RESEARCH ARTICLE

A Null Model for Pearson Coexpression Networks

Andrea Gobbi, Giuseppe Jurman*

Fondazione Bruno Kessler, Trento, Italy

* jurman@fbk.eu

Abstract

Gene coexpression networks inferred by correlation from high-throughput profiling such as microarray data represent simple but effective structures for discovering and interpreting linear gene relationships. In recent years, several approaches have been proposed to tackle the problem of deciding when the resulting correlation values are statistically significant. This is most crucial when the number of samples is small, yielding a non-negligible chance that even high correlation values are due to random effects. Here we introduce a novel hard thresholding solution based on the assumption that a coexpression network inferred by randomly generated data is expected to be empty. The threshold is theoretically derived by means of an analytic approach and, as a deterministic independent null model, it depends only on the dimensions of the starting data matrix, with assumptions on the skewness of the data distribution compatible with the structure of gene expression levels data. We show, on synthetic and array datasets, that the proposed threshold is effective in eliminating all false positive links, with an offsetting cost in terms of false negative detected edges.

OPEN ACCESS

Citation: Gobbi A, Jurman G (2015) A Null Model for Pearson Coexpression Networks. PLoS ONE 10(6): e0128115. doi:10.1371/journal.pone.0128115

Academic Editor: Francisco J. Esteban, University of Jaén, SPAIN

Received: February 11, 2014

Accepted: April 23, 2015

Published: June 1, 2015

Funding: The work has been funded by Fondazione Bruno Kessler. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Introduction

Universally acknowledged by the scientific community as the basic task of systems biology, network inference is the prototypical procedure for moving from the classical reductionist approach to the novel paradigm of data-driven complex systems in the interpretation of biological processes [1].

The goal of all network inference (or network reconstruction) procedures is the detection of the topology (*i.e.*, the wiring diagram) of a graph whose nodes belong to a given set of biological entities, starting from measurements of the entities themselves. In the last fifteen years, the reconstruction of the regulation mechanism of a gene network and of the interactions among proteins from high-throughput data such as expression microarray or, more recently, from Next Generation Sequencing data has become a major line of research for laboratories worldwide. The proposed approaches rely on techniques ranging from deterministic to stochastic, and their number is constantly growing. Nonetheless, network inference is still considered an open, unsolved problem [2]. In fact, in many practical cases, the performance of the reconstruction algorithms are poor, due to several factors limiting inference accuracy [3, 4] to the point of making it no better than a coin toss in some situations [5]. The major problem is the under-determinacy of the task [6], due to the overwhelming number of interactions to be predicted starting from a (usually) small number of available measurements. In general, size and quality of available data are critical factors for all inference algorithms.

In what follows the impact of data size is discussed for one of the simplest inference techniques, *i.e.*, the gene coexpression network, where interaction strength between two genes is a function of the correlation between the corresponding expression levels across the available tissue samples. Despite its simplicity, the number of studies in the literature based on coexpression networks is still a large fraction of all manuscripts in systems biology [7–13], including Next Generation Sequencing data [14]. The underlying biological hypothesis is that functionally related genes have similar expression patterns [15], and thus that coexpression is correlated with functional relationships, although this does not imply causality. This implies for instance that genes that are closer in a biological network are likely to have more similar expression [16, 17]: this hypothesis has been validated to some extent, for example by Jensen *et al.* in [18]. However, a caveat is mandatory here: since correlation is a univariate method, it is unable to capture the relations occurring among genes when the independence hypothesis does not hold, which is a common situation in the -omics datasets [19, 20]. To overcome this problem, characterized by small correlation values between functionally related genes, multivariate approaches are required [21], also in the network case [22].

Notwithstanding this limitation, as highlighted in [23], correlation can help unveil the underlying cellular processes, since coordinated coexpression of genes encode interacting proteins. The Pearson Correlation Coefficient (PCC for short) is used as the standard measure, although alternative correlation measures can be also employed: see for instance [24] for a comparative review of eight well known association metrics. Furthermore, coexpression analysis has been intensively used as an effective algorithm to explore the system-level functionality of genes, sometimes outperforming much more refined approaches [25, 26]. The observation that simpler approaches such as correlation can be superior even on synthetic data has been explained by some authors [27, 28] with the difficulties of a complex algorithm in detecting the subtleties of combinatorial regulation. Coexpression networks can also capture more important features than the conventional differential expression approach [29], and their use has been extended to other tasks, for instance the investigation of complex biological traits [30]. Finally, these networks can be crucial for understanding regulatory mechanisms [31], for the development of personalised medicine [32] or, more recently, in metagenomics [33].

However, as noted in [34], correlation between genes may sometimes be due to unobserved factors affecting expression levels. Moreover, deciding when a given correlation value between two nodes can be deemed statistically significant and thus worthwhile for assigning a link connecting them is a major issue affecting coexpression networks [35]. This translates mathematically into choosing (a function of) a suitable threshold, as in the case of mutual information and relevance networks [36]. As reported in [37], in the literature statistical methods for testing the correlations are underdeveloped, and thresholding is often overlooked even in important studies [38]. The two main approaches known in the literature to solve this problem can be classified as soft or hard thresholding. Soft thresholding is adopted in a well-known framework called Weighted Gene Coexpression Network Analysis (WGCNA) [39], recently used also for other network types [40, 41]. All genes are mutually connected, and the weight of a link is a positive power of the absolute value of PCC, where the exponent is chosen as the best fit of the resulting network according to a scale-free model [42, 43]. This approach, without discarding any correlations, promotes high correlation values and penalises low values. In the hard thresholding approach, instead, only correlation values larger than the threshold are taken into account, and an unweighted link is set for each of these values, so that a binary network is generated (see [44] for one of the earliest references). Clearly, an incorrectly chosen threshold value can jeopardise the obtained results with false negative links (for too strict a threshold) or false positive links (for too loose a threshold). Many heuristics have been proposed for setting the threshold values, such as using the False Discovery Rate [45–47], using a cross-validation strategy [48], the *p*-value of the correlation test [15, 32], employing partial correlation [49, 50], using rank-based techniques [51–53], more complex randomisation techniques [54], or keeping only values exceeding a minimum acceptable strength (MAS) level specified by the threshold [55]. Alternative approaches have been recently studied, evaluating the correlation distribution, both experimentally [56] and theoretically [57]. However, these studies focus on the level of single interaction rather than considering the whole network.

Moreover, in many studies in the literature, the threshold is not chosen according to an algorithmic procedure, but referring to standard choices [58–61], or to heuristics not directly related to the correlation values, but rather with the resulting network topology [62–71]. In [72] a comparison of some coexpression thresholds is shown on a few microarray datasets. Furthermore, each threshold choice yields a compromise between detecting artefacts (false positives) and neglecting existing connections (false negatives) [56], which can be driven by the cost attached to the task studied. For instance, the optimized threshold selection procedure in [73] attains a very low false positive rate by using a permutation-based strategy [74].

Here we propose a new a priori model for the computation of a hard threshold based on the assumption that a random coexpression graph should not have any edges. This threshold follows from the work of Fisher [75] and Bevington [76], and, as a deterministic independent null model, it depends only on the dimensions of the starting data matrix, with assumptions on the skewness of the data distribution compatible with the structure of gene expression levels data [77, 78]. Hence, the procedure to obtain the threshold is not stochastic and thus deeply different from approaches based on permutation tests [74]. Further, this model is non-parametric, because its only input is the data themselves and no additional quantity needs to be tuned. By definition, this threshold is aimed at minimising the possible false positive links, paying a price in terms of false negative detected edges. This characterising property makes this method especially useful when the intrinsic cost function associated with the studied task is biased towards penalising false positives. We conclude by demonstrating the procedure first on a synthetic dataset and then on an ovarian epithelial carcinoma dataset on a large cohort of 285 cases [79, 80].

A Motivating Example

We show hereafter an example of a common situation arising in the small sample size setting, where PCC can reach extremely high values possibly leading to the detection of false positives.

Consider the HepatoCellular Carcinoma (HCC) dataset, first introduced in [81] and later used in [82] and publicly available at the Gene Expression Omnibus (GEO) http://www.ncbi.nlm.nih.gov/geo with accession number GSE6857. The dataset collects 482 tissue samples from 241 patients affected by HCC. For each patient, a sample from cancerous hepatic tissue and a sample from surrounding non-cancerous hepatic tissue were extracted, hybridised on the Ohio State University CCC MicroRNA Microarray Version 2.0 platform consisting of 11,520 probes collecting expressions of 250 non-redundant human and 200 mouse miRNA. After a preprocessing phase including imputation of missing values [83] and discarding probes corresponding to non-human (mouse and controls) miRNA, the resulting dataset HCC includes 240 + 240 paired samples described by 210 human miRNA, with the cohort consisting of 210 male and 30 female patients. In particular, consider now the three microRNA identified as `hsa.mir.010b.precNo1`, `hsa.mir.016a.chr13` and `hsa.mir.016b.chr3`. A search by sequence with the miRBase web-service [84] shows that the two probes `hsa.mir.016a.`

chr13, and hsa.mir.016b.chr3 share a high alignment score (*e*-value: $5 \cdot 10^{-4}$[85]), while no link due to coherent coexpression is known between each of these probes and hsa.mir.010b.precNo1. Consistent with this observation, the absolute PCC $|\rho|$ on the dataset H consisting of the 210 tumoral samples from male patients is

$$\begin{aligned} |\rho_H(\texttt{hsa.mir.016a.chr13}, \texttt{hsa.mir.016b.chr3})| &= 0.969 \\ |\rho_H(\texttt{hsa.mir.010b.precNo1}, \texttt{hsa.mir.016b.chr3})| &= 0.536. \end{aligned}$$

When we restrict the analysis to a suitable small subset of patients, different situations can occur. Consider in fact the following sub-cohort of 10 patients

$$\begin{aligned} \mathrm{S} = &\{03-457, 02-354, 03-146, 03-467, 03-037, \\ &03-033, 03-280, 03-205, 02-421, 02-432\} . \end{aligned}$$

On S, PCC reads as follows:

$$\begin{aligned} |\rho_S(\texttt{hsa.mir.016a.chr13}, \texttt{hsa.mir.016b.chr3})| &= 0.907 \\ |\rho_S(\texttt{hsa.mir.010b.precNo1}, \texttt{hsa.mir.016b.chr3})| &= 0.941, \end{aligned}$$

that is, the correlation between the two unrelated miRNA on S is very high and even higher than the correlation of the two aligned probes. The curves (averaged on 1000 runs) of PCC versus the sample size when the remaining samples are increasingly and randomly added (while S remains the same across the 1000 runs) are shown in Fig 1. When sample size increases, $|\rho($hsa.mir.010b.precNo1,hsa.mir.016b.chr3$)|$ quickly drops down to 0.536, while the correlation of the two aligned probes remains almost constant.

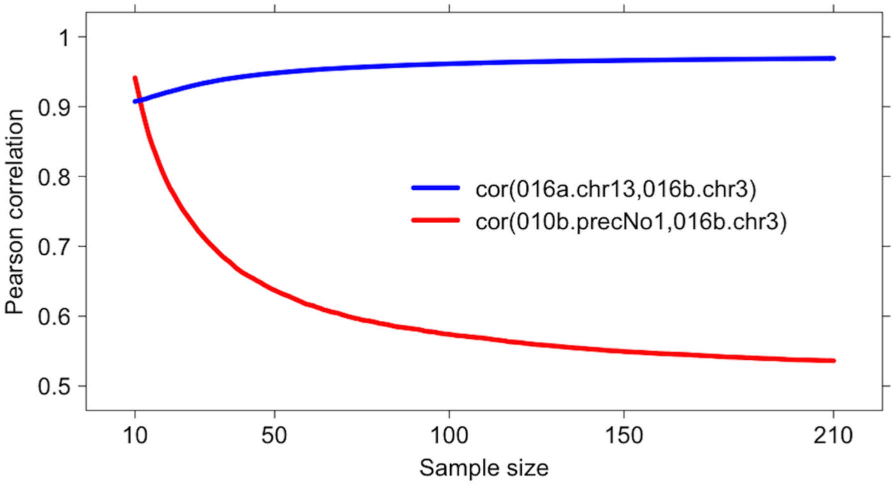


Fig 1. HCC dataset: PCC versus sample size for the coexpression of hsa.mir.016a.chr13 with hsa.mir.016b.chr3 (blue line) and with hsa.mir.010b.precNo1 (red line). When the correlation is computed on the 10-sample set S the correlation between the two uncoexpressed probes is even higher than the correlation between the two probes sharing an almost identical alignment. When more samples are randomly added, the blue line slightly increases, while the red line quickly drops to the final value 0.536 on the whole dataset H. Both curves are averaged over 1000 randomisations of the added samples, keeping the sub-cohort S constant.

doi:10.1371/journal.pone.0128115.g001

Clearly, if the task were to build a PCC coexpression network N_S on S, whatever the chosen hard threshold t, the network N_S would include either a false positive or a false negative link (or even both, for $0.907 \leq t \leq 0.941$). Note that the probability of extracting 10-sample sub-cohorts S′ satisfying

$$|\rho_{S'}(\texttt{hsa.mir.010b.precNo1}, \texttt{hsa.mir.016b.chr3})| >$$
$$|\rho_{S'}(\texttt{hsa.mir.016a.chr13}, \texttt{hsa.mir.016b.chr3})| > 0.9$$

is about 0.03%, but when the constraints are relaxed the probability of detecting artificially high correlation values $|\rho_{S'}(\texttt{hsa.mir.010b.precNo1},\texttt{hsa.mir.016b.chr3})| > 0.75$ raises to about 6.2% for this dataset. Situations like those shown in this example regularly occur when building coexpression networks, supporting the need for an algorithmic solution to the problem of hard thresholding in the small sample size setup.

Methods

Distribution of PCC

Given a real number $0 < p < 1$, define the function

$$F(n,p) = P(|\rho(x,y)| > p) \ , \tag{1}$$

where x and y are two independent normal vectors of length n and ρ is the PCC.

The results in [75, 86–92] and the symmetry of ρ yield that $F(n, p)$ has the following close form:

$$F(n,p) = P(|\rho(x,y)| > p) = \frac{2}{\sqrt{\pi}} \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n-2}{2}\right)} \int_0^{\arccos p} \sin^{n-3}(\vartheta) \mathrm{d}\vartheta \ , \tag{2}$$

where $\Gamma(x)$ is the Gamma function $\Gamma(x) = \int_0^{+\infty} t^{x-1} e^{-t} \mathrm{d}t$. In Text A in S1 Text we also propose a novel proof of Eq (2) by means of an analytic argument stemming from the work of Bevington in [76, Ch. 7]. Moreover, Eq (2) is also a good approximation in the case of a non-normal distribution of x and y when data skewness can be bounded [93], because of the generalization shown in [89, 94–96]. Non-Gaussian asymmetric distributions can occasionally be detected in some array studies [78]: however, techniques for reducing the skewness are routinely applied during preprocessing [77], and thus the aforementioned results can be safely used in the microarray framework.

Coexpression network and threshold selection

The results derived in the previous section are used here to construct a null model for the correlation network, thus yielding a threshold for the inference of a coexpression network from the nodes' data. As mentioned in the Introduction, these correlation networks are subject to the hypothesis of independence between genes, so they are not detecting higher-order relations for which a multivariate method is needed. In Text E in S1 Text we show a few synthetic and -omics examples of relations among genes that are not captured by coexpression networks.

In the general situation, we are measuring the signal (expression) of m probes (genes), in n different instances/conditions (samples), to infer the corresponding correlation networks on m nodes.

Table 1. A subset of values of the secure threshold $\bar{p}$ for different number of samples *n* and genes *m*.

m n	100	500	1000	2000	10000	50000	100000
8	0.95629	0.98520	0.99070	0.99415	0.99800	0.99932	0.99957
15	0.81681	0.89170	0.91323	0.93036	0.95800	0.97456	0.97949
20	0.73825	0.82388	0.85077	0.87330	0.91286	0.93973	0.94852
30	0.62814	0.71776	0.74817	0.77485	0.82534	0.86367	0.87729
50	0.50225	0.58534	0.61513	0.64213	0.69607	0.74036	0.75705
75	0.41647	0.49026	0.51740	0.54238	0.59353	0.63709	0.65394
100	0.36343	0.42999	0.45477	0.47774	0.52537	0.56662	0.58279

doi:10.1371/journal.pone.0128115.t001

Formally, let $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^{m}$ be a set such that $\mathbf{x}_i \in \mathbb{R}^n \; \forall i = 1, \ldots, m$. Then the coexpression p-graph $\mathcal{G}_p = \{V, E_p\}$ is the graph where

$$V = \{v_1, \ldots, v_m\} \qquad \text{and} \qquad (v_i, v_j) \in E_p \Leftrightarrow |\rho(\mathbf{x}_i, \mathbf{x}_j)| > p \ .$$

The first result characterizes the coexpression graphs in terms of null models:

Proposition 1 If the vectors $\mathbf{x}_i$ are sampled from the uniform distribution, the graph $\mathcal{G}_p$ is an Erdös-Rényi model [97] with m nodes and probability $F(n, p)$ as in Eq (2).

The proof follows immediately from the definition of $\mathcal{G}_p$ and Eq (2).

Definition Using the results in the previous section, the *secure threshold* $\bar{p}$ is defined as follows, for an arbitrarily small $\varepsilon > 0$:

$$\bar{p} = \min_{p \in (0,1]} \left\{ F(n, p) \frac{m(m-1)}{2} \leq 1 - \varepsilon \right\} \ , \tag{3}$$

for m nodes measured on n samples. As a consequence of [98], stating that the median of the binomial distribution is the integer part of the mean when $F(n, p) < 1 - \log 2$, the secure threshold $\bar{p}$ is the minimum value of p such that the corresponding random coexpression network is on average an empty graph E_m, *i.e.*, $P(\mathcal{G}_{\bar{p}} = E_m) \geq \frac{1}{2}$.

The underlying hypothesis for Eq (3) is the assumption that in a random dataset no edge is expected, since no relation should occur between nodes. Due to its definition, the secure threshold $\bar{p}$ is biased towards avoiding the false positive links, paying a price in terms of false negatives. In Table 1 a collection of values of $\bar{p}$ is listed for different m and n, while in Fig 2 the contour plot of the function $\bar{p}(n, m)$ is shown first on a large range of values and then zooming on the small sample size area.

As a first example, we compare the secure threshold $\bar{p}$ with the threshold derived by applying the Bonferroni correction procedure [99, 100] to the p-value of PCC. Let now Φ_F and $\Phi_{t,\, n\, -\, 2}$ be the cumulative distribution functions for the Fisher transformation and the Student's t-distribution with $n - 2$ degrees of freedom, respectively. For a given PCC ρ computed between vectors in $\mathbb{R}^n$, let

$$FT_n(\rho) = 2\Phi_F(-|\operatorname{arctanh}(\rho)|\sqrt{n-3}) \qquad tT_n(\rho) = 2\Phi_{t,n-2}\left(-|\rho|\sqrt{\frac{n-2}{1-r^2}}\right) \tag{4}$$

be the functions computing the p-value associated to ρ following either the Fisher approximation FT_n [75, 89, 101] or the Student's t-distribution tT_n [102–106]. Set now the global

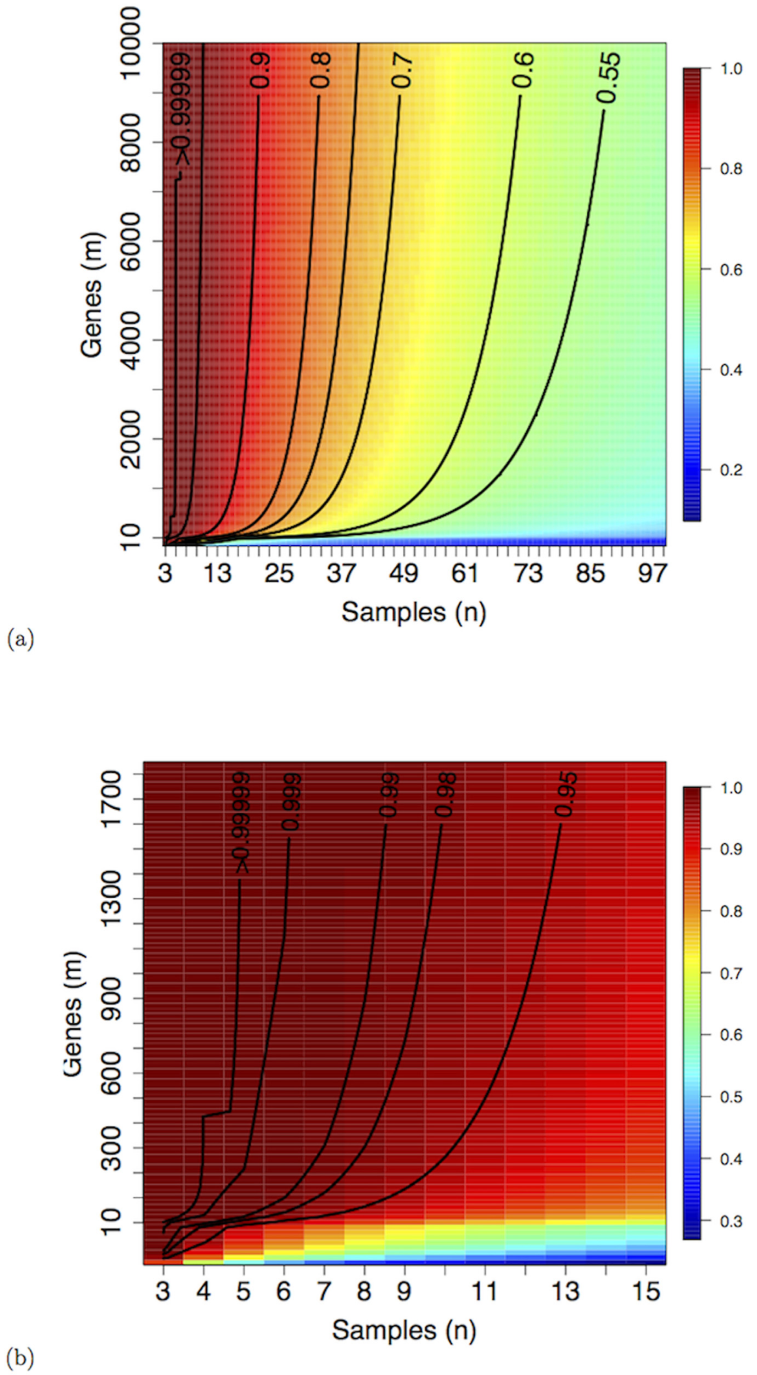


Fig 2. Contour plot of the function $\overline{p}(n, m)$ in the Samples × Genes space on (a) a wide (*n*, *m*) range and (b) zoomed on the small sample size area.

doi:10.1371/journal.pone.0128115.g002

Table 2. The secure threshold and the Bonferroni correction: comparison for different number of samples *n* and genes *m* among the secure threshold $\bar{p}$ and two Bonferroni-derived thresholds $B_{5,\,F}$ and $B_{5,\,t}$.

m n	thr.	5	10	25	50	100	1000
10	$\bar{p}$	0.55	0.71	0.82	0.88	0.92	0.97
	$B_{5,\,F}$	0.79	0.85	0.90	0.92	0.94	0.97
	$B_{5,\,t}$	0.81	0.87	0.92	0.95	0.97	0.99
15	$\bar{p}$	0.44	0.58	0.70	0.77	0.82	0.91
	$B_{5,\,F}$	0.67	0.74	0.80	0.83	0.86	0.92
	$B_{5,\,t}$	0.69	0.76	0.83	0.86	0.89	0.95
20	$\bar{p}$	0.38	0.51	0.62	0.69	0.74	0.85
	$B_{5,\,F}$	0.60	0.66	0.73	0.76	0.79	0.86
	$B_{5,\,t}$	0.61	0.68	0.75	0.79	0.82	0.90
25	$\bar{p}$	0.34	0.46	0.56	0.63	0.68	0.80
	$B_{5,\,F}$	0.54	0.61	0.67	0.71	0.74	0.82
	$B_{5,\,t}$	0.55	0.62	0.69	0.73	0.77	0.85
50	$\bar{p}$	0.24	0.32	0.41	0.46	0.50	0.62
	$B_{5,\,F}$	0.39	0.45	0.50	0.54	0.57	0.66
	$B_{5,\,t}$	0.40	0.45	0.51	0.55	0.59	0.68
100	$\bar{p}$	0.17	0.23	0.29	0.33	0.36	0.45
	$B_{5,\,F}$	0.28	0.32	0.37	0.40	0.43	0.50
	$B_{5,\,t}$	0.28	0.33	0.37	0.40	0.43	0.51

The thresholds $B_{5,\,F}$ and $B_{5,\,t}$ are computed by applying the Bonferroni correction to the *p*-value of PCC as derived by the Fisher approximation ($B_{5,\,F}$) or by the Student's t-distribution ($B_{5,\,t}$), with FWER ≤ 0.05. For each combination of samples *n* and genes *m*, the ranking $\bar{p} < B_{5,F} < B_{5,t}$ consistently holds, and the gaps are narrowing with growing *n* and *m*.

doi:10.1371/journal.pone.0128115.t002

familywise error rate to 5% significance level (FWER < 0.05): for a network with m nodes, by Bonferroni correction, this yields a p-value at most $p_m = \frac{0.1}{n(n-1)}$ for each link in the graph. By using the inverse functions $B_{5,F} = FT_n^{-1}(p_m)$ and $B_{5,t} = tT_n^{-1}(p_m)$ we obtain the corresponding thresholds (Fisher and Student's t, respectively) attaining the desired significance level. In Table 2 the values of $B_{5,\,F}$ and $B_{5,\,t}$ are reported for a selection of pairs (n, m), together with the corresponding secure threshold $\bar{p}$, while in Fig 3 the surface plot of $\bar{p}$ is compared to $B_{5,\,t}$.

In all cases, the relation $\bar{p} < B_{5,F} < B_{5,t}$ holds, and when $n < 10$ both $B_{5,\,F}$ and $B_{5,\,t}$ cannot be defined, since the 5% significance level is not reached by any threshold. Both these results do not come unexpected, since the Bonferroni correction is known to be a very conservative procedure, especially when the number of tests is large (see for instance [107–109]) because it ignores existing relations between measurements. This results in a large False Negative Rate. Thus, in many situations in general and in -omics studies in particular, the Bonferroni correction should be substituted by more refined multiple testing approaches [110].

We show now in Table 3 the comparison on a set of synthetic and array datasets of the secure threshold $\bar{p}$ with other well known hard thresholding methods, the clustering coefficient-based threshold C^* [67] and the statistical thresholds based on the adjusted p-values of 0.01, 0.05 or 0.1, while in Table 4 we compare the secure threshold $\bar{p}$ with the optimal threshold O [73] on a collection of array datasets on brain tissues. The threshold O is optimized for each dataset (*i.e.*, sample size, data quality and biological structure), considering also the effects

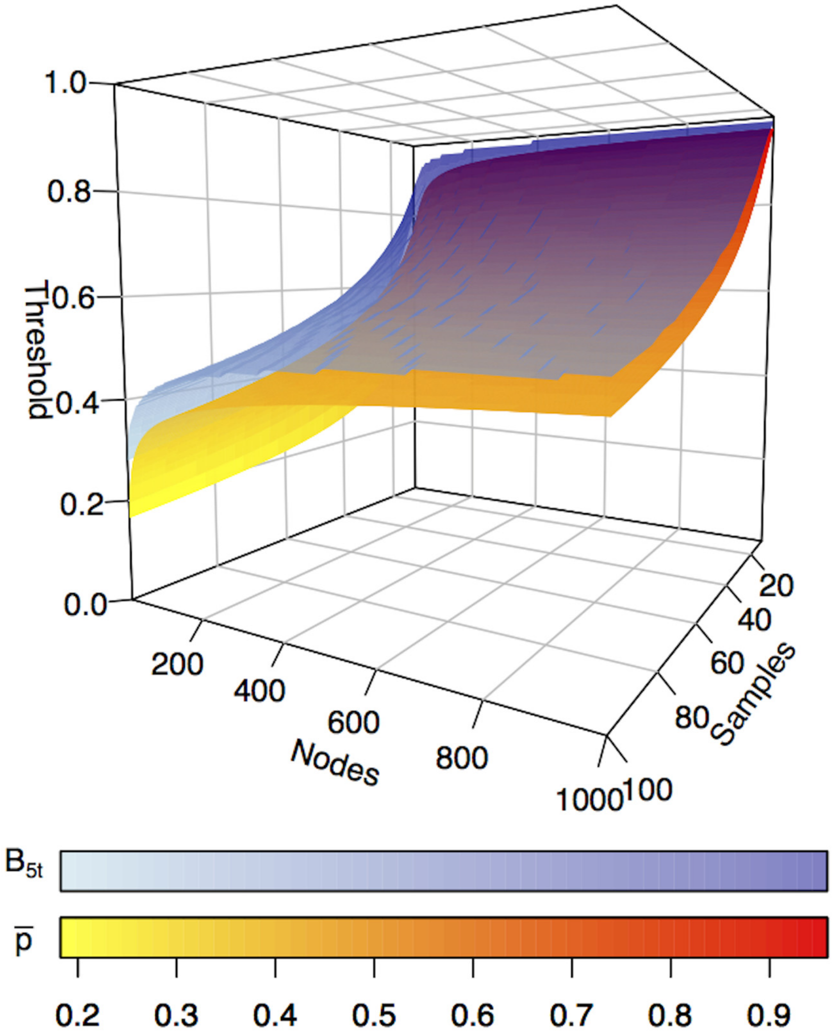


Fig 3. Comparison of the threshold functions $\bar{p}(n, m)$ (yellow-red gradient) and $B_{5,t}$ (blue gradient) in the Samples × Genes space; darker colors correspond to larger threshold values. The relation $\bar{p} < B_{5,t}$ consistently holds.

doi:10.1371/journal.pone.0128115.g003

induced by the influence of subpopulations on network generation, and comparing to permutated data. Note that for a given pair (n, m), the threshold $\bar{p}$ is constant, while the threshold O depends on the particular dataset. In almost all cases, the threshold $\bar{p}$ is the strictest.

As shown in the previous section, for a not very skewed distribution, the good approximation provided by the exact formula for $F(n, p)$ given in Eq (2) guarantees the effectiveness of the secure threshold $\bar{p}$ in detecting actual links between nodes. Nonetheless, whenever a stricter threshold is needed, it is still possible to follow the construction proposed, with the following refinement: the edge-creation process in the Erdös-Rényi model follows a binomial distribution, where n is the number of trials and p the probability associated with the success of a trial. The mean np of this distribution is one of the contributing terms in the definition Eq (3) of the secure threshold. To further restrict the number of falsely detected links, the variance term (np $(1 - p)$ for the binomial distribution) can be added to the original formula through Chebyshev's

Table 3. Comparison of the secure threshold p with the clustering coefficient-based threshold *C [67] and the statistical thresholds based on the adjusted p-values B0.01, B0.05 or B0.1 on a collection of synthetic and array datasets.**

Dataset type	n	m	*C**	B0.01	B0.05	B0.1	$\bar{p}$
Simulated	50	1000	0.57	0.58	0.54	0.52	0.6152
Simulated	25	1000	0.69	0.76	0.72	0.70	0.7956
H-U133P	23	897	0.72	0.78	0.74	0.72	0.8125
H-U133P	10	897	0.78	0.96	0.94	0.93	0.9723
H-U133P	10	675	0.77	0.96	0.93	0.92	0.9681
H-U133P	9	897	0.79	0.97	0.96	0.95	0.9821
H-U133P	8	897	0.81	0.98	0.97	0.96	0.98999
H-U133P	7	897	0.81	0.99	0.99	0.98	0.99558
H-U133P	6	897	0.86	>0.99	>0.99	0.99	0.99872
H-U133P	5	897	0.92	>0.99	>0.99	>0.99	0.99984
H-U133P	4	897	0.99	>0.99	>0.99	>0.99	> 0.9999
H-U133A	4	675	0.99	>0.99	>0.99	>0.99	> 0.9999
H-I6	4	675	0.99	>0.99	>0.99	>0.99	> 0.9999
M-U74	4	401	0.97	>0.99	>0.99	>0.99	0.9999

doi:10.1371/journal.pone.0128115.t003

inequality:

$$P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2} \ ,$$

where μ and σ are the mean and the standard deviation of X. Thus, the definition of secure threshold can be sharpened to $\tilde{p}_k$ as follows, for an arbitrarily small $\varepsilon > 0$:

$$\tilde{p}_k = \min_{p\in(0,1]} \left\{ F(n,p)\frac{m(m-1)}{2} + k\sqrt{(1-F(n,p))F(n,p)\frac{m(m-1)}{2}} \leq 1-\varepsilon \right\} \ .$$

For instance, the binomial distribution, for large values of n, can be approximated as a normal distribution for which 95.45% of the values lie in the interval $(\mu - 2\sigma, \mu + 2\sigma)$. Moreover,

Table 4. Comparison of the secure threshold $\bar{p}$ with the optimal threshold *O* [73] on a collection of array datasets on brain tissues. Note that for a given pair (*n*, *m*), the threshold *p* is constant, while the threshold *O* depends on the particular dataset.

Dataset type	n	m	*O*	$\bar{p}$
H-U133P	30	22277	0.82	0.84570
H-U133P	28	26199	0.85	0.86492
H-U133P	58	29211	0.8	0.68841
H-U133P	56	29211	0.8	0.69741
H-U95av2	50	12453	0.8	0.70261
Agilent	39	12235	0.82	0.76591
H-U133A	44	22383	0.85	0.75203
H-U95av2	47	12453	0.79	0.71857
H-U95av2	46	12453	0.84	0.72411
H-U95av2	50	12453	0.82	0.70261
H-U95av2	59	12453	0.79	0.66011
MOE 430_2	24	25859	0.83	0.89684
MOE 430_2	24	25859	0.78	0.89684
MOE 430_2	24	25859	0.87	0.89684

doi:10.1371/journal.pone.0128115.t004

the Chebyshev's inequality implies that at least 96% of the values lie in the interval ($\mu - 5\sigma$, $\mu + 5\sigma$). Finally, in Text D of S1 Text we show the analogue of Table 1 for $\tilde{p}_2$ and $\tilde{p}_5$, respectively.

Results and Discussion

To conclude, we show the application of the secure threshold $\bar{p}$ in two datasets—synthetic and array—demonstrating its behaviour as a function of data subsampling.

Synthetic dataset

In order to construct a simulated microarray dataset $\mathcal{G}$, we first created a correlation matrix $M_{\mathcal{G}}$ on 20 genes $G_1, \ldots G_{20}$, together with a dataset $\mathcal{G}$ of the corresponding expression G_i^{1000} across 1000 synthetic samples, so that $M_{\mathcal{G}}(i,j) = |\rho(G_i^{1000}, G_j^{1000})|$ is the absolute PCC between the expression of the genes G_i and G_j from $\mathcal{G}$. In particular, $M_{\mathcal{G}}$ has two 10 × 10 blocks highly correlated on the main diagonal, and two 10 × 10 poorly correlated blocks on the minor diagonal, as shown in Fig 4(a).

These blocks are derived from the following generating rule, given uncorrelated starting element G_1^{1000} and G_{11}^{1000}:

$$|\rho(G_k^{1000}, G_j^{1000})| \approx \begin{cases} 1 - 0.03j & \text{for } k = 1, 2 \leq j \leq 10 \\ \\ 0.7 - 0.015j & \text{for } k = 11, 12 \leq j \leq 20 \end{cases} .$$

Outside the two main blocks, all correlation values range between 0.002 and 0.074. In order to use $M_{\mathcal{G}}$ as the ground truth in what follows, all values outside the two blocks on the diagonal are thresholded to zero $(M_{\mathcal{G}})_{i\ 10\ +\ j} = (M_{\mathcal{G}})_{i\ 10\ +\ j} = 0$ for $1 \leq i, j \leq 10$, while entries in the two main blocks are considered as real numbers when evaluating HIM distance and binarized to one when evaluating True/False Positive/Negative links.

In Fig 4(b) we also show the heatmap of the gene expression dataset $\mathcal{G}$. Then a subset of n_s samples is selected from the starting 1000, and the corresponding coexpression networks is built, for the 100 hard threshold values 0.01j, for $1 \leq j \leq 100$. The secure threshold for these cases are respectively 0.799, 0.596 and 0.389. This procedure is repeated 500 times for each value $n_s = 10,20,50$. The same experiment is then repeated adding a 20% and a 40% level of Gaussian noise to the original data: for a given signal s, we build $s + \varepsilon$ with $\varepsilon \in \mathcal{N}(0, \alpha \cdot (\max(s) - \min(s)))$ for $\alpha = 0.2,0.4$ respectively.

Using $M_{\mathcal{G}}$ as the ground truth, for each hard threshold 0.01j we evaluate the ratio of False Positive links (*i.e.*, the quotient between the number of False Positive links and 190, the number of all possible links in a complete undirected network on 20 nodes), the ratio of False Negative links and the Hamming-Ipsen-Mikhailov (HIM) distance from the gold standard. The HIM distance [111, 112] is a metric between networks having the same nodes, ranging from 0 (distance between two identical networks) to 1 (comparison between the complete and the empty graph). The graphs summarizing the experiments are displayed in Fig 5.

In all cases, the secure threshold $\bar{p}$ corresponds to the strictest value yielding a coexpression network with no false positive links included, which is its characterizing property. Moreover, in almost all displayed situations, thresholding at $\bar{p}$ still guarantees an acceptable HIM distance from the ground truth, and a false negative ratio that is always less than 0.4.

Ovarian cancer

The aforementioned results obtained in the synthetic case are then tested here in a large array study on 285 patients of ovarian cancer at different stages [80], recently used in a comparative study on conservation of coexpressed modules across different pathologies [79]. In detail, the

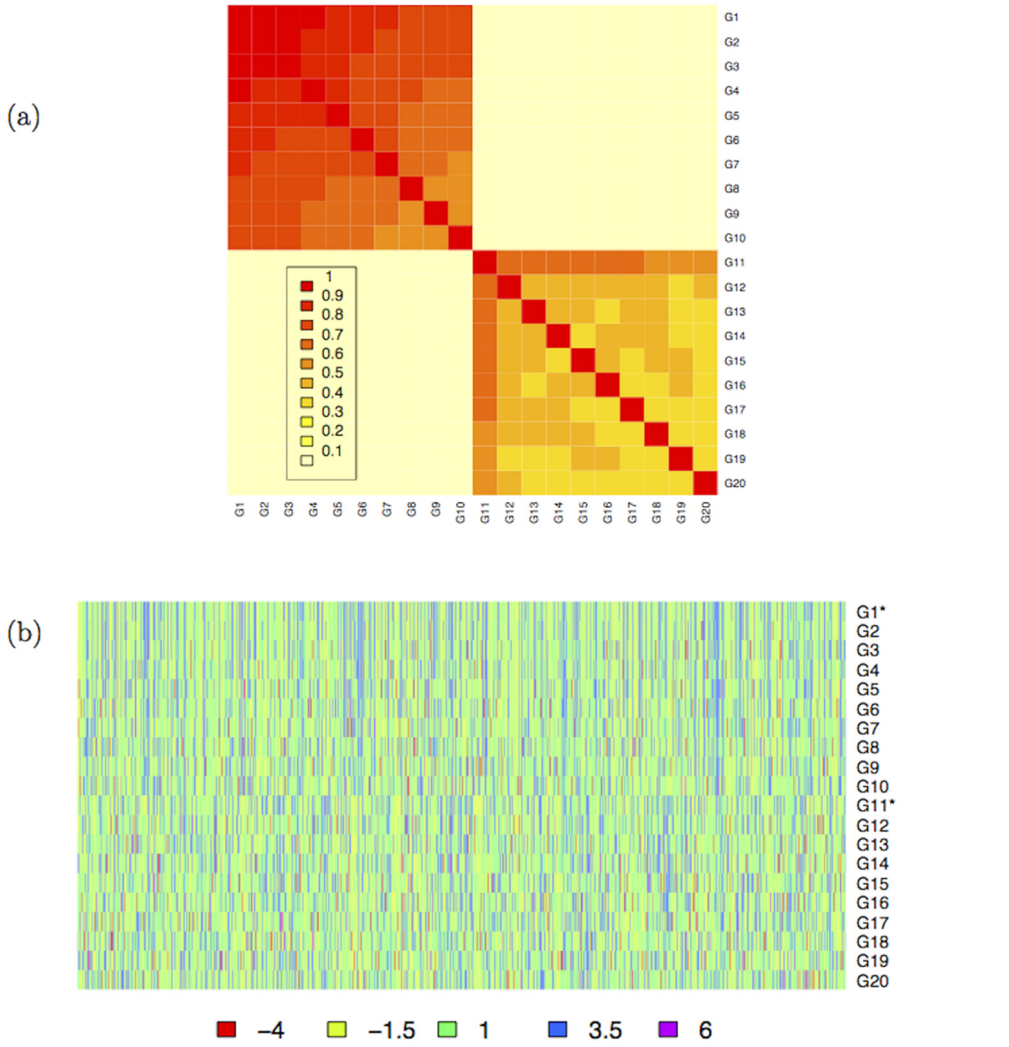
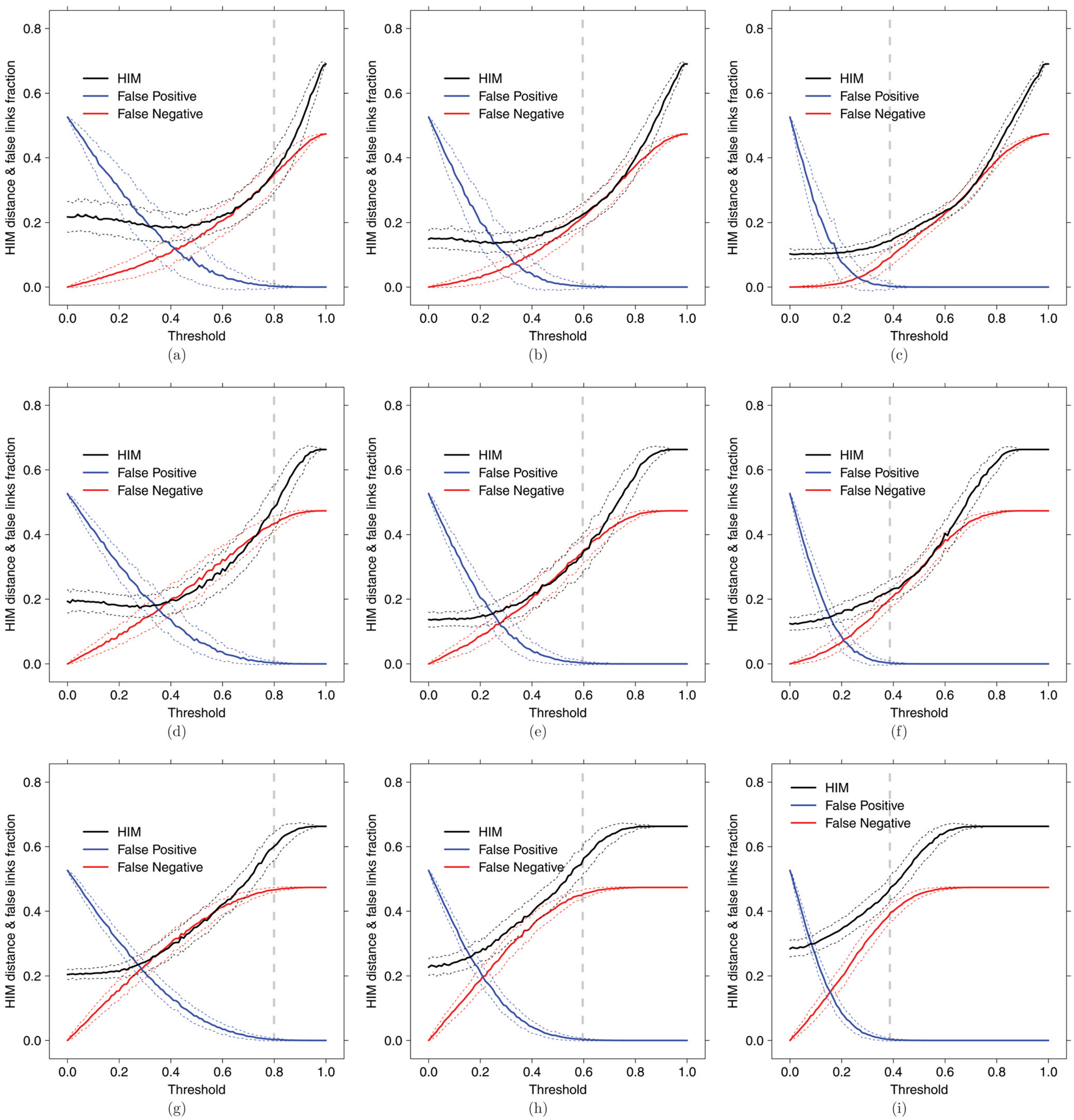


Fig 4. Synthetic dataset $\mathcal{G}$: level plot of the structure of the correlation matrix $M_{\mathcal{G}}$ (a) and heatmap of the dataset $\mathcal{G}$ (b). The generating gene expression vectors G_1^{1000} and G_{11}^{1000} are marked with *.

doi:10.1371/journal.pone.0128115.g004

authors profiled the gene expression of 285 predominately high-grade and advanced stage serous cancers of the ovary, fallopian tube, and peritoneum; the samples were hybridized on the Affymetrix Human Genome HG-U133 Plus 2.0 Array, including 54,621 probes. The goal of the original study was to identify novel molecular subtypes of ovarian cancer by gene expression profiling with linkage to clinical and pathologic features. As a major result, the authors presented two ranked gene lists supporting their claim that molecular subtypes show distinct survival characteristics. The two gene lists characterize the Progression Free Survival (PSF) and the Overall Survival (OS) patients, respectively. In each list genes are ranked according to a score weighting their association to target phenotype (PSF or OS): association is stronger for larger absolute values of the score, while the score's sign is negative for associations with good outcome and positive for associations with poor outcome.

Following the procedure of the previous, synthetic example, first we individuate the sample subset corresponding to the homogeneous cohort of 161 patients with grade three cancer and a set T of 20 genes, with 11 genes strongly associated with good PSF or good OS (EDG7,



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Fig 5. Synthetic dataset $\mathcal{G}$. Coexpression inference of the $M_{\mathcal{G}}$ network from random subsampling of the $\mathcal{G}$ dataset, without noise (a,b,c), with 20% Gaussian noise (d,e,f) and with 40% Gaussian noise (g,h,i), on 10 (a,d,g), 20 (b,e,h) and 50 (c,f,i) samples. Solid lines indicate the mean over 500 replicated instances of HIM distance (black), ratio of False Positive (blue) and ratio of False Negative (red); dotted lines of the same color indicate confidence bars (+/-σ), while grey vertical dashed lines correspond to the secure threshold $\bar{p}$.

doi:10.1371/journal.pone.0128115.g005

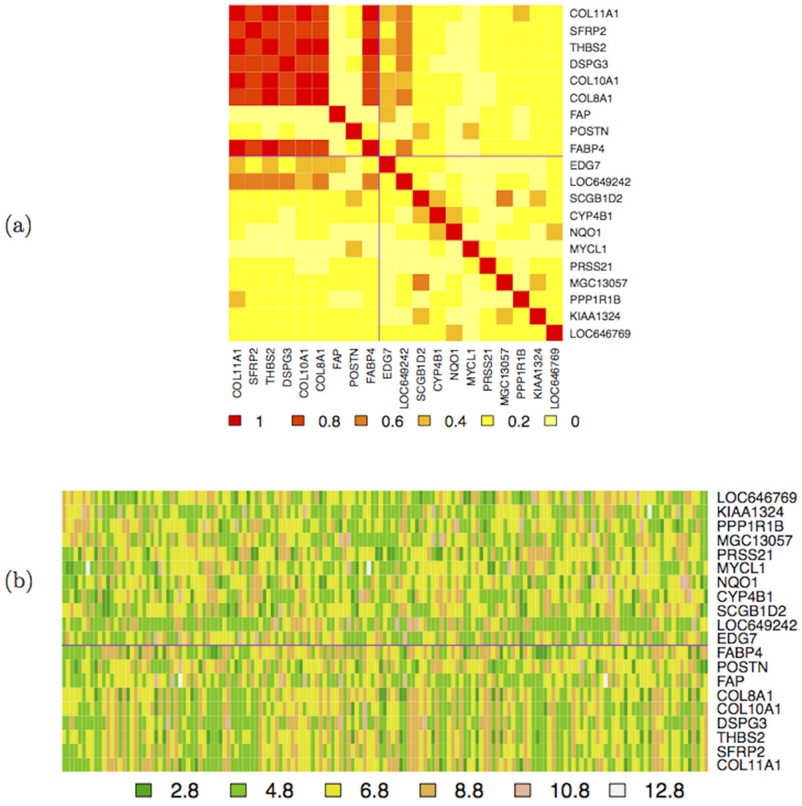


Fig 6. Ovarian cancer dataset $\mathcal{O}$. Level plot of the structure of the correlation matrix O_T (a) and heatmap of the Ovarian dataset $\mathcal{O}_T$ restricted to the set of 20 selected genes T (b). Solid lines separate the group of good and poor PFS/OS top genes.

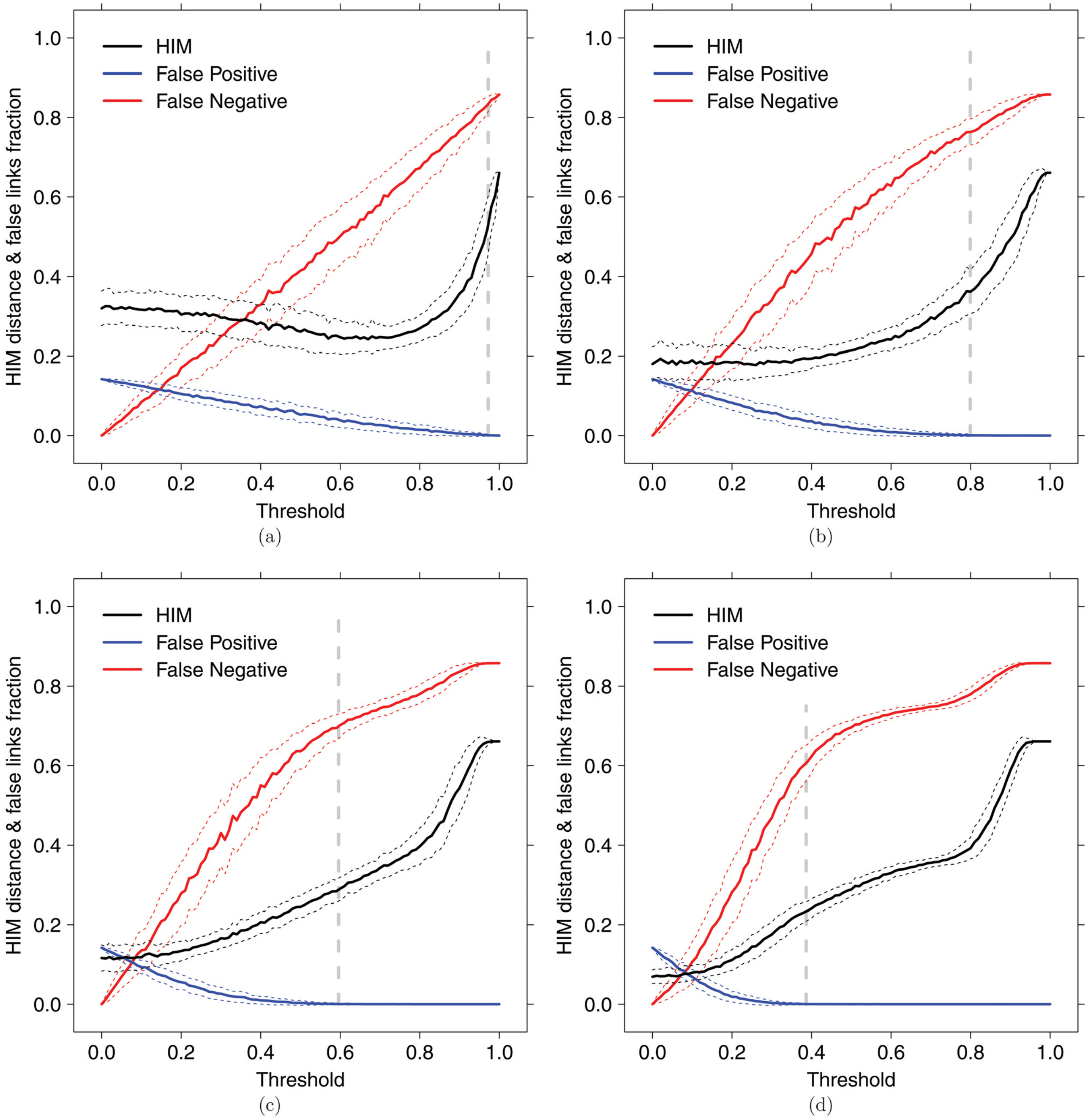
doi:10.1371/journal.pone.0128115.g006

LOC649242, SCGB1D2, CYP4B1, NQO1, MYCL1, PRSS21, MGC13057, PPP1R1B, KIAA1324, LOC646769) and 9 genes strongly associated to poor PFS or poor OS (THBS2, SFRP2, DPSG3, COL11A1, COL10A1, COL8A1, FAP, FABP4, POSTN), thus generating a dataset $\mathcal{O}_T$ of dimension 161 samples and 20 features. The corresponding absolute PCC matrix O_T is then used as the ground truth for the subsampling experiments, thresholding to zero all values smaller than 0.1: the levelplot of O_T and the heatmap of $\mathcal{O}_T$ are displayed in Fig 6.

In these experiments, a random subdataset of n_s samples is extracted from $\mathcal{O}_T$, and the corresponding absolute PCC coexpression network on the nodes T is built, for increasing threshold values. In Fig 7 we report the HIM and the ratio of False Positive and False Negative links for 500 runs of the experiments, separately for $n_s = 5,10,20$ and 50. Again, the secure threshold $\bar{p}$ corresponds to the smallest PCC value warranting that no false positive links are included. Finally, on average, for threshold values greater than $\bar{p}$, the derivative of HIM distance is larger than before $\bar{p}$, while the false negative rate remains under 0.8.

Conclusions

We have proposed a simple a priori, theoretical and non-parametric method for the selection of a hard threshold for the construction of correlation networks. This model is based on the requirements of filtering random data due to noise and reducing the number of false positives,



(a)

(b)

(c)

(d)

Fig 7. Ovarian cancer dataset $\mathcal{O}$. Coexpression inference of the coexpression network from subsampling of $\mathcal{O}_T$, on 5 (a), 10 (b), 20 (c) and 50 (d) samples. Solid lines indicate the mean over 500 replicated instances of HIM distance (black), ratio of False Positive (blue) and ratio of False Negative (red); dotted lines of the same color indicate confidence bars (+/-σ), while grey vertical dashed lines correspond to the secure threshold $\bar{p}$.

doi:10.1371/journal.pone.0128115.g007

and it is implemented by means of analytic properties of PCC. This new approach can be especially useful where there is a small sample size and when the task requires minimising the number of false positive links, probably the most common situation in profiling studies in functional genomics. Finally, when the number of samples increases, coupling this method with soft thresholding approaches, can help recovering false negative links neglected by overly strict thresholds.

Supporting Information

S1 Text. Supplementary Text and Figures.
(PDF)

Author Contributions

Conceived and designed the experiments: AG GJ. Performed the experiments: AG GJ. Analyzed the data: AG GJ. Wrote the paper: AG GJ.

References

1. Barabási AL. The network takeover. Nature Physics. 2012; 8:14–16.
2. Szederkenyi G, Banga J, Alonso A. Inference of complex biological networks: distinguishability issues and optimization-based solutions. BMC Systems Biology. 2011; 5(1):177. doi: 10.1186/1752-0509-5-177 PMID: 22034917
3. He F, Balling R, Zeng AP. Reverse engineering and verification of gene networks: Principles, assumptions, and limitations of present methods and future perspectives. Journal of Biotechnology. 2009; 144:190–203. doi: 10.1016/j.jbiotec.2009.07.013 PMID: 19631244
4. Meyer P, Alexopoulos LG, Bonk T, Califano A, Cho CR, de la Fuente A, et al. Verification of systems biology research in the age of collaborative competition. Nature Biotechnology. 2011; 29(9):811–815. doi: 10.1038/nbt.1968 PMID: 21904331
5. Prill RJ, Marbach D, Saez-Rodriguez J, Sorger PK, Alexopoulos LG, Xue X, et al. Towards a Rigorous Assessment of Systems Biology Models: The DREAM3 Challenges. PLoS ONE. 2010 02; 5(2): e9202. doi: 10.1371/journal.pone.0009202 PMID: 20186320
6. De Smet R, Marchal K. Advantages and limitations of current network inference methods. Nature Reviews Microbiology. 2010; 8(10):717–729. PMID: 20805835
7. Liang M, Zhang F, Jin G, Zhu J. FastGCN: A GPU Accelerated Tool for Fast Gene Co-Expression Networks. PLoS ONE. 2015; 10(1):e0116776. doi: 10.1371/journal.pone.0116776 PMID: 25602758
8. Song Q, Zhao C, Ou S, Meng Z, Kang P, Fan L, et al. Co-expression analysis of differentially expressed genes in hepatitis C virus-induced hepatocellular carcinoma. Molecular Medicine Reports. 2015; 11(1):21–28. doi: 10.3892/mmr.2014.2695 PMID: 25339452
9. Wang S, Pandis I, Johnson D, Emam I, Guitton F, Oehmichen A, et al. Optimising parallel R correlation matrix calculations on gene expression data using MapReduce. BMC Bioinformatics. 2014; 15:35. doi: 10.1186/s12859-014-0351-9
10. Rotival M, Petretto E. Leveraging gene co-expression networks to pinpoint the regulation of complex traits and disease, with a focus on cardiovascular traits. Briefings in Functional Genomics. 2014; 13 (1):66–78. doi: 10.1093/bfgp/elt030 PMID: 23960099
11. Okamura Y, Aoki Y, Obayashi T, Tadaka S, Ito S, Narise T, et al. COXPRESdb in 2015: coexpression database for animal species by DNA-microarray and RNAseq-based expression data with multiple quality assessment systems. Nucleic Acids Research. 2015; 43(D1):D82–D86. doi: 10.1093/nar/gku1163 PMID: 25392420
12. 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. Nature Communications. 2014; 5: Article 3231.
13. López-Kleine L, Leal L, López C. Biostatistical approaches for the reconstruction of gene co-expression networks based on transcriptomic data. Briefings in Functional Genomics. 2013; 12(5):457–467. doi: 10.1093/bfgp/elt003 PMID: 23407269

14. Rau A, Maugis-Rabusseau C, Martin-Magniette ML, Celeux G. Co-expression analysis of high-throughput transcriptome sequencing data with Poisson mixture models. Bioinformatics. 2015; 2015 Jan 5 [Epub ahead of print]:btu845.

15. Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P. Coexpression Analysis of Human Genes Across Many Microarray Data Sets. Genome Research. 2004; 14(6):1085–1094. doi: 10.1101/gr.1910904 PMID: 15173114

16. Lavi O, Dror G, Shamir R. Network-Induced Classification Kernels for Gene Expression Profile Analysis. Journal of Computational Biology. 2012; 19(6):694–709. doi: 10.1089/cmb.2012.0065 PMID: 22697242

17. Rapaport F, Zinovyev A, Dutreix M, Barillot E, Vert JP. Classification of microarray data using gene networks. BMC Bioinformatics. 2007;p. 35. doi: 10.1186/1471-2105-8-35 PMID: 17270037

18. Jansen R, Greenbaum D, Gerstein M. Relating Whole-Genome Expression Data with Protein-Protein Interactions. Genome Research. 2002; 12(1):376. doi: 10.1101/gr.205602

19. Zucknick M, Richardson S, Stronach EA. Comparing the Characteristics of Gene Expression Profiles Derived by Univariate and Multivariate Classification Methods. Statistical Applications in Genetics and Molecular Biology. 2008; 7(1):Article 7. doi: 10.2202/1544-6115.1307 PMID: 18312212

20. Khodakarim S, AlaviMajd H, Zayeri F, Rezaei-Tavirani M, Dehghan-Nayeri N, Tabatabaee SM, et al. Comparison of Univariate and Multivariate Gene Set Analysis in Acute Lymphoblastic Leukemia. Asian Pacific Journal of Cancer Prevention. 2013; 14(3):1629–1633. doi: 10.7314/APJCP.2013.14.3.1629

21. Rahmatallah Y, Emmert-Streib F, Glazko G. Gene Sets Net Correlations Analysis (GSNCA): a multivariate differential coexpression test for gene sets. Bioinformatics. 2014; 30(3):360–368. doi: 10.1093/bioinformatics/btt687 PMID: 24292935

22. Zhi W, Minturn J, Rappaport E, Brodeur G, Li H. Network-based Analysis of Multivariate Gene Expression Data. In: Statistical Methods for Microarray Data Analysis. vol. 972 of Methods in molecular biology. Springer; 2013. p. 121–139.

23. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences. 1998; 95(25):14863–14868. doi: 10.1073/pnas.95.25.14863

24. Kumari S, Nie J, Chen HS, Ma H, Stewart R, Li X, et al. Evaluation of Gene Association Methods for Coexpression Network Construction and Biological Knowledge Discovery. PLoS ONE. 2012; 7(11): e50411. doi: 10.1371/journal.pone.0050411 PMID: 23226279

25. Allen JD, Xie Y, Chen M, Girard L, Xiao G. Comparing Statistical Methods for Constructing Large Scale Gene Networks. PLoS ONE. 2012; 7(1):e29348. doi: 10.1371/journal.pone.0029348 PMID: 22272232

26. Song L, Langfelder P, Horvath S. Comparison of co-expression measures: mutual information, correlation, and model based indices. BMC Bioinformatics. 2012; 13:328. doi: 10.1186/1471-2105-13-328 PMID: 23217028

27. Madhamshettiwar P, Maetschke S, Davis M, Reverter A, Ragan M. Gene regulatory network inference: evaluation and application to ovarian cancer allows the prioritization of drug targets. Genome Medicine. 2012; 4(5):41. doi: 10.1186/gm340 PMID: 22548828

28. Baralla A, Mentzen WI, de la Fuente A. Inferring Gene Networks: Dream or Nightmare? Annals of the New York Academy of Science. 2009; 1158:246–256. PMID: 19348646

29. Carter SL, Brechbühler CM, Griffin M, Bond AT. Gene co-expression network topology provides a framework for molecular characterization of cellular state. Bioinformatics. 2004; 20(14):2242–2250. doi: 10.1093/bioinformatics/bth234 PMID: 15130938

30. Wood LD, Parsons DW, Jones S, Lin J, Sjöblom T, Leary RJ, et al. The Genomic Landscapes of Human Breast and Colorectal Cancers. Science. 2007; 318(5853):1108–1113. doi: 10.1126/science.1145720 PMID: 17932254

31. Carlson M, Zhang B, Fang Z, Mischel P, Horvath S, Nelson S. Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. BMC Genomics. 2006; 7(1):40. doi: 10.1186/1471-2164-7-40 PMID: 16515682

32. Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HYK, Chen R, et al. Personal Omics Profiling Reveals Dynamic Molecular and Medical Phenotypes. Cell. 2012; 148(6):1293–1307. doi: 10.1016/j.cell.2012.02.009 PMID: 22424236

33. Friedman J, Alm EJ. Inferring Correlation Networks from Genomic Survey Data. PLoS Computational Biology. 2012; 8(9):e1002687. doi: 10.1371/journal.pcbi.1002687 PMID: 23028285

34. Furlotte NA, Kang HM, Ye C, Eskin E. Mixed-model coexpression: calculating gene coexpression while accounting for expression heterogeneity. Bioinformatics. 2011; 27(13):i288–i294. doi: 10.1093/bioinformatics/btr221 PMID: 21685083

35. Rider AK, Milenković T, Siwo GH, Pinapati RS, Emrich SJ, Ferdig MT, et al. Networks' characteristics are important for systems biology. Network Science. 2014; 2(02):139–161. doi: 10.1017/nws.2014.13

36. Butte AJ, Kohane IS. Mutual Information Relevance Networks: Functional Genomic Clustering Using Pairwise Entropy Measurements. Pacific Symposium on Biocomputing. 2000; 5:415–426.

37. Wang HQ, Tsai CJ. CorSig: A General Framework for Estimating Statistical Significance of Correlation and Its Application to Gene Co-Expression Analysis. PLoS ONE. 2013; 8(10):e77429. doi: 10.1371/journal.pone.0077429 PMID: 24194884

38. Cho DY, Kim YA, Przytycka TM. Chapter 5: Network Biology Approach to Complex Diseases. PLoS Computational Biology. 2012; 8(12):e1002820. doi: 10.1371/journal.pcbi.1002820 PMID: 23300411

39. Zhang B, Horvath S. A General Framework for Weighted Gene Co-Expression Network Analysis. Statistical Applications in Genetics and Molecular Biology. 2005; 4(1):Article 17. doi: 10.2202/1544-6115.1128 PMID: 16646834

40. Zhang J, Lu K, Xiang Y, Islam M, Kotian S, Kais Z, et al. Weighted Frequent Gene Co-expression Network Mining to Identify Genes Involved in Genome Stability. PLoS Computational Biology. 2012; 8(8):e1002656. doi: 10.1371/journal.pcbi.1002656 PMID: 22956898

41. Gibbs D, Baratt A, Baric R, Kawaoka Y, Smith R, Orwoll E, et al. Protein co-expression network analysis (ProCoNA). Journal of Clinical Bioinformatics. 2013; 3(1):11. doi: 10.1186/2043-9113-3-11 PMID: 23724967

42. de Solla Price DJ. Networks of Scientific Papers. Science. 1965; 149(3683):510–515. doi: 10.1126/science.149.3683.510

43. Barabasi AL, Albert R. Emergence of scaling in random networks. Science. 1999; 286:509–512. doi: 10.1126/science.286.5439.509 PMID: 10521342

44. Davidson GS, Wylie BN, Boyack KW. Cluster Stability and the Use of Noise in Interpretation of Clustering. In: Proceedings of the IEEE Symposium on Information Visualization 2001 INFOVIS'01. IEEE Computer Society; 2001. p. 23.

45. Chen H. Clustering and Network Analysis with Single Nucleotide Polymorphism (SNP) [Ph.D. Thesis]. Stony Brook University; 2011.

46. Numata J, Ebenhöh O, Knapp EW. Measuring correlations in metabolomic networks with mutual information. Genome Informatics. 2008; 20:112–122. doi: 10.1142/9781848163003_0010 PMID: 19425127

47. Fukushima A. DiffCorr: An R package to analyze and visualize differential correlations in biological networks. Gene. 2013; 518(1):209–214. doi: 10.1016/j.gene.2012.11.028 PMID: 23246976

48. Prieto C, Risueño A, Fontanillo C, De Las Rivas J. Human Gene Coexpression Landscape: Confident Network Derived from Tissue Transcriptomic Profiles. PLoS ONE. 2008; 3(12):e3911. doi: 10.1371/journal.pone.0003911 PMID: 19081792

49. Opgen-Rhein R, Strimmer K. From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. BMC Systems Biology. 2007; 1:37. doi: 10.1186/1752-0509-1-37 PMID: 17683609

50. Khanin R, Wit E. Construction of Malaria Gene Expression Network Using Partial Correlations. In: McConnell P, Lin SM, Hurban P, editors. Methods of Microarray Data Analysis V. Springer US; 2007. p. 75–88.

51. Obayashi T, Hayashi S, Shibaoka M, Saeki M, Ohta H, Kinoshita K. COXPRESdb: a database of coexpressed gene networks in mammals. Nucleic Acids Research. 2008; 36(suppl 1):D77–D82. doi: 10.1093/nar/gkm840 PMID: 17932064

52. Ruan J, Dean A, Zhang W. A general co-expression network-based approach to gene expression analysis: comparison and applications. BMC Systems Biology. 2010; 4(1):8. doi: 10.1186/1752-0509-4-8 PMID: 20122284

53. Mistry M, Gillis J, Pavlidis P. Meta-analysis of gene coexpression networks in the post-mortem prefrontal cortex of patients with schizophrenia and unaffected controls. BMC Neuroscience. 2013; 14(1):105. doi: 10.1186/1471-2202-14-105 PMID: 24070017

54. Luo F, Yang Y, Zhong J, Gao H, Khan L, Thompson D, et al. Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory. BMC Bioinformatics. 2007; 8(1):299. doi: 10.1186/1471-2105-8-299 PMID: 17697349

55. Zhu D, Hero AO, Qin ZS, Swaroop A. High throughput screening of co-expressed gene pairs with controlled false discovery rate (FDR) and minimum acceptable strength (MAS). Journal of Computational Biology. 2005; 12(7):1029–1045. doi: 10.1089/cmb.2005.12.1029 PMID: 16201920

56. Scholz M. Approaches to analyse and interpret biological profile data [Ph.D. Thesis]. Potsdam University; 2006.

57. Ma C, Wang X. Application of the Gini Correlation Coefficient to Infer Regulatory Relationships in Transcriptome Analysis. Plant Physiology. 2012; 160(1):192–203. doi: 10.1104/pp.112.201962 PMID: 22797655

58. Caraiani P. Using Complex Networks to Characterize International Business Cycles. PLoS ONE. 2013; 8(3):e58109. doi: 10.1371/journal.pone.0058109 PMID: 23483979

59. Inouye M, Silander K, Hamalainen E, Salomaa V, Harald K, Jousilahti P, et al. An immune response network associated with blood lipid levels. PLoS Genetics. 2010; 6(9):e1001113. doi: 10.1371/journal.pgen.1001113 PMID: 20844574

60. Giorgi FM. Expression-based Reverse Engineering of Plant Transcriptional Networks [Ph.D. Thesis]. Potsdam University; 2011.

61. Usadel B, Obayashi T, Mutwil M, Giorgi FM, Bassel GW, Tanimoto M, et al. Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. Plant, Cell & Environment. 2009; 32(12):1633–1651. doi: 10.1111/j.1365-3040.2009.02040.x

62. Yuan A, Yue Q, Apprey V, Bonney GE. Global pattern of pairwise relationship in genetic network. Journal of Biomedical Science and Engineering. 2010; 3:977–985. doi: 10.4236/jbise.2010.310128 PMID: 21804923

63. Bassel GW, Lan H, Glaab E, Gibbs DJ, Gerjets T, Krasnogor N, et al. Genome-wide network model capturing seed germination reveals coordinated regulation of plant cellular phase transitions. Proceedings of the National Academy of Sciences. 2011; 108(23):9709–9714. doi: 10.1073/pnas.1100958108

64. Zheng ZL, Zhao Y. Transcriptome comparison and gene coexpression network analysis provide a systems view of citrus response to "*Candidatus Liberibacter asiaticus*" infection. BMC Genomics. 2013; 14:27. doi: 10.1186/1471-2164-14-27 PMID: 23324561

65. Stöckel J, Welsh EA, Liberton M, Kunnvakkam R, Aurora R, Pakrasi HB. Global transcriptomic analysis of Cyanothece 51142 reveals robust diurnal oscillation of central metabolic processes. Proceedings of the National Academy of Sciences. 2008; 105(16):6156–6161. doi: 10.1073/pnas.0711068105

66. Dempsey K, Bonasera S, Bastola D, Ali H. A Novel Correlation Networks Approach for the Identification of Gene Targets. In: Proceedings of the 44th Hawaii International Conference on System Sciences—HICSS 2011. IEEE; 2011. p. 1–8.

67. Elo LL, Järvenpää H, Orešič M, Lahesmaa R, Aittokallio T. Systematic construction of gene coexpression networks with applications to human T helper cell differentiation process. Bioinformatics. 2007; 23(16):2096–2103. doi: 10.1093/bioinformatics/btm309 PMID: 17553854

68. Gibson SM, Ficklin SP, Isaacson S, Luo F, Feltus FA, Smith MC. Massive-Scale Gene Co-Expression Network Construction and Robustness Testing Using Random Matrix Theory. PLoS ONE. 2013; 8(2):e55871. doi: 10.1371/journal.pone.0055871 PMID: 23409071

69. Feltus FA, Ficklin SP, Gibson SM, Smith MC. Maximizing capture of gene co-expression relationships through pre-clustering of input expression samples: an *Arabidopsis* case study. BMC Systems Biology. 2013; 7:44. doi: 10.1186/1752-0509-7-44 PMID: 23738693

70. Perkins AD, Langston MA. Threshold selection in gene co-expression networks using spectral graph theory techniques. BMC Bioinformatics. 2009; 10(Suppl 11):1–11. doi: 10.1186/1471-2105-10-S11-S4

71. Stathias V, Pastori C, Griffin TZ, Komotar R, Clarke J, Zhang M, et al. Identifying Glioblastoma Gene Networks Based on Hypergeometric Test Analysis. PLoS ONE. 2014; 9(12):e115842. doi: 10.1371/journal.pone.0115842 PMID: 25551752

72. Borate B, Chesler E, Langston M, Saxton A, Voy B. Comparison of threshold selection methods for microarray gene co-expression matrices. BMC Research Notes. 2009; 2(1):240. doi: 10.1186/1756-0500-2-240 PMID: 19954523

73. Gaiteri C, Sibille E. Differentially expressed genes in major depression reside on the periphery of resilient gene coexpression networks. Frontiers in Neuroscience. 2011; 5:Article 95. doi: 10.3389/fnins.2011.00095 PMID: 21922000

74. Good P. Permutation Tests. Springer; 2000.

75. Fisher RA. Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population. Biometrika. 1915; 10(4):507–521. doi: 10.2307/2331838

76. Bevington PR. Data Reduction and Error Analysis for the Physical Sciences. McGraw-Hill; 1969.

77. Zhang A. Advanced Analysis of Gene Expression Microarray Data. World Scientific; 2006.

78. Casellas J, Varona L. Modeling Skewness in Human Transcriptomes. PLoS ONE. 2012; 7(6):e38919. doi: 10.1371/journal.pone.0038919 PMID: 22701729

79. Doig T, Hume D, Theocharidis T, Goodlad J, Gregory C, Freeman T. Coexpression analysis of large cancer datasets provides insight into the cellular phenotypes of the tumour microenvironment. BMC Genomics. 2013; 14(1):469. doi: 10.1186/1471-2164-14-469 PMID: 23845084

80. Tothill RW, Tinker AV, George J, Brown R, Fox SB, Lade S, et al. Novel Molecular Subtypes of Serous and Endometrioid Ovarian Cancer Linked to Clinical Outcome. Clinical Cancer Research. 2008; 14 (16):5198–5208. doi: 10.1158/1078-0432.CCR-08-0196 PMID: 18698038

81. Budhu A, Jia HL, Forgues M, Liu CG, Goldstein D, Lam A, et al. Identification of Metastasis-Related MicroRNAs in Hepatocellular Carcinoma. Hepatology. 2008; 47(3):897–907. doi: 10.1002/hep.22160 PMID: 18176954

82. Ji J, Shi J, Budhu A, Yu Z, Forgues M, Roessler S, et al. MicroRNA Expression, Survival, and Response to Interferon in Liver Cancer. New England Journal of Medicine. 2009; 361:1437–1447. doi: 10.1056/NEJMoa0901282 PMID: 19812400

83. Troyanskaya OG, Cantor M, Sherlock G, Brown PO, Hastie T, Tibshirani R, et al. Missing value estimation methods for DNA microarrays. Bioinformatics. 2001; 17(6):520–525. doi: 10.1093/bioinformatics/17.6.520 PMID: 11395428

84. Kozomara A, Griffiths-Jones S. miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Research. 2011; 39(suppl 1):D152–D157. doi: 10.1093/nar/gkq1027 PMID: 21037258

85. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology. 1990; 215(3):403–410. doi: 10.1016/S0022-2836(05)80360-2 PMID: 2231712

86. Olkin I, Pratt JW. Unbiased estimation of certain correlation coefficients. Annals of Mathematical Statistics. 1958; 29:201–211. doi: 10.1214/aoms/1177706717

87. Pearson ES, Adyanth ya NK. The distribution of frequency constants in small samples from non-normal symmetrical and skew populations. Biometrika. 1929; 21(1/4):259–286. doi: 10.1093/biomet/21.1-4.259

88. Rider PR. On the distribution of the correlation coefficient in small samples. Biometrika. 1932; 24(3/4):382–403. doi: 10.1093/biomet/24.3-4.382

89. Gayen AK. The Frequency Distribution of the Product-Moment Correlation Coefficient in Random Samples of any Size Drawn from Non-Normal Universes. Biometrika. 1951; 38(1/2):219–247. doi: 10.2307/2332329 PMID: 14848124

90. Kenney JF, Keeping ES. Mathematics of Statistics, Part 2. 2nd ed. Van Nostrand; 1951.

91. Kenney JF, Keeping ES. Mathematics of Statistics, Part 1. 3rd ed. Van Nostrand; 1962.

92. Pugh EM, Winslow GH. The Analysis of Physical Measurements. Addison-Wesley; 1966.

93. Kendall MG, Stuart A. The Advanced Theory of Statistics: Distribution theory. Griffin; 1977.

94. Haldane JBS. A note on non-normal correlation. Biometrika. 1949; 36:467–468. doi: 10.2307/2332685 PMID: 15409360

95. Hey GB. A new method for experimental sampling illustrated in certain non-normal populations. Biometrika. 1938; 30:68–80. doi: 10.2307/2332225

96. Kowalski CJ. On the Effects of Non-Normality on the Distribution of the Sample Product-Moment Correlation Coefficient. Journal of the Royal Statistical Society Series C (Applied Statistics). 1972; 21 (1):1–12.

97. Erdös P, Rényi A. On Random Graphs. I. Publicationes Mathematicae. 1959; 6:290–297.

98. Hamza K. The smallest uniform upper bound on the distance between the mean and the median of the binomial and Poisson distributions. Statistics & Probability Letters. 1995; 23(1):21–25. doi: 10.1016/0167-7152(94)00090-U

99. Bonferroni CE. Teoria statistica delle classi e calcolo delle probabilità. Pubblicazioni del Regio Istituto Superiore di Scienze Economiche e Commerciali di Firenze. 1936; 8:3–62.

100. Miller RGJ. Simultaneous Statistical Inference. Springer; 1981.

101. Fisher RA. On the "probable error" of a coefficient of correlation deduced from a small sample. Metron. 1921; 1:3–32.

102. Sealy Gossetm WS. The probable error of a mean. Biometrika. 1908; 6(1):1–25. doi: 10.2307/2331554

103. Soper HE, Young AW, Cave BM, Lee A, Pearson K. On the distribution of the correlation coefficient in small samples. Appendix II to the papers of "Student" and R. A. Fisher. A co-operative study. Biometrika. 1917; 11:328–413. doi: 10.2307/2331830

104. Fisher RA. Applications of "Student's" distribution. Metron. 1925; 5:90–104.

105. Rahman NA. A Course in Theoretical Statistics. Charles Griffin and Company; 1968.

106. Kendall MG, Stuart A. The Advanced Theory of Statistics, Volume 2: Inference and Relationship. Charles Griffin and Company; 1973.

107. Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. British Medical Journal. 1995; 310(6973):170. doi: 10.1136/bmj.310.6973.170 PMID: 7833759

108. Perneger TV. What's wrong with Bonferroni adjustments. British Medical Journal. 1998; 316 (7139):1236–1238. doi: 10.1136/bmj.316.7139.1236 PMID: 9553006

109. Azuaje FJ. Selecting biologically informative genes in co-expression networks with a centrality score. Biology Direct. 2014; 9:12. doi: 10.1186/1745-6150-9-12 PMID: 24947308

110. Dudoit S, van der Laan MJ. Multiple Testing Procedures with Applications to Genomics. Springer; 2008.

111. Jurman G, Visintainer R, Riccadonna S, Filosi M, Furlanello C. The HIM glocal metric and kernel for network comparison and classification; 2013. ArXiv:1201.2931 [math.CO].

112. Filosi M, Visintainer R, Riccadonna S, Jurman G, Furlanello C. Stability Indicators in Network Reconstruction. PLoS ONE. 2014; 9(2):e89815. doi: 10.1371/journal.pone.0089815 PMID: 24587057