Computers in Biology and Medicine 135 (2021) 104567

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

SEaCorAl: Identifying and contrasting the regulation-correlation bias in RNA-Seq paired expression data of patient groups

Manuela Petti [a], Antonella Verrienti [b], Paola Paci [a], Lorenzo Farina [a,*]

[a] Department of Computer, Control and Management Engineering, Sapienza University of Rome, Italy
[b] Department of Translational and Precision Medicine, Sapienza University of Rome, Italy

ARTICLE INFO

Keywords:
Correlation networks
Correlation analysis
Spurious correlations
RNA-Seq data
Paired data

ABSTRACT

The Cancer Genome Atlas database offers the possibility of analyzing genome-wide expression RNA-Seq cancer data using paired counts, that is, studies where expression data are collected in pairs of normal and cancer cells, by taking samples from the same individual. Correlation of gene expression profiles is the most common analysis to study co-expression groups, which is used to find biological interpretation of -omics big data. The aim of the paper is threefold: firstly we show for the first time, the presence of a "regulation-correlation bias" in RNA-Seq paired expression data, that is an artifactual link between the expression status (up- or down-regulation) of a gene pair and the sign of the corresponding correlation coefficient. Secondly, we provide a statistical model able to theoretically explain the reasons for the presence of such a bias. Thirdly, we present a bias-removal algorithm, called SEaCorAl, able to effectively reduce bias effects and improve the biological significance of correlation analysis. Validation of the SEaCorAl algorithm is performed by showing a significant increase in the ability to detect biologically meaningful associations of positive correlations and a significant increase of the modularity of the resulting unbiased correlation network.

1. Introduction

The advent of RNA-seq studies has revolutionized the field of gene expression data analysis allowing a sequencing-based technology able to provide more precise and reliable quantification of relative RNA levels. Such new technology avoids many limitations of microarrays, such as the possibility of studying alternative splicing and isoform expression with low background noise and a much larger range of values [1]. Gene expression data are often used to detect differentially expressed genes between two conditions to obtain information on the genes that are involved in the biological process of interest.

Recently, the importance of design studies where expression data are collected in pairs, *e.g.,* in normal and cancer cells, by taking samples from the same individual, has been addressed in the relevant literature (see, for example [2], and the references cited therein). Noteworthy, The Cancer Genome Atlas database (TCGA) [3], among others, offers the possibility of analyzing genome-wide expression data (including miRNAs and other non-coding RNAs) using paired counts. The advantages are many and highly significant in biological terms: paired data allows to mitigate the effects of the high biological and technical noise and to better define a patient-specific gene expression profile (signature), potentially able to characterize the specific condition of a single patient by providing valuable information on the disease state and progression [4]. Moreover, it has been proved that paired data studies substantially increase the statistical power of the analysis [5].

Most importantly, a disease can be molecularly characterized with higher accuracy for every single individual by considering its "log-fold change" value along genes, that is the $\log_2$ of the ratio between expression values in cancer and in normal conditions. It is also worth noting that the use of $\log_2$-fold change values can effectively mitigate the bias caused by the compositional nature of RNA-seq data [6], which is another good reason to use paired data in gene expression analysis. Indeed, the disease-specific "gene expression profile" is at the root of precision medicine that, by integrating many sources of omics and clinical data such as the protein-protein interaction network [7], aims to provide a personalized treatment based on such specific molecular signature [8].

In many diverse research areas of life sciences like plant sciences, pharmacology, oncology, etc. a routine analysis is the study of changes in gene expression along different conditions. Such conditions include treatment groups [9], time series kinetics [10,11], cancer development [12], mutant analysis [13], stress response [14] and many others. Here

* Corresponding author.
E-mail address: lorenzo.farina@uniroma1.it (L. Farina).

https://doi.org/10.1016/j.compbiomed.2021.104567
Received 28 April 2021; Received in revised form 27 May 2021; Accepted 8 June 2021
Available online 15 June 2021

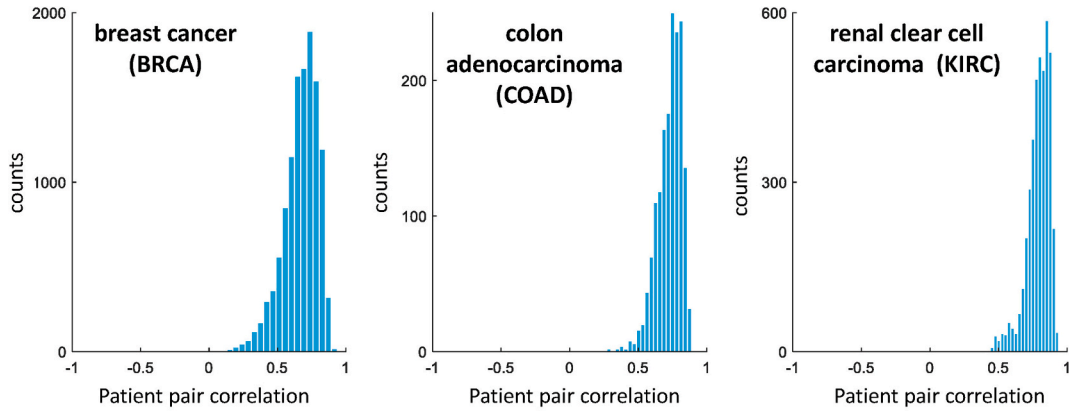

Fig. 1. Patients' expression profiles are positively correlated. Distribution of the Pearson's correlation between pairs of patients in the same pathological condition. Normal and cancer cells for: A) breast cancer (BRCA) B) colon adenocarcinoma (COAD) and C) renal clear cell carcinoma (KIRC). Data are taken from the TCGA database.

we focused on genome-wide gene expression data obtained by groups of patients in two conditions, like, for example, pre- and post-treatment or normal and transformed tissues (as in the TCGA database). The available data are usually organized in rows (genes) and columns (patients), in the two conditions under investigation.

Correlation of gene expression profiles is probably the most common analysis to study co-expression [15] and it is usually computed using Pearson's or Spearman's coefficient on samples taken in several different conditions. It is worth noting that the choice of the appropriate correlation measure for the case of interest is a very important topic. For example, significant improvement of correlation (SIC) method has been successfully obtained by Iqbal et al. [16] for the identification of the effect of a drug on cell image. The construction of large gene co-expression networks [17] is a widely used data analysis technique [18–20] based on the 'guilt-by-association' assumption that a high (positive or negative) correlation value between pairs of gene expression profiles, may indicate the presence of a common underlying mechanism where genes are involved in the same biological process or function [21].

When performing correlation analysis, one must keep in mind a very important distinction between the case of comparing different organisms in the same condition (*e.g.,* patients having the same disease phenotype) and that of comparing different conditions for the same organism (*e.g.,* the same cell line or tissue treated with different compounds). In the first case, we expect the gene expression profile along genes for each patient in the same condition to be positively correlated,

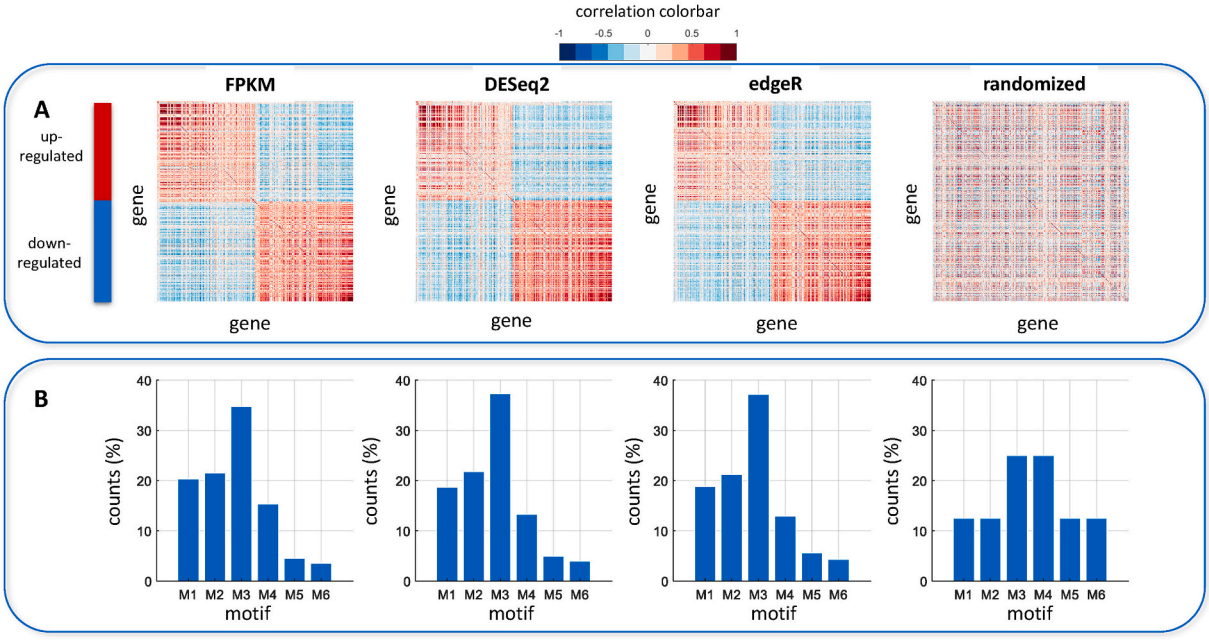

Fig. 2. The correlation matrix heatmap shows a strong relationship between regulation (mean log-fold change) and correlation. A) The four heatmaps represent the Spearman correlation values between genes using three different normalization methods (FPKM, DESeq2, and edgeR) for breast cancer TCGA data and a randomized data matrix. Genes are arranged in the order of decreasing values of M2LFC, so that the first (upper) block represents the upregulated genes and the second (lower) block represents downregulated genes. The picture makes it clear that a strong pattern due to a relationship between M2LFC and correlation, is present. Pairs of up/up or down/down-regulated genes appear to be predominantly positively correlated, whilst pairs of up/down-regulated genes appear to be predominantly negatively correlated. This pattern is independent of the normalization method used for RNA-seq data (FPKM, DESeq2, and edgeR). B) The bar plots report the percentage of the six motifs for each of the four cases considered in panel A. The first three motifs account for about 80% of all the motifs for any normalization method. The comparison with the motif distribution of the randomized matrix case (bottom-right sub-panel) shows that the first three motifs account for 50% of all the motifs distribution. By contrast, the distribution in real data is dominated by the first three motifs.

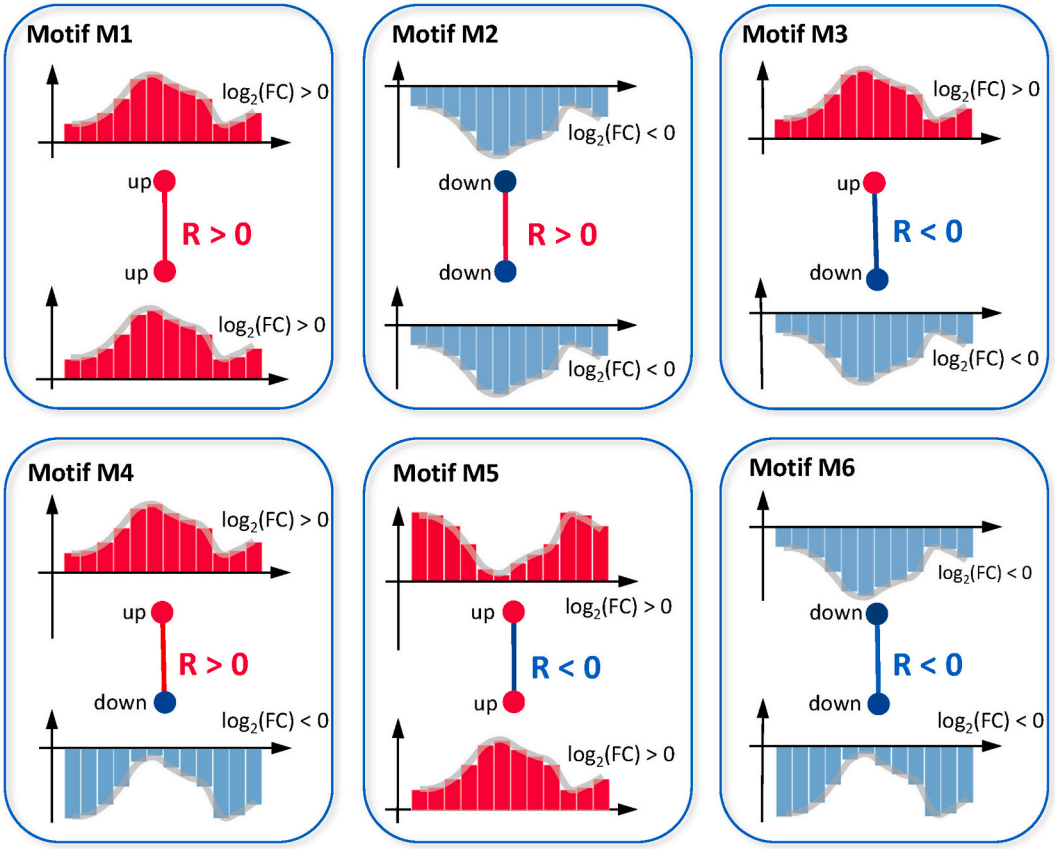

Fig. 3. Differential expression is (*a priori*) independent of correlation. The six motifs M1-M6 of all possible combinations of up/down regulation and positive/negative correlation are represented. Note that it is possible to uniquely classify each link of a correlation network and, therefore, an associated network decomposition can be performed. By contrast, a node may belong to more than one motif.

since the underlying assumption is that there is a "typical" profile (molecular signature) for each specific disease phenotype [22] and that differences among individuals are mainly due to various sources of technical and biological inter-personal variation. By contrast, in the second case, we do not expect a particular sign for the correlation values between pairs of conditions, simply because the assumption is that a different response for each condition is awaited. In other words, when studying – as in our case – groups of patients having the same disease in two conditions (*e.g.,* normal/cancer cells), we can assume the distribution of the correlation between pairs of patients in the same condition to be highly skewed toward positive values. Indeed, this is exactly the case in real data, as shown for the sake of illustration in Fig. 1, for three cancers gene expression values taken from the TCGA database.

The identification of differentially expressed genes (DEGs), *i.e.,* genes with a high and significant absolute value in the $\log_2$ fold change, and the finding of positively correlated subnetworks (often called "modules" [18]), provides valuable information on a biological process under study but from a very different perspective. DEGs are considered "relevant" under the assumption that "large" changes in gene expression, say from normal to cancer, might be good indicators of their active role in the biological process under investigation. By contrast, highly correlated groups of genes may – or may not – be highly expressed, since correlation values depend exclusively on the shape – not on the magnitude – of the gene expression profile across different patients. In other words, correlation reveals biologically meaningful links among genes by exploiting the inherent interpersonal variability, whilst differential expression utility is based on the amplitude (as opposed to shape) of the gene expression value (typically the mean of the $\log_2$ fold changes). Having this important difference in mind, one may ask whether there is a specific relationship between correlation and differential expression (regulation). The general answer is no. As a matter of fact, it is straightforward to think of situations in which all possible combinations of up/down regulated and positive/negative correlated genes are conceivable. In other words, we cannot state *a priori* that expression and co-expression are linked and, therefore, we must assume in general that co-expression (usually measured by correlation) and expression levels of two generic genes are independent. Consequently, the presence of a link between the two is worth investigating from a biological perspective.

As stated above, given the *a priori* independence of differential expression and correlation and, most importantly, their different interpretation, it is very important to study them together to gain synergistic biological insight using their integration. Not surprisingly, the study of the relationship between correlation and expression has recently attracted some researchers [23–30].

As stated above, here we considered RNA-seq paired data and computed the $\log_2$-fold change of cancer *vs.* normal cells for each transcript and each patient. Moreover, we evaluated the relationship (if any) between the mean $\log_2$-fold change (ML2FC) of two given genes and their correlation values using the Spearman's coefficient (details on the data and methods used in this analysis, normalization and preprocessing are reported in the "Methods" section), which is known to be robust to outlier presence and effective also for non-normal distributions. A very simple analysis that can be readily performed is to consider, for example, the top 500 upregulated the top 500 downregulated genes and draw the heatmap of the $\log_2$-fold change (ratio) by sorting genes in order of decreasing ML2FC values. The resulting plots are reported in Fig. 2A.

The heatmaps in the figure point towards the presence in the data of a large bias due to the strong relationship between the ML2FC of two genes and their correlation sign. Precisely, if the two given genes are

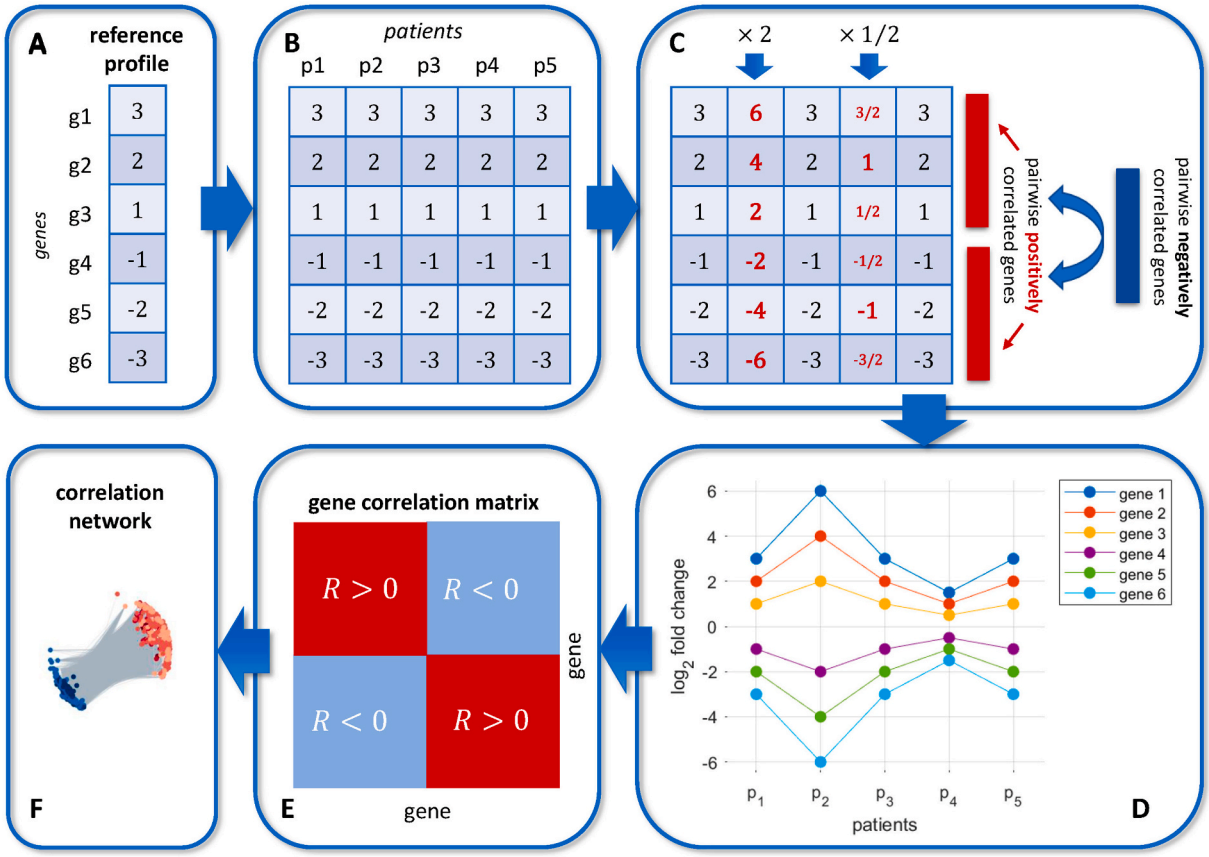

Fig. 4. A toy story: the idealized mechanism of regulation-correlation bias formation. A) the panel shows the prototypical (reference) gene expression profile of the ML2FC using paired data, assuming to have three upregulated genes (positive ML2FC) and three downregulated genes (negative ML2FC). B) Panel B illustrates the building of an ideal patients' group by aggregation, assuming that the resulting *in silico* patients have the same disease phenotype characterized by the profile depicted in panel A. By construction, the correlation between all pairs of patients is equal to one. C) An example of the mechanism producing the bias that affects the data: two patients profiles (columns) are multiplied by a positive constant value, say 2 for the first, and 0.5 for the second one. The effects of this multiplication of the entire column (patient) are to generate correlations between rows, positive values between up- or down-regulated genes, and negative values between up and down-regulated genes. D) The idealized data matrix is plotted on an x/y axis where patients are represented on the x-axis and ML2FC values are represented on the y-axis. The type of correlation among genes described above appears more clearly. E) The panel shows the resulting gene/gene correlation matrix where in position (i,j) the correlation value between gene i and gene j is present. The resulting symmetric correlation matrix pattern is the same as that observed in real data (see Fig. 2A). F) The correlation network: nodes are genes and the red color represents up-regulated genes, the blue color represents down-regulated genes and grey links represent negative correlations.

both up- or both down-regulated, their correlation is likely to be positive. Otherwise, if they have an opposite regulation (one is up the other is down, or *vice versa*), a negative correlation is expected instead. We have called this property the "regulation-correlation bias". Now, inevitably the question arises whether this is a biological feature or an artifact of the data. Here, it is formulated for the first time the hypothesis that this is indeed an artifact, *i.e.*, a bias in the data which is tightly related to the spurious correlation of ratios effect [31], identified by Karl Pearson since 1897 [32]. To support this interpretation, a mechanism that may generate this bias is presented here together with the SEaCorAl (biaS of rEgulAtion-CORrelation removAL) algorithm, able to significantly reduce such bias in RNA-seq paired data of patients' groups. Finally, validation of the SEaCorAl algorithm is performed by showing a significant increase in the ability to detect biologically meaningful associations of positive correlations and a significant increase of the modularity of the resulting unbiased correlation network.

2. Methods

In the previous introductory section, we discussed the properties of the correlation matrix resulting from RNA-seq paired data of groups of patients. The basic features described above can be summarized as follows: 1) patients' profiles are positively correlated and 2) the correlation sign between pairs of genes is tightly linked to their gene expression *status* (up- or down-regulated), *i.e.*, the regulation-correlation bias is present. The first is certainly a feature with strong biological roots. There it is widely agreed that the gene expression profile – although being an incomplete picture in time and space of the cell condition – is highly informative and specific of a given disease state. Indeed, as reported by Ross et al. [22], gene expression of 60 cell lines and approximately 8000 human genes were collected, and cell lines with common tissue of origin showed similar gene expression profiles. Interestingly, cell lines derived from non-small lung carcinoma and breast tumors were scattered across different branches of the dendrogram, suggesting a heterogeneous expression pattern [22]. From a broader perspective, molecular pathology is now become indispensable to inform complex disease diagnosis, prognosis, and therapeutic strategies in day-to-day clinical practice. For example, the use of next-generation sequencing technologies for molecular profiling is having a deep impact in virtually all fields of medical research where physicians are challenged with the complexity of data interpretation.

As regards the second property, *i.e.*, the regulation-correlation bias, here we support the hypothesis that it is the result of a bias in gene expression data, independent of the RNA-Seq normalization used (see Fig. 2A) and resulting in an artifactual relationship between regulation (ML2FC) and correlation. The correlation-regulation bias can be well

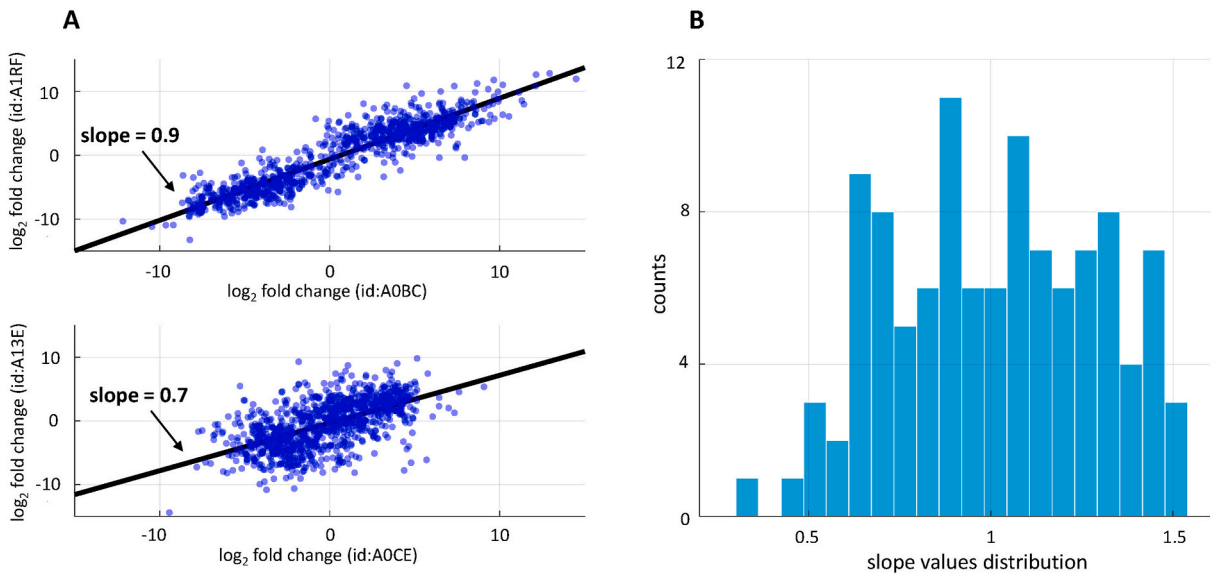

Fig. 5. Different slopes cause the regulation-correlation bias. Using real data (breast cancer, TCGA), the picture shows A) an example of two pairs of patients whose regression lines have different slopes. Such difference can cause the kind of bias discussed here and illustrated by Fig. 2A. B) The panel shows the slope distribution of all patient pairs. It is worth noting that the slopes significantly deviate from the unit value as they range from 0.3 to 1.5.

described using a 2-node network motif. In fact, in a signed network (*i.e.*, links with a sign) composed of signed nodes (the sign of a node is that of its ML2FC), a classification of all the possible network motifs, depicted in Fig. 3, can be obtained as follows:

- **Motif M1.** ↑↑+: upregulated/upregulated/positively correlated.
- **Motif M2.** ↓↓+: downregulated/downregulated/positively correlated.
- **Motif M3.** ↑↓− or ↓↑−: upregulated/downregulated/negatively correlated.
- **Motif M4.** ↑↓+ or ↓↑+: upregulated/downregulated/positively correlated.
- **Motif M5.** ↑↑−: upregulated/upregulated/negatively correlated.
- **Motif M6.** ↓↓−: downregulated/downregulated/negatively correlated

Using these basic building blocks, we can perform a network decomposition by classifying each link in the network as one of the six possible motifs (obviously, nodes may belong to more than one motif). Most notably, as illustrated in Fig. 2B, in a real correlation network, the motifs are usually not equally represented, since some of them (M1, M2, and M3) occur more often than the others.

To prove the existence of this bias in RNA-seq paired data, we present a bias generation statistical consistent with the available experimental evidence previously shown. Using such a model, we define a bias removal (or reducing) algorithm (called SEaCorAl). Finally, to prove its effectiveness, we provide two validations using the biological properties provided by gene annotations (gene ontology biological process, molecular function, and cellular component, KEGG pathways, GSEA collection) and the modularity structure of the correlation network (discussed in the "Results and discussion" section).

Identifying the regulation-correlation bias: the bias generation statistical model.

The artifactual mechanism that may generate the above-discussed bias in RNA-seq paired data (and the corresponding removal algorithm) must be able to explain the following facts highlighted in the previous section: 1) patients' profiles are positively correlated, 2) motifs M1 (↑↑+), M2 (↓↓+) and M3 (↑↓− or ↓↑−) are largely over-represented. To show the hypothesized underlying bias generating mechanism, we start with an idealized process illustrated in Fig. 4.

Such a statistical model hypothesizes that the main cause for the regulation-correlation bias is due to the fact that the fold-change values along all genes (profile) for pairs of patients are not proportional one to the other, with the same proportionality constant. To clarify this key point, one can consider the scatterplot of two pairs of patients: if the slope of the regression line is not equal, the kind of bias shown in Fig. 2A shows up. Indeed, this is exactly what can be seen in real data, as in the illustrative example depicted in Fig. 5.

Most importantly, we note that differences in the proportionality constant between pairs of patients' profiles are biologically implausible. It does not make sense that the profile of a patient is such that its positive expression values (upregulation) are, say, the double of another patient with the same disease and, at the same time, its negative values (downregulation) are the double negative. This line of reasoning immediately leads us to the obvious conclusion that the hypothesis of a "reference" (or prototypical) gene expression profile of a disease phenotype, implies the proportionality constant between any pair of patient's profiles to be equal to one. Indeed, the evidence of a molecular signature characterizing complex diseases, like cancer, auto-immune diseases, and diabetes, suggests that relevant genes must be expressed at the same levels in patients having the same pathological phenotype.

As a final comment, we note that the proposed bias mechanism can simultaneously explain the presence of both positive and negative spurious correlations and the strong relationship between regulation and correlation observed in real data, *i.e.*, what we have called the regulation-correlation bias. Spurious correlations arise from the presence of a different expression level distribution across patients, and the regulation-correlation bias shows up due to the presence of negative and positive values obtained by computing the logarithm of the fold ratio. Upregulated genes result in positive values and, down-regulated ones, result in negative values and, consequently, the sign of the expression change leads to both positive and negative correlation, as pictorially illustrated by Fig. 4.

2.1. *Contrasting the regulation-correlation bias: the SEaCorAl algorithm*

To try and develop a methodology to contrast the impact of the regulation-correlation bias just identified, we need to formalize the idealized process described in the previous section and derive an algorithm for effective bias removal. The formal steps to the generation of

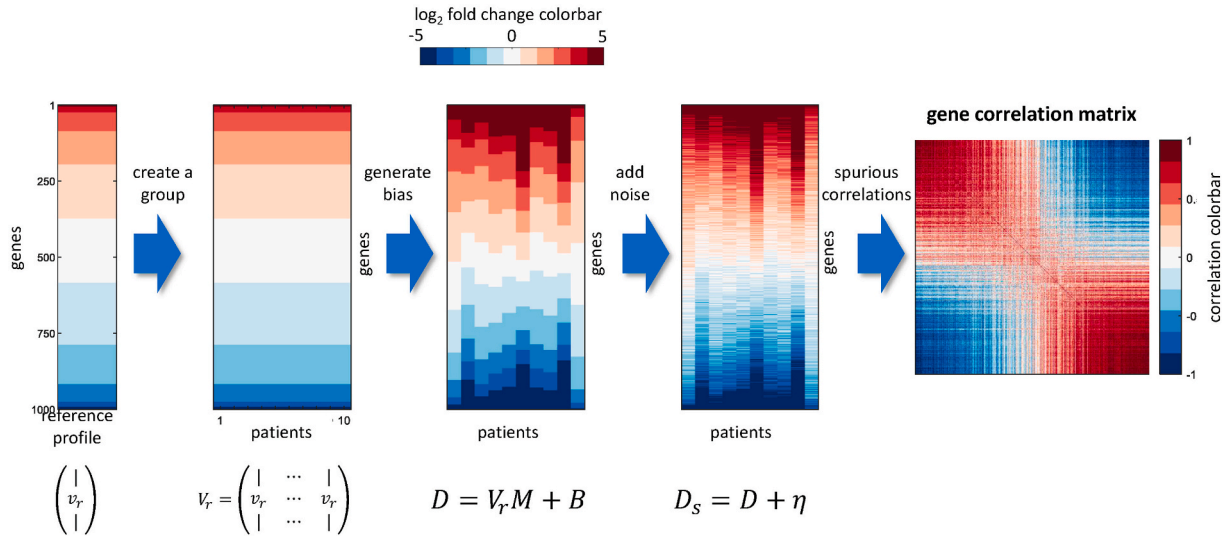

Fig. 6. The bias generation model assumed by the SEaCorAl algorithm. As discussed in the text, the bias generation process starts (step 1) with a single reference profile (or reference pseudo-patient) and obtains the data for a group of patients by aggregation of identical profiles (step 2). Then, each column of the aggregated matrix V_r is multiplied by an (unknown) constant value to be estimated from the data. We considered, to gain generality, a linear affine transformation, *i.e.*, a multiplication of the profile plus a constant value (step 3). Formally, this transformation is defined by multiplying the data matrix V_r by a diagonal matrix M and by adding a diagonal matrix B, where M and B must be estimated from data. Then, a noise term is introduced to account for differences from patients (step four) to get the final *in silico* data matrix. The last panel shows the correlation matrix resulting from the artificial data matrix D_s which resembles those shown in Fig. 2A.

the bias, starting from a reference disease-specific gene expression profile, are illustrated in Fig. 6 and fully described in the next paragraph.

To obtain the SEaCorAl algorithm and try to remove the bias, we first formally defined each step of the process and then derived a formula that provides a correction of the original data aiming to obtain unbiased data to be used for correlation analysis. Let n be the number of genes and p the number of patients available. The reference profile, to be estimated from data, is the following:

$$\begin{pmatrix} | \\ v_r \\ | \end{pmatrix} \in R^{n\times 1}$$

which is a vector of length n and each element represents the expression change characterizing an ideal patient affected by the disease of interest, *i.e.*, its molecular profiling. To model the real situation in which data from several patients are available, we simply use the same vector to represent each of them, aiming to obtain an idealized data matrix (no

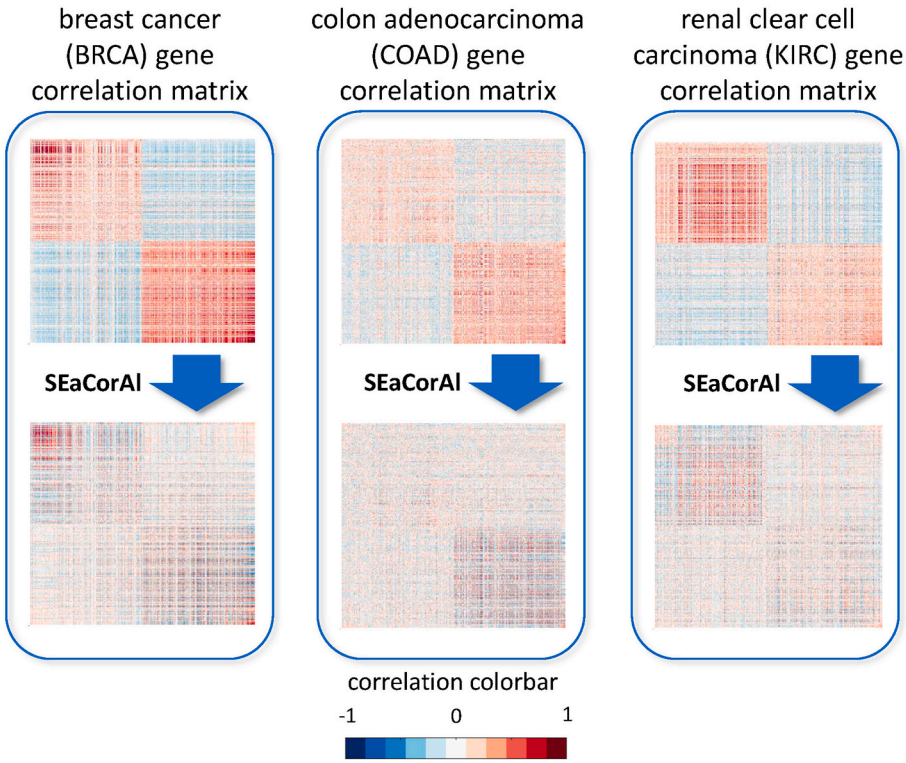

Fig. 7. The "cleaning" effects of the SEaCorAl algorithm on the gene correlation matrix. The figure shows the result of the proposed regulation-correlation bias removal algorithm on three cancers dataset of RNA-Seq paired gene expression values. Precisely, paired fold change data from breast cancer (BRCA), colon adenocarcinoma (COAD), and renal clear cell carcinoma (KIRC) are reported. In all cases, visual inspection immediately confirms that the regulation-correlation relationship is highly reduced by the application of the un-biasing procedure. Two biological validations of the SEaCorAl algorithm for all datasets are provided in the "Results and discussion" section.

interpersonal variation included at this stage). Accordingly, the aggregated reference data matrix is composed of p identical reference profiles vectors as follows:

$$V_r = \begin{pmatrix} | & \cdots & | \\ v_r & \cdots & v_r \\ | & \cdots & | \end{pmatrix} = V_r \in R^{n\times p}$$

thus obtaining a matrix with p identical columns, one for each available patient. Then, we introduce an additive stochastic term η to account for inter-personal variability, a multiplicative coefficient diagonal matrix M and an additive term B to account for differences among columns (patients). Formally, this step can be expressed as an affine linear transformation of the reference data matrix, that is:

$$D_s = V_r M + B + \eta \tag{1}$$

where

$$M = \begin{pmatrix} m_1 & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & m_p \end{pmatrix} \in R^{p\times p},\ B = \begin{pmatrix} b_1 & \cdots & b_p \\ \vdots & \cdots & \vdots \\ b_1 & \cdots & b_p \end{pmatrix} \in R^{n\times p}$$

with $m_1, \cdots, m_p > 0, b_1, \cdots, b_p \geq 0$. Matrices M and B are unknown and must be estimated from the available gene expression data matrix D:

$$D = \begin{pmatrix} | & \cdots & | \\ d_1 & \cdots & d_p \\ | & \cdots & | \end{pmatrix}$$

containing the $\log_2$ fold change paired data. As an estimation of the reference vector v_r, we averaged expression change values of all available patients from data:

$$\overline{v}_r = \frac{1}{p}\sum_{i=1}^{p} d_i$$

Then, we replaced the theoretical data matrix D_s in equation (1) with the real data matrix D to obtain an estimate of the unknown matrices M and B using a least square estimate. Precisely, we considered each column of equation (1):

$$d_i = m_i \overline{v}_r + b_i + \eta_i \tag{2}$$

and obtained least-square estimates of the slopes m_i and of the intercepts b_i. By doing so for any i, we got estimates $\widetilde{M}$ and $\widetilde{B}$ of matrices M and B. Then, using such estimated matrices we could obtain from equation (1) the reference matrix V_r, *i.e.*, the unbiased data matrix D_u. In conclusion, by inverting equation (1), we have:

$$D_u = \left(D - \widetilde{B}\right)\widetilde{M}^{-1} \tag{3}$$

where $\widetilde{M}$ and $\widetilde{B}$ are the (least square) estimates of matrices M and B obtained from equation (2). Formulas (2) and (3) define the SEaCorAl algorithm which consists of solving (1) for any i and obtain matrices $\widetilde{M}$ and $\widetilde{B}$ and then using (3) to get the unbiased data matrix D_u to be used for subsequent correlation analysis. By construction, matrix D_u is such that the proportionality constant between any pair of patients is the same and equal to one. To have a preliminary qualitative view of the Results produced by the algorithm, three illustrative examples of the application of the SEaCorAl algorithm to cancer data from TCGA, are reported in Fig. 7, where the "cleaning" effect of the methodology on the original gene correlation matrix, is visible. The next section is devoted to a systematic validation of the algorithm using 10 cancer datasets from TCGA database.

3. Results and discussion

3.1. First validation: associations of positive correlations with gene function

Many gene expression studies show that positive correlations between profiles are much more common than negative correlations and that they are likely to be associated with functional relatedness [33]. For example, in Ref. [34] 60 large human data sets have been collected and functional relevance of positive correlations has been reliably detected. They found a substantial number of positively correlated expression patterns occurring in multiple independent data sets. Positive correlations between pairs of gene expression profiles indicate the tendency of a "cooperative" behavior that may be due to several reasons. For example, a positive correlation is often observed when gene products are involved in the same biological process, as in the case of enzymes needed to activate a specific metabolic pathway, or when correlated genes code for subunits of the same protein complex.

In this section, given the above mentioned biological significance of positive correlations, we describe a first validation of the SEaCorAl algorithm by considering that, if the un-biasing procedure defined by the SEaCorAl algorithm is effective, a statistically significant increase of the association between positive correlations of a genes pair and their common functional annotations, should be obtained. In other words, the ability of the algorithm to remove spurious correlations can be evaluated by measuring the percentage of positively correlated genes having a common functional annotation. Such percentage must be higher in the unbiased data than in the original one to prove the effectiveness of the SEaCorAl algorithm. To this end, we considered five common annotations: gene ontology [35] biological process (BP), molecular function (MF), cellular component (CC), the KEGG pathways database [36], and the GSEA molecular signature defined by the so-called "hallmark gene set" [37].

The most widely used measure of gene co-expression is the Pearson or Spearman correlation coefficient that quantifies the extent to which genes increase or decrease together across patient expression change values. A positive value is expected when such values increase or decrease in parallel and, a negative value is expected when an opposite behavior is present. Both Pearson and Spearman correlation coefficients measure the strength and direction of association between the gene pair of interest. Here, to measure co-expression, we used the Spearman correlation coefficient which assesses monotonic relationships and, as such, less sensitive to non-normal distributions and outliers. As a first step of the validation procedure, we considered all pairs of genes having a significant Spearman's positive correlation (adjusted p-value less than 0.05) and computed the percentage of such gene pairs with a common annotation, for each of the five types mentioned above. We, therefore, obtained a percentage of successes in predicting a common annotation from a significant Spearman's positive correlation, using the original and the unbiased dataset (SEaCorAl), and computed the associated variation between those two percentages.

As a second step of the validation procedure, to assess the statistical significance of the observed increase, we needed to compare percentages associated with lists of different lengths. Using the same threshold for positive correlations (adjusted p-value less than 0.05), we got (as expected) a larger number of pairs in the original dataset than in the unbiased one. Let T_B be the number of significant Spearman's positive correlations for the original dataset and P_B be the number of those pairs having a common given annotation for the original dataset. Analogously, we defined T_{UB} and P_{UB} for the un-biased (SEaCorAl) dataset, and in all cases considered, we found that $T_{UB} < T_B$ but we obtained better predictions for the un-biased case (SEaCorAl), *i.e.*, $P_{UB}/T_{UB} > P_B/T_B$ for all datasets. To evaluate a p-value, we repeated 100.000 times the random sampling of T_{UB} elements drawn from the original list of T_B elements, obtaining for each iteration i, a number R_i of annotated pairs. Then, we computed the fraction (percentage) of success R_i/T_{UB} and

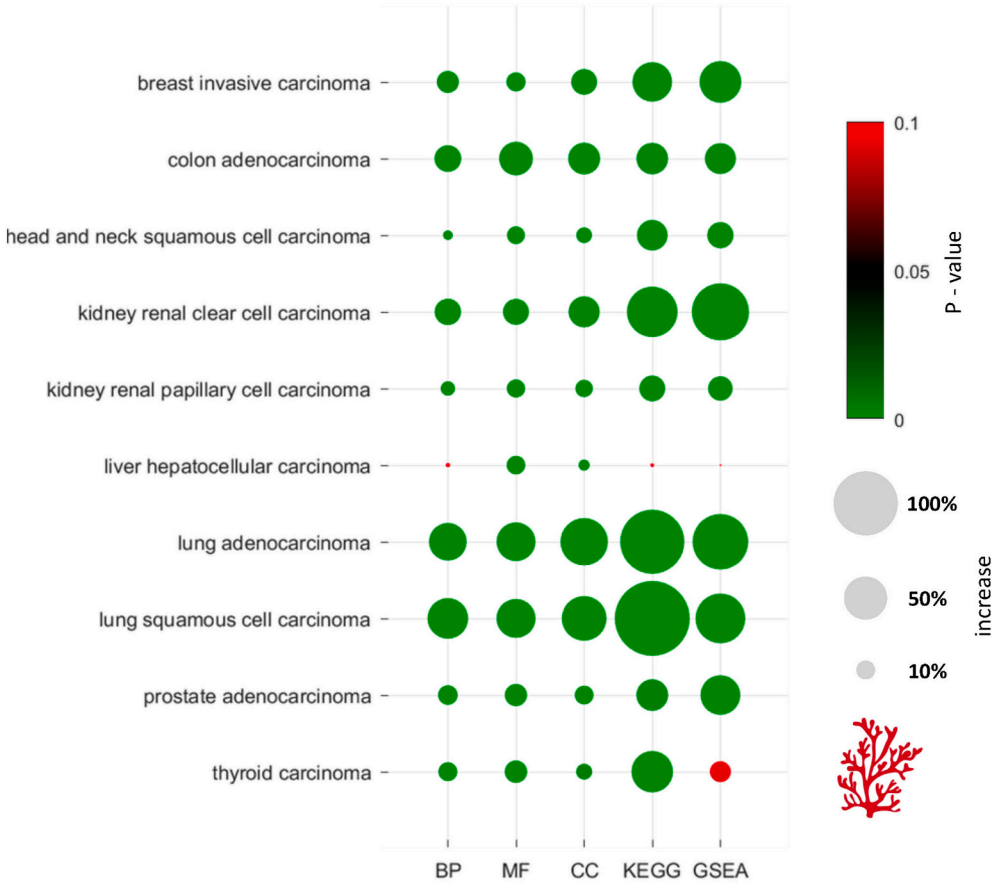

Fig. 8. First validation of SEaCorAl: increased biological significance of Spearman's positive correlations. The figure shows a dot plot where, for each TCGA cancer dataset considered, the percentual increase of annotations found in positive correlations is represented by the size of the dots. The color of the dots represents p-values. The picture makes clear that the SEaCorAl algorithm greatly improves the biologically significant relationship between the presence of a positive correlation between two genes and the presence of common functionality.

therefore obtained a distribution of values from which we derived mean μ and variance σ by sample estimates. Finally, a p-value was computed, assuming a normal distribution, using the usual formula:

$$p = \int_{X}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

where

$$X = \frac{P_{UB}/T_{UB} - \mu}{\sigma}$$

The Results, reported in Fig. 8, clearly show that the percentual variation of successful prediction is always an increase, for all datasets and annotations, *i.e.*, they are significantly greater in the unbiased than in the original dataset. Moreover, for all types of annotations, significances are very high, with only 4 cases out of 50, of non-significant increases (*i.e.*, p-value greater than 0.05).

As a final comment for this section, we note that the un-biasing procedure does not provide good performance for the liver cancer (LIHC) dataset. In our opinion, this may be due to the fact that a large proportion of the patients may have pre-cancerous whole liver damage like cirrhosis, so that the "normal" cells may be heavily affected by the pre-existing disease.

3.2. Second validation: modularity of the correlation network

Modularity is a key feature of living systems. Every cellular event, such as signaling or DNA replication, is the result of the presence of "modules" composed of several molecular devices or regulatory structures, coordinately interacting directly or indirectly [38]. Indeed, at the molecular scale, the presence of modules is often described as an ensemble of gene products highly coordinated at the functional level, interacting physically and subject to co-regulation [39,40]. Moreover, modularity may support evolutionary forces and sustain change. The organization of functions in discrete modules (possibly partially overlapped) provide robustness to change but permit changes by modifications of the interconnections among modules. This is key to allow evolvability in uncertain and noisy environments and, at the same time, maintain adaptability [38,41]. Modularity is an omnipresent property of genomic data of all living systems which can be found in many kinds of experimental datasets, such as protein-protein or protein-DNA interactions, gene expression measurements, and many others [42]. Using network science terminology, modularity is often referred to as having a "community structure", i.e., their vertices are organized into groups, called communities, clusters, or modules. The identification of modules in a network may provide useful information on how it is organized by emphasizing regions with a sort of "degree of autonomy" or "self--organization" within the network.

The co-expression network is usually built using correlation and – as already stated in the introduction – is a very common analysis to infer biological properties from module detection [18]. Modularity of the correlation network reflects the modularity of the structural organization of living systems since modules of correlated gene profiles (as in the case under study here) are associated with common cellular functionality. Given the biological relevance of modularity, in this section, we validated the SEaCorAl algorithm by showing that its application to gene expression profiles of patients' groups, significantly increased the modularity of Spearman's correlation network. In other words, the application of the SEaCorAl un-biasing algorithm resulted in a correlation network that is more consistent with modularity, a fundamental biological feature of the living matter.

The modularity structure of a network and identification of com-

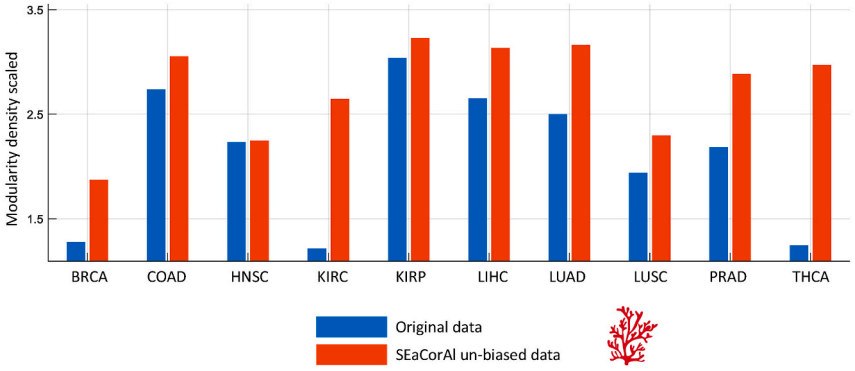

Fig. 9. Second validation of SEaCorAl: increased modularity of the correlation network. The figure shows the density scaled modularity of the correlation networks constructed using the original and the unbiased (SEaCorAl) datasets for ten TCGA cancers. In all cases, the density scaled modularity is larger after the application of the SEaCorAl algorithm. In some cases, this difference is quite consistent (see KIRC and THCA datasets). The differences are statistically significant, as explained in the text.

munities can be formally characterized in many ways. The most widely used one is the "modularity measure" defined as the fraction of edges that belong to the given communities minus the expected fraction whether links were randomly distributed. Community finding algorithms using the modularity measure are based, for example, on maximum likelihood or on a local greedy approach. To quantify the modularity of a network, we referred to the above-mentioned modularity measure due to Newman [43] who defined the following measure Q:

$$Q = \frac{1}{2m}\sum_{vw}\left[A_{vw} - \frac{k_v k_w}{2m}\right]\delta(C_v, C_w)$$

where m is the number of links, k_i is the degree of node i, A_{ij} is the (i,j) element (weight) of the adjacency matrix of the network and C_i indicates the community to which node i belongs. Modularity Q is a property of the entire network and takes values in [-1/2,1]. Large values of Q indicate a high modularity of the network. A commonly used algorithm for finding modules (or communities) is the so-called Louvain algorithm [44] which provides a partition of the network in dense modules by maximizing Q. Since, in our case, also negative weights (correlations) are present, we used the Louvain algorithm to maximize a slightly modified version of the modularity measure Q proposed by Rubinov and Sporns that includes negative weights [45]. The Louvain method for community detection has been designed to identify communities from large networks and it was created by Blondel et al. [44] from the University of Louvain. The method is a greedy optimization method. The basic idea of the algorithm is to initially search for small communities by optimizing modularity on a local base and, then each of them is grouped into a single one and, the first step is repeated until the modularity measure stops growing.

To compare the Spearman's correlation network (adjust p-value threshold less than 0.05) obtained from the original and the un-biased (SEaCorAl) dataset, we took into account that the un-biasing procedure maintains the number of nodes but reduces the number of spurious correlations and, therefore, the number of links in the correlation network. It is known that the modularity measure is heavily affected by link density [46] and, therefore, a considered an adjusted formula to be able to compare networks with different densities. To this end, we resorted to Ref. [47], where it is proved that modularity scales proportionally with the square of link density. Consequently, as a modularity measure to compare networks with different link densities, we considered a *density scaled modularity* defined as follows:

$$Q_a = \frac{Q}{\sqrt{d_l}}$$

where Q is the usual modularity measure and d_l is the links' density of the network.

To assess the statistical significance of the Results, for a given threshold (adjusted p-value less than 0.001), we obtained a random Spearman's correlation network for both the original and the unbiased dataset by randomly rewiring the weighted links, thus preserving both nodes' degree and link density. Degree preserving randomization is a technique widely utilized in network analysis to evaluate whether changes in a network could not be related to a biological property but to its intrinsic topology. Then, we computed the density scaled modularity for the random networks and found that the values obtained by repeating the randomization, for both the original and the unbiased dataset, led to a distribution with zero variance (less than 2.2×10^{-308}) and therefore, every difference from the random case can be considered statistically significant. Accordingly, we considered two correction terms Q_{rnd} and d_l, the first to avoid the bias of the network topology and the second to account for network link density. Accordingly, the corrected modularity measure is the following:

$$Q_a = \frac{Q - Q_{rnd}}{\sqrt{d_l}}$$

where Q_{rnd} is the modularity measure obtained using the degree and link density preserving network randomization (rewiring). The Results, for both the original and the un-biased (SEaCorAl) datasets, are reported in Fig. 9 where in all cases, the density scaled modularity of the correlation network obtained using unbiased data is higher than those using the original dataset.

Table 1
Number of patients available with paired data and number of outlier patients removed from the TCGA datasets described in the paper, using the Grubbs' test, for each dataset.

Dataset	Number of patients available with paired data	Number of patients removed using the Grubbs' test
BRCA	112	2
COAD	41	0
HNSC	43	1
KIRC	72	7
KIRP	31	0
LIHC	50	0
LUAD	57	0
LUSC	49	0
PRAD	52	3
THCA	58	4

3.3. Data preparation

We downloaded RNA-seq raw counts data from the TCGA portal (https://www.cancer.gov/tcga) on February 2021 of 10 cancers having at least 30 patients with paired data: BRCA (breast invasive carcinoma), COAD (colon adenocarcinoma), HNSC (head and neck squamous cell carcinoma), KIRC (kidney renal clear cell carcinoma), KIRP (kidney renal papillary cell carcinoma), LIHC (liver hepatocellular carcinoma), LUAD (lung adenocarcinoma), LUSC (lung squamous cell carcinoma),

PRAD (prostate adenocarcinoma) and THCA (thyroid carcinoma). The raw counts were normalized using the DESeq2 procedure [48]. For each dataset, we first removed genes having a number of zero values greater than 10% of all the available data. Second, we computed the mean of the paired $\log_2$ fold change (ML2FC) on the non-zero values for each gene and select the top 500 up-regulated and the top 500 down-regulated, as to have 1000 genes ordered by decreasing ML2FC. The q-values (adjusted p-values for multiple testing using the Storey method) for all selected genes and all datasets are always less than 10^{-6}. Using this list of genes, for each dataset, we removed those patients having a large number of negative correlations with other patients. Precisely, we computed the Pearson correlation of all pairs of patients and obtained, for each patient, the number of other patients having a negative correlation with an adjusted p-value less than 0.05. Then, we considered the distribution of these values and removed outlier patients using the Grubbs test procedure [[49]]. Table 1 reports the number of removed patients for each dataset.

4. Conclusion

In this paper, we have discussed a computational issue arising in RNA-seq paired data of patients' groups. We found, in real experimental data, a "regulation-correlation bias", which is a relationship between the regulation status of two genes (up or down) and their correlation sign. We hypothesized that the origin of this relationship may not be related to an underlying biological process but to an artifact of the RNA-seq paired data. Accordingly, we have proposed a simple idealized mechanism (analogous to the so-called "spurious correlation of ratios") able to generate the same regulation-correlation pattern observed in real data and proposed a sort of "reverse" procedure, the SEaCorAl algorithm, to remove (or reduce) such bias. To validate our findings, we showed that the known association between positive correlation and function of a gene pair, becomes significantly more evident after the application of the un-biasing procedure proposed in this paper. Moreover, we showed that also modularity of the corresponding correlation network significantly increases. We believe that the same bias may arise also in unpaired data and that this regulation-correlation bias may affect the biological significance of many correlation analyses of gene expression data. To assess these points, further investigations are certainly needed.

Funding

This research was funded by Sapienza University of Rome, "Progetto di Ateneo", grant number: RM11916B88C3E2DE.

Declaration of competing interest

The authors declare that there is no conflict of interest.

References

[1] G.M.Z. Wang, M. Snyder, RNA-Seq: a revolutionary tool for transcriptomics, Nat. Rev. Genet. 10 (2009) 57–63.
[2] Yuanyuan Bian, He Chong, Jie Hou, Jianlin Cheng, Jing Qiu, PairedFB: a full hierarchical Bayesian model for paired RNA-seq data with heterogeneous treatment effects, Bioinformatics 35 (5) (01, March, 2019) 787–797.
[3] J.N. Weinstein, E.A. Collisson, et al., The cancer genome Atlas pan-cancer analysis project, Nat. Genet. 45 (10) (2013) 1113–1120, https://doi.org/10.1038/ng.2764.
[4] C. Sotiriou, M.J. Piccart, Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to patient care? Nat. Rev. Canc. 7 (7) (2007, Jul) 545–553, https://doi.org/10.1038/nrc2173.PMID:17585334.
[5] J.R. Stevens, J.S. Herrick, R.K. Wolff, M.L. Slattery, Power in pairs: assessing the statistical value of paired samples in tests for differential expression, BMC Genom. 19 (1) (2018, Dec, 20) 953, https://doi.org/10.1186/s12864-018-5236-2. PMID: 30572829; PMCID: PMC6302489.
[6] J. Aitchison, The Statistical Analysis of Compositional Data, The Blackburn Press, Caldwell, NJ, 2003.
[7] M. Petti, D. Bizzarri, A. Verrienti, R. Falcone, L. Farina, Connectivity significance for disease gene prioritization in an expanding universe, IEEE ACM Trans. Comput. Biol. Bioinf 17 (6) (2020) 2155–2161, https://doi.org/10.1109/TCBB.2019.2938512. Epub 2020 Dec 8. PMID: 31484130.
[8] K.J. Karczewski, M.P. Snyder, Integrative omics for health and disease, Nat. Rev. Genet. 19 (5) (2018, May) 299–310, https://doi.org/10.1038/nrg.2018.4. Epub.2018.Feb.26. PMID: 29479082; PMCID: PMC5990367.
[9] M.A.J. Smits, K.M. Wong, E. Mantikou, C.M. Korver, A. Jongejan, T.M. Breit, M. Goddijn, S. Mastenbroek, S. Repping, Age-related gene expression profiles of immature human oocytes, Mol. Hum. Reprod. 24 (10) (2018, Oct, 1) 469–477, https://doi.org/10.1093/molehr/gay036.PMID:30257015.
[10] M.C. Palumbo, L. Farina, P. Paci, Kinetics Effects and Modeling of mRNA Turnover. Wiley Interdiscip Rev RNA vol. 6, 2015, pp. 327–336, https://doi.org/10.1002/wrna.1277. Epub 2015 Mar 1. PMID: 25727049.
[11] J. Toppi, M. Petti, F. De Vico Fallani, G. Vecchiato, A.G. Maglione, F. Cincotti, S. Salinari, D. Mattia, F. Babiloni, L. Astolfi, Describing relevant indices from the resting state electrophysiological networks, Annu Int Conf IEEE Eng Med Biol Soc (2012) 2547–2550, https://doi.org/10.1109/EMBC.2012.6346483. PMID: 23366444.
[12] M.S. Kim, D. Kim, J.R. Kim, Stage-dependent gene expression profiling in colorectal cancer, IEEE ACM Trans. Comput. Biol. Bioinf 16 (5) (2019, Sep-Oct) 1685–1692, https://doi.org/10.1109/TCBB.2018.2814043. Epub 2018 Mar 8. PMID: 29994071.
[13] N. Gu, T. Adachi, J. Takeda, N. Aoki, G. Tsujimoto, A. Ishihara, K. Tsuda, K. Yasuda, Sucrase-isomaltase gene expression is inhibited by mutant hepatocyte nuclear factor (HNF)-1alpha and mutant HNF-1beta in Caco-2 cells, J. Nutr. Sci. Vitaminol. 52 (2) (2006, Apr) 105–112, https://doi.org/10.3177/jnsv.52.105. PMID:16802690.
[14] J.B. Garner, A.J. Chamberlain, C. Vander Jagt, T.T.T. Nguyen, B.A. Mason, L. C. Marett, B.J. Leury, W.J. Wales, B.J. Hayes, Gene expression of the heat stress response in bovine peripheral white blood cells and milk somatic cells in vivo, Sci. Rep. 10 (1) (2020, Nov, 5), 19181, https://doi.org/10.1038/s41598-020-75438-2. PMID: 33154392; PMCID: PMC7645416.
[15] J.M. Stuart, E. Segal, D. Koller, S.K. Kim, A gene-coexpression network for global discovery of conserved genetic modules, Science 302 (5643) (2003, Oct, 10) 249–255, https://doi.org/10.1126/science.1087447. Epub 2003 Aug 21. PMID: 12934013.
[16] M.S. Iqbal, I. Ahmad, M. Asif, S.-H. Kim, R.M. Mehmood, Drug investigation tool: identifying the effect of drug on cell image by using improved correlation, Software Pract. Ex. 51 (2021) 260–270, https://doi.org/10.1002/spe.2903.
[17] P. Tieri, L. Farina, M. Petti, L. Astolfi, P. Paci, F. Castiglione, Network Inference and Reconstruction in Bioinformatics, Encyclopedia of Bioinformatics and Computational Biology, Academic Press, 2019, ISBN 9780128114322, pp. 805–813, https://doi.org/10.1016/B978-0-12-809633-8.20290-2.
[18] P. Langfelder, S. Horvath, WGCNA: an R package for weighted correlation network analysis, BMC Bioinf. 9 (2008, Dec, 29) 559, https://doi.org/10.1186/1471-2105-9-559. PMID: 19114008; PMCID: PMC2631488.
[19] P. Paci, T. Colombo, G. Fiscon, A. Gurtner, G. Pavesi, L. Farina, SWIM: a computational tool to unveiling crucial nodes in complex biological networks, Sci. Rep. 7 (2017, Mar, 20), 44797, https://doi.org/10.1038/srep44797. Erratum in: Sci Rep. 2017 Jun 16;7:46843. PMID: 28317894; PMCID: PMC5357943.
[20] R. Falcone, F. Conte, G. Fiscon, V. Pecce, M. Sponziello, C. Durante, L. Farina, S. Filetti, P. Paci, A. Verrienti, BRAFV600E-mutant cancers display a variety of networks by SWIM analysis: prediction of vemurafenib clinical response, Endocrine 64 (2) (2019, May) 406–413, https://doi.org/10.1007/s12020-019-01890-4. Epub 2019 Mar 8. PMID: 30850937.
[21] D.J. Allocco, I.S. Kohane, A.J. Butte, Quantifying the relationship between co-expression, co-regulation and gene function, BMC Bioinf. 5 (2004, Feb, 25) 18, https://doi.org/10.1186/1471-2105-5-18. PMID: 15053845; PMCID: PMC375525.
[22] D.T. Ross, U. Scherf, M.B. Eisen, C.M. Perou, C. Rees, P. Spellman, V. Iyer, S. S. Jeffrey, M. Van de Rijn, M. Waltham, A. Pergamenschikov, J.C. Lee, D. Lashkari, D. Shalon, T.G. Myers, J.N. Weinstein, D. Botstein, P.O. Brown, Systematic variation in gene expression patterns in human cancer cell lines, Nat. Genet. 24 (3) (2000, Mar) 227–235, https://doi.org/10.1038/73432.PMID:10700174.
[23] S.Q. Fang, M. Gao, S.L. Xiong, H.Y. Chen, S.S. Hu, H.B. Cai, Combining differential expression and differential coexpression analysis identifies optimal gene and gene set in cervical cancer, J. Canc. Res. Therapeut. 14 (1) (2018, Jan) 201–207, https://doi.org/10.4103/0973-1482.199787. PMID: 29516986.
[24] Y. Zuo, Y. Cui, C. Di Poto, R.S. Varghese, G. Yu, R. Li, H.W. Ressom, INDEED: integrated differential expression and differential network analysis of omic data for biomarker discovery, Methods 111 (2016, Dec, 1) 12–20, https://doi.org/10.1016/j.ymeth.2016.08.015. Epub 2016 Aug 31. PMID: 27592383; PMCID: PMC5135617.
[25] L.Y. Dong, W.Z. Zhou, J.W. Ni, W. Xiang, W.H. Hu, C. Yu, H.Y. Li, Identifying the optimal gene and gene set in hepatocellular carcinoma based on differential expression and differential co-expression algorithm, Oncol. Rep. 37 (2) (2017, Feb) 1066–1074, https://doi.org/10.3892/or.2016.5333.Epub2016.Dec.23. PMID: 28035405.
[26] E. Pampouille, C. Hennequet-Antier, C. Praud, A. Juanchich, A. Brionne, E. Godet, T. Bordeau, F. Fagnoul, E. Le Bihan-Duval, C. Berri, Differential expression and co-expression gene network analyses reveal molecular mechanisms and candidate biomarkers involved in breast muscle myopathies in chicken, Sci. Rep. 9 (1) (2019, Oct, 17), 14905, https://doi.org/10.1038/s41598-019-51521-1. PMID: 31624339; PMCID: PMC6797748.
[27] R. Anglani, T.M. Creanza, V.C. Liuzzi, A. Piepoli, A. Panza, A. Andriulli, N. Ancona, Loss of connectivity in cancer co-expression networks, PloS One 9 (1) (2014, Jan, 28), e87075, https://doi.org/10.1371/journal.pone.0087075. PMID: 24489837; PMCID: PMC3904972.

[28] J.M. Zamora-Fuentes, E. Hernández-Lemus, J. Espinal-Enríquez, Gene expression and Co-expression networks are strongly altered through stages in clear cell renal carcinoma, Front. Genet. 11 (2020, Nov, 3), 578679, https://doi.org/10.3389/fgene.2020.578679. PMID: 33240325; PMCID: PMC7669746.
[29] M. Drag, R. Skinkyté-Juskiené, D.N. Do, L.J.A. Kogelman, H.N. Kadarmideen, Differential expression and co-expression gene networks reveal candidate biomarkers of boar taint in non-castrated pigs, Sci. Rep. 7 (1) (2017, Sep, 22), 12205, https://doi.org/10.1038/s41598-017-11928-0. PMID: 28939879; PMCID: PMC5610188.
[30] T.W. Lui, N.B. Tsui, L.W. Chan, C.S. Wong, P.M. Siu, B.Y. Yung, DECODE: an integrated differential co-expression and differential expression analysis of gene expression data, BMC Bioinf. 16 (2015, May, 31) 182, https://doi.org/10.1186/s12859-015-0582-4. PMID: 26026612; PMCID: PMC4449974.
[31] D.A. Jackson, K.M. Somers, The spectre of 'spurious' correlations, Oecologia 86 (1) (1991, Mar) 147–151, https://doi.org/10.1007/BF00317404.PMID:28313173.
[32] K. Pearson, Mathematical contributions to the theory of evolution – on a form of spurious correlation which may arise when indices are used in the measurement of organs, Proc. Roy. Soc. Lond. 60 (359–367) (1897) 489–498.
[33] S. van Dam, U. Võsa, A. van der Graaf, L. Franke, J.P. de Magalhães, Gene co-expression analysis for functional classification and gene-disease predictions, Briefings Bioinf. 19 (4) (2018, Jul, 20) 575–592, https://doi.org/10.1093/bib/bbw139. PMID: 28077403; PMCID: PMC6054162.
[34] H.K. Lee, A.K. Hsu, J. Sajdak, J. Qin, P. Pavlidis, Coexpression analysis of human genes across many microarray data sets, Genome Res. 14 (6) (2004, Jun) 1085–1094, https://doi.org/10.1101/gr.1910904. PMID: 15173114; PMCID: PMC419787.
[35] The Gene Ontology Consortium, The gene ontology resource: 20 years and still GOing strong, Nucleic Acids Res. 47 (D1) (2019, Jan, 8) D330–D338, https://doi.org/10.1093/nar/gky1055. PMID: 30395331; PMCID: PMC6323945.
[36] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, K. Morishima, KEGG: new perspectives on genomes, pathways, diseases and drugs, Nucleic Acids Res. 45 (D1) (2017, Jan, 4) D353–D361, https://doi.org/10.1093/nar/gkw1092. Epub 2016 Nov 28. PMID: 27899662; PMCID: PMC5210567.
[37] A. Subramanian, P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette, A. Paulovich, S.L. Pomeroy, T.R. Golub, E.S. Lander, J.P. Mesirov, Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, Proc. Natl. Acad. Sci. U. S. A. 102 (43) (2005, Oct, 25) 15545–15550, https://doi.org/10.1073/pnas.0506580102. Epub 2005 Sep 30. PMID: 16199517; PMCID: PMC1239896.
[38] L.H. Hartwell, J.J. Hopfield, S. Leibler, A.W. Murray, From molecular to modular cell biology, Nature 402 (6761 Suppl) (1999, Dec, 2) C47–C52, https://doi.org/10.1038/35011540.PMID:10591225.
[39] K. Mitra, A.R. Carvunis, S.K. Ramesh, T. Ideker, Integrative approaches for finding modular structure in biological networks, Nat. Rev. Genet. 14 (10) (2013, Oct) 719–732, https://doi.org/10.1038/nrg3552.PMID:24045689. PMCID: PMC3940161.
[40] G.P. Wagner, M. Pavlicev, J.M. Cheverud, The road to modularity, Nat. Rev. Genet. 8 (12) (2007, Dec) 921–931, https://doi.org/10.1038/nrg2267.PMID:18007649.
[41] A. Hintze, C. Adami, Evolution of complex modular biological networks, PLoS Comput. Biol. 4 (2) (2008, Feb) e23, https://doi.org/10.1371/journal.pcbi.0040023. PMID: 18266463; PMCID: PMC2233666.
[42] H.B. Fraser, Coevolution, modularity and human disease, Curr. Opin. Genet. Dev. 16 (6) (2006, Dec) 637–644, https://doi.org/10.1016/j.gde.2006.09.001. Epub 2006 Sep 26. PMID: 17005391.
[43] M.E. Newman, Modularity and community structure in networks, Proc. Natl. Acad. Sci. U. S. A. 103 (23) (2006, Jun, 6) 8577–8582, https://doi.org/10.1073/pnas.0601602103. Epub 2006 May 24. PMID: 16723398; PMCID: PMC1482622.
[44] D Blondel Vincent, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre, Fast unfolding of communities in large networks, J. Stat. Mech. Theor. Exp. (2008, Oct) 1742–5468, 2008.
[45] M. Rubinov, O. Sporns, Weight-conserving characterization of complex functional brain networks, Neuroimage 56 (4) (2011, Jun, 15) 2068–2079, https://doi.org/10.1016/j.neuroimage.2011.03.069. Epub 2011 Apr 1. PMID: 21459148.
[46] Mingming Chen, Tommy Nguyen, Boleslaw K. Szymanski, A new metric for quality of network community structure, ASE Human J. 2 (4) (2013) 226–240.
[47] F. Botta, C. del Genio, Finding network communities using modularity density, J. Stat. Mech. Theor. Exp. (2016), 123402, https://doi.org/10.1088/1742-5468/2016/12/123402.
[48] M.I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2, Genome Biol. 15 (12) (2014) 550, https://doi.org/10.1186/s13059-014-0550-8. PMID: 25516281; PMCID: PMC4302049.
[49] Frank E. Grubbs, Sample criteria for testing outlying observations, Ann. Math. Stat. 21 (1) (1950) 27–58, https://doi.org/10.1214/aoms/1177729885.