Bioinformatics, 36(26), 2020, i700–i708
doi: 10.1093/bioinformatics/btaa803
ECCB2020

Genomes

Exploring chromatin conformation and gene co-expression through graph embedding

Marco Varrone[1,†], Luca Nanni[1,*,†], Giovanni Ciriello[2,3] and Stefano Ceri[1]

[1]Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milan, Italy, [2]Department of Computational Biology, University of Lausanne, 1015 Lausanne, Switzerland and [3]Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Abstract

Motivation: The relationship between gene co-expression and chromatin conformation is of great biological interest. Thanks to high-throughput chromosome conformation capture technologies (Hi-C), researchers are gaining insights on the tri-dimensional organization of the genome. Given the high complexity of Hi-C data and the difficult definition of gene co-expression networks, the development of proper computational tools to investigate such relationship is rapidly gaining the interest of researchers. One of the most fascinating questions in this context is how chromatin topology correlates with gene co-expression and which physical interaction patterns are most predictive of co-expression relationships.

Results: To address these questions, we developed a computational framework for the prediction of co-expression networks from chromatin conformation data. We first define a gene chromatin interaction network where each gene is associated to its physical interaction profile; then, we apply two graph embedding techniques to extract a low-dimensional vector representation of each gene from the interaction network; finally, we train a classifier on gene embedding pairs to predict if they are co-expressed. Both graph embedding techniques outperform previous methods based on manually designed topological features, highlighting the need for more advanced strategies to encode chromatin information. We also establish that the most recent technique, based on random walks, is superior. Overall, our results demonstrate that chromatin conformation and gene regulation share a non-linear relationship and that gene topological embeddings encode relevant information, which could be used also for downstream analysis.

Availability and implementation: The source code for the analysis is available at: https://github.com/marcovarrone/gene-expression-chromatin.

Contact: luca.nanni@polimi.it

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

1 Introduction

The human genome counts approximately more than 20 000 protein-coding genes (International Human Genome Sequencing Consortium *et al.*, 2001, 2004) and their exact number is still unknown (Salzberg, 2018). The availability of gene expression profiling technologies like DNA microarrays and RNA sequencing (Emrich *et al.*, 2007) enables massive studies of gene expression patterns across tissues and clinical conditions. In this context, the study of gene co-expression networks plays a role of major interest.

Co-expression between two genes can indicate a functional relation, the belonging to a shared transcriptional regulatory program or their participation in the same pathway (Stuart *et al.*, 2003). The analysis of co-expression networks have been successfully used to determine gene-disease associations (van Dam *et al.*, 2018) or gene modules associated with a phenotype of interest (Chou *et al.*, 2014; Kogelman *et al.*, 2014; Oh *et al.*, 2015; Yang *et al.*, 2014; Zhao *et al.*, 2010). Co-expression networks can be inferred from expression profiling data using several methods (Zhang and Horvath, 2005), ranging from Pearson correlation (Ala *et al.*, 2008; Langfelder and Horvath, 2008; Stuart *et al.*, 2003) to entropy measurements (Butte butte 2000).

A relevant issue in structural biology is understanding the relationship between gene co-expression and the spatial configuration of the genome (Dekker and Misteli, 2015). Chromosome conformation capture technologies, especially high-throughput chromosome conformation capture (Hi-C) (Lieberman-Aiden *et al.*, 2009), are used to reconstruct the high level three-dimensional architecture of the genome. The human genome reveals a hierarchical structure (Lieberman-Aiden *et al.*, 2009; Szabo *et al.*, 2019); a striking property of genome folding is the existence of sub-megabase regions of strong self-interactions, which were called topologically associating domains (TADs) (Dixon *et al.*, 2012). Initial studies about TADs revealed that genes belonging to the same domain show similar

expression patterns (Dixon *et al.*, 2012; Gonzalez-Sandoval and Gasser, 2016). Another hallmark of genome organization is loops, defined as points of significantly strong interaction in the Hi-C contact matrix, which have been associated with the presence of CTCF binding sites (Rao *et al.*, 2014).

Recent works highlighted the need for integrating gene expression and chromatin conformation to study gene activity (Babaei *et al.*, 2015; Delaneau *et al.*, 2019; Kustatscher *et al.*, 2017) showing how gene activity tends to cluster in cis- and trans-regulatory domains. Therefore, computational tools for the joint study of genome conformation and gene expression are rapidly gaining the interest of the research community (Tian *et al.*, 2019; Zhou *et al.*, 2019). Both the gene chromatin and co-expression networks can be represented by graphs, having genes as nodes and the relationships between genes as edges, weighted either by the strength of physical interaction or by expression correlation. Topological features of each node are modeled by means of features, called *node embeddings*.

In this work, we explore the relationship between chromatin conformation and gene expression using a predictive modeling approach. This enables us to study which feature embedding strategies are best suited for encoding chromatin topology information and to what extent they are able to infer co-expression relations between genes. Specifically, we predict co-expression between two genes from the physical set of interactions derived from a Hi-C experiment. Previous work addressed this problem by computing a set of predefined measures for each gene/node in the Hi-C network to be used as input of the classifier (Babaei *et al.*, 2015); this strategy assures full transparency of the studied topological features, but it hardly captures network topologies and node similarities. We instead explore the use of representation learning (Bengio *et al.*, 2012) for embedding the topological features of genes. Representation learning on graphs (Hamilton *et al.*, 2017) is a rapidly emerging trend in machine learning, recently applied in biology to many network inference tasks (Dai *et al.*, 2015; Du *et al.*, 2019; Nelson *et al.*, 2019; You *et al.*, 2017; Yue *et al.*, 2019). In our work, the features of the nodes are learnt by solving an optimization problem, which defines the embedding strategy of the physical interaction network extracted from Hi-C data. Therefore, the proper choice of the optimization method is critical.

We compare two different node embedding strategies. The first method is based on Matrix Factorization (Yue *et al.*, 2019), while the second exploits a random walk procedure to find similar embeddings for genes in the same neighborhood (Grover and Leskovec, 2016). We then use the learnt embeddings to train a non-linear classifier, based on random forest (Breiman, 2001), and compare the performances between the two embedding strategies and against a set of baselines. We validate our models on an extensive set of tissues, cell-lines and conditions, where Hi-C data and the relative gene expression is available. Our results show that both our embedding methods outperform previous approaches, highlighting the need for more complex gene topological representations. Results are finally validated with a holdout dataset. Automatic learning of genomic features can therefore unravel latent relationships between chromatin organization and gene expression and the learnt representations have the potential to be used for downstream analyses.

2 Materials and methods

2.1 Overview of the method

Our method is composed of three main tasks: generation of the gene chromatin network, learning of the gene embeddings and prediction of co-expression links.

- In the first task, we summarize the chromatin interaction information of genes from Hi-C maps by taking into account also their neighboring regions. This produces a gene network whose edges represent the strength of physical interaction between pairs of genes.
- This network is then used as input of a node embedding algorithm to build a vector representation of each gene. For this task, we propose two embedding strategies, respectively referring to matrix factorization and to the generation of random walks. We also perform a comparative analysis between them and with other known methods.
- The final task of the pipeline takes as input pairs of genes and their embeddings and predicts if they are co-expressed. The prediction algorithm is trained by using as training set a subset of the interactions. For this task, we used a non-linear classifier based on random forest.

The overall framework for gene co-expression prediction is illustrated in Figure 1.

2.2 Generating a gene chromatin network

The first task of our pipeline concerns the definition of a *Gene Chromatin Network* derived from a Hi-C experiment. We binned at 40 KB the Hi-C maps of all experiments and performed iterative correction normalization (Imakaev *et al.*, 2012). Then, for each gene, we extracted its transcription starting site (TSS) coordinates using ENSEMBL. We then associated to each TSS the 40 KB bin of the Hi-C contact matrix overlapping with it. Finally, we extracted a gene × gene matrix where each value (i, j) corresponds to the

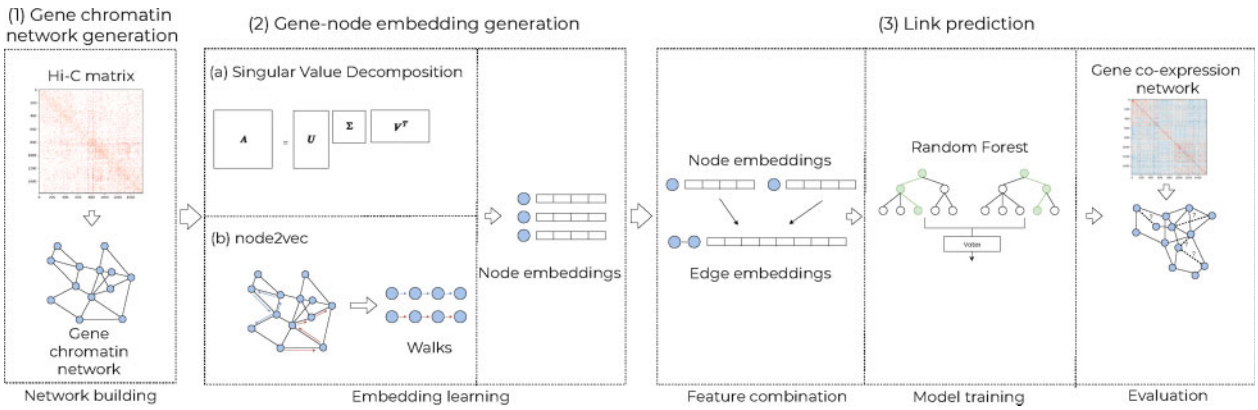

Fig. 1. Schematic representation of the proposed workflow. The pipeline is composed by a three main tasks. Initially, a Gene Chromatin Network is generated by summarizing Hi-C information of genes and their neighborhood for each gene, producing an interaction vector for each gene (1). Then, produce reduced vector representations of genes through network embedding techniques (matrix factorization or random walks) (2). The final step focuses on co-expression prediction, which is done by taking the combined pairs of gene vectors as input for a random forest classifier, trained on a subset of the gene co-expression network (3)

normalized number of contacts between the genomic bins associated with gene i and j; the result is therefore a contact map where each bin maps to at least one TSS. At the end of this procedure, we had a gene chromatin network for each Hi-C experiment.

Since Hi-C contacts are subject to several biases and noise (Yaffe and Tanay, 2011), we then applied a strong threshold on each gene–gene interaction, selecting only those whose number of reads was higher than the 80th percentile of all interactions across chromosomes, after removing the self-interactions along the main diagonal. It must be noted that the ICE normalization performed on the contact map does not normalize the interaction strength by the distance between genomic regions. We opted for this configuration in order to preserve also the local connectivity of the gene chromatin network so to exploit the knowledge of the linear neighbors for each gene during the learning of gene embeddings. We also analyzed the relationship between co-expression and genomic distance between pairs of genes in the same chromosome (Fig. 2) and noticed a dependency between neighboring genes. This finding, in concordance with previous studies (Soler-Oliva *et al.*, 2017), validated our choice to preserve neighboring interactions in the gene chromatin network.

The generation of the gene chromatin network is dependent on the coverage (i.e. the number of reads) of the Hi-C experiment. As a consequence, a proper bin size depends on data quality. Different bin sizes control the extension of the neighboring regions around the TSS of the genes in different ways, therefore capturing interaction patterns at different genomic scales.

We studied the relationship between gene co-expression networks and gene chromatin networks in three different setups, having independent feature generation and model training. We first defined an interaction network for each chromosome, therefore considering only intra-chromosomal interactions, and performed the analysis separately; then, we merged all the single-chromosome networks together, thus defining a unified network with 22 separated components; finally, we used also inter-chromosomal interactions to define a single gene interaction network where genes can potentially interact across different chromosomes.

2.3 Gene chromatin network embedding

Network embedding methods aim at learning, from the adjacency matrix $A \in \mathbb{R}^{n \times n}$ of n nodes, a low-dimensional representation $\mathbf{z} \in \mathbb{R}^k$ for each node of the network (Arsov and Mirceva, 2019; Hamilton *et al.*, 2017). The mapping follows the principle for which similar nodes in the network must have similar representation vectors. The different interpretations of the concept of node similarity pushed the generation of a multitude of network embedding methods. A vector representation enables to leverage the power and speed of the methods designed for data residing in vector spaces. Furthermore, the node embeddings may explicitly highlight functional and structural properties hidden in the network itself. Finally, the dimensionality reduction acts as a noise filter (Nelson *et al.*, 2019).

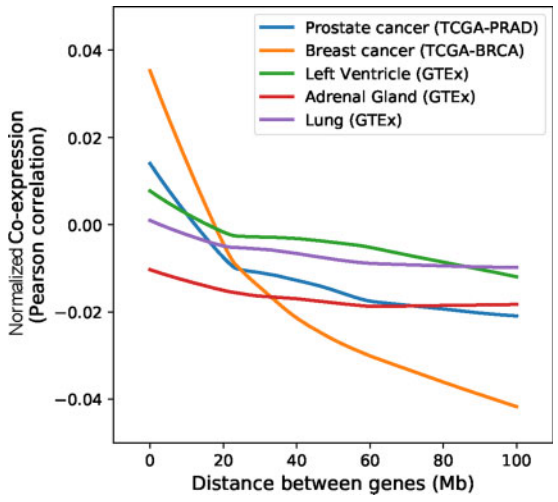

Fig. 2. Average normalized (mean-centered) co-expression between pairs of genes as a function of the distance between their TSS, calculated through lowess regression for five datasets from the studied compendium

Among the various techniques, we considered matrix factorization and random walk. Matrix factorization methods aim at reconstructing the original matrix through the multiplication of two or more small matrices, obtaining a low-rank space for the network; for this approach, we used singular value decomposition (SVD). Random walk-based methods aim at preserving the local structure of a node neighborhood in the transition from the very sparse and high-dimensional space of the adjacency matrix to the dense and low-dimensional space of the embedding; for this approach, we used node2vec. We then compared the co-expression prediction performances of these two methods against a set of baseline embedding methodologies:

- *Random predictor*: since our co-expression prediction task is binary with balanced classes, this corresponds to a 50% prediction accuracy.
- *Distance-based predictor*: gene co-expression can be influenced by the relative distance between genes, where nearby genes tend to have similar expression dynamics (Fig. 2). We assess this property by training a simple classifier using as feature only the genomic distance between pairs of genes.
- *Topological measures*: gene topology can be summarized by a set of engineered features extracted from the Hi-C gene network coming from the graph theory literature.

The topological measures used as baseline embedding are taken from Babaei *et al.* (2015). In their work, they built a feature vector for each pair of genes (g_i, g_j) composed of: shortest path between g_i and g_j, a Jaccard index indicating the proportion of shared connected genes between g_i and g_j and finally the average and absolute difference of degree (i.e. number of connections of a gene), betweenness (i.e. number of shortest paths passing through a gene) and clustering coefficient (i.e. number of connections between the direct neighbors of a gene) of g_i and g_j.

2.3.1 Matrix factorization

SVD is one of the most popular matrix factorization techniques. It factorizes an $m \times n$ matrix A into three distinct matrices, whose multiplication returns an approximation of A itself. These matrices represent the rows and columns of A in terms of a new, low-dimensional space of *latent factors*, therefore capturing high-level similarities between rows and between columns. Depending on the number of latent factors, the resulting matrix decomposition can encode a low or high amount of information.

More specifically, the (truncated) SVD factorizes A as follows:

$$A = U\Sigma V^T$$

where U is an $m \times d$ matrix of the rows expressed in terms of the d latent factors, V^T is a $d \times n$ matrix of the columns expressed in terms of the d latent factors and Σ is a $d \times d$ diagonal matrix of singular values, which are usually ordered by size as they express the importance of each latent factor. The number of latent factors (i.e. the embedding size) depends on the size of Σ.

In our case, the original matrix A is the $n \times n$ adjacency matrix of the gene chromatin network and, thus, rows and columns represent the same set of genes. U and V are of size $n \times d$ and both their rows can be considered as embeddings of the same genes. However, with the current setting, each gene has two embeddings, one from U and one from V. It is useful to obtain a single embedding vector, instead of two, for each gene by further decomposing Σ to express the SVD as binary matrix factorization.

$$A = (U\Sigma^{\frac{1}{2}})(\Sigma^{\frac{1}{2}}V^T) = WC^T.$$

Following Yue *et al.* (2019), the gene embeddings are then obtained by summation of the two factors:

$$E = W + C.$$

2.3.2 Node embedding through random walks

The node2vec algorithm (Grover and Leskovec, 2016) is a random walk-based method to learn the node embeddings of a network, following the principle that similar nodes have similar representations. Similarity in this context is interpreted in two ways:

- *homophily*: two nodes are similar if they belong to similar communities;
- *structural equivalence*: two nodes are similar if they have the same role in the network (e.g. they are both hubs).

The algorithm is able to control the relevance of the two similarities in the generation of the embeddings.

The node2vec algorithm is derived from the work on the Skip-gram model (Mikolov *et al.*, 2013), in which an embedding for a word is learned by predicting the nearby words in the text. The number of surrounding words considered is determined by the window hyper-parameter w. In the case of networks, there is no linear sequence of elements to learn from. For this reason, random walks have been introduced to generate, for each node, sequences of nodes representing its neighborhood.

The algorithm execution is divided into three phases: computation of the transition probabilities, random walk simulation and optimization.

First, it computes the probability of the random walk to transition from a node to another. Depending on the neighbors of a node, the two model's parameters p and q control the probability of transitioning to them, thus, controlling the sampling strategy for generating the random walk.

The sampling strategy for the neighborhood strongly influences the relevance of homophily and structural equivalence for the generation of the embedding. In particular, the higher the value of p, the less likely is the walk to revisit a node, reducing the locality of the generate neighborhood; q controls the balance between exploring inward nodes and outward nodes.

It is important to note that the two sampling methods are not mutually exclusive. The strategy can incorporate both the aspects with different degree depending on the values of p and q.

Then, the transition probabilities can be precomputed and used to simulate, starting from each node in the network, r random walks of fixed length l, resulting in a total of $n \times r$ random walks, where n is the number of nodes in the network.

Given the set of nodes in the network V, let $f : V \rightarrow \mathbb{R}^d$ be the function that takes a node and outputs its embedding, where d is the size of the embedding and let $N_S(u) \subset V$ be the neighborhood of a node u using the sampling strategy S.

The algorithm optimizes the following optimization function through stochastic gradient descent:

$$\max_f \sum_{u \in V} \log Pr(N_S(u)|f(u)).$$

At the end of the training of the Skip-gram model, the embeddings for each node are extracted as the values of the hidden layer associated to the node.

2.4 Co-expression network inference

The final step of our framework concerns the prediction of gene co-expression from the gene topological embeddings extracted from their physical interactions. This problem can be addressed by training a classifier to predict, given two genes and their vector representations, the existence of a link between them *in the co-expression network*. Several strategies are possible for inputing the two embeddings to a classifier, depending on how they are aggregated into a single vector representation, usually called *edge embedding*. In this work, we compared the following node to edge embedding transformations as in Grover and Leskovec (2016):

$$\text{Average} \qquad \frac{f_i(u) + f_i(v)}{2} \tag{1}$$

$$\text{Hadamard} \qquad f_i(u) * f_i(v) \tag{2}$$

$$\text{Weighted} - \text{L1} \qquad |f_i(u) - f_i(v)|, \tag{3}$$

where $f(u)$ and $f(v)$ are the embeddings, respectively, for node u and node v and $f_i(u)$ and $f_i(v)$ represent their ith component. Finally, we trained a random forest binary classifier on the edge embeddings using as label the presence or absence of links in the gene co-expression network.

2.5 Data preparation and preprocessing

To test the consistency of our approach, we performed our analysis on 12 different matched Hi-C and gene expression datasets. We collected Hi-C data from both tissues and cell-lines, healthy and tumor, and matched them with RNA-seq data coming from TCGA (Weinstein *et al.*, 2013) and GTEx (Lonsdale *et al.*, 2013). During the whole analysis and construction of the networks, we excluded the sex chromosomes and considered only the autosomes. In Table 1, we display the various matched experiments we considered in the following analysis.

Table 1. Source datasets used in this study together with their metadata

	Hi-C source	Hi-C reads (millions)	Hi-C type	N. genes	N. samples RNA-seq	N. edges gene co-expression network	N. edges gene chromatin network
Adrenal gland	Schmitt *et al.* (2016)	97.27	Tissue	20 705	264	1 190 506	290 370
Aorta	Schmitt *et al.* (2016)	347.67	Tissue	20 528	438	1 133 207	678 319
Breast cancer	Barutcu *et al.* (2015)	274.0	Cell line (MCF-7)	14 519	1224	609 747	150 762
Breast normal	Le Dily *et al.* (2019)	343.0	Cell line (MCF-10A)	21 353	465	1 285 926	462 855
Hippocampus	Schmitt *et al.* (2016)	103.38	Tissue	20 930	203	1 160 528	377 233
Left ventricle	Schmitt *et al.* (2016)	720.17	Tissue	19 011	438	994 113	433 202
Lung cell line	Rao *et al.* (2014)	1416.12	Cell line (IMR-90)	21 903	584	1 302 634	1 714 703
Lung rep. 1	Schmitt *et al.* (2016)	49.27	Tissue	21 903	584	1 302 634	188 992
Lung rep. 2	Schmitt *et al.* (2016)	70.71	Tissue	21 903	584	1 302 634	288 188
Pancreas rep. 1	Schmitt *et al.* (2016)	69.35	Tissue	20 235	334	1 118 191	354 610
Pancreas rep. 2	Schmitt *et al.* (2016)	46.67	Tissue	20 235	334	1 118 191	141 443
Prostate cancer	Rhie *et al.* (2019)	1000.0	Cell line (22Rv1)	14 643	556	583 446	657 946

2.5.1 Generating co-expression networks

We downloaded RNA-seq datasets from GTEx for normal tissues/cell-lines and from TCGA for tumor tissues/cell-lines; we used, as gene expression values, log2 transformed TPM estimates for the GTEx datasets and log2 transformed RSEM estimates for the TCGA datasets.

Every gene with more than 80% of the samples having expression =0 were excluded. Then, for each pair of genes, we computed the Pearson correlation coefficient of their samples, resulting in a gene×gene matrix where each value (i, j) correspond to the correlation of expression between gene i and gene j.

For both datasets, we evaluated three settings as previously described for the gene chromatin networks. We constructed the intra-chromosomal co-expression networks by applying a threshold, computed as the 90th percentile of correlation across the values of all the 22 co-expression matrices. Then, we generated a single network composed by 22 separated components by combining all the single-chromosome networks. Finally, we considered also inter-chromosomal co-expression relations. In all the cases, self-loops were removed from the networks.

2.5.2 Chromatin interaction data

We downloaded Hi-C data for each of the 12 datasets from the sources reported in Table 1. The data were binned at 40 KB resolution and normalized using the iterative correction method (Imakaev *et al.*, 2012). Notice that, the data from Schmitt *et al.* (2016) did not provide inter-chromosomal contacts, therefore, we could not evaluate the inter-chromosomal chromatin network embedding for them.

2.6 Model evaluation

We compared different embedding methods on the basis of the accuracy of the random forest classifier, which is defined as the fraction of correct predictions out of all the test samples.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}.$$

In order to train the classifier, we fed it with both positive (pairs of co-expressed genes) and negative samples (pairs of not co-expressed genes). To do so, we generated the negative samples by taking pairs of genes, which did not interact in the co-expression network. In addition, we constrained the search of negative pairs to only those whose genes, which were present in the positive set. To balance the training of the classifier, we generated the same number of negative samples as positive ones. All the results for each experimental setting were generated using 5-fold cross-validation.

To maintain balance of the positive and negative classes throughout the evaluation process, the training, validation and test contained a number of negative samples equal to the number of positive samples.

Finally, for each experimental setup, we initially removed 20% of the positive and negative co-expression links as holdout test set. This dataset was used as a final evaluation of our models and was not used during the hyper-parameter search and model definition.

3 Results

3.1 Intra-chromosomal co-expression prediction

In our first setting, we trained a distinct model for each non-sexual chromosome. Embeddings for different chromosomes were generated independently from each other, as well as the training of the classifiers. We set the node2vec hyper-parameters to the default values $p = 1$, $q = 1$, $r = 10$ and $l = 80$. The random forest classifier used 100 trees.

We then studied how the size of the embeddings influenced the prediction performances (Fig. 3, see Supplementary Fig.). We therefore looked at the prediction accuracy as a function of the embedding algorithm, the number of elements in the gene vectors and of the embedding aggregation strategy to build edge representations. We found that accuracy increases monotonically with the embedding size, demonstrating that bigger embeddings can capture more topological information, useful for the prediction task.

To see if this relationship is independent from the chosen edge classifier, we performed the previous analysis using a logistic regression classifier instead of random forest. Given the linear nature of Logistic Regression, we did not expect to match the prediction performances of random forest. On the other hand, the correlation between embedding size and accuracy was preserved. We noted that, the Hadamard product and the average of gene embeddings had the best performance with the random forest classifier in both the left ventricle and adrenal gland datasets.

Interestingly, when comparing node2vec with SVD, they produced slightly different outcomes as a function of embedding size. For small embedding sizes, SVD was more efficient in capturing information about the network topology, but for embedding sizes >8 node2vec achieved the best accuracy, in particular in the adrenal gland dataset. We then decided to use the Hadamard product and an embedding size of 16 elements in the following experiments and in the comparisons with the other baselines.

In all the considered datasets, our embedding strategies outperformed the topological measures and distance-based predictors (Figs 4 and 5). We did not find significant correlation with the Hi-C coverage or the number of gene expression samples used to build the co-expression networks, therefore showing the robustness of the approach. Additionally, we compared results of the two Hi-C replicates of lung and pancreas tissues finding similar results between replicates, highlighting the stability of the method to intrinsic experimental noise (not shown in the figure).

In general, node2vec outperformed SVD. In Figures 4 and 5, we show the box plots, based on a 5-fold cross-validation output for each of the 22 chromosomes, the former on healthy tissues and cell-lines, the latter on two cancer cell-lines.

Interestingly, a predictor purely based on the linear distance between the pair of genes along the chromosomes performs consistently better than the baseline random classifier. This confirms our previous findings (Fig. 2). On average topological measures are able to predict 61–63% of the co-expression links across chromosomes, while SVD and node2vec embeddings, respectively, recover 64–66% and 67%. It must be noticed that although the improvement over previous methodologies is consistent both between chromosomes and across datasets, chromatin data are not sufficient to completely explain co-expression. This is expected and more complex models taking into account transcription factor binding and histone modifications could improve the prediction accuracy. Being our work focused only on chromatin conformation, we deemed the integration of additional data sources out of scope.

3.2 Building a shared model for all chromosomes

We next considered the training on the whole set of chromosomes, therefore generating compatible sets of vector representations for the entire genome. In order to do that, we assembled a single network from each of the 22 single-chromosome networks. The resulting network was therefore composed of 22 independent components. The substantial difference from the previous experimental setup is that both the node embedding strategy and the edge classifier were trained on the entire set of genes and intra-chromosomal interactions. This enabled us to study the generalization capabilities of the models and prevent possible over-fitting issues, which could arise due to the size and gene density of specific chromosomes.

The hyper-parameters p, q, r, l, w of the node2vec algorithm were obtained through a hyper-parameter tuning based on the Bayesian optimization algorithm (Snoek *et al.*, 2012). For each dataset, the optimization process explored around 70 configurations of the hyper-parameters.

The same hyper-parameters were used during the test on the holdout dataset. The results of this test confirmed our previous findings, with the exception of the SVD embedding method, whose performance was significantly lower than in the previous setting, being outperformed also by simple network topology measures (Fig. 6). This is due to the intrinsic design of SVD, which reconstructs the

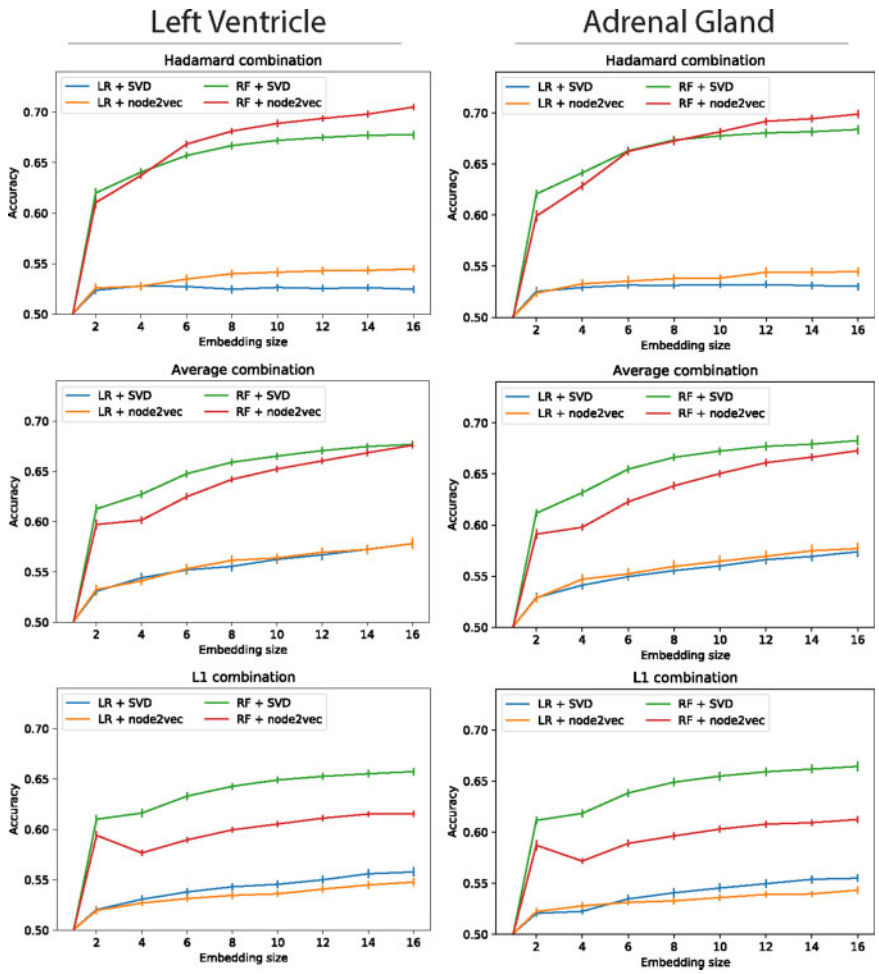

Fig. 3. Accuracy of co-expression prediction as a function of the embedding size in Left Ventricle (left) and Adrenal Gland (right) datasets. The other datasets display similar behavior and are omitted for brevity. We evaluated both SVD and node2vec embeddings using random forest (green and red lines) and logistic regression (blue and orange lines) classifiers. We also explored three different gene embedding combination functions: Hadamard product (top), element-wise average (center) and weighted L1 (bottom). See the Supplementary Figure for the analysis of all the studied datasets

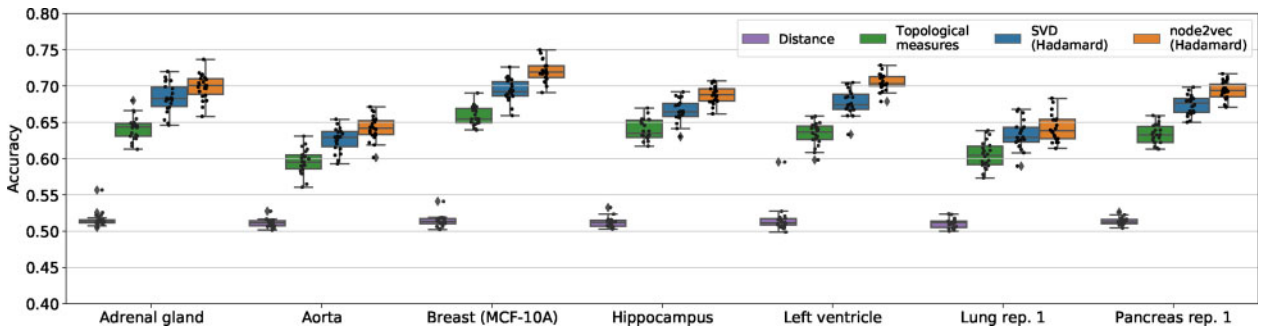

Fig. 4. The 5-fold cross-validation accuracy across the 22 single-chromosome networks of the proposed gene embedding strategies and comparison with baselines. Each box plot is derived from the accuracy measures for each cross-validation fold and for each chromosome. For each matched healthy Hi-C/gene expression dataset, we evaluated a *random predictor*, a pure *distance-based* predictor (purple), a random forests predictor based on manually derived *topological measures* from the gene chromatin network (green) and finally a random forest model leveraging our SVD (blue) and node2vec (orange) embeddings

original adjacency matrix taking into account also the negative interactions (the zeros of the matrix). But in this setting, the adjacency matrix is a block-diagonal matrix with all the inter-chromosomal interactions set to zero. The overwhelmingly sparsity of the matrix thus produces poor quality SVD embeddings. On the other hand, neighborhood-based methods like node2vec show similar results as before, outperforming all other baselines.

3.3 Including inter-chromosomal contacts to create a genome-wide chromatin network

We finally considered also inter-chromosomal contacts in our analysis. Therefore, we built a whole-genome gene chromatin network taking into account also the inter-chromosomal gene interaction profiles. Since the coverage of inter-chromosomal interactions is drastically lower than for intra-chromosomal, we decided to apply a

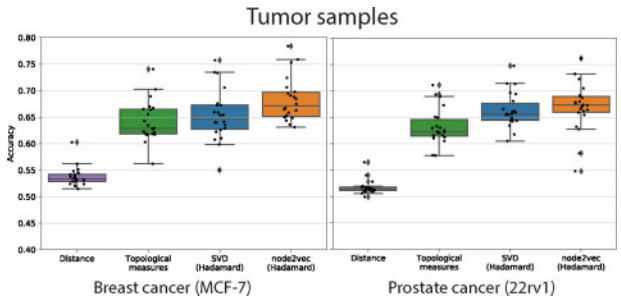

Fig. 5. Same as Figure 4, but focused on the two cancer cell-lines MCF-7 and 22Rv1

may decrease too much the importance of the intra-chromosomal interactions. Additionally, intra- and inter-chromosomal sub-networks differ topologically, since the former has a dependency on genomic distance, while the latter does not.

To prevent the noise coming from inter-chromosomal interactions to affect too much the learning of the node2vec embeddings, thus decreasing their information content with respect to the intra-chromosomal topology, we reduced the probability for a random walk to pass by an inter-chromosomal link to one-tenth of the original probability.

The best hyper-parameters of node2vec, according to the results of the Bayesian optimization, were $p = 1.5, q = 1.2, r = 70, l = 80$, $w = 12$, with a validation accuracy of 0.6922. The process explored 30 different configurations.

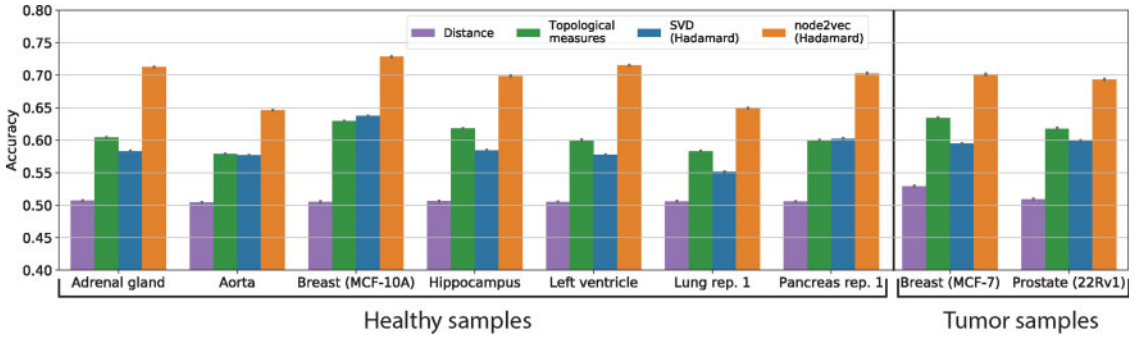

Fig. 6. The 5-fold cross-validation accuracy measures for the aggregated network derived by merging the 22 single-chromosome networks without considering inter-chromosomal interactions. Each bar represents the average accuracy across the 5-folds and the error bars represents the SD. We compared the performances of the same models as in Figure 4

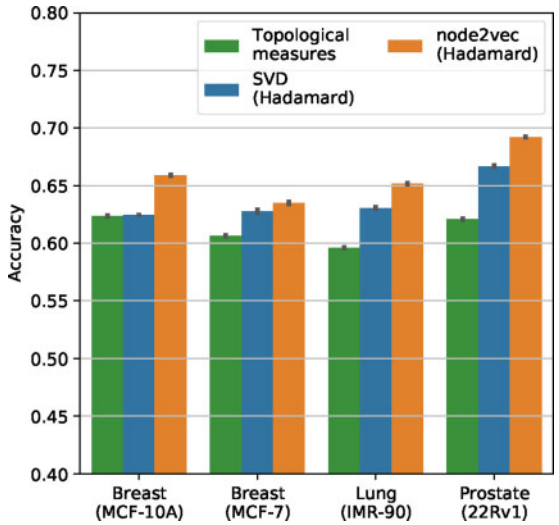

Fig. 7. The 5-fold cross-validation accuracy measures for the complete genome-wide networks derived from both intra- and inter-chromosomal contacts and co-expression links. Each bar represents the average accuracy across the 5-folds and the error bars represent the SDs

more compelling threshold to assign an edge to pairs of genes belonging to different chromosomes: only inter-chromosomal contacts above the 90th percentile were therefore considered. Since only few datasets of our compendium have enough inter-chromosomal reads to perform a meaningful analysis, this analysis was done only on the MCF-10A, MCF-7, IMR-90 and 22Rv1 cell-lines.

Due to their different biological nature, the simultaneous use of both intra- and inter-chromosomal interactions is a challenging task (Lajoie *et al.*, 2015): given their strong unbalance, an excessively sparse inter-chromosomal sub-network may nullify the value of creating a single genome-wide network, while an excessively dense one

For the training of the random forest classifier, we sampled a number of inter-chromosomal links from the whole-genome co-expression network equal to the total number of intra-chromosomal links. In this way, we prevented possible biases of model toward inter-chromosomal co-expression links, which greatly outnumber the intra-chromosomal ones in the gene co-expression network.

In Figure 7, we present the results of this analysis. Since, we tested links, which can span different chromosomes, we could not apply the distance-based predictor. The results confirmed our previous findings. Predictably, SVD embeddings encoded more relevant information than in the previous setup (see Fig. 6) thanks to the more homogeneous network obtained by adding inter-chromosomal contacts. Still, random walk-based approaches outperformed both matrix factorization and manually engineered topological features.

3.4 Evaluation on the holdout dataset

To further validate all our previous findings, we used the holdout dataset of co-expression links which we initially removed from our data collection. In Table 2, we summarize the accuracy of each model on the holdout dataset, showing that the accuracies of the holdout set are similar to those produced by the 5-fold cross-validation. Thus, the models exhibit high generalization power.

4 Discussion

Understanding the relationship between the spatial conformation of the genome and the regulation of gene expression is a fascinating biological question. We defined a general framework for gene co-expression prediction from chromatin conformation data extracted from Hi-C experiments. We first extracted a gene chromatin network by filtering Hi-C interactions at the level of transcription starting sites, therefore associating to each gene its physical interaction profile. We then studied which network embedding strategy best encodes topological information by comparing matrix factorization with a method based on random walks on the gene interaction graph. Finally, we used the gene embedding vectors, which were

Table 2. Accuracy of the models on the holdout test set on all the datasets in the only intra-chromosomal (a), shared intra-chromosomal (b) and intra + inter-chromosomal (c) setups

Dataset	Distance	Top. meas.	SVD	node2vec
Adrenal gland (a)	0.52±0.02	0.65±0.02	**0.69±0.02**	**0.71±0.01**
Adrenal gland (b)	0.51	0.62	0.59	**0.72**
Aorta (a)	0.51±0.01	0.61±0.02	**0.63±0.02**	**0.65±0.02**
Aorta (b)	0.50	0.59	0.58	**0.65**
Breast MCF7 (a)	0.54±0.02	0.65±0.04	**0.66±0.05**	**0.69±0.04**
Breast MCF7 (b)	0.53	0.64	0.60	**0.71**
Breast MCF7 (c)	×	0.61	**0.63**	**0.64**
Breast MCF10A (a)	0.52±0.01	0.67±0.02	**0.70±0.02**	**0.73 ±0.02**
Breast MCF10A (b)	0.51	0.64	**0.65**	**0.74**
Breast MCF10A (c)	×	0.63	**0.64**	**0.67**
Hippocampus (a)	0.51±0.01	0.65±0.02	**0.67±0.02**	**0.70±0.01**
Hippocampus (b)	0.51	0.63	0.59	**0.71**
Left ventricle (a)	0.51±0.01	0.65±0.02	**0.69±0.02**	**0.71±0.01**
Left ventricle (b)	0.51	0.61	0.58	**0.73**
Lung IMR90 (a)	0.51±0.01	0.63±0.02	**0.65±0.02**	**0.67±0.02**
Lung IMR90 (b)	0.51	0.61	0.59	**0.67**
Lung IMR90 (c)	×	0.60	**0.64**	**0.66**
Lung rep. 1 (a)	0.51±0.01	0.62±0.02	**0.64±0.02**	**0.65±0.02**
Lung rep. 1 (b)	0.51	0.59	0.56	**0.66**
Lung rep. 2 (a)	0.52±0.01	0.61±0.02	**0.64±0.02**	**0.65±0.02**
Lung rep. 2 (b)	0.51	0.60	0.57	**0.66**
Pancreas rep. 1 (a)	0.52±0.01	0.64±0.01	**0.68±0.02**	**0.70±0.01**
Pancreas rep. 1 (b)	0.51	0.61	0.61	**0.71**
Pancreas rep. 2 (a)	0.51±0.01	0.65±0.01	**0.68±0.02**	**0.69±0.01**
Pancreas rep. 2 (b)	0.51	0.61	0.57	**0.70**
Prostate 22Rv1 (a)	0.52±0.01	0.64±0.04	**0.67±0.04**	**0.69±0.03**
Prostate 22Rv1 (b)	0.51	0.62	0.60	**0.70**
Prostate 22Rv1 (c)	×	0.63	**0.67**	**0.70**

Note: For the setup (a), we show the values together with their SDs across chromosomes. In the other cases, we simply report the accuracy measure, since it is globally computed from the network. Best performances are shown in bold for each dataset.

learnt according to the two methods for training a binary classifier, based on random forest, to predict co-expression between pairs of genes.

Our results reveal that both matrix factorization and random walk strategies are effective in predicting co-expression between genes belonging to the same chromosome, outperforming simple topological measures and distance-based predictions in a wide set of tissues/cell-lines and healthy/tumor biological conditions. Our model is able to learn gene embeddings also by training on the union of all chromosome networks and by considering inter-chromosomal interactions, showing that the topological properties correlating with co-expression are shared across different chromosomes. Interestingly, we also discovered that a pure distance-based predictor could predict gene co-expression better than a random baseline, thus showing basic correlation relationships at the level of neighboring genes. Significantly, our results suggest that random walk-based models like node2vec outperform matrix factorization approaches, implying that the local topology of the interaction graph has greater predictive power than global topological representations.

The proposed methods can be used as an additional step in co-expression network inference together with more established tools based on gene expression data (Langfelder and Horvath, 2008; Zhang and Horvath, 2005), where chromatin data can be used to refine predictions based on expression correlation. This could enable the discovery of unknown regulatory interactions.

Our framework can be applied to other kinds of chromatin interaction data, like e.g. ChIA-PET or Promoter Capture HiC, thus deriving more specific embeddings for genes. However, the quality of the learnt embeddings is dependent on the size and completeness of the original interaction network, making the learning difficult in the case of very sparse datasets like ChIA-PET. Gene regulation is an extremely complex process, orchestrated and influenced by several biological mechanisms. For this reason, it is clear that for a better prediction of gene co-expression networks also different kind of input data sources must be considered, like the binding of transcription factors and histone modifications. In our study, we did not consider other information beside physical interactions to study how these two systems correlate, but future works could consider to use an integrated vector representation of genes aggregating heterogeneous data sources.

The proposed framework can be seen also as an instance of a more general approach, which summarizes complex interactions and relationships between biological entities (in our case, genes) in a dense vector format. The generality of this approach can be exploited in future work by designing more complete gene representations, taking into account also the binding of transcription factors around genes, the presence of histone modifications and of mutational events. Heterogeneous gene vector representations can be used as input for a clustering algorithm to extract regulatory modules, refining previous annotations and databases, which are usually built from a single data source. More generally, through the use of embeddings, spatial properties (like Hi-C contacts) can be referred either to genes or to genome bins, positioned upon specific genomic regions. In this way, embedding signals can be naturally composed, by means of integrative genomic data analysis languages and tools (e.g. Masseroli *et al.*, 2019; Nanni *et al.*, 2019), with other heterogeneous signals, including variants, gene expression, protein binding sites or methylation intensity, copy number alteration and so on; in a wider perspective, embeddings could be considered as annotations to be associated with specific loci.

Another important future improvement of our work regards the interpretation of gene embeddings extracted from their interaction profile. Given the recent attempts to reconcile gene expression and chromatin topology (Delaneau *et al.*, 2019) and the encouraging results of our work, we can affirm that the learnt embeddings encode relevant information about gene regulation. At the current state-of-the-art, there are no consolidated methods to explore node embeddings learnt through optimization, but the recent interest in these techniques is pushing the development of novel analysis toolboxes (Dalmia and Gupta, 2018). The future deep analysis of the topological embeddings extracted from Hi-C data could reveal important properties of genome folding and their relationship with the studied biological phenomenon, like gene co-expression.

Acknowledgements

We thank all the members of the Data-Driven Genomic Computing research group at Politecnico di Milano and of the Computational Systems Oncology group at University of Lausanne for their continuous support and useful discussion.

Funding

M.V., L.N. and S.C. are supported by the ERC Advanced Grant 693174 'Data-Driven Genomic Computing (GeCo)'.

Conflict of Interest: none declared.

References

Ala,U. *et al.* (2008) Prediction of human disease genes by human-mouse conserved coexpression analysis. *PLoS Comput. Biol.*, **4**, e1000043.

Arsov,N. and Mirceva,G. (2019) Network Embedding: An Overview. arXiv preprint arXiv:1911.11726.

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**, 214.

Bengio,Y. *et al.* (2012) Representation Learning: A Review and New Perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, **35**, 1798–1828. 10.1109/TPAMI.2013.50

Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.

Buttebutte,A.K. I. (2000) Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. Pac Symp Biocomput, **2000**, 418

Chou,W.-C. *et al.* (2014) Visual gene-network analysis reveals the cancer gene co-expression in human endometrial cancer. *BMC Genomics*, **15**, 300.

Dai,W. *et al.* (2015) Matrix factorization-based prediction of novel drug indications by integrating genomic space. *Comput. Math. Methods Med.*, **2015**, 1–9.

Dalmia,A. and Gupta,M. (2018) Towards interpretation of node embeddings. In:*Companion Proceedings of the Web Conference 2018*. pp. 945–952.

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

Delaneau,O. *et al.* (2019) Chromatin three-dimensional interactions mediate genetic effects on gene expression. *Science*, **364**, eaat8266.

Dixon,J.R. *et al.* (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**, 376–380.

Du,J. *et al.* (2019) Gene2vec: distributed representation of genes based on co-expression. *BMC Genomics*, **20**, 82.

Emrich,S.J. *et al.* (2007) Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Res.*, **17**, 69–73.

Gonzalez-Sandoval,A. and Gasser,S.M. (2016) On TADs and LADs: spatial control over gene expression. *Trends Genet.*, **32**, 485–495.

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*. pp. 855–864. ACM.

Hamilton,W.L. *et al.* (2017) Representation learning on graphs: methods and applications. *arXiv preprint arXiv: 1709.05584*.

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

International Human Genome Sequencing Consortium *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860.

International Human Genome Sequencing Consortium *et al.* (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931.

Kogelman,L.J. *et al.* (2014) Identification of co-expression gene networks, regulatory genes and pathways for obesity based on adipose tissue RNA sequencing in a porcine model. *BMC Med. Genomics*, **7**, 57.

Kustatscher,G. *et al.* (2017) Pervasive coexpression of spatially proximal genes is buffered at the protein level. *Mol. Syst. Biol.*, **13**, 937.

Lajoie,B.R. *et al.* (2015) The Hitchhiker's guide to Hi-C analysis: practical guidelines. *Methods*, **72**, 65–75.

Langfelder,P. and Horvath,S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, **9**, 559.

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

Lieberman-Aiden,E. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.

Lonsdale,J. *et al.* (2013) The genotype-tissue expression (GTEx) project. *Nat. Genet.*, **45**, 580–585.

Masseroli,M. *et al.* (2019) Processing of big heterogeneous genomic datasets for tertiary analysis of next generation sequencing data. *Bioinformatics*, **35**, 729–736.

Mikolov,T. *et al.* (2013) Efficient estimation of word representations in vector space. *arXiv preprint arXiv: 1301.3781*.

Nanni,L. *et al.* (2019) PyGMQL: scalable data extraction and analysis for heterogeneous genomic datasets. *BMC Bioinformatics*, **20**, 560.

Nelson,W. *et al.* (2019) To embed or not: network embedding as a paradigm in computational biology. *Front. Genet.*, **10**, 381.

Oh,E.-Y. *et al.* (2015) Extensive rewiring of epithelial-stromal co-expression networks in breast cancer. *Genome Biol.*, **16**, 128.

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.

Salzberg,S.L. (2018) Open questions: how many genes do we have? *BMC Biol.*, **16**, 94.

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.

Snoek,J. *et al.* (2012) Practical Bayesian optimization of machine learning algorithms. In:*Advances in Neural Information Processing Systems*. pp. 2951–2959.

Soler-Oliva,M.E. *et al.* (2017) Analysis of the relationship between coexpression domains and chromatin 3D organization. *PLoS Comput. Biol.*, **13**, e1005708.

Stuart,J.M. *et al.* (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science*, **302**, 249–255.

Szabo,Q. *et al.* (2019) Principles of genome folding into topologically associating domains. *Sci. Adv.*, **5**, eaaw1668.

Tian,D. *et al.* (2020) MOCHI enables discovery of heterogeneous interactome modules in 3D nucleome. Genome Research, **30**, 227–238. 10.1101/gr.250316.119

van Dam,S. *et al.* (2018) Gene co-expression analysis for functional classification and gene-disease predictions. *Brief. Bioinformatics*, **19**, 575–592.

Weinstein,J.N. *et al.*; The Cancer Genome Atlas Research Network. (2013) The cancer genome atlas pan-cancer analysis project. *Nat. Genet.*, **45**, 1113–1120.

Yaffe,E. and Tanay,A. (2011) Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat. Genet.*, **43**, 1059–1065.

Yang,Y. *et al.* (2014) Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. *Nat. Commun.*, **5**, 3231.

You,Z.-H. *et al.* (2017) An improved sequence-based prediction protocol for protein-protein interactions using amino acids substitution matrix and rotation forest ensemble classifiers. *Neurocomputing*, **228**, 277–282.

Yue,X. *et al.* (2019) Graph embedding on biomedical networks: methods, applications, and evaluations. *arXiv preprint arXiv: 1906.05017*.

Zhang,B. and Horvath,S. (2005) A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.*, **4**, 17.

Zhao,W. *et al.* (2010) Weighted gene coexpression network analysis: state of the art. *J. Biopharm. Stat.*, **20**, 281–300.

Zhou,N. *et al.* (2019) Hierarchical Markov Random Field model captures spatial dependency in gene expression, demonstrating regulation via the 3D genome. *bioRxiv*.