HIERARCHICAL MARKOV RANDOM FIELD MODEL CAPTURES SPATIAL DEPENDENCY IN GENE EXPRESSION, DEMONSTRATING REGULATION VIA THE 3D GENOME

A PREPRINT

Naihui Zhou
Bioinformatics and Computational Biology Graduate Program
Iowa State University
Ames, IA, USA
nzhou@iastate.edu

Iddo Friedberg
Department of Veterinary Microbiology and Preventive Medicine
Iowa State University
Ames, IA, USA
idoerg@iastate.edu

Mark S. Kaiser
Department of Statistics
Iowa State University
Ames, IA, USA
mskaiser@iastate.edu

December 16, 2019

ABSTRACT

HiC technology has revealed the eukaryotic genome's complex 3D architecture. It has been shown that the genome is separated into organizational structures which are associated with gene expression. However, to the best of our knowledge, no studies have quantitatively measured the level of gene expression in the context of the 3D genome.

Here we present a novel model that integrates data from RNA-seq and HiC experiments, and determines how much of the variation in gene expression can be accounted for by the genes' spatial locations. We used Poisson hierarchical Markov Random Field (PhiMRF), to estimate the level of spatial dependency among protein-coding genes in human IMR90 cells. The inference of PhiMRF follows a Bayesian framework, and we introduce the Spatial Interaction Estimate (SIE) to measure the strength of spatial dependency in gene expression.

We find that the quantitative expression of genes in human chromosomes 1, 4, 5, 6, 8, 9, 12, 19, 20 , 21 and X all show meaningful positive intra-chromosomal spatial dependency. Moreover, the spatial dependency is much stronger than the dependency based on linear gene neighborhoods, suggesting that 3D chromosome structures such as chromatin loops and Topologically Associating Domains (TADs) are indeed strongly correlated with gene expression levels. In some chromosomes the spatial dependency in gene expression is only detectable when the spatial neighborhoods are confined within TADs, suggesting TAD boundaries serve as insulating barriers for spatial interactions in the genome. We also report high inter-chromosomal spatial correlations in the majority of chromosome pairs, as well as the whole genome. Some functional groups of genes show strong spatial dependency in gene expression as well, providing new insights into the regulation mechanisms of these molecular functions. This study both confirms and quantifies the spatial correlation in gene expression. In addition, PhiMRF improves upon the stochastic modelling of gene expression that is currently widely used in differential expression analyses.

PhiMRF is available at: `https://github.com/ashleyzhou972/PhiMRF` as an R package.

***K*eywords** HiC · nuclear organization · gene expression · Markov random field · RNA-seq

1 Introduction

It has long been hypothesized that the 3D genome organization plays an important role in gene expression through various mechanisms[1, 2, 3, 4]. Of special interest is how genes in close spatial proximity coordinate expression. Several molecular models involving different organizational hierarchies have been proposed to explain this phenomenon [4]. One hypothesis is *transcription factories*, where RNA polymerase II is significantly enriched to allow efficient transcription of multiple genes at the same foci [5, 6, 7]. Even outside the context of transcription factories, several well-defined multi-gene complexes exist. The *Hox* gene cluster is a good example, where the genes dynamically organize into transcriptionally active compartment [8]. It has been shown that the TNFα-induced multigene complex regulated by NF-κB is disrupted once the chromatin loops for the complex are cleaved [9]. This effect might be explained by the fact that these genes are dependent on NF-κB-responsive enhancers [10]. Indeed, another molecular model for spatial gene clusters hypothesized that the spatial cluster are brought together to allow their promoters to interact with enhancers [11, 12]. Studies into promoter interaction networks suggest genes sharing common regulatory elements are found to be spatially co-localized and with correlated expression levels [13].

Many of the aforementioned studies are made possible following the advent of Chromosomal Conformation Capture (3C) [14] and subsequent 4C [8, 15], HiC [16, 10, 17] and Capture HiC [13, 12] technology. The new advances not only quantifies interaction frequency for any local genomic loci, but also allows for a global overview of the genomic architecture. HiC has enabled or confirmed discoveries of a hierarchy of organizational structures of the eukaryotic

genome, from Topologically Associating Domains (TAD) [18], to A/B compartments [16] and Chromosomal Territories (CT) [19]. General association between these structures and gene expression has been observed [3, 20].

While these hypotheses and models all propose that gene expression is indeed at least partially regulated by the spatial location of genes, they either lack global coverage, or do not directly quantify gene expression levels. Several whole-genome computational studies have attempted to untangle the relationship between gene co-expression and their spatial organization. One study found that inter-chromosomal co-expression is significantly enhanced for genes with spatial contacts in yeast [21]. Another study showed that the gene interaction network predicts co-expression [22]. One latest study also found that gene pair functional similarity are correlated with spatial distance [23]. However, these studies represented the expression levels of each gene as a pairwise property, either co-expression or co-functionality. Even though the data represent multiple experimental samples of RNA-seq data, the study subjects were no longer individual genes but gene pairs. One non-pairwise study has also looked at the relationship between functionality and 3D gene clusters, and found that functional (Protein-protein Interaction, complexes and pathways) gene groups are spatially concentrated using a permutation test based scheme [24]. In summary, these computational studies confirmed that there is a general trend of co-localization for co-regulated, co-expressed or co-functional genes. However, none of them provide a probabilistic model for gene expression levels within a genome.

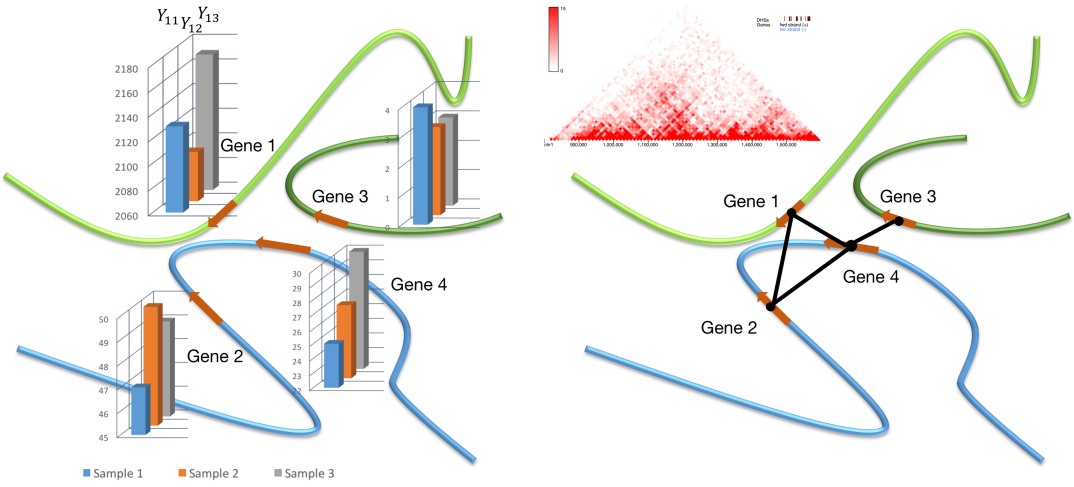
At the same time, many routine analyses, especially in RNA-seq data, rely heavily on a good probabilistic model of gene expression. Most common model for gene expression count is some kind of Poisson mixture model, including negative binomial [25, 26, 27, 28]. The variation they are modelling are the between sample variations to detect whether gene expression level for one particular gene has changed non-randomly between two or more experimental conditions. Some dispersion parameter or scaling factor usually exist for each sample, regardless of gene, to adjust for bias across biological replicates. Empirical Bayes models like limma [25] borrow information from the probabilistic space of the genome to inform the model for each gene. But this is not enough to inform the model of actual between-gene variation, which can be further modeled given more data. Indeed, as we have explained so far, gene expression levels between genes are regulated by a number of factors such as spatial location. So accounting for such correlation between the genes in one sample could improve the performance of Differential Expression analyses tools.

In this paper, we provide a probabilistic model dubbed PhiMRF (Poisson hierarchical Markov Random Field) for gene expression that takes into account the spatial location of these genes. We introduce the *Spatial Interaction Estimate* (SIE) whose value indicates the strength of the spatial dependency. Using PhiMRF, we not only confirm the current consensus that gene expression is positively spatially correlated, we also quantify the level of such correlation for the each chromosome, the whole genome as well as selected gene groups. This work will help molecular biologists in their quest to understand gene regulation and computational biologists to improve upon on current statistical models representing the gene expression landscape.

2 Results

2.1 Hierarchical Markov Random Field Model

The goal of our statistical model is to represent the probability space of gene expression count levels in the genome, while attributing part of the variance in this distribution to gene spatial location. Gene expression varies greatly for different genes in the genome in order to perform different functions for the cell. As shown in Figure 1, k samples of biological replicates of RNA-seq quantification for gene expression could be observed at each gene location. PhiMRF attempts to attribute the variation in gene expression to the genes' 3D locations using an autoregression-based model. Autoregressive (AR) models are traditionally used to explain variation for data observed in spatial or temporal settings [29] by modeling the distribution of an observation as dependent on its past (time series) or on its spatial neighbors. Limiting the scope of spatial neighbors to immediate neighbors, produces a Markov Random Field (MRF) model. We extract a neighborhood network for gene locations in 3D space from HiC data, where a gene is mapped to all of its overlapping loci (Figure S1), and its spatial proximity to another gene is calculated as the summary of all of the loci pairs that the two gene covers. An edge is inferred between them if the resulting proximity is greater than a threshold. These two genes are then considered as immediate neighbors. See Methods for details of the network inference algorithm.



(a) (b)

Figure 1: **Overall Scheme of the PhiMRF model applied to RNA-seq and HiC data. (a)** Replicates of RNA-seq quantification can be observed at each gene. **(b)** A spatial gene network is inferred from HiC data. Each gene is treated as a node in the network. An edge exists between two genes if the spatial interaction frequency between loci overlapping with the two genes is higher than a threshold. See Methods for detailed description of the network inference. The triangular HiC heat map is generated using the 3D genome browser [30] with data from Rao *et al* [17].

Our Poisson Hierarchical Markov Random Field (PhiMRF) model is briefly described below. Let Y_{ik} be the random variable connected with the RNA-seq count for gene i (located at location s_i) from sample k, $i = 1, 2, \ldots, n$;

$k = 1, 2, \ldots, M$. It is customary to model count Y_{ik} with a Poisson distribution [31], with it's parameter λ_i, i.e. $Y_{ik} \sim$ Poisson(λ_i). Let $w_i = log(\lambda_i)$. We **conditionally** specify the distribution for w_i as,

$$w_i | \boldsymbol{w}(N_i) \sim N(\mu_i, \tau^2) \tag{1}$$

where N_i is the set of locations neighboring s_i: $N_i = \{s_j : s_j \text{ is a neighbor of } s_i\}$ and $\boldsymbol{w}(N_i) = \{w_j : s_j \text{ is a neighbor of } s_i\}$.

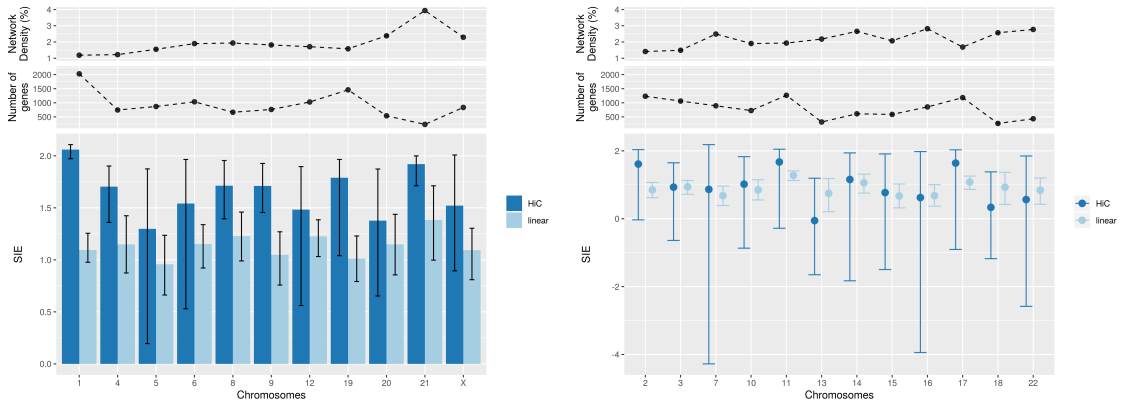
What sets Equation 1 apart from the Gaussian assumption for a simple linear regression is that the distribution for w_i is conditionally specified, as opposed to the marginally specified model, e.g. $Y_i \sim N(\mu_i, \tau^2)$. Moreover, in a simple linear regression, the next step of the model is to regress the Gaussian mean against independent variable X (e.g. $\mu_i = X_i\beta$), while in our *autoregressive* model, the Gaussian mean is regressed against its immediate neighbors $\boldsymbol{w}(N_i)$, shown as follows.

$$\mu_i = \alpha + \eta \sum_{j \in N_i} \frac{1}{|N_i| + |N_j|}(w_j - \alpha). \tag{2}$$

The parameter η here is analogous to the regression coefficient β in a simple linear regression, that symbolizes effect of the spatial autocorrelation. $\frac{1}{|N_i|+|N_j|}$ is a normalizing constant that adjusts the effect of different number of neighbors for different gene i's. Throughout this study, we will be looking at the posterior distribution of η to understand the strength of the spatial dependency present in our genome. The main properties of the posterior η distribution that we care about are the mean ($\hat{\eta}$) and the 95% credible interval, which is obtained as the 2.5% and 97.5% quantiles of the simulated posterior distribution. We will refer to the *estimate* of posterior η ($\hat{\eta}$) as the **Spatial Interaction Estimate (SIE)** hereinafter. If the 95% credible interval of posterior η does not contain 0, we say that there is meaningful spatial dependency, whereas if the interval contains zero, we conclude that meaningful spatial dependency is not detected. The two other unknown parameters in this model are α and τ^2, where α is a parameter analogous to the intercept in a simple linear regression model that is connected with a basal expression rate for all genes. τ^2 is the parameter for the conditional Gaussian variance, which accounts for any more remaining variance in gene expression within a sample, not explained by spatial locations. The same properties (mean, 2.5% and 97.5% quantiles) are used to summarize their posterior distributions. Note that both parameters are a rather crude oversimplification of the complicated mechanism of gene expression and could be further specified to account for other effects on gene expression. However, as we are focused on the effects of spatial location, this simplification is preferred. In summary, we have modelled gene expression counts with a conditional Poisson-lognormal mixture, where the Poisson parameter λ_i follows a conditional log-normal distribution. We then applied a autoregressive model for w_i with a parameter η that is connected with spatial dependency. We applied Bayesian inference that allowed us the simulate from the posterior distributions of α, η and τ^2, resulting in the Spatial Interaction Estimate (SIE) that symbolizes the strength of the spatial dependency. We provide a general purpose R package called PhiMRF for Poisson HIerarchical Markov Random Field models. The package is applicable to any count data collected in space, whether on a regular or irregular lattice.

2.2 Intra-chromosomal dependency

Within each chromosome. We ran PhiMRF on all genes in each of the 23 human chromosomes (Y chromosome excluded), with the spatial gene networks inferred from intra-chromosomal HiC data with 10kb resolution[17]. We also implemented a linear baseline for each chromosome. It takes the same gene expression data for each gene but the spatial network is simply inferred from genes within 10k baseparis of each other in the linear chromosome. By comparing our data with this baseline dataset, we can observe whether the long-range, non-linear interactions do play a role in gene expression. Strong evidence of positive spatial dependency is found in eleven chromosomes: 1, 4, 5, 6, 8, 9,



(a) (b)

Figure 2: **Spatial Interaction Estimate (SIE) for whole chromosomes.** Top panels depicts the network properties of each chromosome. Network density is defined as the percentage of actual edges versus number of possible edges. **(a)** Chromosomes with 95% credible interval above zero. Height of bar is SIE, and error bars are the 2.5% and 97.5% quantiles from the posterior distributions. **(b)** Chromosomes with 95% credible interval including 0. Error bars are the 2.5% and 97.5% quantiles from the posterior distributions.

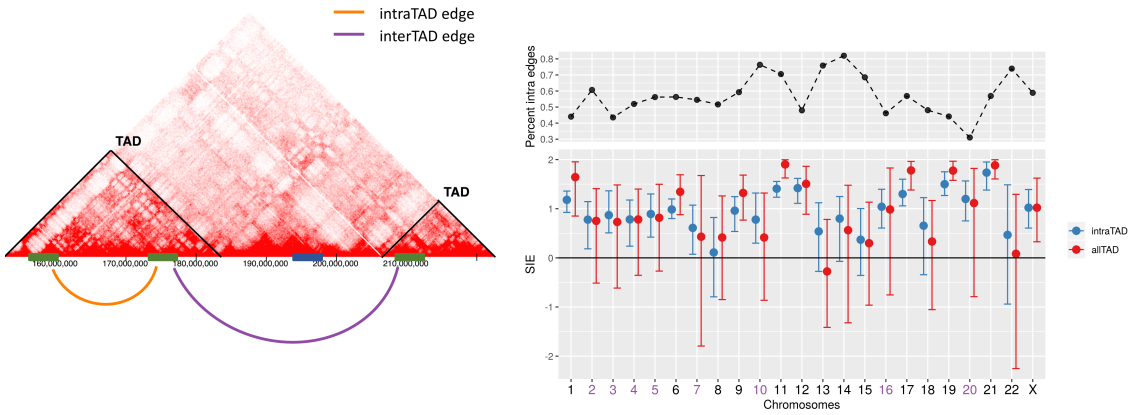
12, 19, 20, 21 and X, with SIEs higher than the linear baseline (Figure 2a). This suggested genes in these chromosomes are co-dependent on their neighbors when it comes to gene expression, proving that genes that are spatially close, not just on the linear chromosome, have coordinated expression patterns on a global scale, regardless of the function of the genes. We noticed that although in all these eleven chromosomes, the 95% credible interval of η is greater than zero, their sizes vary greatly. Although these SIEs are all higher than their linear counterpart, the intervals do not suggest that the difference between HiC and linear is statistically significant, except in Chromosome 1, 9 and 21. However, the differences in these SIEs should not be quantitatively compared (e.g. measuring the difference or ratio of these SIE's), as they all have different number of genes and connectivity. Top panel of Figure 2a shows the connectivity of each chromosome as measured by the number of genes in the network, and network density. There does not seem to be correlation between the parameter estimates from the PhiMRF model and the network size or connectivity.

Twelve chromosomes did not show meaningful spatial dependency in gene expression (Figure 2b). Despite only half of the chromosomes showing 3D spatial dependency, all 23 chromosomes show meaningful positive linear dependency in gene expression. In other words, the expression level of a gene is predictive of the expression levels of (at most) two other genes that are within 10k basepairs upstream or downstream from it. This is a strong confirmation for the efficacy

of our model as it suggests that the linear dependency in gene expression is stable and detectable. This is in line with our existing perception of the transcription mechanism.

Table S1 reports the full parameter estimates for basal expression rate (α) and conditional variance (τ^2). In general, all chromosomes have comparable basal expression rates, around 4, and conditional variance around 20. The conditional variance is indeed very large for a Poisson-lognormal mixture, which is indicative of the large variation in gene expression within a chromosome.

Within Topologically Associating Domains. Topologically Associating Domains (TAD) are megabase-sized spatial structures in the chromosomes observed from HiC data, manifested as significantly more frequent interactions within those domains than outside[18] (Figure 3a). It has been shown that enhancer-promoter interactions are constrained within TADs [10], and genes within TADs are more active in transcription than genes in TAD boundaries [18]. However, It is unclear how TAD structures causally affects gene expression levels, especially on a global scale[32, 33]. Here we investigate the level of gene expression spatial dependency for genes located in and out of TAD boundaries. Several algorithms are available for the computational identification of TADs based on HiC data with varying levels of accuracy[34], we chose Arrowhead[17] as the TAD caller. In total, 65.1% of genes are located within TAD boundaries out of all genes in the genomes, with Chromosome X having the lowest TAD gene percentage at 40.4%, and Chromosome 19 having the highest percentage at 73.1% (Table S3). We first tested the hypothesis that genes located within TADs have more connectivity than genes outside of TADs. Indeed, with a few exceptions, genes within TADs show higher clustering, as measured by the clustering coefficient [35], despite having lower network density (Figure S2). However, as to the maximum degree a gene can achieve, although one TAD gene in chromosome 11 is connected with 200 other genes, genes within TADs do not seem to exhibit extreme hub nodes versus non-TAD genes. We now turn



(a) (b)

Figure 3: **Spatial dependency in TAD gene expression. (a)** Illustration of intra-TAD edges. **(b)** SIE of TAD genes using intraTAD versus allTAD edges. Blue: only includes edges within one TAD (intraTAD). Green: includes all edges connecting all TAD genes (allTAD).

our attention exclusively to genes located within TADs.

We set out to isolate those edges that are located within TADs, i.e. interactions that connect two genes within the same TAD. First, we found that about half of the degree for genes located within TADs are intra-TAD edges (Figure 3b, top). Then we ran PhiMRF to detect spatial dependency of gene expression on these TAD genes, using only the edges within each individual TAD. Consequently, the network is essentially made up of a group of connected components that represent each TAD. Seventeen chromosomes, with the exception of 8, 13, 14, 15, 18, 22 show meaningful spatial dependency when only including intra-TAD edges, while only nine chromosomes show positive spatial dependency when including both intra-TAD and inter-TAD edges. Eight out of these nine chromosomes are part of the 11 chromosomes that show positive dependency as a whole in Figure 3b (bottom), when considering all genes and all edges in each chromosome. We also ran the model on TAD genes with only interTAD edges, to rule out the possibility that the difference in SIE is due the number of edges in the network. The dataset using interTAD edges display similar results as the dataset using all edges (Figure S3). We can therefore conclude that for some chromosomes, limiting the interactions to within each TAD results in a more detectable effect of spatial dependency. Therefore, it is reasonable to conclude that these intra-TAD interactions are more important than inter-TAD interactions when it comes to regulating coordinated gene expression levels. Such effects might be a result of TAD boundaries acting as insulating barriers.

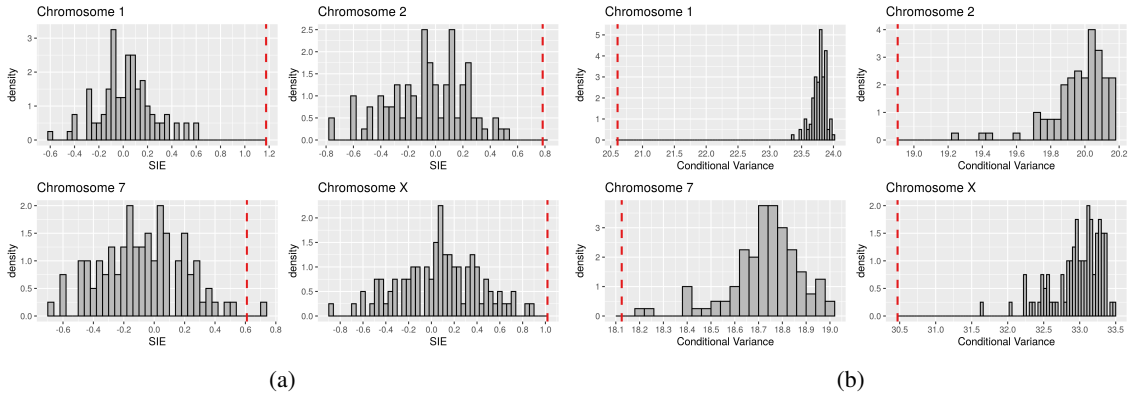


Figure 4: **Permutation test for spatial dependency in TAD gene expression.(a)** Histogram of SIE for 100 randomly permuted networks for all TAD genes in four chromosomes. Red dashed line is the observed SIE from the non-random HiC network. **(b)** Histogram of $\hat{\tau^2}$ for 100 randomly permuted networks for all TAD genes in four chromosomes. Red dashed line is the observed $\hat{\tau^2}$ from the non-random HiC network. The permutation is carried out using an Erdős-Rényi [36] algorithm with equal probability for any possible edge to be sampled. Total number of actual edges in each random graph is equal to the number of intra-TAD edges in the observed HiC network.

To completely rule out the size effect, and to further validate our model, in the next step, we carried out an *in silico* experiment to disrupt the TAD boundaries by randomly sampling edges in our HiC network. In other words, we randomly permuted the order of the edges in each HiC network to create a reference distribution. For each of these 100 random samples, edges are randomly designated between any two genes. The total number of random edges is equal to the number of intra-TAD edges. We then fit PhiMRF for all 100 networks and obtain the SIE's, as shown in Figures 4a, 4b and S4. For most chromosomes, the SIE from our observed model is significantly higher than our

reference distribution built from 100 randomly sampled networks. Moreover, our observed PhiMRF model often reports significantly lower posterior conditional variance when compared with randomly permuted networks, as some of the variance are accounted for by the spatial dependency, through the observed HiC network. The randomly permuted networks cannot account for the spatial dependency, hence having higher estimated variance. This serves as a validation that our model is picking up real spatial dependency signal instead of potentially picking up noise in the data. More importantly, the permutations could be viewed as an artificial disruption of TAD boundaries, connecting genes located in different TADs together. The results of such disruption demonstrated that TADs are indeed important structures in the non-random distribution of gene expression.

2.3 Inter-chromosomal dependency

Similar to intra-choromosomal HiC networks, we can derive gene networks from inter-chromosomal HiC. Many HiC studies are focused on intra-chromosomal interactions as chromatin looping and TADs are important mechanisms for gene regulation[10, 17]. Moreover, it was initially observed that different chromosomes tend to occupy different territories in the nucleus with rare inter-chromosomal interactions[37]. A single-cell HiC study has found that there are about 5-10% of chromosome intermingling[38]. However, it is suggested that the chromosome intermingling has a strong correlation with gene expression [39, 40]. We first set out to interrogate the relationship of inter-chromosomal HiC contacts and gene expression by fitting PhiMRF to all pairs of chromosomes. The inter-chromosomal HiC gene networks are derived the same way as the intra-chromosomal ones, but only using inter-chromosomal HiC interactions. In order to isolate the inter-chromosomal effect, even though for each dataset, genes from two chromosomes are present, no two genes from the same chromosome are connected. For a total of 253 pairs of chromosomes, only 32 (12.64%) pairs do not show meaningful positive spatial dependency. From Figure 5 we can observe that the lower the SIE, the larger the intervals, since spatial dependency is less and less detectable in those cases, and it manifested in both the effect sizes and the statistical significance. The chromosome pair with the highest SIE is Chromosome 9 and Chromosome 21, followed by Chromosome 9 and Chromosome 13. Some studies show that Chromosome 9 is located in the center of the nuclei [41], which may contribute to the high inter-chromosomal spatial dependency of its gene expression. Another chromosome that appeared twice in the top ten list is chromosome 19, which has been shown to locate in the center of the nuclei, and is very gene-rich[42]. Chromosome 1 is the only chromosome that appeared three times in the top ten list, which might be attributed to its exceptional length and gene count. We then incorporated all inter-chromosomal and intra-chromosomal HiC edges to all 19631 genes in the genome. Among the total 2,368,756 edges, 176,282 (7.44%)are intra-chromosomal edges. The PhiMRF shows that the global gene network has a SIE of 2.561(2.551, 2.570), confirming spatial dependency of gene expression as a global effect, observed in the whole genome. The global dataset has a basal expression rate of 3.039 (2.909, 3.176) and conditional variance of 22.337 (21.607, 23.227).

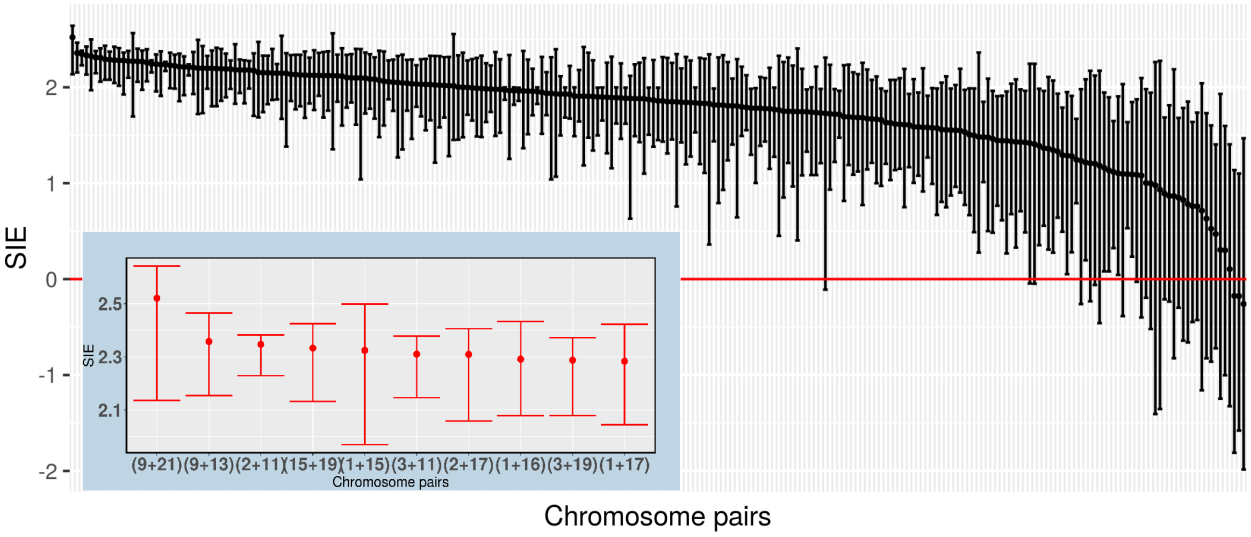


Figure 5: **SIE of chromosome pairs using only inter-chromosomal HiC interactions.** For each pair of chromosomes, PhiMRF ran on a dataset where all genes in both chromosomes are included, while only inter-chromosomal edges are included. From left to right, chromosome pairs ranked by highest SIE to lowest. **Background:** SIE and 95% credible interval of all 253 chromosome pairs. **Zoomed overlay:** SIE and 95% credible interval of the top ten chromosome pairs with the highest SIE.

2.4 Functional gene groups

We are interested in whether spatial dependency of gene expression is correlated with the function of the genes. In bacteria and archaea, some co-functioning genes are arranged in operons where their sequences are often transcribed into one mRNA transcript. In eukaryotes, genes are much more sparsely located in the linear genome, but evidence has been found to suggest that co-functioning genes also tend to cluster[24, 21]. The function of gene products is closely correlated with its expression levels, since co-functioning genes are often co-transcribed[43]. The latest Critical Assessment of Functional Annotation (CAFA3) reported that the best computational model in predicting the proteins associated with a particular function takes advantage of expression data [44]. While PhiMRF serves as a statistical model to understand the relationship between expression and spatial organization, it would be interesting to see whether such relationship is particularly strong for certain functional groups of genes, which could potentially explain the regulation mechanism for such functions.

We took advantage of the Gene Ontology (GO) consortium as the controlled vocabulary for functional annotations [45], where the functions of gene products in three aspects, biological process, molecular function and cellular component are described in a hierarchical way. We picked the top 15 GO terms in the Biological Process Ontology (BPO) that annotate the largest number of proteins in humans and collected all the genes associated with each GO term (Table S5). These terms include functions like transcription regulation, signal transduction, protein phosphorylation and so on. The size of the subsets of genes involved with each function varies in size, from 504 genes to 167 genes.

We found that the groups of genes associated with **positive/ negative regulation of transcription by RNA polymerase II** (GO:0045944, GO:0000122), **G protein-coupled receptor signaling pathway** (GO:0007186), **neutrophil degran-**

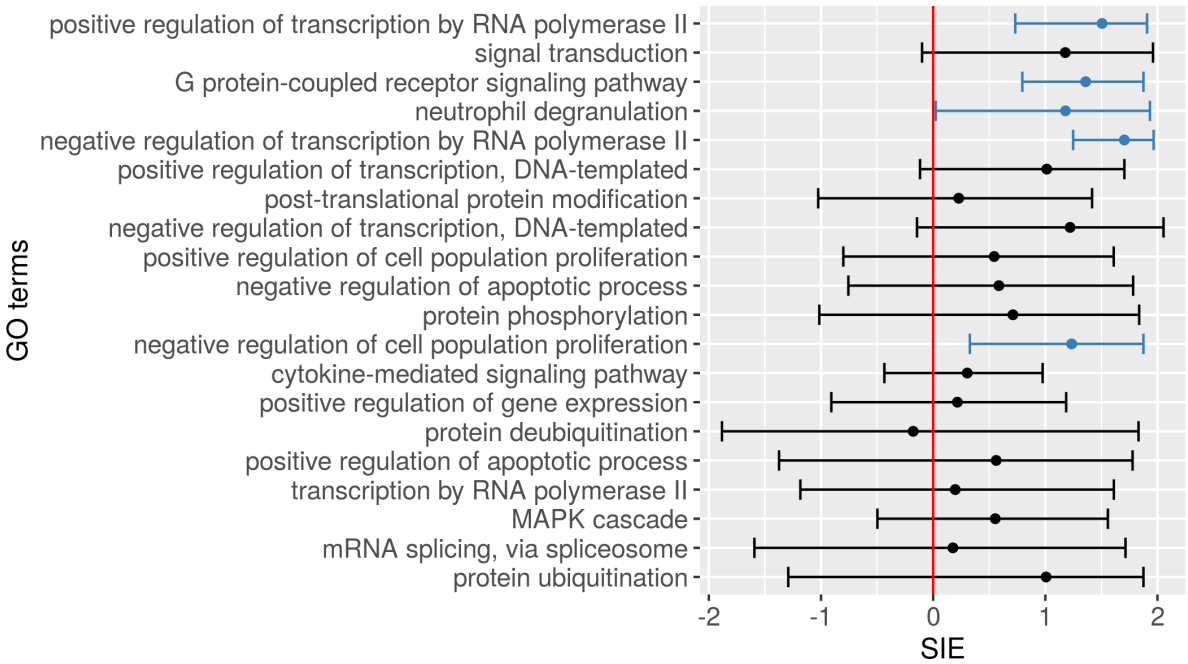


Figure 6: **SIE of functional gene groups.** SIEs and 95% credible intervals obtained for each group of genes associated with the top 20 GO BPO terms. For each group of genes, both intra-chromosomal and inter-chromosomal HiC interactions are included. The GO terms are ranked by the number of genes the annotate in decreasing order from top to bottom. Meaningful spatial dependencies are marked in blue.

ulation (GO:0043312), and **negative regulation of cell population proliferation** (GO:0008285) have statistically meaningful spatial dependency (Figure 6). The full set of parameter estimates can be found in Table S6.

The interpretation of these positive results is the interplay of the three important factors of a gene: its spatial location, its expression level and its function. A positive SIE for a functional gene group means that genes that are part of this particular biological process have co-dependent expression levels based on their spatial locations. This further means that in order for these genes to carry out the function in a coordinated fashion, regulation of the expression of these genes takes advantage of the 3D genome structure. For example, in GPCR signalling pathways, it is not surprising that we observe such an effect as the function involves protein complexes, hence having coordinated expression. What's surprising is the spatial aspect, where the gene locations can explain why these genes have the coordinated expression, hinting that the 3D genome is part of the regulation mechanism.

3 Methods

The gene expression data used in this study comes from the ENCODE project [46] (`https://www.encodeproject.org/`) with the following identifiers, ENCFF353SBP and ENCFF496RIW. These are two biological replicates of a total RNA-seq experiment on the IMR90 cell. Genes are mapped to Ensembl stable IDs with coordinates from Ensembl release 90 [47], which uses the GRCh38.p10 human genome assembly.

3.1 Bayesian Inference

The goal of our Bayesian framework is to simulate from the posterior distributions $p(\alpha|\boldsymbol{y})$, $p(\eta|\boldsymbol{y})$ and $p(\tau^2|\boldsymbol{y})$. The properties of these distributions directly answer the biological questions from which we abstracted the stochastic model. The overall strategy is to simulate from the joint posterior distribution of $p(\alpha, \eta, \tau^2, \boldsymbol{w}|\boldsymbol{y})$ using a Gibbs sampler[48], where we sequentially simulate from each of the full conditional posteriors of our parameters as follows.

$$
\begin{aligned}
p(\alpha|\boldsymbol{w}, \eta, \tau^2, \boldsymbol{y}) &\propto \pi(\alpha) g(\boldsymbol{w}, \alpha, \eta, \tau^2), \\
p(\eta|\boldsymbol{w}, \alpha, \tau^2, \boldsymbol{y}) &\propto \pi(\eta) g(\boldsymbol{w}, \alpha, \eta, \tau^2), \\
p(\tau^2|\boldsymbol{w}, \alpha, \eta, \boldsymbol{y}) &\propto \pi(\tau^2) g(\boldsymbol{w}, \alpha, \eta, \tau^2), \\
p(\boldsymbol{w}|\alpha, \eta, \tau^2, \boldsymbol{y}) &\propto g(\boldsymbol{w}|\alpha, \eta, \tau^2) f(\boldsymbol{y}|\boldsymbol{w}),
\end{aligned}
$$

where $\pi(\alpha)$, $\pi(\eta)$ and $\pi(\tau^2)$ are Uniform prior distributions, $f(\boldsymbol{y}|\boldsymbol{w})$ is the Poisson distribution for observed data $\boldsymbol{y}$, and $g(\boldsymbol{w}|\alpha, \eta, \tau^2)$ is the marginal distribution for $\boldsymbol{w}$. This marginal distribution is not readily available from our model specification, since $\boldsymbol{w}$ is only conditionally specified. However, it can be constructed from the conditional distributions using a negpotential function [49]. See Section 1 in the Supplemental Materials for details. Moreover, the negpotential introduces an intractable constant to the posterior that cannot be dropped, so we use the double Metropolis-Hastings algorithm [50] to simulate from these posteriors (Section 2 in Supplemental Materials).

Prior distributions The prior used for α is a Uniform distribution $U(-10, 10)$. The prior used for η is a Uniform distribution over the parameter space of η. The parameter space of η is directly calculated from the neighborhood adjacency matrix, as the inverse maximum and minimum of its eigenvalues[29]. See Section 3 in Supplemental Materials for details. We assume that τ follows a prior uniform distribution $U(0, 10)$ and derive the prior distribution for τ^2 in the model as $\pi(\tau^2) = \frac{1}{20\tau}\mathbb{1}(0, 100)$.

Iterations For each intra-chromosomal (including the linear baseline) and pairwise inter-chromosomal HiC dataset, we ran the double Metropolis-Hastings algorithm through 5000 iterations, with 1000 burn-in iterations. All TAD related datasets, the whole-genome dataset and the functional datasets went through 2000 iterations with 400 burn-in. The variances for the jump proposal distributions (See Section 4 in Supplemental Materials for details) were chosen through multiple rounds of initial testing to ensure that the jump frequencies fall within 15% to 40% for randomly selected datasets in each group of datasets, and that the MCMC simulations converge within the number of iterations used.

3.2 HiC Data processing

Normalization Raw observed HiC data for the lung IMR90 cell from Rao *et al*[17] are used, with 10kb resolution. The KR normalization technique was applied. The goal of the normalization is to remove one-dimensional bias in HiC counts. On Chromosome 9 at this resolution, the KR algorithm did not converge on that particular matrix, this is likely

due to sparsity of the matrix[17]. In this case raw counts are used. T Interactions between the same locus are eliminated to avoid bias towards neighboring genes (see below for how the loci are mapped to genes)

Gene Mapping Interactions are observed for every consecutive 10k bins (loci) on the entire human genome (except chromosome Y), while expression are generally considered for genes located intermittently on the genome. Because each bin is 10k in size, one gene is often mapped to multiple bins. The interaction between two genes is then decided by all of the interacting bins that overlap with the pair of genes. An overlap is when a bin shares more than 10% of base pairs with a gene, as shown in Figure S1. If gene 1 overlaps with bins A and bin B, while gene 2 overlaps with bins C, D and E, then the interaction between gene 1 and gene 2 are considered as the pool of interactions A-C, A-D, A-E, B-C, B-D and B-E. The goal of such gene-bins mapping is to inform a gene network where edges connect gene pairs that are close in 3D space while eliminating bias from particular loci. We found that most gene pairs overlap with a pool of less than five bins pairs, for both intra-chromosomal and inter-chromosomal gene pairs. Therefore, we adopted simple metrics to summarize these pools of interactions instead of using more complicated parametrizations (e.g. a t test). Four metrics are considered, mean, median, max and min to summarize this pool of interactions. The computed metric for each pair of genes is then compared to a threshold (described in the next section) to decided whether two genes are neighbors.

Soft threshold The mean (median, max or min) of the pool of interactions for a gene pair is compared with a threshold to determine whether there is an edge between the gene pair. The higher the interaction score, the closer in proximity. The threshold is determined as a 90% quantile of *all* locus-locus interactions. Therefore, the threshold changes from chromosome to chromosome, eliminating chromosomal bias. Naturally, out of the four metrics, the min metric is the most conservative, as it requires that all interactions in the pool for that gene pair be larger than the threshold, to consider the gene pair to be neighbors, while the max metric only requires that one interaction out of the pool be larger than the threshold. Mean is our main metric, and all subsequent studies presented in the main text uses the mean metric. Intra-chromosomal datasets are repeated using the other three metrics (median, min and max) as well and presented in Tables S8, S9 and S10.

3.3 Functional Annotations

We used the UniProt Gene Ontology Annotation (downloaded on September 16, 2019) [51] to extract all Gene Ontology BPO terms annotating human gene products. We only extracted the annotations with experimental evidence, with the following evidence codes: EXP, IDA, IPI, IMP, IGI, IEP, TAS and IC. We decided to *not* propagate the annotations through the hierarchical structure of the ontology. An example of a propagation is where the ontology structure specifies that **carbohydrate phosphorylation** (GO:0046835) "is a" **phosphorylation** (GO:0016310). If a protein is annotated with the former, a more informative term, then it is automatically annotated with the latter, a less informative term. Since we are ranking the GO terms by the number of proteins they annotate, if the annotations are propagated, the top ones will be the very general functions like "cellular process", that are in general not of interest to researchers. Therefore,

we have elected to rank the GO terms by the number of proteins they annotate without propagating them. These leaf annotations are directly curated from published experiments, which is evidence that these terms are of interest to some researchers. Right column of Table S5 shows the top 15 GO terms by count after propagation.

4 Discussion

We have developed a novel probabilistic model for gene expression incorporating spatial dependency. By applying this model to RNA-seq and HiC data, we get our first peek into how the transcriptome is shaped by the 3D location of the genes. The different expression levels of genes can be partially explained by the spatial location of the genes. The quantitative measurement of such effects provides unprecedented insight into the relationship and interplay among chromosomal organization, gene expression and functionality.

Insulating effects of TADs It has been hypothesized that TADs to play a role in the regulation of gene expression [18]. For example, the expression profiles of genes whose promoters are located within the same TAD are more correlated during cell differentiation than those of genes not in the same TAD[40]. One hypothesis of that the the regulation is weakened by the insulating effects of TAD boundaries. TAD boundaries are enriched with CTCF proteins and CTCF binding sites, which are known to help shape the structure of the genome [18]. Our study confirms that the expression of intra-TAD genes shows spatial dependency on a global scale, strengthening the hypothesis that TADs are an integral part of the gene expression regulation mechanism. Interestingly, since we are able to examine the 3D spatial interactions within each TAD, we discovered that for eight chromosomes, spatial dependency is only detectable when the interactions are confined to TADs (Figure 3b). If we include only interTAD edges (interactions between two genes on two different TADs), we observe meaningful spatial dependency only in five chromosomes, compared with seventeen chromosomes for intra-TAD edges (Figure S3). All edges present in the HiC gene network represent spatial proximity of these genes in the 3D space. However, if some of these edges are insulated or blocked by boundary elements such as CTCF proteins, then it makes sense that the spatial dependency of gene expression is no longer detectable when we include these edges. Even though genes connected by these interTAD edges are still spatially close, their expression are no longer coordinated due to the insulating effect of TAD boundaries.

High inter-chromosomal dependency The rapid pace of development of Chromosome Conformation Capture (3C) technology enables us to obtain detailed understanding of intra-chromosomal architecture that shapes the regulatory landscape of the genome, but inter-chromosomal interactions and their functions are studied less and still poorly understood [52]. Many known inter-chromosomal interactions are not detected by HiC experiments, although they can be as stable as intra-chromosomal contacts [53]. Maass *et al* reasoned that the lack of significant detection of the inter-chromosomal interactions is due to the different distance scale of the inter– versus the intra-chromosomal ones [53]. We acknowledge the bias that the HiC technique has against inter-chromosomal interactions. To mitigate this bias, we used a high resolution (10kb) HiC interaction map, together with the soft thresholding of HiC interactions. The high resolution map enabled us to pool information from several loci for one gene, while the soft thresholding allowed

us to infer a high interaction frequency based on individual chromosome pairs and not any absolute distance across the genome. However, even with such measures, network density is still lower for gene networks of inter-chromosomal pairs than for intra-chromosomal gene networks (Figure S5), demonstrating the need for caution when using and interpreting HiC data. Despite the relative low network density for inter-chromosomal HiC gene networks, we were still able to observe high inter-chromosomal spatial dependency in gene expression for the majority (87.35%) of chromosomal pairs (Figure 5). These results strongly suggest extremely long-range and intra-chromosomal gene interactions on different chromosomes as a commonly occurring regulation mechanism for gene expression. Therefore, in addition to the more commonly studied intra-chromosomal regulation mechanism, we advocate more attention to be paid to studying inter-chromosomal gene regulation.

Application to genomes and gene groups PhiMRF can also be applied to RNA-seq and HiC data from non-human genomes. HiC data in this model are represented as a spatial network, where two genes are connected if they are close in physical space. Even without HiC data, as long as any distance between each pair of genes could be represented in a network, our model is still applicable. The distance does not necessarily have to be from 3D physical space, but could be from any other types of pairwise gene-gene interactions. The interpretation of such results depends on the specific type of gene-gene interaction used as the underlying network.

We also demonstrated that PhiMRF model can be used to interrogate gene groups that are of particular interest; for example, genes that are annotated with the same function. Several functional gene groups show meaningful spatial dependency (Figure 6). PhiMRF can be further applied to gene groups that are functionally enriched, genes in the same biological pathway, genes that are causal for a certain disease or phenotype, and so on. In doing so, we are essentially looking at the regulation mechanism for such gene groups to perform a function or set of functions. A meaningful spatial dependency would indicate that the regulation of such function or disease involves the 3D genome architecture as one of the regulation mechanisms, allowing biologists to explore new directions when studying the functions, pathways or diseases of interest.

Co-expression and extending PhiMRF It is important to note that PhiMRF is not modelling co-expression, since it does not account for the variability in gene expression that comes from multiple experimental conditions or samples. PhiMRF is an autoregressive model that could take one or more observations at each location. Multiple observations at one location are considered conditionally independent of each other, while co-expression (and differential expression) analyses focus on the probability space for one gene among multiple samples. Although what we find can be described as genes having *coordinated* expression based on spatial proximity, the term ”co-expression” is not applicable at this stage. However, PhiMRF can be useful in co-expression (differential expression) analyses by accounting for the spatial effect in the between-sample variation.

PhiMRF also does not account for other effects that may be at play in gene expression. The basal expression level parameter α, and the variance parameter τ^2 are assumed to be the same for all genes in the genome. This assumption is crude and does not reflect the nature of the transcriptome. For example, assigning a different α parameter to

housekeeping and non-housekeeping genes could potentially greatly reduce variation in the model since those two groups are known to have very different expression levels. Our argument for the simple model specification is that it is hard to separate these other effects from the spatial effects. At the same time, the spatial effect (close proximity of certain genes) might as well be the result of some other molecular mechanisms that also modify the gene expression profile, e.g. enhancer-promoter interactions at hub-enhancers[54]. It is not our goal to isolate the spatial effect from the other effects, rather we use the spatial effect as a proxy for other mechanisms that might be at play.

Producing valid posteriors It is challenging to reproduce MCMC based results as doing so involves specific tuning of an algorithm. Our model is also restrictive in the parameter space of η, where it is dictated by the adjacency matrix structure and not a fixed value. PhiMRF should not be used on small datasets. (< 100 nodes). It is always good practice to check the convergence of the MCMC, and tune the parameters so that the jump rates of the MCMC fall within a reasonable range.

In summary, we have developed a hierarchical Markov random field (PhiMRF) model to explain the variation in gene expression. We took advantage of the recent HiC data to construct a spatial gene network. We found that for certain whole chromosomes, TADs and functional gene groups, there exists meaningful spatial dependency for gene expression levels. We also found this effect for the whole human genome, accounting for both intra-chromosomal and inter-chromosomal spatial interactions.

5 Availability

The code for PhiMRF is available here, `https://github.com/ashleyzhou972/PhiMRF` under GPL-2.

The code for data processing is available at `https://github.com/ashleyzhou972/bioMRF`.

Full intermediate and final results are available at `https://doi.org/10.6084/m9.figshare.11357321.v3`.[55].

References

[1] Christian Lanctôt, Thierry Cheutin, Marion Cremer, Giacomo Cavalli, and Thomas Cremer. Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions. *Nature Reviews Genetics*, 8(2):104, 2007.

[2] Robert Schneider and Rudolf Grosschedl. Dynamics and interplay of nuclear architecture, genome organization, and gene expression. *Genes & development*, 21(23):3027–3043, 2007.

[3] Boyan Bonev and Giacomo Cavalli. Organization and function of the 3d genome. *Nature Reviews Genetics*, 17(11):661, 2016.

[4] Anton Krumm and Zhijun Duan. Understanding the 3d genome: Emerging impacts on human disease. In *Seminars in cell & developmental biology*, volume 90, pages 62–77. Elsevier, 2019.

[5] Konstantinos Sofiadis and Argyris Papantonis. Transcription factories as spatial and functional organization nodes. In *Nuclear Architecture and Dynamics*, pages 283–296. Elsevier, 2018.

[6] Cameron S Osborne, Lyubomira Chakalova, Karen E Brown, David Carter, Alice Horton, Emmanuel Debrand, Beatriz Goyenechea, Jennifer A Mitchell, Susana Lopes, Wolf Reik, et al. Active genes dynamically colocalize to shared sites of ongoing transcription. *Nature genetics*, 36(10):1065, 2004.

[7] Peter R Cook. A model for all genomes: the role of transcription factories. *Journal of molecular biology*, 395(1):1–10, 2010.

[8] Daan Noordermeer, Marion Leleu, Erik Splinter, Jacques Rougemont, Wouter De Laat, and Denis Duboule. The dynamic architecture of hox gene clusters. *Science*, 334(6053):222–225, 2011.

[9] Stephanie Fanucchi, Youtaro Shibayama, Shaun Burd, Marc S Weinberg, and Musa M Mhlanga. Chromosomal contact permits transcription between coregulated genes. *Cell*, 155(3):606–620, 2013.

[10] Fulai Jin, Yan Li, Jesse R Dixon, Siddarth Selvaraj, Zhen Ye, Ah Young Lee, Chia-An Yen, Anthony D Schmitt, Celso A Espinoza, and Bing Ren. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature*, 503(7475):290, 2013.

[11] Robert A Beagrie, Antonio Scialdone, Markus Schueler, Dorothee CA Kraemer, Mita Chotalia, Sheila Q Xie, Mariano Barbieri, Inês de Santiago, Liron-Mark Lavitas, Miguel R Branco, et al. Complex multi-enhancer contacts captured by genome architecture mapping. *Nature*, 543(7646):519, 2017.

[12] Biola M Javierre, Oliver S Burren, Steven P Wilder, Roman Kreuzhuber, Steven M Hill, Sven Sewitz, Jonathan Cairns, Steven W Wingett, Csilla Várnai, Michiel J Thiecke, et al. Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. *Cell*, 167(5):1369–1384, 2016.

[13] Stefan Schoenfelder, Mayra Furlan-Magaril, Borbala Mifsud, Filipe Tavares-Cadete, Robert Sugar, Biola-Maria Javierre, Takashi Nagano, Yulia Katsman, Moorthy Sakthidevi, Steven W Wingett, et al. The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome research*, 25(4):582–597, 2015.

[14] Job Dekker, Karsten Rippe, Martijn Dekker, and Nancy Kleckner. Capturing chromosome conformation. *science*, 295(5558):1306–1311, 2002.

[15] Marieke Simonis, Petra Klous, Erik Splinter, Yuri Moshkin, Rob Willemsen, Elzo De Wit, Bas Van Steensel, and Wouter De Laat. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4c). *Nature genetics*, 38(11):1348, 2006.

[16] Erez Lieberman-Aiden, Nynke L Van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science*, 326(5950):289–293, 2009.

[17] Suhas SP Rao, Miriam H Huntley, Neva C Durand, Elena K Stamenova, Ivan D Bochkov, James T Robinson, Adrian L Sanborn, Ido Machol, Arina D Omer, Eric S Lander, et al. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680, 2014.

[18] Jesse R Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S Liu, and Bing Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376, 2012.

[19] Thomas Cremer and Marion Cremer. Chromosome territories. *Cold Spring Harbor perspectives in biology*, 2(3):a003889, 2010.

[20] Przemyslaw Szalaj and Dariusz Plewczynski. Three-dimensional organization and dynamics of the genome. *Cell biology and toxicology*, 34(5):381–404, 2018.

[21] Dirar Homouz and Andrzej S Kudlicki. The 3d organization of the yeast genome correlates with co-expression and reflects functional relations between genes. *PloS one*, 8(1):e54699, 2013.

[22] Sepideh Babaei, Ahmed Mahfouz, Marc Hulsman, Boudewijn PF Lelieveldt, Jeroen de Ridder, and Marcel Reinders. Hi-c chromatin interaction networks predict co-expression in the mouse cortex. *PLoS computational biology*, 11(5):e1004221, 2015.

[23] Li Liu, Qian-Zhong Li, Wen Jin, Hao Lv, and Hao Lin. Revealing gene function and transcription relationship by reconstructing gene-level chromatin interaction. *Computational and structural biotechnology journal*, 17:195–205, 2019.

[24] Annelyse Thévenin, Liat Ein-Dor, Michal Ozery-Flato, and Ron Shamir. Functional gene groups are concentrated within chromosomes, among chromosomes and in the nuclear space of the human genome. *Nucleic acids research*, 42(15):9854–9861, 2014.

[25] Gordon K Smyth. Limma: linear models for microarray data. In *Bioinformatics and computational biology solutions using R and Bioconductor*, pages 397–420. Springer, 2005.

[26] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

[27] Thomas J Hardcastle and Krystyna A Kelly. bayseq: empirical bayesian methods for identifying differential expression in sequence count data. *BMC bioinformatics*, 11(1):422, 2010.

[28] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):550, 2014.

[29] Noel Cressie. Statistics for spatial data. *Terra Nova*, 4(5):613–617, 1992.

[30] Yanli Wang, Fan Song, Bo Zhang, Lijun Zhang, Jie Xu, Da Kuang, Daofeng Li, Mayank NK Choudhary, Yun Li, Ming Hu, et al. The 3d genome browser: a web-based browser for visualizing 3d genome organization and long-range chromatin interactions. *Genome biology*, 19(1):151, 2018.

[31] Mélina Gallopin, Andrea Rau, and Florence Jaffrézic. A hierarchical poisson log-normal model for network inference from rna sequencing data. *PloS one*, 8(10):e77503, 2013.

[32] Alexandra Despang, Robert Schöpflin, Martin Franke, Salaheddine Ali, Ivana Jerkovic, Christina Paliou, Wing-Lee Chan, Bernd Timmermann, Lars Wittler, Martin Vingron, et al. Functional dissection of tads reveals non-essential and instructive roles in regulating gene expression. 2019.

[33] Darío G Lupiáñez, Katerina Kraft, Verena Heinrich, Peter Krawitz, Francesco Brancati, Eva Klopocki, Denise Horn, Hülya Kayserili, John M Opitz, Renata Laxova, et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, 161(5):1012–1025, 2015.

[34] Mattia Forcato, Chiara Nicoletti, Koustav Pal, Carmen Maria Livi, Francesco Ferrari, and Silvio Bicciato. Comparison of computational methods for hi-c data analysis. *Nature methods*, 14(7):679, 2017.

[35] Stanley Wasserman, Katherine Faust, et al. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

[36] Paul Erdős and Alfréd Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60, 1960.

[37] Thomas Cremer and Christoph Cremer. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nature reviews genetics*, 2(4):292, 2001.

[38] Tim J Stevens, David Lando, Srinjan Basu, Liam P Atkinson, Yang Cao, Steven F Lee, Martin Leeb, Kai J Wohlfahrt, Wayne Boucher, Aoife O'Shaughnessy-Kirwan, et al. 3d structures of individual mammalian genomes studied by single-cell hi-c. *Nature*, 544(7648):59, 2017.

[39] Miguel R Branco and Ana Pombo. Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. *PLoS biology*, 4(5):e138, 2006.

[40] Elphège P Nora, Bryan R Lajoie, Edda G Schulz, Luca Giorgetti, Ikuhiro Okamoto, Nicolas Servant, Tristan Piolot, Nynke L van Berkum, Johannes Meisig, John Sedat, et al. Spatial partitioning of the regulatory landscape of the x-inactivation centre. *Nature*, 485(7398):381, 2012.

[41] S Kozubek, E Lukášová, A Marečková, M Skalnikova, M Kozubek, E Bartova, V Kroha, E Krahulcova, and J Šlotová. The topological organization of chromosomes 9 and 22 in cell nuclei has a determinative role in the induction of t (9, 22) translocations and in the pathogenesis of t (9, 22) leukemias. *Chromosoma*, 108(7):426–435, 1999.

[42] Andreas Bolzer, Gregor Kreth, Irina Solovei, Daniela Koehler, Kaan Saracoglu, Christine Fauth, Stefan Müller, Roland Eils, Christoph Cremer, Michael R Speicher, et al. Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS biology*, 3(5):e157, 2005.

[43] David Eisenberg, Edward M Marcotte, Ioannis Xenarios, and Todd O Yeates. Protein function in the post-genomic era. *Nature*, 405(6788):823, 2000.

[44] Naihui Zhou, Yuxiang Jiang, Timothy R Bergquist, Alexandra J Lee, Balint Z Kacsoh, Alex W Crocker, Kimberley A Lewis, George Georghiou, Huy N Nguyen, Md Nafiz Hamid, et al. The cafa challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *bioRxiv*, page 653105, 2019.

[45] Gene Ontology Consortium. The gene ontology resource: 20 years and still going strong. *Nucleic acids research*, 47(D1):D330–D338, 2018.

[46] Carrie A Davis, Benjamin C Hitz, Cricket A Sloan, Esther T Chan, Jean M Davidson, Idan Gabdank, Jason A Hilton, Kriti Jain, Ulugbek K Baymuradov, Aditi K Narayanan, et al. The encyclopedia of dna elements (encode): data portal update. *Nucleic acids research*, 46(D1):D794–D801, 2017.

[47] Fiona Cunningham, Premanand Achuthan, Wasiu Akanni, James Allen, M Ridwan Amode, Irina M Armean, Ruth Bennett, Jyothish Bhai, Konstantinos Billis, Sanjay Boddu, et al. Ensembl 2019. *Nucleic acids research*, 47(D1):D745–D751, 2018.

[48] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6):721–741, 1984.

[49] Mark S Kaiser and Noel Cressie. The construction of multivariate distributions from markov random fields. *Journal of Multivariate Analysis*, 73(2):199–220, 2000.

[50] Faming Liang. A double metropolis–hastings sampler for spatial models with intractable normalizing constants. *Journal of Statistical Computation and Simulation*, 80(9):1007–1022, 2010.

[51] Emily C Dimmer, Rachael P Huntley, Yasmin Alam-Faruque, Tony Sawford, Claire O'Donovan, Maria J Martin, Benoit Bely, Paul Browne, Wei Mun Chan, Ruth Eberhardt, et al. The uniprot-go annotation database in 2011. *Nucleic acids research*, 40(D1):D565–D570, 2011.

[52] Philipp G Maass, A Rasim Barutcu, and John L Rinn. Interchromosomal interactions: A genomic love story of kissing chromosomes. *The Journal of Cell Biology*, 218(1):27–38, 2019.

[53] Philipp G Maass, A Rasim Barutcu, Catherine L Weiner, and John L Rinn. Inter-chromosomal contact properties in live-cell imaging and in hi-c. *Molecular cell*, 69(6):1039–1045, 2018.

[54] Zhenhua Hu and Wee-Wei Tee. Enhancers and chromatin structures: regulatory hubs in gene expression and diseases. *Bioscience reports*, 37(2):BSR20160183, 2017.

[55] Naihui Zhou. *Supplementary Data for Detecting Spatial Dependency in Gene Expression*. figshare, Dec 2019.