.4 Graph attention layer

Given a graph G, a vanilla GCN takes an adjacency matrix $\mathbf{A}$ of G and node features $\mathbf{X}$ as input. Here, G is a gene contact network built above. GCN generalizes the convolution to the graph-structure data. It learns the representation of a node by aggregating its neighbors' representations in G with equal weights (Wu *et al.*, 2021).

Graph Attention Network (GAT) (Velickovic *et al.*, 2018) can be regarded as an advanced GCN, since it assigns different attention weights for the neighbors of a central node during the neighborhood aggregation (Dzmitry *et al.*, 2015). By adopting graph attention in the model, we can distinguish the importance of different neighboring genes when characterizing a central gene.

The graph attention layer is the only component in a GAT model, and we employ one graph attention layer in HiCoEx. Following Velickovic *et al.* (2018), we denote a set of node representations by $\mathbf{H}^0 = \{\boldsymbol{h}_1^0, \boldsymbol{h}_2^0, \ldots, \boldsymbol{h}_N^0\}$, where $\boldsymbol{h}_i^0 \in \mathbb{R}^d$ is the hidden

embedding of node v_i at this graph attention layer and d is the dimension of node features. Here, $\boldsymbol{x}_i \in \mathbf{X}$ is randomly initialized for gene v_i as the input node feature, i.e. $\boldsymbol{h}_i^0 = \boldsymbol{x}_i$.

Suppose $\mathcal{N}_i$ is the set of neighbors of gene v_i, and let us take gene $v_j \in \mathcal{N}_i$ as an example. The attention coefficient e_{ij}^0, measuring the importance of v_j's embedding for v_i at this attention layer, is computed by an attention mechanism. Here, the attention mechanism is parameterized by $\boldsymbol{a} \in \mathbb{R}^{2d'}$ and transformed by LeakyReLU. Hence, the attention coefficient is:

$$e_{ij}^0 = \text{LeakyReLU}(\boldsymbol{a}^T[\mathbf{W}^0\boldsymbol{h}_i^0 || \mathbf{W}^0\boldsymbol{h}_j^0]), \tag{1}$$

where $\mathbf{W}^0 \in \mathcal{R}^{d' \times d}$ is the matrix of learned parameters of this layer $\cdot^T$ and $[\cdot||\cdot]$ denote the operations of transposition and concatenation respectively. The attention coefficient is then normalized across all the neighbors of gene v_i by the softmax function (Fig. 1C)

$$\alpha_{ij}^0 = \text{softmax}_j(e_{ij}^0) = \frac{\exp(e_{ij}^0)}{\sum_{k \in \mathcal{N}_i} \exp(e_{ik}^0)}. \tag{2}$$

Therefore, as illustrated in Figure 1B, the latent representation of gene v_i after this layer can be calculated by weighted aggregation as below:

$$\boldsymbol{h}_i^1 = \sigma\left(\sum_{j \in \mathcal{N}_i} \alpha_{ij}\mathbf{W}^0\boldsymbol{h}_j^0\right), \tag{3}$$

where $\sigma(\cdot)$ is the activation function at a graph attention layer. Then, we obtain the embedding of an edge between two genes by taking the element-wise product of their embeddings as follows:

$$\boldsymbol{h}_{ij}^1 = [\boldsymbol{h}_i^1 \cdot \boldsymbol{h}_j^1]. \tag{4}$$

2.5 Feed-Forward layer

A Feed-Forward layer can increase the nonlinearity of latent embedding and be used as the classifier. For a gene pair (v_i, v_j), the predicted gene co-expression can be calculated by:

$$\hat{q}_{ij} = p(\mathbf{W}_{ff}\boldsymbol{h}_{ij}^1 + b_{ff}), \tag{5}$$

where ($\mathbf{W}_{ff}$ and b_{ff}) are parameters of the feed-forward layer, and $p(\cdot)$ is the activation function.

Furthermore, we adopt cross entropy for our binary classification and therefore our objective is to minimize:

$$\mathcal{L}(\Theta) = \frac{1}{|\mathcal{C}|}\sum_{(v_i,v_j)\in\mathcal{C}} (q_{ij}\log(\hat{q}_{ij}) + (1-q_{ij})\log(1-\hat{q}_{ij})). \tag{6}$$

q_{ij} from gene co-expression matrix $\mathbf{Q}$ is the real co-expression between v_i and v_j. Θ represents all the training parameters in the model and $\mathcal{C} = \{(v_i, v_j)\}$ is the set of gene pairs for training.

2.6 Experimental setup

2.6.1 Baselines

We compare our model with the following baseline models, named according to their feature extraction methods and classifiers. For all the models (except Topology-RF), we first obtain the embedding for each gene and then compute the embedding of each edge by the element-wise product of the embeddings of the two genes connected through the edge. The first four models have been implemented and compared in Varrone *et al.* (2020), and we re-use their code for fair comparison here.

- Distance-RF: This method takes the 1D genomic distance between two genes as the input feature for a random forest (RF) classifier.
- Topology-RF: This method encodes the features of a gene pair based on five topological properties of the gene contact network, which are shortest path length, Jaccard index, degree centrality, betweenness centrality and clustering coefficient Babaei *et al.* (2015). Then, a random neural network (Babaei *et al.*, 2015) or a RF (Varrone *et al.*, 2020) is used to predict gene co-expression.
- Singular vector decomposition (SVD)-RF: The gene contact network can be embedded through matrix factorization. Using SVD, the adjacency matrix of a gene contact network is decomposed into two factors, and then gene embeddings are obtained by summing up the two factors.
- Node2vec-RF: Genes with similar roles or within the same communities should have similar representations. Based on this assumption, the node2vec algorithm is used in Varrone *et al.* (2020) to learn random walks for each node in the gene contact network to get the corresponding gene's embedding.
- GCN: This is a vanilla model of graph convolutional network, which automatically learns the gene embeddings and predicts gene co-expression with a feed-forward layer similar to HiCoEx. GCN assigns the same weights to the neighbors of a node during the node embedding propagation.

2.6.2 Implementation details

We evaluate all the methods based on Accuracy. We randomly sample gene pairs in each dataset to balance the sizes of positive and negative samples. Then, we split each of our datasets by the proportion 7:1:2 for training, validation and testing.

In our model, we use one graph attention layer and exponential linear units (Clevert *et al.*, 2016) as the activation function in Equation (3), and we use a softmax as the activation function in Equation (5). The Adam algorithm (Kingma and Ba, 2015) is used as the optimizer. We linearly reduce the learning rate by half in each iteration from the initial value 1e-3 until the maximum epoch 100. The batch size is set to 64, and the latent embedding size is 16, the same as in Varrone *et al.* (2020). Batch normalization is used for regularization, and the dropout rate is set to 0.5. We run each experiment three times to compute the mean and SD of accuracy, using an Nvidia Tesla V100 GPU.

2.7 Explanations of HiCoEx

Recently, explainable AI techniques have been proposed for GNNs (Schlichtkrull *et al.*, 2021; Ying *et al.*, 2019), and one of the explanation methods is to explore the structural information encoded in the learned graph embeddings. Here, We employed a method in Jin *et al.* (2021) to explain the gene embeddings trained from HiCoEx. This method is based on the hypothesis that nodes near each other in the embedding space have similar structural properties.

For a gene in a Hi-C contact network, we first selected its k-nearest neighbors (k-NNs) by using Euclidean distance (we have also tried Cosine distance) in the embedding space. Then, we calculated six types of topological properties, namely Degree Centrality, Between Centrality, Clustering Coefficient, PageRank, Jaccard Index and Shortest path length. For Jaccard Index and Shortest path length, since they are properties for gene pairs, we calculated their values for a single gene by adding the corresponding values of all gene pairs that include this gene. For each property, we also calculated the average value over the k-NNs of a gene, as the property value of the neighborhood of the gene. Finally, PCC about each topological property is computed between the genes and their neighborhoods in the embedding space. If the PCC about a topological property is higher, it means that this topological property has been better encoded in the gene embeddings and thereby, it makes a more significant contribution to the prediction of gene co-expression than other topological properties.

3 Results

3.1 Predicting gene co-expression by chromosome-specific gene contacts

To assess the performance of our model, we first conducted experiments for each chromosome separately. Here, we set the node

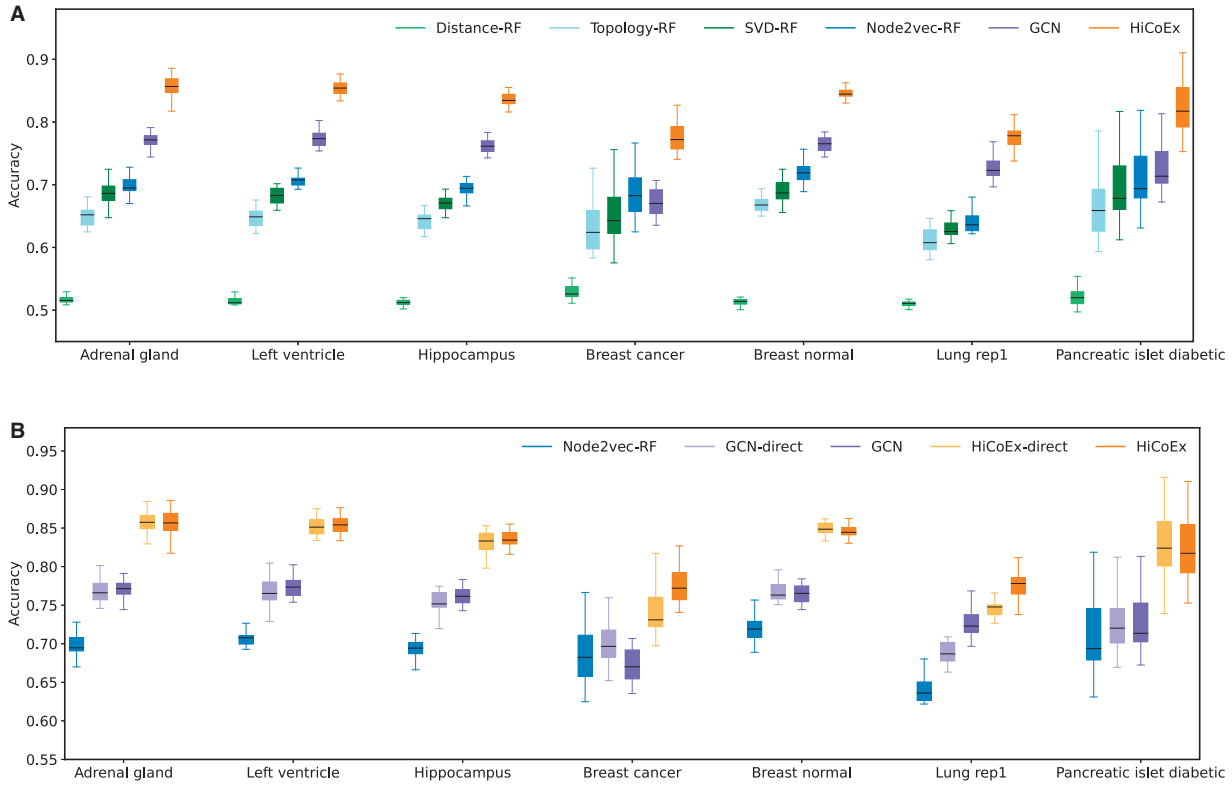

Fig. 2. Performance comparison of methods using intra-chromosomal gene contact networks for 22 autosomes separately on several datasets. (**A**) Overall accuracy of all methods on seven datasets. (**B**) Accuracy of the methods after removing the feed-forward layer of two GNN-based models. Each boxplot describes the distribution of the prediction results over 22 chromosomes for a specific tissue or cell line

embedding size to 16 as in Varrone *et al.* (2020) and replicated each experiment three times to obtain the average accuracy. For each type of tissue or cell line, the prediction results on 22 autosomes are integrated within a boxplot (Fig. 2A). See Supplementary Figure S1A and Table S1 for complete results.

Measured by classification accuracy, distance-RF has the worst performance, although it still exceeds 0.5, the random accuracy of a binary classifier. This result suggests that the genomic distance is not a key factor that affects gene co-expression. The two GNN-based methods (i.e. GCN and HiCoEX) outperform the three feature engineering-based methods (i.e. Topology-RF, SVD-RF and Node2vec-RF). HiCoEx outperforms other models on all datasets.

To examine whether the good performance of GCN and HiCoEx depends on the choice of classifier, we removed the feed-forward layer from the two GNN-based methods. The modified GNN models learn node embeddings with size 2, and the final prediction is directly calculated by the dot product between the embeddings of a pair of genes. In this way, we implemented new GNN models (named *-direct as in Fig. 2B and Supplementary Fig. S1B). Then, we trained the two models on all the datasets and compared them to Node2vec-RF, the best non-GNN baseline model. We observed that the removal of the feed-forward layer leads to only small changes in the prediction accuracy. As shown in Figure 2B, HiCoEx-direct had comparable results with HiCoEx and even performed better than HiCoEx on the breast normal dataset. Moreover, the four GNN-based methods still outperformed Node2vec-RF among all the datasets. These results suggest that the graph representation learning is crucial for gene co-expression prediction.

3.2 Predicting gene co-expression from combined intra-chromosomal contacts

After co-expression prediction for each individual chromosome, we conducted similar experiments on the set of all autosomes. For each type of tissue or cell line, we constructed a network containing all intra-chromosomal gene contacts as the input data. Following the experimental procedure of Varrone *et al.* (2020), we randomly chose some gene pairs as the training dataset. We kept the same hyperparameters as before, except that we used a larger batch size and removed all regularization terms during the training of the GNN-based models. For each tissue or cell line, since there is only one dataset, instead of 22 datasets for the autosomes, we used bar graphs instead of boxplots to summarize the results (Fig. 3 and Supplementary Fig. S2). When compared with the non-GNN baselines, the two GNN-based models perform better on all the datasets, and HiCoEx outperforms GCN on 12 out of 14 datasets. Interestingly, the two GNN models have accuracies close to each other on the lung rep.1 dataset, and GCN has higher accuracy than HiCoEx on the breast cancer dataset. This indicates that the attention mechanism as used in HiCoEx may not help too much when all intra-chromosomal contacts are taken as input.

3.3 Parameter analysis

Several parameters are important for constructing gene contact networks and gene co-expression networks, such as the bin size (i.e. resolution) of Hi-C data, thresholds for binarizing gene–gene contacts and gene–gene co-expression values. For each parameter, we selected different values to analyze its impact on the prediction performance.

We used Hi-C contact maps at different resolutions on the breast cancer dataset as input data. The resolution is set to 40, 100, 250 kb and 1 Mb, respectively. The results in Figure 4A show that HiCoEx outperforms other baselines at four types of Hi-C resolutions. Moreover, we found that, as the resolution decreased from 40 kb to 1 Mb, the accuracies of most models also decreased. HiCoEx achieves the best performance at the finest resolution of 40 kb, probably because Hi-C data at a finer resolution can identify more contacts between proximal genes, which helps predict gene co-expression.

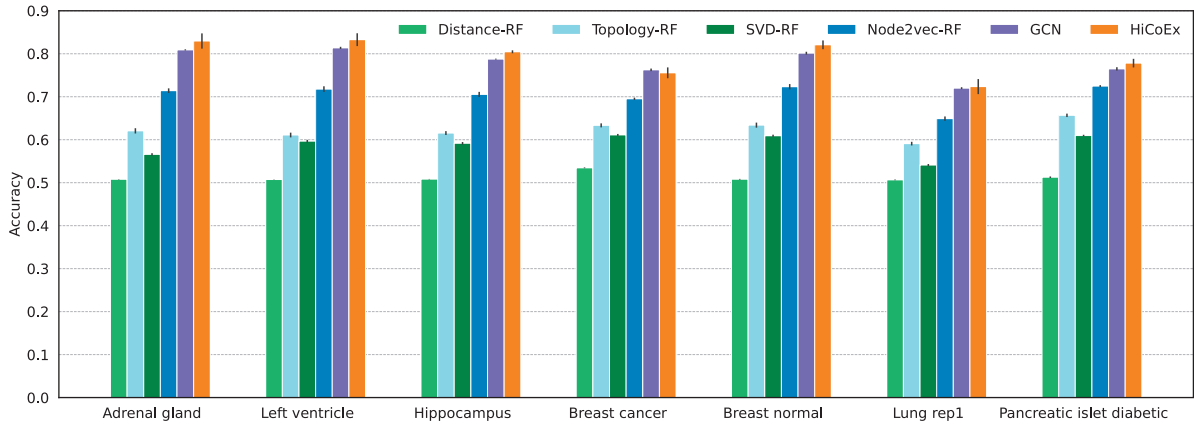

Fig. 3. Performance comparison of methods using intra-chromosomal gene contacts combined from 22 autosomes. The height of each bar represents the average accuracy of three replicated experiments and the vertical line corresponds to the SD

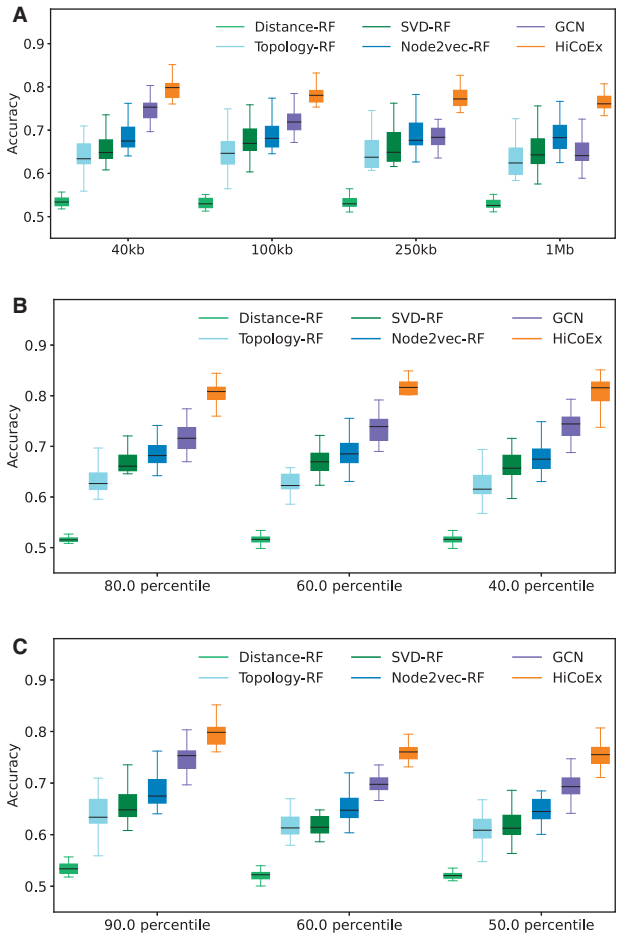

Fig. 4. Performance comparison of different Hi-C resolutions, Hi-C thresholds and co-expression thresholds. (**A**) Accuracy of using different Hi-C resolutions on the breast cancer dataset. (**B**) Accuracy of using different Hi-C thresholds on the prostate cancer dataset. (**C**) Accuracy of using different co-expression thresholds on the breast cancer dataset. Each boxplot describes the distribution of the prediction results over 22 chromosomes

Next, we conducted experiments of preserving Hi-C contacts above different threshold values on the prostate cancer dataset. The threshold value for Hi-C contact is set to 80th, 60th and 40th percentile, respectively (Fig. 4B). Decreasing the Hi-C threshold affects none of the models in their prediction accuracy significantly, which means that preserving more contacts cannot help improve the prediction performance. Additionally, we conducted experiments of preserving gene co-expression above different threshold values on the breast cancer dataset. The threshold for co-expression value is set to 90th, 60th and 50th percentile, respectively. The result (Fig. 4C) shows that our model still performs the best with the three cutoff values. Besides, we noticed that, as the threshold decreased from 90th to 50th percentile, the accuracies of all models decreased as well. This is probably because more weakly COPs are preserved as the threshold for gene co-expression becomes lower. These gene pairs might be assigned to wrong classes and therefore mislead the model training.

3.4 Learned gene embeddings explain predictions of HiCoEx

To better understand what types of structural properties HiCoEx could encode, we analyzed the Pearson correlations between some topological features (part of which are extracted from [1]) and gene embeddings learned by HiCoEx.

Let us take Chromosome 7 of breast cancer dataset as an example (Fig. 5A and Supplementary Fig. S3). Among all the topological properties, the correlation about Jaccard Index is the highest and the correlation about Degree Centrality is the second highest. Similar results have been observed on most other autosomes (13 of 21). These results suggest that, among the six topological properties, Jaccard Index is encoded by the gene embeddings from HiCoEx better than all other topological properties. In a graph, Jaccard Index is the ratio of the number of common neighbors shared by two nodes to the number of all the neighbors connected to at least one of the two nodes. The high correlation about the Jaccard Index suggests that, in a gene contact network, the common neighbors of two central genes could be important for learning gene embeddings and predicting the co-expression between the two genes.

To further test which gene pair-based properties are encoded in the embeddings, we performed a prediction task similar to that in Dalmia *et al.* (2018) and Jin *et al.* (2021). The embedding of a pair of genes (i.e. the edge embedding) is calculated by the element-wise product of the embeddings of the two genes. Given the edge embedding of a gene pair as input, we used a k-NN regression model to predict the value of a pair-based topological property for the gene pair. We used 5-fold cross-validation and took the mean Root mean square error (RMSE) across the 5 folds as the prediction error for each chromosome. Here, we predicted four pair-based topological properties, namely Jaccard Index, Resource Allocation Index, Shortest Path Length and Pairwise Node Connectivity, using the edge embeddings. In Figure 5B, the prediction of Jaccard Index exhibits the smallest errors, indicating that Jaccard Index is the most correlated with the learned edge embeddings among the four topological properties. Besides, the errors of predicting both Jaccard Index and Resource Allocation Index are significantly smaller than the errors of predicting the other two properties. Since the former

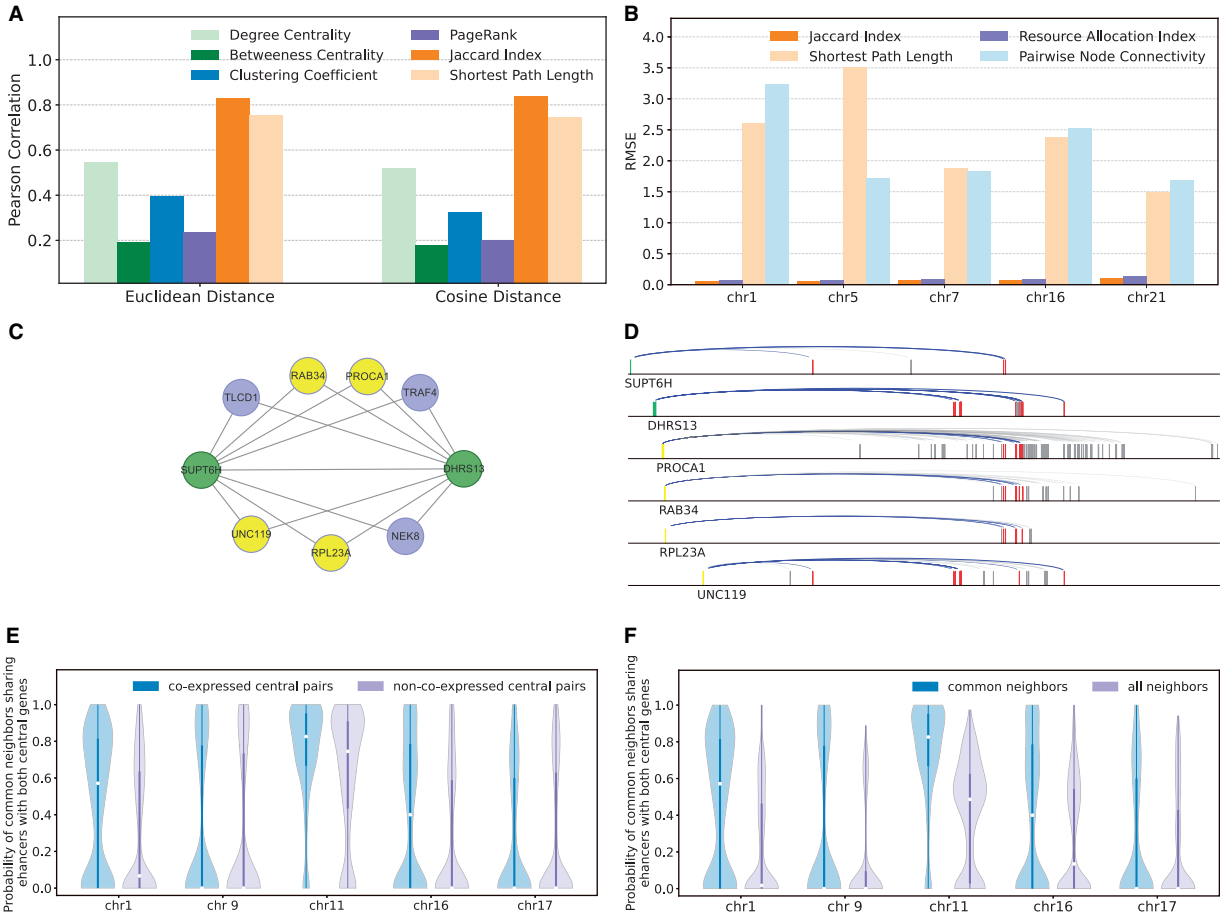

Fig. 5. Explanation of gene embeddings learned from HiCoEx on the breast cancer dataset. (**A**) Correlation of topological properties between genes and their k-NNs, where k-NNs are selected by Euclidean distance or Cosine distance in the embedding space. The height of each bar represents the PCC calculated on Chromosome 7. (**B**) Comparison of prediction errors of gene pair-based topological properties by the learned edge embeddings. (**C**) The intersection subgraph of a central gene pair (SUPT6H, DHRS13) on Chromosome 17. RAB34, PROCA1, UNC119 and RPL23A are the common neighbors sharing enhancers with both SUPT6H and DHRS13, and TLCD1, TRAF4 and NEK8 are the common neighbors not sharing enhancers with the two central genes. (**D**) A diagram of gene–enhancer interactions. Shared enhancers are plotted by highlighted bars, and interactions between a gene and shared enhancers are plotted by highlighted curves. The highlighted bars vertically aligned refer to the identical location and DNA sequence of the enhancers shared between one of the central genes and the common neighbors. (**E**) Comparison of distributions between COPs and non-COPs. The distribution is over the intersection subgraphs of the central gene pairs with at least 10 common neighbors in a chromosome. (**F**) Comparison of distributions for COPs with common neighbors and with all neighbors. The distribution is over the intersection subgraphs of the COPs with at least 10 common neighbors in a chromosome. Each violin plot in (**E**) and (**F**) describes a distribution of the probability that, in an intersection subgraph, a common neighbor shares enhancers with both central genes

two properties are both based on the number of common neighbors shared by two central genes, this result again shows the importance of the common neighboring genes for the prediction of the two central genes' co-expression.

Based on the above observations, we extracted one-hop subgraphs for two genes respectively (because one GAT layer used in our model integrates first-order neighbors for a gene during message passing) and took the intersection between the two subgraphs. Hereafter, we call it the intersection subgraph of the central gene pair. Then we explored the contribution of the common neighbors to the co-expression of the two genes. As one of the most important regulatory elements, an enhancer could interact with some target genes and is critical to coordinating their transcriptional activities. Therefore, we examined whether enhancers are correlated with the common neighbors of two genes. Here, we only consider the gene pairs which are both co-expressed and interacting with each other. We searched gene–enhancer interactions from PantherDB (Thomas *et al.*, 2003). From the intersection subgraph of each gene pair, we found that common neighbors are more likely to share enhancers with both central genes. Figure 5C displays the intersection subgraph of SUPT6H and DHRS13 on Chromosome 17, plotted using Cytoscape. Four out of seven common neighbors share enhancers with both SUPT6H and DHRS13. Figure 5D is a diagram describing the gene–enhancer interactions. Here, we only show the gene–enhancer interactions for the two central genes and four of their common neighbors (i.e. the four nodes colored in yellow in Fig. 5C) that share enhancers with both central genes. In Figure 5D, we can see that the two central genes, SUPT6H and DHRS13, do not directly share any enhancer with each other, and thus their co-expression might be driven by the enhancers shared with their common neighbors.

We tested whether the common neighbors of COPs are more likely to share enhancers with central genes than those of non-COPs. Here, all the central gene pairs we analyzed interact with each other. For a pair of central genes, we estimated the probability that, in a subgraph, a common neighboring gene (i.e. it contacts both central genes) shares enhancers with both central genes. Note that the two central genes may or may not share any enhancer between themselves. We selected 14 autosomes for analysis, because Jaccard Index values are the most correlated with the learned node embeddings on each of these chromosomes. Since the calculation of the probability needs a sufficient number of common neighbors for a gene pair, to estimate the probability more accurately, we first filtered out the COPs and non-COPs with not enough common neighbors (i.e., the number of common neighbors is less than a threshold, which is set to 10) for all the 14 autosomes. We then picked only the autosomes with at least 100 COPs and 100 non-COPs each with an enough number of common neighbors, i.e. chr1, chr9, chr11, chr16 and chr17, for comparison. We downloaded the data of gene–enhancer interactions of breast cancer cells from the database of

Enhancer Atlas V2.0 (Gao and Qian, 2019). Figure 5E shows that on three of the five autosomes, the common neighbors of COPs are more likely to share enhancers with both central genes than those of non-COPs (with t-test P values < 0.05). This indicates that through sharing enhancers, the common neighboring genes might contribute to the co-expression of the two central genes.

To further examine whether such common neighbors contribute to gene co-expression, we compared the probability distributions of enhancer sharing between neighboring and central genes for COPs. The probabilities were estimated on two sets of neighboring genes. One is the set of common neighbors, and the other is the set of all neighbors (i.e. all the genes that contact at least one of the two central genes). Similar to Figure 5E, we performed the comparisons on the five autosomes. Figure 5F shows that the probability for a common neighbor to share enhancers with both central genes (which are co-expressed and interacting with each other) is significantly higher than that of a randomly sampled neighbor, and such a tendency is consistent across the five autosomes (with t-test P values < 0.05). In other words, compared to non-common neighbors, the common neighbors are more likely to share enhancers with the co-expressed and contacted central genes. This demonstrates the importance of spatially proximal genes for gene co-expression.

4 Discussion

Exploring the relationship between spatial genome organization and gene co-expression can shed light on the epigenomic mechanisms of transcriptional regulation. To this end, we presented HiCoEx, a novel machine learning framework based on GNN for explainable prediction of gene co-expression using Hi-C data, which comprises a predictive model followed by an explanation technique. HiCoEx is able to automatically capture important patterns in the Hi-C and RNA-seq data for the prediction of co-expression from chromosomal contacts between genes, and visualize the gene–gene interactions for mechanistic exploration.

The accurate predictions of our model can be explained by uncovering the structural information encoded in the learned gene embeddings. Our analyses suggest that the GNN-based model could automatically capture some topological properties, especially Jaccard Index on most chromosomes. This also means that the topology of the gene contact network may be important for gene co-expression, which is consistent with the previous research (Babaei *et al.*, 2015). According to the above findings, for gene pairs which are co-expressed and interacted with each other, we further visualized the 1-hop union subgraphs of these gene pairs. We found that the common proximal genes, sharing the same enhancers with two central genes, may be responsible for two genes' co-expression. By conducting the above explanations, we showed that our framework could be utilized to aid the exploration of the mechanism of gene regulations. Here, we have explored several topological features. In the future, we may include more types of topological features for analysis, such as Closeness Centrality derived from genes and Local Path Index (Lü *et al.*, 2009) derived from gene pairs. Moreover, other regulatory elements (e.g. transcription factor binding sites) could be also analyzed (Chepelev *et al.*, 2012; Ribeiro *et al.*, 2021a,b), e.g. whether they are also shared by central genes and their spatially proximal genes, and how these regulatory elements work together to coordinate transcriptional regulation.

In this article, we predicted gene co-expression using spatial contacts between genes inferred from Hi-C data. Our explanation emphasizes the importance of common neighboring genes and their functional interactions with enhancers for gene co-expression. Therefore, in the future, ChIA-PET data are needed to identify such functional interactions and characterize the complex relations between genes and their regulatory elements (Avsec *et al.*, 2021; Cao *et al.*, 2017). Additionally, Hi-C data with finer resolutions are also needed to distinguish precise gene–-gene contacts. By combining the two types of spatial interactions above, we can construct a more comprehensive network, so as to further decipher the relationship between 3D chromatin organization and gene regulation mechanism.

Acknowledgements

We would like to thank Frank Alber, Allison Furterer and Asli Yildirim at UCLA and Hanhui Ma at ShanghaiTech University for their advice about data pre-processing and valuable discussions. We acknowledge the Pancreatic ß-Cell Consortium for providing the context in which this research is performed.

Financial Support: none declared.

Conflict of Interest: none declared.

Data availability

Hi-C datasets used in this study are available under accession numbers `TSTSR043623`, `GSE87112`, `GSE66733`, `GSE63525` and `GSE118629`. RNA-seq datasets are available from TCGA for cancer cell lines and GTEx for normal tissues and cell lines except that pancreatic islet is from `GSE50244`.

References

Ahn,J.H. *et al.* (2021) Phase separation drives aberrant chromatin looping and cancer development. *Nature*, **595**, 591–595.

American Diabetes Association. (2010) Standards of medical care in diabetes—2010. *Diabetes Care*, **33** (Suppl. 1), S11–S61.

Avsec,Z. *et al.* (2021) Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods*, **18**, 1196–1203.

Babaei,S. *et al.* (2015) Hi-c chromatin interaction networks predict co-expression in the mouse cortex. *PLoS Comput. Biol.*, **11**, e1004221.

Barutcu,A.R. *et al.* (2015) Chromatin interaction analysis reveals changes in small chromosome and telomere clustering between epithelial and breast cancer cells. *Genome Biol.*, **16**, 1–14.

Bhat,P. *et al.* (2021) Nuclear compartmentalization as a mechanism of quantitative control of gene expression. *Nat. Rev. Mol. Cell Biol.*, **22**, 653–618.

Cao,Q. *et al.* (2017) Reconstruction of enhancer–target networks in 935 samples of human primary cells, tissues and cell lines. *Nat. Genet.*, **49**, 1428–1436.

Cao,Q. *et al.* (2020) A unified framework for integrative study of heterogeneous gene regulatory mechanisms. *Nat. Mach. Intell.*, **2**, 447–456.

Chepelev,I. *et al.* (2012) Characterization of genome-wide enhancer-promoter interactions reveals co-expression of interacting genes and modes of higher order chromatin organization. *Cell Res.*, **22**, 490–503.

Clevert,D. *et al.* (2016) Fast and accurate deep network learning by exponential linear units (ELUs). In: *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2–4, 2016, Conference Track Proceedings*.

Dalmia,A. *et al.* (2018) Towards interpretation of node embeddings. In: *Companion of the The Web Conference 2018 on The Web Conference 2018, WWW 2018, Lyon, France, April 23–27, 2018*, pp. 945–952.

Dekker,J. and Misteli,T. (2015) Long-range chromatin interactions. *Cold Spring Harb. Perspect. Biol.*, **7**, a019356.

Dong,X. *et al.* (2010) Human transcriptional interactome of chromatin contribute to gene co-expression. *BMC Genomics*, **11**, 704–705.

Dzmitry,B. *et al.* (2015) Neural machine translation by jointly learning to align and translate. In: Bengio,Y. and LeCun,Y. (eds.) *International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings*.

Fadista,J. *et al.* (2014) Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism. *Proc. Natl. Acad. Sci. USA.*, **111**, 13924–13929.

Fout,A. *et al.* (2017) Protein interface prediction using graph convolutional networks. In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4–9, 2017, Long Beach, CA, USA*, pp. 6530–6539.

Gao,T. and Qian,J. (2019) EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. *Nucleic Acids Res.*, **48**, D58–D64.

Greenwald,W.W. *et al.* (2019) Pancreatic islet chromatin accessibility and conformation reveals distal enhancer networks of type 2 diabetes risk. *Nat. Commun.*, **10**, 1–12.

Grover,A. and Leskovec,J. (2016) node2vec: Scalable feature learning for networks. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13–17, 2016*, pp. 855–864.

Ibn-Salem,J. *et al.* (2017) Co-regulation of paralog genes in the three-dimensional chromatin architecture. *Nucleic Acids Res.*, **45**, 81–91.

Imakaev,M. *et al.* (2012) Iterative correction of hi-c data reveals hallmarks of chromosome organization. *Nat. Methods*, **9**, 999–1003.

Jin,J. *et al.* (2021) Toward understanding and evaluating structural node embeddings. *ACM Trans. Knowl. Discov. Data*, **16**, 1–32.

Kingma,D.P. and Ba,J. (2015) Adam: a method for stochastic optimization. In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings*.

Kipf,T.N. and Welling,M. (2017) Semi-supervised classification with graph convolutional networks. In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24–26, 2017, Conference Track Proceedings*.

Lanchantin,J. and Qi,Y. (2020) Graph convolutional networks for epigenetic state prediction using both sequence and 3d genome data. *Bioinformatics*, **36**, i659–i667.

Le Dily,F. *et al.* (2019) Hormone-control regions mediate steroid receptor–dependent genome organization. *Genome Res.*, **29**, 29–39.

Lü,L. *et al.* (2009) Similarity index based on local paths for link prediction of complex networks. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, **80**, 046122.

Niepert,M. *et al.* (2016) Learning convolutional neural networks for graphs. In: *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19–24, 2016*, Vol. 48, pp. 2014–2023.

Rao,S.S. *et al.* (2014) A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**, 1665–1680.

Rhie,S.K. *et al.* (2019) A high-resolution 3D epigenomic map reveals insights into the creation of the prostate cancer transcriptome. *Nat. Commun.*, **10**, 1–12.

Ribeiro,D.M. *et al.* (2021a) The molecular basis, genetic control and pleiotropic effects of local gene co-expression. *Nat. Commun.*, **12**, 1–13.

Ribeiro,D.M. *et al.* (2021b) Shared regulation and functional relevance of local gene co-expression revealed by single cell analysis. *bioRxiv*. https://doi.org/10.1101/2021.12.14.472573.

Sandhu,K.S. *et al.* (2012) Large-scale functional organization of long-range chromatin interaction networks. *Cell Rep.*, **2**, 1207–1219.

Schlichtkrull,M.S. *et al.* (2021) Interpreting graph neural networks for NLP with differentiable edge masking. In: *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3–7, 2021*.

Schmitt,A.D. *et al.* (2016) A compendium of chromatin contact maps reveals spatially active regions in the human genome. *Cell Rep.*, **17**, 2042–2059.

Thibodeau,A. *et al.* (2017) Chromatin interaction networks revealed unique connectivity patterns of broad h3k4me3 domains and super enhancers in 3d chromatin. *Sci. Rep.*, **7**, 1–12.

Thomas,P.D. *et al.* (2003) Panther: a library of protein families and subfamilies indexed by function. *Genome Res.*, **13**, 2129–2141.

Tian,D. *et al.* (2020) Mochi enables discovery of heterogeneous interactome modules in 3d nucleome. *Genome Res.*, **30**, 227–238.

Varrone,M. *et al.* (2020) Exploring chromatin conformation and gene co-expression through graph embedding. *Bioinformatics*, **36**, i700–i708.

Velickovic,P. *et al.* (2018) Graph attention networks. In: *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30–May 3, 2018*, Conference Track Proceedings.

Wu,Z. *et al.* (2021) A comprehensive survey on graph neural networks. *IEEE Trans. Neural Netw. Learn. Syst.*, **32**, 4–24.

Ying,Z. *et al.* (2019) Gnnexplainer: Generating explanations for graph neural networks. In: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8–14, 2019, Vancouver, BC, Canada*, pp. 9240–9251.

Yu,M. and Ren,B. (2017) The three-dimensional organization of mammalian genomes. *Annu. Rev. Cell Dev. Biol.*, **33**, 265–289.

Zhang,R. and Ma,J. (2020) Matcha: probing multi-way chromatin interaction with hypergraph representation learning. *Cell Syst.*, **10**, 397–407.