MDPI

Article

Spatial organization of the Gene Regulatory Program: An information theoretical approach to breast cancer transcriptomics

Guillermo de Anda-Jáuregui [1], Jesús Espinal-Enriquez [1,2] and Enrique Hernández-Lemus [1,2,*]

[1] Computational Genomics Division, National Institute of Genomic Medicine
[2] Centro de Ciencias de la Complejidad, Universidad Nacional Autónoma de México
* Correspondence: ehernandez@inmegen.gob.mx (E.H-L.)

Version January 28, 2019 submitted to Entropy

Abstract: Gene regulation may be studied from an information-theoretic perspective. Gene Regulatory programs are representations of the complete regulatory phenomenon associated to each biological state. In diseases such as cancer, these programs exhibit major alterations, which have been associated to the spatial organization of the genome into chromosomes. In this work, we analyze intrachromosomal, or *cis*-, and interchromosomal, or *trans*- Gene Regulatory programs in order to assess the differences that arise in the context of breast cancer. We find that using Information Theoretic approaches, it is possible to differentiate *cis*- and *trans*- regulatory programs in terms of the changes that they exhibit in the breast cancer context, indicating that in breast cancer there is a loss of *trans*-regulation. Finally, we use these programs to reconstruct a possible spatial relationship between chromosomes.

Keywords: Gene Regulatory Program; Mutual Information; Markov Random Field; Spatial Dependency Structures; Cancer Transcriptomics

1. Introduction

1.1. The Gene Regulatory Program

In order to respond to external stimuli, maintain the basal function, or adapt to new environments, the cell triggers a sophisticated mechanism to produce the specific class and amount of elements responsible for carrying out the particular tasks involved in a cellular context. Processes such as development, cell differentiation and homeostasis are driven and controlled by a set of genes interacting in time and space to respond to the changing environment. We will call to said set of genes and the manner in which they interact as Gene Regulatory Program (GRP).

In the eukaryotic cell, DNA is packed forming structural units called chromosomes. The human cell contains 23 chromosomes. Chromosomes are composed by the already mentioned DNA molecule (in which the genetic information is encoded), and structural proteins called histones, which attach the DNA molecule to them. These elements form the chromatin fiber, which in turn is coiled to generate the structure of a chromosome.

To initiate the gene transcription (production of an RNA molecule from DNA), and the consequent gene regulatory program, the chromosome must be "open", i.e., DNA should be visible to the proteins which will carry out the transcriptional process. Opening of DNA is a highly coordinated event that

allow the simultaneous production of RNA molecules in different sections of the chromosome, but also in different chromosomes. This co-regulated production of genes is one of the most important factors to generate a GRP. Fom now on, intra-chromosomal regulation will be termed *cis-* regulation whereas we will refer to inter-chromosomal *trans-* regulation. In this, we are somewhat extending or borrowing the classical concepts of cis- and trans- regulation (Stergachis et al. 2014) instead of the more verbose *intra-chromosomal* and *inter-chromosomal* terms.

1.2. Spatial anomalies in cancer-associated GRPs

The whole GRP determine the phenotype. Since gene regulation is keystone for the correct functioning of the living cell, abnormal performance of the way in which genes are co-regulated in time and space give place to aberrant phenotypes. A paradigmatic example of this is cancer.

During the rise and development of a cancerous phenotype, several abnormal signals of gene regulation are triggered. This set of signals produce a faster cell growth, cell duplication and proliferation, evasion of of immune system, and others. The majority of said hallmarks of cancer (Hanahan and Weinberg 2011) are produced by genes in which mutations, different expression patterns or epigenetic signals appear. This altered gene expression pattern can be studied by means of next generation sequencing (NGS) techniques such as RNA-Seq, which allow to have at the genome-wide level the information of the amount of any RNA transcript from a given sample (person).

NGS opened the possibility to have the information regarding the gene expression of the whole genome of several samples. The large data corpus allow to increase the statistical power and observe general behavior of a cancerous phenotype and compare it with a non-cancerous one. Other approaches to this problem have been developed including 3C, 4C (Aguilar-Arnal and Sassone-Corsi 2015) and Hi-C chromosome capture techniques (Dryden et al. 2014) as well as ultra-microscopy (Cremer et al. 2017), among others.

With the aforementioned in mind, a simple and direct point of investigation is the observation of a GRP in cancer and compare it with a normal (non-cancerous) program. This is, comparing at the genome-wide level, the whole set of gene interactions between these two phenotypes (cancer and normal).

Previously (De Anda-Jaúregui et al. 2018; Espinal-Enriquez et al. 2017; García-Cortés et al. 2018), we observed that in breast cancer, *trans-* (inter-chromosome) gene interactions are more scarce and weaker in cancer samples compared to the healthy phenotype. Furthermore, in breast cancer, *cis-* interactions become stronger but between physically close genes, and this gene correlation strength decays with the distance. Said effect is not present in the normal phenotype.

In order to characterize in a quantitative manner the qualitative differences observed between the two phenotypes, in this work we have implemented an information theoretical approach, by constructing a series of indicators that, as it will be shown later, allow the classification for the distinctive patterns of both GRPs.

1.3. An information theoretical approach to gene regulatory programs

A paradigmatic question in contemporary computational biology, is the probabilistic inference of the *best* set of regulatory interactions between genes starting from a large –but incomplete– data corpus Ω. This is, being able to found the maximum-likelihood or maximum-entropy solution to the deconvolution of the GRP of the cells starting from data sampled in, say RNA sequencing experiments over whole genome transcriptomes. Such deconvolution involves the inverse problem of large scale

probabilistic inference over an incomplete and noisy sample space.

A paramount solution to this extremely difficult task is founded on the tenets of Information Theory (Hernández-Lemus and Rangel-Escareño 2011), as we will show in what follows.

Let $X_i = \{X_1, X_2, \ldots, X_N\}$ be a set of N random variables, representing the expression levels of N genes in a transcriptome. For each duplex $\mathbb{D}_{i,j} = (i, j)$ (i.e. a pair of genes), the mutual information function $I(X_i, X_j)$ is given by (Cover and Thomas 2012):

$$I(X_i, X_j) = \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}} P(X_i, X_j) \log \frac{P(X_i, X_j)}{P(X_i)\, P(X_j)} \tag{1}$$

$\mathcal{I}$ and $\mathcal{J}$ are the complete gene expression sampling spaces for genes i and j respectively –i.e. the sets of all possible values of the experimentally measured gene expression levels X_i, and X_j, within a large experimental data corpus Ω. $P(X_i, X_j)$ is the joint probability distribution of X_i and X_j and $P(X_i)$ and $P(X_j)$ are the marginal probability distributions of X_i and X_j, respectively. As it is widely know, the mutual information function $I(X_i, X_j)$ quantifies the statistical dependence between two given random variables X_i and X_j (Cover and Thomas 2012).

We can also define the *off-diagonal mutual information*, $I^\dagger$ as follows :

$$I^\dagger(X_i, X_j) = I(X_i, X_j) \cdot (1 - \delta_{ij}) \tag{2}$$

δ_{ij} is Kronecker's delta. The purpose of $I^\dagger(X_i, X_j)$ is to eliminate self-information from our calculations. From now on, we will drop the † superscript and we will always refer to the off-diagonal mutual information in all of our further calculations.

A GRP encompasses the full set of interactions among genes that gives rise to a transcriptional phenotype. Within the context of the theoretical and experimental settings we have just described, let us define what the solution of a GRP Deconvolution problem is.

Following previous work (De Anda-Jaúregui et al. 2018), we define a Gene Regulatory Program (GRP) as a graph $\mathcal{G}[I(X_i, X_j)]$ of all the mutual information functions for a given empirical transcriptomics sampling space Ω. It can be shown that $\mathcal{G}[I(X_i, X_j)]$ is indeed a Markov Random Field (Kindermann 1980; Moussouris 1974) considering mutual information distributions under the pairwise sufficiency assumption (Merchan and Nemenman 2016).

We will consider both *cis* and *trans* GRPs $\mathcal{G}^{k,l}[I(X_i, X_j)]$, here $k, l = \{1, 2, \ldots, 22, x\}$ are indexes working as the chromosome label. $k = l$ implies associations between genes i and j located in the same chromosome (*cis-GRPs*, $\mathcal{G}[I(X_i, X_j)]^{cis}$) , whereas $k \neq l$ are statistical dependencies in different chromosomes (*trans-GRPs*, $\mathcal{G}[I(X_i, X_j)]^{trans}$). Partitions of the global GRP $\mathcal{G}[I(X_i, X_j)]$ into its cis- and trans- constituents, are called subregulatory programs from now on.

2. Analysis

2.1. Data

The inference of the GRPs $\mathcal{G}[I(X_i, X_j)]$ is based the RNA sequencing of basal breast cancer patients and healthy samples from the Cancer Genome Atlas (TCGA) collaboration (TCGA 2012) data acquired, and pre-processed as described in (Espinal-Enriquez et al. 2017). Briefly, we used 142 Basal-like subtype breast cancer samples, as well as 101 solid-tissue normal samples. Each sample contains 15,642 annotated genes, after removal of low-counts transcripts ($<$ 5 per sample). This set

of un-paired data were pre-processed, normalized and bias-reduced, to have a comparable set of expression data between cancer and normal samples.

2.2. GRP inference

GRPs for tumors and controls were obtained by calculating MI values for every pair of genes i, j in the genome as measured in the aforementioned RNA sequencing data. These calculations were performed using an in-house (Tovar et al. 2015) parallel implementation based on the ARACNE (Margolin et al. 2006) engine.

2.3. Measures of change in MI between health and disease

In order to characterize in a quantitative manner the qualitative differences observed between the mutual information distributions making up for the statistical dependence structure behind the different conditions, we have implemented a series of indicators that, as it will be shown later allow the classification for the distinctive patterns or features of the GRPs.

Consider two GRPs $\mathcal{G}^{k,l}_{tumor}$ and $\mathcal{G}^{k,l}_{control}$ representing the set of interactions among genes in a phenotype. We may define a difference matrix $\mathcal{G}^{k,l}_{\Delta}$ as follows:

$$\Delta\mathcal{G}^{k,l} = \mathcal{G}^{k,l}_{tumor} - \mathcal{G}^{k,l}_{control}$$

$\Delta\mathcal{G}^{k,l}$ describes the changes in the interactions among genes, in terms of MI, between the two phenotypes.

2.3.1. Gain Loss Score

The first indicator that we define is the Gain Loss Score, $\mathcal{GLS}$ an aggregated measure of the direction of MI changes in a GRP.

$$\mathcal{GLS}^{k,l} = \frac{|(\Delta[I(X_i, X_j)] \in \mathcal{G}^{k,l}_{\Delta}) > 0| - (\Delta[I(X_i, X_j)] \in \mathcal{G}^{k,l}_{\Delta}) < 0|}{|(\Delta[I(X_i, X_j)] \in \mathcal{G}^{k,l}_{\Delta})|} \tag{3}$$

Basically, $\mathcal{GLS}$ is the difference between the number of gene pairs that exhibit a gain in MI values minus the number of gene pairs that exhibit a loss in MI values between the two phenotypes, divided by the total number of gene pairs. This indicator will be positive if there are more gains, and negative if there are more losses.

2.3.2. Gain Loss Ratio

The second indicator we define is the Gain Loss Ratio, $\mathcal{GLR}$ which is an aggregated measure of the magnitude of the losses and gains of MI. Basically, it is the ratio of the absolute mean value of MI gains over the absolute mean value of MI losses.

$$\mathcal{GLR}^{k,l} = \frac{absMean(\Delta\mathcal{G}^{k,l} > 0)}{absMean(\Delta\mathcal{G}^{k,l} < 0)} \tag{4}$$

The $\mathcal{GLR}$ indicator will be larger than 1 if the average value of MI gains is larger than the average value of MI losses, and will be smaller than 1 otherwise.

2.4. Comparison of GRPs between control and cancer conditions

To assess the changes in the overall behavior of GRPs between both conditions, we used the Kolmogorov-Smirnov (KS) test. We performed the KS test between cancer GRP^t_{ij} and control GRP^c_{ij}

GRPs to quantify the distance metric between the MI distributions. The null distribution of this statistic is calculated under the null hypothesis that the samples are drawn from the same distribution.

2.5. Comparison between cis- and trans-GRPs in each condition

To assess differences between *cis*- and *trans*- GRPs within the same biological condition, we again made use of the KS test. In each phenotype (tumor or control), we performed the KS test to compare, for each chromosome k, the difference between the $cis - GRP_{kk}$ and every $trans - GRP_{kl}$ regulatory programs.

Additionally, we decided to compare, in both biological conditions, each $cis - GRP_{kk}$ to every $trans - GRP_{kl}$ for each chromosome k by using the *Hellinger distance*, $\mathcal{H}_2(X, Y)$. The Hellinger distance $\mathcal{H}_2$ is a semi-quadratic form of f-divergence to measure the difference between two probability functions. Unlike the KS metric –already introduced– that considers maximum deviations (as given by the supremne difference), we may think of $\mathcal{H}_2(X, Y)$ as a weighted average of the odds ratio given by a probability distribution X which is absolutely continuous respect to another probability distribution Y. For the case of the sub-regulatory programs we have the following expression:

$$\mathcal{H}_2(\delta\mathcal{G}^{k,k}, \delta\mathcal{G}^{k,l}) = \frac{1}{\sqrt{2}} \left\| \sqrt{\delta\mathcal{G}^{k,k}} - \sqrt{\delta\mathcal{G}^{k,l}} \right\|_2 \tag{5}$$

Here $\|\cdot\|_2$ is the Euclidean norm. $\delta\mathcal{G}^{k,k}$ is the probability density of the cis-GRP for chromosome k and $\delta\mathcal{G}^{k,l}$ is the probability density of the trans-GRP involving chromosomes k and l.

3. Results and Discussion

3.1. Intra- and Inter-chromosome interactions exhibit differences in MI changes

We have previously observed that intra and inter-chromosome interactions behave differently in breast cancer and regular breast tissue; if a threshold is established based on MI values, as to generate sparse graphs, the observed effect may be thought of as a loss of trans-regulation in breast cancer, as compared to healthy breast tissue (Espinal-Enriquez et al. 2017). By considering full *cis*- and *trans*-GRPs it is possible to further assess the way in which these types of interactions change.

In Figure 1 we observe the changes of *cis*- GRPs ($\Delta\mathcal{G}^{cis}$) for every chromosome as well and *trans*-GRPs ($\Delta\mathcal{G}^{trans}$) for every pair of chromosomes, in terms of two indicators: $\mathcal{GLS}$, a measure of the direction of MI changes, and $\mathcal{GLR}$, a measure of the magnitude of losses and gains in MI.

It may be seen that *trans* interactions between any two pairs of chromosomes exhibit overall more losses than gains in terms of MI, with higher MI drops than MI gains. On the other hand, (*cis*) interactions in each chromosome have more varied behaviors: a) either they also exhibit losses, but both their frequency and magnitude are lower than the one observed in *trans* interactions (this is the case for chromosomes 1, 2, 5, 6, 11, 17, 19 and X); or b) they exhibit more losses than gains, but the average magnitude of the gains is higher than the average magnitude of the losses (the case for chromosomes 3, 4, 7, 8, 9, 10, 12, 13, 14, 15, 16, 18, 19, 20, and 22); the behavior of chromosome 21 is the only one where there are more gains, and gains have a higher magnitude.

$\mathcal{GLS}$ and $\mathcal{GLR}$ are proportional. As it can be observed from Figure 1, an increase in g/l scrore, is accompanied with an increase in the $\mathcal{GLR}$.

3.2. cis- patterns depend on the chromosome size

The structure of a chromosome is composed by two arms: the p (short) and q arms, separated by a centromere (Supplementary Figure S2). Based on the position of the centromere, the chromosomes are

classified into *metacentric*, where the centromere is placed in the middle of the chromosome, *acrocentric*, where the centromere is placed closer to the extreme of the chromosome, and *submetacentric*, which the centromere is not in the center of the chromosome, but neither at the extreme position.

There is a direct relationship between the structure and number of genes in the chromosomes: metacentric and submetacentric chromosomes contain more genes than acrocentric chromosomes. Chromosome 1, 19, or 2, which are metacentric chromosomes, contain around 2,000 genes; meanwhile chromosome 21, 22, 13 or 14 contain around 300 genes.

Interestingly, the $\mathcal{GLS}$ and $\mathcal{GLR}$ in the *cis*- GRPs exhibit a different pattern depending on the size of the chromosomes: the larger chromosomes show lower $\mathcal{GLS}$ and $\mathcal{GLR}$ than acrocentric and smaller chromosomes. Supplementary Figure 1 and supplementary table 1 provide a more detailed description of this phenomenon.

This apparently functional behavior appears to be highly related to the structure of the chromosome. This is, during cancer the loss of information observed in terms of mutual information, depends on the number of genes in the chromosome, which is in turn related to the size of the chromosome. A possible explanation to this behavior could be related to the closeness between genes inside the chromosome. Meanwhile chromosomes 1, 17, 11 or 19 present a high density of genes, chromosomes 21, 18, or 13 are less dense and present less genes.

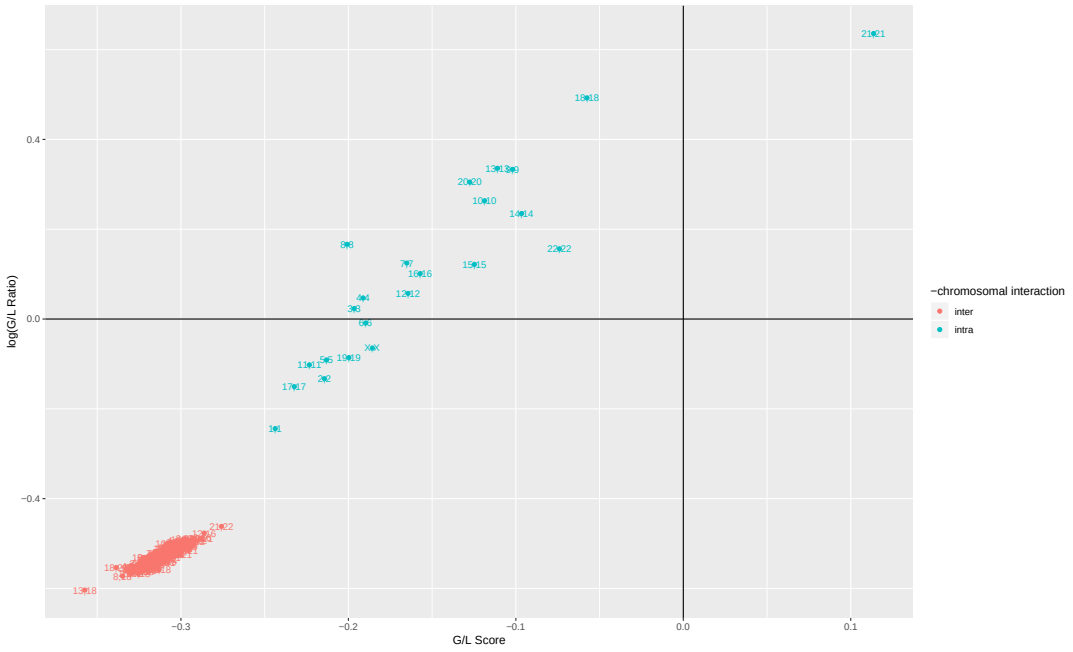

Figure 1. A scatterplot, where each point represents a subregulatory program for a pair of chromosomes, comprised of all MI values for each pair of genes in Chromosome i and Chromosome j. By comparing the MI between gene pairs in tumor and control, in terms of $\mathcal{GLS}$ (whether there are more losses or gains in MI) and $\mathcal{GLR}$ (whether MI losses or MI gains have a higher magnitude), we identify that interchromosomal interactions between genes in any pair of chromosomes have more losses than gains of MI in disease, with an average MI loss greater than the average MI gain. Meanwhile, intrachromosomal interactions may exhibit three different behaviors: i) they have more losses with higher average MI loss, although with higher $\mathcal{GLS}$ and $\mathcal{GLR}$ values than the interchromosomal interactions (chromosomes 1, 2, 5, 6, 11, 17, 19, X); ii) they have more losses, but the average MI gain is higher (chromosomes 3, 4, 7, 8, 9, 10, 12, 13, 14, 15, 16, 18, 19, 20, 22) or iii) they have more gains, with a higher average MI gain (21).

3.3. Cis-GRPs are more similar in health and disease than trans-GRPs

Based on previous observations regarding the changes in gene regulation observed in breast cancer, the observed phenomenon may obey to one of the following: *trans*-regulation becoming weaker, *cis*-regulation becoming stronger, or a combination of the two. By comparing whole GRPs between health and disease, it is possible to have a complete assessment of this phenomenon.

In figure 2, a heatmap is presented in which the color intensity is proportional to the log negative Kolmogorov-Smirnov (KS) distance between GRPs in tumors and the corresponding GRPs in health ($-\log(ks_{tc})$). As it is kown KS distance ks_{ij} arises from an uniparametric test to compare probability distributions $ks_{ij} = \sup_x |F_{i,n}(x) - F_{j,m}(x)|$ where $F_{i,n}$ and $F_{j,m}$ are the corresponding cummulative distributions. In the central diagonal, the KS distances between *cis*-GRPs may be found, while KS distances between *trans*-GRPs are found elsewhere. It may be seen that *cis*-GRPs are closer between health and disease (ranging from 0.07 to 0.18) than *trans*-GRPs which are notably farther.

These observations, along with those mentioned in section 2.1, may be pointing to a phenomenon in which *trans*-regulation in fact becomes weaker, whereas the *cis*-regulation is less severely affected, and therefore prevails as the main component of the regulatory phenomenon.

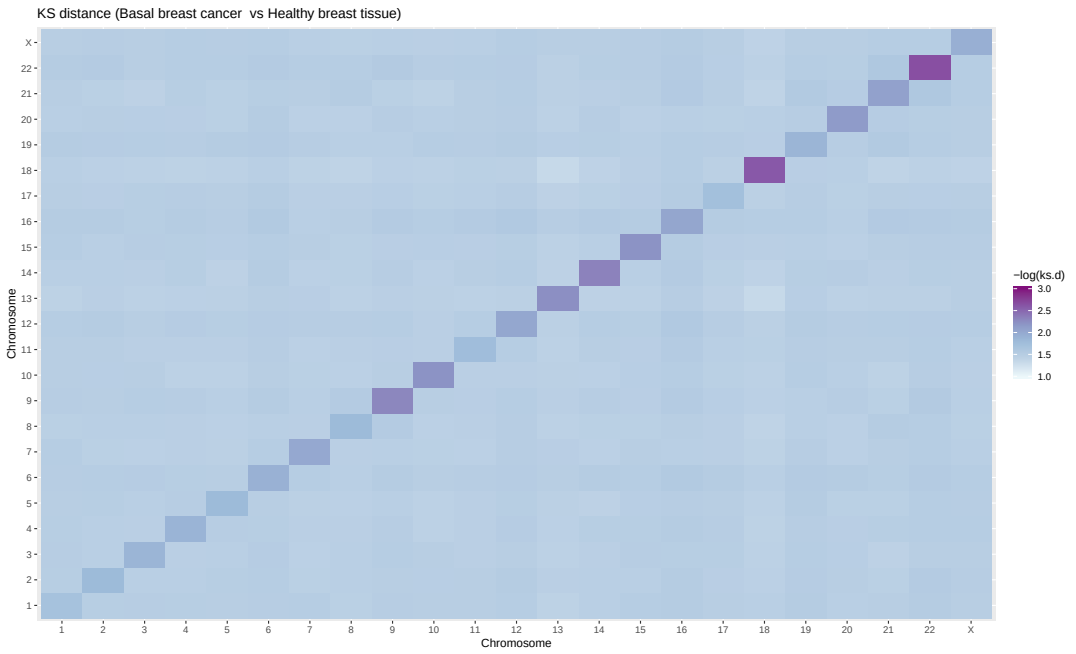

Figure 2. A heatmap showing the differences between GRPs in health and disease. In each square, the color intensity is proportional to $-\log(ks_{tc})$, the Kolmogorov-Smirnov distance between the subregulatory program for ij in cancer vs the subregulatory program for ij in control. We may observe that in general, the distances between *trans*- GRPs in control and cancer are greater than the distances between *cis*- GRPs in health and disease.

3.4. Differences in cis- and trans-GRPs in health and disease

A final question is to observe whether *cis*- and *trans*- regulation behaves differently within the same phenotype. We may evaluate this through the use of GRPs. We do so by comparing, for each chromosome k, the $\mathcal{G}^{k,k}$ with each $\mathcal{G}^{k,l}$ through the use of the KS test.

In figure 3 we show two heatmaps, one for tumors (panel A) and one for controls (panel B). In each heatmap, the color intensity is proportional to the (negative log) KS distance between the $\mathcal{G}^{k,k}$ (*cis*) and the $\mathcal{G}^{k,l}$ (*trans*). The figure clearly illustrates how, in the case of cancer, *trans*-GRPs involving

a given chromosome are virtually equidistant to the corresponding *cis*-GRP for said chromosome. Meanwhile, in healthy breast tissue, each *trans*-GRP has a unique distance from the corresponding *cis*-GRP. Furthermore, in all chromosomes in cancer, KS values are lower than those for the healthy phenotype.

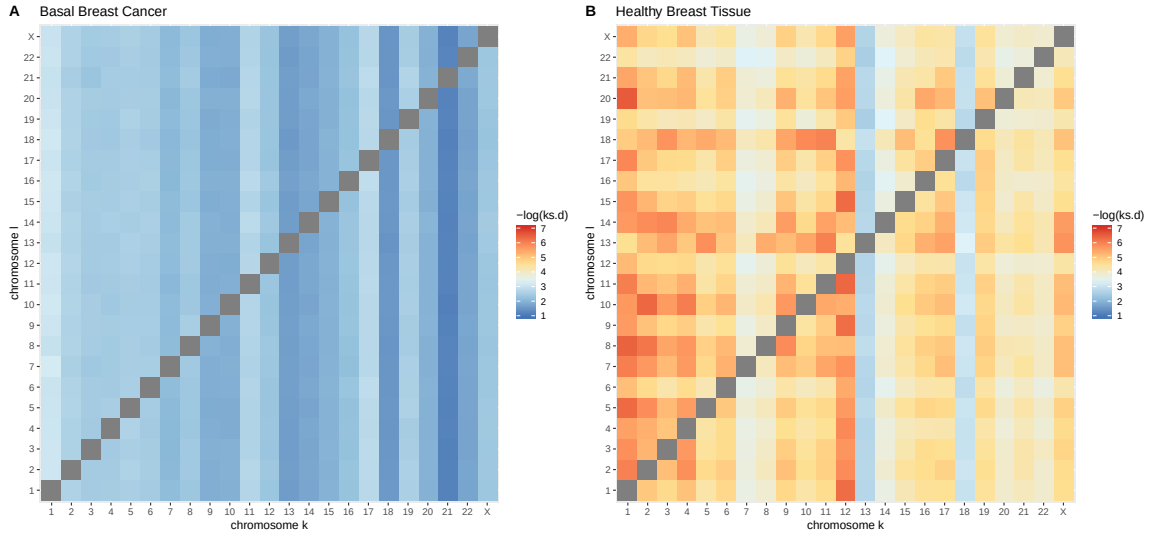

Figure 3. A heatmap showing the differences between $cis - GRP_k$ and $trans - GRP_{kl}$ in terms of KS statistic. Notice that in tumors, each $trans - GRP_{kl}$ is almost equidistant to $cis - GRP_{kk}$ (that is, each column has virtually the same color intensity in all rows), which is not the case in controls.

3.4.1. Reconstructing a spatial dimension of gene regulation through information theoretic approaches

As we have mentioned, the difference between *cis*- and *trans*- regulation is at its core, a spatial difference, as chromosomes are fundamentally units of biological organization in localized space. Therefore, the differences observed through these Information Theoretical approaches may be reflecting this spatial organization.

To illustrate this, we used another distance matrix: Hellinger distance between the Probability Density Functions (PDFs) associated to each GRP. For each chromosome k, the Hellinger distance between $PDF(\mathcal{G}^{k,k})$ and each $PDF(\mathcal{G}^{k,l})$ was calculated, in the cancer and healthy phenotypes.

In figure 4, we show network visualizations (panel A, cancer, panel B, healthy) in which each node represents a chromosome, and the links represent the aforementioned Hellinger distance between $PDF(\mathcal{G}^{k,k})$ and each $PDF(\mathcal{G}^{k,l})$. Through this, we may use a force-directed layout to organize these chromosomes in space. In these visualizations, the thickness and color intensity of the edges is higher if the distance between the PDFs is smaller. Furthermore, in the case of cancer, we may observe that the layout (based on the aforementioned Hellinger distances) *pushes together* certain pairs of chromosomes (such as chromosomes 2 and X, or 3 and 8). This is a phenomenon that is not observed in health.

It is important to mention that this observation by itself is not revealing a true spatial orientation of chromosomes in the cell nucleus space. However, based on the relationship that exists between information-theoretic based correlations in gene expression, and the spatial organization of genes, it may be indicative of specific spatial arrangements observed in each phenotype. In the end, ours is a descriptive method that may serve as a hypothesis generator; ultimately, experimental validation is needed. Our results could serve as an starting point for experimental explorations using novel technologies such as hi-C, ultra-microscopy, and future related techniques.

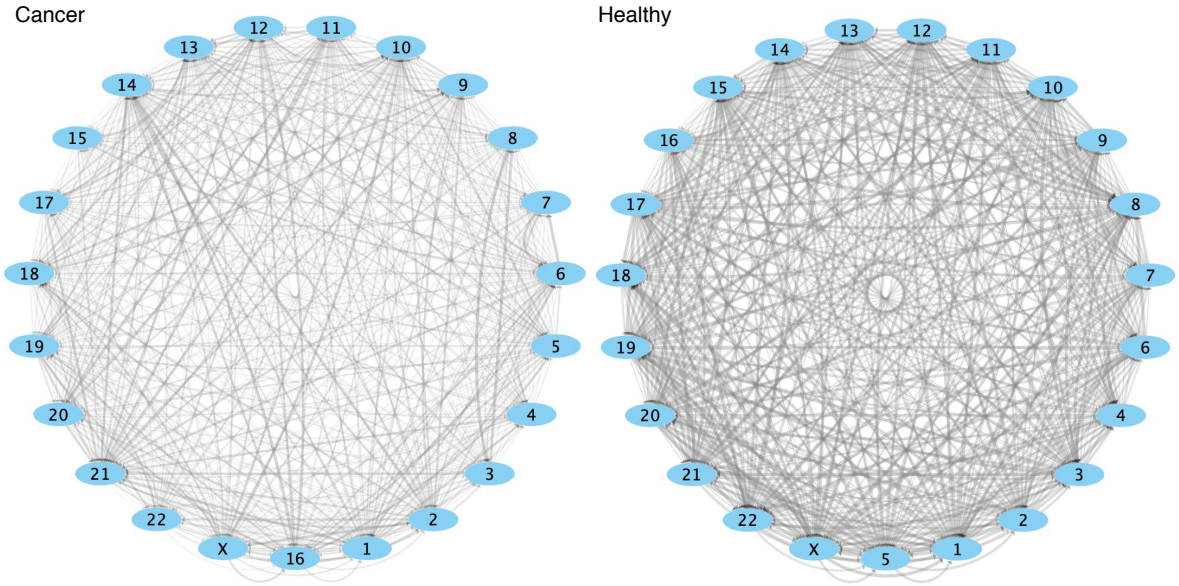

Figure 4. A network visualization of spatial behavior in terms of MI. Each node represents a chromosome. Each directed link has as weight the Hellinger distance (as calculated with the *textmineR* R package) between the PDFs of $cis - GRP_{kk}$ and $trans - GRP_{kl}$. The intensity of each link (transparency and thickness) is inversely proportional to the Hellinger distance. The nodes are arranged using a prefuse force-directed layout algorithm, considering the inverse of the Hellinger distance. This pushes nodes where $cis - GRP_{kk}$ and $trans - GRP_{kl}$ are similar together. Notice that the position of chromosomes is different in tumors and controls. Also notice that, overall, links are thicker (that is, PDFs are closer) in controls. Supplementary Figure 1 provides a *force-directed* visualization that shows some cases in tumor networks where chromosomes are "pushed together" (such as 2 and X, or 3 and 8).

4. Conclusions

Despite the enormous effort that has been devoted to dissect and analyze the molecular origins of breast cancer, the complex gene regulatory mechanisms behind this terrible disease still constitute a conundrum challenging diagnostic and therapeutic interventions. With the advent of high throughput experimental approaches (and the big data provided by them), information theoretical tools have allowed us to analyze at an extremely detailed level such complex gene regulatory programs.

Here we have analyzed how cancer-associated gene regulatory programs present a robust phenomenon of spatial organization, associating mechanistic features of gene regulation with the three dimensional structure of genomes and its influence on the transcriptional machinery. In brief, we have discovered how the global regulatory patterns diverge from health. How some relationships are lost, and few are gained. *Cis*-regulation becomes the norm, while *trans*-regulation becomes undifferentiated. A new spatial organization thus emerges.

A number of questions and hypotheses arise from this study, namely

- To what extent changes in gene regulation are relevant to breast cancer evolution?
- What are the possible consequences (functional or otherwise) of regulatory localization?
- Why different chromosomes behave differently? Including, but not limited to size effects.
- Are these patterns different in different cancers? Are they similar?

Rigurous quantitative studies, firmly grounded on the tenets of information theory will no doubt continue shedding light on the phenomenology of complex diseases, thus providing pivotal insights to the advancement of medical science.

Supplementary Materials: The following supplementary materials are available online:

Figure S1: **Classification of Chromosme types.** Classifications of Chromosomes: I Telocentric Centromere placement very close to the top, p arms barely visible if visible at all. II Acrocentric q arms are still much longer than the p arms, but the p arms are longer than those in telocentric. III Submetacentric p and q arms are very close in length but not equal. IV Metacentric p and q arms are equal in length. A: Short arm (p arm) B: Centromere C: Long arm (q arm) D: Sister Chromatids. Figure used under CC BY-SA 4.0 licensing. Source: https://commons.wikimedia.org/w/index.php?curid=49028965
http://www.mdpi.com/1099-4300/xx/1/5/s1

Figure S2: **Force-directed chromosome-wise GRP network visualizations.**
http://www.mdpi.com/1099-4300/xx/1/5/s2

GRP for control and cancer. The whole Gene Regulatory Program for Basal and healthy phenotype are available upon request.

Author Contributions: Conceptualization, G.D.J and E.H.L.; methodology, G.D.J. and J.E.E.; software, G.D.J.; validation, G.D.J., J.E.E. and E.H.L.; formal analysis, E.H.L.; investigation, G.D.J., J.E.E. and E.H.L.; resources, G.D.J., J.E.E.; data curation, G.D.J., J.E.E.; writing–original draft preparation, G.D.J. and E.H.L.; writing–review and editing, G.D.J., J.E.E. and E.H.L.; visualization, G.D.J., J.E.E.; supervision, E.H.L.; project administration, E.H.L.; funding acquisition, J.E.E. and E.H.L.

Funding: This research was funded by Conacyt grant number 285544/2016 and 2115/2016 as well as Federal funding from the National Institute of Genomic Medicine. J.E.E. is an awardee of the Miguel Alemán-Valdés Foundation Fellowship for Medical Research. E.H.L. is an awardee of the Marcos Moshinsky Foundation Fellowship for the Physical Sciences.

Acknowledgments: The authors want to acknowledge Diana Garcia-Cortes and Cristobal Fresno for helpful discussions

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

GRP	Gene Regulatory Program
MI	Mutual Information function
PDF	Probability Distribution Function
TCGA	the Cancer Genome Atlas

References

Aguilar-Arnal, Lorena and Paolo Sassone-Corsi. 2015. Chromatin dynamics of circadian transcription. *Current molecular biology reports 1*(1), 1–9.

Cover, Thomas M and Joy A Thomas. 2012. *Elements of information theory*. John Wiley & Sons.

Cremer, Marion, Volker J Schmid, Felix Kraus, Yolanda Markaki, Ines Hellmann, Andreas Maiser, Heinrich Leonhardt, Sam John, John Stamatoyannopoulos, and Thomas Cremer. 2017. Initial high-resolution microscopic mapping of active and inactive regulatory sequences proves non-random 3d arrangements in chromatin domain clusters. *Epigenetics & chromatin 10*(1), 39.

De Anda-Jaúregui, Guillermo, Diana García-Cortés, Cristobal Fresno, Jesus Espinal-Enríquez, and Enrique Hernández-Lemus. 2018. Intra-chromosomal regulation decay in breast cancer. *Applied Mathematics and Nonlinear Sciences*, In Press.

Dryden, Nicola H, Laura R Broome, Frank Dudbridge, Nichola Johnson, Nick Orr, Stefan Schoenfelder, Takashi Nagano, Simon Andrews, Steven Wingett, Iwanka Kozarewa, et al.. 2014. Unbiased analysis of potential targets of breast cancer susceptibility loci by capture hi-c. *Genome research*, gr–175034.

Espinal-Enriquez, Jesus, Cristobal Fresno, Guillermo Anda-Jauregui, and Enrique Hernandez-Lemus. 2017, May. Rna-Seq based genome-wide analysis reveals loss of inter-chromosomal regulation in breast cancer. *Scientific Reports 7*, 1760. doi:10.1038/s41598-017-01314-1.

García-Cortés, Diana, Guillermo de Anda-Jáuregui, Cristobal Fresno, Enrique Hernandez-Lemus, and Jesús Espinal-Enriquez. 2018. Loss of trans regulation in breast cancer molecular subtypes. *bioRxiv*, 399253.

Hanahan, Douglas and Robert A Weinberg. 2011. Hallmarks of cancer: the next generation. *cell 144*(5), 646–674.

Hernández-Lemus, Enrique and Claudia Rangel-Escareño. 2011. The role of information theory in gene regulatory network inference. *Information Theory: New Research*, 109–144.

Kindermann, Ross. 1980. *Markov random fields and their applications*. American Mathematical Society.

Margolin, Adam A, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Dalla Favera, and Andrea Califano. 2006, March. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics 7 Suppl 1*, S7. doi:10.1186/1471-2105-7-S1-S7.

Merchan, Lina and Ilya Nemenman. 2016. On the sufficiency of pairwise interactions in maximum entropy models of networks. *Journal of Statistical Physics 162*(5), 1294–1308.

Moussouris, John. 1974. Gibbs and Markov random systems with constraints. *Journal of Statistical Physics 10*(1), 11–33.

Stergachis, Andrew B, Shane Neph, Richard Sandstrom, Eric Haugen, Alex P Reynolds, Miaohua Zhang, Rachel Byron, Theresa Canfield, Sandra Stelhing-Sun, Kristen Lee, et al.. 2014. Conservation of trans-acting circuitry during mammalian regulatory evolution. *Nature 515*(7527), 365.

TCGA. 2012. Comprehensive molecular portraits of human breast tumours. *Nature 490*(7418), 61.

Tovar, Hugo, Rodrigo García-Herrera, Jesús Espinal-Enríquez, and Enrique Hernández-Lemus. 2015. Transcriptional master regulator analysis in breast cancer genetic networks. *Computational biology and chemistry 59*, 67–77.

Sample Availability: The code associated with this research is available at the following repository: https://github.com/CSB-IG/regulaciontrans-pipeline