

RESEARCH

Functional and Transcriptional Connectivity of Communities in Breast Cancer Co-expression Networks

Guillermo de Anda-Jáuregui¹, Sergio Antonio Alcalá-Corona², Jesús Espinal-Enríquez^{1,3} and Enrique Hernández-Lemus^{1,3*}

*Correspondence:

ehernandez@inmegen.gob.mx

¹Computational Genomics,
Instituto Nacional de Medicina
Genómica, Periférico Sur 4809,
Ciudad de México, México
Full list of author information is
available at the end of the article

Abstract

Transcriptional co-expression networks represent the concerted gene regulation programs by means of statistical inference of co-expression patterns. The rich phenomenology of transcriptional processes behind complex phenotypes such as cancer, is often captured (at least partially) in the connectivity structure of transcriptional co-expression networks. By analyzing the community structure of these networks, we may develop a deeper understanding of that phenomenology. We identified the modular structure of a transcriptional co-expression network obtained from breast cancer gene expression as well as a non-cancer adjacent breast tissue network as a control. We then analyzed the biological functions associated to the resulting communities by means of enrichment analysis. We also generated two projected networks for both, tumor and control networks: The first one is a projection to a network in which nodes are communities and edges represent topologically adjacent communities, indicating co-expression patterns between them. For the second projection, a bipartite network was generated containing a layer of modules and a layer of biological processes, with links between modules and the functions in which they are enriched; from this bipartite network, a projection to the community layer was obtained. From the analysis of the communities and projections, we were able to discern distinctive patterns of regulation between tumors and controls. Even though the connectivity structure of transcriptional co-expression networks is quite different, the topology of the projected networks is somehow similar, indicating functional compartmentalization, in both tumor and control conditions. However, the biological functions represented in the corresponding modules resulted notably different, with the tumor network comprising functional modules enriched for well-known hallmarks of cancer.

Keywords: Breast cancer networks; Modularity; Bipartite networks; Functional enrichment

Background

Co-expression networks are graph-theoretical constructs that represent global-level regulatory interactions and expression patterns of genes. These are well-defined mathematical structures amenable for systematic analysis of its global and local properties, as well as its dynamics and functionality. The case of said networks related to complex phenotypes such as cancer has been an area of interest in recent times [1–3]. Modular structure [4–7] is a quite relevant feature of co-expression networks, since it may provide some clues as to what are the actual biological

mechanisms in complex phenotypes [8]. In the case of breast cancer deregulation, functional biological organization has been shown to be related to network modularity [9, 10]. Such *community structure* of gene organization is characteristic of the different breast cancer molecular subtypes [9], so that particularities of the molecular phenotypes are well represented in the modular partition of the network [11].

Transcriptional co-expression networks can be probabilistically inferred from high-throughput gene expression data [12–15], and provide a representation of the expression landscape of a given phenotype. These type of regulatory networks consist of nodes representing genes and links representing co-expression (i.e. strong statistical dependency) between said genes. Given a Co-Expression Network $\mathcal{G}$, composed of gene nodes and links representing co-expression between genes, it is possible to detect non-overlapping *co-expression modules* (communities) due to its topology.

The gene set of each module M_i may be tested for association to known gene-sets of biological interest, such as biological functions, using enrichment analysis. These associations may be represented as a bipartite graph $\mathcal{B}$, with a set of module nodes M and a set of biological functions F , with links between modules and the functions in which they are enriched. With this in mind, it is possible to project $\mathcal{G}$ and $\mathcal{B}$ into two new graphs $\mathcal{GP}$ and $\mathcal{BP}$ (see methods) where nodes correspond to modules detected in the original graph $\mathcal{G}$.

These two projections recover two distinct types of relationships between groups of genes: on the one hand, whether different groups of genes have a level of *co-expression* that may be driven by biological factors, such as co-regulation; and on the other hand, whether different groups of genes are involved in the *control* of biological functions that are necessary for a given biological context (for instance, a phenotype). An interesting case is that of modules that are co-regulated and connected through shared biological functions.

In this work, we analyze two coexpression networks derived from *basal breast cancer* (tumors) and *healthy breast tissue* (controls), and explore the two modular projections described. We identify the differences in modular structure between the two phenotypes, and how these different modular structures differ in terms of the two types of intermodular relationships that we have described.

Methods

Network inference

Co-expression networks were reconstructed from gene expression data. Basal breast cancer gene expression data, along with adjacent normal expression data, were obtained from the Cancer Genome Atlas [16]. Data acquisition, and pre-processing is described in [17]. Briefly, we used 142 Basal-like subtype breast cancer samples, along with 101 solid-tissue adjacent normal samples. 15,642 annotated genes were included in each sample, after removal of low-counts transcripts (< 5 per sample). This set of un-paired data were pre-processed, normalized and bias-reduced, to have

a comparable set of expression data between cancer and control samples.

Mutual Information (MI) was computed using an implementation of the ARACNE algorithm for all gene pairs [18]. A suitable MI threshold was selected based on the following criteria:

- At least 80% of nodes in the genome (out of 15,642) must be present in the network by being connected to at least one other gene
- The network must have a giant connected component (i.e., the largest connected component with more than half of the nodes)
- The highest (most restrictive) MI threshold must be selected

We evaluated different MI threshold values related to quantiles of the MI distribution. Generated networks were imported as `igraph` for [R] objects. `igraph` version 0.7.1 and R version 3.5.1 were used.

Mutual information is the maximum entropy/maximum likelihood estimate of statistical dependence between two random variables [19]. It is indeed a symmetrized version of the Kullback-Leibler divergence between the joint probability distribution for two variables and the product of their marginals [20] (i.e. the joint probability distribution under independence conditions). Being a maximum entropy estimate it needs the least number of assumptions on the probability distributions. Indeed the only needed assumption is that these distributions have compact support. Other correlation measures assume identically distributed variables, linearity or rank ordering among them, etc. Such assumptions are often not compliant with the nature of gene expression data such as nonlinearity, 'delays' (i.e. correlation shifts), and so on. For these reasons, mutual information has been thoroughly used for the inference of (large) gene co-expression networks. Another advantage of the use of mutual information measures to deconvolute gene regulatory networks from massive gene expression data is the fact that, in most cases (whenever Hammersley-Clifford conditions apply), the resulting graphs meet the requirements to belong to the family of Markov random fields, something that under some scenarios may be quite useful [21].

The major drawback for the use of the mutual information approach is the fact that one needs a way to reconstruct the probability distributions from empirical data. Even under the relatively 'soft' conditions imposed by Glivenko-Cantelli convergence, this means that one still have to have a somehow large number of samples (more than approx. 100 for the case of gene expression data) for the empirical distribution to be useful in order to minimize the number of false positives. These conditions are fulfilled here.

Module detection and enrichment

Modules were detected using the Infomap [22–24] implementation for `igraph`, using 1,000 iterations to achieve convergence. We ~~then associated the sets of genes that are identified through the module detection algorithm, to biological functions, a process known as enrichment analysis~~ have chosen the Infomap algorithm, since it has proven to be highly efficient compared to other methods. Based on benchmarks,

Infomap was the best- ranked method in runtime, accuracy and performance [25], as it was assessed in terms of the LFR benchmark [26].

The field of enrichment analysis includes a wide variety of techniques [?]. In this work we used an Over Representation Analysis, in which a hypergeometric (or Fischer exact) test is used to identify a statistically significant association between each module's gene set, and the sets of genes involved in biological functions as described by the Gene Ontology (GO) database [27].

Each module gene set was tested for enrichment of ~~GO~~ Gene Ontology [27] terms via hypergeometric testing using the HTSanalyzer [28] package for R. GO terms were considered enriched if they had an adjusted Benjamini-Hochberg [29] p-value smaller than 0.05. Enrichment relationships found were represented as a bipartite network, with a layer of modules and a layer of GO terms.

Figure 1 presents a pictorial abstraction of this process. Panel A represents module detection of $\mathcal{G}$ using Infomap. In panel B modules detected in panel A become nodes in the $\mathcal{GP}$ projection; links represent intermodule connections. Enrichment of modules (i.e. the $\mathcal{B}$ network) detected in panel A is presented in panel C. The three modules are connected to turquoise diamonds, which represent biological processes associated to said modules. Panel D shows a projection $\mathcal{BP}$ of $\mathcal{B}$ in which nodes are modules linked if they share a biological process.

GP and BP projections

Topological and functional neighborhoods define two projections $\mathcal{GP}$ and $\mathcal{BP}$ as previously mentioned. The first projection, $\mathcal{GP}$, is a graph where nodes are modules M , and links exist between modules M_i and M_j if there are links in $\mathcal{G}$ between genes in M_i and genes in M_j : we say these modules are *topologically adjacent* in the original network.

The second projection, $\mathcal{BP}$, is a graph where nodes are modules M and links exist between M_i and M_j if there is overlap in the neighborhoods of M_i and M_j in $\mathcal{B}$: we say that these modules are *functionally adjacent*.

Results

Co-expression networks for breast cancer and adjacent normal

Networks were generated from the tumors and control datasets. After scanning different threshold values for mutual information (see Additional File 1) the highest threshold for MI that covered our criteria was found at the 0.999 quantile. These networks are described in Additional File 2. Figure 2 illustrates how different the tumor and control networks are; ~~this-nodes are colored by the module to which they belong. It can be observed in the tumor network, modules with nodes of the same color, whereas in the control network, modules are not observable and colors are less separated. This~~ is further supported by the different degree distributions (~~insets of Figure 2~~ Figure 3).

Said differences can be identified even by a quick glance at the node-degree distributions (~~bottom of Figure 2~~ Figure 3; χ^2 testing for differences in discrete

distribution was performed, with the following results: χ^2 -statistic = 1074170, p -value = $4.99e - 10$), as well as by the observation of the force-directed network visualization. The control network is characterized by a mono-scaled regime (the degree distribution appears to follow a power-law with a single scaling exponent over the whole range of degree values) whereas the tumor network presents an evident crossover phenomenon leading to multi-scaling, i.e. the degree distribution does not follow a power-law with a single scaling exponent, but rather it seems to have several different scaling regimes, with regions containing inflection points in so-called *crossover* regions.

Modular structure of breast cancer and healthy breast networks

Distinct modular structures were found in each network, in agreement with previous results from our group [10]. The partition for the tumor network has a smaller description length L [23] value ($L = 8.268641$) than the control network ($L = 11.80941$). In the control network, we identified 981 modules, whereas for the tumor network we found 910 modules. ~~Figure 2 and 3, Figure 4~~ As it can also be observed in Figure 2 and 3, Figure 4 shows histograms of the different module sizes, showing ~~larger~~ the largest modules in control network (χ^2 testing for differences in discrete distribution was performed, with the following results: χ^2 -statistic = 40324.45, p -value = $4.99e - 10$).

For each transcriptional co-expression network, we projected the modules identified in it to a $\mathcal{GP}$ network where adjacent modules in the original network are found. These $\mathcal{GP}_t$ (for tumors) and $\mathcal{GP}_c$ (for controls) are depicted in Figure 45, and described in Additional file 3 (modular projection parameters). There are three main differences between these networks that may be observed: i) a characteristic degree distribution for each projection (~~bottom of Figure 4~~ Figure 6 χ^2 testing for differences in discrete distribution was performed, with the following results: χ^2 -statistic = 49532.28, p -value = $4.99e - 10$), ii) the higher edge density in $\mathcal{GP}_c$, which is also related to iii) the higher link/node ratio in $\mathcal{GP}_c$.

Functional Enrichment

We identified a set of biological functions described as GO terms associated to modules detected in the tumor and control networks. We represented these functional associations as bipartite graphs $\mathcal{B}_t$ (for tumors) and $\mathcal{B}_c$ (for controls) that are represented in Figure 5A-7A and B, with parameters described in Additional file 4.

We identified 665 GO terms associated to $\mathcal{M}_t$ and 827 GO terms associated to $\mathcal{M}_c$. It is important to notice that not all modules were enriched in biological processes; in fact, only 110 enriched modules are found in $\mathcal{B}_t$ and 82 enriched modules were found in $\mathcal{B}_c$. Furthermore, the set of enriched GO terms $\mathcal{B}_c$ and $\mathcal{B}_t$ are different (with a Jaccard index of 0.34).

The projections of modules based on functional adjacency $\mathcal{BP}_c$ and $\mathcal{BP}_t$ are shown in ~~figure 6.~~ Figure 8A and 8B. In the figure, modules are connected if they

share at least one enriched process. Node size represents the module degree. Edge width is proportional to the number of shared enriched processes between modules. In both cases there are some modules that share several enriched processes. In $\mathcal{BP}_t$ (8B), there are clusters of modules sharing GO terms, whereas in 8A the compartmentalization is less evident.

Additional file 5 shows some of the relevant parameters for these projections. It may be observed that these projections are very sparse in terms of edges: only 51 of $\mathcal{M}_c$ are connected to other modules, whereas for $\mathcal{M}_t$ the number is 70. Importantly, there are modules in tumor and control networks (40 and 29, respectively) that are associated to GO terms not shared with any other module.

Discussion

Most central modules in the GP projection are the largest ones

The most central modules in both the $\mathcal{GP}_c$ and $\mathcal{GP}_t$ projections are also the largest ones. In $\mathcal{GP}_t$, this central module has 231 genes and 5437 intra-modular links. It is connected to 99 other modules. The most central module in $\mathcal{GP}_c$, has 1000 genes and 17583 intra-modular links. It is connected to 742 other modules.

Interestingly such highly central modules are not particularly notable in terms of their functional associations. In controls, the largest module is enriched in 6 processes of nucleic acid regulation; it is linked through processes (i.e. in $\mathcal{BP}_c$) to 6 other modules. For tumors, the largest module shows no statistically significant enrichment, and therefore is not linked to any other module in the $\mathcal{BP}_t$ projection.

Functional compartmentalization in health and disease

The bipartite graphs $\mathcal{B}$ are topologically similar between tumor and control; however, the enriched functions in each network are different. In both cases, the structures show star-like motives (Figure 5A and 5B 7A and 7B), which indicate mostly unique processes associated to a given gene module. We interpret this as evidence of compartmentalization of regulation, where each module is controlling the activity of independent sets of biological processes.

We observe important differences in terms of the biological processes associated to the most connected (i.e., most enriched) modules. The two most connected modules (with 146 and 121 neighbors, respectively) in $\mathcal{B}_c$ are associated to metabolism and cell cycle processes, as illustrated Figure 7A as well as in Additional file 6A and 6B; meanwhile, ~~immunity and regulation of gene expression are the associated processes~~ ~~immunity-related processes are associated~~ for the two most connected modules (with 95 and 81 neighbors, ~~respectively~~ ~~respectively~~) in $\mathcal{B}_t$, which we illustrate in Additional file 6C and 6D.

As it may be observed, associated processes in $\mathcal{B}_c$ are for maintenance, meanwhile the processes associated to the $\mathcal{B}_t$ are well-known *hallmarks of cancer* [30]. The identification of *hallmark* processes in breast cancer co-expression networks derived from high-throughput data is consistent with recent reports by our own group.

Most connected modules through functional adjacency are similar in health and disease

The way modules are connected through processes is similar between health and disease, even though the modules and functions are different. The most enriched modules are not, however, the ones that are more connected to other modules in terms of functional adjacency. These are, as seen in the $\mathcal{BP}$ projections for both controls or tumors, of comparable sizes: 86 and 74 genes, with 356 and 361 intra-modular links respectively. In controls, this module is enriched in 20 processes. Through these processes, it is connected to 18 modules. It is also connected through co-expression links, as seen in the $\mathcal{BP}_c$ projection, to 123 other modules. Meanwhile, the comparable module in tumors is enriched in 81 different processes, but through these is linked to only 20 other modules. Through co-expression links, it is connected to 123 other modules. Interestingly, again there is little overlap in the processes associated to these modules, sharing only one function, *Membrane protein complex*, a general homeostatic event.

Connections between modules through functional and topological adjacency are seldom found

By comparing the set of links in the $\mathcal{GP}$ and $\mathcal{BP}$ projections, we may observe that there are very few links between modules appearing in both projection. In the case of tumors, $\mathcal{GP}_t$ and $\mathcal{BP}_t$ have 37 shared links (Additional file 7), whereas in controls, $\mathcal{GP}_t$ and $\mathcal{BP}_t$ have 51 shared links. As such, we may observe that both in health and disease, the connectivity patterns among gene modules in terms of co-expression and functionality are quite different.

Conclusions

Networks of gene regulation are known to exhibit a modular behavior. The co-expression of gene modules is a form in which cellular processes are regulated. In this work, we demonstrate that modules in transcriptional co-expression networks have different ways to interact, either through co-expression or through jointly regulating functional processes. There are instances in which modules are connected both transcriptionally and functionally, but these are rare. transcriptional co-expression networks of cancer have a more modular structure than those found in health. Modules found in the health network have higher degrees, whereas modules in the breast cancer network are less likely to have transcriptional relationships to other modules.

We observe that the set of biological functions associated to gene modules are vastly different in breast cancer and health, with gene modules of cancer associated to functions that drive disease, whereas gene modules in health are linked to functions associated to the maintenance of homeostasis. However, we may observe that the connectivity patterns formed by associations of gene modules and biological functions are similar in both health and disease, which indicates that compartmentalization of functional regulation through gene expression remains, even though the processes that are being regulated change.

The behaviors in terms of transcriptional and functional connectivity that gene modules in transcriptional co-expression networks exhibit, may allow for the identification of important modules in terms of either transcriptional, or functional, importance associated to biological conditions of importance, such as cancer.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

GDJ contributed analytical methods, performed calculations, participated in the discussion, collaborated in writing the manuscript, SAAC contributed analytical methods, participated in the discussion, JEE participated in the biological discussion, contributed to the visualization of results, collaborated in writing the manuscript, EHL contributed analytical methods, participated in the discussion, collaborated in writing the manuscript, oversaw the project. All authors reviewed and approved the manuscript.

Acknowledgements

This work was supported by CONACYT (grants no.285544/2016, Ciencia Básica and 2115/2016, Fronteras de la Ciencia, JEE), as well as by federal funding from the National Institute of Genomic Medicine (Mexico). Additional support has been granted by the National Laboratory of Complexity Sciences (grant no. 232647/2014 CONACYT, EHL). EHL is a recipient of the 2016 Marcos Moshinsky Fellowship in the Physical Sciences.

Available code

All the code used for the present work is available in our repository:
<https://www.github.com/guillermodeandajauregui/meap>

Author details

¹Computational Genomics, Instituto Nacional de Medicina Genómica, Periférico Sur 4809, Ciudad de México, México. ²Department of Ecology & Evolution, University of Chicago, 1103 East 57th St., 60637 Chicago, IL, USA. ³Centro de Ciencias de la Complejidad, Universidad Nacional Autónoma de México, Ciudad Universitaria, Ciudad de México, México.

References

- De Craene, B., Berx, G.: Regulatory networks defining EMT during cancer initiation and progression. *Nature Reviews Cancer* **13**(2), 97 (2013)
- Dang, C.V., O'Donnell, K.A., Zeller, K.I., Nguyen, T., Osthus, R.C., Li, F.: The c-myc target gene network. In: *Seminars in Cancer Biology*, vol. 16, pp. 253–264 (2006). Elsevier
- de Anda-Jáuregui, G., Velázquez-Caldelas, T.E., Espinal-Enríquez, J., Hernández-Lemus, E.: Transcriptional network architecture of breast cancer molecular subtypes. *Frontiers in Physiology* **7**, 568 (2016)
- Girvan, M., Newman, M.E.: Community structure in social and biological networks. *Proceedings of the National Academy of Sciences* **99**(12), 7821–7826 (2002)
- Newman, M.E.: Modularity and community structure in networks. *Proceedings of the National Academy of Sciences* **103**(23), 8577–8582 (2006)
- Palla, G., Derényi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **435**(7043), 814 (2005)
- Palla, G., Farkas, I.J., Pollner, P., Derényi, I., Vicsek, T.: Directed network modules. *New Journal of Physics* **9**(6), 186 (2007)
- Alcalá-Corona, S.A., Velázquez-Caldelas, T.E., Espinal-Enríquez, J., Hernández-Lemus, E.: Community structure reveals biologically functional modules in mef2c transcriptional regulatory network. *Frontiers in Physiology* **7**, 184 (2016)
- Alcalá-Corona, S.A., de Anda-Jáuregui, G., Espinal-Enríquez, J., Hernández-Lemus, E.: Network modularity in breast cancer molecular subtypes. *Frontiers in Physiology* **8**, 915 (2017)
- Alcalá-Corona, S.A., de Anda-Jáuregui, G., Espinal-Enríquez, J., Tovar, H., Hernández-Lemus, E.: Network modularity and hierarchical structure in breast cancer molecular subtypes. In: *International Conference on Complex Systems*, pp. 352–358 (2018). Springer
- Alcalá-Corona, S.A., Espinal-Enríquez, J., de Anda Jáuregui, G., Hernández-Lemus, E.: The hierarchical modular structure of her2+ breast cancer network. *Frontiers in Physiology* **9**, 1423 (2018)
- Basso, K., Margolin, A.A., Stolovitzky, G., Klein, U., Dalla-Favera, R., Califano, A.: Reverse engineering of regulatory networks in human b cells. *Nature Genetics* **37**(4), 382 (2005)
- Margolin, A.A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., Califano, A.: Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. In: *BMC Bioinformatics*, vol. 7, p. 7 (2006). BioMed Central
- Hernández-Lemus, E., Siqueiros-García, J.M.: Information theoretical methods for complex network structure reconstruction. *Complex Adaptive Systems Modeling* **1**(1), 8 (2013)
- Hernández-Lemus, E., Rangel-Escareño, C.: The role of information theory in gene regulatory network inference. *Information Theory: New Research*, 109–144 (2011)
- ~~Computational methods for gene regulatory networks reconstruction and analysis: A review. Artificial intelligence in medicine (2018)~~
- ~~Network~~, C.G.A., ~~et al.~~: Comprehensive molecular portraits of human breast tumours. *Nature* **490**(7418), 61 (2012)

17. ~~Espeñal-Enríquez, J., Fresno, C., Anda-Jáuregui, G., Hernández-Lemus, E.: Rna-seq based genome-wide analysis reveals loss of inter-chromosomal regulation in breast cancer. Scientific Reports 7(1), 1760 (2017)~~
~~Crosstalknet: A visualization tool for differential co-expression networks and communities. Cancer research, 1383 (2018)~~
18. de Anda-Jáuregui, Espinal-Enríquez, J., Hernández-Lemus, E.: Spatial organization of the gene regulatory program: An information theoretical approach to breast cancer. Entropy **SI: Information Theory in Complex Systems**, 1 (2019)
19. Chow, C., Liu, C.: Approximating discrete probability distributions with dependence trees. IEEE Transactions on Information Theory **14**(3), 462–467 (1968)
20. ~~(Kullback, S., Leibler, R.A.: On information and sufficiency. The Annals of Mathematical Statistics 22(1), 79–86 (1951))~~
21. ~~(Dobruschin, P.: The description of a random field by means of conditional probabilities and conditions of its regularity. Theory of Probability & Its Applications 13(2), 197–224 (1968))~~
22. Rosvall, M., Bergstrom, C.T.: An information-theoretic framework for resolving community structure in complex networks. Proceedings of the National Academy of Sciences **104**(18), 7327–7331 (2007)
23. Rosvall, M., Bergstrom, C.T.: Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences **105**(4), 1118–1123 (2008)
24. Rosvall, M., Axelsson, D., Bergstrom, C.T.: The map equation. The European Physical Journal Special Topics **178**(1), 13–23 (2009)
25. ~~(Lancichinetti, A., Fortunato, S., Kertész, J.: Detecting the overlapping and hierarchical community structure in complex networks. New Journal of Physics 11(3), 033015 (2009))~~
26. Lancichinetti, A., Fortunato, S., Radicchi, F.: Benchmark graphs for testing community detection algorithms. Physical review E **78**(4), 046110 (2008)
27. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., *et al.*: Gene ontology: tool for the unification of biology. Nature genetics **25**(1), 25 (2000)
28. Wang, X., Terfve, C., Rose, J.C., Markowitz, F.: Htsanalyzer: an r/bioconductor package for integrated network analysis of high-throughput screens. Bioinformatics **27**(6), 879–880 (2011)
29. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological), 289–300 (1995)
30. Hanahan, D., Weinberg, R.A.: Hallmarks of cancer: the next generation. cell **144**(5), 646–674 (2011)

Figures

Figure 1 Graphical description of the workflow presented here. A) Module detection of tumor and control networks using infomap. In this figure, three modules are detected. B) Modules detected in A) become nodes in the $\mathcal{GP}$ projection; the links represent intermodule connections. C) Enrichment of modules detected in A). In this case, the three modules are connected to light-blue diamonds, which represent biological processes associated to said modules. D) Projection of C) In this final case, nodes are modules linked if they share a biological process. Notice that B and D networks are not connected in the same way, despite they have the same nodes.

Figure 2 Regulatory networks corresponding to the control (panel A) and tumor (Panel B) phenotypes. ~~Bottom of Nodes are colored according to the figure present the degree distributions for both networks module to which each node belongs.~~ Notice that in B, a visible modular structure appears, ~~reinforced by the distribution of colors,~~ meanwhile in A the network link distribution looks more ~~homogeneoushomogeneous.~~

Figure 3 Degree distributions for both networks. Red dots form the histogram of tumor network, meanwhile black dots take account for control network. Red dots appear to have two different regimes, with a crossover phenomenon. Black dots, on the other hand, appear to follow a power-law with a single scaling exponent.

Figure 4 Histograms of module sizes in tumor and control networks. As it can be observed, the largest modules correspond to the control network (black dots in the upper left part the figure). Also notice the different concavities in red and black curves.

Figure 5 The modular network structure in tumor and control. In this case, nodes are modules and the connections represent intermodule-inter-module genes connected in the original network. A) control module network. B) Tumor module network. ~~The bottom of the figure shows the degree distribution of said networks.~~

Figure 6 Degree distributions for module networks of figure 5. Red dots represent the tumor network, meanwhile black dots are for controls.

~~Bipartite graph of GO term enrichment in network modules. A) Control network. B) Tumor network. Notice the similarity between these two structures (non-enriching modules and GO terms not shown).~~

Figure 7 Bipartite graph of GO term enrichment in network modules. A) Control network. B) Tumor network. In both networks, grey diamonds represent the module that have enriched GO terms. Colored circles represent GO categories enriched for the linked modules. In some cases, GO categories are connected to more than one module. Colors of GO categories represent a higher category in which each GO term belongs. Colors in A and B are not related. Notice that the categories in A are mainly related to maintenance, meanwhile in B (tumor bipartite network) the majority of categories are related to immunity a well known hallmark of cancer.

Figure 8 Projection of breast cancer modules linked by shared enriched GO terms. A) Control projection. B) Tumor projection. Modules which show enrichment, but are not connected to other through shared enriched GO terms (29 in control, 40 in tumors) are not shown.

Additional Files

Additional file 1 — ZIP file with Analysis results for different thresholds

Includes two TXT files with tables of the analysis parameters for both tumors and controls at different MI thresholds. The files include the threshold quantile, number of nodes, edges, components, and largest component sizes for each quantile threshold tested.

Additional file 2 — Excel file with Topological parameters for co-expression networks

Additional file 3 — Excel file with Topological parameters for topologically adjacent module networks

Additional file 4 — Excel file with Topological parameters for bipartite networks

Additional file 5 — Excel file with Topological parameters for functionally adjacent networks

Additional file 6 — Subgraphs of bipartite networks highlighting important functions

Additional file 7 — Excel file with Shared Links between Adjacent and Functional Projections

Excel file with two sheets with lists, for each phenotype, of links between modules found in both projections