Comparing multiple networks using the Co-expression Differential Network Analysis (CoDiNA)

Deisy Morselli Gysi\textsuperscript{1,2,*}, Tiago de Miranda Fragoso\textsuperscript{3}, Volker Buskamp\textsuperscript{4}, Eivind Almaas\textsuperscript{5,6}, and Katja Nowick\textsuperscript{7,*}

\textsuperscript{1}Department of Computer Science, Interdisciplinary Centre of Bioinformatics, University of Leipzig, Leipzig, D-04109, Leipzig.
\textsuperscript{2}Swarm Intelligence and Complex Systems Group, Faculty of Mathematics and Computer Science, University of Leipzig, Leipzig, D-04109, Leipzig.
\textsuperscript{3}Fundação Cesgranrio, Rio de Janeiro, 20261-903, Brazil.
\textsuperscript{4}Technische Universität Dresden, DFG Research Center for Regenerative Therapies, 01307 Dresden, Germany
\textsuperscript{5}Department of Biotechnology, NTNU - Norwegian University of Science and Technology, Trondheim, Norway.
\textsuperscript{6}K.G. Jebsen Centre for Genetic Epidemiology, NTNU - Norwegian University of Science and Technology, Trondheim, Norway.
\textsuperscript{7}Human Biology Group, Institute for Biology, Department of Biology, Chemistry, Pharmacy, Freie Universitaet Berlin, Koenigin-Luise-Str. 1-3, D-14195 Berlin, Germany.
\textsuperscript{*}corresponding authors: deisy@bioinf.uni-leipzig.de, katja.nowick@fu-berlin.de

ABSTRACT

Biomedical sciences are increasingly recognising the relevance of gene co-expression-networks for analysing complex-systems, phenotypes or diseases. When the goal is investigating complex-phenotypes under varying conditions, it comes naturally to employ comparative network methods. While approaches for comparing two networks exist, this is not the case for multiple networks. Here we present a method for the systematic comparison of an unlimited number of networks: Co-expression Differential Network Analysis (CoDiNA) for detecting links and nodes that are common, specific or different to the networks. Applying CoDiNA to a neurogenesis study identified genes for neuron differentiation. Experimentally overexpressing one candidate resulted in significant disturbance in the underlying neurogenesis' gene regulatory network. We compared data from adults and children with active tuberculosis to test for signatures of HIV. We also identified common and distinct network features for particular cancer types with CoDiNA. These studies show that CoDiNA successfully detects genes associated with the diseases.

Main

Complex systems, exemplified by biological pathways, social interactions, and financial markets, can be expressed and analysed as systems of multi-component interactions\textsuperscript{1}. In systems biology, it is necessary to develop a thorough understanding of the interactions between factors, such as genes or proteins. Gene co-expression networks have been especially effective in identifying those interactions\textsuperscript{2–5}. In gene co-expression networks, nodes represent genes and a weighted link between a pair of genes represents their connection, often calculated as a correlation. The sign of the relation may suggest an up- or down-regulation of one factor by the other\textsuperscript{6}. It has been shown that different conditions have different underlying regulatory patterns and will therefore lead to different networks even for a single system\textsuperscript{1,7,8}.

While the analysis of expression differences allows for identification of genes that are significantly differentially expressed between two or more conditions\textsuperscript{4,9–12}, it does not enable the investigation of more complex patterns, such as changes in the rewiring of the regulatory relationships of genes. Differential network analyses are able to capture changes in gene relationships and are thus exceptionally suitable for understanding complex phenotypes and diseases\textsuperscript{4}.

In order to distinguish similarities and differences between system-level activity patterns, it has become common practice to compare multiple networks, derived from different conditions. This is usually followed by an analysis of links that are shared between all networks or are chiefly specific to some of the networks. Classification of links according to the concepts of being present, different or absent in some networks is essential to understanding how different environments influence particular interactions. Several methods for such comparisons exist for pairs of networks\textsuperscript{9,13–21}. However, in many situations it is of
great interest to compare more than just two networks simultaneously. Such a set of networks could e.g. be gene-regulatory networks arising from different species, tissues or diseases, or co-existence networks from different environments. For instance, an evolutionary study conducted a pairwise comparison between humans, chimpanzees and rhesus macaques to pinpoint similarities and differences in the prefrontal cortex\textsuperscript{22}. In a recent medical study, the authors compared enriched gene functions using the Gene Ontology\textsuperscript{23} instead of comparing the networks of the multiple cancers. Another study generated a network that involved only differentially expressed genes\textsuperscript{24} extracted from their multiple networks. These studies could have profited extensively from applying a method capable of systematically comparing multiple networks simultaneously.

Kuntal and collaborators\textsuperscript{1} proposed a method, CompNet, that may address the comparison of multiple networks. However, the focus of CompNet is on the visualisation of the union, intersections and exclusive links of the analysed networks. Thus, a method that is capable of comparing both links and nodes of any number of networks is still lacking. Here, we propose a novel method for that purpose, CoDiNA (Co-expression Differential Network Analysis for n Dimensions), implemented as an R package. This package also includes an interactive tool for network visualisation. Our method was first applied to find common, specific or different links and nodes in a study of neurogenesis of induced Pluripotent Stem Cells (iPSC) with or without the presence of the micro RNA miR-124. CoDiNA identified modules of genes that were present in miR-124 knockout or wildtype cells and a hub-gene for the time point of highest network differences. Overexpressing this hub gene during neurogenesis resulted in a clear repression of neuronal differentiation. Here we further demonstrate the power and versatility of our method with two example applications. In one example we compare expression data from tuberculosis (TB) patients with or without HIV. The co-infection of HIV and TB is of strong medical importance since, to this date, it is challenging to detect TB in the presence of an HIV infection. In the other showcase, we compare three types of cancers to understand which molecular signatures they have in common and how they differ.

Results and Discussion

To perform a comparison of co-expression networks, CoDiNA requires as input a set of networks to be assessed. The networks can be constructed using a correlation method, but should only contain links that are statistically significant given a predefined \( p \)-value threshold. We classify each link into one of three \( \Phi \) categories based on its weight: a link is said to be an \( \alpha \) link if it is present in all networks with the same sign, i.e., it is an interaction that is common to all networks. A link is called \( \beta \) if it is present in all networks but with different signs of the link’s weight, i.e., it represents a different kind of interaction in at least one network. The biological interpretation of this category is that a particular gene changed its function, so that a gene that up-regulates another gene in one condition down-regulates the same gene in another condition (or vice versa). A link is considered a \( \gamma \) link, if it is present in some networks but not all, regardless of the sign of the link’s weight, i.e., this link is specific to at least one network. To further characterise how a particular link is different or specific, we are assigning a sub-category, which we refer to as \( \tilde{\Phi} \). This subcategory clarifies to which condition a link is specific or in which condition it has changed.

In order to avoid false associations, i.e. the incorrect inference that a particular gene is associated with a specific condition, we require that all investigated \( \textit{nodes} \) are present in all networks: If a node is absent in at least one network, we remove all of its links from the networks in which it is present. This does not apply to nodes that are present but have no significant links. In this case, it is assigned a (weight) value of zero, thus allowing all measured nodes to be included in the analysis, even when they only have not significant links.

The weight value of the link between gene \( i \) and \( j \), denoted by \( \rho_{ij} \), is defined in the interval \([-1, 1]\). To denote links as positive, negative, or neutral we divide this interval into three equal parts. To compare networks whose intervals might vary, one option is to normalise the data inside the interval to \([-1, 1]\); we refer to this approach as \textit{stretch}. This step is particularly important to compare networks that were not measured under the same experimental conditions. For a set of \( W \) networks of \( N \) nodes, each one of the \( E \) links has a specific link weight \( \rho_{ij,k} \) connecting nodes \( i \) and \( j \) in network \( k \). Each link weight \( \rho_{ij,k} \) is categorised into 1, 0 or \(-1\) depending on its presence or absence in all, some, or none of the networks. If a particular link is absent in all of the \( W \) networks, this link is removed from subsequent analyses. If a link has a signal different than the link of the reference network, its \( \tilde{\Phi} \) category will be \( \tilde{\Phi}_{N_1} \), where \( N_1 \) is a set listing the networks for which the signal changes (assuming the first network as the reference). Accordingly, if a link is specific, the category is \( \gamma_{N_2} \), where \( N_2 \) is a set listing the networks for which the link exists.

After defining the \( \Phi \) categories, we score the links to identify those that are stronger. To this end, for every node \( i = 1, 2, \ldots, N \), we interpret the array of link weights \((\rho_{1i}, \ldots, \rho_{Wi})\) as a point in a \( W \)-dimensional Euclidean space (Fig. 1, panel II). Thus, all points are contained in the cube determined by the Cartesian product \([-1, 1]^W\). As such, a link that is closer to the centre of a \( W \)-dimensional cube is weaker than a link closer to the surface. This can be represented as \( \Delta_{ij} \), the Euclidean distance of links to the origin. To have representative values of all \( \tilde{\Phi} \), we normalise \( \Delta_{ij} \) by each \( \Phi \) and \( \tilde{\Phi} \) category. We denote the normalised distance as \( \Delta_{ij}^\ast \). The \( \Delta_{ij}^\ast \) normalised by \( \Phi \) is denoted as \( \Delta_{\Phi} \). We also calculate what we term the internal score,
denoted as \( \Delta_3 \), a categorical weight defined as the distance of each link to the best well-defined theoretical point that the link belongs to, \( \tilde{\rho}_{i,j} \). This measure assures that a particular link is clustered in the correct \( \tilde{\Phi} \) category. The ratio between \( \Delta_\tilde{\Phi} \) and \( \Delta_\tilde{\rho} \) simultaneously scores how well clustered and strong a link is: The higher this ratio, the better defined a link is. Links with a ratio smaller than unity should be removed from posterior analysis.

![Network 1](image1)
![Network 2](image2)
![Network 3](image3)

**Figure 1.** Visual representation of the CoDiNA method for a 3-network comparison. Panel I displays three independent networks to be compared. Red links represent positively correlated gene-pairs, and blue links negatively correlated ones. Node-size is relative to node strength. Panel II shows the geometrical representation of CoDiNA: a three-dimensional scatter-plot that is derived from plotting the weights from each link in the three networks. Different \( \Phi \) regions: green for \( \alpha \) (Common) links; \( \beta \) (Different) are red; blue corresponds to \( \gamma \) (Specific) links. Scores are represented as arrows: arrows away from centre to the point correspond to \( \tilde{\Phi} \) scores (\( \Delta_\tilde{\Phi} \)), while arrows from the point towards the centre represent the internal scores (\( \Delta_\tilde{\rho} \)). Panel III shows the relationship between the scores. Panel IV exhibits the CoDiNA network filtered for a score-ratio greater than unity. Link-width corresponds to link strength, and colour represents the \( \tilde{\Phi} \) class. Note that, not all links appear in the filtered network.

The classification of links is not sufficient to describe a network, which is why it is also necessary to categorise the nodes. We use a \( \chi^2 \) goodness-of-fit test to test if the frequency of the links in each category is different from the overall expected frequency in that category. If the null hypothesis is rejected, the \( \Phi \)-category with the maximum number of links is assigned to that particular node. Similarly for the \( \tilde{\Phi} \) node categorisation. More detailed information on CoDiNA can be found in the Online Methods.

**Comparing CoDiNA to other methods**
Evaluating multiple co-expression network methodologies for comparing networks is considerably challenging due to the lack of a gold standard network for multiple tissues\(^{26} \), of which all links are experimentally detected. Therefore, we are only able to identify similarities and differences among the methodologies.
To the best of our knowledge, only one other tool, CompNet\textsuperscript{1}, allows for the comparison of more than two networks. The focus of CompNet is on the visualisation of pairwise Jaccard-similarities from the union, intersections and exclusive links of those networks. It includes features such as pie-nodes and links to allow the user to identify key elements of the network. Elements are identified by providing a distribution of global graph properties, such as the amount of nodes, links, density, clustering coefficient, average path length and diameter, of the networks. Even though, building a visualisation tool, is not the focus of CoDiNA, we also incorporated an interactive tool for visualisation of the final network, and CoDiNA provides summary statistics of the network, such as the total number of links and nodes as well as how many links and nodes have been classified as common, different or specific to each category.

Lichtblau et. al. (2017)\textsuperscript{26} compared ten differential network analysis methods that are able to perform pairwise comparison. The authors split the methods into two main categories: Local search and Global search. Global methods focus on changes in the network topology while local methods search for changes in the nodes. CoDiNA combines both: it first searches for changes in the topology of the networks and then for specificity of the nodes. This allows investigating both features with one tool. Changes in network topology indicate alterations in affected pathways or regulatory relationships, while changes concerning specific nodes can evaluate the importance of particular genes for the network and suggest genes that might be responsible for the topology differences. Together, the local and global changes are crucial for understanding the functional effects of network changes.

**CoDiNA: an R package to compare co-expression networks**

To make the proposed methodology publicly available, an R package called CoDiNA was developed (2), where all the presented steps are implemented. The R package also includes an interactive visualisation tool. The main functions of the package are MakeDiffNet, which categorises all the links into $\Phi$ and $\Phi^c$ categories, calculates the internal scores and the normalised scores, and ClusterNodes, which clusters the nodes into the different categories. The visualisation tool can be accessed using the function plot.

**Pairwise version of CoDiNA: iPSC cells undergoing Neurogenesis had its process unravelled by CoDiNA**

An earlier version of CoDiNA for a pairwise network comparison was already successfully applied to an expression dataset of human induced Pluripotent Stem Cells (iPSC)\textsuperscript{7} (in press). The iPSCs were induced to undergo neuronal differentiation within four days. We compared expression patterns of wildtype and miR-124 deleted iPSCs over the time course of differentiation with the goal to uncover the function of miR-124 during neurogenesis. The experiment was conducted in seven replicates to facilitate the construction of a co-expression network for each day for the wildtype and the knockout cells. Using the pairwise version of CoDiNA, we revealed strong network differences between wildtype and knockout cells on day 3 of neuronal differentiation. For that day, CoDiNA classified the transcription factor (TF) ZNF787 as the TF with most specific links, suggesting it as one of the drivers of miR-124 induced network changes. Since ZNF787 is more highly expressed in miR-124 knock-out compared to wildtype iPSC, we overexpressed ZNF787 in the wildtype cells to experimentally validate our result. This overexpression still resulted in neuronal differentiation but was associated with some alterations in expression of genes underlying the specificity of ZNF787 in the CoDiNA network. These results strongly suggest ZNF787 as one repressor of neuronal features. Importantly, the experiment demonstrated that predictions found by CoDiNA could be experimentally verified.

**Example applications for situation of more than two networks**

Here we present two example applications of the CoDiNA method. In the first example, we use CoDiNA to analyse two tuberculosis (TB) studies with patients with and without human immunodeficiency virus (HIV) infection. The second showcase uses data from a study with patients with three types of glioma.

**CoDiNA applied to a tuberculosis and HIV study**

**Database**

The dataset contains expression data from peripheral blood of children and adults from two TB and HIV studies. In this application, our aim is to identify similarities and differences in TB and HIV in both age groups. Both studies are available at GEO; the first one (GSE39941\textsuperscript{27}) contains gene expression data from 192 children with TB from Kenya, South Africa and from Malawi (HIV$^+$ $n = 69$, HIV$^-$ $n = 123$); the second one (GSE37250\textsuperscript{28}) contains gene expression data from 197 adults with TB from South Africa and Malawi (HIV$^+$ $n = 99$, HIV$^-$ $n = 98$). Both studies aimed to define transcriptional signatures for detection of TB patients with and without HIV. We used the raw data provided at GEO\textsuperscript{29,30}, pre-processed and normalised them and performed quality control using R and the R package lumi\textsuperscript{31-33}.
Figure 2. Workflow process of the CoDiNA R package. Input data for the CoDiNA R package can be any network, filtered for only significant links. Edge list is a list containing all the links and its weights. To links for which the $p$-value is not significant, the user can assign a weight of zero. A minimum input requirement is of two networks with at least one link each. The function MakeDiffNet() clusters the links into the $\Phi$ and $\tilde{\Phi}$ categories, calculates and normalises the scores. Its output is used as input for clustering the nodes into categories by the function ClusterNodes(). The plot() function can be used on the output from MakeDiffNet() and automatically calls the function ClusterNodes().
We found 123 genes associated with HIV+ children, 208 to HIV− children. We compared the full gene co-expression network of HIV− children with the remaining nodes that were not classified into any of these categories. There were no significant \( p \)-values using the \( \omega \) method for positive and negative interactions. Importantly, among the unclassified nodes, 573 were classified as HIV− children, 33 were set to zero. Finding a large absolute \( \omega \)-value for a pair of genes means that the expression patterns of both genes are strongly (positively or negatively) correlated. The \( \omega \) package \( \omega \) was used for this calculation. The parameters used to build the networks were Pearson product-moment correlation coefficient and bootstrap re-sampling method.

Building the networks

Networks were generated separately for adults and children that are HIV+ or HIV− using the weighted Topological Overlap (wTO) method for positive and negative interactions\(^ {5,34} \) for all the 13,817 genes with 1,000 bootstraps. wTO values \( \omega_{ij} \), that were not significant (\( p \)-value \( \leq 0.001 \)) or with \( |\omega_{ij}| < 0.33 \) were set to zero. Finding a large absolute wTO-value for a pair of genes means that the expression patterns of both genes are strongly (positively or negatively) correlated. The \( \mathcal{R} \) package \( \omega \) was used for this calculation. The parameters used to build the networks were Pearson product-moment correlation coefficient and bootstrap re-sampling method.

Defining the Gene vs Disorder Enrichment

In order to define the disorders enriched in each \( \Phi \) category, we test if the amount of genes associated to HIV or TB classified in each category is different than random using an exact Fisher’s test and a proportion test. We combine both \( p \)-values using the Fisher’s method and use this as a weight to filter the results. The list of genes associated with HIV Infections, AIDS (Acquired Immunodeficiency Syndrome) or sAIDS (Simian Acquired Immunodeficiency Syndrome) or Tuberculosis was retrieved using the tool Gene 2 Disease tool (GS2D)\(^ {36} \). For our disease enrichment analyses we only considered those genes that were measured in the final CoDiNa networks.

Comparing the Tuberculosis networks using CoDiNa

This dataset gave rise to three main comparisons of interest: (i) Adults with TB: HIV− vs HIV+; (ii) Children with TB: HIV− vs HIV+; (iii) Adults and Children with TB: HIV− vs HIV+.

In order to associate the \( \Phi \) links to each one of the genes, we used the CoDiNa networks filtered for values where the ratio of the \( \Delta_{\Phi} \) and \( \Delta_{\Psi} \) is greater than unity. This means that we only present links that are the most distant to the centre: links with highest scores of being highly specific, highly different or highly common, and are most well clustered. After assigning the gene \( \Phi \) category, we performed an enrichment test for disorders among high scoring nodes.

Comparing the HIV− vs HIV+ networks

We compared first the full gene co-expression network of HIV− and HIV+ adults. In this comparison, CoDiNa was able to identify 80,509 links and 3,786 nodes. From those nodes, 455 are \( \alpha \) type, 172 \( \gamma_{HIV}^− \) and 1,948 \( \gamma_{HIV}^+ \) (Figure 3, panel I), while the remaining nodes were not classified into any of these categories. There were no \( \beta \) nodes, although \( \beta \) links existed. Importantly, among the nodes not classified as \( \alpha \), \( \beta \) or \( \gamma \), our enrichment analysis showed an over-representation of TB. This is to be expected since all individuals were infected with TB.

When comparing the networks from the data regarding children, CoDiNa identified 24,3645 links and 6,763 nodes. From those nodes, 573 are of \( \alpha \), 3,546 of \( \gamma_{HIV}^− \) and 926 of \( \gamma_{HIV}^+ \) category, while 1718 were unclassified (Figure 3, panel II). Our enrichment analysis found over-representation of genes related to AIDS and sAIDS for the HIV positive children. The HIV negative group is enriched for genes related to TB.

Our last comparison included data from both children and adults. This final network identified 35,683 links connecting 4,254 nodes. The nodes were classified as 77 of \( \alpha \) type, 44 specific to adults, 33 to HIV− adults and 430 to HIV+ adults. We found 123 genes associated to children, 208 to HIV+ children and 1,351 to HIV− children. Only 28 genes are common between HIV+ in adults and children, and only 11 to HIV− of both age groups (Figure 3, panel III). Among the unclassified genes, we again find enrichment for TB, similar to the network for adults. We further find an association to AIDS in HIV+ positive children and sAIDS in adults and children infected by HIV (Table 1).

Thus, CoDiNa was able to successfully identify an enrichment of known genes associated with HIV infections among the specific nodes, providing support for the ability of CoDiNa to retrieve biological meaningful results. Importantly, we were also able to pinpoint modules of genes related to each one of the co-infections.
Figure 3. Comparing children and adults with tuberculosis and tested for HIV: The three panels show the categories of links and nodes for each CoDiNA networks as well as the final CoDiNA network for HIV related genes for Panel I adults, Panel II children, and Panel III adults and children. Note that, in the adults network (Panel I) we mainly see genes that are related to HIV+ state, while for children (Panel II), there is a big loss of gene functions as judged from the specificity of genes and links to HIV− state. This is combined in the full CoDiNA network of adults and children (Panel III).
CoDiNA applied to a cancer study

Database

For the second showcase of our method, we used the gene-expression data from GSE4290, a study of patients with glioma. The dataset contains 157 brain tumour samples of three types (26 astrocytomas, 50 oligodendrogliomas, and 81 glioblastomas).

The data was downloaded from the GEO website, pre-processed and normalised by ourselves. The cancer expression profiles were normalised with the controls. We used micro-array data, which was analysed using the R environment and the affy package from the Bioconductor.

The probe expression levels (RMA expression values) and MAS5 detection p-values were computed, and only probesets significantly detected in at least one sample ($p$-value $\leq 0.05$) were considered. After quality control and probe normalisation, the probes that were not specific to only one gene were removed. If one gene was bound by more than one probeset, the average expression was computed.

Building the networks

Because TF deregulation is central to disease progression in many disease states, and particularly in cancers, we focused on a comparison of the TF co-expression networks between the three different kinds of tumours. To this end, we calculated the wTO network of the TFs for each tumour. We computed the wTO network for each cancer dataset and the controls separately using only the set of 3,229 unique TF symbols from the Gene Regulatory Factors (GRF)-Catalogue, filtered by genes with proteins that also are included in the ENSEMBL protein dataset.

Links with Benjamini-Hochberg adjusted $p$-values smaller than 0.01 were kept, and links with larger $p$-values were set to zero. Setting the non-significant wTO values to zero ensures that all nodes that were measured can be present in the final CoDiNA networks instead of being removed in the first step of the approach.

Defining the Gene vs Disorder Enrichment

To verify the enrichment of disorders in each one of the $\Phi$ classes, we test if the amount of genes associated to each one of the gliomas under study is different than random using an exact Fisher’s test and a proportion test. Both $p$-values are combined into one using the Fisher’s method and the resulting $p$-value was used as a weight to filter the results. The association of genes to disorders was retrieved using the tool GS2D. To perform the enrichment test, we used as background only the genes of that list that were expressed in the samples.

Comparing the Networks

In total, the CoDiNA network contains 2,209 nodes and 206,856 links above the score ratio threshold. According to the GS2D tool (weight $< 0.10$) 8 TFs are described in literature to be associated to Astrocytoma, 3 with Oligodendroglioma and 51 with Glioblastoma (Table 2). In our CoDiNA network we identified one of the 8 known Astrocytoma TFs to be associated with Astrocytoma. Two of the known Oligodendroglioma and 45 of the known Glioblastoma TFs were associated with the respective glioma types by CoDiNA, providing strong support for the validity of our comparative network approach.

In addition, we identified several TFs specifically associated with Astrocytoma that were not previously linked to this type of cancer (Fig. 4, panel I). The TFs with the 10 strongest associations are: FGD1, TCEAL4, ZNF628, TBPL1, BMP5, MYPOP, HMGA2, PRR3, MIS18BP1, BMP7. Among these, HMGA2, TBPL1, BMP5 and BMP7 were previously described as associated to neoplasm and neoplasm metastasis. The most strongly differentiated TFs associated with Oligodendroglioma not previously described (Fig. 4, panel II) were: SMARCE1, ZNF274, NRG1, ZNF232, UBE2I, TXK, TAF11, PLXNB2, HLX and SAP30BP. Of these, SMARCE1, NRG1, PLXNB2 and UBE2I were previously described as associated to Neoplasm Invasiveness and neoplastic cellular transformation. The TFs most specific for Glioblastoma (not previously described) (Fig. 4, panel III) were: ZNF558, PTBP1, XRN2, RNF114, ZNF45, ZNRD1, KHDRBS2, RFXANK, NIFK and ZNF540. Here, the genes PTBP1, RNF114, XRN2 and ZNRD1 are described as associated with other neoplasms. This suggests, that CoDiNA can be applied to detect novel candidate genes for specific cancer types. We were able to identify TFs previously associated with other neoplasms, but not the types of glioma under study, in important roles in the differential glioma network, indicating that those TFs are also deregulated in those disorders.

---

Table 2. Disease Enrichment Analysis for each $\Phi$ group in a Cancer-Glioma Study.

| Disease            | Phi   | Expected | Observed | weight |
|--------------------|-------|----------|----------|--------|
| Astrocytoma        | $\Phi$| 8        | 1        | 0.06   |
| Oligodendroglioma  | $\Phi$| 3        | 2        | 0.06   |
| Glioblastoma       | $\Phi$| 51       | 45       | 0.10   |

---

8/16
Figure 4. TF-TF CoDiNA networks for each of the Glioma types: CoDiNA identified TFs with specific co-regulation changes to each cancer, Panel I Astrocytoma, Panel II Glioblastoma, Panel III Oligodendroglioma. Panel I: We can see that mostly glioma and Astrocytoma TFs are present as specific to this network. Panel II: There is an overlap of TFs involved in other gliomas, but the biggest amount of nodes refers to specific changes in co-expression in Glioblastoma. Panel III: Even though most links and genes are specific to Oligodendroglioma, we can identify overlapping TFs of this cancer and Astrocytoma.

Conclusion
We presented a novel method that allows for the systematic comparison of multi-dimensional data in different conditions and the representation of the analysis as a single network. In particular, our method identifies links and nodes that are common to all networks under consideration, specific to at least one network, or have different signs among the compared networks. To evaluate CoDiNA, we applied it to a neurogenesis study where it identified genes for neuron differentiation, of which which one was experimentally modified, confirming our assessment of it’s importance. Applying our method to an HIV and TB study, CoDiNA retrieved networks that were enriched for genes involved in HIV. With multiple glioma-type cancer datasets, we identified network signatures that are specific to each type of glioma. More importantly, we were able to identify genes previously associated with disorders and identify new genes interacting with those. This suggests that our method produces biologically meaningful results. We expect that our method will be helpful for many diverse studies comparing network data generated from multiple conditions, such as different diseases, tissues, species or experimental treatments.

Availability
CoDiNA is open source and freely available from CRAN https://cran.r-project.org/web/packages/CoDiNA/ under the GPL-2 Open Source License. It is platform independent.

Methods
How the CoDiNA method works
Let \( \mathcal{W} \) be a set of networks constructed using a similarity score, \( p_{i,j} \) with scalar range \([-1, 1]\), e.g. correlation methods, and each network contains only links with weights found to be statistically significant given a predefined threshold value. For each link, we classify its presence in \( \mathcal{W} \) into one of the three following \( \Phi \) categories based on its link weights (one link-weight value per network): a link is said to be an \( \alpha \) link, if it is present in all networks with the same sign on its weight, i.e., this link is common to all networks. A link is considered as \( \beta \) type if it is present in all networks but with different signs of the link weights, i.e., the links represents a qualitatively different kind of interaction in at least one network. This interaction might indicate,
e.g. that a particular gene changed its function. Finally, a link is considered to be of $\gamma$ type if it is present (i.e. statistically significant non-zero weight) in some networks but not all, regardless of the sign of the link weight. The $\gamma$ link-type is specific to at least one network.

We may further sub-divide the $\Phi$ categorisation of each link into $\Phi$ according to the networks containing said link. This sub-classification is mainly important for understanding the behaviour of the $\beta$ and $\gamma$ links. Since we work with networks typically constructed using correlations (and thus setting the weight of a link as a correlation measure), the link weight value $\rho_{ij} \in [-1, 1]$. We partition, by default, the interval in three equal parts ($\tau = 1/3$), which will be denoted as corresponding to a positive link, negative link or neutral link. If the compared networks have different link-weight ranges, these may be normalised by using a multiplicative (stretch) parameter.

In order to avoid false associations, an important step to be aware of is that a node should be present in all networks; if a node is absent in at least one network, we remove all of its links in the networks where this node is present. This step is implemented to prevent the erroneous inference that a particular node is associated with a specific condition, when in fact that specific node possibly was not measured in the other conditions. If a link weight is found not to be significant, we assign it a (weight) value of zero, thus allowing all measured nodes to be included in the analysis even when only some of its links are significant.

For a system consisting of a set $\mathbb{N}$ of nodes, where $N$ is the cardinality of $\mathbb{N}$ in the $\mathbb{W}$ networks, let $\rho_{ij,k}$ be a specific link weight between nodes $i$ and $j$ in network $k$. Each link is categorised as follow

$$\tilde{\rho}_{ij,k} = \begin{cases} -1 & \text{if } \rho_{ij,k} < -\tau \\ 1 & \text{if } \rho_{ij,k} > \tau \\ 0 & \text{otherwise,} \end{cases}$$

where $\tilde{\rho}_{ij,k}$ is an integer transformation of the link weight based on the threshold, and $\tau$, in the standard implementation of CoDiNA, is $1/3$, which allows the space to be divided equally in 3 parts. If a particular link categorical weight $\tilde{\rho}_{ij,k}$ is zero in all the $\mathbb{W}$ networks, this link is removed from posterior analyses.

After the correlation values are coded into the categorical variables $\tilde{\rho}$, each link is assigned to an additional group category that shows in which condition the link is present and what its sign is, if present. This classification step is particularly important for links that are classified as $\beta$ or $\gamma$ type, because it is straightforward to clearly identify in which network(s) the link is specific or different. The maximum number of groups is $G_{\text{max}} = (3^W - 1)$, where $W$ denotes the cardinality or number of networks in the set. Note that, the group for which all category values are equal to zero is removed from analyses. Our approach is to assign an $\alpha$, $\beta$ or $\gamma$ classification to each of the links, by defining $\Phi$ as follow

$$\Phi_{ij} = \begin{cases} \alpha_{ij} & \text{if } \sum_{k=1}^{W} |\tilde{\rho}_{ij,k}| = W \land \sum_{k=1}^{W} |\tilde{\rho}_{ij,k}| = #(W) \\ \beta_{ij} & \text{if } \sum_{k=1}^{W} |\tilde{\rho}_{ij,k}| = W \land \sum_{k=1}^{W} |\tilde{\rho}_{ij,k}| < #(W) \\ \gamma_{ij} & \text{otherwise.} \end{cases}$$

Each link receives a sub-category, $\tilde{\Phi}_{ij}$, based on the pattern of networks in which that link exists. This makes it more straightforward to interpret the links in each of the categories, and as a result this improves our ability to identify links that are specific to a subset of networks and the subset of networks it has a different behaviour.

To illustrate the concept of sub-category, assume the following $\tilde{\rho}$ of a particular link in 3 networks: Network A value is 1, Network B is 1 and Network C is also 1. Because the value 1 is common in the three networks, this $\Phi$ category is clearly $\alpha$, and no further explanation is needed. Now, take as a second example, Network A has the value 1, Network B, -1, and Network C 1. Its $\Phi$ class is $\beta$, but this class cannot help us understand where the change occurs, therefore, the $\Phi$ is needed. Its $\Phi$ class is $\beta_{\text{Network}2}$. Important to note is that CoDiNA assumes the first network to be the reference network. As a final example, assume that the $\tilde{\rho}$ weight of the three networks are 0, 1 and 1 for Networks A, B and C, respectively. This link does not occur in network A, so it is a $\gamma$ links, that is specific to networks B and C. But this is not possible to understand only by reading that its category is $\gamma$, therefore, its $\Phi$ category is $\gamma_{\text{NetworkB,NetworkC}}$.

Let, $E$, be the set of links and $E$ the cardinality of $E$. When all $E$ links are assigned a $\Phi_{ij}$ category and further sub-categorised as $\tilde{\Phi}_{ij}$, we score the links to identify those that are stronger. For every node $i = 1, 2, \ldots, N$, we interpret the array of link weights $(\rho_{1i}, \ldots, \rho_{Ni})$ as a point in a $W$-dimensional Euclidean space. In particular, as each link weight is bounded, all points are contained in the cube determined by the Cartesian product $[-1, 1]^W$.

As such, a link that is closer to the centre of the $W$-dimensional cube is weaker than a link closer to the links. Based on that, the Euclidean distance, $\Delta$, to the origin of the space is calculated for all links $E_{ij}$.
However, since links closer to corners will trivially have a larger $\Delta$ compared to the others, all distances are penalised by the maximum theoretical distance a link can assume in its category. Consequently, we define a normalised distance, $\Delta_{ij}$ in Equation (2) so that it is in the unit interval

$$\Delta_{ij} = \sqrt{\frac{\sum_{k=1}^{W} P_{ij,k}^2}{\sum_{k=1}^{W} |\rho_{ij,k}|}}.$$  \hfill (2)

We test if some link-clusters are closer to the border of the cube than the others by a regression model, where the distance to the centre is the dependent variable and the categories are the covariates of the model. Indeed, if statistical differences between clusters and distance is detected, we only select those nodes that belong to one particular cluster of links: the cluster that is the furthest away from the centre. Normalising the distance can be a way to overcome this challenge, by use of Equation (3)

$$\Delta^*_{ij} = \frac{\Delta_{ij} - \min(\Delta_{ij})}{\max(\Delta_{ij}) - \min(\Delta_{ij})}.$$  \hfill (3)

Three different approaches may be applied to the normalisation:

- Normalise all the links together: Here, we do not consider if a complete cluster is situated near the surface or closer to the centre of the cube;

- Normalise links according to their $\Phi_{ij}$ and $\tilde{\Phi}_{i,j}$ class: In this alternative, all the categories are a part of the final output. This means that if one of the $\Phi$ groups lies inside the cube closer to its centre compared to the other $\Phi_{ij}$ categories, it will be possible to see links that belong to this category in the final network.

Another important Score calculated by CoDiNA, called internal Score, denoted by $\Delta_\Phi$, measures the distance from the link $ij$ to the theoretical best well clustered link in that particular $\Phi$ category. In other words, if a link is considered an $\alpha$ with all positive links, we calculate its distance to the point $(1,1,1)$. This score allows us to identify links that are most well defined for each $\Phi$ category.

Because the two scores $\Delta_\Phi$ and $\Delta_\tilde{\Phi}$ are highly negative correlated, the ratio between them also gives us a measure of the very best well defined links. For a well defined CoDiNA network, this ratio should be greater or equal than 1.

Knowing only the links classification is not sufficient to describe a network; we are also interested in the nodes’ classification. To define the $\Phi$ category of a particular node, we make a frequency table of how many times each node had a link in each $\Phi$ category and sub-category. We test, using a $\chi^2$ goodness-of-fit test, if the links of a node are distributed equally in all categories. If the null hypothesis is rejected, the $\Phi$-category with the maximum number of links is assigned to that particular node. Similarly, the same is done for the $\tilde{\Phi}$.

\textbf{Algorithm 1 Description of the RemoveNodes procedure.}

\textbf{Input:} Set of $\mathbb{N}$ nodes that belongs to the set of $\mathcal{W}$ networks;

\textbf{Output:} Set of common nodes to all networks

1: \textbf{procedure} \textsc{RemoveNodes}(N$_1$, \ldots , N$_W$)
2: \hspace{1cm} Set_Nodes = $\bigcap_{j=1}^{W} \text{Nodes}_j$
3: \hspace{1cm} \textbf{return} Set_Nodes.
4: \textbf{end procedure}
Algorithm 2 Description of the links categorisation algorithm.

**Input:** Set of \( W \) networks with \( N \) nodes (\( W \geq 2; N \geq 2 \));

**Output:** Network with links weight categorised into -1, 0 or 1;

1: Set \( \tau > 0 \);
2: By default \( \tau \leftarrow 1/3 \);
3: procedure ASSIGNCLASSES
4: for \( \rho \ j \leftarrow 1 \) to \( W \) do
5: for \( \rho \ j \ i \leftarrow 1 \) to \( E \) do
6: if \( \rho_{ij} < \tau \) then
7: \( \tilde{\rho}_{ij} \leftarrow -1 \);
8: else if \( j > \tau \) then
9: \( \tilde{\rho}_{ij} \leftarrow 1 \);
10: else
11: \( \tilde{\rho}_{ij} \leftarrow 0 \);
12: end if
13: end for
14: end for
15: end procedure

Implementation: The CoDiNA R package

To make the proposed methodology publicly available, an R package, called CoDiNA, was developed, where all the presented steps are implemented. The R package also includes an interactive visualisation tool. The functions included in the package are:

- **normalize:** Normalise a variable according to Equation (3);
- **OrderNames:** Reorder the names of the nodes for each link in alphabetical order;
- **MakeDiffNet:** Categorise all the links into \( \Phi \), \( \tilde{\Phi} \) and Group the categories. Calculate the normalised Scores;
- **plot:** Classifies the nodes into \( \Phi \) and \( \tilde{\Phi} \) following a user-defined cutoff for the chosen distance and plots the network in an interactive graph, where nodes and links can be dragged, clicked and chosen according to its group or classification. The size of a node is relative to its degree. Nodes and links that belong to the \( \alpha \) (common) group are coloured in shades of green and have a triangle shape. Nodes belonging to the \( \beta \) (different) group are coloured in shades of red and have a square shape. Nodes of the \( \gamma \) (specific) group are coloured in blue and their shape is a star. Nodes have a category for group and \( \Phi \) or \( \tilde{\Phi} \), according to a \( \chi^2 \)-goodness of fit test as defined above. If a node is group-undetermined it is diamond-shaped, and if a particular node is cluster-undetermined it is grey. The user can also choose a layout for the network visualisation from those available in the igraph package\(^{46}\). It is further possible to cluster nodes, using the parameter MakeGroups, and the user may select among the following clustering algorithms: “walktrap”\(^{47}\), “optimal”\(^{48}\), “spinglass”\(^{49-51}\), “edge.betweenness”\(^{52,53}\), ‘fast_greedy’\(^{54}\), “infomap”\(^{55,56}\), “louvain”\(^{57}\), “label_prop”\(^{58}\) and “leading_eigen”\(^{59}\). These algorithms are implemented in the igraph package\(^{46}\).
- The AST data.table contains the nodes and the weighted topological overlap (wTO) of Transcription Factors (TFs), from GSE4290\(^{37}\) for astrocytomas;
- The GLI data.table contains the nodes and the wTO of TFs, from GSE4290\(^{37}\) for glioblastomas;
- The OLI data.table contains the nodes and the wTO of TFs, from GSE4290\(^{37}\) for oligodendrogliomas;
- And the CTR data.table contains the nodes and the wTO of TFs, from GSE4290\(^{37}\) for controls.

Acknowledgements

This work was supported partially by a doctoral grant from the Brazilian government’s Science without Borders program (GDE 204111/2014 – 5). This work was partly developed during a D.M.G. internship at the Almaas Lab and she expresses her gratitude to members of this group for inputs and great discussions. The authors would also like to thank Alvaro Perdomo-Sabogal for sharing the GRF manually curated database. The authors would further like to thank Wesley Bertoli for proofreading the manuscript.
Algorithm 3 Description of the Φ algorithm.

Input: Set of \( \mathbb{W} \) networks with \( \mathbb{N} \) nodes (\( W \geq 2; N \geq 2 \));
Output: Network with links categorised into \( \alpha, \beta \) or \( \gamma \)

1: Set \( \tau > 0 \);
2: procedure \textsc{PhiLinks}
3: for \( \bar{\rho} i j \leftarrow 1 \) to \( E \) do
4: for \( \bar{\rho} i j, k \leftarrow 1 \) to \( W \) do
5: if \( \sum_j W \bar{\rho} i j, k = 0 \) then
6: remove link;
7: else if \( \sum_j W \bar{\rho} i j, k \) \& \( \sum_j W \bar{\rho} i j, k = W \) then
8: \( \Phi i j \leftarrow \alpha \);
9: else if \( (\sum_j W \bar{\rho} i j, k \) \& \( \sum_j W \bar{\rho} i j, k) \neq W \) then
10: \( \Phi i j \leftarrow \beta \);
11: else
12: \( \Phi i j \leftarrow \gamma \);
13: end if
14: if \( \bar{\rho} i j, k = 1 \) then
15: Positive\(_{ij} \leftarrow 1 \)
16: else if \( \bar{\rho} i j, k = -1 \) then
17: Negative\(_{ij} \leftarrow 1 \)
18: end if
19: \( \Delta^*_i j \leftarrow \text{Equation (2)} \).
20: end for
21: end for
22: end procedure

Author contributions statement
Conceptualisation, E.A. and D.M.G; Methodology, E.A., D.M.G, K.N. T.M.F.; Software, D.M.G.; Validation, D.M.G., K.N. and V.B.; Formal Analysis, D.M.G., K.N. and V.B.; Resources, D.M.G., E.A, K.N., and V.B.; Data Curation, D.M.G.; Writing – Original Draft, D.M.G and K.N.; Writing – Review and Editing, all authors; Visualisation, D.M.G.; Supervision, E.A. and K.N.; Funding Acquisition, E.A., K.N., and V.B.

Additional information
Correspondence and requests for materials should be addressed to D.M.G. (email: deisy@bioinf.uni-leipzig.de) or K.N (email: katja.nowicki@fu-berlin.de)
Algorithm 4 Description of the node-categorisation algorithm.

**Input:** Set of $N$ nodes with $E$ links ($E \geq 2; N \geq 2$);

**Output:** Node classified as $\alpha$, $\beta$ or $\gamma$ type

1: procedure PHIODES
2: \hspace{1em} for $i \leftarrow 1$ to $N$ do
3: \hspace{2em} $\Phi_\alpha = \text{Count } \alpha$;
4: \hspace{2em} $\Phi_\beta = \text{Count } \beta$;
5: \hspace{2em} $\Phi_\gamma = \text{Count } \gamma$;
6: \hspace{1em} Test if $\Phi_\alpha = \Phi_\beta = \Phi_\gamma$
7: \hspace{1em} end for
8: end procedure

Algorithm 5 Description of the CoDiNA algorithm.

1: Call: REMOVEODES
2: End Call
3: Call: ASSIGNCLASSES
4: End Call
5: Call: PHI LINKS
6: End Call
7: Call: PHI NODES
8: End Call

References

1. Kuntal, B. K., Dutta, A. & Mande, S. S. CompNet: a GUI based tool for comparison of multiple biological interaction networks. *BMC Bioinformatics*. 17, 185, DOI: 10.1186/s12859-016-1013-x (2016).
2. Barabasi, A.-L. & Oltvai, Z. N. Network biology: understanding the cell’s functional organization. *Nat. Rev. Genet.* 5, 101–113 (2004).
3. Bansal, M., Belcastro, V., Ambesi-Impiombato, A. & Di Bernardo, D. How to infer gene networks from expression profiles. *Mol. Syst. Biol.* 3, 78 (2007).
4. Furlong, L. I. Human diseases through the lens of network biology. *Trends Genet.* 29, 150–159 (2013).
5. Dempsey, K. et al. On mining biological signals using correlation networks. In *Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on*, 327–334 (IEEE, 2013).
6. van Dam, S., Vösa, U., van der Graaf, A., Franke, L. & de Magalhães, J. P. Gene co-expression analysis for functional classification and gene–disease predictions. *Briefings Bioinformatics*. bbw139 (2017).
7. Berto, S., Perdomo-Sabogal, A., Gerighausen, D., Qin, J. & Nowick, K. A consensus network of gene regulatory factors in the human frontal lobe. *Front. Genet.* 7, 31 (2016).
8. Gysi, D. M., Voigt, A., Fragoso, T. d. M., Almaas, E. & Nowick, K. wTO: an R package for computing weighted topological overlap and consensus networks with an integrated visualization tool. *arXiv preprint arXiv:1711.04702* (2017). 1711.04702.
9. Fukushima, A. Diffcorr: an R package to analyze and visualize differential correlations in biological networks. *Gene* 518, 209–214 (2013).
10. Del Sol, A., Balling, R., Hood, L. & Galas, D. Diseases as network perturbations. *Curr. opinion biotechnology* 21 (2010).
11. Bandyopadhyay, S. et al. Rewiring of genetic networks in response to dna damage. *Science* 330, 1385–1389 (2010).
12. De la Fuente, A. From ‘differential expression’to ‘differential networking’–identification of dysfunctional regulatory networks in diseases. *Trends genetics* 26, 326–333 (2010).
13. Chen, M. & Hofestädt, R. PathAligner: metabolic pathway retrieval and alignment. *Appl Bioinforma.* 3, DOI: 10.2165/00822942-200403040-00006 (2004).
14. Liang, Z., Xu, M., Teng, M. & Niu, L. NetAlign: a web-based tool for comparison of protein interaction networks. *Bioinformatics* 22, DOI: 10.1093/bioinformatics/btl287 (2006).
15. Watson, M. Coxpress: differential co-expression in gene expression data. *BMC bioinformatics* **7**, 509 (2006).
16. Tian, Y., McEachin, R. C., Santos, C., States, D. J. & Patel, J. M. SAGA: a subgraph matching tool for biological graphs. *Bioinformatics* **23**, DOI: 10.1093/bioinformatics/btl571 (2007).
17. Wu, J. *et al.* Integrated network analysis platform for protein-protein interactions. *Nat. methods* **6**, 75 (2008).
18. Dost, B. *et al.* QNet: a tool for querying protein interaction networks. *J Comput. Biol* **15**, DOI: 10.1089/cmb.2007.0172 (2008).
19. Lena, P. D., Wu, G., Martelli, P. L., Casadio, R. & Nardini, C. MIMO: an efficient tool for molecular interaction maps overlap. *BMC Bioinforma.* **14**, DOI: 10.1186/1471-2105-14-159 (2013).
20. Bass, J. I. F. *et al.* Using networks to measure similarity between genes: association index selection. *Nat. methods* **10**, 1169 (2013).
21. Voigt, A., Nowick, K. & Almaas, E. A composite network of conserved and tissue specific gene interactions reveals possible genetic interactions in glioma. *PLoS Comput. Biol.* **13**, e1005739 (2017).
22. Berto, S. & K. N. Species-specific changes in a primate transcription factor network provide insights into the molecular evolution of the primate prefrontal cortex. *Genome Biol. Evol.* (2018).
23. Yang, Y. *et al.* Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. *Nat. communications* **5**, 3231 (2014).
24. Yang, Q. *et al.* Identification of the key genes implicated in the transformation of olp to oscc using ma-sequencing. *Oncol. reports* **37**, 2355–2365 (2017).
25. Manuscript_DG.
26. Lichtblau, Y. *et al.* Comparative assessment of differential network analysis methods. *Briefings bioinformatics* **18**, 837–850, DOI: 10.1093/bib/bbw061 (2017).
27. Anderson, S. T. *et al.* Diagnosis of childhood tuberculosis and host RNA expression in Africa. *New Engl. J. Medicine* **370**, 1712–1723 (2014).
28. Kaforou, M. *et al.* Detection of tuberculosis in HIV-infected and-uninfected African adults using whole blood RNA expression signatures: a case-control study. *PLoS medicine* **10**, e1001538 (2013).
29. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research* **30**, 207–210 (2002).
30. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets—update. *Nucleic acids research* **41**, D991—D995 (2012).
31. Du *et al.* lumi: a pipeline for processing Illumina microarray. *Bioinformatics* (2008).
32. P. D. *et al.* Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinforma.* (2010).
33. Lin *et al.* Model-based Variance-stabilizing Transformation for Illumina Microarray Data. *Nucleic Acids Res* (2008).
34. Nowick, K., Gernat, T., Almask, E. & Stubbs, L. Differences in human and chimpanzee gene expression patterns define an evolving network of transcription factors in brain. *Proc. Natl. Acad. Sci.* **106**, 22358–22363 (2009).
35. Gysi, D. M. *et al.* wTO: Computing Weighted Topological Overlaps (wTO) & Consensus wTO Network (2017).
36. Fontaine, J. F. & Andrade-Navarro, M. A. Gene set to Diseases (GS2D): disease enrichment analysis on human gene sets with literature data. *Genomics Comput. Biol.* **2**, e33—e33 (2016).
37. Sun, L. *et al.* Neuronal and glioma-derived stem cell factor induces angiogenesis within the brain. *Cancer cell* **9**, 287–300 (2006).
38. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2016).
39. Gautier, L., Cope, L., Bolstad, B. M. & Irizarry, R. A. affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* **20**, 307–315, DOI: http://dx.doi.org/10.1093/bioinformatics/btg405 (2004).
40. Lee, T. I. & Young, R. A. Transcriptional regulation and its misregulation in disease. *Cell* **152**, 1237–1251 (2013).
41. Bhagwat, A. S. & Vakoc, C. R. Targeting transcription factors in cancer. *Trends cancer* **1**, 53–65 (2015).
42. Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. & Barabási, A.-L. Hierarchical organization of modularity in metabolic networks. *Science* **297**, 1551–1555 (2002).

43. Carlson, M. R. J. *et al.* Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. *BMC Genomics* **7**, 40 (2006).

44. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet. Mol Biol* **4**, Article17, DOI: 10.2202/1544-6115.1128 (2005).

45. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. royal statistical society. Ser. B (Methodological)* 289–300 (1995).

46. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal, Complex Syst* **1695**, 1–9 (2006).

47. Pons, P. & Latapy, M. Computing communities in large networks using random walks. *J. Graph Algorithms Appl.* **10**, 191–218 (2006).

48. Brandes, U. *et al.* On modularity clustering. *IEEE transactions on knowledge data engineering* **20**, 172–188 (2008).

49. Reichardt, J. & Bornholdt, S. Statistical mechanics of community detection. *Phys. Rev. E* **74**, 16110 (2006).

50. Newman, M. E. J. & Girvan, M. Finding and evaluating community structure in networks. *Phys. review E* **69**, 26113 (2004).

51. Traag, V. A. & Bruggeman, J. Community detection in networks with positive and negative links. *Phys. Rev. E* **80**, 36115 (2009).

52. Freeman, L. C. Centrality in social networks conceptual clarification. *Soc. Networks* **1**, 215–239 (1978).

53. Brandes, U. A faster algorithm for betweenness centrality. *J. Math. Sociol.* **25**, 163–177 (2001).

54. Clauset, A., Newman, M. E. J. & Moore, C. Finding community structure in very large networks. *Phys. review E* **70**, 66111 (2004).

55. Rosvall, M. & Bergstrom, C. T. Maps of information flow reveal community structure in complex networks. *arXiv preprint physics.soc-ph/0707.0609* (2007).

56. Rosvall, M., Axelsson, D. & Bergstrom, C. T. The map equation. *The Eur. Phys. Journal-Special Top.* **178**, 13–23 (2009).

57. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. statistical mechanics: theory experiment* **2008**, P10008 (2008).

58. Raghavan, U. N., Albert, R. & Kumara, S. Near linear time algorithm to detect community structures in large-scale networks. *Phys. review E* **76**, 36106 (2007).

59. Newman, M. E. J. Finding community structure in networks using the eigenvectors of matrices. *Phys. review E* **74**, 36104 (2006).