frontiers in Oncology

ORIGINAL RESEARCH
published: 17 November 2021
doi: 10.3389/fonc.2021.726493

Gene Co-Expression in Breast Cancer: A Matter of Distance

Alfredo González-Espinoza[1,2], *Jose Zamora-Fuentes*[2], *Enrique Hernández-Lemus*[2,3] *and Jesús Espinal-Enríquez*[2,3]*

[1] Department of Biology, University of Pennsylvania, Philadelphia, PA, United States, [2] Computational Genomics Division, National Institute of Genomic Medicine, Mexico City, Mexico, [3] Centro de Ciencias de la Complejidad, Universidad Nacional Autónoma de México, Mexico City, Mexico

Gene regulatory and signaling phenomena are known to be relevant players underlying the establishment of cellular phenotypes. It is also known that such regulatory programs are disrupted in cancer, leading to the onset and development of malignant phenotypes. Gene co-expression matrices have allowed us to compare and analyze complex phenotypes such as breast cancer (BrCa) and their control counterparts. Global co-expression patterns have revealed, for instance, that the highest gene-gene co-expression interactions often occur between genes from the same chromosome (*cis*-), meanwhile inter-chromosome (*trans*-) interactions are scarce and have lower correlation values. Furthermore, strength of *cis*- correlations have been shown to decay with the chromosome distance of gene couples. Despite this *loss of long-distance co-expression* has been clearly identified, it has been observed only in a small fraction of the whole co-expression landscape, namely the most significant interactions. For that reason, an approach that takes into account the whole interaction set results appealing. In this work, we developed a hybrid method to analyze whole-chromosome Pearson correlation matrices for the four BrCa subtypes (Luminal A, Luminal B, HER2+ and Basal), as well as adjacent normal breast tissue derived matrices. We implemented a systematic method for clustering gene couples, by using eigenvalue spectral decomposition and the *k*–medoids algorithm, allowing us to determine a number of clusters without removing any interaction. With this method we compared, for each chromosome in the five phenotypes: *a)* Whether or not the gene-gene co-expression decays with the distance in the breast cancer subtypes *b)* the chromosome location of *cis*- clusters of gene couples, and *c)* whether or not the *loss of long-distance co-expression* is observed in the whole range of interactions. We found that in the correlation matrix for the control phenotype, positive and negative Pearson correlations deviate from a random null model independently of the distance between couples. Conversely, for all BrCa subtypes, in all chromosomes, positive correlations decay with distance, and negative correlations do not differ from the null model. We also found that BrCa clusters are distance-dependent, meanwhile for the control phenotype, chromosome location does not determine the clustering. To our knowledge, this is the first time that a dependence on distance is reported for gene

OPEN ACCESS

Edited by:
Maria Rodriguez Martinez, IBM Research—Zurich, Switzerland

Reviewed by:
Sheng Liu, Indiana University, United States
Noemi Eiro, Jove Hospital Foundation, Spain
Kimberly Glass, Brigham and Women's Hospital and Harvard Medical School, United States

***Correspondence:**
Jesús Espinal-Enríquez
jespinal@inmegen.gob.mx

Specialty section:
This article was submitted to Cancer Genetics, a section of the journal Frontiers in Oncology

Received: *17 June 2021*
Accepted: *26 October 2021*
Published: *17 November 2021*

Citation:
González-Espinoza A, Zamora-Fuentes J, Hernández-Lemus E and Espinal-Enríquez J (2021) Gene Co-Expression in Breast Cancer: A Matter of Distance. Front. Oncol. 11:726493. doi: 10.3389/fonc.2021.726493

clusters in breast cancer. Since this method uses the whole *cis*- interaction geneset, combination with other -omics approaches may provide further evidence to understand in a more integrative fashion, the mechanisms that disrupt gene regulation in cancer.

Keywords: eigenvalue decomposition, gene co-expression clustering, loss of long-distance co-expression, co-expression matrices, breast cancer molecular subtypes

1 INTRODUCTION

1.1 Breast Cancer: A Complex Disease

Breast cancer is the first cancer-related cause of death in women worldwide. It is also, according to the most recent data (1), the most diagnosed neoplasm in the world. Breast cancer is also the malignant neoplasm with the highest incidence (1). Its diagnosis, response to treatment, relapse, and outcome are strongly determined by the molecular profile underlying the disease (2–4). The PAM50 classifier is among the most relevant methods of classification for breast cancer molecular subtypes (5). This molecular classification is based on the expression signature of 50 genes relevant to the oncogenic phenotype (5–7).

Publicly available massive cohorts of genomic and clinical data in the study of cancer, have allowed the analysis of an immeasurable amount of information. The latter has contributed to a better understanding of the oncogenic process (8). Based on gene expression of hundreds-to-thousands of samples, now it is possible to study such vast experimental information to infer and analyze the whole-genome co-expression landscape, aiming to highlight similarities and differences between cancer and non-cancer samples. Among these efforts, The Cancer Genome Atlas (TCGA) has contributed in an outstanding way (9).

1.2 Gene Co-Expression Networks

The study of Cancer within the framework of complex networks has become increasingly relevant in the last years (10–20). Given its size and complexity, genome-wide regulation may include a large number of features (all the genes), potentially inducing a fully connected network, with contributions of very different relevance and certainty. For this reason, several approaches to reduce its dimensionality have been implemented, including the use of threshold methods, to look for the most significant co-expression relationships (18, 21). In particular, in the case of breast cancer molecular subtype networks, the most significant co-expressed pairs have been used as connected nodes in biologically relevant modules (22–25).

Further approaches to determine the optimal network size may analyze a wide range of network scales (13, 26, 27) or backbone-related threshold networks (28), and even use gene co-expression subsets of clinical/biological relevance (29).

In the attempt of reducing the dimensionality of a fully-connected network, identification of groups of genes that behave in a similar way –indicating that their expression profiles are correlated– is a relevant problem and is still an open challenge in network biology (29, 30). The latter point is closely related to the so-called graph sparsification problem in graph theory. The choose of a significance threshold then becomes relevant.

For instance, in a recent study by Kimura et al. (31), an approach was developed to select parameters in genetic networks by computational methods (mainly Machine Learning and Artificial Intelligence). Other approaches have used the complete set of interactions in order to construct a network backbone (28). There, the authors used the complete matrix of interactions to obtain the most important relationships, preserving those edges with statistically significant deviations with respect to a null model for the local edge's weight assignment.

1.3 Gene Co-Expression Is Distance Dependent

In cancer, gene co-expression networks have been used to uncover genes and relationships that may represent crucial elements to determine differences between phenotypes (32). In particular, in breast cancer and breast cancer molecular subtypes (4), gene co-expression networks have been useful to identify the phenomenon of *loss of long-range co-expression* (10, 12, 14, 33): this is, a property observed in cancer networks in which the most significant gene co-expression relationships occur between genes that belong to the same chromosome, i.e., *cis*- interactions. Conversely, inter-chromosome (*trans*-) interactions are often weak in cancer.

Furthermore, the loss of long-range co-expression is not only observed at the level of genes located on different chromosomes. Regarding *cis*- (intra-chromosome) gene interactions, there is an exponential decay of strength of correlations (14) as genes become more distant. This situation could be related to a diminishing of the *accessibility* that a certain region of the genome may have of its environment during the carcinogenic process. Importantly, this lack of accessibility can be attributed to several factors, among which we can mention aberrant expression of transcription factors, copy number alterations, incorrect binding to CTCF, or changes in Topologically Associated Domains (TADs). All of these factors have the potential to alter, both, the structure of DNA and gene expression.

Despite this phenomenon has been discovered not only in breast cancer, but also in clear cell renal carcinoma (13), lung adenocarcinoma and squamous cell lung carcinoma (12), loss of long-range co-expression has been determined for the top highest interactions: a small subset of the most co-expressed gene-gene interactions (tens-to-hundreds of thousands) of the

whole co-expression landscape is observed to be biased to *cis*- interactions.

Since the strength of intra-chromosome interactions have been observed to be the highest ones, it becomes important to evaluate the behavior of the whole intra-chromosome landscape of cancer networks. In these terms, network clustering may provide us with information related to, for example, sets of genes constrained by physical restrictions in certain regions of the genome, genes that act in tandem, events related with the transcriptional process, etc.

To address the questions above, we performed a data-driven clustering analysis using a hybrid algorithm that involves eigenvalue decomposition and *k*–medoids from correlation matrices of each chromosome. These matrices were inferred from RNA-Seq-based gene expression. We evaluated whether or not the loss of long-range co-expression is preserved, by studying all chromosomes for the four breast cancer subtypes as compared with normal tumor-adjacent tissue as control.

With this approach, we constructed co-expression matrices for all chromosomes in adjacent normal breast tissue network, as well as in all four breast cancer subtypes. We analyzed the statistics for their clustering nearest neighbor distributions within each chromosome, comparing each breast cancer molecular subtype as well as the adjacent normal tissue. Additionally, for all phenotypes, we constructed a null model to provide statistical robustness to our analyses. With this, we present a systematic method for intra-chromosome gene clustering, which allows to compare the whole co-expression landscape between a cancerous phenotype with its control counterpart.

2 MATERIALS AND METHODS

2.1 Data Acquisition

Gene expression data of breast invasive carcinoma was collected from The Cancer Genome Atlas (TCGA) (34). 735 tumor and 113 non-cancerous (adjacent normal), samples were considered, see **Table 1**. Illumina HiSeq RNASeq samples were filtered (biotype, expression mean >10), pre-processed, and log_2 normalized gene expression values as described in (10). We performed data corrections for transcript length, GC content and RNA composition. Tumor expression values were classified using PAM50 algorithm into the respective intrinsic breast cancer sub-types (Luminal A, Luminal B, Basal, and HER2-Enriched) using the Permutation-Based Confidence for Molecular Classification (35) as implemented in the pbcmc R package (36).

Tumor samples with a non-reliable breast cancer sub-type call were removed from the analysis. To avoid overlapping patterns among subtype expression values, multidimensional noise reduction was performed using ARSyN R implementation (37), and a multidimensional Principal Component Analysis (PCA) was implemented to confirm noise reduction (14).

Since a crucial part of this work lies in having a highly-confident set of matrices, it is necessary to obtain as many well-characterized samples as possible, for each molecular subtype. Due to this fact, we decided to include all the available samples with a molecular subtype classification i.e., those samples with a molecular subtype label from the original source. Further investigations must be conducted with even more stringent inclusion and exclusion criteria, such as histologically confirmed diagnosis, histopathologically-assessed axillary lymph nodes, metastatic disease at presentation, adjuvant treatment, etc.

In order to provide all the information to reproduce our results, the clinical information about histological data by subtype-samples is now included in the **Supplementary Material S1**. There, for each breast cancer subtype sample we describe: 1) availability of historical adjuvant treatment, 2) lymph node assessment existence, 3) histological type of tumor and 4) axillary lymph-node-stage method type.

To show that those samples with the same molecular subtype are indeed properly classified in their molecular profiles to be included in our correlation matrices, we performed a Principal Component Analysis (PCA) for each subtype (**Supplementary Figure S1**). The PCA groups samples based on the main eigenvalues of the expression profiles. In this case, we present the two main principal components (X and Y axes of the **Supplementary Figure S1**) -though the calculations were made with the full eigenvalue spectra of the matrices. Hence, the PCA could indicate those samples that are not similar to the rest of their class (if any) or if there is any "confounded" or misclassified sample.

As it can be noticed in the **Supplementary Figure S1**, all subtype samples are clearly separated based on the molecular classification. All samples are grouped by its subtype (color). Hence, constructing correlation matrices by using these subtype-separated samples, certainly improves the statistical significance without adding a clear source of noise.

2.2 Correlation Matrices

We built intra-chromosomal cross-correlation matrices by estimating the Pearson correlation coefficient between the expression of two genes *i* and *j*, defined as follows:

$$C_{ij} = \frac{Cov(g_i, g_j)}{\sigma_{gi}\sigma_{gj}} = \frac{1}{N_s}\sum_{s=1}^{N_s}\frac{(g_{is} - \mu_{gi})(g_{js} - \mu_{gj})}{\sigma_{gi}\sigma_{gj}}, \quad (1)$$

where g_i is the set of N_s expression samples for gene *i*. By definition, a correlation matrix is symmetric ($C_{ij} = C_{ji}$),

TABLE 1 | Samples for each subtype.

Control	Basal	Her2	LumA	LumB
113	221	105	217	192

the elements in the diagonal are 1 ($C_{ii} = 1, \forall i$), and its values are bounded to $-1 \leq C_{ij} \leq 1$, where $C_{ij} = 1$ corresponds to perfect correlations, $C_{ij} = -1$ corresponds to perfect anticorrelations, and $C_{ij} = 0$ corresponds to uncorrelated gene pairs.

We calculated Pearson correlation between all genes for each chromosome for the five phenotypes. The code for calculation of Pearson correlations can be found in (38).

2.3 Spectral Decomposition

Pearson correlation matrices for each chromosome were calculated in order to analyze their spectral properties. Previous works on correlation matrices have shown that their spectral properties carry information about the structure and dynamics of the system (39–48).

For example, in stock market data, the first eigenvectors correspond to clusters of related industries (49, 50). In Electroencephalography measurements, these eigenvectors correspond to different functional regions in the brain (51). However, not all of the eigenvalues carry relevant information about the system. It has been shown that the smallest ones are the most sensitive to noise and some of them correspond to weak interrelations between small components from different clusters (47, 48). To distinguish how many eigenvalues contain useful information to identify clusters, we compared the spectral properties of the empirical correlation matrix to a null model represented by an ensemble of random matrices.

This ensemble of random matrices, is obtained by doing non-biased shuffling over the gene expression values for each sample (in this way, the original distribution of the data is preserved while its correlations will be destroyed) and computing the correlation matrix of each randomized data as in equation 1, we generated an ensemble of $n_m = 100$ random matrices for each chromosome and phenotype.

The k deviating eigenvalues of the empirical matrix from the randomized data $max(\lambda_R) < \{\lambda_1, \ldots, \lambda_k\}$ are the ones containing correlations that cannot be attributed to either the noise in the system or data randomization. It is worth noticing that instead of using the eigenvectors from the spectral decomposition, which can be difficult to separate into independent clusters (52) (see **Supplementary Figure S2**), we used the number of k deviating eigenvalues as the number of independent clusters for a different clustering method.

2.4 Clustering Analysis

We implemented a clustering analysis based on the *k-medoids* algorithm. In a similar fashion to *k-means*, *k-medoids* clustering attempts to minimize the *distance* between the elements inside a cluster but one element is designated as the center of the cluster. The *k-medoids* algorithm works not exclusively with Euclidean distances, but with general pairwise interactions, this means we can use the correlation values we have estimated for each intra-chromosome matrix. Since correlation values are signed and their magnitude goes from –1 to 1, we define the pairwise interactions between genes i and j as:

$$D_{i,j} \equiv 1 - \left|C_{i,j}\right|, \tag{2}$$

with $0 \leq D \leq 1$, high correlation or anti-correlation values mean close distance between points, while small correlation values will give higher distances. Finally, for the parameter k in the clustering algorithm, we considered the number of deviating eigenvalues as obtained from the spectral decomposition.

Given the stochastic nature of the *k-medoids* algorithm, we did $n_r = 100$ realizations for each clustering computation to ensure statistical significance ($p < 0.01$), choosing the output configuration as the one with the minimum mean distance between the centroids and the elements in each cluster.

In order to compare the clustering results between the control phenotype and any other cancer subtype in a given chromosome, we constructed the intra-cluster Nearest Neighbor Distance (NND) distribution for each subtype. The NND of a given gene i in a cluster k is defined as:

$$D^i_{nn} \equiv \min\left(|j - i|\right) \; \forall \; j \in C_k, \tag{3}$$

where C_k refers to the cluster k. To quantify the difference between the clustering in adjacent normal and cancer subtypes we compute Shannon's entropy $H(x) = -\Sigma_{x \in \chi} p(x) \log\left(p(x)\right)$ for the NND distributions, which in this case can be interpreted as how localized or how spread are the genes within each cluster in the chromosome. We also computed the Kolmogorov-Smirnov distance between the adjacent normal case and each of the Cancer subtypes. Given two cumulative distribution functions (CDF) the Kolmogorov-Smirnov distance is defined as:

$$D_{KS}(F_n, F_m) = \sup_x |F_n(x) - F_m(x)|, \tag{4}$$

where the functions F_n and F_m are the CDFs for two samples n and m.

3 RESULTS

A correlation matrix of the sort just described, can be visualized as a heatmap as shown in **Figure 1** where correlation matrices for adjacent normal and basal subtype samples in the chromosome 1 are displayed. The axis represent the genes ordered by their physical location in the chromosome. The clearest difference between both matrices seems to be the lowest value of absolute correlation for genes that are physically distant in the basal subtype case. The heatmaps for each chromosome in the five phenotypes can be observed in **Supplementary Materials S2**–**S6**.

The effect of loss in long range co-expression is consistent with previous works of regulatory networks in breast cancer (10, 12–14, 33, 53). The block-type structure of the basal subtype matrix suggests the utility of clustering analysis to compare the structural properties of the correlation matrices. In what follows, we will present results for these clustering analyses. Through the manuscript, the presented figures will show different chromosomes for the five phenotypes. This has been done, in

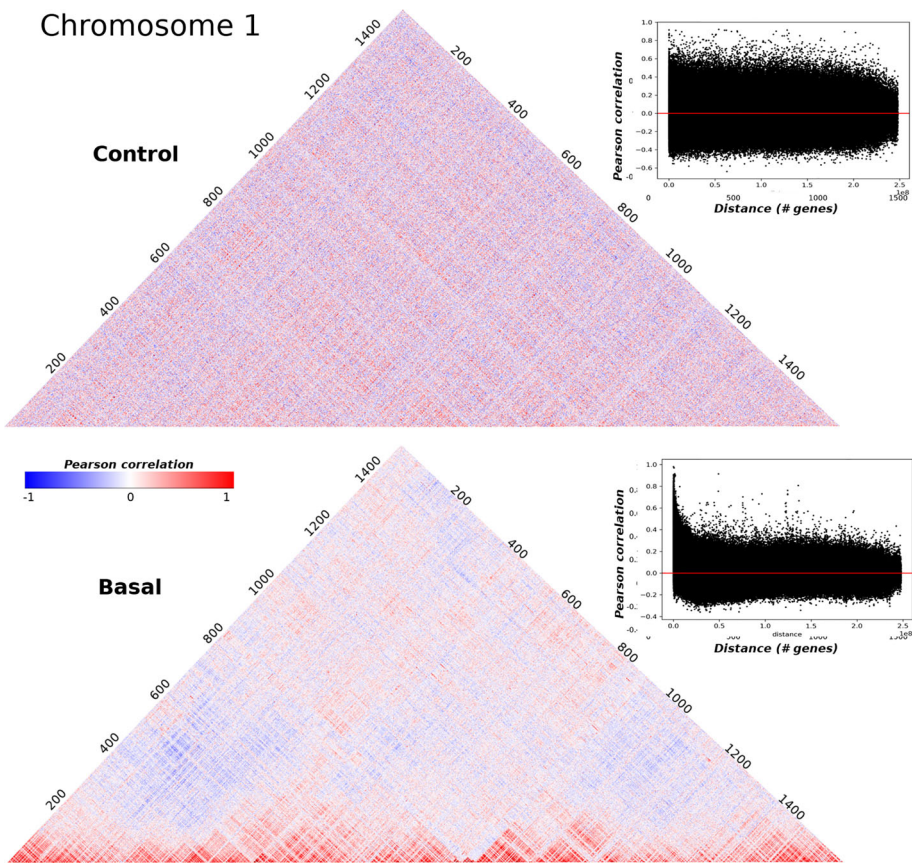

FIGURE 1 | Pearson correlation square matrices for chromosome 1 in control samples (up) and basal breast cancer subtype (down). Genes are placed according to their physical location on the chromosome. Colors represent the correlation value: red corresponds to positive values, meanwhile negative correlations are depicted in blue. The inserts in the right part of both matrices correspond to the scatterplot of Pearson correlations versus distance. The horizontal red line corresponds to Pearson correlation = 0.

order to illustrate the universal nature of the gene clustering in breast cancer molecular subtypes, compared with the adjacent normal tissue.

3.1 Co-Expression Decays in All Chromosomes in All Subtypes

We observed a common pattern of distance dependency in all chromosomes in all breast cancer phenotypes. The decay in gene co-expression corresponds exclusively to positive correlations. In the case of negative correlations, such effect is not observed. Conversely, in adjacent normal chromosomes, there is no dependency of distance neither in negative nor positive interactions. Interestingly, this effect is observed in all chromosomes in the four breast cancer molecular subtypes and not observed in adjacent normal breast tissue-derived correlations (**Figure 2** and **Supplementary Materials S7–S11**).

In order to evaluate the differences between the empirical data and the null model, we performed a non-parametric hypothesis test (Kolmogorov-Smirnov) for the correlation values distributions (in all tumor subtypes and adjacent normal tissue) versus phenotype-specific null models. Additionally we implemented their corresponding significance tests (obtained *via* bootstrap/permutation analysis). The results of the KS test can be observed in **Figure 3**. The results for the rest of chromosomes, as well as their significance p-values, are presented in **Supplementary Materials S12**, **S13**).

Notice that at short distances, the cancer phenotypes have larger values than the adjacent normal correlations. However, at larger distances, KS for adjacent normal network are larger than those for cancerous phenotypes. The p-values shown in the upper right part of the figure, represent the average of all set of distances.

Based on a null model that lacks the linear correlations from the original data (see Methods), we observed that in adjacent normal chromosomes, positive and negative correlation values seem to be independent of the distance between genes, having significantly higher absolute values when compared with the null model at any distance. In the case of cancer subtypes, negative correlations are non-significant, but a few small regions in specific chromosomes (See **Supplementary Figure S3**).

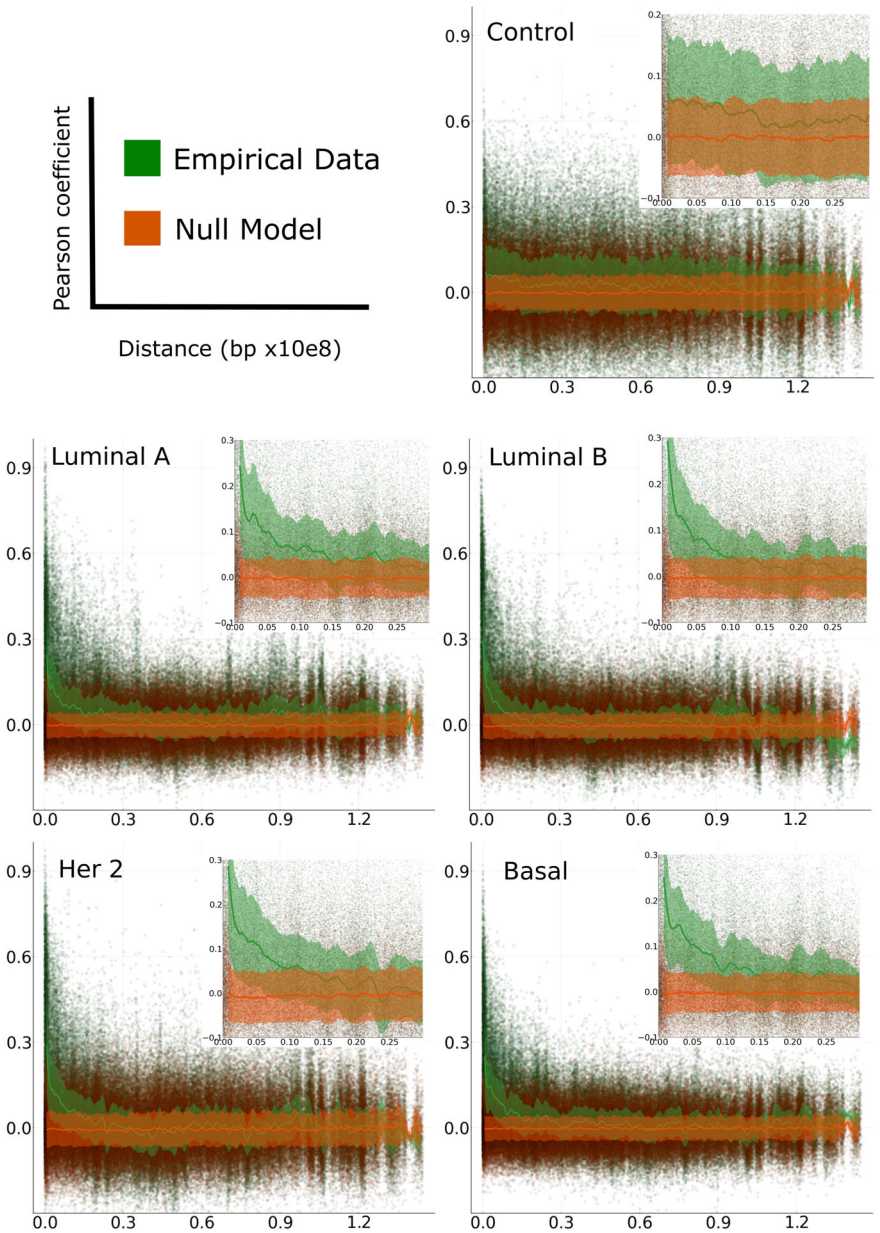

FIGURE 2 | Pearson correlation of gene-gene expression *versus* distance. Plots for adjacent normal and cancer subtypes of chromosome 8 (green) and their respective null model (orange). The solid lines represent the median of a moving average in the distribution of correlation values over each window and the shaded area is the range from its first and third quartiles.

3.2 Eigenvalue Decomposition Defines the Number of Clusters in All Phenotypes

We generated an ensemble of (N = 100) random matrices and compare the eigenvalue distributions from both, original and random matrices (see *Materials and Methods*). The left panel of **Figure 4** shows the eigenvalue distribution for the ensemble of random matrices, where its shape is the well-known Marchenko-Pastur distribution from random matrix theory (54). Overlapped eigenvalue distributions for the original matrix of chromosome 17 and the ensemble of surrogates are shown in the right panel of **Figure 4**. A set of significant eigenvalues was determined by random matrix permutations ($p < 0.01$) (see box in **Figure 4**).

3.3 Gene Clustering Is Distance Dependent in Breast Cancer

With the method referred in Section 3.2 we obtained the full set of clusters for each chromosome in all phenotypes. **Figure 5**

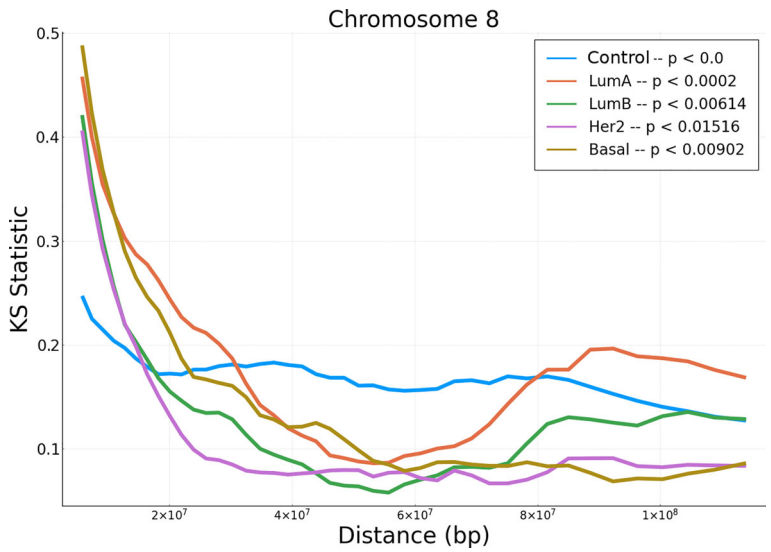

FIGURE 3 | KS hypothesis test between empirical data from chromosome 8 and null model for the five phenotypes. This plot represents the KS statistic *versus* distance for all phenotypes in chromosome 8. The *p*– value for the control phenotype is smaller than 10^{-5}.

shows Chromosome 19 clusters with genes sorted by gene start base pair position. In the adjacent normal chr19 figure (upper part) we cannot discern a pattern in cluster colors. The distribution of clusters does not seem to depend on the distance between genes. Meanwhile, in basal breast cancer, we can observe cluster panels of colors, clearly detached. In the same figure, in the right panels, we plot the cumulative distribution of genes for each cluster. The larger the slope, the more often contiguous genes belong to the same cluster. All clusters for the five phenotypes in all chromosomes can be found in **Supplementary Material S14**. Cumulative distribution for all clusters can be found in **Supplementary Material S15**.

Cumulative distributions for the Nearest Neighbor Distance (NND) of two different chromosomes are shown in **Figure 6**, which can be interpreted as the probability distribution of the minimum distance between two genes in the same cluster. The behavior seen in the previous **Figure 5** holds: genes from the same cluster are more likely to be close to each other.

Results for the entropies for the NND distributions are shown on the left panel of **Figure 7**, where a clear trend with the value $H(x)$ can be identified: Luminal A, Luminal B, HER2+, Basal. It is worth noticing that the aforementioned order coincides with survival rates and metastatic behavior (14, 55, 56). The subtypes with the lowest survival rates and more metastatic behavior also present lower entropy values.

The latter is in agreement with a previously observed trend for the top 0.1% gene co-expression interactions for the four phenotypes: The most aggressive phenotype (basal) has the lowest number of inter-chromosome interactions, meanwhile the Luminal A subtype, which is considered the one with the best prognosis, contains a much larger fraction of interactions between genes from different chromosomes (14).

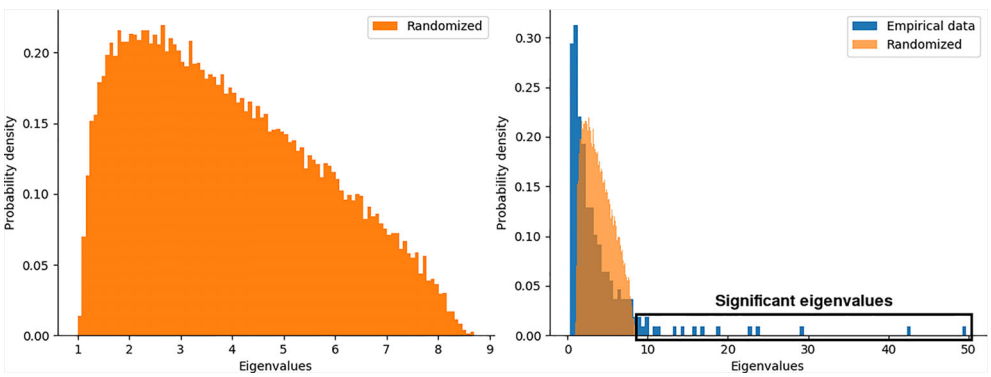

FIGURE 4 | Probability distributions of eigenvalues for a) the ensemble of random matrices, b) random and empirical data for the chromosome 17 in the Basal sub-type.

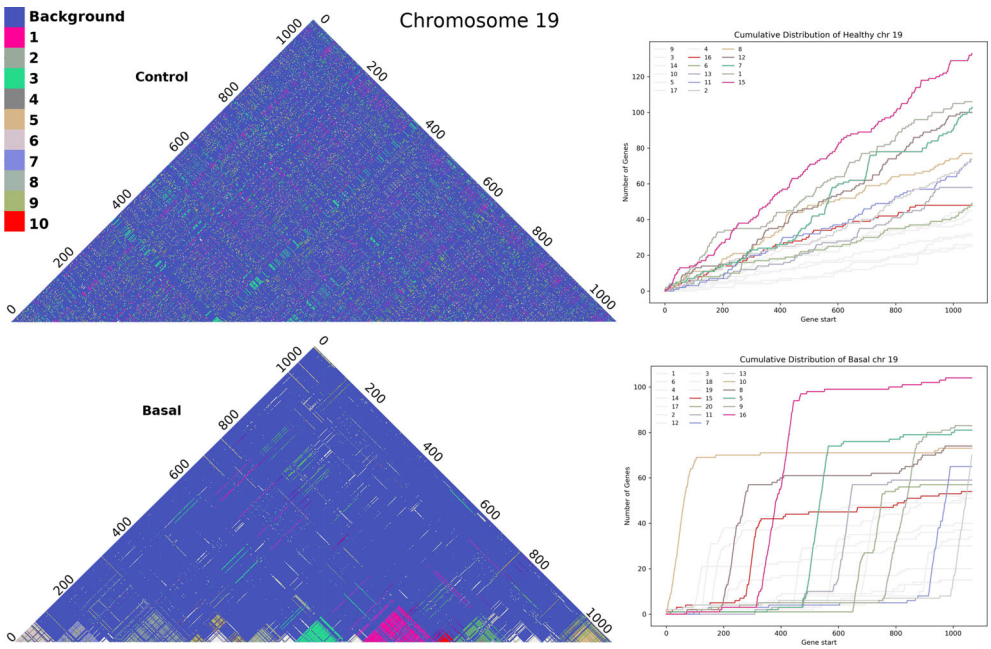

FIGURE 5 | Cluster assembly of Chromosome 19 for adjacent normal-tissue and Basal breast cancer matrices. Upper part: Heatmap for all clusters in the adjacent normal matrix. Lower part: analog heatmap for Basal breast cancer network. The right panels correspond to the cumulative distribution for each cluster in chromosome 19. Colors represent the top-10 largest clusters.

The decay in the entropy for the NND distribution presents further evidence that in the cancer subtypes, genes co-express in tighter patterns, in contrast with genes in the control phenotype that co-express at broadly scattered distances over the chromosome. A similar trend holds for the *KS* distance between the control phenotype and each subtype in the right panel of **Figure 7**, where higher values indicate a larger difference in the spatial organization of the clusters. The difference in spatial organization within the clusters in the chromosome is evident with both measures and it is correlated with the survival rate and metastasis of the subtype (14).

4 DISCUSSION

Cancer research increasingly requires comprehensive computational analysis tools. In the search for relevant biological information, it is essential to be able to find selective patterns of individual or collective gene expression. In this sense, clustering methods are becoming a pivotal computational tool.

In this work, we studied the co-expression of genes in breast cancer molecular subtypes. We implemented a method to find the optimal clustering between genes that are co-expressed. We observed a grouping pattern in the case of cancer phenotypes

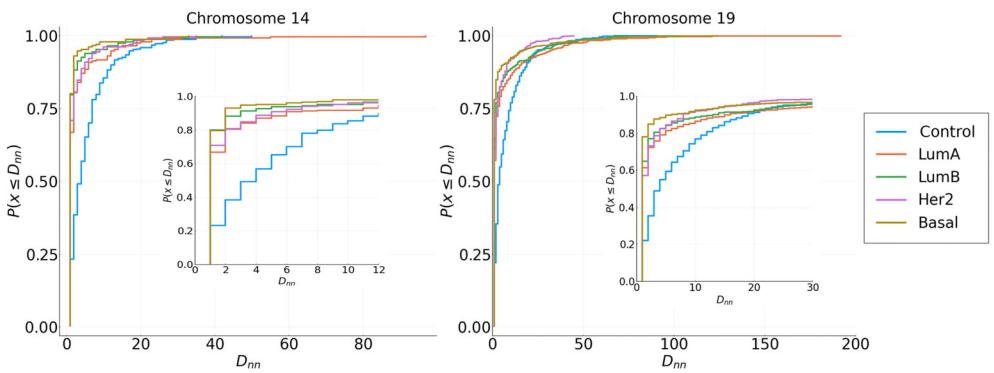

FIGURE 6 | Cumulative NND distributions for each subtype in Chromosomes 14 and 19. Distributions of Cancer subtypes have different behavior in short and long distances compared with the adjacent normal tissue.

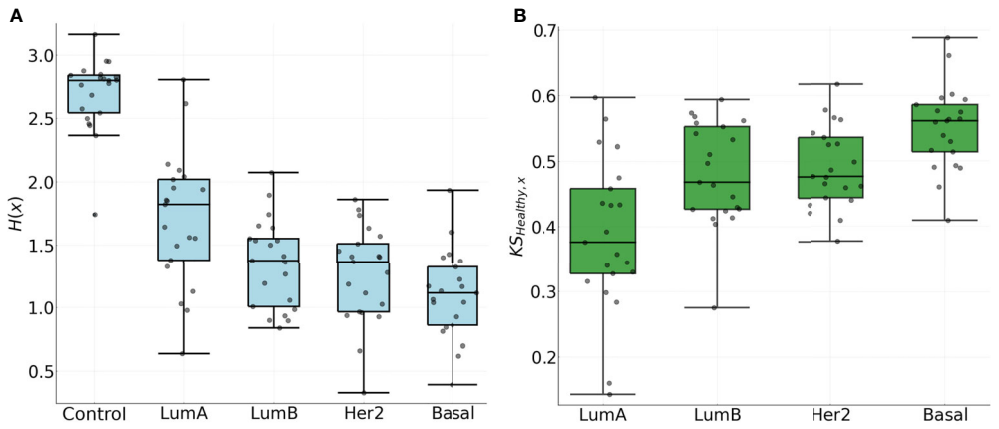

FIGURE 7 | **(A)** Entropy for the NND distribution in all chromosomes. **(B)** Kolmogorov-Smirnov distance between the CDF of the NND in the adjacent normal phenotype and each of the Cancer subtypes for each chromosome.

with respect to the adjacent normal group. These patterns in the genome indicate that in cancer, physically close genes are co-expressed (*cis*- interactions), while for distant genes (*trans*-interactions) the clustered co-expression is, to a large extent, lost.

The piece-wise Kolmogorov-Smirnov tests for all tumor subtypes and adjacent normal tissue versus phenotype-specific null models and their corresponding significance tests (obtained *via* bootstrap/Permutation analysis) (included in the **Supplementary Materials S12**, **S13**), show that correlation values at short distances are much more significant for all chromosomes in any cancer phenotype than the adjacent normal network.

It is worth noticing that the significance of the KS tests also decays with the distance for all chromosomes in any cancer phenotype. Conversely, for the adjacent normal network, distance does not exert a considerable influence in the significance of the KS test. Finally, the KS test also show that the significance of differences between the correlation of our empirical data with the null model is unique for each chromosome and for each phenotype.

The fact that genes are highly co-expressed in groups with close positions, may be due to a favored number of nearby transcription sites or to the strong presence of transcription factors. It has been observed for instance, that in Luminal A breast cancer gene co-expression networks, co-factors, CTCF binding sites (57, 58) or copy number alterations (59, 60), may remodel chromatin making it more or less accessible, thus allowing gene transcription of local neighborhoods, resulting in the concomitant high co-expression between those neighboring genes (53). On the other hand, TFs influence more often inter-chromosome edges, meanwhile intra-chromosome interactions are less affected by them (53).

Physical interactions such as CTCF binding sites have captured attention in recent years (61, 62), and more importantly, in breast cancer (63).

For instance, in (53), we constructed an intra-chromosome gene-co-expression network for Luminal A breast cancer samples. There, a community detection method was performed to determine whether CTCF binding sites appeared in the borders of those communities. We observed that there is no link between CTCF binding sites and the border of intra-chromosome communities. In that sense, we argued that, at least for Luminal A breast cancer gene co-expression networks, CTCF binding sites are not determinant for network structure.

Transcription factor (TF) regulation is, of course, one of the central mechanisms for gene regulation. With respect to the role that TFs may exert on gene clustering, we have previously shown that TFs influence genes in a trans- fashion, i.e. TFs from a given chromosome regulate genes from different chromosomes. We have shown that in terms of Master Regulators in breast cancer (64, 65), but also in Luminal A breast cancer networks (53). Conversely, for intra-chromosome genes, TFs influence is much less evident.

Finally, Copy Number Variations (CNVs) have been considered as a crucial factor in the rise and development of breast cancer (59). In fact, a correlation between CNVs, protein levels and mRNA gene expression has also been reported previously (66). Hence, high correlations between clusters of physically closed genes appear to be related to copy number alterations.

We have used TCGA-derived CNV data and compared the amplification/deletion peaks with LumA network communities. Interestingly, the community with more overexpressed genes, composed of genes such as FOXM1, HJURP, or CENPA, presented large regions of deletions. The apparently contradictory result suggested that the copy number alterations do not influence the structure of that community. On the other hand, a gene community formed by HLA family genes, presented a common pattern of amplification, but those genes were not differentially expressed (53).

With the aforementioned in mind, we argue that CNVs are not as relevant as one could expect in terms of the gene clustering shown here. Moreover, CNVs influence may be at the expression level, but said effect is more limited regarding co-expression. However, further investigation is necessary to clarify these ideas.

The structure of clustered genes in physically close neighborhoods resembles the images obtained by Hi-C methods (67–69).

In recent times, there has been an increased interest as to how chromosome conformation capture experiments such as Hi-C may lead to relevant clues towards our understanding of further effects in connection with transcriptional regulation. Indeed, we are currently conducting research along these lines in our group. Work is ongoing, however, we can advance that there seems to be important correlations between loss/gain of statistically significant chromosomal contacts and co-expression relationships between genes in the associated genomic regions. It remains to be determined however whether said correlations are significant *via* proper assessment of null models and, more importantly, to determine what may be the biological consequences of these associations.

Preliminary findings from our Hi-C analysis in breast cancer indicate that more relevant contacts are mostly (but not exclusively) on close genomic regions. This is not unlike what we have observed with MI-based gene co-expression networks in which there is a preponderance of co-expression interactions in shorter distances for tumors. Future work undoubtedly will focus on the comparison between the network clusters constructed by this method and those from Hi-C. In particular, the zones/genes between gene groups. The assessment and comparison of both structures will provide us more information regarding the structural alterations during the carcinogenic process.

In brief, after revising the evidence about other mechanisms of gene regulation, we may hypothesize that the ultimate cause of the distance-dependent gene clustering is not a single mechanism, but instead, it could be a non-linear combination of different phenomena. In particular, regarding gene clustering, we have evaluated for the first time the whole set of gene interactions, and the loss of long-distance co-expression remains, which is more evident in the most aggressive subtypes.

Homogeneity/redundancy promotes higher entropy. Systems with redundancy are less likely to fail to catastrophic events. In other words, it seems there are mechanisms that give robustness to gene regulation in a control phenotype. It is still uncertain whether the loss of long range (or gain in short range) gene co-expression is a consequence of cancer, forcing the system to work in a less entropic configuration, but it seems that this preference for a less entropic configuration is common in all cancer subtypes and is consistently progressive with subtype aggressiveness.

As a summary of findings, we may establish the following:

- We used tools previously implemented in time series analysis in the stock market and neuroscience settings (49–51) to develop a systematic, data-driven method for intra-chromosomal gene expression clustering. Using spectral decomposition and a null model, we were able to determine the number of co-expressed group of genes to perform *k*–medoids algorithm calculations and determine the most accurate clustering configuration. This method allowed us to have significant results, avoiding to set an *a priori* threshold for co-expression values.
- In the adjacent normal phenotype matrices, negative and positive correlations are significant throughout the entire chromosome. On the other hand, in breast cancer, negative correlations are observed in the same rank than those from a null model (see **Figure 2**); furthermore, the positive ones are only out of the null model cloud over short distances.
- In cancer, clustering mostly occurs between nearby genes, unlike what happens in the adjacent normal phenotype matrices. This is a representation of high co-expression over short distances. This fact coincides and corroborate previous results on mutual information-based co-expression networks in these and other types of cancer (10, 12–14).
- The intra-cluster Nearest Neighbor Distance (NND) clearly decays from the adjacent normal network to those cancerous ones. Additionally, the NND for breast cancer networks also decays according to the aggressiveness of the subtype: Luminal A, Luminal B, HER2+ and Basal.
- Analogously to the last point, Kolmogorov-Smirnov (KS) distance between the Cumulative Distribution Function of the NND in the adjacent normal and each breast cancer subtype network, increases with the aggressiveness of the subtype, thus indicating that the larger value of the KS distance, the larger difference between adjacent normal and breast cancer phenotypes' networks.

Clustered genes may be subject to further analyses to reveal, for instance, statistical enrichment of functional categories revealing certain biological functions, additional patterns of coordinated activity, etc. This in turn may lead to the generation of hypotheses to be tested *via* more narrowly targeted assays and interventions.

A closer look at matrices' patterns generated by other type of sorting methods may shed some light on possible mechanisms behind the regulatory changes in co-expression and perhaps even in the establishment of the tumor phenotypes. This is, indeed, still ongoing work.

Further steps towards the understanding of co-expression patterns and the differences in clustering among adjacent normal and cancerous phenotypes may be also based on the usage of multi-layer approaches (11, 70).

There are remaining questions prompted by this study. For example, while it is evident that there is a decay in the strength of correlations depending on the distance in all chromosomes, it is not fully clear what is the origin of the differences in the slope of the aforementioned decays. Also, the negative correlations in adjacent normal network are significant, independently of the position in the chromosome. Is the anti-correlation between genes a possible mechanism of negative feedback? Is that mechanism disrupted in breast cancer? Another important question regarding the clustering in cancer network is the size of the clusters. Is there an "optimal" cluster

size for cancer networks? If so, what is the rationale behind such number?

Finally, the fact that other types of cancer, which have been analyzed in terms of gene co-expression interactions, such as clear cell renal carcinoma (13), lung adenocarcinoma, or lung squamous cell carcinoma (12) have been reported to have the same bias in short-distance interactions, a remaining question is whether the clustering behavior observed in breast cancer subtype networks is a conserved phenomenon along other cancer types.

The above mentioned questions, together with the acquired knowledge on cancer networks, will be eventually answered and that will bring us with complementary information to have a broader point of view on gene regulation in cancer.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://github.com/josemaz/gene-matrices.

AUTHOR CONTRIBUTIONS

AG-E and JZ-F performed computational analyses, developed methods and implemented programming code, performed pre-processing and low-level data analysis, participated in the writing of the manuscript. EH-L contributed to the design of the study, co-supervised the project, contributed and supervised the writing of the manuscript. JE-E conceived and designed the project, performed calculations, supervised the project, drafted the manuscript. All authors contributed to the article and approved the submitted version. AG-E and JZ-F contributed equally as first author.

FUNDING

This work was supported by the Consejo Nacional de Ciencia y Tecnología [SEP-CONACYT-2016-285544 and FRONTERAS-2017-2115], and the National Institute of Genomic Medicine, México. Additional support has been granted by the Laboratorio Nacional de Ciencias de la Complejidad, from the Universidad Nacional Autónoma de México.

ACKNOWLEDGMENTS

The results shown here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2021.726493/full#supplementary-material

Supplementary Figure 1 | Principal component analyses of RNA-seq data clustered by breast cancer subtypes.

Supplementary Figure 2 | Comparison of the squared components of the four largest eigenvectors from the correlation matrix b in **Figure 1**. It can be seen that there is an overlap between the eigenvectors that does not allow to separate the components into clusters.

Supplementary Figure 3 | Outliers of positive correlations (red dots) in some chromosomes of the breast cancer subtype networks. Plots for chromosomes 1, 9, 19 and 17 for the four subtypes. Arrows indicate sets of positive correlations considered outliers. Notice that the outliers in each plot form almost vertical lines, indicating that those interactions present approximately the same distance. Additionally we can see a null model overlapping (blue).

Supplementary Material S1 | Excel file containing cross tables between subtype-samples and histological variables.

Supplementary Material S2 | Heatmaps of Pearson correlation for each chromosome in the adjacent normal phenotype. The color code is the same than in **Figure 1**.

Supplementary Material S3 | Heatmaps of Pearson correlation for each chromosome in the Luminal A phenotype.

Supplementary Material S4 | Heatmaps of Pearson correlation for each chromosome in the Luminal B phenotype.

Supplementary Material S5 | Heatmaps of Pearson correlation for each chromosome in the HER2+ phenotype.

Supplementary Material S6 | Heatmaps of Pearson correlation for each chromosome in the Basal phenotype.

Supplementary Material S7 to S11 | Pearson distribution scatter plots for normal adjacent tissue, Basal HER2+, Luminal A and Luminal B, respectively. These plots show correlations sorted by gene start position for the four cancer phenotypes and the adjacent normal network per chromosome.

Supplementary Material S12 | Kolmogorov-Smirnov significance tests for all chromosomes in the five phenotypes. As in **Figure 3**, the figures represent the KS statistics of 50 equal-area (same number of data-points) intervals for each chromosome. In the figures, the phenotypes are represented by different colors. The associated p-value for the complete set of correlations are depicted in the upper right part of the figures.

Supplementary Material S13 | Piece-wise permutation p-values of the KS statistics, calculated for all bins obtained in **Supplementary Material S8**, in every chromosomal region for each phenotype.

Supplementary Material S14 | Clusters by chromosome for the five phenotypes, obtained by eigenvalue decomposition and k-medoids method. The figures are depicted as in the manuscript. Additionally, this material contains files for clusters including the name of the gene, the cluster that the gene belong to, the assignment cost function value, the chromosome location of the gene, and the gene start position of said gene.

Supplementary Material S15 | Cumulative distribution for the four cancer phenotypes and the adjacent normal network per chromosome. Color code coincides with clusters in **Supplementary Material S7**. For clarity, only the top-ten clusters were colored.

REFERENCES

1. Siegel RL, Miller KD, Jemal A. Cancer Statistics, 2020. *CA: A Cancer J Clin* (2020) 70:7–30. doi: 10.3322/caac.21590
2. Kittaneh M, Montero AJ, Glück S. Molecular Profiling for Breast Cancer: A Comprehensive Review. *Biomarkers Cancer* (2013) 5:BIC–S9455. doi: 10.4137/BIC.S9455
3. Liu R, Wang X, Chen GY, Dalerba P, Gurney A, Hoey T, et al. The Prognostic Role of a Gene Signature From Tumorigenic Breast-Cancer Cells. *N Engl J Med* (2007) 356:217–26. doi: 10.1056/NEJMoa063994
4. de Anda-Jáuregui G, Velázquez-Caldelas TE, Espinal-Enríquez J, Hernández-Lemus E. Transcriptional Network Architecture of Breast Cancer Molecular Subtypes. *Front Physiol* (2016) 7:568. doi: 10.3389/fphys.2016.00568
5. Perou CM, Sørlie T, Eisen MB, Van De Rijn M, Jeffrey SS, Rees CA, et al. Molecular Portraits of Human Breast Tumours. *Nature* (2000) 406:747–52. doi: 10.1038/35021093
6. Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, et al. Gene Expression Patterns of Breast Carcinomas Distinguish Tumor Subclasses With Clinical Implications. *Proc Natl Acad Sci* (2001) 98:10869–74. doi: 10.1073/pnas.191367098
7. Guedj M, Marisa L, De Reynies A, Orsetti B, Schiappa R, Bibeau F, et al. A Refined Molecular Taxonomy of Breast Cancer. *Oncogene* (2012) 31:1196–206. doi: 10.1038/onc.2011.301
8. Gao GF, Parker JS, Reynolds SM, Silva TC, Wang LB, Zhou W, et al. Before and After: Comparison of Legacy and Harmonized TCGA Genomic Data Commons' Data. *Cell Syst* (2019) 9:24–34.e10. doi: 10.1016/j.cels.2019.06.006
9. Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, et al. Cell-Of-Origin Patterns Dominate the Molecular Classification of 10, 000 Tumors From 33 Types of Cancer. *Cell* (2018) 173:291–304.e6. doi: 10.1016/j.cell.2018.03.022
10. Espinal-Enríquez J, Fresno C, Anda-Jáuregui G, Hernández-Lemus E. RNA-Seq Based Genome-Wide Analysis Reveals Loss of Inter-Chromosomal Regulation in Breast Cancer. *Sci Rep* (2017) 7:1760. doi: 10.1038/s41598-017-01314-1
11. Dorantes-Gilardi R, García-Cortés D, Hernández-Lemus E, Espinal-Enríquez J. Multilayer Approach Reveals Organizational Principles Disrupted in Breast Cancer Co-Expression Networks. *Appl Network Sci* (2020) 5:47. doi: 10.1007/s41109-020-00291-1
12. Andonegui-Elguera SD, Zamora-Fuentes JM, Espinal-Enríquez J, Hernández-Lemus E. Loss of Long Distance Co-Expression in Lung Cancer. *Front Genet* (2021) 12. doi: 10.3389/fgene.2021.625741
13. Zamora-Fuentes JM, Hernández-Lemus E, Espinal-Enríquez J. Gene Expression and Co-Expression Networks Are Strongly Altered Through Stages in Clear Cell Renal Carcinoma. *Front Genet* (2020) 11:578679. doi: 10.3389/fgene.2020.578679
14. García-Cortés D, de Anda-Jáuregui G, Fresno C, Hernández-Lemus E, Espinal-Enríquez J. Gene Co-Expression Is Distance-Dependent in Breast Cancer. *Front Oncol* (2020) 10:1232. doi: 10.3389/fonc.2020.01232
15. Yang Y, Han L, Yuan Y, Li J, Hei N, Liang H. Gene Co-Expression Network Analysis Reveals Common System-Level Properties of Prognostic Genes Across Cancer Types. *Nat Commun* (2014) 5:1–9. doi: 10.1038/ncomms4231
16. Anglani R, Creanza TM, Liuzzi VC, Piepoli A, Panza A, Andriulli A, et al. Loss of Connectivity in Cancer Co-Expression Networks. *PloS One* (2014) 9: e87075. doi: 10.1371/journal.pone.0087075
17. Deng SP, Zhu L, Huang DS. Predicting Hub Genes Associated With Cervical Cancer Through Gene Co-Expression Networks. *IEEE/ACM Trans Comput Biol Bioinf* (2015) 13:27–35. doi: 10.1109/TCBB.2015.2476790
18. Tang J, Kong D, Cui Q, Wang K, Zhang D, Gong Y, et al. Prognostic Genes of Breast Cancer Identified by Gene Co-Expression Network Analysis. *Front Oncol* (2018) 8:374. doi: 10.3389/fonc.2018.00374
19. Liao Y, Wang Y, Cheng M, Huang C, Fan X. Weighted Gene Coexpression Network Analysis of Features That Control Cancer Stem Cells Reveals Prognostic Biomarkers in Lung Adenocarcinoma. *Front Genet* (2020) 11:311. doi: 10.3389/fgene.2020.00311
20. Yu X, Cao S, Zhou Y, Yu Z, Xu Y. Co-Expression Based Cancer Staging and Application. *Sci Rep* (2020) 10:1–10. doi: 10.1038/s41598-020-67476-7
21. Altay G, Emmert-Streib F. Inferring the Conservative Causal Core of Gene Regulatory Networks. *BMC Syst Biol* (2010) 4:1–13. doi: 10.1186/1752-0509-4-132
22. Alcalá-Corona SA, Velázquez-Caldelas TE, Espinal-Enríquez J, Hernández-Lemus E. Community Structure Reveals Biologically Functional Modules in Mef2c Transcriptional Regulatory Network. *Front Physiol* (2016) 7:184. doi: 10.3389/fphys.2016.00184
23. Alcalá-Corona SA, de Anda-Jáuregui G, Espinal-Enríquez J, Hernández-Lemus E. Network Modularity in Breast Cancer Molecular Subtypes. *Front Physiol* (2017) 8:915. doi: 10.3389/fphys.2017.00915
24. de Anda-Jáuregui G, Espinal-Enríquez J, Drago-García D, Hernández-Lemus E. Nonredundant, Highly Connected Micrornas Control Functionality in Breast Cancer Networks. *Int J Genomics* (2018) 2018. doi: 10.1155/2018/9585383
25. Velazquez-Caldelas TE, Alcalá-Corona SA, Espinal-Enríquez J, Hernandez-Lemus E. Unveiling the Link Between Inflammation and Adaptive Immunity in Breast Cancer. *Front Immunol* (2019) 10:56. doi: 10.3389/fimmu.2019.00056
26. Liesecke F, De Craene JO, Besseau S, Courdavault V, Clastre M, Vergès V, et al. Improved Gene Co-Expression Network Quality Through Expression Dataset Down-Sampling and Network Aggregation. *Sci Rep* (2019) 9:1–16. doi: 10.1038/s41598-019-50885-8
27. Alcalá-Corona SA, Espinal-Enríquez J, De Anda Jáuregui G, Hernandez-Lemus E. The Hierarchical Modular Structure of Her2+ Breast Cancer Network. *Front Physiol* (2018) 9:1423. doi: 10.3389/fphys.2018.01423
28. Serrano MA, Boguna M, Vespignani A. Extracting the Multiscale Backbone of Complex Weighted Networks. *Proc Natl Acad Sci* (2009) 106:6483–8. doi: 10.1073/pnas.0808904106
29. Perkins AD, Langston MA. Threshold Selection in Gene Co-Expression Networks Using Spectral Graph Theory Techniques. *BMC Bioinf* (2009) 10. doi: 10.1186/1471-2105-10-s11-s4
30. Tieri P, Farina L, Petti M, Astolfi L, Paci P, Castiglione F. Network Inference and Reconstruction in Bioinformatics. *Encyclopedia Bioinf Comput Biol (Elsevier)* (2019) 2:805–13. doi: 10.1016/b978-0-12-809633-8.20290-2
31. Kimura S, Fukutomi R, Tokuhisa M, Okada M. Inference of Genetic Networks From Time-Series and Static Gene Expression Data: Combining a Random-Forest-Based Inference Method With Feature Selection Methods. *Front Genet* (2020) 11:595912. doi: 10.3389/fgene.2020.595912
32. de Anda-Jáuregui G, Alcalá-Corona SA, Espinal-Enríquez J, Hernández-Lemus E. Functional and Transcriptional Connectivity of Communities in Breast Cancer Co-Expression Networks. *Appl Network Sci* (2019) 4:22. doi: 10.1007/s41109-019-0129-0
33. Dorantes-Gilardi R, García-Cortés D, Hernández-Lemus E, Espinal-Enríquez J. K-Core Genes Underpin Structural Features of Breast Cancer. *Sci Rep* (2021) 11:16284. doi: 10.1038/s41598-021-95313-y
34. Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): An Immeasurable Source of Knowledge. *Contemp Oncol (Poznan Poland)* (2015) 19:A68–77. doi: 10.5114/wo.2014.47136
35. Fresno C, González GA, Merino GA, Flesia AG, Podhajcer OL, Llera AS, et al. A Novel Non-Parametric Method for Uncertainty Evaluation of Correlation-Based Molecular Signatures: Its Application on PAM50 Algorithm. *Bioinf (Oxford England)* (2017) 33:693–700. doi: 10.1093/bioinformatics/btw704
36. Fresno C, González GA, Llera AS, Fernández EA. Pbcmc: Permutation-Based Confidence for Molecular Classification. *R Package version* (2016) 1 2. doi: 10.18129/B9.bioc.pbcmc
37. Nueda MJ, Ferrer A, Conesa A. ARSyN: A Method for the Identification and Removal of Systematic Noise in Multifactorial Time Course Microarray Experiments. *Biostatistics (Oxford England)* (2012) 13:553–66. doi: 10.1093/biostatistics/kxr042
38. (2021). Available at: https://github.com/josemaz/gene-matrices/blob/master/Notebooks/CorrelationVsDistance.ipynb.
39. Vinayak, Prosen T, Buča B, Seligman TH. Spectral Analysis of Finite-Time Correlation Matrices Near Equilibrium Phase Transitions. *Epl* (2014) 10820006. doi: 10.1209/0295-5075/108/20006
40. Vinayak V, Seligman TH. Time Series, Correlation Matrices and Random Matrix Models. *AIP Conf Proc* (2014) 1575:196–217. doi: 10.1063/1.4861704
41. Gopikrishnan P, Rosenow B, Plerou V, Stanley HE. Quantifying and Interpreting Collective Behavior in Financial Markets. *Phys Rev E - Stat Physics Plasmas Fluids Related Interdiscip Topics* (2001) 64:4. doi: 10.1103/PhysRevE.64.035106
42. Luo F, Zhong J, Yang Y, Zhou J. Application of Random Matrix Theory to Microarray Data for Discovering Functional Gene Modules. *Phys Rev E - Stat Nonlinear Soft Matter Phys* (2006) 73:1–5. doi: 10.1103/PhysRevE.73.031924

43. Fossion R. A Time-Series Approach to Dynamical Systems From Classical and Quantum Worlds. *AIP Conf Proc* (2014) 1575:89–110. doi: 10.1063/1.4861700
44. Fossion R, Vargas GT, Vieyra JC. Random-Matrix Spectra as a Time Series. *Phys Rev E - Stat Nonlinear Soft Matter Phys* (2013) 88:1–4. doi: 10.1103/PhysRevE.88.060902
45. Laloux L, Cizeau P, Bouchaud JP, Potters M. Noise Dressing of Financial Correlation Matrices. *Phys Rev Lett* (1999) 83:1467–70. doi: 10.1103/PhysRevLett.83.1467
46. Zhong J, Gao H, Thompson DK, Luo F, Zhou J, Khan L, et al. Constructing Gene Co-Expression Networks and Predicting Functions of Unknown Genes by Random Matrix Theory. *BMC Bioinf* (2007) 8:299. doi: 10.1186/1471-2105-8-299
47. Rummel C, Müller M, Baier G, Amor F, Schindler K. Analyzing Spatio-Temporal Patterns of Genuine Cross-Correlations. *J Neurosci Methods* (2010) 191:94–100. doi: 10.1016/j.jneumeth.2010.05.022
48. Müller M, Jiménez YL, Rummel C, Baier G, Galka A, Stephani U, et al. Localized Short-Range Correlations in the Spectrum of the Equal-Time Correlation Matrix. *Phys Rev E - Stat Nonlinear Soft Matter Phys* (2006) 74:041119. doi: 10.1103/PhysRevE.74.041119
49. Utsugi A, Ino K, Oshikawa M. Random Matrix Theory Analysis of Cross Correlations in Financial Markets. *Phys Rev E - Stat Physics Plasmas Fluids Related Interdiscip Topics* (2004) 70:11. doi: 10.1103/PhysRevE.70.026110
50. Plerou V, Gopikrishnan P, Rosenow B, Amaral LA, Guhr T, Stanley HE. Random Matrix Approach to Cross Correlations in Financial Data. *Phys Rev E - Stat Physics Plasmas Fluids Related Interdiscip Topics* (2002) 65:1–18. doi: 10.1103/PhysRevE.65.066126
51. Rummel C. Quantification of Intra- and Inter-Cluster Relations in Nonstationary and Noisy Data. *Phys Rev E - Stat Nonlinear Soft Matter Phys* (2008) 77:016708. doi: 10.1103/PhysRevE.77.016708
52. Rummel C, Müller M, Schindler K. Data-Driven Estimates of the Number of Clusters in Multivariate Time Series. *Phys Rev E - Stat Nonlinear Soft Matter Phys* (2008) 78:1–12. doi: 10.1103/PhysRevE.78.066703
53. García-Cortés D, Hernández-Lemus E, Espinal-Enríquez J. Luminal a Breast Cancer Co-Expression Network: Structural and Functional Alterations. *Front Genet* (2021) 12:629475. doi: 10.3389/fgene.2021.629475
54. Marchenko VA, PL A. Distribution of Eigenvalues for Some Sets of Random Matrices. *Mat Sb. (N.S.)* (1967) 1:457–83. doi: 10.1070/SM1967v001n04ABEH001994
55. Kennecke H, Yerushalmi R, Woods R, Cheang MCU, Voduc D, Speers CH, et al. Metastatic Behavior of Breast Cancer Subtypes. *J Clin Oncol* (2010) 28:3271–7. doi: 10.1200/JCO.2009.25.9820
56. Fallahpour S, Navaneelan T, De P, Borgo A. Breast Cancer Survival by Molecular Subtype: A Population-Based Analysis of Cancer Registry Data. *CMAJ Open* (2017) 5:E734. doi: 10.9778/cmajo.20170030
57. Achinger-Kawecka J, Valdes-Mora F, Luu PL, Giles KA, Caldon CE, Qu W, et al. Epigenetic Reprogramming at Estrogen-Receptor Binding Sites Alters 3d Chromatin Landscape in Endocrine-Resistant Breast Cancer. *Nat Commun* (2020) 11:1–17. doi: 10.1038/s41467-019-14098-x
58. Corces MR, Corces VG. The Three-Dimensional Cancer Genome. *Curr Opin Genet Dev* (2016) 36:1–7. doi: 10.1016/j.gde.2016.01.002
59. Inaki K, Menghi F, Woo XY, Wagner JP, Jacques PÉ, Lee YF, et al. Systems Consequences of Amplicon Formation in Human Breast Cancer. *Genome Res* (2014) 24:1559–71. doi: 10.1101/gr.164871.113
60. Myhre S, Lingjærde OC, Hennessy BT, Aure MR, Carey MS, Alsner J, et al. Influence of DNA Copy Number and mRNA Levels on the Expression of Breast Cancer Related Proteins. *Mol Oncol* (2013) 7:704–18. doi: 10.1016/j.molonc.2013.02.018
61. Achinger-Kawecka J, Clark SJ. Disruption of the 3D Cancer Genome Blueprint. *Epigenomics* (2017) 9:47–55. doi: 10.2217/epi-2016-0111
62. Pugacheva EM, Kubo N, Loukinov D, Tajmul M, Kang S, Kovalchuk AL, et al. Ctcf Mediates Chromatin Looping *via* N-Terminal Domain-Dependent Cohesin Retention. *Proc Natl Acad Sci* (2020) 117:2030–31. doi: 10.1073/pnas.1911708117
63. Fiorito E, Sharma Y, Gilfillan S, Wang S, Singh SK, Satheesh SV, et al. Ctcf Modulates Estrogen Receptor Function Through Specific Chromatin and Nuclear Matrix Interactions. *Nucleic Acids Res* (2016) 44:10588–602. doi: 10.1093/nar/gkw785
64. Tovar H, García-Herrera R, Espinal-Enríquez J, Hernández-Lemus E. Transcriptional Master Regulator Analysis in Breast Cancer Genetic Networks. *Comput Biol Chem* (2015) 59:67–77. doi: 10.1016/j.compbiolchem.2015.08.007
65. Tapia-Carrillo D, Tovar H, Velazquez-Caldelas TE, Hernandez-Lemus E. Master Regulators of Signaling Pathways: An Application to the Analysis of Gene Regulation in Breast Cancer. *Front Genet* (2019) 10:1180. doi: 10.3389/fgene.2019.01180
66. Lachmann A. PhD Thesis, Columbia University, New York *Confounding Effects in Gene Expression and Their Impact on Downstream Analysis (Columbia University)*. [PhD Thesis]. New York: Columbia University (2016).
67. Soler-Oliva ME, Guerrero-Martínez JA, Bachetti V, Reyes JC. Analysis of the Relationship Between Coexpression Domains and Chromatin 3d Organization. *PloS Comput Biol* (2017) 13:e1005708. doi: 10.1371/journal.pcbi.1005708
68. Varrone M, Nanni L, Ciriello G, Ceri S. Exploring Chromatin Conformation and Gene Co-Expression Through Graph Embedding. *Bioinformatics* (2020) 36:i700–8. doi: 10.1093/bioinformatics/btaa803
69. Beesley J, Sivakumaran H, Marjaneh MM, Lima LG, Hillman KM, Kaufmann S, et al. Chromatin Interactome Mapping at 139 Independent Breast Cancer Risk Signals. *Genome Biol* (2020) 21:1–19. doi: 10.1186/s13059-019-1877-y
70. Ochoa S, de Anda-Jáuregui G, Hernández-Lemus E. An Information Theoretical Multilayer Network Approach to Breast Cancer Transcriptional Regulation. *Front Genet* (2021) 12:232. doi: 10.3389/fgene.2021.617512

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.