A strategy to incorporate prior knowledge into correlation network cutoff selection

Elisa Benedetti[1,12], Maja Pučić-Baković[2], Toma Keser[3], Nathalie Gerstner[1], Mustafa Büyüközkan[1,12], Tamara Štambuk[3], Maurice H.J. Selman[4], Igor Rudan[5], Ozren Polašek [6,7], Caroline Hayward[8], Hassen Al-Amin[9], Karsten Suhre[9,10], Gabi Kastenmüller[1,11], Gordan Lauc[2,3], Jan Krumsiek[1,12*]

[1] Institute of Computational Biology, Helmholtz Zentrum München - German Research Center for Environmental Health, 85764 Neuherberg, Germany

[2] Genos Glycoscience Research Laboratory, 10000 Zagreb, Croatia

[3] Faculty of Pharmacy and Biochemistry, University of Zagreb, 10000 Zagreb, Croatia

[4] Leiden University Medical Center, 2333 ZA Leiden, The Netherlands

[5] Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, EH8 9AG Edinburgh, UK

[6] University of Split School of Medicine, 21000 Split, Croatia

[7] Gen-info Ltd., 10000 Zagreb, Croatia

[8] Medical Research Council Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, EH4 2XU Edinburgh, UK

[9] Department of Psychiatry, Weill Cornell Medicine in Qatar, Education City, Doha, Qatar

[10] Department of Physiology and Biophysics, Weill Cornell Medical College in Qatar, Education City, Doha, Qatar

[11] Institute of Bioinformatics and Systems Biology, Helmholtz-Zentrum München, 85764 Neuherberg, Germany

[12] Institute for Computational Biomedicine, Englander Institute for Precision Medicine, Department of Physiology and Biophysics, Weill Cornell Medicine, New York, USA

* Corresponding Author

Abstract

Correlation networks are commonly used to statistically extract biological interactions between omics markers. Network edge selection is typically based on the significance of the underlying correlation coefficients. A statistical cutoff, however, is not guaranteed to capture biological reality, and heavily depends on dataset properties such as sample size. We here propose an alternative, innovative approach to address the problem of network reconstruction. Specifically, we developed a cutoff selection algorithm that maximizes the agreement to a given ground truth. We first evaluate the approach on IgG glycomics data, for which the biochemical pathway is known and well-characterized. The optimal network outperforms networks obtained with statistical cutoffs and is robust with respect to sample size. Importantly, we can show that even in the case of incomplete or incorrect prior knowledge, the optimal network is close to the true optimum. We then demonstrate the generalizability of the approach on an untargeted metabolomics and a transcriptomics dataset from The Cancer Genome Atlas (TCGA). For the transcriptomics case, we demonstrate that the optimized network is superior to statistical networks in systematically retrieving interactions that were not included in the biological reference used for the optimization. Overall, this paper shows that using prior information for correlation network inference is superior to using regular statistical cutoffs, even if the prior information is incomplete or partially inaccurate.

Keywords

Correlation cutoff / Correlation Networks / Gaussian Graphical Models / Network inference / Prior knowledge

Introduction

Network inference, i.e. the reconstruction of biological networks from high-throughput data, has become a popular field in systems biology[1–3]. Interactions among biomolecules extracted from the analysis of large datasets can represent known and predict novel biological mechanisms[4,5], in particular enzymatic reactions in molecular pathways[6–8].

Virtually all network inference methodologies require the definition of a parameter that determines which molecular interactions should be included in the network and which should be discarded. The construction of correlation-based networks commonly requires a series of simple steps (Figure 1A). First, pairwise correlations between variables are estimated from the data, for which a wide variety of methods is available. The next step is to determine which correlation coefficients are statistically different from zero using a hypothesis test, which produces p-values associated with each correlation coefficient. These p-values are then compared to a given significance level α, typically 0.01 or 0.05, with appropriate multiple hypothesis testing correction. Finally, significant correlations can be visualized and further analyzed as a network, where nodes represent the variables in the dataset and edges represent significant correlations.

However, this straightforward network inference pipeline has two major pitfalls that are usually overlooked and substantially affect the robustness and reproducibility of correlation-based network inference. First, for most correlation measures, the resulting network will vary substantially depending on the number of observations available in the dataset. In general, the bigger the sample size, the lower the p-values. This means that with increasing sample size, weaker correlations become significant and the corresponding network becomes denser (Figure 1B). Second, different multiple testing correction methods (e.g., Bonferroni[9] or Benjamini-Hochberg[10]) have different underlying assumptions, such as controlling for the familywise error rate versus the false discovery rate (FDR), respectively. However, in practice, the choice of one method over another is usually not scrutinized adequately. Thus, depending on the arbitrary choice of error correction and significance level, one may obtain vastly different networks (Figure 1B) which are all statistically sound, but that do not necessarily represent relevant underlying biological mechanisms.

We here address the problem of correlation-based network inference from a different perspective. Instead of a statistically-driven cutoff selection, we propose to choose the correlation cutoff that produces the correlation network with the highest agreement to a given ground truth (Figure 1C and D), hereafter referred to as 'biological reference'. That is, we search for the network that shows the highest overlap with the known underlying biology, thereby avoiding the above-mentioned arbitrarily determined cutoffs for p-values.

We postulate that even a coarse, incomplete, or partially incorrect biological reference is suitable for this approach, as long as a sufficient amount of correct biological knowledge is covered. In many cases, the molecular networks regulating the system under study are not fully known, which results in an only partial biological reference being available. For example, often only few of the pathways of the system under study are well-characterized, and for some systems, detailed biochemical information is not available at all. In these cases, we will demonstrate that one can still use the available prior knowledge as a biological reference and obtain a cutoff that is close to the global optimum.

In this paper, we first show that, under certain circumstances, statistical significance selection is indeed substantially influenced by the dataset size. We then apply the prior-knowledge based cutoff optimization approach to plasma Immunoglobulin G (IgG) glycomics measurements. In this particular case, we have a well-characterized, supposedly complete biochemical synthesis pathway, which we use as gold standard biological reference to test our optimization approach. We show that the optimal correlation cutoff is unique and sample size independent. Moreover, even when the optimization procedure is performed with only a fraction of the original biological reference, the resulting optimum remains the same. We demonstrate the generalizability of the algorithm by applying it to metabolomics data, for which a full detailed prior knowledge is not available. We show that different sources of partial and coarse prior information lead to the same optimal network. Finally, we consider RNA-sequencing data from The Cancer Genome Atlas (TCGA), and show that the optimized networks are systematically superior to statistical ones in identifying known molecular interactions not included in the optimization procedure. This proves that partial prior knowledge can be exploited to infer a data-driven correlation networks that represent true although possibly unknown biological interactions better than regular statistically-inferred networks.

Results

Correlation cutoffs of most correlation measures depend on sample size

For most correlation measures, the larger the sample size, the lower the resulting correlation cutoff at a given significance level. In other words, increasing the number of subjects measured in a study automatically results in a denser correlation network. To quantitatively investigate this effect, we analyzed IgG glycomics measurements from four large Croatian cohorts (see Methods). In the following, the results for one of the four cohorts (Korčula 2013) are shown, while the other three cohorts were used for replication. The discovery dataset included 669 samples and 50 glycan structures measured. Data were normalized, log-transformed and corrected for age and gender prior to analysis.

We subsampled the glycomics dataset without replacement to simulate different sample sizes, from 10 to 669 samples. For each subsample, we computed the glycan correlation matrix and applied a 0.01 FDR cutoff using the Benjamini-Hochberg method as an exemplary approach for multiple testing correction. Results would be qualitatively identical with other methods (e.g. Bonferroni) and α levels. We considered two correlation measures commonly used in the field of computational biology: Classical pairwise Pearson correlation and partial correlation, which accounts for the presence of confounders (see Methods). We included two different estimators for partial correlation: Exact partial correlations obtained from the inversion of the covariance matrix (referred to as *parcor*), and a shrinkage-based regularization approach, which has been shown to give more stable estimates and still works in datasets with less samples than variables (*GeneNet*[11]).

As expected, for both Pearson correlation and parcor, the significance cutoff, i.e. the smallest still-significant correlation coefficient (in absolute value), decreases with increasing sample size and does not converge even for larger sample sizes (Figure 2A, red and blue curves, respectively). Interestingly, partial correlations estimated with GeneNet do not show the same behavior, as the statistical correlation cutoff is fairly stable across the considered sample sizes (Figure 2A, black line). This is also reflected in the total number of edges in the resulting network: While for Pearson correlation and parcor the number of significant coefficients included in the network systematically increases with the sample size, the network estimated with GeneNet maintains a roughly constant number of edges (Figure 2B). As an example, when considering twice as many samples, from 200 to 400, the GeneNet network remains stable with around 60 edges, while the Pearson correlation network increases by a factor of roughly 1.2 (from 655 to 790) and the parcor network increases by a factor 1.5 (from 95 to 155). Analogous results were obtained in the three replication cohorts (Figure S1).

This first analysis showed that indeed there is a strong dependence of network density (number of significant correlation) on sample size of the dataset for both Pearson and partial correlations. GeneNet did not show this behavior, which is most likely an effect of the p-value estimation method used in the algorithm (see Methods). Moreover, GeneNet gave rise to a considerably more stable network, essentially independent of the sample size.

Reference-based cutoff optimization

We applied our reference-based network inference approach to IgG glycomics data, for which the pathway of synthesis is well characterized (Figure 3A). We have previously shown that edges in a partial correlation network represent single enzymatic reaction in the IgG glycosylation pathway[8]. In this first step, we tested how our method compares to regular statistical cutoffs.

As a quantitative measure of overlap, we used Fisher's exact test based on the overlap contingency table, which classifies glycan-glycan pairs based on whether an edge between them appears both in the correlation network and in the biological reference (true positives), only in the correlation network (false positives), only in the biological reference (false negatives), or in neither (true negatives). Thus, higher the overlap between correlation network and biological reference, the lower this p-value will be (see Methods). The cutoff that produces the maximum overlap to the biological reference is hereafter referred to as the "optimal cutoff" and the corresponding network as the "optimal network". As expected, regular Pearson correlation performed poorly in comparison to parcor, as it does not account for confounding factors, while GeneNet was the overall best performing method (Figure 3B). In this case, the optimal GeneNet network yields only a minor improvement over the network obtained with FDR 0.05, which turned out to be the statistical cutoff closest to the optimum. However, the performance of any statistical cutoffs cannot be predicted a priori and depends on the specific case under investigation. The analysis of the replication cohorts showed similar results (Figure S2). This first analysis proves that biological prior knowledge improves the choice of a network cutoff, and that the optimal network is identifiable and unique for all correlation measures considered.

To assess whether the optimal network obtained with our procedure depends on sample size, as statistical cutoffs, we again performed the optimization procedure on subsamples of the original dataset (Figure 3C). For GeneNet, the optimal cutoff turned out to be sample size-independent, as expected. This indicates that, by optimizing the cutoff with our approach even with a relatively small sample size (roughly 160 observations), we still obtain the same optimal network that we would get with a much larger dataset (669 observations). Strikingly, even for parcor and Pearson correlation, for which statistical cutoffs showed strong sample size dependence, the optimal cutoff appeared to be sample size-independent over 300 samples (Figure S3 and S4, respectively), although the overall performance was lower than GeneNet. In conclusion, using prior information to optimize the correlation cutoff allowed to infer the same *optimal* network regardless of the sample size of the dataset.

Incomplete, incorrect, or coarse biological references

Our optimization approach determines the correlation cutoff at which the data-driven network best represents the biological reference. However, IgG glycan synthesis is a well-characterized process, while in most other practical cases a reference that describes the system in accurate detail is not available. We postulate that even with an incomplete or partially incorrect biological reference, we will obtain a close-to-optimal network. To this end, we considered the performance obtained from the optimization procedure when comparing the full biological reference with an artificially incomplete, incorrect or coarse version of it, as described in the following.

Scenario 1: Incomplete biological reference. Since for many biological systems the full biochemical pathway of synthesis is not available, we simulated a case in which only a percentage of the IgG glycosylation pathway is known. To this end, we randomly constructed incomplete pathways by selecting a fraction (10% to 90% in increments of 10%) of the edges in the IgG glycosylation pathway shown in Figure 3A. For each percentage, we generated 100 different incomplete pathways and used each of them to optimize the correlation cutoff (Figure 4A). Obviously, due to the increase in false positives, the fewer edges from the original reference we consider, the lower the overlap to the correlation network becomes. Importantly however, the optimum is highly conserved across the curves, yielding the same optimal cutoff (0.23) regardless of the amount of prior information available. This means that if we only knew, e.g., 50% of the reactions in the IgG glycosylation pathway shown in Figure 3A, we would still obtain the identical optimal network as we would by using the full pathway.

Scenario 2: Partially wrong biological reference. In many cases, our understanding of how a biological system works might be partially incorrect. Therefore, we considered the possibility of our reference to include wrong information, i.e. a given number of wrong edges. We simulated an increasing number of edge swaps in the IgG glycosylation pathway until we reached full randomization. For each condition, we generated 100 different pathways and performed the optimization procedure on them (Figure 4B). Again, while the overall performance decreased as expected, the shape of the curve clearly leads to the same optimal cutoff as the original pathway for up to 20 swaps. This means that even when starting with a substantially incorrect prior, as long as partial truth is contained in the reference network, the optimized network will still produce the same network as the one obtained with the complete biological reference.

Scenario 3: Coarse biological reference. Sometimes no detailed biochemical mechanisms are known, but only general biological properties of the molecules in the dataset. For example, we know that glycan processing occurs when the sugar chain is already bound to the protein. In our datasets, we have the measurements of three different protein isoforms (IgG1, IgG2 and IgG3 together, and IgG4). Therefore, we can constrain the set of possible biochemical reactions only to glycans pairs *within the same IgG isoform* (adjacency matrix 1 in Figure 4C). Moreover, we know that glycosylation enzymes can only add a *single monosaccharide at a time* during glycan synthesis. Hence, we can further reduce the possible reactions to those between glycan pairs that differ of a single sugar unit (adjacency matrix 2 in Figure 4C). When comparing the optimization results carried out starting from these biological references to that of the full biochemical pathway (adjacency matrix 3 in Figure 4C), we observe that, while the overall performance varies, the optimal values are close to each other, thus producing similar networks. Therefore, even when biochemical details are not available for the system under study, other sources of information can be used for the optimization and lead to the same optimum as the complete biological reference.

The three scenarios' results replicated for the other cohorts (Figure S5 and S6).

In conclusion, for various cases of incomplete prior knowledge, our approach still leads to a close-to-globally-optimal network.

Application to metabolomics data

In order to test whether our approach can be generalized to other data types, we applied the algorithm to untargeted urine metabolomics dataset (Table 2). The dataset consisted of 95 samples with 1,021 measured metabolites, and is therefore significantly more complex than the glycomics dataset considered so far, which only included 50 variables. Data were normalized, log-transformed, imputed and corrected for age, gender, and BMI prior to analysis (see Methods).

Since current pathway databases cover only a part of the blood metabolites measured in a typical mass-spectrometry-based analysis, we had to rely on partial prior information: (1) Enzymatic reactions connecting the measured metabolites were obtained from the RECON2 database[12]. In addition, as a weak informative prior, we considered two block adjacency matrices, allowing interactions among molecules (2) within the same biological pathway (in the following referred to as *sub-pathway*) or (3) within the same general molecular class (referred to as *super-pathway*).

We inferred GeneNet-based networks using these three priors as biological references (Figure 5). Although the absolute performances varied significantly depending on the chosen prior, the maxima were still remarkably close to each other. This means that the corresponding resulting optimal networks will be similar. Also in this case, the statistical cutoff of FDR 0.05 was found to perform comparably to the optimized cutoff.

For reference, the performances of Pearson and parcor correlation measures can be found in Figure S7.

In conclusion, we demonstrated that our approach can be generalized to metabolomics data, where a full biological reference is unavailable. Partial prior information can be used from different sources and the optima obtained with different priors are highly consistent.

Application to transcriptomic data

To evaluate the approach on a substantially different type of omics data not based on mass-spectrometry, we analyzed RNA-sequencing data from The Cancer Genome Atlas[13] (TCGA, Table 3). After preprocessing, the dataset included expression measurements of 11,993 genes from 3,571 samples across 12 different cancer types (see Methods). This dataset is much more complex than the glycomics and the metabolomics datasets, and is thus an informative test case to evaluate the flexibility of our approach. Expression values were corrected for age, gender and cancer type prior to analysis.

The analysis was performed separately for transcripts in 311 different pathways, as defined in the Reactome database[14,15] (see Methods for details). For each of these pathways, we defined two independent biological references: (1) Protein complexes from the CORUM database[16], and (2) Other protein-protein interactions as described in the STRING database[17,18]. In order to assure independence of the two biological priors, we removed all interactions contained in the CORUM reference from the STRING reference. Consequently, neither reference contained any of the connections included in the respective other reference. If our optimization approach is truly able to identify a biologically meaningful cutoff, a network optimized on one reference will still be able to recover significant amount of information included in the second, independent prior.

We first optimized the network cutoff based on the STRING reference. Since we tested 311 pathways in this analysis, we used a conservative significance threshold for the Fisher's p-value of $0.01/311 = 3.21 \cdot 10^{-5}$, which yielded 46 pathways with a significant optimum (Figure S8). In this first phase, the optimized networks showed a systematically higher overlap with the corresponding STRING reference when compared to their statistical counterparts, as expected by construction (Figure 6A).

When we computed the overlap of the STRING-optimized and the statistical network to the CORUM reference, the former was still systematically superior to the latter (Figure 6B). An interactive version of Figure 6 is provided as supplement (File S1). This finding indicates that the optimized network is able to represent true interactions that were not used in the optimization process better than statistically determined networks, proving its full inference potential.

As a showcase of how this difference in correlation cutoff translates into differences of the inferred networks, we compared the partial correlation network obtained at FDR 0.05 and with our optimization approach for the "Axon guidance” pathway (Figure 7). A network comparison for all pathways with a significant optimum is provided as an interactive R-Markdown file (File S2). Notice that PPI networks are much denser than biochemical pathways (see glycomics and metabolomics analysis), which was not reflected in partial correlation network estimated with statistical cutoffs.

Discussion

Correlation network inference often relies on correlation cutoffs based on p-values, which are however known to be substantially affected by sample size and are subject to an arbitrary choice of significance level and multiple testing correction procedure. We showed that an exception to this general observation is GeneNet[11], which exhibits a remarkable robustness to sample size, but is still subject to choice of a proper statistical cutoff. While several other network inference approaches that do not rely on a p-value-based threshold exist, it is worth mentioning that most these methods still require the assignment of a cutoff parameter, for example the lambda parameter in the graphical Lasso[19], which suffers from the same problem of the p-value based cutoff. Other methodologies that do not rely on any cutoff parameter, like for example the weighted network approach of WGCNA[20], produce a fully connected network and, although powerful in identifying clusters or modules of co-regulated genes or proteins, are unsuitable to identify single enzymatic steps in synthesis pathways.

The approach presented here overcomes the problem of the cutoff choice by establishing a biologically optimal correlation cutoff for network inference. The procedure ranges over the correlation cutoff value until an optimal overlap with a given, possibly incomplete, biological reference is achieved. We benchmarked the approach on LC-ESI-MS IgG glycosylation data from four large Croatian cohorts. For this type of data, the full synthesis pathway has been established and thus served as a gold standard for method evaluation. We showed that for the GeneNet partial correlation method, the resulting optimization curve lead to a well-determined and unique optimum, regardless of sample size and p-value cutoffs. The other investigated correlation-based methods performed inferior compared to GeneNet.

The approach was then applied to the more realistic case of partial prior knowledge, i.e. the case where a full, detailed and correct biological reference is not available. We considered three different scenarios: 1. Only a fraction of the biochemical pathway of synthesis is known; 2. The biochemical pathway contains incorrect information; 3. Only relations between classes of variables are known. In all three cases, we obtained nearly optimal networks despite the reduced biological knowledge that was available. This means that even only marginally informative priors are sufficient to obtain a reasonable approximation of the true network optimum.

We further demonstrated the applicability of the approach to metabolomics and transcriptomics data, for which only partial prior knowledge is available. The three partial biological references used for the metabolomics data, based either on metabolic reactions or molecular annotations, yielded very similar optima, supporting the claim that partial knowledge from different sources, like sub- and super-pathway annotations, can be used to optimize the correlation cutoff.

Interestingly, for both the glycomics and the metabolomics dataset, the statistical 0.05 FDR cutoff was very close to the optimum. However, this good performance of FDR is coincidental and cannot be generalized to other datasets or data types. This was corroborated by the analysis of transcriptomics data, for which FDR cutoffs were found to significantly overestimate the optimal cutoffs in many cases. This proves that the performance of FDR cannot be known *a priori* and varies substantially depending on the chosen significance level and data type, and thus does not guarantee an optimal network.

To validate the potential of the optimized networks to infer new biological interactions not included in the biological reference used for the optimization process, we optimized the cutoff of transcriptomics networks based on known protein-protein interactions from the STRING database. We then tested how well the optimized networks represented a different source of information, namely protein complexes from the CORUM database. We constructed the two biological references so that they included complementary information, with no redundancy, and hence were completely independent. Our results show that the STRING optimized network had a significant overlap with the CORUM reference, and that this overlap was much higher than that obtained from statistically inferred networks, meaning that optimizing on partial prior knowledge still allows to correctly infer unknown biology.

The procedure described in this paper requires a quantitative overlap measure to perform the cutoff optimization. We chose Fisher's exact test p-value as a proxy for the agreement between calculated correlation network and prior knowledge. It is to be noted that more conventional machine learning measures exist for classification problems. As an example, the popular F_1-score[21], does not account for true negatives and was therefore disregarded here. Interestingly, Matthews correlation coefficient[22], another popular measure that uses all values in the contingency table, is actually related to the Fisher's p-value. Its absolute value is proportional to the square root of the chi-square statistic, which is asymptotically equivalent to that of the Fisher's exact test[23].

Cutoff optimization as presented in this paper is a very flexible and generalizable inference strategy. Most other methods that account for prior knowledge integrate the biological reference directly into a specific network inference or regression framework[24–30], for example by penalizing or enhancing specific edges according to the biological reference. On the contrary, our approach uses prior knowledge as an external reference system to optimize the purely data-driven association matrix. This allows applying the same concept to different association measures, for example mutual information[31] or other non-linear association quantities, in future studies.

In conclusion, we propose a novel approach for network inference, by optimizing a correlation-based cutoff to prior pathway knowledge. The impact of our study lies in the demonstration that the same optimal network can be obtained when the available prior knowledge is incomplete, partially wrong, or only provides information on the overall relationship across classes of molecules (Figure 8). Consequently, a network fitted to partial priors can be used to enrich the knowledge with new, previously unknown interactions, or to rectify incorrect links, and can therefore serve as a valuable tool to infer biological interactions when a direct experimental validation is unavailable or unfeasible.

Materials and Methods

Glycomics datasets

Plasma samples from four Croatian cohorts[32] were analyzed (see Table 1 for details). For this paper, we only considered unrelated individuals. To this end, kinship coefficients[33] were estimated based on identity-by-state, which were computed using genotyped SNP data with the ibs function in the GenABEL package[34] for R. Unrelated individuals were obtained by selecting all pairs of individuals whose kinship coefficient was higher than 0.0312, which removed all individuals that were first degree cousins or closer. Samples with missing values were excluded from the analysis.

IgG was isolated from plasma using affinity chromatography with 96-well protein G monolithic plates, as reported previously[35]. IgG Fc glycopeptides were extracted by trypsin digestion and measured by LC-ESI-MS, which allows the separation of different IgG glycoforms. In Caucasian populations, the tryptic Fc glycopeptides of IgG2 and IgG3 have identical peptide moieties[36,37] and so cannot be separately identified using the profiling method. Furthermore, only 10 glycoforms of IgG4 were detectable due to the low abundance of this IgG subclass in human plasma. A detailed description of the experimental procedure can be found in Selman *et al.* (2012)[38] and in Huffman *et al.* (2014) [39].

For each IgG subclass, the LC-ESI-MS raw ion counts were normalized using probabilistic quotient normalization, which was originally introduced for metabolomics measurements[40]. The reference sample was calculated as the median value of each glycan abundance across all measured samples. For each sample, a vector of quotients was then obtained by dividing each glycan measure by the corresponding value in the reference sample. The median of these quotients was then used as the sample's dilution factor, and the original sample values were subsequently divided by that value. This procedure was repeated for each sample in the dataset. The data were log transformed and corrected for age and gender prior to statistical analysis.

The Croatian cohorts received ethical approval of the ethics committee of the University of Split School of Medicine, as well as the South East Scotland Research. Written informed consent was obtained from each participant.

Metabolomics dataset

Metabolomic samples were taken from an antipsychotics study conducted in Qatar[41]. Urine samples were analyzed using ultra-high-performance liquid-phase chromatography and gas-chromatography separation, coupled with tandem mass spectrometry by Metabolon, Inc. Data were runday-median scaled, normalized using probabilistic quotient normalization[40] and log-transformed. From the original data matrix, we first excluded metabolites with more than 20% missing values, and then samples with more than 10% missing values. Samples with missing covariates were subsequently excluded from the analysis. The filtered data matrix contained 97 samples and 1,021 metabolites (527 known structures and 494 unknown), see Table 2. Remaining missing values were imputed with a KNN-based method with variable pre-selection[42]. Data were corrected for age, gender and BMI prior to analysis.

All participants have given written informed consent and the local ethics committees have approved the studies.

Transcriptomics dataset

RNA-seq data were downloaded from The Cancer Genome Atlas[13] (TCGA) and initially included 3,599 samples and 16,115 genes from 12 cancer types: Acute myeloid leukemia, bladder urothelial carcinoma, breast invasive carcinoma, colon adenocarcinoma, glioblastoma multiforme, head & neck squamous cell carcinoma, kidney clear cell carcinoma, lung adenocarcinoma, lung squamous cell carcinoma, ovarian serous cystadeno-carcinoma, rectum adenocarcinoma, and uterine corpus endometrioid carcinoma[13], see Table 3.

For each cancer type, genes with more than 20% of missing values were excluded. Missing values in the remaining genes were imputed using a KNN-based method with variable pre-selection[42]. Values were corrected for age and gender, and samples without this information were excluded. We only considered genes present in all cancer types after preprocessing and we further corrected for cancer type. The final dataset included 3,571 samples and 11,993 genes.

Correlation analysis

Three measures of correlation were used in this study: 1) Classical Pearson correlation, which represents the linear relation between two variables. 2) Partial correlation, which allows to account for the presence of confounders or covariates and is calculated as the Pearson correlation coefficient corrected for the presence of all other variables[43]. Analytically, a partial correlation matrix can be obtained by inverting and normalizing the covariance matrix. In this paper, we refer to this technique as parcor. This estimation procedure is efficient but unstable for low sample sizes. 3) A more stable estimate of the partial correlation matrix can be obtained with the GeneNet algorithm[11], where a shrinkage parameter is optimized to correct the covariance matrix prior to inversion. Moreover, the algorithm fits a mixture model to the partial correlation matrix to compute the p-values[11], which results in a more robust p-value estimation. We attribute to this particular step the observed independence from the data sample size of the partial correlation networks calculated with GeneNet.

For the metabolomics data analysis, partial correlations were corrected for unknown variables. Note that these unknowns were then excluded from the overlap evaluation during the optimization procedure, which was only based on identifiable metabolites. Statistical cutoffs were based on the full correlation matrix, including unknowns.

Biological references

Glycomics data. The biological reference reflects the current understanding of the IgG glycosylation pathway, as established in Benedetti *et al.* (2017)[8]. Glycans can be modified by the addition of one monosaccharide at a time, but only selected reactions are enzymatically feasible, as shown in Figure 3A.

Metabolomics data. There is no established complete biochemical pathway to consider as biological reference for metabolomics data. Known metabolic reactions were imported from the RECON2 database[12] and included in one of the adjacencies. As a more coarse type of biological references, we used sub- and super-pathway annotations provided with the metabolites measurements by Metabolon, Inc., from which adjacency matrices were created by connecting all metabolites within the same sub- or super-pathway, respectively.

Transcriptomics data.

Pathway annotations were imported from the Reactome database[14,15]. We restricted the analysis to pathways containing at least 50 genes and at most 1,000 genes and with at least 50% of the genes in the pathway measured in the TCGA data. These constraints led to a total of 311 Reactome pathways being selected. For each pathway, protein-protein interactions were downloaded from the STRING database[17,18], while protein complexes were taken from the CORUM database[16]. CORUM interactions were subsequently removed from the STRING prior, creating two independent references. The resulting modified STRING prior was then used as biological reference for the optimization, while the CORUM prior was used for validation.

Overlap estimation

The overlap between biological reference and correlation network was calculated using Fisher's exact tests, which evaluate whether two categorical variables are statistically independent[44], with low p-values indicating a lack of independence. For the purposes of this analysis, we treated the Fisher's p-value as a measure of the overlap between the reference and the calculated correlation network. Lower p-values indicate a better overlap. Therefore, the optimal cutoff for a given correlation is defined as the cutoff at which the Fisher's p-value is the lowest. Specifically, all correlation coefficients are first classified in a contingency table, according to significant yes/no and whether the corresponding variable pair is connected by an edge in the biological reference adjacency matrix (Figure 3A). The p-value of Fisher's exact test is calculated according to the hypergeometric distribution as:

$$p = \frac{\binom{TP+FP}{TP}\binom{FN+TN}{FN}}{\binom{TP+FP+FN+TN}{TP+FN}}$$

Data Availability

Preprocessed glycan concentrations, as well as values corrected for age and gender, are available from the figshare database with the DOI: 10.6084/m9.figshare.5335861.

The metabolomics data cohort are available upon request and after compliance with the policies and procedures of Weill Cornell Medicine-Qatar and Qatar National Research Fund for data sharing. Requests can be submitted to Hassen Al-Amin at haa2019@qatar-med.cornell.edu.

The transcriptomics data are available online through the TCGA Research Network portal at https://cancergenome.nih.gov.

Acknowledgements

The CROATIA_Vis, CROATIA_Korčula, and CROATIA_Split studies were funded by grants from the Medical Research Council (UK), European Commission Framework 6 project EUROSPAN (Contract No. LSHG-CT-2006-018947), FP7 contract BBMRI-LPC (grant No. 313010), Croatian Science Foundation (grant 8875) and the Republic of Croatia Ministry of Science, Education and Sports (216-1080315-0302). We would like to acknowledge the staff of several institutions in Croatia that supported the field work, including but not limited to The University of Split and Zagreb Medical Schools, Institute for Anthropological Research in Zagreb and the Croatian Institute for Public Health. This work was funded in part by grants from the German Federal Ministry of Education and Research (BMBF), by BMBF Grant No. 01ZX1313C (project e:Athero-MED), by the European Commission MIMOmics (contract #305280), IMForFuture (contract #721815) and CarTarDis (contract #602936) grants, and by the European Structural and Investment Funds grant to the Croatian National Centre of Research Excellence in Personalized Healthcare (#KK.01.1.1.01.0010) and Croatian National Centre of Competences in Molecular Diagnostics (#KK.01.2.2.03.0006). The Antipsychotic cohort was funded by funded by Qatar National Research Fund (QNRF) (www.qnrf.org) (grant No. NPRP 4-268-3-086). K.S. is supported by the Biomedical Research Program at Weill Cornell Medicine in Qatar, a program funded by the Qatar Foundation.

Authors Contributions

E.B. and J.K. conceived and designed the project. M.P.-B., T.K, T.S., M.J.H.S., I.R., O.P., C.H., H.A., K.S., G.K., G.L. contributed the data. E.B. and N.G. performed the analyses on the glycomics and metabolomics data, M.B. performed the analysis on the transcriptomics data. E.B. and J.K. wrote the primary manuscript. All authors approved the final manuscript.

Conflict of Interest

G.L. declares that he is a founder and owner of Genos, a private research organization that specializes in high-throughput glycomics and has several patents in the field. M.P.-B. is an employee of Genos. The remaining authors declare no competing financial interests.

References

1. Albert, R. Network inference, analysis, and modeling in systems biology. *Plant Cell* **19**, 3327–3338 (2007).

2. Carter, H., Hofree, M. & Ideker, T. Genotype to phenotype via network analysis. *Curr. Opin. Genet. Dev.* **23**, 611–621 (2013).

3. Rider, A. K. *et al.* Networks' Characteristics Matter for Systems Biology HHS Public Access. *Netw Sci* **213**, (2014).

4. Barabási, A.-L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.* **12**, 56–68 (2011).

5. Yang, Y. *et al.* Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. *Nat. Commun.* **5**, 3231 (2014).

6. Krumsiek, J., Suhre, K., Illig, T., Adamski, J. & Theis, F. J. Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC Syst. Biol.* **5**, 21 (2011).

7. Krumsiek, J. *et al.* Mining the Unknown: A Systems Approach to Metabolite Identification Combining Genetic and Metabolic Information. *PLoS Genet.* **8**, e1003005 (2012).

8. Benedetti, E. *et al.* Network inference from glycoproteomics data reveals new reactions in the IgG glycosylation pathway. *Nat. Commun.* **8**, 1483 (2017).

9. Dunn, O. J. Estimation of the Medians for Dependent Variables. *Ann. Math. Stat.* **30**, 192–197 (1959).

10. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).

11. Schäfer, J. & Strimmer, K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat. Appl. Genet. Mol. Biol.* **4**, Article32 (2005).

12. Swainston, N. *et al.* Recon 2.2: from reconstruction to model of human metabolism. *Metabolomics* **12**, 109 (2016).

13. Hoadley, K. A. *et al.* Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin. *Cell* **158**, 929–944 (2014).

14. Croft, D. *et al.* The Reactome pathway knowledgebase. *Nucleic Acids Res.* **42**, D472--D477 (2014).

15. Fabregat, A. *et al.* The Reactome Pathway Knowledgebase. *Nucleic Acids Res.* **46**, D649--D655 (2018).

16. Giurgiu, M. *et al.* CORUM: the comprehensive resource of mammalian protein complexes—2019. *Nucleic Acids Res.* **47**, D559–D563 (2019).

17. Szklarczyk, D. *et al.* STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **43**, D447--D452 (2015).

18. Szklarczyk, D. *et al.* The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Res.* **45**, D362--D368 (2017).

19. Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *J. R. Stat. Soc. Ser. B* **58**, 267–288 (1996).

20. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).

21. Powers, D. M. W. EVALUATION: FROM PRECISION, RECALL AND F-MEASURE TO ROC, INFORMEDNESS, MARKEDNESS & CORRELATION. *J. Mach. Learn. Technol. ISSN* **2**, 2229–3981 (2011).

22. Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta - Protein Struct.* **405**, 442–451 (1975).

23. Camilli, G. The relationship between Fisher's exact test and Pearson's chi-square test: A Bayesian perspective. *Psychometrika* **60**, 305–312 (1995).

24. Wang, Z., Xu, W., San Lucas, F. A. & Liu, Y. Incorporating prior knowledge into Gene Network Study. *Bioinformatics* **29**, 2633–2640 (2013).

25. Linde, J., Schulze, S., Henkel, S. G. & Guthke, R. Data- and knowledge-based modeling of gene regulatory networks: an update. *EXCLI J.* **14**, 346 (2015).

26. Pei, B. & Shin, D.-G. Reconstruction of Biological Networks by Incorporating Prior Knowledge into Bayesian Network Models. *J. Comput. Biol.* **19**, 1324–1334 (2012).

27. Zuo, Y., Cui, Y., Yu, G., Li, R. & Ressom, H. W. Incorporating prior biological knowledge for network-based differential gene expression analysis using differentially weighted graphical LASSO. *BMC Bioinformatics* **18**, 99 (2017).

28. Ante, M., Wingender, E. & Fuchs, M. Integration of gene expression data with prior knowledge for network analysis and validation. *BMC Res. Notes* **4**, 520 (2011).

29. Li, Y. & Jackson, S. A. Gene Network Reconstruction by Integration of Prior Biological Knowledge. *G3 Genes|Genomes|Genetics* **5**, 1075 (2015).

30. Stavrakas, V., Melas, I. N., Sakellaropoulos, T. & Alexopoulos, L. G. Network Reconstruction Based on Proteomic Data and Prior Knowledge of Protein Connectivity Using Graph Theory. *PLoS One* **10**, e0128411 (2015).

31. Shannon, C. E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).

32. Rudan, I. *et al.* "10 001 Dalmatians:" Croatia Launches Its National Biobank. *Croat. Med. J.* **50**, 4–6 (2009).

33. Astle, W. & Balding, D. J. Population Structure and Cryptic Relatedness in Genetic Association Studies. *Stat. Sci.* **24**, 451–471 (2009).

34. Aulchenko, Y. S., Ripke, S., Isaacs, A. & van Duijn, C. M. GenABEL: an R library for genome-wide association analysis. *Bioinformatics* **23**, 1294–1296 (2007).

35. Pucić, M. *et al.* High throughput isolation and glycosylation analysis of IgG-variability and heritability of the IgG glycome in three isolated human populations. *Mol Cell Proteomics* **10**, M111.010090 (2011).

36. Jefferis, R. & Lefranc, M.-P. Human immunoglobulin allotypes: possible implications for immunogenicity. *MAbs* **1**, 332–338 (2009).

37. Balbin, M., Grubb, A., de Lange, G. G. & Grubb, R. DNA sequences specific for Caucasian G3m(b) and (g) allotypes: allotyping at the genomic level. *Immunogenetics* **39**, 187–193 (1994).

38. Selman, M. H. J. *et al.* Fc specific IgG glycosylation profiling by robust nano-reverse phase HPLC-MS using a sheath-flow ESI sprayer interface. *J. Proteomics* **75**, 1318–1329 (2012).

39. Huffman, J. E. *et al.* Comparative performance of four methods for high-throughput glycosylation analysis of immunoglobulin G in genetic and epidemiological research. *Mol. Cell. Proteomics* **13**, 1598–1610 (2014).

40. Dieterle, F., Ross, A., Schlotterbeck, G. & Senn, H. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in1H NMR metabonomics. *Anal. Chem.* **78**, 4281–4290 (2006).

41. Hammoudeh, S. *et al.* The prevalence of metabolic syndrome in patients receiving antipsychotics in Qatar: a cross sectional comparative study. *BMC Psychiatry* **18**, 81 (2018).

42. Do, K. T. *et al.* Characterization of missing values in untargeted MS-based metabolomics data and evaluation of missing data handling strategies. *bioRxiv* 260281 (2018). doi:10.1101/260281

43. Baba, K., Shibata, R. & Sibuya, M. Partial correlation and conditional correlation as measures of conditional independence. *Aust. N. Z. J. Stat.* **46**, 657–664 (2004).

44. Routledge, R., Routledge & Rick. Fisher's Exact Test. in *Encyclopedia of Biostatistics* (John Wiley & Sons, Ltd, 2005). doi:10.1002/0470011815.b2a10020

Figures

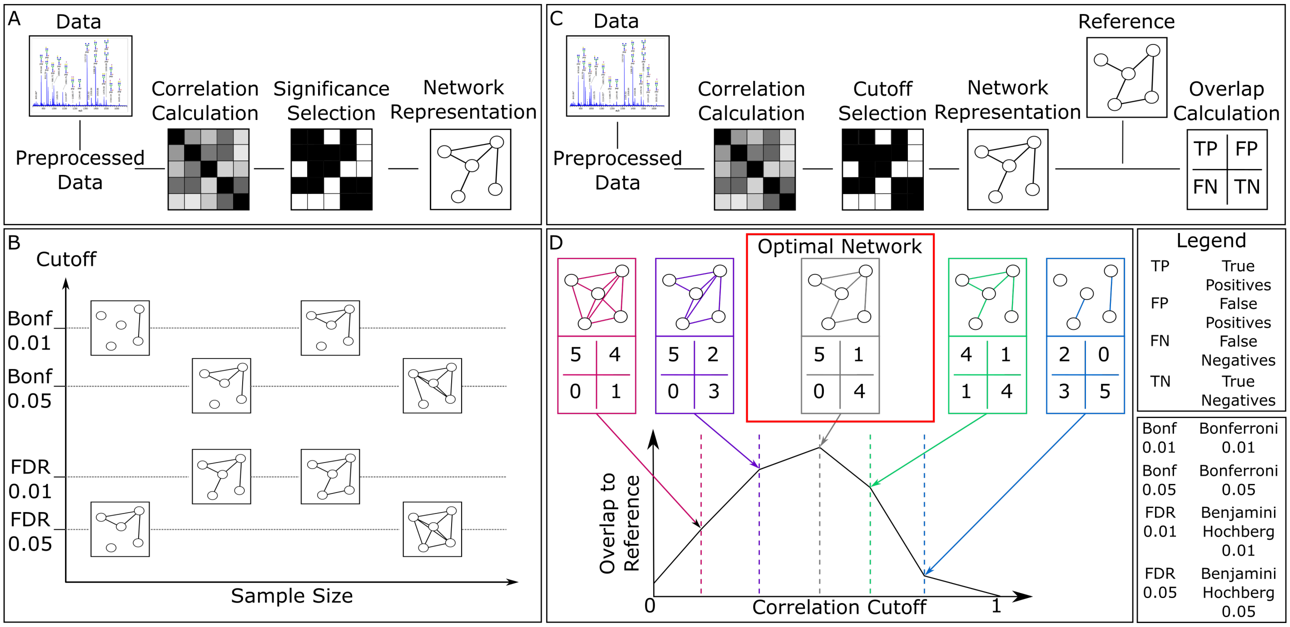

Figure 1: Pipeline of network inference and workflow of the paper.

A Typical pipeline of correlation network inference. A correlation matrix is estimated from the preprocessed data. A significance selection step identifies correlations that are statistically different from zero. Significant correlations are commonly visualized as a network. **B** Schematic representation of the dependence of the correlation network on sample size and statistical cutoff. Note that, despite looking substantially different, all resulting networks can be considered statistically correct. **C** Prior knowledge-based network overlap estimation. The correlation network is compared to a prior knowledge network, where the overlap is quantified using true positives (TP), false positives (FP), false negatives (FN) and true negatives (TN). Based on these values, a quality overlap measure between data-driven correlation matrix and biological reference is computed. **D** Prior knowledge-based network inference approach. We discard the p-value-based significance selection, and instead analyze how the overlap between correlation network and biological reference varies depending on the correlation cutoff. We then define optimal the correlation cutoff at the point where the overlap is maximal.

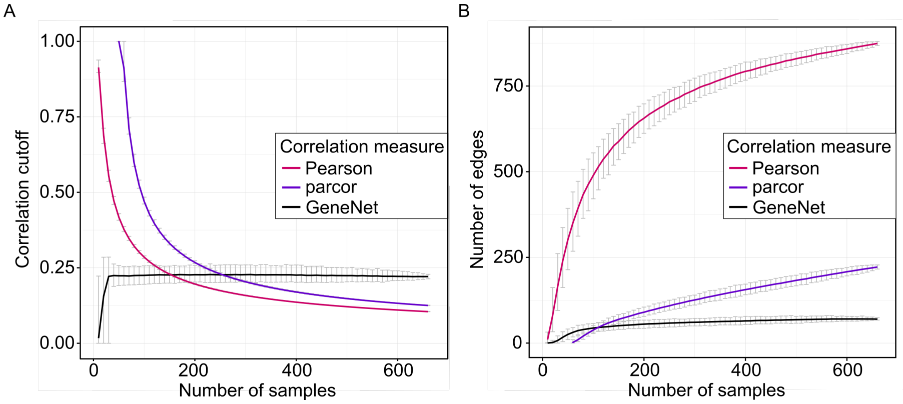

Figure 2: Correlation cutoff as a function of the sample size.

A Correlation cutoff (0.01 FDR) as a function of the dataset sample size for the three correlation measures considered: Pearson correlation (red), exact partial correlation (purple), shrinkage partial correlation based on GeneNet (black). Error bars represent 95% confidence intervals from 1,000 bootstrapping samples. **B** Number of edges in the correlation network after applying a 0.01 FDR cutoff as a function of the dataset sample size. Error bars represent 95% confidence intervals of 1,000 bootstrapping samples. Note that for parcor, correlation coefficients can only be estimated for a sample size greater than or equal to the number of variables, in this case 50.

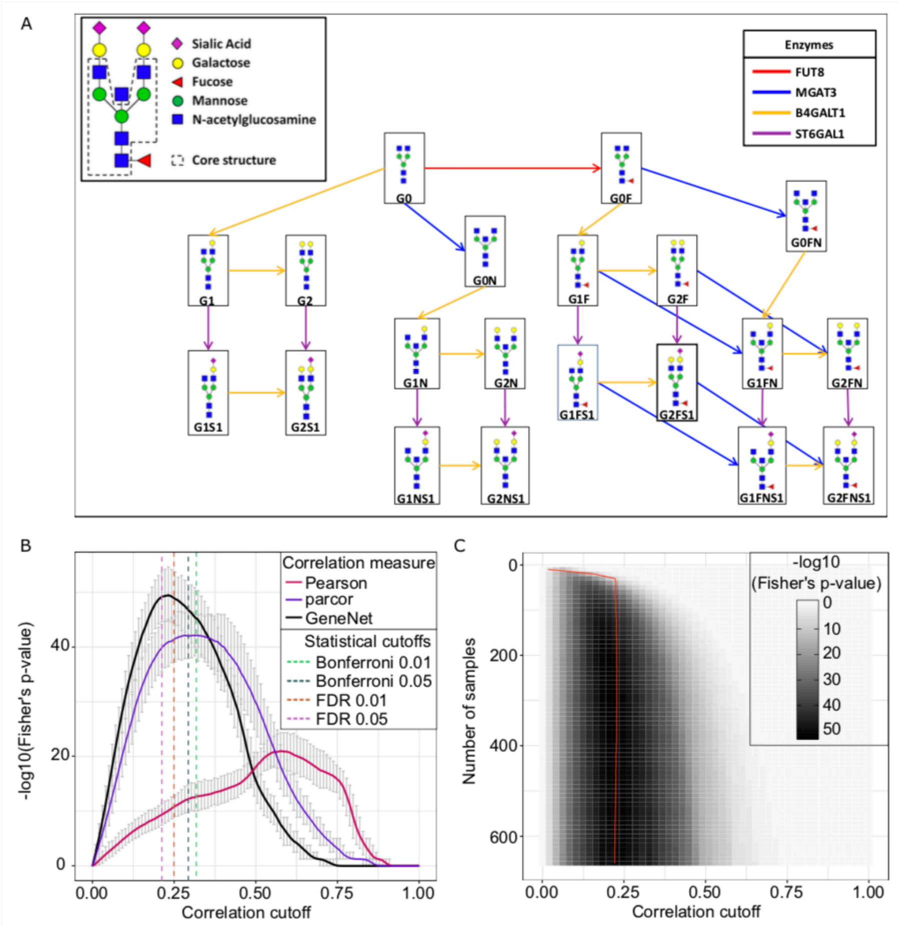

Figure 3: Network quality as a function of the correlation cutoff.

A IgG glycan structures and synthesis pathway. The figure was adapted from Benedetti et al. (2017) and represents the IgG glycosylation pathway, with nodes representing glycan structures and arrows representing single enzymatic reactions in the synthesis process. **B** Fisher's exact test p-values as a function of the correlation cutoff calculated for three correlation estimators: Pearson correlation (pink), exact partial correlation (purple), GeneNet partial correlation (black). For each correlation cutoff, the original dataset was bootstrapped 1,000 times. Error bars represent the 95% confidence intervals of the bootstrapping results. Dashed lines represent the statistical cutoffs for GeneNet on the original data matrix. **C** Fisher's exact test p-value for partial correlations estimated with GeneNet, as a function of both sample size and correlation cutoff. Shade represents the mean across 1,000 bootstrapping samples, while the red line represents the mean of the 0.01 FDR cutoff.

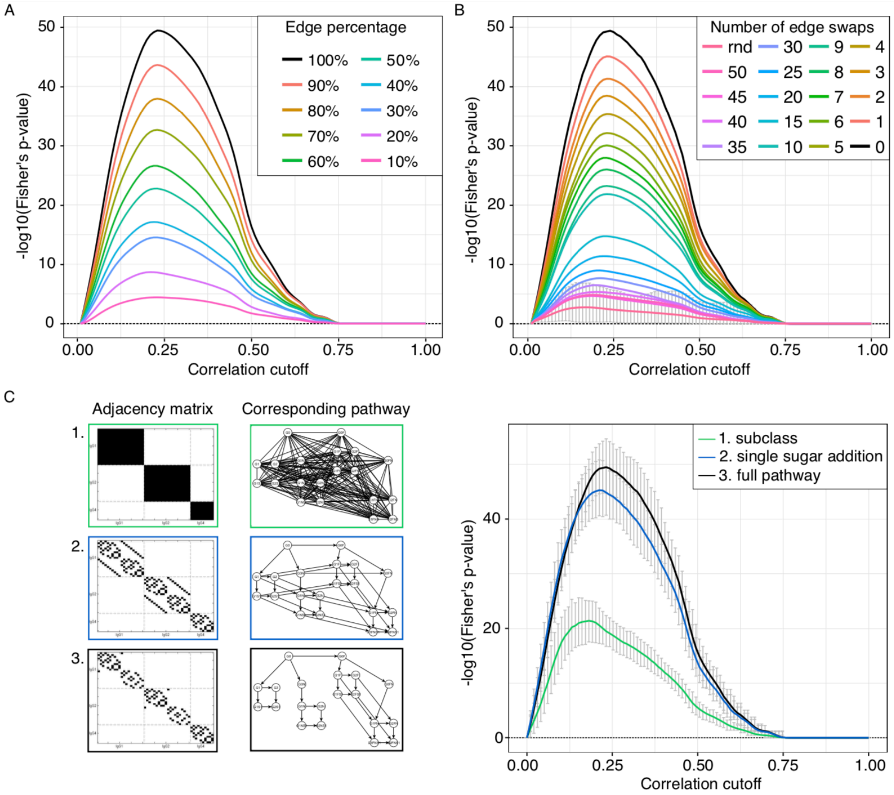

Figure 4: Cutoff optimization with partial knowledge.

A Incomplete biological reference. For each percentage, 100 different adjacency matrices were generated by randomly removing edges from the IgG glycosylation pathway. The curves in the figure represent the means of the 1,000 bootstrapping resamplings on each adjacency matrix. **B** Incorrect biological reference. Edges in the IgG glycosylation pathway were randomly swapped to simulate incorrect information in the biological reference. For each number of swaps, 100 adjacency matrices were generated and the averages over those curves and over the 1,000 bootstrapping resamplings are shown. Here, the red curve represents 100 fully randomized adjacency matrices (rnd). The error bars on this curve represent the 95% confidence interval of the bootstrapping. Any signal that falls within these intervals should be regarded as noise **C** Coarse biological reference. For IgG glycomics data we know that only enzymatic reactions between glycans attached to the same IgG isoform are feasible (adjacency matrix 1) and, in addition, that only they can be modified by the addition of one sugar unit at a time (adjacency matrix 2). The black curve corresponds to the optimization performed on the full reference (adjacency matrix 3) for comparison. The curves in the figure represent the means of the 1,000 bootstrapping resamplings and the different considered adjacency matrices. In all plots, the black curve corresponds to the optimization performed on the full reference.

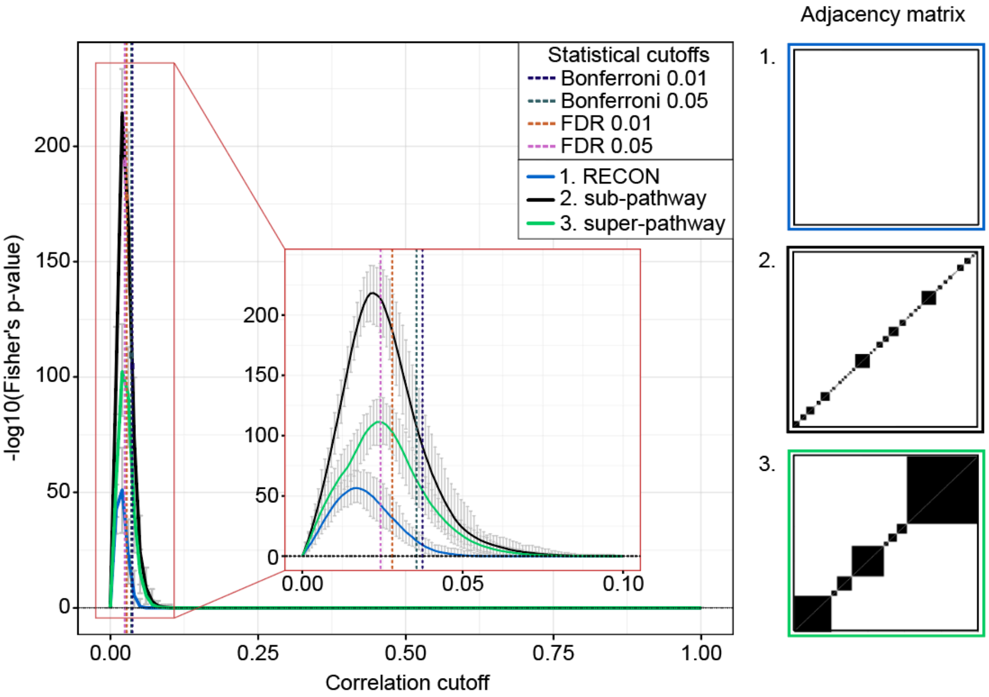

Figure 5: Cutoff optimization for metabolomics data.

We used biochemical reactions from the RECON database as partial prior knowledge (adjacency matrix 1), as well as sub- and super-pathway annotations (adjacency matrices 2 and 3, respectively). Curves in the figure represent the average over 100 bootstrapping resamplings, and error bars show the corresponding 95% confidence intervals. Vertical lines represent the statistical cutoffs computed on the original data matrix.

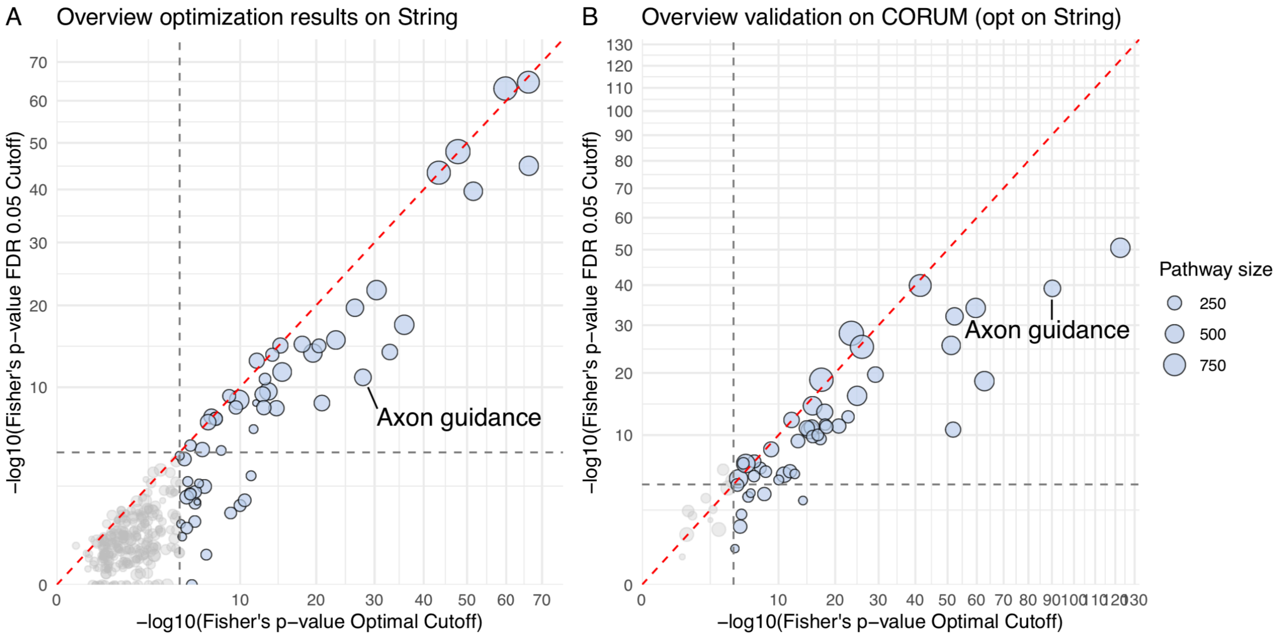

Figure 6: Overview of the TCGA transcriptomics data analysis.

A Comparison overlap of statistical (FDR 0.05) and STRING-optimized network with STRING prior. Each dot represents one of the 311 analyzed pathways, where the size and color of the dot codes for the pathway size. **B** Comparison of the overlap to the CORUM prior of the statistical (FDR 0.05) and STRING-optimized network. Each dot represents one of the pathways that resulted significant during optimization (blue dots in A), where the size and color of the dot codes for the pathway size. The gray dashed lines represent the significance threshold of $3.21 \cdot 10^{-5}$. In both cases, the STRING-optimized networks display a systematic better overlap to the priors than the statistically inferred networks.

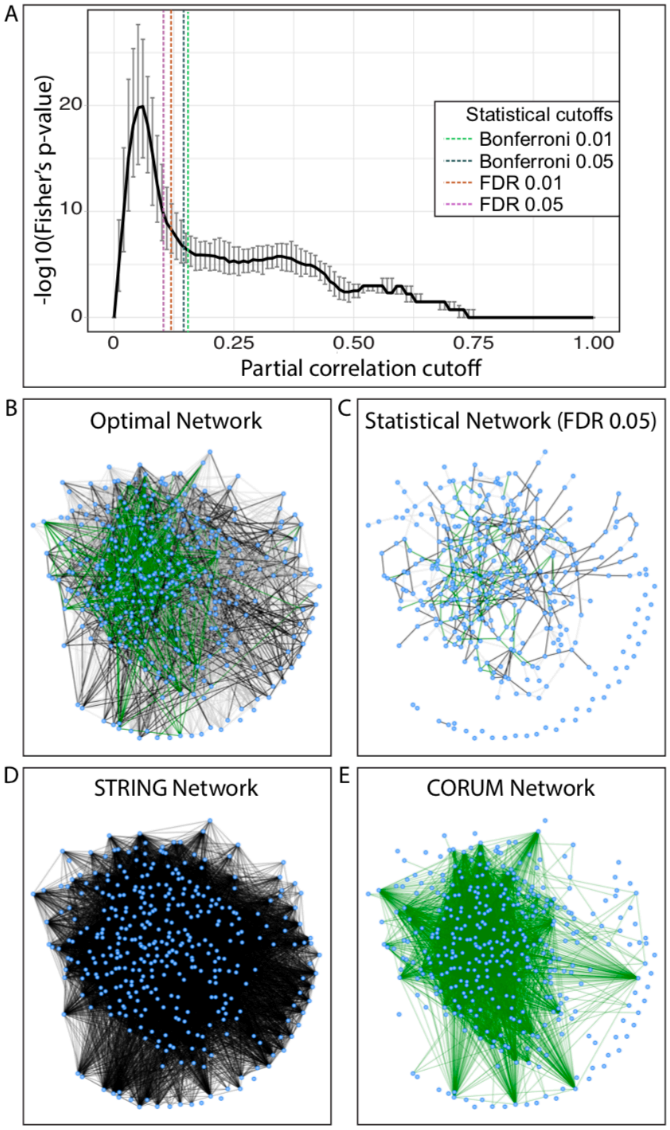

Figure 7: TCGA transcriptomics analysis results for the 'Axon guidance' pathway.

This example illustrates how statistical cutoffs can fail to identify a biological optimum. **A** Cutoff optimization. A protein-protein interaction network from STRING was used as reference. The black curve represents the average over 100 bootstrapping resamplings, and the error bars show the corresponding 95% confidence intervals. Vertical lines indicate the statistical cutoffs computed on the original data matrix. **B** Partial correlation network obtained with our optimization procedure. **C** Partial correlation network obtained with a 0.05 FDR cutoff. **D** Biological reference used for the optimization (PPI network from STRING without CORUM interactions). **E** Biological reference used for validation (protein complexes from CORUM). Black edges represent connections found in the biological reference used for optimization (STRING), green edges represent connections found in the reference used for validation (CORUM), while gray edges indicate connections not included in the prior knowledge.

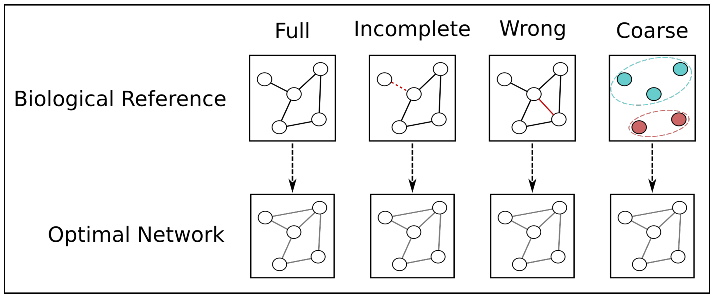

Figure 8: Main conclusions of the paper.

We have proven that biochemical pathways can be used to optimize a correlation cutoff to produce the network that best reflects known biological interactions. However, in most concrete cases, a full biological reference is not available. Our approach allows to retrieve the same optimal network even when the prior knowledge available is incomplete, wrong or only provides information on the relationship among classes of molecules. The optimized network will still provide the best correlation-based representation of the underlying molecular interactions.

Tables

Table 1. Characteristics of the four analyzed glycomics cohorts.

<table>
<tr><th colspan="6"></th><th colspan="6">Korčula
2013
Discovery</th><th colspan="6">Korčula
2010
Replication</th><th colspan="6">Split
Replication</th><th colspan="6">Vis
Replication</th></tr>
<tr><td colspan="6">Number of measured glycans</td><td colspan="6">50</td><td colspan="6">50</td><td colspan="6">50</td><td colspan="6">50</td></tr>
<tr><td colspan="2">IgG1</td><td colspan="2">IgG2</td><td colspan="2">IgG4</td><td colspan="2">20</td><td colspan="2">20</td><td colspan="2">10</td><td colspan="2">20</td><td colspan="2">20</td><td colspan="2">10</td><td colspan="2">20</td><td colspan="2">20</td><td colspan="2">10</td><td colspan="2">20</td><td colspan="2">20</td><td colspan="2">10</td></tr>
<tr><td colspan="6">Total number of samples</td><td colspan="6">695</td><td colspan="6">951</td><td colspan="6">994</td><td colspan="6">780</td></tr>
<tr><td colspan="3">Males</td><td colspan="3">Females</td><td colspan="3">277</td><td colspan="3">418</td><td colspan="3">339</td><td colspan="3">612</td><td colspan="3">390</td><td colspan="3">604</td><td colspan="3">326</td><td colspan="3">454</td></tr>
<tr><td colspan="6">Samples with no missing values</td><td colspan="6">669</td><td colspan="6">849</td><td colspan="6">980</td><td colspan="6">729</td></tr>
<tr><td colspan="6">Unrelated samples</td><td colspan="6">669</td><td colspan="6">504</td><td colspan="6">980</td><td colspan="6">395</td></tr>
<tr><td colspan="3">Males</td><td colspan="3">Females</td><td colspan="3">271</td><td colspan="3">398</td><td colspan="3">156</td><td colspan="3">348</td><td colspan="3">386</td><td colspan="3">594</td><td colspan="3">152</td><td colspan="3">243</td></tr>
<tr><td colspan="6">Age range
(mean, standard deviation)</td><td colspan="6">18–88
(53, 16)</td><td colspan="6">18–90
(56, 14)</td><td colspan="6">18–85
(50, 14)</td><td colspan="6">18–91
(55, 15)</td></tr>
</table>

Table 2. Characteristics of the metabolomics cohort.

<table>
<tr><th colspan="2">Antipsychotics urine
preprocessed</th><th colspan="2">Metabolon
Validation</th></tr>
<tr><td colspan="2">Number of total metabolites</td><td colspan="2">1,021</td></tr>
<tr><td colspan="2">Number of known structures</td><td colspan="2">527</td></tr>
<tr><td colspan="2">Number of samples</td><td colspan="2">95</td></tr>
<tr><td>Males</td><td>Females</td><td>35</td><td>60</td></tr>
<tr><td colspan="2">Age range
(mean, standard deviation)</td><td colspan="2">21-60
(36, 8)</td></tr>
<tr><td colspan="2">BMI range
(mean, standard deviation)</td><td colspan="2">17-46
(28, 5)</td></tr>
</table>

Table 3. Characteristics of the transcriptomics cohort.

<table>
<tr><th colspan="2">TCGA</th><th colspan="2">Validation</th></tr>
<tr><td colspan="2">Original number of transcripts</td><td colspan="2">16,115</td></tr>
<tr><td colspan="2">Original number of samples</td><td colspan="2">3,599</td></tr>
<tr><td>Males</td><td>Females</td><td>1,319</td><td>2,279</td></tr>
<tr><td colspan="2">Age range</td><td colspan="2">18-90</td></tr>
<tr><td colspan="2">Cancer types</td><td colspan="2">12</td></tr>
<tr><td colspan="2">Number of transcripts after preprocessing</td><td colspan="2">11,993</td></tr>
<tr><td colspan="2">Number of samples after preprocessing</td><td colspan="2">3,571</td></tr>
</table>