Applied Network Science

RESEARCH Open Access

Multilayer approach reveals organizational principles disrupted in breast cancer co-expression networks

Rodrigo Dorantes-Gilardi[1], Diana García-Cortés[1], Enrique Hernández-Lemus[1,2] and Jesús Espinal-Enríquez[1,2*]

*Correspondence:
jespinal@inmegen.gob.mx
[1]Department of Computational Genomics, National Institute of Genomic Medicine, Periférico sur 4809, Arenal Tepepan, Mexico City, Mexico
[2]Centro de Ciencias de la Complejidad (C3), Universidad Nacional Autónoma de México, Circuito de la Investigación Científica S/N, 04510 Mexico City, Mexico

Abstract

The study of co-expression programs in the context of cancer can help to elucidate the genetic mechanisms that are altered and lead to the disease. The identification of gene co-expression patterns, unique to healthy profiles (and absent in cancer) is an important step in this direction. Networks are a good tool for achieving this as they allow to model local and global structural properties of the gene co-expression program. This is the case of gene co-expression networks (GCNs), where nodes or vertices represent genes and an edge between two nodes exists if the corresponding genes are co-expressed. Single threshold co-expression networks are often used for this purpose. However, important interactions in a broader co-expression space needed to unravel such mechanisms may be overlooked. In this work, we use a multilayer network approach that allows us to study co-expression as a discrete object, starting at weak levels of co-expression building itself upward towards the top co-expressing gene pairs. We use a multilayer GCNs (or simply GCNs), to compare healthy and breast cancer co-expression programs. By using the layers of the gene co-expression networks, we were able to identify a structural mechanism unique in the healthy GCN similar to well-known preferential attachment. We argue that this mechanism may be a reflection of an organizational principle that remains absent in the breast cancer co-expression program. By focusing on two well-defined set of nodes in the top co-expression layers of the GCNs—namely hubs and nodes in the main core of the network—we found a set of genes that is well conserved across the co-expression program. Specifically, we show that nodes with high inter-connectedness as opposed to high connectedness are conserved in the healthy GCN. This set of genes, we discuss, may partake in several different functional pathways in the regulatory program. Finally, we found that breast cancer GCN is composed of two different structural mechanisms, one that is random and is composed by most of the co-expression layers, and another non-random mechanism found only in the top co-expression layers.
Overall, we are able to construct within this approach a portrait of the whole transcriptome co-expression program, thus providing a novel manner to study this complex biological phenomenon.

Keywords: Gene co-expression network, Multilayer network, Breast cancer, *k*-core

Introduction

Identifying the mechanisms underlying gene regulation is paramount to understand the system-wide development and functioning of the cell. Particularly in cancer, it can unravel the regulatory causes of cell malfunction (Hernández-Lemus et al. 2019). Similarly, the detection of regulatory patterns exclusive to healthy cells is a main topic of interest in cancer research.

Thanks to accessible high-throughput data, it is now possible to study whole-genome gene expression and to represent genetic profiles of different tissue cells (Weinstein et al. 2013). For a given tissue and the genetic profiles of several samples we can then calculate the co-expression values of each pair of genes using a measure of statistical dependence (Margolin et al. 2006a; Hernández-Lemus et al. 2009; Marbach et al. 2012). It follows that highly co-expressed genes are likely related to a biological activity, while the rest of co-expressing values could represent systemic noise lacking any biological interpretation.

In this context, a gene co-expression network (GCN) is a fit theoretical framework to study the mechanisms of co-expression in a cell, and can be used as a model of inference to the more general gene regulatory program (Marbach et al. 2012). In the specific case of cancer, GCNs can be used to classify structural and functional properties of healthy and cancer expression profiles to differentiate their co-expression mechanisms (Yang et al. 2014; de Anda-Jáuregui et al. 2016; Tovar et al. 2015; de Anda-Jáuregui et al. 2019; Madhamshettiwar et al. 2012; Liu et al. 2015).

Un-weighted GCNs are defined by a single co-expression threshold value that is selected to be astringent in order to capture only strongly co-expressed gene pairs—the most likely interactions to be related at a biological level (Marbach et al. 2012). This approach neglects most of the interactions between co-expressing gene pairs, including those overlooked at lower levels of co-expression that may be biologically relevant (Zhang and Horvath 2005). Indeed, even if the actual gene interactions were known, these would likely be scattered across a long range of co-expression values. In order to overcome this limitation, the approach of a multilayer network results appealing (Boccaletti et al. 2014; Kivelä et al. 2014). Particularly, in the context of cancer, the analysis of a whole-transcriptome network divided by layers of co-expression may unveil differences between a healthy and a cancerous phenotype at several levels-of-depth, and thus provide alternative manners for studying the disease.

By integrating the ten thousand highest co-expression interactions into a single-layer GCN, previous works have showed different interesting inter/intra-chromosomal connectivity patterns between breast cancer and healthy networks. Moreover, the inter-chromosomal connectivity was close to random mixing in healthy tissues, as opposed to breast cancer where co-expressing genes shared almost exclusively the same chromosome (Espinal-Enriquez et al. 2017; de Anda-Jáuregui et al. 2019; de Anda-Jáuregui et al. 2019).

The fact that connectivity patterns differ at the top co-expression layer in GCNs of breast cancer and healthy tissues, prompts the question of to what extent this holds true in other co-expression layers, and if connectivity indeed differs at a multilayer scale, is there a structural co-expression mechanism unique to the layers of the healthy GCN that is lost in breast cancer? The answers of these questions, are central issues in current research on co-expression networks with possible implications in identifying the mechanisms underlying the development of cancer.

To tackle these questions, we assembled the multilayer gene co-expression networks (GCNs, for simplicity) of Basal-like breast cancer (one of the most aggressive breast cancer subtypes), and healthy tissues (Fig. 1). First, we compared co-expression layers in terms of their structural properties, namely their degree mean and standard deviation, and their attribute and degree assortativity coefficients (Methods). Second, by means of two different sets of well-defined nodes (hubs and the main k-core of the network), we obtained the most relevant genes in terms of structural development in both networks across all layers. We then observed the conservation rate of those genes across layers, and finally compared the results between both phenotypes.

We found that inter-chromosomal connectivity is almost constant and close to random mixing across the layers of the healthy GCN. On the other hand, the inter-chromosomal connectivity in breast cancer has an abrupt change around the top few layers. This contrasts with the structural development of the layers in both networks, where the healthy GCN has a higher variation from random early on in its lower co-expression layers, and breast cancer shows a close-to-random structure up until its top 5 layers, where structural values change abruptly.

In the healthy GCN, degree standard deviation and assortative mixing are both sensitive to a subtle decrease in inter-chromosomal connectivity. This sensitivity is accelerated across layers of strong co-expression, suggesting that a set of nodes could be gaining edges faster than others across layers. Indeed highly-connected nodes (hubs) at the top co-expression layer separate rapidly from the rest of genes in terms of degree across layers, similar to a preferential attachment mechanism.

This structural mechanism is not present in the Basal GCN, where hubs from the top layer do not differentiate from the rest of the nodes in the lower co-expression layers in terms of degree. However, top-layer hubs do differentiate within the top five co-expression layers, thus showing a contrast in the co-expression distribution between these top layers and the rest in the Basal GCN, as previously noted.

Interestingly, in the healthy GCN highly inter-connected nodes (nodes in the main core of the top layer), gain connections faster than hubs and are furthermore well conserved across cores in subsequent layers. This high rate of conservation of top-layer core nodes in spite of the fact that edges belong to exactly one layer, lead us to argue that these genes may be a reflection of the existence of an organizational principle in the gene co-expression program. This set of nodes may apparently link other genes across several pathways at different levels of co-expression.

The dramatic loss of said principles in the cancer phenotype, as well as the intra-chromosomal structure of the core nodes in the top layer, could indicate that the aforementioned organizational principles of gene co-expression favor long-distance correlations, leaving short-distance (intra-chromosomal) as the main source of interactions in the cancerous phenotype.

With this approach we were able to observe that a larger co-expression space from classical approaches is important as it contains relevant information to a deeper understanding of the co-expression program of a cell. Additionally, we show that core genes maintain the overall co-expression landscape in the healthy phenotype, whereas this is not the case in the breast cancer network. The loss of organizational principles in cancer could

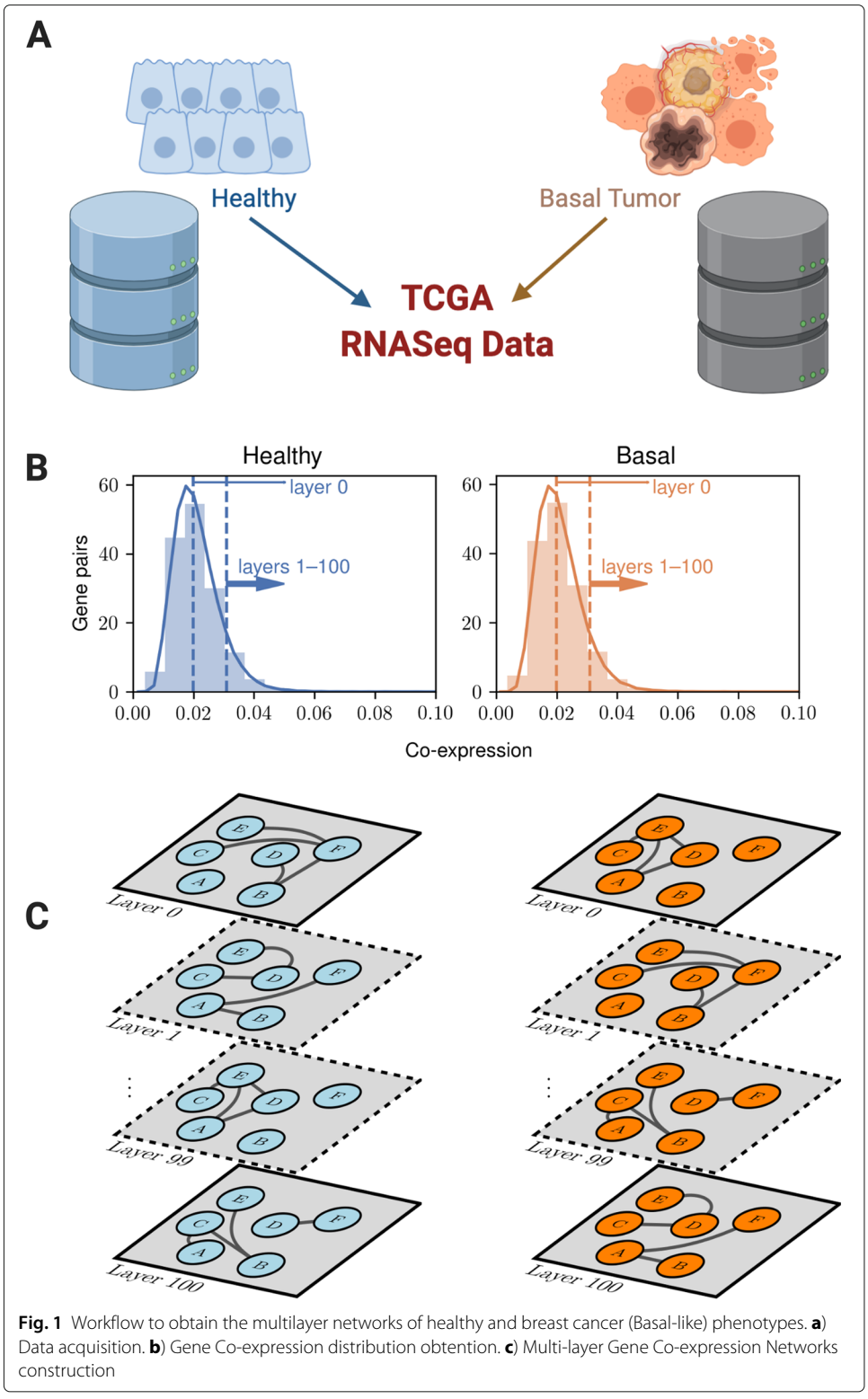

Fig. 1 Workflow to obtain the multilayer networks of healthy and breast cancer (Basal-like) phenotypes. **a**) Data acquisition. **b**) Gene Co-expression distribution obtention. **c**) Multi-layer Gene Co-expression Networks construction

then trigger further hallmarks of cancer. This is an instance of how a network-theory-based approach may help to elucidate the underlying biological mechanisms behind complex phenotypes.

Methods

Workflow

The general data workflow to obtain the multilayer gene co-expression networks is shown in Fig. 1.

Data acquisition

A collection of The Cancer Genome Atlas (TCGA) breast invasive carcinoma datasets were used in this work (Tomczak et al. 2015). The datasets analysed for this study can be found in the GDC repository, `https://portal.gdc.cancer.gov/repository`. The samples belong to the TCGA-BRCA project and were downloaded on January 9th, 2019. Briefly, 105 Basal breast-tumor and 113 normal IlluminaHiSeq RNASeq samples were acquired and pre-processed. The log_2-normalized gene expression values were obtained as described in Espinal-Enriquez et al. (2017). We decided to focus on Basal subtype because it is usually the most aggressive and has the poorest prognosis; additionally, it exhibites the most remarkable differences relative to the healthy phenotype (García-Cortés et al. 2020; de Anda-Jáuregui et al. 2019).

Network construction

Gene regulatory network deconvolution from experimental data has been extensively used to unveil co-regulatory interactions between genes by looking out for patterns in their experimentally-measured mRNA expression levels. A number of correlation measures have been used to deconvolute transcriptional interaction networks based on the inference of the corresponding statistical dependency structure in the associated gene expression patterns (Nielsen and Jensen 2009; Friedman et al. 2000; Wang et al. 2005; Emmert-Streib et al. 2012; Hernández-Lemus and Rangel-Escareño 2011). It has long been known that the maximum likelihood estimator of statistical dependency is mutual information (MI) (Emmert-Streib et al. 2012; Hernández-Lemus and Rangel-Escareño 2011; Margolin et al. 2006b; Basso et al. 2005). ARACNE (Margolin et al. 2006) is a widely used algorithm that quantifies the degree of statistical dependence between pairs of genes. In a nutshell, the algorithm calculates the *Mutual Information (MI)*—a non-parametric measure that captures non-linear dependencies between variables—in a relatively fast implementation.

ARACNE was executed using the following parameters: p-value $= 1$, data processing inequality (DPI) tolerance = 1, MI threshold = 0; to keep all interactions. The fixed bandwidth algorithm was used setting the kernel width to be determined by the program. For the Basal network, the kernel width was set to 0.143809 and for the healthy network to 0.169105.

Gene co-expression network

The advancement of deep-sequencing technologies has made the characterization of whole-genome transcriptomic profiles, and their accessibility a current reality (Wang et al. 2009). Here, we use the transcriptomic profile data from healthy and Basal breast cancer tissues to compute the co-expression values of all gene pairs to obtain their co-expression programs.

A gene co-expression network (GCN) is a network $G = (V, E, l)$ where genes are represented by the set of nodes V and two nodes u and v are connected if their co-

expression value (measured by mutual information) is within a given semi-closed real interval $l = [a, b)$. Let MI be the function assigning each pair of genes its mutual information, and let u and v be two genes. We have that if $MI(u, v) \in (a, b)$ then u and v are connected by an edge in G. The set of all such edges E represents the co-expression of genes within a layer l, and V is the set of genes. The co-expression value of two genes is taken as the mutual information of their expression across sample data.

Here, we consider 101 equally-sized co-expression intervals containing 0.1% of the co-expression values each. The top 100 layers are taken from an equal-size partition of the top 10% of the co-expression distribution. Layer 0 corresponds to the interval between the 50–50.1% of the ordered co-expression distribution, that is it corresponds to taking all the values between the 50 and 50.1 percentile of the distribution and accounts to noise in the distribution as explained below. All layers have thus the same number of interactions but may vary in the number of nodes.

The multilayer approach

Usually un-weighted GCNs are defined by a co-expression threshold that sets the number of edges (Marbach et al. 2012). In general, this limits the value space of co-expression studied to less than the top 0.1% of co-expressing gene pairs and can neglect useful information about the co-expression program. In order to circumvent this limitation, we adapt the GCN to be so-called multi-layered (Boccaletti et al. 2014; Kivelä et al. 2014). Specifically, we consider 100 subsequent layers in the top 10% values of co-expression where each has around 100,000 non-overlapping edges, and a same-size layer belonging to the median of the co-expression distribution. We take the $90th$ percentile as a lower bound to limit the assumptive noise found at lower levels of co-expression. However, we let explicitly a layer around the median of the distribution to account for this noise (see below for a full explanation). Finally, layers are numbered from 0 to 100 by increasing order of co-expression (0 being the layer around the median).

Using layers we then consider the distribution as a discrete mathematical object, where each layer is mapped to an interval of the co-expression distribution. The reason why we do not consider a weighted network is to treat the distribution as discrete. Each layer encapsules a subset of gene-pairs with similar co-expression level, thus considering implicitly its weight in the distribution (for a comprehensive review on weighted networks see Barthélemy et al. (2005)).

Structural parameters

To evaluate similarities and differences between GCN layers and between phenotypes, we calculated the gene (node) mean degree (mean number of neighbors of nodes), the degree standard deviation, and the degree assortativity coefficient (DAC).

Since the starting point of the comparison between healthy and breast-cancer networks is their different inter-chromosomal connectivity, we additionally calculated the attribute assortativity coefficient (AAC). An $AAC = 1$ means that all interactions of any given gene are intra-chromosomal, an $AAC = -1$ imposes that all connections are inter-chromosomal, and an $AAC = 0$ suggests a random mixing.

Median layer as noise

In order to justify our use of the bottom layer (representing the median of the distribution) as noise, we tested the structural parameters of its network to those of random networks preserving the same inter/intra-chromosomal connectivity. To do so, we computed the DAC, AAC, size, mean and standard deviation of the degree of 1000 random networks using stochastic block models. These random networks have for each pair of chromosomes, the same proportion of edges as in the network of layer 0.

We only considered the parameters that showed a Gaussian distribution (using the Shapiro–Wilk test Shapiro and Wilk (1965) implemented in Seabold and Perktold (2010)) among the random networks, namely: DAC, AAC, and the mean and standard deviation of the degree. The size of the network is omitted in this case, as it does not show a normal distribution.

For each parameter, we then performed a Z-test with $\alpha = 0.05$ in which the null hypothesis is that the parameter of the network of the bottom layer is not likely to be part of the random distribution. If the hypothesis is rejected, then the parameter is likely to be part of the random distribution. This happens when the Z-score of the real parameter (Z_0) respects $1.97 > Z_0$ (where 1.97 is the Z-score of corresponding to α). All four parameters tested rejected the hypothesis, therefore suggesting that the parameters were likely to belong to a random network. The standard deviation of the degree has a Z-score of 1.4, the greatest Z-score of all (supplementary Figure S1). Moreover, the hypothesis was tested in both the healthy and the cancer bottom layers confirming that median co-expression levels represent noise in both tissues, as expected.

Stochastic block model

As chromosome connectivity appears as the fundamental characteristic that differs between phenotypes, for the assessment of structural parameters of both sets of GCNs, we constructed a null model usin the stochastic block model with the following features:

Given an ordered list of 23 chromosomes $(1, 2, \ldots, 23)$, and a GCN $G = (V, E, c)$, we can define the chromosome density matrix D of G as :

$$D_{ij} = \frac{|N_{ij}|}{\binom{|i|+|j|}{2}}. \tag{1}$$

Where N_{ij} is the set of neighbors between chromosomes i and j and $|N_{ij}|$ its the number of neighbors between them, and $\binom{|i|+|j|}{2}$ is the maximum number of potential neighbors between i and j.

In order to construct a random network with the same *cis-* gene-pairs ratio as a given GCN, we use the stochastic block model (Holland et al. 1983; Nowicki and Snijders 2001). For our case, each block is a chromosome and edges between blocks have probability equal to the number of edges between the two blocks over the total number of edges in the GCN.

k-core and hub genes

After calculating the structural parameters of all layers in both phenotypes, we obtained the most relevant genes for each layer of GCNs by two different approaches: hub genes (degree) per layer and k-core genes. Specifically, here we define a hub such that its degree is greater than the mean degree of the network plus two standard deviations. On the other

hand, the *k-core* is defined as follows: given a network $G = (V, E)$ with set of nodes V and edges E, a k-core is a maximal subgraph of G such that every node in the subgraph has degree at least k.

To observe whether or not k-core and hub genes are preserved through the different layers in both phenotypes, we defined the cumulative conservation rate (*ccr*). The *ccr* of the top layer (ccr_{100}) is always equal to 1, as both hubs and the core are taken from layer 100. Let S_i be the set of nodes in the group of interest (hubs or nodes in the core) of layer i, we define $ccr_{99} = |S_{99} \cap S_{100}|/|S_{100}|$, where '$|S|$' denotes the number of nodes in set S. In general, the cumulative conservation rate of layer $i \in \{98, 97, \ldots 1, 0\}$ is defined as:

$$ccr_i = \frac{|S_{100} \cap S_{99} \cdots \cap S_{i+1} \cap S_i|}{|S_{100}|}.$$

For any two layers i, j, where $j < i$, we have that $ccr_j \leq ccr_i$. The slower the decay of *ccr* from the top layer to the bottom (noise) layer, the more conserved is the initial group of nodes. The conservation of a set of nodes across co-expression layers, could indicate its relevance in cell regulatory processes at different co-expression scales, and possibly in several functional pathways.

In the case of the k-core nodes, for each tissue, we also computed the *ccr* values of 1000 random multi-layer networks, in which each layer is a stochastic block model of the real layer. To compare the *ccr* values of the real network to the control network, we first obtained the mean *ccr* for each layer in the random networks, call it ccr_r and then calculated the average mean square error between ccr_r and each of the random networks. The mean square error between two same-size distributions $\{a\}_n$ and $\{g\}_n$ is given by:

$$\frac{1}{n}\sum_{i=1}^{n}\sqrt{(g_i - a_i)^2}.$$

Therefore, the average mean square error between a distribution $\{a\}_n$ and k distributions $\{g_k\}_n$ is obtained with:

$$\frac{1}{kn}\sum_{i=1}^{k}\sum_{j=1}^{n}\sqrt{(a_j - g_{ji})^2} \tag{2}$$

Degree of hubs and cores

To compare the degrees of highly-connected nodes in layer 100 across other layers to that of a random group, we (*i*) randomly picked a same-size set of nodes (once for core nodes and once for hubs), (*ii*) computed their degree in all layers and (*iii*) repeated 1000 times. To test weather highly-connected nodes were comparable to random nodes, we used a Z-test (for $p = 0.05$) in which we defined the hypothesis that highly-connected nodes and random nodes were likely to have the same degrees across layers other than layer 1000.

Assortativity

The assortativity measures used here are taken from the definitions by Newman (2002) and implemented in the python library networkx (Hagberg et al. 2008). One measure (degree assortativity) is defined for continuous attributes and the other (attribute assortativity) for categorical ones.

For a given network, define D as the set of all degrees found in the network. For any two different degrees $d_1, d_2 \in D$, let $e_{d_1 d_2}$ be the proportion of edges connecting two nodes with degrees d_1 and d_2. Finally, for each degree $d \in D$ let a_d be the proportion of incident edges to a node with degree d. The definition of the degree assortativity DAC is as follows:

$$\mathrm{DAC} = \frac{1}{\sigma_a^2} \sum_{i,j \in D} ij(e_{ij} - a_i a_j) \tag{3}$$

Where σ_a is the standard deviation of the distribution $\{a_d\}_{d \in D}$. The attribute assortativity coefficient (AAC) can be defined using the same notation in the case when D is a set of categorical vertex attributes as follows:

$$\mathrm{AAC} = \frac{\sum_{i \in D} e_{ii} - a_i^2}{1 - \sum_{i \in D} a_i^2} \tag{4}$$

Code

All code was written in Python using libraries `numpy`, `pandas`, `sklearn`, and `networkx` (Hagberg et al. 2008; McKinney 2012; Van Der Walt et al. 2011; Pedregosa et al. 2011). The code can be found at `https://github.com/CSB-IG/pychromnet`.

Results and discussion

Inter-chromosomal connectivity varies differently between healthy and cancer GCNs

Previous works have used GCNs to reveal topological and functional differences between cancer and healthy genetic profiles (Espinal-Enriquez et al. 2017; Alcalá-Corona et al. 2017; de Anda-Jáuregui et al. 2019). The work by Teschendorff and collaborators (West et al. 2012; Teschendorff and Severini 2010) for instance, has presented how information theoretical entropies based on local probability measures may be able to discern some functional features. Their results show that cancer-related networks are likley characterized by higher information theoretical entropies, possibly due to graph-structure configurational contributions. One remarkable topological difference found between healthy and breast cancer GCNs lies in terms of inter/intra-chromosomal connectivity (Espinal-Enriquez et al. 2017; García-Cortés et al. 2020; de Anda-Jáuregui et al. 2019; de Anda-Jáuregui et al. 2019). Edges in breast cancer GCNs seldom connect nodes between two different chromosomes while edges in the healthy GCN mostly link nodes in different chromosomes.

As previously mentioned, inter-chromosomal connectivity can be measured using the attribute assortativity coefficient (AAC), where a value of 1 is found in networks never connecting nodes in two different chromosomes, -1 in networks when only nodes in different chromosomes are connected, and 0 when the mixing is random (see Methods for a formal definition of assortativity). Single-layer breast cancer GCN has a very low inter-chromosomal connectivity (AAC close to 1) while single-layer healthy GCN has an expected or random inter-chromosomal connectivity (AAC close to 0).

As we can see in Fig. 2, the GCNs belonging to Basal (breast cancer) and healthy tissues exhibit different inter-chromosomal connectivity patterns across co-expression layers. We see a steady decrease in inter-chromosomal connectivity in the GCN of Basal tissue (as shown by a steady increase in AAC), while healthy GCN presents an almost constant close to random (0-valued AAC) inter-chromosomal connectivity throughout the layers.

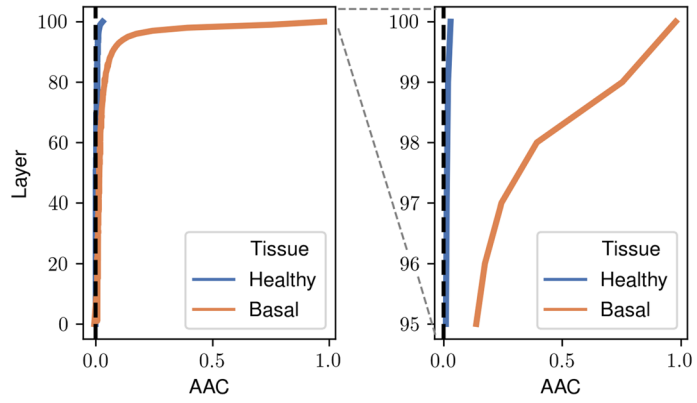

Fig. 2 Inter/intra-chromosomal connectivity as measured by the attribute assortativity coefficient (AAC, where a value close to 1 imply weak inter-chromosomal connectivity and values close to 0 imply random-like inter-chromosomal connectivity) of healthy and breast cancer (Basal subtype) gene co-expression networks. Left plot shows all 101 co-expression layers while right plot zooms-in on the top 6 layers

On average, AAC equals 0.004 and 0.05 in healthy and Basal GCNs, i.e. Basal GCN presents a much weaker inter-chromosomal connectivity as compared to healthy tissue. To discern the extent of their AAC deviation from random mixing, we calculated the mean square error (MSE) from 0 of AAC values. Healthy GCN has a very small MSE (3×10^{-5}) proving to be closer in terms of inter-chromosomal connectivity to a random network. Despite its almost constant form in healthy tissue, the AAC slightly increases when approaching the top co-expression layers. Meaning that inter-chromosomal connectivity decreases by a small amount when approaching top layers: the MSE of healthy GCN in the top 6 layers is an order of magnitude greater than the overall MSE (3×10^{-4}).

In breast cancer on the other hand, the inter-chromosomal connectivity drops abruptly when reaching the last few layers, going from an overall MSE of 0.018 to 0.3 in the top 6 layers. One of the advantages of modeling the co-expression program as a multi-layer GCN is that we can think of a co-expression program as a discrete process, where each layer is a co-expression 'picture' representing a co-expression range ending in the top co-expressed layer.

Under this framework, we can state that layers in both networks lose inter-chromosomal edges as they approach the final layer but not at the same rate nor magnitude. Indeed, the deviation in inter-chromosomal connectivity of a random state in the healthy GCN is subtle but clear in the last few layers, while the one of the Basal GCN has a fluctuation layer (around layer 95) ending in the disconnection of chromosomes in the network (right panel of Fig. 2). The bottom layer, close to the median of the co-expression distribution, reinforces the assumption of noise in most of the co-expression program as both GCNs have an AAC close to 0 here, in agreement with the fact that only a handful of all possible gene pairs actually interact in a biological regulation task (Consortium 2004).

If co-expression is thought of as a discrete process with the top co-expressed layer as the endpoint, the underlying structural evolution mechanism that happens across layers in the GCN becomes a central question to the understanding of the genetic co-expression program. A first step in this direction is to evaluate the structural or topological evolution

of the layers in both GCNs, first to see whether this mechanism is related to the evolution in terms of inter/intra-chromosomal connectivity across the co-expression layers. Second, to evaluate this structural process with the help of known network-generating mechanisms.

Structural evolution of healthy GCN absent in breast cancer

A first observation from looking at the structural evolution of the GCN of breast cancer is an almost constant structural nature. All structural values stagnate on a value for most layers, only to abruptly explode in the last few (Fig. 3a). Its degree assortativity coefficient (DAC) is 0 for most layers, implying a random assortative mixing in terms of the degree of nodes. In the case of the healthy GCN, the closest layer with a DAC of 0 is the noisy bottom layer with a value of -3×10^{-3}.

To quantify the extent of the variation from noisy co-expression in both GCNs, we computed for each structural measure, the absolute sum of the differences between the layers'

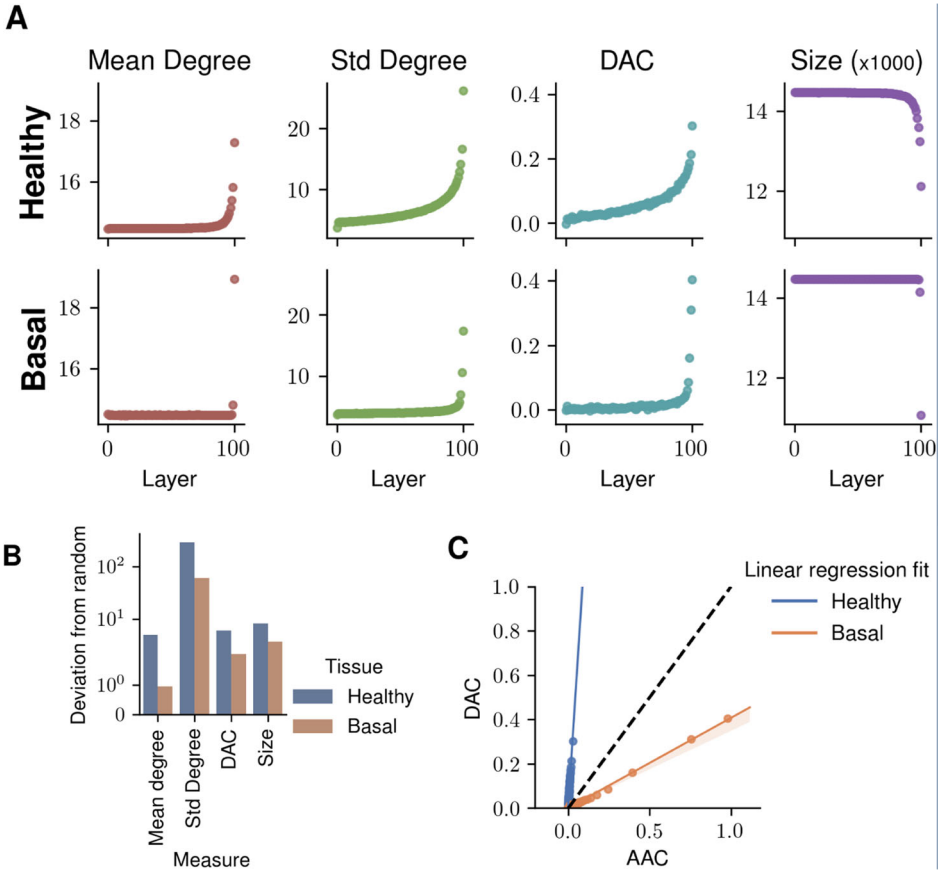

Fig. 3 Structural evolution of the gene co-expression networks of Basal and healthy tissues. **a** Each column is a different structural measure, namely the mean and standard deviation of the degree, the degree assortativity coefficient (DAC), and the size or number of nodes (in thousands of nodes). **b** Structural deviation from random of the GCN of each tissue, where random, for each structural measure is taken as the structural value of layer 0, representing co-expression noise. The deviation is the sum of the absolute differences of each layers' structural value from random. **c** Relation between the (chromosome) attribute assortativity (AAC), and degree assortativity (DAC) coefficients of the GCN of healthy and Basal tissues. Blue and orange lines are simple regression models of each relation, where shadowed background represents the 95% confidence interval of the model. The black dotted line is the identity line (y=y), and is added for reference

values and the value of layer 0 (Fig. 3b). Structurally, the healthy GCN deviates considerably more from the network structure of noise co-expression (layer 0 in the GCN), than breast cancer. This contrasts to its intra-chromosomal connectivity, where the AAC of healthy GCN evolves closely around random mixing across most layers (Fig. 2), with a slight increase in the top co-expression layers. The fact that the structural sensitivity to co-expression changes in the healthy GCN is lost in breast cancer, could suggest two independent/different structural development mechanisms underlying their co-expression programs.

The above discussion indicates that the difference in structural sensitivity is best depicted by the relation between chromosomal and degree assortativity, AAC and DAC respectively. This relation is quite close to be linear in both GCNs (Fig. 3c), and well-fitted using a simple regression model as defined in James et al. (2013). This claim is supported by the small mean square errors of both models (1×10^{-4} for the healthy GCN and 1.5×10^{-5} for the Basal GCN), where both models have an intercept value close to 0, congruent with the fact that a random network has random mixing in terms of chromosomes and degrees.

Moreover, the slope varies greatly between both models: in the healthy case, the slope is 27.5 times greater than in breast cancer (11.3 and 0.4 respectively), suggesting a different structural sensitivity—both in terms of network structure and chromosome rearrangement—and an almost linear relation in each GCN between these two. In other words, the structural changes happening across the layers of the healthy GCN are steady but do not seem to be related to inter-chromosomal connectivity, as in breast cancer. The mechanism underlying the healthy co-expression program, starts from an early layer a structural differentiation from noise that is non-existent in breast cancer, which follows an almost exclusively noisy path. This structural difference could be a lead into an organizational principle that may be necessary for a healthy co-expression program.

Preferential attachment-like evolution in healthy GCN is lost in breast cancer

A popular model for network generation of heavy-tailed degree distributions is preferential attachment (Barabási and Albert 1999; Price 1965). Roughly, preferential attachment happens when nodes with higher degree (many neighbors) gain edges *faster* than less connected nodes. This phenomenon of network creation appears in several different networks like the internet, scientific citations networks, movie actors networks, among others (Price 1976; Lehmann et al. 2004; Newman 2001; 2003; Jeong et al. 2003). This model was thought around the concept of a scale-free network, whose degree distribution function follows a power-law, $Pr(x) = x^{\alpha}$.

Commonly, a power-law distribution is characterized on a log-log scale as a linear shape (Jeong et al. 2003). This linear shape is found in the degree distribution of the top layer of the healthy GCN (Fig. 4b). Interestingly, the *evolution* of the degree distributions of the healthy layers starts at least 50 layers downstream and gradually takes the linear shape seen at the top co-expression layer (Fig. 4a). This supports the hypothesis of the inclusion of an organizational factor in the structure of the healthy GCN across layers. As we will see, such a factor seems to be based on a preferential attachment-like mechanism.

It is worth mentioning that the ubiquity and even the existence of power laws in networks (therefore its relation with preferential attachment), remains an active (and

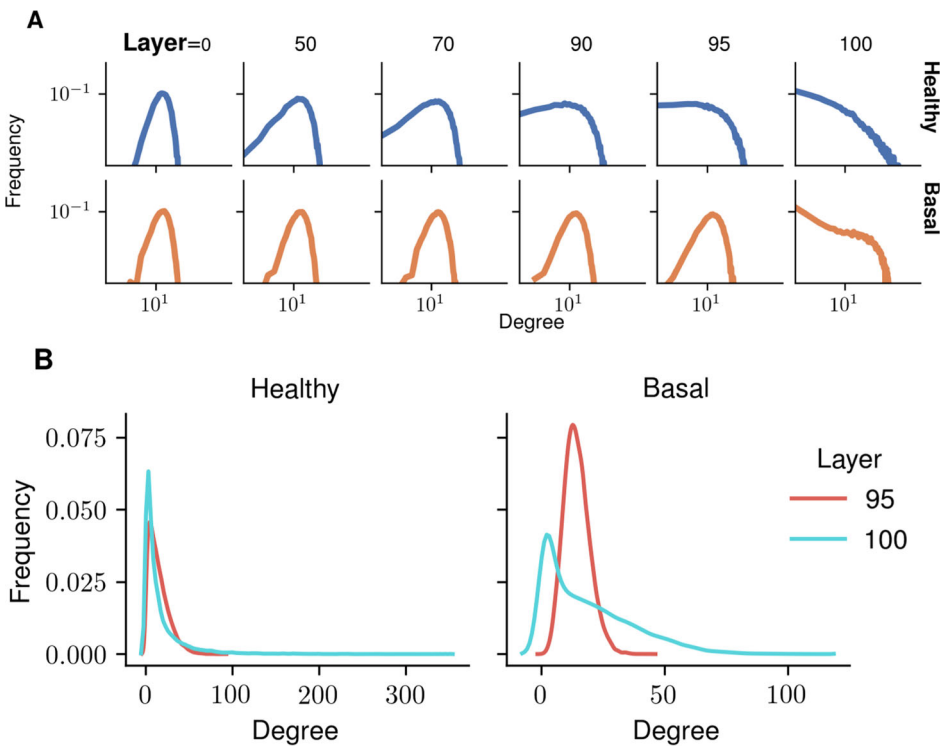

Fig. 4 **a** The evolution of degree distribution across 6 layers of the GCN of healthy and breast cancer (Basal-like) tissue. **b** Degree distribution of layer 100 (top layer) and layer 95 in both GCNs

somewhat polemic) topic of research (Broido and Clauset 2019). In our case, only the tail of the degree distribution of the top layer follows a power-law after statistical testing (Alstott et al. 2014). The rest of the distribution is best fitted by a log-normal law than by a power-law. Although the true statistical nature of the heavy-tailed distribution is out of the scope of the paper, here we make use of the concept of emergent, as opposed to statistical, scale-free power-law network (Holme 2019).

In breast cancer, the degree distribution of layers shows a different evolution from the healthy GCN: most layers have a degree distribution similar to that of layer 0, representing a random network (Fig. 4a). There is no gradual development towards the top co-expression layer, only an abrupt change in the degree distribution around layer 95. The top layer has a degree distribution that is also heavy tailed but without the power law form of the one of the healthy GCN (Fig. 4b). With this in consideration, the network generator model of preferential attachment does not appear suited for the breast cancer co-expression program.

Top layer hubs and core nodes are structurally relevant in the healthy GCN but not in cancer

Taken as a general concept, preferential attachment relates to the fact that a group of exclusive nodes gain faster connections than the rest. Here, we do not make reference to the mathematical model of preferential attachment (as coined in Barabási and Albert (1999)), but to the general concept of network generation. To see whether the preferential attachment concept can hold in the healthy GCN, we consider two different sets of highly connected nodes: hub genes and core genes. The idea is to take a highly connected

group of nodes in the top layer to look at their degree evolution from noise to highest co-expression layer.

Hub nodes, or simply hubs, are nodes that have significantly more neighbors in the network than the average node (Barabási and et al 2016). The degree threshold for hubs in healthy GCN is 69.6 and 53.8 in Basal GCN, which defines a set of 533 and 534 nodes, respectively. Their degree distributions are displayed in the rightmost panel of Fig. 6a, in red. To compare their distributions to the rest of the network, we constructed a control set of nodes, generated by taking at random (100 times) a set of nodes with the same size as the hubs (see Methods). The degree distribution of this control group is displayed in blue across layers. In the top layer, the mean degree of the control group in healthy GCN is 17.5 and 111.5 for the hubs. In the case of Basal GCN these numbers are 18.9 and 63.9, respectively.

For each tissue, we test the hypothesis H_0 that degrees of hubs in the top layer are comparable to a same-size random group across 6 layers (layers 0, 50, 70, 90, 95, and 100). In the case of the healthy tissue, H_0 is rejected not surprisingly in layer 100, were the hubs were taken from. In layers 90 and 95 H_0 is also rejected with probability $p = 0.95$, and in layer 70 H_0 is rejected with $p = 0.93$ and in layer 50 with $p = 0.87$. Finally, H_0 is not rejected in layer 0, showing that degrees of top-layer hubs are confused with random nodes in the median of the co-expression distribution. On the Basal-like GCN, H_0 is only rejected in layer 100, providing evidence of the random-like structure of the network starting at layer 95.

Interestingly, the degree evolution of hubs seems to separate from the control group early on in the healthy co-expression program (top panel Fig. 6a). The higher the layer, the larger the difference in mean degree between the two groups. Furthermore, the mean degree of the control group takes values that are close to the mean degree of layer 0 (16.5), across all the layers. On the other hand, the degrees of the top layer hubs are more connected on average from the control group, at least 50 co-expression layers downstream. These results support a preferential attachment-like mechanism in the healthy GCN: the group of hubs in the top-layer network gain gradually more connections across layers, and increasingly faster, than the average node.

In the case of breast cancer, this mechanism is completely lost (bottom panel Fig. 6a). Indeed, the degree distribution of the top layer hubs closely resembles that of the control group for most layers. There is no gradual gain of connections of the hubs at an early layer, suggesting that top layer hubs are average nodes in the rest of the layers. This is congruent with the overall degree distribution of Basal GCN shown previously, where most layers seem to have a distribution similar to the one found in the noise co-expression layer. Top layer hubs in breast cancer only start to differentiate to the control group around layer 96 (supplementary Figure S2), suggesting a two structural mechanisms across its layers, one random and constant from layer 0 to 95 and a different one starting at layer 96, which is strongly dictated by the intra-chromosomal connectivity.

In terms of spreading information on a network, hubs may not be the ideal set of nodes to consider, but the nodes in the main core of the network (Kitsak et al. 2010). In the context of a gene co-expression network, spreader genes could be an important part of the formation of functional pathways within a cell and thus important to be considered in this study. The core of a network is a concept from graph theory: the k-core of a network is the largest subnetwork $H = (V_k, E_k)$, such that every node in V_k has degree at least k.

The core or main core of a network is a k-core such that the $(k+1)$-core does not exist (Seidman 1983).

The core is a strongly inter-connected subnetwork: each node in the top layer core of the healthy GCN shares at least 58 connections with other nodes in the core (it is a 58-core), and 47 in the case of the Basal GCN. The total number of nodes is 141 and 42 in the healthy and Basal GCN, respectively (Fig. 5). As it can be observed in the top panel of Fig. 6b, the degree evolution of the core in the healthy GCN is similar to the one of hubs: its nodes gradually gain connections faster than the rest, starting at an early co-expression layer. The mean degree of the healthy top layer core is 142.9, while the degree of the control group is 17.5, a greater difference than in the case of hubs. In this case, H_0 is rejected with $p = 0.95$ for layers 100, 95 and 90, with $p = 0.94$ for layer 70 and $p = 0.89$ for layer 50. It is not rejected in layer 0.

In a similar fashion, the degree evolution of the top layer core in breast cancer GCN is stagnant and close to the control group on 5 of the 6 layers, similarly to the top layer hubs (bottom panel of Fig. 6b). The differentiation to the control group happens even later than in top layer hubs, in layer 98 (supplementary Figure S3). As in the case of hubs, H_0 is only rejected in layer 100. In general, the nodes in the top layer core have a degree evolution parallel to that of top layer hubs, with a slightly stronger differentiation to control nodes in the case of the healthy GCN and, conversely, a closer resemblance to the control group in the breast cancer GCN.

Organizational principles in the healthy co-expression program are lost in cancer

A preferential attachment-like mechanism in the healthy gene co-expression program could imply the existence of an underlying organizational principle in the co-expression program, where a set of genes could act as an interface of regulatory functional pathways. So far, we have seen that two groups of nodes convey such a mechanism: the top layer core nodes and the hubs. A high conservation rate of these nodes throughout the layers of the

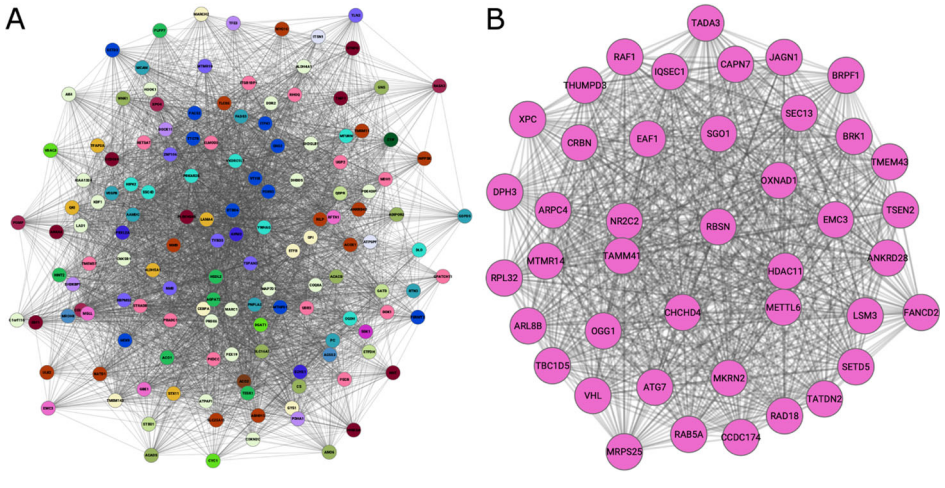

Fig. 5 Main k-core (or simply, cores) of **a** healthy, and **b** Basal GCNs. The color of genes is associated with the chromosome to which they belong. In the case of the Basal main core, all genes blong to chromosome 3, meanwhile in the healthy core, its 142 genes are shared between all chromosomes

healthy GCN could indicate that such an organizational principle exists and is dependent upon them.

Results show that such a similar principle is not likely to be found in breast cancer given the structural properties of its co-expression program. The random-like nature of the structure of the Basal GCN in most layers, suggests that an order of any kind is not to be expected with the exception of the top five layers. We also highlight the existence of another non-random mechanism in breast cancer (first seen in a single-layer network in Espinal-Enriquez et al. (2017)), that appears to be strongly related to a molecular phenomenon leading to changes in inter-chromosomal connectivity. This other mechanism is apparent only within the top five co-expression layers, where structural values deviate from random, and where highly connected nodes in the top layer gain connections faster than the average nodes (Figs. 4 and 6). Further experimental and theoretical research is needed to identify the causes of this change.

The cumulative conservation rate (*ccr*, see Methods) of nodes belonging to the top layer hubs and core nodes of the healthy GCN across layers is shown in Fig. 7. In sum, *ccr* is a measure of how well-conserved is a class of top layer nodes in the set of the same class

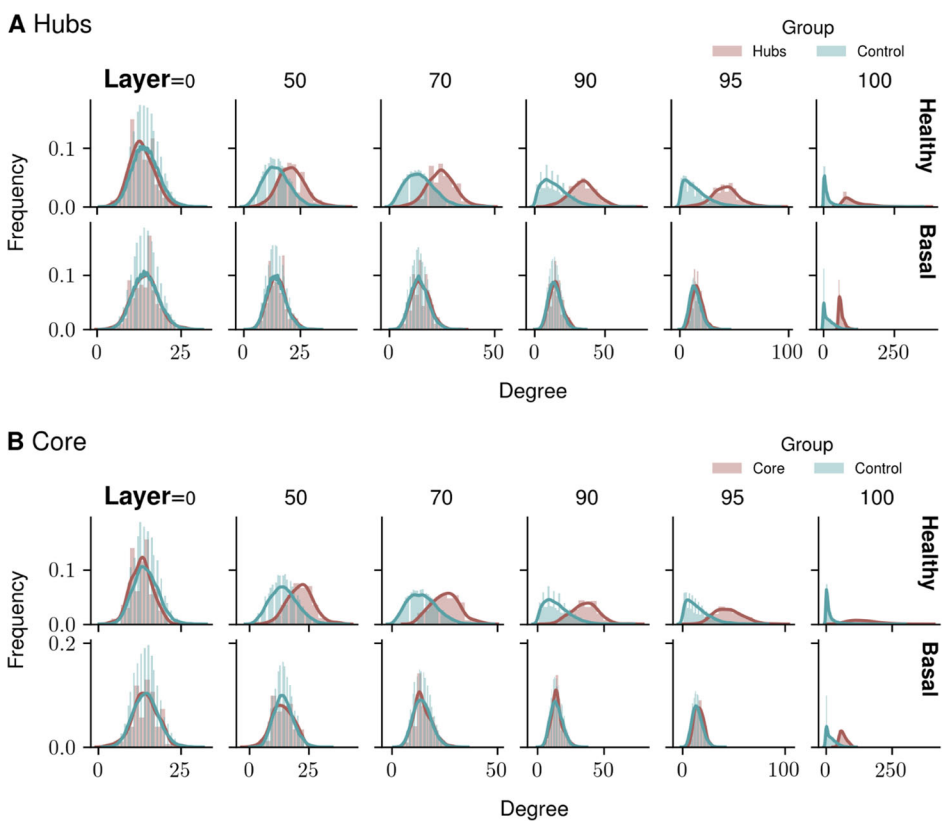

Fig. 6 The evolution of degree of highly connected nodes in the top co-expression layer. Two groups of highly connected nodes are considered: hub nodes and the nodes in the main core of the top co-expression layer. **a** The hub nodes are defined as the nodes with a degree greater than two standard deviations from the mean ($d_{hub} \geq \mu + 2\sigma$). **b** The core of a network is a well-defined set of highly inter-connected nodes (see for example (Seidman 1983)). For each GCN and each layer, degree distribution of each group is compared to that of a control set of nodes using a *Z*-test. This control group is obtained by picking at random 100 set of nodes of the same size. The cardinality of the hub sets is 533 and 534 for the healthy and Basal GCNs, respectively; and 141 and 133 in the case of the core

of the rest of layers. Figure 7 displays the *ccr* decay for healthy and Basal GCNs using a radial plot where the co-expression layers are displayed on the circumference, starting at layer 100 at the left and finishing at the same position with layer 0. The radius of the plot represents the value of *ccr* at any given layer. Values at the center of the circle are equal to 0 while values on the circumference are equal to 1. The *ccr* values of both hubs and nodes in the core are represented, as well as the values of a control core group (see *ccr* in Methods).

Hubs have been shown to be relevant in protein-protein networks in terms of both lethality and evolutionary adaptability of a cell (Jeong et al. 2001; Helsen et al. 2019). However, in the case of both GCNs, hubs are not specially well-conserved across co-expression layers (Fig. 7). Even if they are preferentially connected as we have seen, a large number of neighbors alone appears not to be a sufficient property for conservation at different levels of co-expression. In the case of the Basal GCN, the conservation is almost null, with non-null values only found previous to layer 96. The only little conservation happens on the top 4 layers, congruent with the double structural mechanism divided at around layer 95 in the Basal GCN. The healthy GCN shows a greater conservation of the top layer hubs, but it decay happens fast: $ccr_{99} = 0.84$, and $ccr_{81} = 0.03$; only 3% of the hubs in the top layer are conserved in the preceding 20 layers.

The control core group of Fig. 7 is obtained, for each layer, by producing one thousand random networks with the same inter/intra-chromosomal connectivity as said layer (as in a stochastic block model, see Methods), and then obtaining its main core. The *ccr* values of the control group are then computed for the complete set of one thousand iterations, and averaged. This way, we have a random *ccr* for reference to compare the significance of the conservation rate of core nodes in the real GCNs.

A striking result is the high *ccr* of nodes in the core of the top layer of the healthy GCN (Fig. 7). There is a good conservation rate from the top layer core nodes until the last layer $ccr_1 = 0.32$, moreover, there is a slow decay of conservation with $ccr_{81} = 0.93$. In the control core GCN, these values are 0.001 and 0.06, respectively. The cumulative conservation of core nodes, starting at the top-layer and across the subsequent co-expression-layer

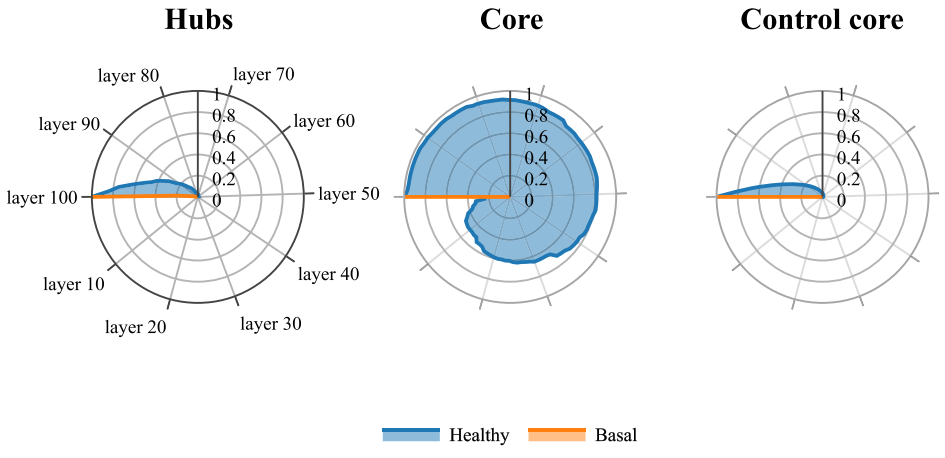

Fig. 7 Cumulative conservation ratio (*ccr*, see Methods) across the layers of healthy and Basal GCNs of two sets of highly connected nodes of the top layer: hubs and nodes in the core of the network. The rightmost figure represents the *ccr* of nodes in the core of the top co-expression layer of a randomly generated multilayer network using stochastic block models (see Methods)

cores is in general much greater than what would be expected at random. Indeed, in the set of 141 genes comprising the top-layer core of the healthy GCN, 90% of them are conserved in the cores of the other subsequent 30 layers' cores. This, considering a different pattern of interconnectedness across any two cores, as edges do not repeat across layers.

The different conservation rates in core nodes between cancer and healthy tissues hold true when considering layer 94 as the reference to compute *ccr* (supplementary Figure S4). In this case, the conservation ratio of the cores in the healthy network is lower with respect to the case when layer 100 was the reference, indicating that the set of cores at the top 6 layers maximize conservation and cores in subsequent layers add substantial irrelevant genes. However, the difference in *ccr* is still significant with respect to the breast-cancer GCN. The conservation ratio ten layers below the reference (ccr_{84}) is 0.6 and 0.42 for healthy and cancer GCNs, respectively. Ten further layers downstream these values are 0.46 and 0.12, finally ccr_{64} values are 0.37 and 0.05 in healthy and Basal GCNs, respectively. More generally, avoiding the top 5 layers shows a close relation between the conservation ratio of the breast cancer and the control group, congruent with the random-like behavior found in layer 94 and below in breast cancer.

In both cases, the *ccr* values are unlikely to be obtained from a random control group in the case of the healthy tissue. The average mean-square-error (MSE) between the control *ccr* values and those of 1000 random networks is equal to 8×10^{-5}, which is five orders of magnitude smaller than 0.5, the MSE between control and real *ccr* values. In the case of breast cancer the MSE between the random group and the real values is 0, specifically $ccr_{99} = 1$ in both cases, meaning that the vertiginous increase in intra-chromosomal connections from layer 99 to layer 100 is responsible for this MSE.

Results above are an effect of taking layer 100 as a reference for *ccr*: when instead we take layer 94 as the reference the MSE value between control and breast cancer *crr* is 6×10^{-3}, suggesting a close relation between the structure of random networks and that of breast cancer GCN for layers below 94.

Nodes in the core of networks have been shown to be good information spreaders (Al-garadi et al. 2017; Liu et al. 2015). One possible reason of this conservation could be the fact that genes in this set take part of several different functional pathways, each one with a complementary set of other genes. However, further research is needed for the identification of the functional relevance of this gene set.

The strong variance in the structural properties of healthy and Basal breast cancer GCNs in terms of conservation, leads us to argue that the organizational principle that determines the co-expression landscape in a healthy phenotype is abolished during the oncogenic process. While it is an already known factor that co-expression between physically close genes is higher than distant gene pairs in general (Hurst et al. 2004; Wang et al. 2011; Hurst 2017), in this work we have observed that the distant interactions between inter-chromosomal genes are central in the process that shapes the network landscape in the healthy phenotype, even if strong interactions between close genes also appear.

The loss of inter-chromosomal co-expression at the top layers in the cancer GCN, apparently leaves the intra-chromosomal interactions as the main (if not the only) mechanism for shaping the co-expression landscape. As intra-chromosomal interactions are the strongest and more abundant in the top layers, the lower layers have less of them and are connected similarly to a random network. The separation of the layers into inter

and intra chromosomal layers, reflects on the two different structural mechanisms found in the breast cancer GCN found in this investigation.

Functional implications of the loss of inter-chromosomal co-expression in basal breast cancer

The biological implications of this phenomenon may have impact in the way we understand gene expression and co-expression in cancer. In the healthy GCN, we observe a far-from-random degree distribution from the lowest layers. Conversely, almost all the structure in the Basal breast cancer GCN is similar to a randomly-generated network, but for the top layers, in which we observe a high intra-chromosomal connectivity. This effect in cancer GCN might be attributed to a loss of those mechanisms that orchestrate the gene co-expression landscape.

After a disruption in the organizational principles involved in maintaining the co-expression landscape, the possible remaining solution for a damaged cancer cell could be an elevated co-expression between close genes. The regulatory elements of gene transcription may behave in an *operon-like* fashion: the RNA polymerase complex may transcribe large sections of a given chromosome (perhaps due to an incorrectly open 3D structure of DNA). These sections could be flanked similarly in several breast cancer patients (probably due to methylation marks, CTCF binding sites, or stop signals of transcription). The resulting sections will have similar expression patterns, thus allowing high co-expression values in a large portion of the cancer genome.

Gene somatic copy number alterations (SCNAs) are one of the most documented genomic modifications in breast cancer. SCNAs are also known to affect gene expression over clusters of genes (Zhou et al. 2003; Inaki et al. 2014; Menghi et al. 2016). A specific breast cancer molecular subtype called HER2+, is actually defined by an amplification (located at Chr17q12) and protein over-expression of ERBB2 and neighbor genes (Slamon et al. 1987; Sørlie et al. 2001). Another breast cancer-associated amplification is located at region 17q25.3. This alteration occurs in BRCA1 mutated triple negative breast cancer, HER2+, or Luminal B breast cancer subtype (Toffoli et al. 2014).

Recently (García-Cortés et al. 2020), it was reported that in breast cancer subtypes GCNs, highly dense intra-cytoband hotspots coincide with commonly amplified regions. Intra-cytoband clusters of genes such as the aforementioned region Chr17q12 or Chr8q24.3 form highly dense connected components. The presence of co-expression clusters in cancer GCNs allows us to suggest that—additionally with CNAs—several other mechanisms may influence the co-expression landscape in breast cancer: non-coding RNAs, epigenetic modifications, 3D structure of DNA, CTCF binding sites alterations, etc. To establish the validity of the previous lines, it is necessary to analyze and integrate other *-omic* technologies (high-throughput genetical data) in a global framework. This will allow us to define the structures responsible for the correct maintenance of the genome-wide co-expression landscape.

Concluding remarks

The genetic co-expression program is often used as a proxy to the more general gene regulatory program (Marbach et al. 2012). Differences in the gene co-expression net-

works (GCNs) of cancer and healthy cells are thus a focal interest in cancer research. Un-weighted GCNs are defined by a single threshold of co-expression that selects a small subset of strongly co-expressed gene pairs.

A notorious difference between GCNs of healthy and breast cancer tissues found previously is related to inter-chromosomal connectivity (Espinal-Enriquez et al. 2017). The co-expression of breast cancer presents almost no inter-chromosomal interactions (gene pairs are almost exclusively in the same chromosome), while healthy tissue shows values similar to random assortative mixing as introduced in Newman (2002). To broad the scope of the segment of the co-expression program investigated, we use the concept of a multilayer GCN by splitting co-expression in 101 equal-sized intervals. This results in a multilayer network that allows us to study co-expression as a discrete process, starting at weak levels of co-expression building itself upward towards the strongest co-expressing gene pairs.

A first look at the multilayer GCNs of healthy and breast cancer (Basal-like subtype) tissues, shows a consistent difference between the two across all layers: healthy GCN has an almost constant but slightly decreasing inter-chromosomal connectivity, while breast cancer GCN abruptly looses inter-chromosomal connectivity around top co-expressing layers. This contrasts with the network structural changes across layers in both GCNs: in healthy tissue, the assortativity coefficient (how often high-degree nodes connect to other high-degree nodes), and degree variance increase disproportionately to the decrease in inter-chromosomal connectivity; while in breast cancer, inter-chromosomal connectivity decreases accordingly to the structural changes across layers.

A closer look shows that a preferential-attachment-like mechanism—where a subset of nodes gains edges faster than the rest of nodes—is behind this result in the healthy GCN. This explains the disproportionate structural changes, as a small subset of nodes gains edges in detriment to other nodes, and in turn decreases inter-chromosomal connectivity. This became apparent after looking at the hub nodes (highly connected nodes) of the top layer, as they significantly gain connections faster across co-expression layers than the rest of the nodes.

A mechanism similar to preferential attachment in breast cancer is, on the other hand, non-existent across most co-expression layers. A few of the strongest co-expressed layers present a mechanism resembling preferential attachment where a group of nodes gains connections slightly faster than the rest. However, this is lost quickly after showing the disconnectedness in terms of network structure between top layers and the rest of the breast cancer GCN.

In healthy tissue, it is perhaps the case of an organizational principle that is reflected by a preferential attachment mechanism, where a small group of preferential nodes is conserved across layers by keeping interconnected at different stages of the co-expression program. We claim that such a group exists by showing the strong conservation rate of nodes in the top layer main core—strongest interconnected nodes of a network—across the main cores of all other co-expression layers.

This is an important result as the presentation of a structural mechanism similar to preferential attachment is, to our knowledge, a novel property of the healthy co-expression program. Furthermore, the organizational principle that comes with the conservation of nodes in the main core of the strongest co-expression layer, opens the possibility of future

work dedicated to the identification of the biological regulatory processes underlying this organization.

Finally, the importance of this work is emphasized on the result showing the loss of an organizational principle in breast cancer. Interestingly, it appears that the co-expression program of breast cancer seems to be mostly noise. This random process ends at the top five co-expression layers in breast cancer, as they follow a principle of their own that is not fully identified here. Further analysis on this fact is needed to better understand the mechanism that abruptly interrupts the noise in the co-expression program of breast cancer, leading to a strong decrease in inter-chromosomal connectivity.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1007/s41109-020-00291-1.

Additional file 1: Supplementary Figure 1: Distribution of each structural parameter in 1000 random GCNs conserving chromosomal connectivity of layer 0. The real values in layer 0 of healthy and Basal GCNs are added as vertical dotted lines. Supplementary Figure 2: Degree distribution of layers 95 to 100 of healthy and Basal GCNs. Supplementary Figure 3: Degree distribution of core nodes of layer 100 and a control group in layers 95 to 100 of healthy and Basal GCNs. Supplementary Figure 4: Cumulative conservation rates (ccr) in core nodes of Basal and healthy tissues considering layer 94 as the reference to compute ccr.

Abbreviations

GCN: Gene co-expression network; TCGA: The Cancer Genome Atlas; MI: Mutual information; DAC: Degree assortativity coefficient; AAC: Attribute assortativity coefficient

Authors' contributions

RDG devised the project's methodological approach, performed the computational analyses, developed and implemented programming code, made the figures, drafted the manuscript. DGC performed pre-processing and low-level data analysis, constructed the networks, contributed to the writing of the manuscript. EHL co-supervised the project, contributed to the writing of the manuscript. JEE conceived and designed the overall project, co-supervised the project, took the lead in the biological analyses, drafted the manuscript. All authors read and approved the final version of the manuscript.

Funding

This work was supported by CONACYT (721450 student grant to D.G.C; 285544/2016, and 2115/2018 to J.E.E.), as well as by federal funding from the National Institute of Genomic Medicine (Mexico). Additional support has been granted by the National Laboratory of Complexity Sciences (232647/2014 CONACYT). J.E.E. is recipient of the 2018 Miguel Alemán Fellowship in Health Sciences. E.H.L. is a recipient of the 2016 Marcos Moshinsky Fellowship in the Physical Sciences. D.G.C is a doctoral student from Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México (UNAM). This work is part of her PhD Thesis. The funding institutions had no role in the design of the study and collection, analysis, and interpretation of data, nor in writing the manuscript.

Availability of data and materials

The datasets analyzed for this study can be found in the GDC repository, https://portal.gdc.cancer.gov/repository.

Competing interests

The authors declare that they have no competing interests.

Received: 27 February 2020 Accepted: 24 July 2020
Published online: 05 August 2020

References

Al-garadi MA, Varathan KD, Ravana SD (2017) Identification of influential spreaders in online social networks using interaction weighted k-core decomposition method. Phys A Stat Mech Appl 468:278–288

Alcalá-Corona SA, de Anda-Jáuregui G, Espinal-Enríquez J, Hernández-Lemus E (2017) Network modularity in breast cancer molecular subtypes. Front Physiol 8:915

Alstott J, Bullmore E, Plenz D (2014) powerlaw: a python package for analysis of heavy-tailed distributions. PLoS ONE 9(1):85777

Barabási A-L, Albert R (1999) Emergence of scaling in random networks. Science 286(5439):509–512

Barabási A-L, et al (2016) Network Science. Cambridge university press, Cambridge

Barthélemy M, Barrat A, Pastor-Satorras R, Vespignani A (2005) Characterization and modeling of weighted networks. Phys A Stat Mech Appl 346(1-2):34–43

Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A (2005) Reverse engineering of regulatory networks in human B cells,. Nat Genet 37(4):382–90. https://doi.org/10.1038/ng1532

Boccaletti S, Bianconi G, Criado R, Del Genio CI, Gómez-Gardenes J, Romance M, Sendina-Nadal I, Wang Z, Zanin M (2014) The structure and dynamics of multilayer networks. Phys Rep 544(1):1–122
Broido AD, Clauset A (2019) Scale-free networks are rare. Nat Commun 10(1):1–10
Consortium GO (2004) The gene ontology (go) database and informatics resource. Nucleic Acids Res 32(suppl_1):258–261
de Anda-Jáuregui G, Alcalá-Corona SA, Espinal-Enríquez J, Hernández-Lemus E (2019) Functional and transcriptional connectivity of communities in breast cancer co-expression networks. Appl Netw Sci 4(1):22
de Anda-Jáuregui G, Espinal-Enriquez J, Hernández-Lemus E (2019) Spatial organization of the gene regulatory program: An information theoretical approach to breast cancer transcriptomics. Entropy 21(2):195
de Anda-Jáuregui G, Fresno C, García-Cortés D, Enríquez JE, Hernández-Lemus E (2019) Intrachromosomal regulation decay in breast cancer. Appl Math Nonlinear Sci 4(1):223–230
de Anda-Jáuregui G, Velázquez-Caldelas TE, Espinal-Enríquez J, Hernández-Lemus E (2016) Transcriptional network architecture of breast cancer molecular subtypes. Front Physiol 7:568
Emmert-Streib F, Glazko GV, Altay G, de Matos Simoes R (2012) Statistical inference and reverse engineering of gene regulatory networks from observational expression data. Front Genet 3:8. https://doi.org/10.3389/fgene.2012.00008
Espinal-Enriquez J, Fresno C, de Anda-Jáuregui G, Hernández-Lemus E (2017) Rna-seq based genome-wide analysis reveals loss of inter-chromosomal regulation in breast cancer. Sci Rep 7(1):1760
Friedman N, Linial M, Nachman I, Pe'er D (2000) Using Bayesian networks to analyze expression data,. J Comput Biol J Comput Mol Cell Biol 7(3-4):601–20. https://doi.org/10.1089/106652700750050961
García-Cortés D, de Anda-Jáuregui G, Fresno C, Hernandez-Lemus E, Espinal-Enriquez J (2020) Gene co-expression is distance-dependent in breast cancer. Front Oncol 10:1232. https://www.frontiersin.org/article/10.3389/fonc.2020.01232, doi:https://doi.org/10.3389/fonc.2020.01232. 2234-943X
Hagberg A, Swart P, S Chult D (2008) Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States)
Helsen J, Frickel J, Jelier R, Verstrepen KJ (2019) Network hubs affect evolvability. PLoS Biol 17(1):3000111
Hernández-Lemus E, Rangel-Escareño C (2011) The role of information theory in gene regulatory network inference. Inf Theory New Res:109–144
Hernández-Lemus E, Reyes-Gopar H, Espinal-Enríquez J, Ochoa S (2019) The many faces of gene regulation in cancer: A computational oncogenomics outlook. Genes 10(11):865
Hernández-Lemus E, Velázquez-Fernández D, Estrada-Gil JK, Silva-Zolezzi I, Herrera-Hernández MF, Jiménez-Sánchez G (2009) Information theoretical methods to deconvolute genetic regulatory networks applied to thyroid neoplasms. Physica A Stat Mech Appl 388(24):5057–5069
Holland PW, Laskey KB, Leinhardt S (1983) Stochastic blockmodels: First steps. Soc Netw 5(2):109–137
Holme P (2019) Rare and everywhere: Perspectives on scale-free networks. Nat Commun 10(1):1–3
Hurst LD (2017) It's easier to get along with the quiet neighbours. Mol Syst Biol 13(9):943. https://doi.org/10.15252/msb.20177961
Hurst LD, Pál C, Lercher MJ (2004) The evolutionary dynamics of eukaryotic gene order. Nat Rev Genet 5(4):299
Inaki K, Menghi F, Woo XY, Wagner JP, Jacques P.-É., Lee YF, Shreckengast PT, Soon WW, Malhotra A, Teo AS, et al (2014) Systems consequences of amplicon formation in human breast cancer. Genome Res 24(10):1559–1571
James G, Witten D, Hastie T, Tibshirani R (2013) An Introduction to Statistical Learning, vol. 112. Springer, Switzerland
Jeong H, Mason SP, Barabási A-L, Oltvai ZN (2001) Lethality and centrality in protein networks. Nature 411(6833):41–42
Jeong H, Néda Z, Barabási A-L (2003) Measuring preferential attachment in evolving networks. EPL (Europhys Lett) 61(4):567
Kitsak M, Gallos LK, Havlin S, Liljeros F, Muchnik L, Stanley HE, Makse HA (2010) Identification of influential spreaders in complex networks. Nat Phys 6(11):888
Kivelä M, Arenas A, Barthelemy M, Gleeson JP, Moreno Y, Porter MA (2014) Multilayer networks. J Complex Netw 2(3):203–271
Lehmann S, Jackson AD, Lautrup B (2004) Life, death and preferential attachment. EPL (Europhys Lett) 69(2):298
Liu R, Cheng Y, Yu J, Lv Q-L, Zhou H-H (2015) Identification and validation of gene module associated with lung cancer through coexpression network analysis. Gene 563(1):56–62
Liu Y, Tang M, Zhou T, Do Y (2015) Core-like groups result in invalidation of identifying super-spreader by k-shell decomposition. Sci Rep 5:9602
Madhamshettiwar PB, Maetschke SR, Davis MJ, Reverter A, Ragan MA (2012) Gene regulatory network inference: evaluation and application to ovarian cancer allows the prioritization of drug targets. Genome Med 4(5):41
Marbach D, Costello JC, Küffner R, Vega NM, Prill RJ, Camacho DM, Allison KR, Aderhold A, Bonneau R, Chen Y, et al (2012) Wisdom of crowds for robust gene network inference. Nat Methods 9(8):796
Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, Califano A (2006) Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics 7:7. BioMed Central
Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, Califano A (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context,. BMC bioinformatics 7 Suppl 1:7. https://doi.org/10.1186/1471-2105-7-S1-S7
Margolin AA, Wang K, Lim WK, Kustagi M, Nemenman I, Califano A (2006) Reverse engineering cellular networks,. Nat Protoc 1(2):662–71. https://doi.org/10.1038/nprot.2006.106
McKinney W (2012) Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython. O'Reilly Media, Inc., California
Menghi F, Inaki K, Woo X, Kumar PA, Grzeda KR, Malhotra A, Yadav V, Kim H, Marquez EJ, Ucar D, et al (2016) The tandem duplicator phenotype as a distinct genomic configuration in cancer. Proc Natl Acad Sci 113(17):2373–2382
Newman ME (2001) Clustering and preferential attachment in growing networks. Phys Rev E 64(2):025102
Newman ME (2002) Assortative mixing in networks. Phys Rev Lett 89(20):208701
Newman ME (2003) The structure and function of complex networks. SIAM Rev 45(2):167–256
Nielsen TD, Jensen FV (2009) Bayesian Networks and Decision Graphs. Springer, Switzerland
Nowicki K, Snijders TAB (2001) Estimation and prediction for stochastic blockstructures. J Am Stat Assoc 96(455):1077–1087
Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al (2011) Scikit-learn: Machine learning in python. J Mach Learn Res 12(Oct):2825–2830

Price DJDS (1965) Networks of scientific papers. Science:510–515
Price D. d. S. (1976) A general theory of bibliometric and other cumulative advantage processes. J Am Soc Inf Sci 27(5):292–306
Seabold S, Perktold J (2010) Statsmodels: Econometric and statistical modeling with python. In: Proceedings of the 9th Python in Science Conference, vol. 57. p 61. Scipy
Seidman SB (1983) Network structure and minimum degree. Soc Netw 5(3):269–287
Shapiro SS, Wilk MB (1965) An analysis of variance test for normality (complete samples). Biometrika 52(3/4):591–611
Slamon DJ, Clark GM, Wong SG, Levin WJ, Ullrich A, McGuire WL (1987) Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene,. Science (New York, N.Y.) 235(4785):177–82. https://doi.org/10.1126/science.3798106
Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Lønning PE, Børresen-Dale AL (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications,. Proc Natl Acad Sci USA 98(19):10869–74. https://doi.org/10.1073/pnas.191367098
Teschendorff AE, Severini S (2010) Increased entropy of signal transduction in the cancer metastasis phenotype. BMC Syst Biol 4(1):104
Toffoli S, Bar I, Abdel-Sater F, Delrée P, Hilbert P, Cavallin F, Moreau F, Van Criekinge W, Lacroix-Triki M, Campone M, Martin AL, Roché H, Machiels JP, Carrasco J, Canon JL (2014) Identification by array comparative genomic hybridization of a new amplicon on chromosome 17q highly recurrent in BRCA1 mutated triple negative breast cancer. Breast Cancer Res 16(1). https://doi.org/10.1186/s13058-014-0466-y
Tomczak K, Czerwińska P, Wiznerowicz M (2015) The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge,. Contemp Oncol(Poznan, Poland) 19(1A):68–77. https://doi.org/10.5114/wo.2014.47136
Tovar H, García-Herrera R, Espinal-Enríquez J, Hernández-Lemus E (2015) Transcriptional master regulator analysis in breast cancer genetic networks. Comput Biol Chem 59:67–77
Van Der Walt S, Colbert SC, Varoquaux G (2011) The numpy array: a structure for efficient numerical computation. Comput Sci Eng 13(2):22
Wang Z, Gerstein M, Snyder M (2009) Rna-seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10(1):57
Wang Z, Hou Z, Xin H (2005) Internal noise stochastic resonance of synthetic gene network. Chem Phys Lett 401(1-3):307–311. https://doi.org/10.1016/j.cplett.2004.11.064
Wang GZ, Lercher MJ, Hurst LD (2011) Transcriptional coupling of neighboring genes and gene expression noise: Evidence that gene orientation and noncoding transcripts are modulators of noise. Genome Biol Evol 3(1):320–331. https://doi.org/10.1093/gbe/evr025
Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM, Network CGAR, et al (2013) The cancer genome atlas pan-cancer analysis project. Nat Genet 45(10):1113
West J, Bianconi G, Severini S, Teschendorff AE (2012) Differential network entropy reveals cancer system hallmarks. Sci Rep 2:802
Yang Y, Han L, Yuan Y, Li J, Hei N, Liang H (2014) Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. Nat Commun 5:3231
Zhang B, Horvath S (2005) A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol 4(1). https://doi.org/10.2202/1544-6115.1128
Zhou Y, Luoh SM, Zhang Y, Watanabe C, Wu TD, Ostland M, Wood WI, Zhang Z (2003) Genome-wide identification of chromosomal regions of increased tumor expression by transcriptome analysis. Cancer Res 63(18):5781–5784

Publisher's Note