Identifying communities from multiplex biological networks

Gilles Didier[1], Christine Brun[2,3] and Anaïs Baudot[1]

[1] Aix Marseille Université, CNRS, Centrale Marseille, I2M UMR 7373, Marseille, France
[2] Aix Marseille Université, Inserm, TAGC UMR_S1090, Marseille, France
[3] CNRS, Marseille, France

ABSTRACT

Various biological networks can be constructed, each featuring gene/protein relationships of different meanings (e.g., protein interactions or gene co-expression). However, this diversity is classically not considered and the different interaction categories are usually aggregated in a single network. The multiplex framework, where biological relationships are represented by different network layers reflecting the various nature of interactions, is expected to retain more information. Here we assessed aggregation, consensus and multiplex-modularity approaches to detect communities from multiple network sources. By simulating random networks, we demonstrated that the multiplex-modularity method outperforms the aggregation and consensus approaches when network layers are incomplete or heterogeneous in density. Application to a multiplex biological network containing 4 layers of physical or functional interactions allowed recovering communities more accurately annotated than their aggregated counterparts. Overall, taking into account the multiplexity of biological networks leads to better-defined functional modules. A user-friendly graphical software to detect communities from multiplex networks, and corresponding C source codes, are available at GitHub (https://github.com/gilles-didier/MolTi).

INTRODUCTION

Biological macromolecules do not act in isolation, but rather interact with each other to perform their cellular functions. In particular, thousands of interactions are observed between proteins, forming the basis of cellular biological processes. Thanks to the scaling of the experimental techniques allowing interaction discovery, and to the development of centralized databases, recent years have witnessed the accumulation of thousands of physical and functional interactions of various nature (*Sharan & Ideker, 2006*). For instance, data obtained from yeast two-hybrid and affinity purification-mass spectrometry experiments inform on the physical interactions between proteins and their organization as molecular complexes; mRNA expression correlations identify the genes that are co-expressed in a set of experimental conditions, and signaling pathway data describe the cascades of reactions transmitting signals within and between cells.

Interaction data are usually represented as networks, i.e., graphs in which vertices correspond to genes or proteins, and edges to interactions. One of the leading approaches to extract functional knowledge from biological networks is their clustering into communities—or modules—of tightly linked genes/proteins (*Aittokallio & Schwikowski, 2006*). These modules are expected to represent the building blocks of the cells (*Hartwell et al., 1999*). Clustering algorithms developed for community detection are frequently based on graph topological properties, such as density or modularity (for reviews, *Wang et al., 2010*; *Pizzuti & Rombo, 2014*). They have been thoroughly assessed on biological networks (*Brohée & Van Helden, 2006*), and are widely used in biology (*Arroyo et al., 2015*; *Wan et al., 2015*; *Huttlin et al., 2015*; *Chapple et al., 2015*; *Katsogiannou et al., 2014*). However, they have been applied on single networks whereas recently *multiplex* (alt. *multi-layer* or *multi-slice*) networks (*Kivelä et al., 2014*) have been introduced in biology. These multiplex networks are sets of networks sharing the same nodes, but in which edges belong to different categories or represent interactions of different nature. As biological interaction sources are diverse, biological multiplex networks contain sparse layers of high relevance (e.g., curated signaling pathway networks), while other layers contain thousands or even millions of interactions (e.g., co-expression networks). In order to extract relevant biological information, it is thus important to assess the approaches able to detect communities from these multiplex biological networks.

As the majority of current clustering approaches takes as input a single network, the interactions from the different sources are classically aggregated into a unique network, herein called *monoplex*, in which all the edges are equally considered, regardless of their molecular nature. For instance, in the chaperome network (*Brehme et al., 2014*), protein–protein interactions and co-expression associations were aggregated and the link-community clustering algorithm (*Ahn, Bagrow & Lehmann, 2010*) was applied on the resulting monoplex network to identify communities. However, considering all the interaction categories as equivalent likely excludes important information, because the gene/protein relationships have different meanings or relevance (*Battiston, Nicosia & Latora, 2014*). On another hand, approaches known as *consensus clustering* aggregate *a posteriori* the communities obtained from several independent networks (*Senbabaoglu, Michailidis & Li, 2014*; *Lancichinetti & Fortunato, 2012*). Finally, more recently, different topological measures have been adapted to multiplex networks (*Kivelä et al., 2014*; *Battiston, Nicosia & Latora, 2014*), and a handful of community detection algorithms can take multiple networks as input (*Mucha et al., 2010*; *Shiga & Mamitsuka, 2012*; *Papalexakis, Akoglu & Ience, 2013*; *Bennett et al., 2015*). To our knowledge, these 3 sets of approaches, namely network aggregations, consensus clustering and multiplex approaches, have not been extensively compared and evaluated.

In order to do so, we carefully explored modularity-based approaches by identifying communities from multiplex networks using (i) the union-, intersection-, or sum-aggregations of the layers of the multiplex network, (ii) the consensus of the community obtained on the individual layers and (iii) a natural extension of the modularity, called here multiplex-modularity, that allows identifying communities directly from multiplex networks (similar to *Bennett et al. (2015)*). Importantly, the strong assumption that all

the network sources share the same community structure underlies all these approaches. More generally, this assumption is implicit for any approach returning a single community structure from a multiplex network.

We investigated whether combining multiple network sources could improve the detection of communities, in particular when noise impairs the inference of the community structure from individual networks. More particularly, we assessed the influence of important features characterizing biological networks, such as their heterogeneous densities and the weight of missing data. We first ran extensive simulations on random multiplex networks generated with a community structure, and mimicking the sparsity, the heterogeneity and the incompleteness of biological networks. We then showed that (i) the aggregation of the layers of the multiplex networks through their union and intersection ultimately erases the community structure, and (ii) the multiplex-modularity clustering outperforms the aggregation and consensus approaches. Importantly, the multiplex-modularity approach is more reliable than all other approaches when the set of studied networks are incomplete or heterogeneous in density, i.e., when networks display features of real biological networks.

We then studied a real-case of multiplex biological network composed of 4 interaction layers. Communities obtained with the multiplex-modularity are more associated to significantly enriched biological processes compared to their aggregated counterparts. We detailed, as an example, a module that clusters the proteins implicated in the Coffin-Siris syndrome. These proteins are clustered with proteins involved in other syndromes displaying overlapping clinical features.

Finally, we provide *MolTi*, a standalone graphical interface allowing the users to cluster multiple input networks, annotate the obtained clusters and parse the resulting communities.

MATERIAL AND METHODS

Multiplex networks, modularity and community detection
Monoplex and multiplex networks
Let X be a (monoplex) network, i.e., an undirected graph. We use X for its incidence matrix too, where the entry $X_{i,j}$ indicates the presence ($X_{i,j} = 1$) or the absence ($X_{i,j} = 0$) of an edge between the vertices i and j.

A *multiplex network* is a collection of undirected graphs $(X^{(g)})_g$ sharing the same set of vertices (*Kivelä et al., 2014*). A basic way to apply standard graph approaches on a given multiplex network is to first aggregate all its graph layers into a monoplex network. We considered here three types of aggregation, the intersection $X^\cap$, the union $X^\cup$ and the sum X^Σ of the graphs in $(X^{(g)})_g$, respectively defined as:

$$X_{i,j}^\cap = \bigwedge_g X_{i,j}^{(g)}, X_{i,j}^\cup = \bigvee_g X_{i,j}^{(g)} \text{ and } X_{i,j}^\Sigma = \sum_g X_{i,j}^{(g)}$$

for all pairs of vertices i and j, where $\wedge$ and $\vee$ denote the logical conjunction and the (inclusive) disjunction respectively. Note that $X^\cap$ and $X^\cup$ are unweighted graphs whereas X^Σ is a weighted graph.

Modularity

Under the assumption that a network X has a community structure $\mathbf{c}$ and by putting $\mathbf{c}_i$ for the community of the vertex i, the Newman–Girvan modularity (*Newman & Girvan, 2004*) is defined as:

$$Q(X,\mathbf{c}) = \frac{1}{2m} \sum_{\substack{\{i,j\} \\ i \neq j}} \left(X_{i,j} - \frac{k_i k_j}{2m} \right) \delta_{\mathbf{c}_i, \mathbf{c}_j} \tag{1}$$

$$= \sum_a \left[\frac{\sum_{\substack{\{i,j\} \\ i \neq j, \mathbf{c}_i = \mathbf{c}_j = a}} X_{i,j}}{2m} - \frac{\sum_{\substack{\{i,j\} \\ i \neq j, \mathbf{c}_i = \mathbf{c}_j = a}} k_i k_j}{(2m)^2} \right] \tag{2}$$

where m is the total number of edges of X, k_i is the number of edges containing the vertex i (i.e., the degree of i) and $\delta_{\mathbf{c}_i, \mathbf{c}_j} = \begin{cases} 1 & \text{if } \mathbf{c}_i = \mathbf{c}_j \\ 0 & \text{otherwise} \end{cases}$.

The modularity was initially conceived as a measure of the strength of a community structure (*Newman & Girvan, 2004*). In the original article, it is defined as the sum over all the communities a, of the proportions of within-community edges (first term of Eq. (2)) minus the expectation of these proportions in random networks with the same vertex degrees (second term of Eq. (2)).

This modularity has been extended to deal with weighted graphs, by considering the sum of the weights instead of the number of edges (*Newman, 2004*).

Multiplex-modularity

In order to extend the modularity of *Newman & Girvan (2004)* to multiplex networks, we measured the strength of a given community structure in all the graphs of a multiplex network $(X^{(g)})_g$ by considering, for all communities a, the sum of the proportions of within-community edges over all the graphs minus the expectation of this sum under the same model as in *Newman & Girvan (2004)*. Since the expectation of the sum of random variables is the sum of their expectations, we obtained the Eq. (3). Reorganizing this sum shows that it is equal to the sum of the individual modularities of the graphs of the multiplex with regard to the same community partition Eq. (5). The *multiplex-modularity* of the multiplex network $(X^{(g)})_g$ is defined as:

$$Q^M((X^{(g)})_g, \mathbf{c}) = \sum_a \left[\sum_g \frac{\sum_{\substack{\{i,j\} \\ i \neq j, \mathbf{c}_i = \mathbf{c}_j = a}} X_{i,j}^{(g)}}{2m^g} - \sum_g \frac{\sum_{\substack{\{i,j\} \\ i \neq j, \mathbf{c}_i = \mathbf{c}_j = a}} k_i^g k_j^g}{(2m^g)^2} \right] \tag{3}$$

$$= \sum_g \frac{1}{2m^g} \sum_{\substack{\{i,j\} \\ i \neq j}} \left(X_{i,j}^{(g)} - \frac{k_i^g k_j^g}{2m^g} \right) \delta_{\mathbf{c}_i, \mathbf{c}_j} \tag{4}$$

$$= \sum_g Q(X^{(g)}, \mathbf{c}) \tag{5}$$

Didier et al. (2015), *PeerJ*, DOI 10.7717/peerj.1525

4/20

where m^g is the total number of edges of the graph $X^{(g)}$ and k_i^g is the degree of the vertex i in the graph $X^{(g)}$.

The multiplex-modularity could easily be adapted to multiplex networks containing weighted edges.

Clustering algorithms for community detection

The network modularity was initially designed to measure the quality of a partition into communities (*Newman & Girvan, 2004*). Several clustering algorithms were subsequently developed to identify the communities by optimizing the modularity (*Blondel et al., 2008*; *Shiokawa, Fujiwara & Onizuka, 2013*). Finding the partition optimizing the modularity is NP-complete (*Brandes et al., 2008*) and thus requires meta-heuristic approaches when dealing with large graphs such as biological networks. We applied the Louvain algorithm (*Blondel et al., 2008*). The Louvain algorithm starts from the community structure that separates all vertices. Next, it tries to move each vertex from its community to another, keeping only the changes increasing the modularity, until no change increases the modularity. It then replaces the vertices by the detected communities and iterates the same operations on the new graph obtained, until stability. This algorithm was applied to the networks aggregated as $X^\cap$, $X^\cup$ and X^Σ. Adapting Louvain to the multiplex networks required replacing the optimization of the modularity by the optimization of the multiplex-modularity defined previously.

Resolution parameter

A general issue with modularity-based clustering approaches is that they may fail to detect communities smaller than a certain size, depending on the graph size. This phenomena is known as the *resolution limit* (*Fortunato & Barthélemy, 2007*). This issue is overcame by adding a *resolution parameter* γ to the modularity formula (*Reichardt & Bornholdt, 2006*). The *parametrized modularity* Q_γ is defined as:

$$Q_\gamma(X, \mathbf{c}) = \frac{1}{2m} \sum_{\substack{\{i,j\} \\ i \neq j}} \left(X_{i,j} - \gamma \frac{k_i k_j}{2m} \right) \delta_{\mathbf{c}_i, \mathbf{c}_j}. \tag{6}$$

The greater the resolution parameter γ, the smaller the communities maximizing the corresponding modularities. The *parametrized multiplex-modularity* Q_γ^M of a multiplex network is defined accordingly.

Consensus clustering

Consensus clustering approaches compute a unique community structure from the community structures obtained on each graph independently. We implemented a consensus clustering approach inspired from *Lancichinetti & Fortunato (2012)*. It first computes the community structures of each individual graph with the standard Louvain algorithm. Next, each community structure is transformed into a new graph in which an edge connects two vertices if they belong to the same community. We obtained the same number of new graphs as the initial number of graphs composing the multiplex network. These new graphs form a new multiplex network from which we infer a single community structure by using the sum-aggregation approach and the Louvain algorithm.

Simulated multiplex networks
Random graph models

Stochastic block models (SBMs, also known as planted partition models) are among the most intuitive models of random graphs with community structures (*Holland, Laskey & Leinhardt, 1983*). In SBMs, the number of communities is set *a priori* and each vertex *i* is assigned to a unique community c_i. The presence of an edge between two vertices *i* and *j* is drawn independently conditionally on the respective community assignments of *i* and *j*, and with a probability that only depends on these assignments. Note that edges are independent conditionally on their communities, although not independent in an absolute sense. For instance, two edges involving vertices of the same community are more likely to be both present than edges from different communities. Hence, the parameters of the SBMs are the number of vertices, the community structure and the matrix π in which the entry π_{ql} is the probability of observing an edge between a vertex of the community *q* and one of the community *l*.

A natural way to generate multiplex networks $(X^{(g)})_g$ with a community structure is to draw independent realizations of SBMs sharing the same community structure. This means that all the graphs are drawn independently and, in each graph, all the edges are also drawn independently, conditionally on the community assignments of the vertices. The parameters of the multiplex SBMs are the same as classical SBMs, except that instead of one matrix of edge probabilities, we have a collection $(\pi^{(g)})_g$, the *g*th matrix corresponding to the *g*th graph of the multiplex network. To ease the interpretation of the simulations, we parametrized edge probability matrices with only two values: the probability p_I of an internal edge and the probability p_E of an external one. All diagonal entries are equal to p_I and all non-diagonal entries are equal to p_E.

Network incompleteness and missing data

We simulated missing data by setting an arbitrary probability for a vertex to be absent in a graph, and using it to randomly remove some vertices. The community detection algorithm was then applied to the incomplete multiplex networks. In practice, whenever a vertex is missing in a graph layer, it is considered as not interacting in this graph.

Biological multiplex networks

Four biological networks were constructed from different sources of interactions between human genes or proteins. A network of physical binary protein–protein interactions was created from the PSICQUIC portal (*Del-Toro et al., 2013*) and the CCSB Interactome database (*Rolland et al., 2014*), and contains 60,669 direct interactions between 12,110 proteins. A co-expression network was constructed from RNA-Seq data downloaded from the Human Protein Atlas (http://www.proteinatlas.org). Spearman correlations of expression were computed between all genes based on FPKM values in 27 tissues and 44 cell lines, and correlation ≥ 0.7 were selected as interactions (1,107,547 interactions between 9,212 genes). A pathway network was constructed from Biocarta (http://www.biocarta.com), Spike (*Paz et al., 2011*), Kegg (*Kanehisa et al., 2008*), PID (*Schaefer et al., 2009*) and Reactome (*Croft et al., 2014*), using the R package *graphite* (*Sales, Calura & Romualdi, 2014*). It contains 166,761 interactions between 8,839 genes. A network

of complexes was constructed from the CORUM database (*Ruepp et al., 2009*) using the matrix model, and contains 36,762 interactions between 2,528 genes. For detailed protocol and options, the markdown is available at GitHub (https://github.com/gilles-didier/MolTi). Data were fetched and networks created in May, 2015. Networks in the Fig. 4 are represented with Cytoscape (*Saito et al., 2012*).

The 4 biological networks correspond to the different layers of the multiplex biological network, sharing the same set of vertices (i.e., the complete set of human genes or proteins, considered here equally). Each layer is incomplete and provides information for a subset of genes/proteins only, the information for the remaining nodes being unknown, and therefore considered as missing data.

Assessment of the detected communities
Adjusted Rand index
The comparison between community partitions is performed by computing the adjusted Rand index (*Santos & Embrechts, 2009*), a widely used measure of the similarity between two partitions. The closer two community partitions, the greater their adjusted Rand index.

Annotation enrichment
The actual biological communities are unknown, but the accuracy of the detected communities can be assessed by verifying their consistency with known biological processes. We performed exact Fisher tests (*Fisher, 1954*) to search for enrichment in Gene Ontology Biological Process (GOBP) terms (*Ashburner et al., 2000*) associated with the genes/proteins of each community. The Bonferroni correction for multiple testing was applied on the Fisher p-values by taking into account both the number of tested ontology terms and communities. Similarly, the communities were screened for enrichment in disease genes extracted from the CTD database (*Davis et al., 2014*) and filtered to select the curated gene-disease associations (marker/mechanisms/therapeutic).

MolTi software
We developed *MolTi*, a standalone user-friendly graphical software. *MolTi* detects communities from multiplex networks provided by the users, by optimizing the multiplex-modularity with the adapted Louvain algorithm. *MolTi* further optionally performs Fisher tests of annotation enrichment for the detected communities, and allows an easy exploration of the results. *MolTi* and corresponding source code are freely available at Github (https://github.com/gilles-didier/MolTi).

RESULTS
The multiplex-modularity recovers accurate communities from simulated multiplex networks
We used stochastic block models (SBMs) to randomly simulate multiplex networks with a reference community structure, and composed of 1–9 graph layers (Methods, Section 'Random graph models'). To reflect some of the characteristics of biological networks in the simulated networks, we generated multiplex networks composed of sparse, dense and

mixed graphs. In addition, the incompleteness of biological networks was simulated by node withdrawal.

We chose the Louvain algorithm that optimizes the network modularity (*Blondel et al., 2008*) to identify communities. The multiplex-modularity is defined as the sum of the individual network layer modularities with regard to the same community partition (Methods, Section 'Multiplex-modularity'). We adapted the Louvain algorithm to optimize this multiplex-modularity (Methods, Section 'Clustering algorithms for community detection'). The Louvain algorithm was directly applied to the multiplex networks aggregated through their union, intersection or sum (Methods, Section 'Clustering algorithms for community detection'). Finally, we also computed a consensus clustering of the communities detected on individual networks (Methods, Section 'Consensus clustering'). The adjusted Rand index, which measures the similarity between two partitions (Methods, Section 'Adjusted Rand index'), was used to compare the communities identified with the different approaches to the reference community structure used to generate the networks.

The intersection and union aggregations, and consensus clustering, fail to recover the community structure

We first observed that the identification of communities from the intersection of the graphs composing the multiplex network eventually leads to detect less accurate community structures (Fig. 1). An exception occurs for the simulation of dense graphs, for which we observed an improvement with the intersection of multiplex networks composed of 2 and 3 graph layers (Fig. 1C-1). In all others cases, the accuracy of the communities detected by the intersection aggregation decreases with the number of graphs composing the multiplex networks. Indeed, as the probability of observing an edge between two given vertices in all the graphs decreases with the number of graphs, the intersection tends to be empty.

A similar behavior is expected for the identification of communities from the union of the graphs composing the multiplex network, as the union aggregated network tends to be complete. This is observed for the simulation of dense and mixed-densities multiplex networks (Figs. 1C-1,2 and 1B-1,2). However, the adjusted Rand index increases with the number of graph layers on sparse simulations (Fig. 1A-1, 2). This unexpected behavior comes from the fact that these simulated networks are so sparse that their unions do not saturate, given the number of graphs considered here. In summary, both the union and intersection aggregations ultimately lead to lose the community structure.

The performance of the consensus clustering increases slowly when considering more than 2 graph layers, showing that a general community structure may eventually emerge from that detected on the individual graphs (Figs. 1A-1, 1B-1 and 1C-1). However, the accuracy of the consensus clustering never outperformed nor got close to the Multiplex-modularity accuracy in our simulations.

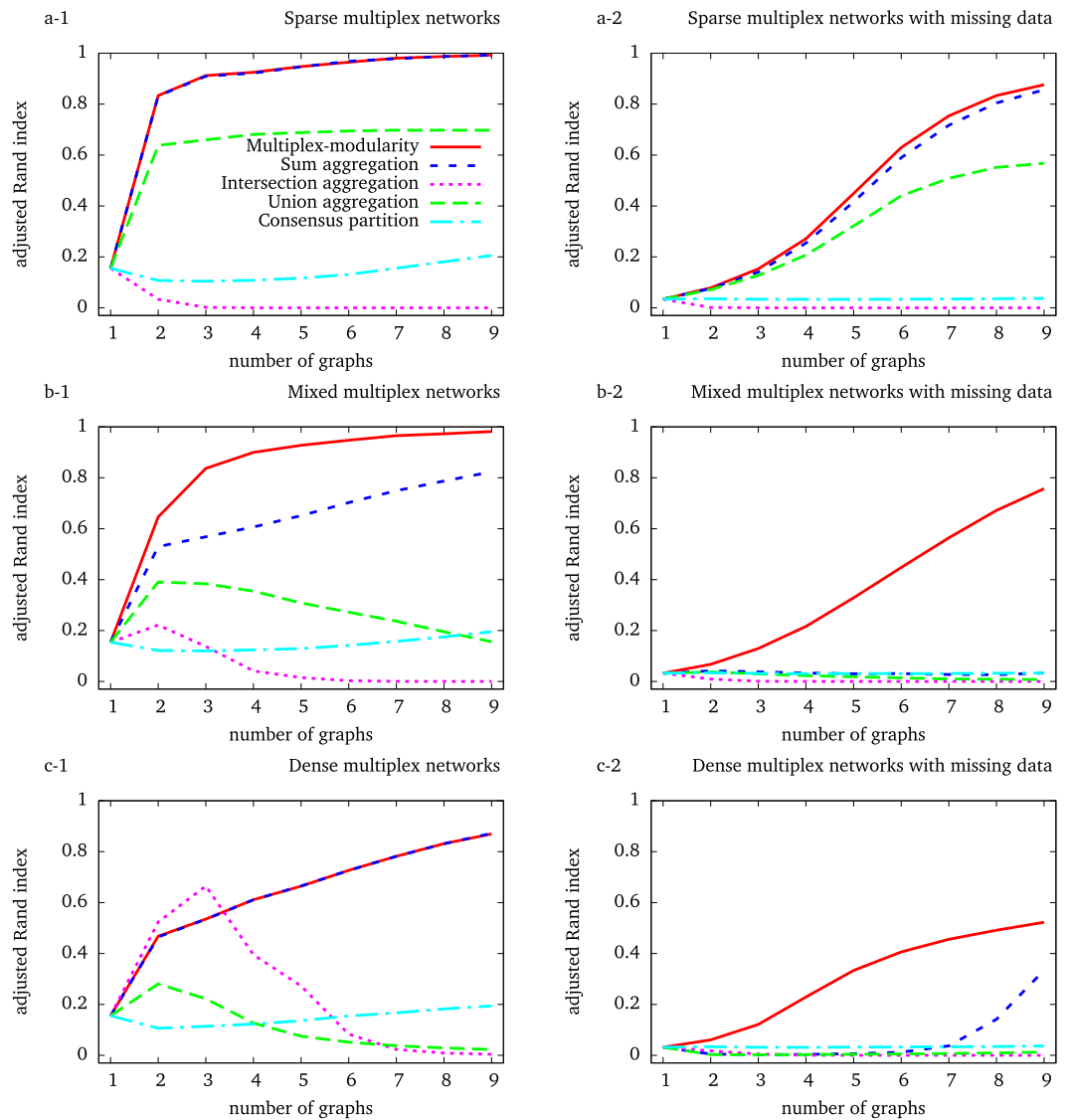Didier et al. (2015), *PeerJ*, DOI 10.7717/peerj.1525

8/20

Figure 1 Adjusted Rand indexes between the reference community structure used to generate the random multiplex networks and the communities detected by the different approaches. Multiplex networks contain from 1 to 9 graph layers. The indexes are averaged over 1,000 random multiplex networks of 1,000 vertices and 20 balanced communities. Sparse (resp. dense) multiplex networks are simulated with 0.1/0.01 (resp. 0.5/0.2) internal/external edge probability matrix. Mixed multiplex networks are simulated by uniformly sampling among these two matrices. Each vertex is withdrawn from each graph with a probability 0.5 to generate missing data.

The multiplex-modularity better recovers communities in heterogeneous density and missing data contexts

In our simulations, the accuracy of the communities detected by the sum aggregation and the multiplex-modularity approach is consistently better than the accuracy of the union and intersection aggregations, or the consensus clustering. Importantly, the accuracy of the communities detected by the sum aggregation and the multiplex-modularity approach increases with the number of graph layers considered. Although the community structure

seems undetectable from a single graph, it is almost perfectly inferred by the multiplex-modularity approach when a sufficient number of graphs is reached. For instance, communities are recovered with Rand Indexes over 0.9 with 3 graphs for sparse multiplex networks (Fig. 1A-1), 4 graphs for mixed multiplex networks (Fig. 1B-1) or 9 graphs for dense graphs and sparse multiplex networks with missing data (Figs. 1C-1 and 1A-2).

The sum aggregation and multiplex-modularity approaches show very similar performances on sparse and dense homogeneous multiplex networks (Figs. 1A and 1C). However, these two approaches do differ for mixed-densities multiplex networks and when missing data are simulated to account for network incompleteness (Fig. 1, Methods, Section 'Network incompleteness and missing data'). In these cases, the accuracy of the multiplex-modularity approach is better than the sum aggregation. As expected, in a missing data context, the community structures are more difficult to detect. All the approaches based on the aggregation of layers composing a multiplex network are sensitive to missing data. Nonetheless, the multiplex-modularity approach still identifies adequate communities when the number of graphs is sufficiently large, and is particularly more accurate than the sum aggregation approach. Additionally, we compared our multiplex-modularity approach to GenLouvain, a general approach to detect communities from multiplex networks incorporating inter-slice connections (Mucha et al., 2010). In most simulated cases, GenLouvain has a behavior similar to the the sum aggregation approach (Fig. S1).

Multiplex-modularity identifies relevant modules from a biological multiplex network

The identification of communities from 4 biological networks and the issue of the resolution parameter

We constructed 4 biological networks from physical and functional interaction data: a protein–protein interaction network, a network of pathways, a network of protein complex, and a mRNA co-expression network (Methods, Section 'Biological multiplex networks').

These networks are sparse, heterogeneous in densities and incomplete. Their densities range from 0.001 to 0.023, with an average of 0.01. They are also prone to missing data as their sizes range from 2,528 to 12,110 nodes (Table S1). Missing data is indeed an important feature in biological networks since insufficient depth and coverage of the interaction space leads to incomplete interaction datasets (Venkatesan et al., 2009; Braun et al., 2009).

Altogether, the 4 biological networks contain 17,003 nodes and 1,371,739 edges, among which 1,338,086 appear in 1 network only (Table S1). The Louvain algorithm adapted to the multiplex-modularity was applied to identify communities from this biological multiplex network. The classical Louvain algorithm was applied to the 4 networks individually, and to their union, and sum aggregations. Considering their weak performance in the simulations, the intersection-aggregation and consensus clustering approaches were not taken into account in the following analyses.

Didier et al. (2015), *PeerJ*, DOI 10.7717/peerj.1525

10/20

As some of the biological networks considered here are very large, we used the resolution parameter γ to reduce the sizes of the detected communities (Methods, Section 'Resolution parameter'). We tested the clustering of the 4 individual networks, sum and union aggregations, and multiplex-modularity approach with γ parameter values ranging from 1 to 15. We next elected the parameter value $\gamma = 5$ that gives module average sizes lower than 50 proteins for the sum, union and multiplex partitions (Fig. S3). The number of modules obtained from each individual network ranges from 85 communities for the Pathway partition to 1,320 communities for the Co-expression network partition (Fig. S2). Using this value of resolution parameter, more than 800 communities are obtained from the sum and union aggregation networks, and 350 from the multiplex-modularity approach. The communities obtained with the multiplex-modularity approach display a median size of 17 proteins per community as opposed to a median sizes of 2 proteins per community for the sum or union aggregations (Fig. S4). The multiplex-modularity communities are hence more balanced and easier to interpret biologically.

The multiplex-modularity clustering identifies communities closer to the individual partitions

We computed the adjusted Rand Indexes between the community partitions obtained from the 4 individual networks and from their sum and union aggregations, and the multiplex-modularity approach (Methods, Section 'Adjusted Rand Index'). We observed that, on average, the multiplex-modularity partition is closer to the individual network partitions than the aggregations (Fig. 2).

The multiplex-modularity identifies communities enriched in gene ontology terms

To assess the consistency of the communities, we tested their enrichment in Gene Ontology Biological Process (GOBP) terms (*Ashburner et al., 2000*, Methods, Section 'Annotation enrichment'). The communities detected by the multiplex-modularity approach are more often associated to at least one significant (q-value < 0.05) GOBP term than the sum or union aggregation communities (Fig. 3). For $\gamma = 5$, for instance, 38% of the multiplex-modularity communities are annotated, compared to 14% and 10% of the sum and union communities. Overall, these results suggest that the multiplex-modularity approach is extracting relevant biological communities.

Clustering of proteins involved in the Coffin-Siris syndrome and other syndromes with overlapping clinical features

As an illustration, we screened the modules obtained with the multiplex-modularity approach ($\gamma = 5$) for enrichment in disease-related proteins (Methods, Section 'Annotation enrichment'). Twenty-nine communities are significantly associated to at least one disease. Among those, one module contains the 6 gene products known to be associated to the Coffins-Siris syndrome (q-value $= 5 \times 10^{-10}$, Fig. 4, Methods, Section 'Annotation enrichment'). This syndrome is a rare congenital disorder characterized by specific cranofacial and digital features, microcephaly and intellectual disability, among other phenotypes (*Kosho & Okamoto, 2014*; *Vergano & Deardorff, 2014*).
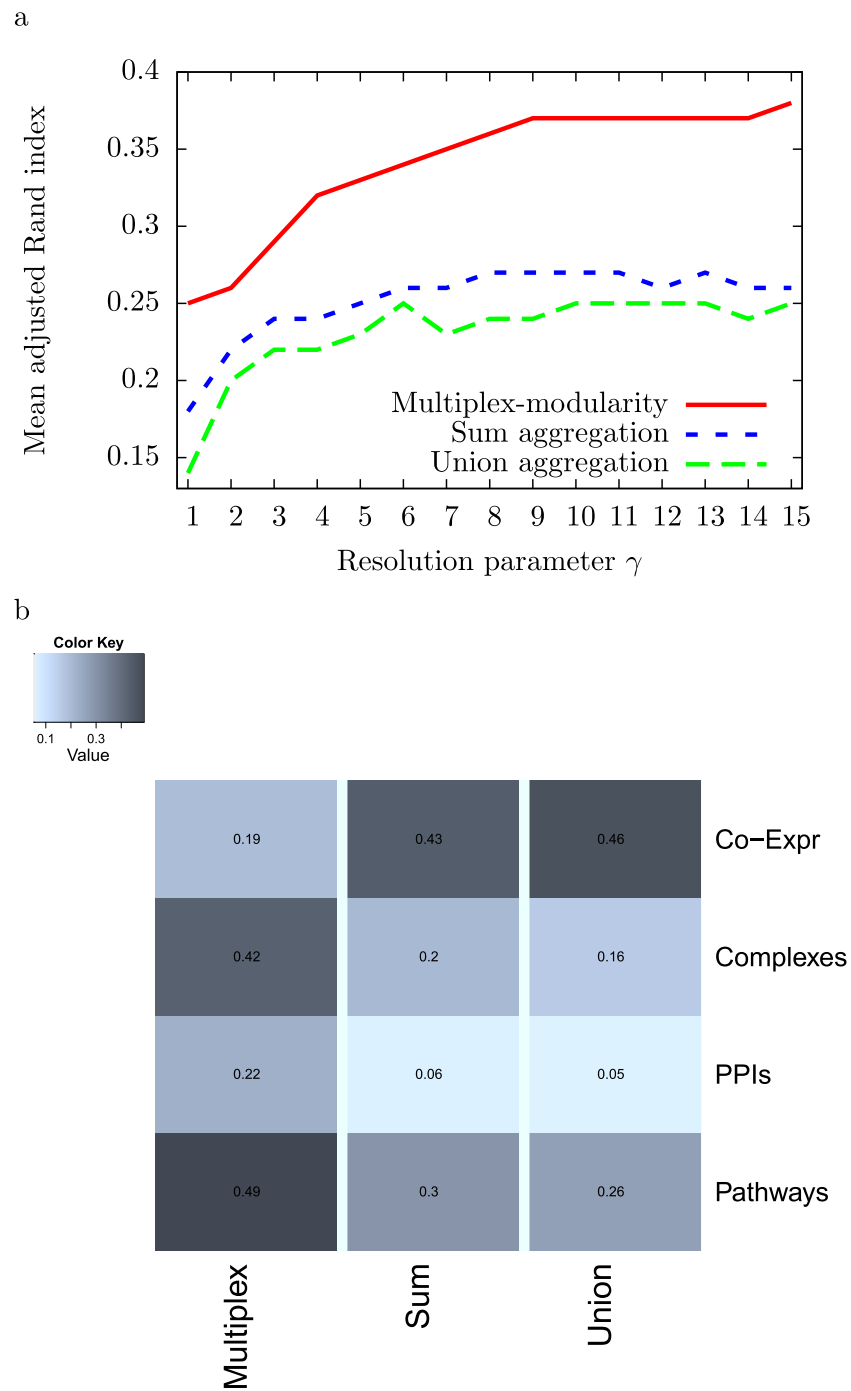
Didier et al. (2015), *PeerJ*, DOI 10.7717/peerj.1525

11/20

a

b

Figure 2 **Adjusted Rand Indexes between the partitions of the 4 individual networks and their sum-aggregation, union-aggregation, and the multiplex-modularity approach.** (A) Mean adjusted rand indexes for $\gamma = 1$ to $\gamma = 15$. (B) detailed adjusted rand indexes for $\gamma = 5$.
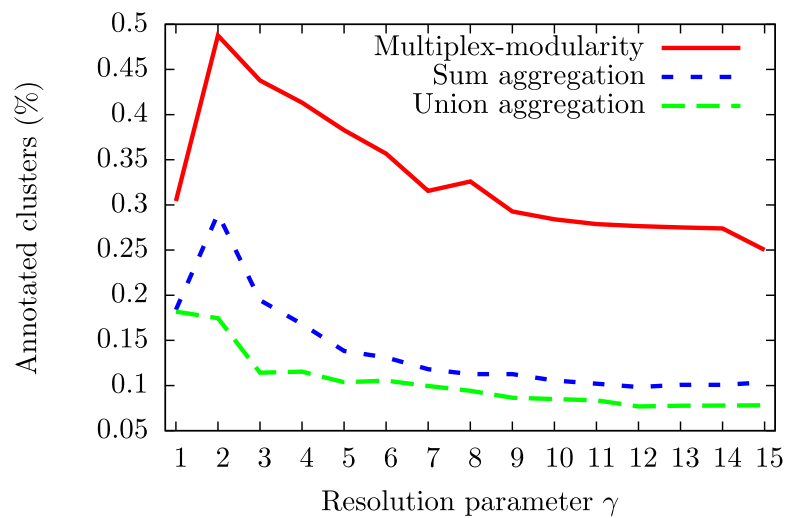
Figure 3 Percentage of communities associated to at least one significant GOBP term, for $\gamma = 1$ to $\gamma = 15$.
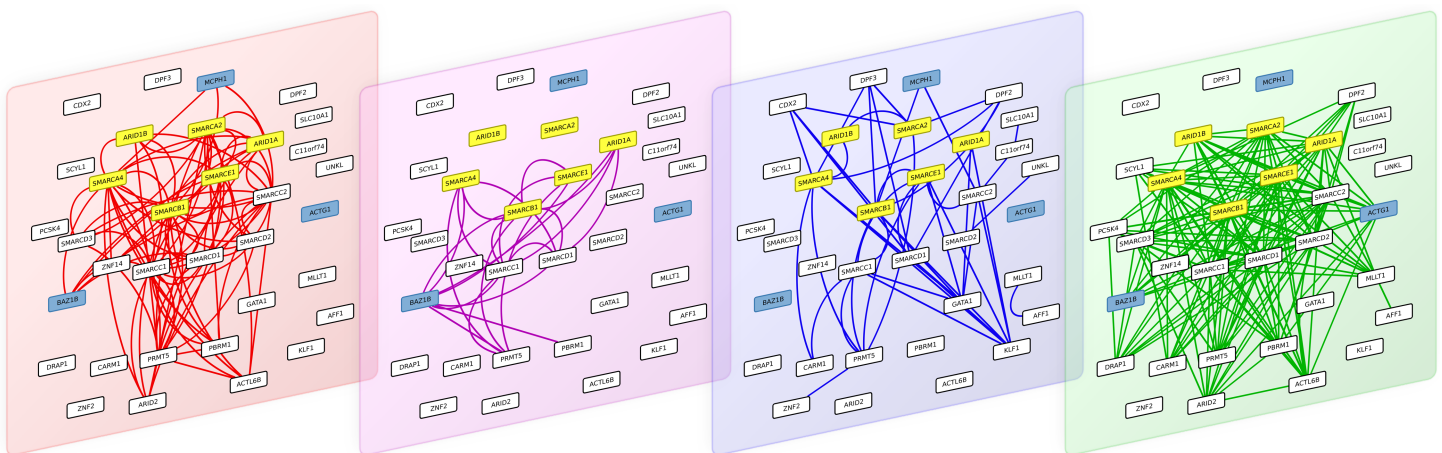

Figure 4 The 4 interaction layers of a module obtained after partitioning the multiplex biological networks with the multiplex-modularity approach. From left to right: pathways, Co-expression, PPIs and Complexes networks. Proteins involved in the Coffin-Siris syndrome are highlighted in yellow, and protein related to other syndrome with shared clinical features are highlighted in blue.

The module is enriched in proteins involved in transcription regulation, and participating in different cellular processes such as chromatin remodeling, transcription from RNA-polII promoter and DNA repair. The 6 Coffin-Siris syndrome proteins are components of the SWI/SNF chromatin remodeling complex (*Tsurusaki et al., 2012*). Interestingly, the module also contains other gene products implicated in other syndromes, which clinical features overlap with the Coffin-Siris syndrome. MCPH1 is involved in primary autosomal recessive microcephaly (*Venkatesh & Suresh, 2014*), BAZ1B is involved in the William-Beuren syndrome that also presents neuropsychological deficits (*Fusco et al., 2014*; *Xiao et al., 2009*) and ACTG1 mutations are involved in the Baraitser-Winter syndrome, another developmental disorder associated to intellectual disability

(*Rivière et al., 2012*). It is to note that all these syndrome-related proteins are clustered only when the multiplex-approach is used. Indeed, when the different network layers are aggregated, only 4 out of the 6 Coffin-Siris syndrome proteins belong to the same module.

DISCUSSION

Interactions between genes and proteins can be of various nature and are identified by diverse experimental approaches. We focused here on different facets of protein cellular functions by taking into account (i) physical interactions between proteins, extracted from experiments such as yeast 2-hybrid screens, (ii) interactions derived from the belonging of proteins to complexes, (iii) functional interactions extracted from pathway databases and (iv) co-expression correlations derived from mRNA expression in different cellular contexts. However, other types of functional relationships between genes or proteins, such as their implication in the same disease (*Goh et al., 2007*) or their targeting by the same drug (*Yildirim et al., 2007*) could be considered. Consequently, many networks can depict gene/protein functional relationships. Each of them represents a different layer of a multiplex biological network, and each layer has its own meaning, size, relevance and bias. For instance, the network of pathways is derived from information provided by several pathway databases containing heavily curated expert knowledge rather than large-scale datasets prone to noise. Conversely, the network of correlations computed from RNA-seq expression data contains a lot of links, some of which are undoubtedly noise. Furthermore, the different experimental approaches sourcing the interaction data capture different parts of the interaction space, therefore leading to incomplete networks. It is estimated for instance, that only 20% of the human protein–protein interaction space has been currently deciphered (*Menche et al., 2015*). Similarly, curated interactions in pathway databases are limited to the proteins for which such information exists, letting aside a potential wealth of data yet to be deciphered. In this context, taking into account diverse network sources when deriving functional modules is expected to cover a more comprehensive picture of protein cellular functioning. The underlying reason is to combine informative-but-sparse and plenty-but-noisy information to strengthen each other, and improve the clustering performance (*Papalexakis, Akoglu & Ience, 2013*).

We explored here different approaches based on modularity to identify communities from a set of networks. These approaches are all based on the assumption that the layers of a multiplex network have the same community structure. For biological networks, this means that the different categories of gene/protein functional relationships are considered as instances or realizations of the real underlying functional modules. In our work, this assumption is also fulfilled by the simulated random multiplex networks, as they are generated from the same reference community structure. The aggregation of the network layers into a monoplex network, or the consensus of communities are classically applied to biological networks. The use of a natural extension of the modularity metric to multiplex networks is newer (*Bennett et al., 2015*, this work).

While performing these analyses, *Bennett et al. (2015)* published an extension of the modularity to multiplex networks equivalent to our multiplex-modularity. The originality

of our work lies in the extensive comparisons we performed with the aggregation and consensus approaches. In particular, we used random simulated networks to precisely study the effects of different network topological features. This is particularly important for the identification of communities from multiplex biological networks, as their density and incompleteness can vary. Overall, when simulating networks with heterogeneous densities and missing data, the multiplex-modularity approach outperforms all its aggregated counterparts. This better performance is also observed in a real-case multiplex biological network, for which the multiplex-modularity approach identifies more balanced and annotated communities.

Other clustering approaches on multiplex networks were not considered because they cannot readily be applied to detect communities from large-scale multiplex biological networks with missing data (*Shiga & Mamitsuka, 2012*; *Papalexakis, Akoglu & Ience, 2013*). The GenLouvain approach (*Mucha et al., 2010*), a generalization of the modularity to any kind of multi-slice networks, requires parameters difficult to set for biological networks, such as inter-layer coupling (*Mucha et al., 2010*). Hence, we applied GenLouvain only to the simulated multiplex networks (Fig. S1).

The large-scale interaction networks available nowadays constitute both a fantastic source of information to study proteins functioning in their cellular context, and a huge challenge calling for the development of methods to mine knowledge from such complex data. We established here that the multiplex framework is suited to combine interaction networks of different nature and allows recovering relevant functional modules.

ACKNOWLEDGEMENTS

ADDITIONAL INFORMATION AND DECLARATIONS

Competing Interests

Anaïs Baudot is an Academic Editor for PeerJ. The authors declare there are no competing interests.

Didier et al. (2015), *PeerJ*, DOI 10.7717/peerj.1525

15/20

Author Contributions

- Gilles Didier and Anaïs Baudot conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.
- Christine Brun conceived and designed the experiments, analyzed the data, wrote the paper, reviewed drafts of the paper.

Data Availability

Supplemental Information

REFERENCES

Ahn Y-Y, Bagrow JP, Lehmann S. 2010. Link communities reveal multiscale complexity in networks. *Nature* **466**(**7307**):761–764 DOI 10.1038/nature09182.

Aittokallio T, Schwikowski B. 2006. Graph-based methods for analysing networks in cell biology. *Briefings in Bioinformatics* **7**(**3**):243–255 DOI 10.1093/bib/bbl022.

Arroyo R, Suñé G, Zanzoni A, Duran-Frigola M, Alcalde V, Stracker TH, Soler-López M, Aloy P. 2015. Systematic identification of molecular links between core and candidate genes in breast cancer. *Journal of Molecular Biology* **427**(**6**):1436–1450 DOI 10.1016/j.jmb.2015.01.014.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. 2000. Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nature Genetics* **25**(**1**):25–29 DOI 10.1038/75556.

Battiston F, Nicosia V, Latora V. 2014. Structural measures for multiplex networks. *Physical Review E* **89**:032804 DOI 10.1103/PhysRevE.89.032804.

Bennett L, Kittas A, Muirhead G, Papageorgiou LG, Tsoka S. 2015. Detection of composite communities in multiplex biological networks. *Scientific Reports* **5**:10345 DOI 10.1038/srep10345.

Blondel V, Guillaume J, Lambiotte R, Mech E. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* **10**:P10008.

Brandes U, Delling D, Gaertler M, Goerke R, Hoefer M, Nikoloski Z, Wagner D. 2008. On modularity clustering. *IEEE Transactions on Knowledge and Data Engineering* **20**(**2**):172–188 DOI 10.1109/TKDE.2007.190689.

Braun P, Tasan M, Dreze M, Barrios-rodiles M, Lemmens I, Yu H, Sahalie JM, Murray RR, Roncari L, Smet ASD, Venkatesan K, Rual J-F, Cusick ME, Pawson T, Hill DE,

Tavernier J, Wrana JL, Roth FP, Vidal M. 2009. An experimentally derived confidence score for binary protein–protein interactions. *Nature Methods* **6**(1):91–97 DOI 10.1038/nmeth.1281.

Brehme M, Voisine C, Rolland T, Wachi S, Soper J, Zhu Y, Orton K, Villella A, Garza D, Vidal M, Ge H, Morimoto R. 2014. A chaperome subnetwork safeguards proteostasis in aging and neurodegenerative disease. *Cell Reports* **9**(3):1135–1150 DOI 10.1016/j.celrep.2014.09.042.

Brohée S, Van Helden J. 2006. Evaluation of clustering algorithms for protein–protein interaction networks. *BMC Bioinformatics* **7**:488 DOI 10.1186/1471-2105-7-488.

Chapple CE, Robisson B, Spinelli L, Guien C, Becker E, Brun C. 2015. Extreme multifunctional proteins identified from a human protein interaction network. *Nature Communications* **6**(May):7412 DOI 10.1038/ncomms8412.

Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, Caudy M, Garapati P, Gillespie M, Kamdar MR, Jassal B, Jupe S, Matthews L, May B, Palatnik S, Rothfels K, Shamovsky V, Song H, Williams M, Birney E, Hermjakob H, Stein L, D'Eustachio P. 2014. The Reactome pathway knowledgebase. *Nucleic Acids Research* **42**(**Database issue**):D472–D477 DOI 10.1093/nar/gkt1102.

Davis AP, Grondin CJ, Lennon-Hopkins K, Saraceni-Richards C, Sciaky D, King BL, Wiegers TC, Mattingly CJ. 2014. The comparative toxicogenomics database's 10th year anniversary: update 2015. *Nucleic Acids Research* **43**(**October 2014**):D914–D920.

Del-Toro N, Dumousseau M, Orchard S, Jimenez RC, Galeota E, Launay G, Goll J, Breuer K, Ono K, Salwinski L, Hermjakob H. 2013. A new reference implementation of the PSICQUIC web service. *Nucleic Acids Research* **41**(**Web Server issue**):W601–W606 DOI 10.1093/nar/gkt392.

Fisher R. 1954. Statistical methods for research workers. In: *Number 5 in biological monographs and manuals.* 12th edition. Edinburgh: Oliver and Boyd rev. edition.

Fortunato S, Barthélemy M. 2007. Resolution limit in community detection. *Proceedings of the National Academy of Sciences of the United States of America* **104**(1):36–41 DOI 10.1073/pnas.0605965104.

Fusco C, Micale L, Augello B, Teresa Pellico M, Menghini D, Alfieri P, Cristina Digilio M, Mandriani B, Carella M, Palumbo O, Vicari S, Merla G. 2014. Smaller and larger deletions of the Williams Beuren syndrome region implicate genes involved in mild facial phenotype, epilepsy and autistic traits. *European Journal of Human Genetics* **22**(1):64–70 DOI 10.1038/ejhg.2013.101.

Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabási A-L. 2007. The human disease network. *Proceedings of the National Academy of Sciences of the United States of America* **104**(21):8685–8690 DOI 10.1073/pnas.0701361104.

Hartwell LH, Hopfield JJ, Leibler S, Murray AW. 1999. From molecular to modular cell biology. *Nature* **402**(6761 Suppl):C47–C52 DOI 10.1038/35011540.

Holland PW, Laskey KB, Leinhardt S. 1983. Stochastic blockmodels: first steps. *Social Networks* **5**(2):109–137 DOI 10.1016/0378-8733(83)90021-7.

Huttlin EL, Ting L, Bruckner RJ, Gebreab F, Gygi MP, Szpyt J, Tam S, Zarraga G, Colby G, Baltier K, Dong R, Guarani V, Vaites LP, Ordureau A, Rad R, Erickson

Didier et al. (2015), *PeerJ*, DOI 10.7717/peerj.1525

17/20

BK, Wühr M, Chick J, Zhai B, Kolippakkam D, Mintseris J, Obar RA, Harris T, Artavanis-Tsakonas S, Sowa ME, De Camilli P, Paulo JA, Harper JW, Gygi SP. 2015. The BioPlex network: a systematic exploration of the human interactome. *Cell* **162**(2):425–440 DOI 10.1016/j.cell.2015.06.043.

Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, Yamanishi Y. 2008. KEGG for linking genomes to life and the environment. *Nucleic Acids Research* **36**(1):480–484.

Katsogiannou M, Andrieu C, Baylot V, Baudot A, Dusetti NJ, Gayet O, Finetti P, Garrido C, Birnbaum D, Bertucci F, Brun C, Rocchi P. 2014. The functional landscape of Hsp27 reveals new cellular processes such as DNA repair and alternative splicing and proposes novel anticancer targets. *Molecular & Cellular Proteomics* **13**(12):3585–3601 DOI 10.1074/mcp.M114.041228.

Kivelä M, Arenas A, Barthelemy M, Gleeson JP, Moreno Y, Porter MA. 2014. Multilayer networks. *Journal of Complex Networks* **2**:203–271 DOI 10.1093/comnet/cnu016.

Kosho T, Okamoto N. 2014. Genotype-phenotype correlation of Coffin-Siris syndrome caused by mutations in SMARCB1, SMARCA4, SMARCE1, and ARID1A. *American Journal of Medical Genetics Part C: Seminars in Medical Genetics* **166**(August):262–275 DOI 10.1002/ajmg.c.31407.

Lancichinetti A, Fortunato S. 2012. Consensus clustering in complex networks. *Scientific Reports* **2**:336 DOI 10.1038/srep00336.

Menche J, Sharma A, Kitsak M, Ghiassian SD, Vidal M, Loscalzo J, Barabási A-L. 2015. Uncovering disease–disease relationships through the incomplete interactome. *Science* **347**:6224 DOI 10.1126/science.1257601.

Mucha PJ, Richardson T, Macon K, Porter MA, Onnela J-P. 2010. Community structure in time-dependent, multiscale, and multiplex networks. *Science* **328**(5980):876–878 DOI 10.1126/science.1184819.

Newman MEJ. 2004. Analysis of weighted networks. *Physical Review E* **70**(5):056131 DOI 10.1103/PhysRevE.70.056131.

Newman MEJ, Girvan M. 2004. Finding and evaluating community structure in networks. *Physical Review E* **69**(2):026113 DOI 10.1103/PhysRevE.69.026113.

Papalexakis EE, Akoglu L, Ience D. 2013. Do more views of a graph help? Community detection and clustering in multi-graphs. In: *Information fusion (FUSION), 2013 16th international conference on, 9-12 July 2013*, 899–905.

Paz A, Brownstein Z, Ber Y, Bialik S, David E, Sagir D, Ulitsky I, Elkon R, Kimchi A, Avraham KB, Shiloh Y, Shamir R. 2011. SPIKE: a database of highly curated human signaling pathways. *Nucleic Acids Research* **39**(Database issue):D793–D799 DOI 10.1093/nar/gkq1167.

Pizzuti C, Rombo SE. 2014. Algorithms and tools for protein–protein interaction networks clustering, with a special focus on population-based stochastic methods. *Bioinformatics* **30**(10):1343–1352 DOI 10.1093/bioinformatics/btu034.

Reichardt J, Bornholdt S. 2006. Statistical mechanics of community detection. *Physical Review E* **74**:016110 DOI 10.1103/PhysRevE.74.016110.

Rivière J-B, Bon V, Bregje WM, Hoischen A, Kholmanskikh SS, O'Roak BJ, Gilissen C, Gijsen S, Sullivan CT, Christian SL, Abdul-Rahman OA, Atkin JF, Chassaing N, Drouin-Garraud V, Fry AE, Fryns J-P, Gripp KW, Kempers M, Kleefstra T, Mancini GMS, Nowaczyk MJM, Ravenswaaij-Arts V, Conny MA, Roscioli T, Marble M, Rosenfeld JA, Siu VM, Vries D, Bert BA, Shendure J, Verloes A, Veltman JA, Brunner HG, Ross ME, Pilz DT, Dobyns WB. 2012. De novo mutations in the actin genes ACTB and ACTG1 cause Baraitser-Winter syndrome [Letter]. *Nature Publishing Group* **44**(**4**):440–444 DOI 10.1038/ng.1091.

Rolland T, Taşan M, Charloteaux B, Pevzner S, Zhong Q, Sahni N, Yi S, Lemmens I, Fontanillo C, Mosca R, Kamburov A, Ghiassian S, Yang X, Ghamsari L, Balcha D, Begg B, Braun P, Brehme M, Broly M, Carvunis A-R, Convery-Zupan D, Corominas R, Coulombe-Huntington J, Dann E, Dreze M, Dricot A, Fan C, Franzosa E, Gebreab F, Gutierrez B, Hardy M, Jin M, Kang S, Kiros R, Lin G, Luck K, MacWilliams A, Menche J, Murray R, Palagi A, Poulin M, Rambout X, Rasla J, Reichert P, Romero V, Ruyssinck E, Sahalie J, Scholz A, Shah A, Sharma A, Shen Y, Spirohn K, Tam S, Tejeda A, Trigg S, Twizere J-C, Vega K, Walsh J, Cusick M, Xia Y, Barabási A-L, Iakoucheva L, Aloy P, DeLasRivas J, Tavernier J, Calderwood M, Hill D, Hao T, Roth F, Vidal M. 2014. A proteome-scale map of the human interactome network. *Cell* **159**(**5**):1212–1226 DOI 10.1016/j.cell.2014.10.050.

Ruepp A, Waegele B, Lechner M, Brauner B, Dunger-Kaltenbach I, Fobo G, Frishman G, Montrone C, Mewes HW. 2009. CORUM: the comprehensive resource of mammalian protein complexes-2009. *Nucleic Acids Research* **38**(**SUPPL.1**):497–501 DOI 10.1093/nar/gkp914.

Saito R, Smoot ME, Ono K, Ruscheinski J, Wang P-L, Lotia S, Pico AR, Bader GD, Ideker T. 2012. A travel guide to Cytoscape plugins. *Nature Methods* **9**(**11**):1069–1076 DOI 10.1038/nmeth.2212.

Sales G, Calura E, Romualdi C. 2014. *graphite: GRAPH interaction from pathway topological environment*. R package version 1.12.0.

Santos J, Embrechts M. 2009. On the use of the adjusted rand index as a metric for evaluating supervised classification. In: Alippi C, Polycarpou M, Panayiotou C, Ellinas G, eds. *Artificial neural networks—ICANN 2009*. *Lecture notes in computer science*, vol. 5769. Berlin Heidelberg: Springer, 175–184.

Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, Buetow KH. 2009. PID: the pathway interaction database. *Nucleic Acids Research* **37**(**Database issue**):D674–D679 DOI 10.1093/nar/gkn653.

Senbabaoglu Y, Michailidis G, Li JZ. 2014. Critical limitations of consensus clustering in class discovery. *Scientific Reports* **4**:6207 DOI 10.1038/srep06207.

Sharan R, Ideker T. 2006. Modeling cellular machinery through biological network comparison. *Nature Biotechnology* **24**(**4**):427–433 DOI 10.1038/nbt1196.

Shiga M, Mamitsuka H. 2012. A variational bayesian framework for clustering with multiple graphs. *IEEE Transactions on Knowledge and Data Engineering* **24**(**4**):577–590 DOI 10.1109/TKDE.2010.272.

Shiokawa H, Fujiwara Y, Onizuka M. 2013. DesJardins M, Littman ML, eds. *Fast algorithm for modularity-based graph clustering*. Palo Alto: AAAI Press.

Tsurusaki Y, Okamoto N, Ohashi H, Kosho T, Imai Y, Hibi-Ko Y, Kaname T, Naritomi K, Kawame H, Wakui K, Fukushima Y, Homma T, Kato M, Hiraki Y, Yamagata T, Yano S, Mizuno S, Sakazume S, Ishii T, Nagai T, Shiina M, Ogata K, Ohta T, Niikawa N, Miyatake S, Okada I, Mizuguchi T, Doi H, Saitsu H, Miyake N, Matsumoto N. 2012. Mutations affecting components of the SWI/SNF complex cause Coffin-Siris syndrome. *Nature Genetics* **44**(4):376–378 DOI 10.1038/ng.2219.

Venkatesan K, Rual J-f, Vazquez A, Lemmens I, Hirozane-kishikawa T, Hao T, Xin X, Goh KI, Yildirim MA, Heinzmann K, Gebreab F, Sahalie JM, Simon C, Smet A-SD, Dann E, Vinayagam A, Yu H, Szeto D, Dricot A, Klitgord N, Murray RR, Lin C, Lalowski M, Timm J, Rau K, Boone C, Braun P, Michael E, Roth FP, Hill DE, Tavernier J, Wanker EE, Barabási L, Vidal M. 2009. An empirical framework for binary interactome mapping. *Nature Methods* **6**(6):83–90 DOI 10.1038/nmeth.1280.

Venkatesh T, Suresh PS. 2014. Emerging roles of MCPH1: expedition from primary microcephaly to cancer. *European Journal of Cell Biology* **93**(3):98–105 DOI 10.1016/j.ejcb.2014.01.005.

Vergano SS, Deardorff MA. 2014. Clinical features, diagnostic criteria, and management of Coffin-Siris syndrome. *American Journal of Medical Genetics Part C: Seminars in Medical Genetics* **166**(August):252–256 DOI 10.1002/ajmg.c.31411.

Wan C, Borgeson B, Phanse S, Tu F, Drew K, Clark G, Xiong X, Kagan O, Kwan J, Bezginov A, Chessman K, Pal S, Cromar G, Papoulas O, Ni Z, Boutz DR, Stoilova S, Havugimana PC, Guo X, Malty RH, Sarov M, Greenblatt J, Babu M, Derry WB, Tillier ER, Wallingford JB, Parkinson J, Marcotte EM, Emili A. 2015. Panorama of ancient metazoan macromolecular complexes. *Nature* **525**:339–344 DOI 10.1038/nature14877.

Wang J, Li M, Deng Y, Pan Y. 2010. Recent advances in clustering methods for protein interaction networks. *BMC Genomics* **11**(Suppl 3):S10 DOI 10.1186/1471-2164-11-S3-S10.

Xiao A, Li H, Shechter D, Ahn SH, Fabrizio LA, Erdjument-Bromage H, Ishibe-Murakami S, Wang B, Tempst P, Hofmann K, Patel DJ, Elledge SJ, Allis CD. 2009. WSTF regulates the H2A.X DNA damage response via a novel tyrosine kinase activity. *Nature* **457**(7225):57–62 DOI 10.1038/nature07668.

Yildirim MA, Goh K-I, Cusick ME, Barabási A-L, Vidal M. 2007. Drug-target network. *Nature Biotechnology* **25**(10):1119–1126 DOI 10.1038/nbt1338.