METHODOLOGY

Functional networks inference from rule-based machine learning models

Nicola Lazzarini[1], Paweł Widera[1], Stuart Williamson[2], Rakesh Heer[3], Natalio Krasnogor[1] and Jaume Bacardit[1*]

Abstract

Background: Functional networks play an important role in the analysis of biological processes and systems. The inference of these networks from high-throughput (-omics) data is an area of intense research. So far, the similarity-based inference paradigm (e.g. gene co-expression) has been the most popular approach. It assumes a functional relationship between genes which are expressed at similar levels across different samples. An alternative to this paradigm is the inference of relationships from the structure of machine learning models. These models are able to capture complex relationships between variables, that often are different/complementary to the similarity-based methods.

Results: We propose a protocol to infer functional networks from machine learning models, called FuNeL. It assumes, that genes used together within a rule-based machine learning model to classify the samples, might also be functionally related at a biological level. The protocol is first tested on synthetic datasets and then evaluated on a test suite of 8 real-world datasets related to human cancer. The networks inferred from the real-world data are compared against gene co-expression networks of equal size, generated with 3 different methods. The comparison is performed from two different points of view. We analyse the enriched biological terms in the set of network nodes and the relationships between known disease-associated genes in a context of the network topology. The comparison confirms both the biological relevance and the complementary character of the knowledge captured by the FuNeL networks in relation to similarity-based methods, and demonstrates its potential to identify known disease associations as core elements of the network. Finally, using a prostate cancer dataset as a case study, we confirm that the biological knowledge captured by our method is relevant to the disease and consistent with the specialised literature and with an independent dataset not used in the inference process.

Availability: The implementation of our network inference protocol is available at: http://ico2s.org/software/funel.html

Keywords: machine learning; biological knowledge extraction; network inference; functional networks

Background

The inference of biological networks is a highly relevant and challenging task in systems biology and integrative bioinformatics. Biological networks are graphs in which nodes represent genes or proteins, and a connection between them indicates some kind of biological relationship, e.g. regulatory or functional. The network inference is, in an essence, an attempt to reverse engineer the biological relationships from the high-throughput biological data [1].

Most biological network inference methods focus on the definition of gene regulatory networks, in which edges represent direct regulatory interactions between

*Correspondence: jaume.bacardit@newcastle.ac.uk

[1] Interdisciplinary Computing and Complex BioSystems (ICOS) research group, School of Computing Science, Newcastle University, Newcastle upon Tyne, UK

Full list of author information is available at the end of the article

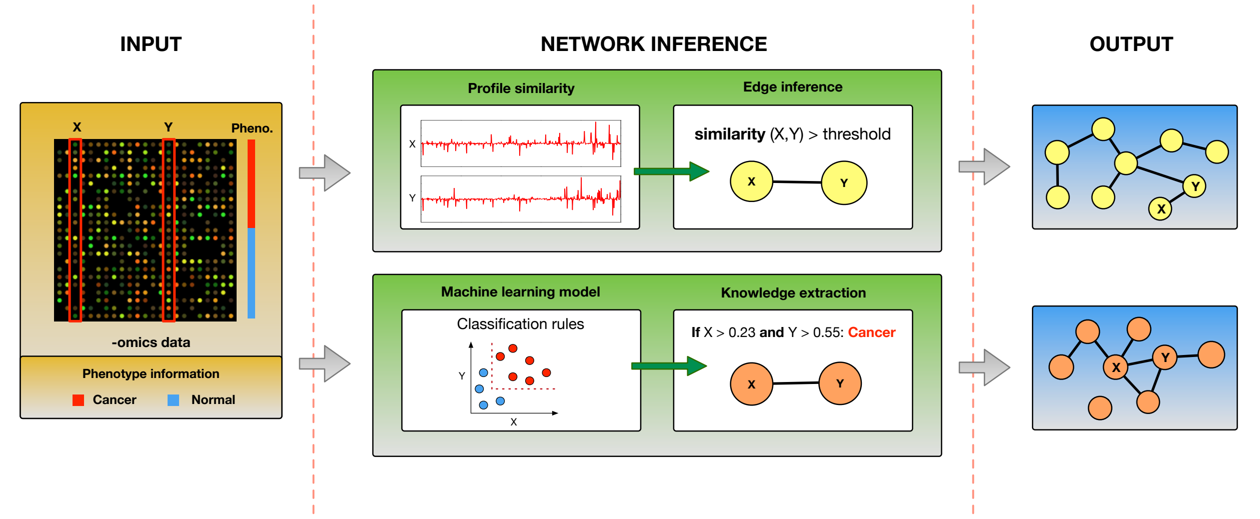


Fig 1 Two approaches to functional network inference: one based on the expression profile similarity and the other based on the extraction of knowledge from machine learning models. The similarity-based methods construct a new network edge $X \leftrightarrow Y$, when the similarity between the expressions of genes X and Y across the samples is above a threshold. Methods based on machine learning, first build a predictive model, in this example a rule-based model, using the samples phenotype information (class labels) and then construct a network edge $X \leftrightarrow Y$, when genes X and Y are used together within that model to classify the samples. As these two approaches lead to different functional networks, it is possible that they capture complementary knowledge.

genes [2–4]. Far less effort has been put into the design of methods to build functional networks in which a connection indicates a functional relationship, e.g. membership in the same pathway or protein complex. One of the typical uses of these networks is the identification of functional modules (subset of genes with multiple internal connections and a few connections with genes outside the module that describe, explain or predict a biological process or phenotype.).

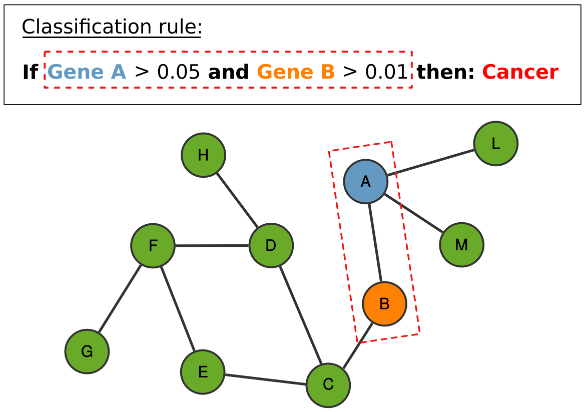
One of the earliest (but still widely used) approach to infer functional networks is the "guilt-by-association" principle [5]. That is, if two genes show *similar* expression profiles, it is assumed they are also functionally related (via a direct or indirect interaction). Initially, this paradigm was applied to infer networks from transcriptomics data, and this is why in most of the literature it is known as the *co-expression* network inference principle. Nevertheless, it is abstract enough to be applied to all kinds of biological data. It has been demonstrated that co-expression networks are able to effectively identify pathways and candidate biomarkers [6] or reveal gene modules representing a biological process perturbed in a disease [7], just to name a few examples, and the similarity-based approach remains the dominant method of functional network inference today, with many recent examples: [8–12].

A different approach that is recently gaining popularity, is the use of machine learning techniques to infer biological networks. Due to the wide range of knowledge representations used within machine learning methods (e.g. classification rules, decision trees, artificial neural networks, SVM kernels, etc.), they can discover more complex and diverse relationships, and overcome the limitations of the similarity-based methods. This is possible since within machine learning models the attributes are associated not because they are similar (e.g. have similar expression profiles), but because together they detect strong patterns. In addition, if learning is supervised, it can take advantage of the additional phenotype information (class labels of the samples, e.g. case and control) available with the data. Therefore, by mining the complex machine learning models, it should be possible to uncover new and different (biological) knowledge, that is likely to escape the traditional approaches. Figure 1 illustrates these differences between the two approaches (similarity-based methods vs. knowledge extraction from the machine learning models).

Alternative strategies exist to infer networks using machine learning. One approach is to train machine learning models that directly predict network edges [13], but this process requires an experimentally verified "ground truth" of known interactions and suitable controls. A different approach, which is the focus of this work, is to generate machine learning models from the biological data and then *mine* the structure of the models to infer networks. Several types of machine learning have been successfully applied to this task: unsupervised learning in the form of association rules [14], supervised learning using regression (model trees [15]) or classification (random forest [16]).

The specific focus of this paper is the network inference from rule-based machine learning models. Such models have been successfully applied before to extract knowledge from genetic data [17] and identify disease risk factors in a bladder cancer study [18]. The methods presented in these works share some pipeline components with our current work, such as the permutation test and a 2-phase learning strategy. In our previous works we applied rule-based machine learning to transcriptomics [19, 20], proteomics [21], lipidomics [22] and protein structure data [23]. We formulated a paradigm called *co-prediction* (in opposition to the classic co-expression) in which the prediction rules of a classification algorithm, in our case BioHEL [24], are used to identify relationships between genes.

Co-prediction is based on the assumption that attributes (e.g. genes) within the same classification rules, due to their co-operation in predicting the sample class, have an increased likelihood of being functionally related to the biological process in question (Figure 2). Differently than co-expression, the co-prediction approach exploits the phenotype information of the data (class labels) to detect functional relations.


Classification rule:

If Gene A > 0.05 **and Gene B** > 0.01 **then: Cancer**

Fig 2 Co-prediction paradigm. Association between the genes is inferred from their co-occurrence in classification rules.

However, from a methodological perspective, many questions remained unanswered. Can the co-prediction approach identify known genetic relationships? How can we quantify the biological significance of the co-prediction networks? What is the impact of data preprocessing on the generated networks? Is this methodology able to capture knowledge that escapes other methods? Are the discovered functional relationships meaningful in the human disease context?

To address these questions, we propose in this article a new network inference protocol, called FuNeL (Functional Network Learning). FuNeL substantially extends our previous work [19] by incorporating: (1) statistical filtering of inferred functional relationships via permutation tests, (2) a multi-stage network generation to maximise the knowledge extraction, and (3) a configurable feature selection stage to control the size of the generated networks.

We first tested FuNeL's ability to correctly identify functional relationships using a set of synthetic datasets. Then, we evaluated FuNeL on 8 real-world transcriptomics datasets related to different types of cancer. For each dataset we tested 4 different configurations of the protocol and compared the inferred networks to co-expression networks of equivalent size. In order to have an extensive evaluation of our approach, we employed 3 different methods to generate co-expression networks. We systematically looked at the differences between co-prediction and co-expression networks from two points of view: (1) the enriched biological terms and (2) the relationships between the genes known to be associated with a particular type of cancer. Finally, we used a prostate cancer dataset as a case study and performed a more detailed biological analysis of the enriched terms and the disease related genes. We looked at the largest hubs and the most central nodes in the prostate cancer co-prediction networks and studied their involvement in the disease. We found literature support for the association between these topologically important genes and prostate cancer, and we further confirmed it with an independent transcriptomics dataset (not used as a source in the inference process). Overall, we found that the FuNeL inferred networks: (1) capture relevant biological knowledge that is complementary to the knowledge captured by different co-expression networks, and (2) more adequately represent the relationships between genes associated with the disease targeted by each dataset.

Materials and Methods

In this section we describe the proposed network inference protocol, the datasets from which we inferred the networks and the experimental design we used to evaluate it.

The functional network inference protocol

The stages of the co-prediction inference protocol are illustrated in Figure 3. Two of these stages are optional

(1 and 4), they lead to a total of 4 different protocol configurations. If the first optional stage (feature selection) is performed, the original dataset is reduced to the most relevant attributes. In the second stage a rule-based machine learning is used to infer a network. This network is statistically refined in Stage 3, in which a permutation test is used to filter out non-significant nodes. The final stage, in which the network generation is repeated for the second time, is again optional. A complete time complexity analysis of the FuNeL protocol is available in Section 2 of the Supplementary Material.

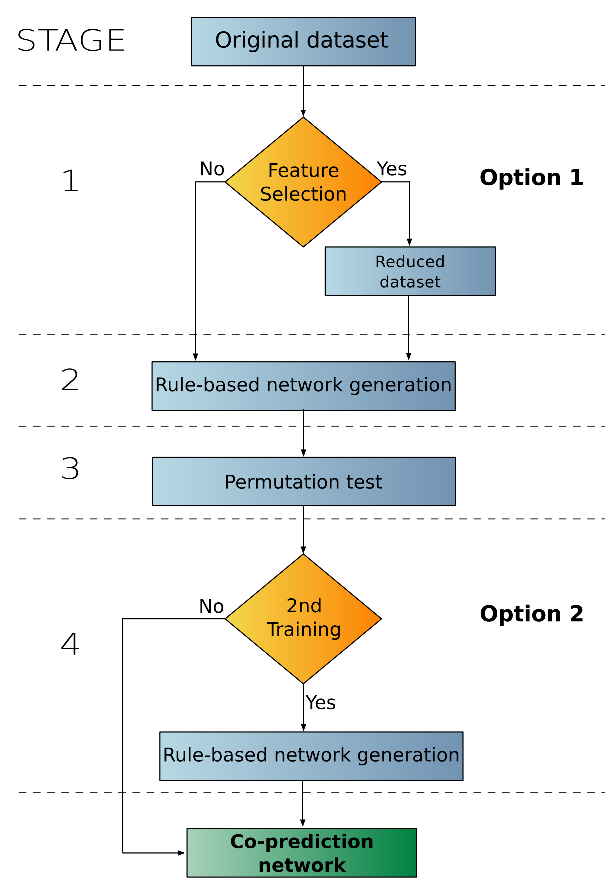


Fig 3 Stages of the functional network inference protocol.

Feature selection (stage 1) When datasets contain a large number of attributes, some might be irrelevant to the prediction target and discarding them helps the classification algorithm to focus its learning effort on the attributes that matters. Therefore, the feature selection is the first stage of the inference process. To pick the relevant attributes we used the support vector machine recursive feature elimination (SVM-RFE) [25]. We opted for the SVM algorithm with a linear kernel as our preliminary studies suggested that it can eliminate as much as 90% of the original dataset attributes, without losing much of the classification accuracy (see Section 1 in Supplementary Material).

Rule-based network inference (stage 2) To infer the rule-based classification models we used BioHEL [24]. It generates sets of classification rules using a genetic algorithm and is able to work with large datasets. Due to the stochastic nature of BioHEL's learning process, each of its runs generates a different rule set. We leverage this fact by creating a large number of alternative hypotheses of functional relationships via multiple runs of the algorithm. For each dataset we run BioHEL 10 000 times and infer the network from the *consensus* of all the generated rule sets. To do that, we use all the pairs of attributes that appear together in the same classification rule as the network edges (co-prediction paradigm). Then, we score each network node (attribute) by counting how many times it has been used in the rules (node score).

Permutation test (stage 3) Given a list of edges (attribute-attribute associations) extracted from the rule sets, we try to filter out the non-significant nodes. To determine the node significance, we follow a statistical analysis procedure based on a permutation test, similar to the one described in [17]. We generate 100 permutated datasets by randomly shuffling the class labels. Next, we infer the co-prediction networks (as in Stage 2) from these permutated datasets. Then, for each node, we calculate a distribution of scores across the 100 networks generated from the permutated datasets. Using a one-tailed permutation test, we assign to each node a p-value, to estimate how likely it is to draw its score from the calculated distribution. With this process we make sure that the nodes with high scores are really tied to the classes present in the data, and that the network truly represents functional relationships. To decide if a node is statistically significant we use a typical $\alpha = 0.05$ threshold.

After preliminary experiments we realised, that using significant nodes alone leads to small and dense networks. To counter that, we relaxed the node pruning to also keep all direct neighbours of the significant nodes.

Network construction (stage 4) There are two ways to interpret the result of the statistical test (option 2 in Figure 3). The first approach is to use the significant nodes as a filter for the inferred relationships (edges) and remove all the edges between two non-significant nodes. The second approach is to use the permutation test as a further feature selection and build a new rule-based machine learning model using only the significant nodes. This second run of the learning algorithm is then focused only on the statistically important genes and creates the final network.

Protocol configurations As a result of two independent optional stages in the FuNeL protocol, there are 4 different configurations that it can run with (see Table 1). We decided to test them all and infer four networks from each dataset, one per configuration.

Table 1 Protocol configurations used in the experiments.

Config.	Description
C_1	reduced dataset + 1 stage of network generation
C_2	original dataset + 1 stage of network generation
C_3	reduced dataset + 2 stages of network generation
C_4	original dataset + 2 stages of network generation

Datasets

Synthetic datasets

To verify if FuNeL is able to correctly identify functional relationships we tested it on a set of synthetic datasets. Although there are several generators that model expression data with genetic relationships, such as GNW used in several DREAM challenges [26], they generate unlabeled samples (without phenotype information, e.g. case vs. control) and the class labels are necessary to perform the supervised learning at the core of FuNeL.

For that reason, we decided to use GAMETES instance generator [27], designed to create genetic datasets with multi-locus disease associations, where no fewer than n loci can predict a phenotype (disease status). GAMETES generates genotype data (rather than gene expression data) based on models with specific genetic constraints, e.g. different heritabilities or frequencies of the SNPs.

To generate the synthetic datasets, we used a set of 2-locus configurations similar to what was employed in a recent work of Li et al. [28] to evaluate permuted random forest networks of gene interactions. Specifically, the genetic models varied in terms of heritability (0.001-0.4) and number of attributes (5-25), with fixed allele frequency of 0.2 and 2000 samples per dataset. For each configuration, we selected from 100 000 random models, two models with extreme value of the ease of detection metric (EDM) (the least and the most difficult). Finally, for each selected model we generated 50 datasets, obtaining 4000 datasets in total.

Real-world datasets

We used 8 publicly available human cancer microarray datasets (see Table 2). These datasets represent a broad range of characteristics in terms of biological information (different types of cancers), number of samples (patients) and number of attributes (genes). For each dataset the attributes were defined by the probes used in the microarray experiment. Generally, a gene can be represented by more than one probe and extra post-processing step is needed to merge the information and generate networks where nodes truly represent genes. We used MADGene [29] to map the Affymetrix probe IDs into HUGO gene IDs, then for all probes mapped to the same gene, we merged the probes and their connections. If a probe was unmapped it was removed from the network.

Table 2 Description of the source datasets used to infer networks.

Name	Attributes	Samples	Class labels
Dlbcl [30]	2647	77	Dlbcl; Follicular lymphoma
CNS [31]	7129	60	Survivor; Failures
Leukemia [32]	7129	72	AML; ALL
Lung-Michigan [33]	7129	96	Tumor; Normal
Lung-Harvard [34]	12534	181	Mesothelioma; ADCA
Prostate [35]	12600	102	Tumor; Normal
AML [36]	12625	54	Remission; Relapse
Colon-Breast [37]	22283	52	Colon cancer; Breast cancer

While in this instance we focused on transcriptomics datasets only, the FuNeL protocol is general and can be applied to other types of biological data too (proteomics, lipidomics, etc.).

Co-expression networks

In this paper we are comparing our FuNeL networks against co-expression networks. The co-expression paradigm identifies similarity of gene expression pattern under different experimental conditions. Co-expression edges are an abstraction of functional relationships between genes and do not represent physical binding as in protein interaction or gene regulatory networks. Two genes are considered to be functionally related (co-expressed), if their transcript levels are similar across a set of samples.

In here we employed three well known methods to infer co-expression networks, each one uses a different metric to assess gene expressions similarity: Pearson correlation coefficient, ARACNE [2] and MIC [38]. In the following subsections we briefly present those methods, for more details check the cited original papers.

Pearson correlation coefficient

Pearson's correlation coefficient (PCC) is a well known measure of linear dependence between two variables.

Applied to gene expression profiles, it measures the similarity in the direction of gene response across samples. Its main disadvantages are the lack of distributional robustness (it assumes data normality) and the sensitivity to outliers. We generated the PCC-based co-expression networks using the *SciPy* Python library [39].

ARACNE: Algorithm for the Reconstruction of Gene Regulatory Networks

The ARACNE method [2] measures the dependence between two gene expression profiles using mutual information. Mutual information $I(X;Y)$ estimates entropy to quantify the amount of information that Y contains about X (measured in bits). In contrast to correlation, it is able to detect non-linear dependencies. ARACNE calculates $I(X;Y)$ for every pair of gene expression profiles X and Y, and applies the data processing inequality to remove the majority of indirect dependencies. For each triplet X, Y and Z the weakest link is removed, e.g. the edge between X and Y is removed if $I(X;Y) \leq \min(I(X;Z),\ M(Z;Y)) - \epsilon$. The tolerance threshold ϵ is used to adjust for the variance of the mutual information estimator. To generate the ARACNE based networks we used the *minet* R package [40] with the following parameters: *mi.empirical* estimator, *equalwidth* distance and $\epsilon = 0$.

MIC: Maximal Information Coefficient

The MIC [38] is a recently proposed measure of the strength of association between two variables, closely related to mutual information. Instead of using a single discretisation strategy to bin the compared variables, it chooses individual bins for each variable, such that value of mutual information $I(X;Y)$ is maximised. Compared to standard estimation of $I(X;Y)$ value used in ARACNE, the optimised estimation provided by MIC is able to detect a wider range of non-linear associations. To generate MIC based networks we used the *minepy* Python library [41] with the following parameters: $\alpha = 0.6$ and $c = 15$.

Inference of the co-expression networks counterparts

To fairly compare the co-prediction and co-expression networks generated from the same data, we had to make sure they match in size. To do that, for every co-prediction network C with m edges and n nodes, we created two co-expression counterparts:

- $SE(C)$: co-expression network with m edges
- $SN(C)$: co-expression network with n nodes

PCC and MIC methods directly compute the pairwise similarity between the gene expressions. Given that, we generated $SE(C)$ using m gene pairs with the highest similarity coefficient. To build $SN(C)$ we used as many top gene pairs as needed, to reach at least n nodes (as we included all pairs tied on the similarity value, sometimes we end up with a few nodes more).

ARACNE uses a pruning procedure and generates a weighted network, not a list of pairwise similarities. When the resulting network was smaller than m edges or n nodes, we increased the default tolerance threshold ϵ to obtain a large enough network. This was the case for the *CNS* ($\epsilon = 0.002$) and the *Dlbcl* datasets ($\epsilon = 0.043$). Then we used the edge weights to select top gene pairs, as in the case of PCC and MIC methods.

Several examples of inferred co-prediction networks and corresponding co-expression networks are visualised in Section 7 of the Supplementary Material and are accompanied, in there, by an initial analysis of selected topological properties in Section 3.

Enrichment analysis

To understand the biological information captured by the generated networks we conducted an enrichment analysis. This is a statistical method of checking whether a set of genes have common characteristics. In our study, the set is defined by the nodes of the generated functional network and is analysed with PANTHER [42]. Because many statistical tests are performed (one for each term) at the same time, PATHER uses Bonferroni correction for multiple testing with $\alpha = 0.05$. We searched for two categories of biological knowledge: Gene Ontology (GO) terms and PANTHER pathways (176 primarily signalling pathways). From the set of GO term, we selected only the manually curated annotations that were supported by experimental evidence.

Disease association analysis

To evaluate the predictive power of the generated networks, and to assess their relevance within a cancer-related context, we analysed the relationships between known disease-associated genes. We used two sources for the disease associations: Malacards (a meta-database of human maladies consolidated from 64 independent sources) [43] and the union of several manually curated databases (OMIM [44], Orphanet [45], Uniprot [46] and CTD [47]). We looked at two properties: (1) the proximity of the disease-associated

Table 3 FuNeL success rate in identification of disease-predicting SNPs. The datasets differed with respect to heritability, number of SNPs and detection difficulty (L-EDM models were the hardest, H-EDM the easiest).

	5 SNP		10 SNP		15 SNP		20 SNP		25 SNP	
Her.	**L-EDM**	**H-EDM**	**L-EDM**	**H-EDM**	**L-EDM**	**H-EDM**	**L-EDM**	**H-EDM**	**L-EDM**	**H-EDM**
0.001	6 %	16 %	8 %	18 %	4 %	10 %	4 %	12 %	12 %	16 %
0.005	8 %	82 %	0 %	86 %	6 %	80 %	2 %	82 %	8 %	72 %
0.01	8 %	96 %	8 %	100 %	8 %	100 %	12 %	100 %	14 %	100 %
0.05	14 %	100 %	60 %	100 %	42 %	100 %	34 %	100 %	34 %	100 %
0.1	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %
0.2	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %
0.3	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %
0.4	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %

genes within a network and (2) the number of triangles in a network, containing one or more disease-associated genes.

Higher proximity represents stronger functional relationship between genes involved in the disease. Triangles represent groups of attributes used together across different prediction rules, and therefore indicate strong mutual relationship between the genes (useful in the discovery of potential new disease associations). Triangles are also the smallest non-trivial motifs that can be found in a complex network and over-represented motifs usually identify functional units of biological processes in cells [48].

The proximity of disease-associated genes was measured using the average shortest path length (SPL). The proximity was defined as a ratio of two distances: average SPL between all pairs of the non-associated genes and average SPL between all pairs of disease-associated genes A:

$$\frac{1}{n}\sum_{i=1}^{n} w_i \frac{\overline{SPL}(CC_i \setminus A)}{\overline{SPL}(A)}, \text{where } w_i = \frac{|CC_i|}{\sum_{j=1}^{n} |CC_j|}$$

As the generated networks often were disconnected (had more than 1 connected component), we introduced a weight w_i that represents the relative size of a connected component CC_i. Components with less than 3 nodes or disease-associated genes were not used in the calculation.

Results

The main results described in this section are based on the analysis of 8 real-world datasets. The only exception is the subsection below, which reports the test results on synthetic datasets.

Identification of predefined relationships in synthetic datasets

To verify how well FuNeL is able to identify functional relationships, we tested it first on synthetic datasets generated using GAMETES. We used 80 different model configurations that varied in heritability, number of SNPs and ease of detection, and tested the success rate on 50 datasets per model. Given the small number of attributes in the synthetic datasets, we used only the C_2 protocol configuration in the tests (no feature selection, single learning phase). The percentage of successfully identified relationships for each model is reported in Table 3. We counted as success the presence of an edge between the interacting pair of SNPs in the inferred network.

As expected, a higher success rate was obtained for models where relationships were easy to detect (H-EDM). The performance increased with higher values of heritability and 100% success rate was obtained for heritability values above 0.05 regardless of model difficulty. The overall results are similar to those reported in [28], or even slightly better, as FuNeL's success rate was unaffected by the increase in the number of SNPs.

Complementarity of the enriched terms

To test how unique are the biological terms (GO terms and pathways) over-represented in the inferred FuNeL networks, we measured an overlap between terms found for each type of network. We defined the overlap between terms enriched for networks inferred using configurations C_a and C_b as:

$$O(C_a, C_b) = \frac{c}{u_a + u_b + c}$$

where c is the number of common terms, u_a is the number of unique terms for C_a and u_b is the number of unique terms for C_b.

Table 4 summaries the pair-wise overlap between the 4 different FuNeL configurations. For GO terms we reported the average overlap between the biological process, cellular component and molecular function categories. Although configurations that operate on the same dataset (C_1/C_3 and C_2/C_4) shared the most terms/pathways, the overlap is quite far from 100%. The observed difference is a result of the second training stage. Configurations used on different datasets (i.e. different set of attributes) resulted in networks sharing less than 40% GO terms and 20% pathways.

Table 4 Average overlap of enriched GO terms and pathways between different FuNeL configurations. The overlap was averaged across all 8 datasets.

	Gene Ontology				Pathways			
	C_1	C_2	C_3	C_4	C_1	C_2	C_3	C_4
C_1	—	0.353	0.749	0.405	—	0.186	0.513	0.183
C_2		—	0.321	0.701		—	0.095	0.591
C_3			—	0.364			—	0.104
C_4				—				—

Similarly, we analysed the term overlap between co-prediction and co-expression by comparing the C_i networks with their co-expression counterparts $SE(C_i)$ and $SN(C_i)$ generated with different approaches (see Table 5). We found the percentage of overlap to be similar across the different inference methods. The overlap in enriched terms was never higher than 62% (still leading to a difference around 40%) and was the largest for configuration not using feature selection (C_2 and C_4). In general the percentages were lower for biological pathways with a minimum of only 10% of shared terms. Low values of terms overlap indicate that the co-prediction and the co-expression approaches can be seen as complementary. Despite starting from the same dataset, they generate networks expressing different biological information.

Table 5 Average overlap of enriched GO terms and pathways between the co-prediction and co-expression networks. Each co-expression network C_i was compared to the corresponding co-expression networks $SE(C_i)$ and $SN(C_i)$. The overlap was averaged across all 8 datasets.

		Co-expression (SE)				Co-expression (SN)			
Method	Cat.	C_1	C_2	C_3	C_4	C_1	C_2	C_3	C_4
PCC	GO	0.280	0.414	0.297	0.432	0.315	0.576	0.367	0.488
	path.	0.223	0.260	0.258	0.190	0.264	0.400	0.175	0.287
ARACNE	GO	0.348	0.621	0.272	0.565	0.333	0.612	0.277	0.535
	path.	0.126	0.463	0.139	0.479	0.085	0.423	0.016	0.356
MIC	GO	0.316	0.513	0.283	0.487	0.300	0.614	0.289	0.527
	path.	0.097	0.339	0.142	0.315	0.112	0.469	0.080	0.352

Quantifying the amount of captured biological knowledge

The amount of biological knowledge (number of enriched terms) captured by a network is related to its size (number of nodes). To fairly compare the networks of different sizes we used the normalised Enrichment Score (ES):

$$ES = \frac{number\ of\ enriched\ terms}{number\ of\ nodes}$$

The score assesses if a network contains biologically related nodes. The higher it is, the larger is a biological similarity between the nodes of a network.

To have a global view of the performances of each inference method in term of ES, we performed a two-step analysis for each enrichment category. First, using the ES, we ranked the networks generated by each method in order to identify the best performing one. See Section 4 of the Supplementary Material for the complete analysis.

Once we identified the best network for each method, we ranked them together by ES and calculated their average rank across the datasets. The results of this analysis are reported in Table 6. MIC performed best when ES was calculated using the GO terms (it was ranked first in each of those categories). When ES was calculated using the biological pathways, C_4 and ARACNE $SE(C_1)$ shared the highest rank.

Table 6 Average ranks based on the Enrichment Score for the best performing networks of each inference method. For each category and for each method, we report the network used in the analysis. The ranks (in brackets) were averaged across all 8 datasets, and the highest ranks are shown with bold font. The last row reports the average ranks across all the biological categories. The following abbreviations were used for GO categories: biological process (BP), molecular function (MF) and cellular component (CC).

Category	FuNeL	PCC	ARACNE	MIC
GO BP	C4 (3)	**SE(C3) (1.5)**	SN(C3) (4)	**SN(C3) (1.5)**
GO MF	C3 (3.5)	SN(C3) (3.5)	SN(C3) (2)	**SN(C3) (1)**
GO CC	C3 (4)	SN(C1) (3)	SN(C3) (2)	**SN(C3) (1)**
Patwhays	**C4 (1.5)**	SN(C2) (3.5)	**SE(C1) (1.5)**	SE(C3) (3.5)
Average	3	2.88	2.38	**1.75**

Table 6 shows that the best performing networks for each method were mostly C_3 co-expression counterparts, in particular $SN(C_3)$. This is consistent with the result of the topological analysis in Section 3 of the Supplementary Material were these networks were found to have the lowest number of nodes, and suggests that smallest networks tend to be more enriched. The difference in performance between the FuNeL configurations is mainly a result of the application of the second machine learning phase (the best networks were C_3 and C_4).

In Supplementary Table 6 we reported the results of a similar analysis where we compared the similarity-based inference methods against FuNeL (ranks in there range from 1 to 12: 4 C_i + 4 $SE(C_i)$ + 4 $SN(C_i)$). In this pairwise analysis, FuNeL networks performed similarly to PCC and ARACNE. We did not observe any consistent winner across all the enrichment categories. MIC seems to have better results than FuNeL only for GO categories, as emerged from Table 6, while FuNeL networks tend to be more enriched for biological pathways.

Evaluation of the networks in a disease context

To verify if the topology of the inferred networks is biologically meaningful, we analysed how it defines the relationships between genes that are known to be associated with a disease targeted by each dataset. We expected the disease-associated genes to be more closely connected than other genes and to be present in functional units, such as triangle motifs. We measured the proximity of the disease-associated genes (i.e. how closely connected they are compared with non-disease-associated genes) and counted the number of triangular relationships present in each network (i.e. the percentage of triangles containing one, two or three disease-associated genes). We repeated the two-step analysis as presented in Section *Quantifying the amount of captured biological knowledge* by using the gene-disease metrics for the ranking. The results are reported in Table 7. The detailed results for each inference method are available in Section 5 of the Supplementary Material.

Table 7 Average ranks based on the disease-associations for the best performing networks of each inference method. For each category and for each method we report the network used for the analysis. The ranks (in brackets) were averaged across all 8 datasets, and the highest ranks are shown with bold font. The last row reports the average ranks across all the categories. The number of disease-associated genes participating in a triangle is denoted as 1A, 2A and 3A.

Source	Cat.	FuNeL	PCC	ARACNE	MIC
Curated	**1A**	**C2 (1)**	SN(C2) (4)	SN(C3) (2.5)	SN(C2) (2.5)
	2A	**C3 (1)**	SN(C3) (2)	SE(C2) (3)	SN(C2) (4)
	3A	C1 (2)	SN(C1) (3)	SE(C4) (4)	**SE(C2) (1)**
	Proximity	**C2 (1)**	SN(C3) (2.5)	SE(C4) (2.5)	SE(C2) (4)
	Average	**1.25**	2.88	3	2.88
Malacards	**1A**	**C2 (1)**	SN(C2) (4)	SN(C4) (3)	SE(C4) (2)
	2A	**C2 (1.5)**	SN(C4) (4)	**SE(C4) (1.5)**	SN(C2) (3)
	3A	C3 (2)	SN(C4) (3)	SE(C2) (4)	**SN(C2) (1)**
	Proximity	**C2 (1)**	SE(C4) (4)	SE(C4) (3)	SE(C2) (2)
	Average	**1.78**	3.75	2.88	2

The average ranks, for both sources of disease associations, suggest that co-prediction outperforms the other inference paradigms. The proximity of the disease-associated genes was in general higher in C_2 network. Therefore, the co-prediction paradigm has identified the core elements of the network more accurately. This result highlights the benefits of including functional information, whenever these are available, in the network inference process (FuNeL is using the class labels assigned to the samples of the dataset), in contrast to the co-expression approach solely based on gene expression similarity (unsupervised).

There is also a clear difference in the number of disease-associated genes participating in the triangles; co-prediction networks were ranked higher than the co-expression networks. The only category in which MIC had a higher rank was $3A$. However, considering that there were not many triangles with disease-associated genes, many ties affected the ranks in this category. Overall, these results demonstrate the higher predictive potential of the FuNeL networks in identifying new disease associations.

Prostate cancer case study: enriched terms

To compare in detail the difference in biological knowledge captured by the co-prediction and co-expression networks, we followed our global analysis with a case study focused on a dataset targeting a single disease — prostate cancer [35]. We were especially interested in specific knowledge captured by one paradigm but not the other.

In Figures 4 and 5 we compared the co-prediction and PCC co-expression networks inferred from the prostate cancer dataset. We focused on unique GO terms and pathways, enriched only in one type of networks. For the sake of readability we filtered out the generic GO terms (with depth $<$ 9 in the GO hierarchical structure). C_2 was the network with the largest number of unique terms, followed by C_4 and $SN(C_2)$. We found 16 GO terms and 21 pathways unique to co-prediction networks and only 3 GO terms and 4 pathways unique to co-expression networks. A similar disproportion in favour of the co-prediction networks was found in comparison with MIC and ARACNE networks (see Supplementary Figures 2 and 3).

We found several of the unique GO terms enriched in the co-prediction networks to be related to prostate cancer. The role of the *Protein ubiquination* in prostate cancer was recently analysed and showed an impact for its treatments [49]. *ERK* pathway is involved in the motility of prostate cancer cells [50]. Prostate cancer cells seems to alter the nature of their *calcium* influx to promote growth and acquire *apoptotic* resistance [51]. Furthermore, the role of *calcium homeostasis* in the majority of the cell-signaling pathways involved in carcinogenesis has been well established, prostate cancer included [52].

A number of enriched pathways specific to co-prediction networks are also highly relevant to the prostate cancer. Several studies demonstrated the involvement of the *JAK/STAT pathway* in the prostate cancer development [53, 54]. There is multiple evidence suggesting that one of the major aging-associated influences on prostate carcinogenesis is *oxidative stress* and its cumulative impact on DNA damage [55, 56]. Finally, *FAS* (also called Apo1 or CD95) plays a central role in the physiological regulation of programmed cell death

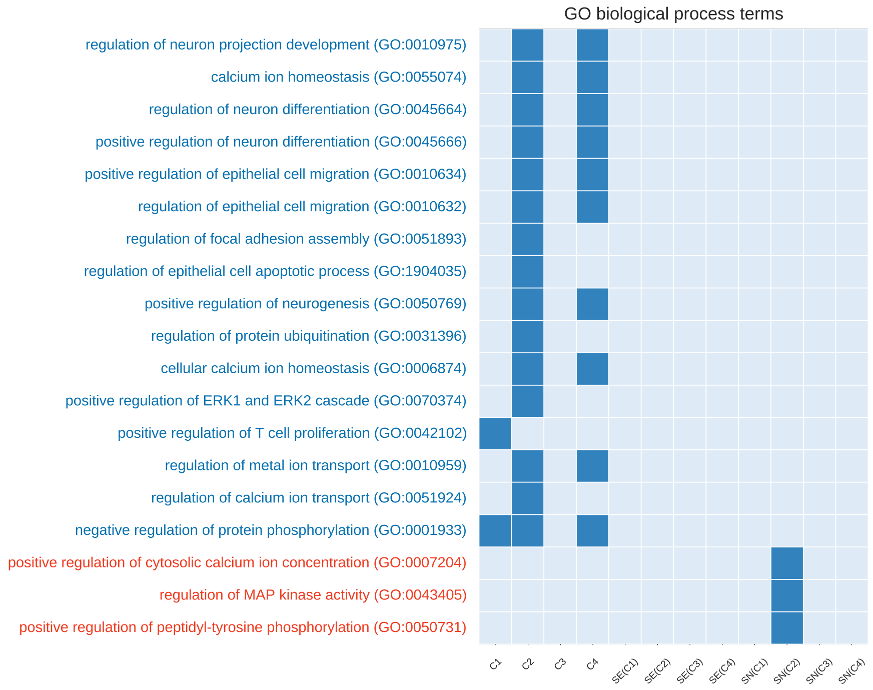


Fig 4 Number of unique enriched GO terms (biological process) for each network configuration (generated from the prostate cancer dataset). On the x-axis we show the 12 investigated networks. On the y-axis we show the names of enriched terms unique to co-prediction or PCC co-expression networks. Red terms are associated with co-expression networks, blue with co-prediction. Empty columns indicate networks with no unique terms.

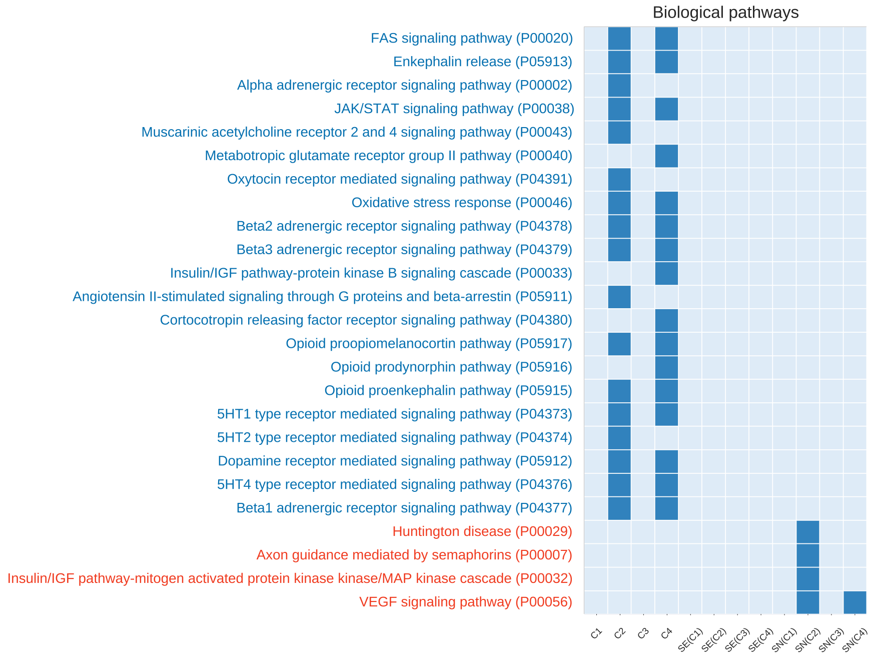


Fig 5 Number of unique enriched biological pathways for each network configuration (generated from the prostate cancer dataset). On the x-axis we show the 12 investigated networks. On the y-axis we show the names of enriched pathways unique to co-prediction or PCC co-expression networks. Red terms are associated with co-expression networks, blue with co-prediction. Empty columns indicate networks with no unique pathways.

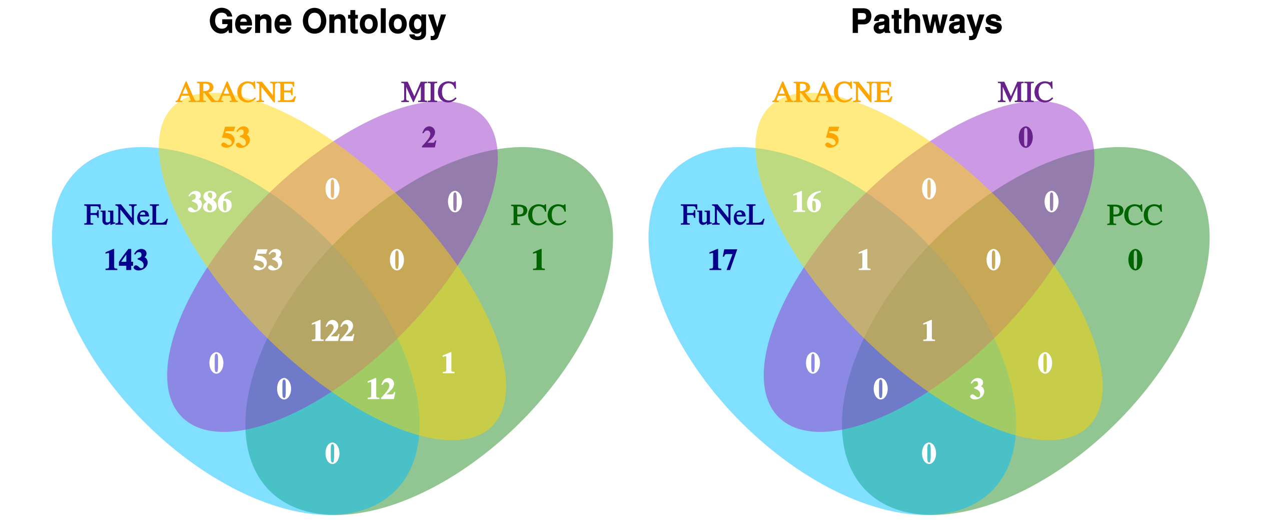


Fig 6 Overlap of enriched terms between the best performing networks in the disease-association analysis (curated databases). On the left the overlap of GO terms (including all 3 categories: BP, CC and MF), on the right the overlap of pathways.

and has been implicated in the pathogenesis of various malignancies and diseases of the immune system including prostate cancer [57].

We also performed an additional analysis of the biological terms related to the hubs (highly connected nodes) of the inferred networks. A node v was considered to be a hub if its degree was at least one standard deviation above the mean network degree. To compare the networks, we used the 10 most frequent Gene Ontology terms (biological processes with at least depth 10) shared among each network's hubs. We found 16 unique terms for co-prediction networks, 19 unique terms for PCC co-expression networks and 11 common terms. These results further highlight that some biological terms are exclusively associated either with co-prediction or co-expression networks. The complete analysis (method by method) is available in the Supplementary Material (Supplementary Figures 4, 5 and 6).

A further analysis of term overlap was conducted using only the best performing networks in the curated disease-association analysis (namely C_2 for FuNeL, $SN(C_3)$ for PCC, $SE(C_4)$ for ARACNE and $SE(C_2)$ for MIC, see Section 5 of the Supplementary Material for details). In Figure 6 we show the overlap of GO terms (including all three GO categories) and pathways across networks from different inference algorithms. In both categories FuNeL had much larger number of unique terms than the co-expression methods and it shared the largest number of terms with ARACNE. In total 122 common GO terms were found between all the methods, while there was only 1 common pathway. Figure 6 further highlights the complementarity between the co-prediction and co-expression approaches in terms of captured biological knowledge.

Prostate cancer case study: disease associations

We searched the literature and the public cancer databases (not used in the inference process), to verify if key nodes in the generated networks are associated with prostate cancer. As a measure of node importance we used the node degree (number of connections) and the betweenness centrality (number of shortest paths between all pair of nodes pass through a given node).

Literature analysis We picked the top 3 most connected nodes (hubs) for each of the four co-prediction networks. The set contained six genes: *GSTM2*, *NELL2*, *CFD*, *PTGDS*, *PAGE4* and *LMO3*. All the genes from this set, except *LMO3*, were also found to be the most central nodes (with highest betweenness centrality).

Almost all these genes are related with prostate cancer:

- *NELL2* contributes to alterations in epithelial-stromal homeostasis in benign prostatic hyperplasia and codes for a novel prostatic growth factor [58], and is also an indicator of expression changes in cancer samples [59],
- *CFD* (adipsin gene) is over expressed in PP periprostatic adipose tissue of prostate cancer patients [60],

- *PTGDS* (and other 2 genes) are expressed at consistently lower levels in clinical prostate cancer tissues and form a signature that predicts biochemical relapse [61],
- *PAGE4* modulates androgen receptor signaling, promoting the progression to advanced lethal prostate cancer [62], and has a significantly lower expression level in patients with prostate recurrent disease [63],
- *LMO3* interacts with *p53*, a well known gene tumour suppressor in prostate cancer [64].

The only gene without literature support was *GSTM2*. It might represent a good target for further experimental verification.

Validation on independent data To further validate the biological significance of the inferred networks, we used an independent prostate cancer dataset [65] from the cBioPortal for Cancer Genomics [66]. We analysed the top 10 hubs (nodes with highest degree) and the top 10 central nodes (with highest betweenness centrality) in the co-prediction network that better performed in the gene-disease association analysis using the curated databases: C_2 (see Supplementary Table 8a). The genes with highest degree were: *PTGDS*, *PAGE4*, *NELL2*, *GSTM2*, *PARM1*, *MAF*, *LMO3*, *COL4A6*, *RBP1* and *ABL1*. For the betweenness centrality, the set was almost identical, only *RBP1* was replaced by *MYH11*. On average the expression in samples was altered in 31.8% cases for hubs and in 35.6% cases for central nodes. The most altered genes were found to be downregulated at the mRNA level: *COL4A6* (65%), *MYH11* (58%), *PARM1* (53%) and *GSTM2* (52%). In addition, genomic alterations in several key genes have been found to be strongly cooccurent (e.g. *PTGDS* – *GSTM2*, *PAGE4* – *COL4A6*, *PAGE4* – *RBP1*, etc.).

When we repeated this analysis for the co-expression networks that were best ranked in the gene-disease analysis using the curated databases ($SN(C_3)$ for PCC, $SE(C_4)$ for ARACNE and $SE(C_2)$ for MIC), we found that on average the alteration level was consistently lower, at most half of the co-prediction key genes. The percentages of alterations are represented as boxplots in Figure 7, while the average alterations are reported in Table 8. As Figure 7 shows, our method is able to identify many more genes with higher percentage of alteration than other methods. Therefore, the topologically important nodes in the best coprediction network represent genes more strongly related to the prostate cancer, with over two times more frequent genomic alterations.

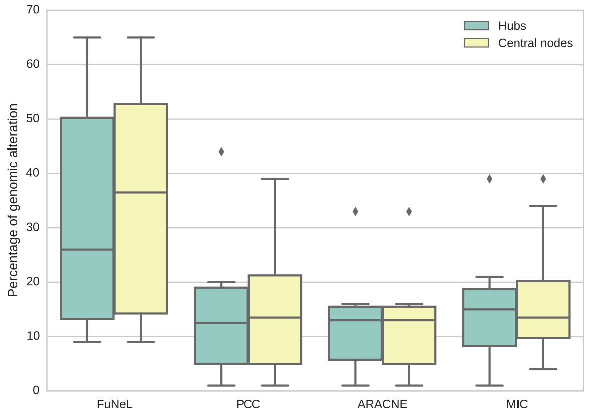


Fig 7 Distribution of the percentage genomic alteration in the samples of an independent dataset for top 10 hubs and central nodes. The topologically important genes were selected from the best performing networks in the disease-association analysis on curated datasets.

Table 8 Average percentage of genomic alteration for top hubs and central nodes in the independent dataset.

Genes	FuNeL	PCC	ARACNE	MIC
Hubs	31.8 %	14.2 %	12.3 %	15.2 %
Central nodes	35.6 %	14.7 %	12.2 %	17.1 %

The detailed list of genomic alterations for top 10 hubs and top 10 central nodes for each analysed network is shown in Section 6 of the Supplementary Material (Figures 7–14).

Discussion

We proposed FuNeL, a protocol to infer functional networks based on the *co-prediction* paradigm where the structure of a rule-based machine learning model (in this paper the rules of a classification algorithm called BioHEL) is used to identify relationships between genes. We tested FuNeL on synthetic datasets and obtained a high success rate in identifying pairwise relationships between attributes. Encouraged by this result, we hypothesised that a rule-based machine learning model, with its complex knowledge representation, might be used to identify biologically meaningful relationships that escape the standard inference methods.

To test this hypothesis, we evaluated 4 different configurations of the inference protocol using 8 cancer-related transcriptomics datasets. We compared FuNeL with other 3 co-expression inference methods by using networks of matching size generated from the same data. We looked at the differences, between co-prediction and co-expression, from three points of view: basic topological properties, enriched biological terms and relationships between known disease-associated genes.

The comparison of networks topology (see Section 3 of the Supplementary Material) revealed the influence of the protocol options. Not surprisingly, both the feature selection and the second training phase reduced the size of the networks, but at the same time, increased the clustering coefficient and the number of connections. The clustering coefficient was found to be lower in almost all the ARACNE networks, probably due to the pruning procedure, it was also lower in many MIC networks. Moreover, when feature selection was applied, the resulting networks had higher clustering coefficient than PCC co-expression networks with the same number of edges. Interestingly, all co-expression networks were less compact, with up to 3 times higher diameter for PCC and ARACNE and up to 7 times higher for MIC.

The differences in networks topology translated to differences in contained biological information. The overlap between enriched GO terms and pathways across protocol configurations was generally low, indicating that different configurations infer networks that capture different biological knowledge. The same terms overlap between the co-prediction networks and their equivalent co-expression counterparts was even lower, never exceeding 62%. We interpret that as evidence, that the biological knowledge captured by the two paradigms is not completely redundant, but in a large part complementary.

The most apparent differences between the networks were observed during the analysis of the connections between genes known to be related to a specific disease. The disease-associated genes were more closely connected (higher proximity) in the co-prediction networks, which means that the disease-related nodes of the network were closer to its core. We also found that the number of functional units (triangle motifs), that can identify new gene-disease associations, was higher in the co-prediction networks. Therefore, we conclude that the co-prediction networks better capture the abstract concept of functional relationship.

The prostate cancer case study further confirmed this conclusion. We found enriched GO terms and biological pathways, unique to the co-prediction networks, to be reported in the literature as related to prostate cancer. Furthermore, FuNeL generated networks enriched with knowledge totally missed by all the co-expression networks when using the prostate cancer dataset. We also found that genes corresponding to the topologically important nodes in the co-prediction networks: (1) were altered in a high percentage of tumour samples in an independent cancer transcriptomic study, and (2) were already associated with prostate cancer according to the specialised literature. Therefore, the co-prediction networks not only capture biological knowledge complementary to the co-expression networks, but also highlights better the important genes involved in the disease process.

The superior performance of FuNeL networks in identifying the disease-associated genes is likely a result of effective use of the class labels of the samples, which the similarity-based methods ignore. Although it would be tempting to attribute this performance difference entirely to the use of supervised learning in FuNeL, it would be an overstatement, as the knowledge of explicit links between genes and diseases is not available to it in training. Our hypothesis is that this is rather a result of differences in expression values of the disease-associated genes, which taken together are able to discriminate between sample phenotypes.

Given that our co-prediction networks were found to be not only biologically meaningful, but also complementary to similarity-based functional networks, we believe that network inference based on machine learning models deserves to be studied in more detail in the future. In here we only touched the subject of feature selection and network post-processing, and although we now know they indeed influence the network topology and its biological interpretation, there are many strategies to choose from in that respect.

At the same time, the machine learning step in the FuNeL protocol does not have to be limited to the rule-based machine learning methods. We can imagine unsupervised methods, such as the Apriori algorithm for association rule learning, or other supervised methods, such as decision tree algorithms (e.g. C4.5 or random forest), replacing BioHEL in the FuNeL protocol. Some adjustment would be necessary to extract the knowledge from a different model representation, but the rest of the protocol could remain unchanged. For example in the case of the decision trees, relationships could be inferred between attributes that share the same path from the root to the leaves of a tree. This potential flexibility in the choice of a learning algorithm, together with the ability to apply the protocol to different types of data, becomes important in the context of results correctness. As has been discussed to a great length in [67], when methods or data used in the network inference process are tightly controlled, some results will replicate more easily than others not because they are correct, but due to a replicable bias. Therefore a diversity in methods and data is a necessary condition to be able to converge on the scientific truth.

Finally, in terms of testing new functional networks, there is a limit of how thorough and complete a manual literature analysis can be, which leads to a great need of synthetic or experimentally validated benchmarks, similar to those proposed for protein-protein interaction networks or gene regulatory networks. Although we understand that this would be a difficult and challenging task, we see this as a necessary step on the way to refining the functional inference methods.

Conclusions

We presented FuNeL: a protocol for the inference of functional networks from rule-based machine learning models. FuNeL is based on the co-prediction paradigm, which hypothesises that genes used together with a rule-based machine learning model, are more likely to be functionally related. We verified that FuNeL correctly identifies relationships in synthetic datasets and we thoroughly compared FuNeL to three co-expression inference methods: PCC, ARACNE and MIC, on 8 real-world datasets. We contrasted the different approaches by looking at the inferred networks topology, enriched biological terms and the relationships between genes associated with cancer. We found that FuNeL networks capture relevant biological knowledge that is complementary to what is captured by the co-expression approaches, and demonstrated that FuNeL networks are better at identifying relationships between genes with known disease associations.

Availability of data and material

The datasets used for the analysis were collected from the following public domain resources:

Dlbcl	http://ico2s.org/datasets/microarray.html
CNS	http://datam.i2r.a-star.edu.sg/datasets/krbd/NervousSystem/NervousSystem.html
Leukemia	http://datam.i2r.a-star.edu.sg/datasets/krbd/Leukemia/ALLAML.html
Lung-Michigan	http://datam.i2r.a-star.edu.sg/datasets/krbd/LungCancer/LungCancer-Michigan.html
Lung-Harvard	http://datam.i2r.a-star.edu.sg/datasets/krbd/LungCancer/LungCancer-Harvard2.html
Prostate	http://datam.i2r.a-star.edu.sg/datasets/krbd/ProstateCancer/ProstateCancer.html
AML	http://www.biolab.si/supp/bi-cancer/projections/info/AMLGSE2191.html
Colon-Breast	http://www.biolab.si/supp/bi-cancer/projections/info/BC_CCGSE3726_frozen.html

The FuNeL source code is publicly available:

Project name:	FuNeL
Project home page:	http://ico2s.org/software/funel.html
Operating system(s):	GNU/Linux
Programming language:	Python, R
License:	GNU GPLv3

Author's contributions

NL, PW, NK and JB designed the experiments. NL and PW performed the experiments. NL, PW, NK and JB analysed the data. RH and SW contributed to the biological validation of the results. NL, PW, NK and JB wrote the paper. All authors read and approved the final manuscript.

Acknowledgements

We thank Joseph Mullen for the help provided in finding the gene-disease associations. This work made use of the facilities of N8 HPC Centre of Excellence, provided and funded by the N8 consortium and EPSRC [EP/K000225/1]. The Centre is co-ordinated by the Universities of Leeds and Manchester.

Funding

This work was supported by the Engineering and Physical Sciences Research Council [EP/L001489/2, EP/J004111/2, EP/I031642/2, EP/N031962/1]. The funding agency was not involved with the design of the study, analysis and interpretation of data or in the writing of the manuscript.

Author details

[1] Interdisciplinary Computing and Complex BioSystems (ICOS) research group, School of Computing Science, Newcastle University, Newcastle upon Tyne, UK. [2] Clinical and Experimental Pharmacology Group, Cancer Research UK Manchester Institute, University of Manchester, Manchester, UK. [3] Northern Institute for Cancer Research, Medical School, Newcastle University, Newcastle upon Tyne, UK.

References

1. Bansal, M., Belcastro, V., Ambesi-Impiombato, A., di Bernardo, D.: How to infer gene networks from expression profiles. Molecular Systems Biology **3**(1) (2007). doi:10.1038/msb4100120
2. Margolin, A.A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., Califano, A.: ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics **7**(Suppl 1), 1–15 (2006). doi:10.1186/1471-2105-7-S1-S7
3. Barzel, B., Barabási, A.-L.: Network link prediction by global silencing of indirect correlations. Nature biotechnology **31**(8), 720–725 (2013). doi:10.1038/nbt.2601
4. Huynh-Thu, V.A., Irrthum, A., Wehenkel, L., Geurts, P.: Inferring regulatory networks from expression data using tree-based methods. PLoS ONE **5**(9), 1–10 (2010). doi:10.1371/journal.pone.0012776
5. Childs, K.L., Davidson, R.M., Buell, C.R.: Gene Coexpression Network Analysis as a Source of Functional Annotation for Rice Genes. PLoS ONE **6**(7), 22196 (2011). doi:10.1371/journal.pone.0022196
6. Presson, A.P., Sobel, E.M., Papp, J.C., Suarez, C.J., Whistler, T., Rajeevan, M.S., Vernon, S.D., Horvath, S.: Integrated weighted gene co-expression network analysis with an application to chronic fatigue syndrome. BMC Systems Biology **2**(1), 1–12 (2008). doi:10.1186/1752-0509-2-95
7. Ray, M., Jianhua, R., Weixiong, Z.: Variations in the transcriptome of Alzheimer's disease reveal molecular networks involved in cardiovascular diseases. Genome Biology **9**(10), 148 (2008). doi:10.1186/gb-2008-9-10-r148
8. Ransbotyn, V., Yeger-Lotem, E., Basha, O., Acuna, T., Verduyn, C., Gordon, M., Chalifa-Caspi, V., Hannah, M.A., Barak, S.: A combination of gene expression ranking and co-expression network analysis increases discovery rate in large-scale mutant screens for novel Arabidopsis thaliana abiotic stress genes. Plant Biotechnology Journal **13**(4), 501–513 (2015). doi:10.1111/pbi.12274
9. Yang, Y., Han, L., Yuan, Y., Li, J., Hei, N., Liang, H.: Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. Nature Communications **5** (2014). doi:10.1038/ncomms4231
10. Kommadath, A., Bao, H., Arantes, A.S., Plastow, G.S., Tuggle, C.K., Bearson, S.M.D., Luo Guan, L., Stothard, P.: Gene co-expression network analysis identifies porcine genes associated with variation in Salmonella shedding. BMC Genomics **15**(1), 1–15 (2014). doi:10.1186/1471-2164-15-452
11. Wei, S.-N., Zhao, W.-J., Zeng, X.-J., Kang, Y.-M., Du, J., Li, H.-H.: Microarray and co-expression network analysis of genes associated with acute doxorubicin cardiomyopathy in mice. Cardiovascular Toxicology **15**(4), 377–393 (2015). doi:10.1007/s12012-014-9306-7

12. Silva, A.T., Ribone, P.A., Chan, R.L., Ligterink, W., Hilhorst, H.W.M.: A predictive co-expression network identifies novel genes controlling the seed-to-seedling phase transition in Arabidopsis thaliana. Plant Physiology **170**(4), 2218–2231 (2016). doi:10.1104/pp.15.01704
13. Mordelet, F., Vert, J.-P.: ProDiGe: Prioritization Of Disease Genes with multitask machine learning from positive and unlabeled examples. BMC Bioinformatics **12**(1), 1–15 (2011). doi:10.1186/1471-2105-12-389
14. Martínez-Ballesteros, M., Nepomuceno-Chamorro, I.A., Riquelme, J.C.: Inferring gene-gene associations from Quantitative Association Rules. In: ISDA, pp. 1241–1246 (2011). doi:10.1109/ISDA.2011.6121829
15. Nepomuceno-Chamorro, I.A., Aguilar-Ruiz, J.S., Riquelme, J.C.: Inferring gene regression networks with model trees. BMC bioinformatics **11**(1), 517 (2010). doi:10.1186/1471-2105-11-517
16. Yoshida, M., Koike, A.: SNPInterForest: a new method for detecting epistatic interactions. BMC bioinformatics **12**(1), 469 (2011). doi:10.1186/1471-2105-12-469
17. Urbanowicz, R.J., Granizo-Mackenzie, A., Moore, J.H.: An Analysis Pipeline with Statistical and Visualization-Guided Knowledge Discovery for Michigan-Style Learning Classifier Systems. IEEE Comp. Int. Mag. **7**(4), 35–45 (2012). doi:10.1109/MCI.2012.2215124
18. Urbanowicz, R.J., Andrew, A.S., Karagas, M.R., Moore, J.H.: Role of genetic heterogeneity and epistasis in bladder cancer susceptibility and outcome: a learning classifier system approach. Journal of the American Medical Informatics Association : JAMIA **20**(4), 603–612 (2013). doi:10.1136/amiajnl-2012-001574
19. Bassel, G.W., Glaab, E., Marquez, J., Holdsworth, M.J., Bacardit, J.: Functional Network Construction in Arabidopsis Using Rule-Based Machine Learning on Large-Scale Data Sets. The Plant Cell Online **23**(9), 3101–3116 (2011). doi:10.1105/tpc.111.088153
20. Glaab, E., Bacardit, J., Garibaldi, J.M., Krasnogor, N.: Using Rule-Based Machine Learning for Candidate Disease Gene Prioritization and Sample Classification of Cancer Gene Expression Data. PLoS ONE **7**(7), 39932 (2012). doi:10.1371/journal.pone.0039932
21. Swan, A.L., Hillier, K.L., Smith, J.R., Allaway, D., Liddell, S., Bacardit, J., Mobasheri, A.: Analysis of mass spectrometry data from the secretome of an explant model of articular cartilage exposed to pro-inflammatory and anti-inflammatory stimuli using machine learning. BMC musculoskeletal disorders **14**(1), 349 (2013). doi:10.1186/1471-2474-14-349
22. Fainberg, H.P., Bodley, K., Bacardit, J., Li, D., Wessely, F., Mongan, N.P., Symonds, M.E., Clarke, L., Mostyn, A.: Reduced Neonatal Mortality in Meishan Piglets: A Role for Hepatic Fatty Acids? PLoS ONE **7**(11), 49101 (2012). doi:10.1371/journal.pone.0049101
23. Bacardit, J., Widera, P., Marquez-Chamorro, A., Divina, F., Aguilar-Ruiz, J.S., Krasnogor, N.: Contact map prediction using a large-scale ensemble of rule sets and the fusion of multiple predicted structural features. Bioinformatics **28**(19), 2441–2448 (2012). doi:10.1093/bioinformatics/bts472
24. Bacardit, J., Burke, E.K., Krasnogor, N.: Improving the scalability of rule-based evolutionary learning. Memetic Computing **1**(1), 55–67 (2009). doi:10.1007/s12293-008-0005-4
25. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using Support Vector Machines. Machine Learning **46**(1-3), 389–422 (2002). doi:10.1023/A:1012487302797
26. Schaffter, T., Marbach, D., Floreano, D.: GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods. Bioinformatics **27**(16), 2263–2270 (2011). doi:10.1093/bioinformatics/btr373
27. Urbanowicz, R.J., Kiralis, J., Sinnott-Armstrong, N.A., Heberling, T., Fisher, J.M., Moore, J.H.: GAMETES: a fast, direct algorithm for generating pure, strict, epistatic models with random architectures. BioData Mining **5**(1), 1–14 (2012). doi:10.1186/1756-0381-5-16
28. Li, J., Malley, J.D., Andrew, A.S., Karagas, M.R., Moore, J.H.: Detecting gene-gene interactions using a permutation-based random forest method. BioData Mining **9**(1), 1–17 (2016). doi:10.1186/s13040-016-0093-5
29. Baron, D., Bihouee, A., Teusan, R., Dubois, E., Savagner, F., Steenman, M., Houlgatte, R., Ramstein, G.: MADGene: retrieval and processing of gene identifier lists for the analysis of heterogeneous microarray datasets. Bioinformatics **27**(5), 725–726 (2011). doi:10.1093/bioinformatics/btq710
30. Shipp, M.A., Ross, K.N., Tamayo, P., Weng, A.P., Kutok, J.L., Aguiar, R.C., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G.S., Ray, T.S., Koval, M.A., Last, K.W., Norton, A., Lister, T.A., Mesirov, J., Neuberg, D.S., Lander, E.S., Aster, J.C., Golub, T.R.: Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nature medicine **8**(1), 68–74 (2002). doi:10.1038/nm0102-68
31. Pomeroy, S.L., Tamayo, P., Gaasenbeek, M., Sturla, L.M., Angelo, M., McLaughlin, M.E., Kim, J.Y.H., Goumnerova, L.C., Black, P.M., Lau, C., Allen, J.C., Zagzag, D., Olson, J.M., Curran, T., Wetmore, C., Biegel, J.A., Poggio, T., Mukherjee, S., Rifkin, R., Califano, A., Stolovitzky, G., Louis, D.N., Mesirov, J.P., Lander, E.S., Golub, T.R.: Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature **415**(6870), 436–442 (2002). doi:10.1038/415436a
32. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S.: Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science **286**(5439), 531–537 (1999). doi:10.1126/science.286.5439.531
33. Beer, D.G., Kardia, S.L., Huang, C.-C.C., Giordano, T.J., Levin, A.M., Misek, D.E., Lin, L., Chen, G., Gharib, T.G., Thomas, D.G., Lizyness, M.L., Kuick, R., Hayasaka, S., Taylor, J.M., Iannettoni, M.D., Orringer, M.B., Hanash, S.: Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nature medicine **8**(8), 816–824 (2002). doi:10.1038/nm733
34. Gordon, G.J., Jensen, R.V., Hsiao, L.-L., Gullans, S.R., Blumenstock, J.E., Ramaswamy, S., Richards, W.G., Sugarbaker, D.J., Bueno, R.: Translation of Microarray Data into Clinically Relevant Cancer Diagnostic Tests Using Gene Expression Ratios in Lung Cancer and Mesothelioma. Cancer Research **62**(17), 4963–4967 (2002)
35. Singh, D., Febbo, P.G., Ross, K., Jackson, D.G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A.A., D'Amico, A.V., Richie, J.P., Lander, E.S., Loda, M., Kantoff, P.W., Golub, T.R., Sellers, W.R.: Gene expression correlates of clinical prostate cancer behavior. Cancer Cell **1**(2), 203–209 (2002). doi:10.1016/S1535-6108(02)00030-2
36. Yagi, T., Morimoto, A., Eguchi, M., Hibi, S., Sako, M., Ishii, E., Mizutani, S., Imashuku, S., Ohki, M., Ichikawa, H.: Identification of a gene expression signature associated with pediatric AML prognosis. Blood **102**(5), 1849–1856 (2003). doi:10.1182/blood-2003-02-0578
37. Chowdary, D., Lathrop, J., Skelton, J., Curtin, K., Briggs, T., Zhang, Y., Yu, J., Wang, Y., Mazumder, A.: Prognostic gene expression signatures can be measured in tissues collected in RNAlater preservative. The Journal of Molecular Diagnostics **8**(1), 31–39 (2006). doi:10.2353/jmoldx.2006.050056
38. Reshef, D.N., Reshef, Y.A., Finucane, H.K., Grossman, S.R., McVean, G., Turnbaugh, P.J., Lander, E.S., Mitzenmacher, M., Sabeti, P.C.: Detecting novel associations in large data sets. Science **334**(6062), 1518–1524 (2011). doi:10.1126/science.1205438
39. Jones, E., Oliphant, T., Peterson, P., et al.: SciPy: Open source scientific tools for Python (2001–). http://www.scipy.org/
40. Meyer, P.E., Lafitte, F., Bontempi, G.: minet: A R/Bioconductor package for inferring large transcriptional networks using mutual information. BMC Bioinformatics **9**(1), 1–10 (2008). doi:10.1186/1471-2105-9-461
41. Albanese, D., Filosi, M., Visintainer, R., Riccadonna, S., Jurman, G., Furlanello, C.: minerva and minepy: a C engine for the MINE suite and its R, Python and MATLAB wrappers. Bioinformatics **29**(3), 407–408 (2013). doi:10.1093/bioinformatics/bts707
42. Thomas, P.D., Campbell, M.J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A., Narechania, A.: PANTHER: A library of protein families and subfamilies indexed by function. Genome Research **13**(9), 2129–2141 (2003). doi:10.1101/gr.772403
43. Rappaport, N., Nativ, N., Stelzer, G., Twik, M., Guan-Golan, Y., Iny Stein, T., Bahir, I., Belinky, F., Morrey, C.P., Safran, M., Lancet, D.: MalaCards: an integrated compendium for diseases and their annotation. Database **2013** (2013). doi:10.1093/database/bat018

44. Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A., McKusick, V.A.: Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Research **33**(suppl 1), 514–517 (2005). doi:10.1093/nar/gki033
45. INSERM: Orphanet: an online database of rare diseases and orphan drugs (1997). `http://www.orpha.net/`
46. Magrane, M., Consortium, U.: UniProt Knowledgebase: a hub of integrated protein data. Database **2011** (2011). doi:10.1093/database/bar009
47. Davis, A.P., Grondin, C.J., Lennon-Hopkins, K., Saraceni-Richards, C., Sciaky, D., King, B.L., Wiegers, T.C., Mattingly, C.J.: The Comparative Toxicogenomics Database's 10th year anniversary: update 2015. Nucleic Acids Research **43**(D1), 914–920 (2015). doi:10.1093/nar/gku935
48. Tran, N.H., Choi, K.P., Zhang, L.: Counting motifs in the human interactome. Nature communications **4** (2013). doi:10.1038/ncomms3241
49. Chen, Z., Lu, W.: Roles of Ubiquitination and SUMOylation on Prostate Cancer: Mechanisms and Clinical Implications. International Journal of Molecular Sciences **16**(3), 4560–4580 (2015). doi:10.3390/ijms16034560
50. McCubrey, J.A., Steelman, L.S., Chappell, W.H., Abrams, S.L., Wong, E.W.T., Chang, F., Lehmann, B., Terrian, D.M., Milella, M., Tafuri, A., Stivala, F., Libra, M., Basecke, J., Evangelisti, C., Martelli, A.M., Franklin, R.A.: Roles of the Raf/MEK/ERK pathway in cell growth, malignant transformation and drug resistance. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research **1773**(8), 1263–1284 (2007). doi:10.1016/j.bbamcr.2006.10.001. Mitogen-Activated Protein Kinases: New Insights on Regulation, Function and Role in Human Disease
51. Monteith, G.R.: Prostate cancer cells alter the nature of their calcium influx to promote growth and acquire apoptotic resistance. Cancer cell **26**(1), 19–32 (2014). doi:10.1016/j.ccr.2014.06.015
52. Flourakis, M., Prevarskaya, N.: Insights into Ca2+ homeostasis of advanced prostate cancer cells. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research **1793**(6), 1105–1109 (2009). doi:10.1016/j.bbamcr.2009.01.009. 10th European Symposium on Calcium
53. Kwon, E.M., Holt, S.K., Fu, R., Kolb, S., Williams, G., Stanford, J.L., Ostrander, E.A.: Androgen metabolism and JAK/STAT pathway genes and prostate cancer risk. Cancer Epidemiology **36**(4), 347–353 (2012). doi:10.1016/j.canep.2012.04.002
54. Barton, B.E., Karras, J.G., Murphy, T.F., Barton, A., Huang, H.F.-S.: Signal transducer and activator of transcription 3 (STAT3) activation in prostate cancer: Direct STAT3 inhibition induces apoptosis in prostate cancer lines. Molecular Cancer Therapeutics **3**(1), 11–20 (2004)
55. Minelli, A., Bellezza, I., Conte, C., Culig, Z.: Oxidative stress-related aging: A role for prostate cancer? Biochimica et Biophysica Acta (BBA) - Reviews on Cancer **1795**(2), 83–91 (2009). doi:10.1016/j.bbcan.2008.11.001
56. Khandrika, L., Kumar, B., Koul, S., Maroni, P., Koul, H.K.: Oxidative stress in prostate cancer. Cancer Letters **282**(2), 125–136 (2009). doi:10.1016/j.canlet.2008.12.011
57. Drewa, T., Wolski, Z., Skok, Z., Czajkowski, R., Wiśniewska, H.: The FAS-related apoptosis signaling pathway in the prostate intraepithelial neoplasia and cancer lesions. Acta Poloniae Pharmaceutica **63**(4), 311–315 (2006)
58. DiLella, A.G., Toner, T.J., Austin, C.P., Connolly, B.M.: Identification of Genes Differentially Expressed in Benign Prostatic Hyperplasia. Journal of Histochemistry and Cytochemistry **49**(5), 669–670 (2001). doi:10.1177/002215540104900517
59. Luo, J., Dunn, T.A., Ewing, C.M., Walsh, P.C., Isaacs, W.B.: Decreased gene expression of steroid 5 alpha-reductase 2 in human prostate cancer: Implications for finasteride therapy of prostate carcinoma. The Prostate **57**(2), 134–139 (2003). doi:10.1002/pros.10284
60. Ribeiro, R., Monteiro, C., Silvestre, R., Castela, A., Coutinho, H., Fraga, A., Príncipe, P., Lobato, C., Costa, C., Cordeiro-da-Silva, A., Lopes, J.M., Lopes, C., Medeiros, R.: Human periprostatic white adipose tissue is rich in stromal progenitor cells and a potential source of prostate tumor stroma. Experimental Biology and Medicine **237**(10), 1155–1162 (2012). doi:10.1258/ebm.2012.012131
61. Thompson, V.C., Day, T.K., Bianco-Miotto, T., Selth, L.A., Han, G., Thomas, M., Buchanan, G., Scher, H.I., Nelson, C.C., Greenberg, N.M., Butler, L.M., Tilley, W.D.: A gene signature identified using a mouse model of androgen receptor-dependent prostate cancer predicts biochemical relapse in human disease. Int J Cancer **131**(3), 662–672 (2012). doi:10.1002/ijc.26414
62. Sampson, N., Ruiz, C., Zenzmaier, C., Bubendorf, L., Berger, P.: PAGE4 Positivity Is Associated with Attenuated AR Signaling and Predicts Patient Survival in Hormone-Naive Prostate Cancer. The American Journal of Pathology **181**(4), 1443–1454 (2012). doi:10.1016/j.ajpath.2012.06.040
63. Shiraishi, T., Terada, N., Zeng, Y., Suyama, T., Luo, J., Trock, B., Kulkarni, P., Getzenberg, R.: Cancer/Testis antigens as potential predictors of biochemical recurrence of prostate cancer following radical prostatectomy. Journal of Translational Medicine **9**(1), 153 (2011). doi:10.1186/1479-5876-9-153
64. Larsen, S., Yokochi, T., Isogai, E., Nakamura, Y., Ozaki, T., Nakagawara, A.: LMO3 interacts with p53 and inhibits its transcriptional activity. Biochemical and Biophysical Research Communications **392**(3), 252–257 (2010). doi:10.1016/j.bbrc.2009.12.010
65. Taylor, B.S., Schultz, N., Hieronymus, H., Gopalan, A., Xiao, Y., Carver, B.S., Arora, V.K., Kaushik, P., Cerami, E., Reva, B., Antipin, Y., Mitsiades, N., Landers, T., Dolgalev, I., Major, J.E., Wilson, M., Socci, N.D., Lash, A.E., Heguy, A., Eastham, J.A., Scher, H.I., Reuter, V.E., Scardino, P.T., Sander, C., Sawyers, C.L., Gerald, W.L.: Integrative Genomic Profiling of Human Prostate Cancer. Cancer Cell **18**(1), 11–22 (2010). doi:10.1016/j.ccr.2010.05.026
66. Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksoy, B.A., Jacobsen, A., Byrne, C.J., Heuer, M.L., Larsson, E., Antipin, Y., Reva, B., Goldberg, A.P., Sander, C., Schultz, N.: The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. Cancer Discovery **2**(5), 401–404 (2012). doi:10.1158/2159-8290.CD-12-0095
67. Verleyen, W., Ballouz, S., Gillis, J.: Positive and negative forms of replicability in gene network analysis. Bioinformatics **32**(7), 1065–1073 (2016). doi:10.1093/bioinformatics/btv734

Supplementary Material

Functional networks inference from machine learning models

Nicola Lazzarini, Paweł Widera, Stuart Williamson, Rakesh Heer, Natalio Krasnogor and Jaume Bacardit

1 Classification accuracy under feature selection

To choose the default percentage of attributes retained in the feature selection procedure, we performed a preliminary analysis using all 8 transcriptomic datasets from the main article. We evaluated how the BioHEL classification accuracy changes with the number of selected features (using linear SVM-RFE). The accuracy was measured using a standard 10 cross-fold validation. The full experiment (not reported here) used $100\%, 90\%, 80\%, ..., 10\%$ of the original dataset attributes.

We found that even when only 10% of the attributes are retained, the classification accuracy remains almost unchanged. Specifically, with 10% of the original attributes the accuracy increased for 2 datasets, slightly decreased for 3 datasets and remained unaltered for the other datasets. The exact results are reported in Supplementary Table 1 below:

Supplementary Table 1: BioHEL classification accuracy for each dataset, in 10-fold cross-validation experiments on the original and reduced set of attributes (before and after the feature selection). Linear SVM-RFE was used to select best 10% of the attributes.

dataset	**all attributes**	**10% attributes**
Dlbcl	0.871	0.886
CNS	0.473	0.451
Leukemia	0.945	0.945
Lung-Michigan	0.980	0.980
Lung-Harvard	0.978	0.964
Prostate	0.892	0.892
AML	0.637	0.592
Colon-Breast	0.903	0.940

Given that we were able to maintain good classification accuracy despite large reduction in number of used attributes, we decided to use 10% attributes as a default setting for the FuNeL feature selection procedure. However, this FuNeL parameter is under the user control and the default setting can be changed.

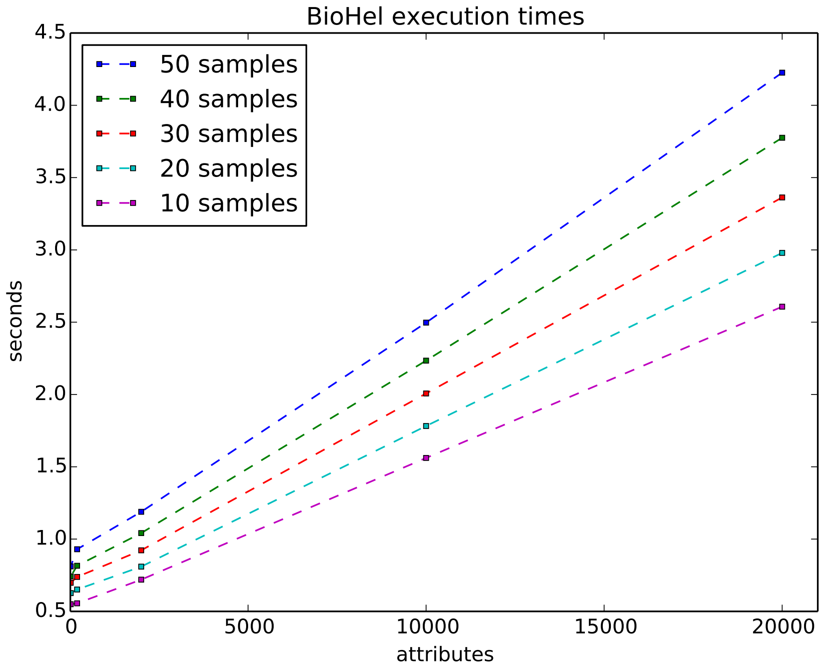
2 Time complexity

The FuNeL protocol has four stages (see Figure 2 in the main article): (1) feature selection (optional), (2) rule-based network generation, (3) permutation test and (4) second rule-based network generation (optional).

The running time for the whole pipeline depends on the rule set generation time (execution time of BioHEL), as the optional feature selection stage can be seen as running in constant time. Two main factors that influence the rule set generation time are: (1) the **number of attributes** and (2) the **number of samples**.

We performed an execution time analysis of BioHEL using the largest (in terms of number of attributes) Colon-Breast dataset (Chowdary *et al.*, 2006). In the feature selection stage we retained: 20, 200, 2000, 10 000 and 20 000 attributes. From each of these 5 datasets we generated 100 random subsets of 50, 40, 30, 20 and 10 samples. Finally, we ran BioHEL 1000 times to obtain 1000 rule sets for each dataset.

Supplementary Figure 1 shows the running times averaged across 100 000 runs (1000 runs for each of the 100 datasets).



Supplementary Figure 1: Average execution times of a single BioHEL run for a given number of samples and attributes.

The total execution time of FuNeL configurations C_1 and C_2 is calculated as:

$$T_1 = (rule\_sets \times t(atts_1, samples)) + (permutation\_runs \times t(atts_1, samples)) \tag{1}$$

where $rule\_sets$ is the number of inferred rule sets, $permutation\_runs$ is the number of randomised datasets used in the permutation test and $t(atts_1, samples)$ represents execution time of a single BioHEL run, that linearly depends on the size of a dataset measured in number of attributes and samples.

Configurations C_3 and C_4 require an additional run of BioHEL (step 4), and their total execution time is:

$$T_2 = T_1 + (rule\_sets \times t(atts_2, samples)) \tag{2}$$

where $atts_2$ is the number of attributes after the permutation test ($atts_1 \leq atts_2$).

It is important to notice that each run of BioHEL is independent, thus the generation of the rule sets can be trivially parallelised without any extra overhead. Given n computational cores, the total execution times could be reduced to:

$$T_{real_1} = \frac{T_1}{n} \qquad T_{real_2} = \frac{T_2}{n} \tag{3}$$

3 Comparison of networks topological properties

The network topology refers to the spatial arrangements of its elements. The analysis of topological properties tells us how different nodes are connected to each other and how their communication paths look like. There are many aspects and characteristics that can be evaluated in a network. For simplicity we report just four metrics: number of nodes, number of edges, clustering coefficient and diameter.

The *clustering coefficient* is a measure of degree to which nodes in a network tend to cluster together. It expresses the likelihood that any two nodes with a common neighbour are themselves connected. The *diameter* indicates the maximum distance between two nodes in the network.

We compared the topology of the networks built with two different approaches: co-prediction and co-expression. For each generated network we calculated the topological properties described above. We compare the FuNeL networks with co-expression networks inferred with different methods. The Supplementary Tables 2 to 4 below show the results for the PCC, ARACNE and MIC networks.

Supplementary Table 2: Topological properties for co-prediction and PCC co-expression networks generated for all 8 datasets.

		Co-prediction				**Co-expression (SE)**				**Co-expression (SN)**			
Dataset	**Cat.**	C_1	C_2	C_3	C_4	$SE(C_1)$	$SE(C_2)$	$SE(C_3)$	$SE(C_4)$	$SN(C_1)$	$SN(C_2)$	$SN(C_3)$	$SN(C_4)$
Leukemia	**Nodes**	421	1480	294	988	683	873	843	941	422	1482	293	979
	Edges	1529	2294	2154	2646	1529	2294	2154	2646	680	7145	409	2870
	Clust.Coef.	0.712	0.155	0.589	0.33	0.333	0.348	0.354	0.341	0.323	0.388	0.303	0.344
	Diameter	5	6	4	6	24	18	19	22	16	18	9	20
LungH	**Nodes**	429	1419	382	1030	578	930	955	1214	432	1413	384	1027
	Edges	1068	2317	2398	3410	1068	2317	2398	3410	617	4302	476	2650
	Clust.Coef.	0.344	0.298	0.43	0.404	0.356	0.373	0.376	0.372	0.341	0.386	0.296	0.376
	Diameter	5	8	5	7	10	23	23	21	6	22	6	23
LungM	**Nodes**	91	919	48	247	76	280	59	119	90	915	50	248
	Edges	134	1858	78	410	134	1858	78	410	224	13574	64	1510
	Clust.Coef.	0.379	0.262	0.418	0.457	0.465	0.525	0.446	0.514	0.539	0.523	0.493	0.511
	Diameter	3	5	3	3	6	11	6	6	5	14	6	12
CNS	**Nodes**	501	4257	494	3538	945	2152	1616	2607	501	4261	488	3532
	Edges	4302	25069	12769	40840	4302	25069	12769	40840	1553	171052	1502	90395
	Clust.Coef.	0.743	0.255	0.521	0.302	0.354	0.389	0.367	0.400	0.346	0.427	0.35	0.421
	Diameter	4	7	4	6	21	15	23	13	12	13	14	12
Dlbcl	**Nodes**	201	1699	201	1617	207	1411	1238	1790	200	1699	200	1614
	Edges	848	10471	7351	33170	848	10471	7351	33170	832	24280	832	17865
	Clust.Coef.	0.872	0.574	0.642	0.453	0.508	0.438	0.411	0.51	0.501	0.504	0.501	0.481
	Diameter	3	5	3	5	2	16	17	14	2	14	2	15
GSE2191	**Nodes**	890	4802	846	3561	837	1848	1239	1750	897	4799	839	3553
	Edges	3290	13424	6469	12074	3290	13424	6469	12074	3711	90410	3292	47806
	Clust.Coef.	0.488	0.082	0.317	0.291	0.377	0.409	0.394	0.4	0.382	0.415	0.377	0.417
	Diameter	5	9	5	9	25	23	19	21	21	13	25	15
GS3726	**Nodes**	668	2077	524	1170	879	1300	992	1367	759	2300	1739	3440
	Edges	1761	3255	2051	3502	1761	3255	2051	3502	1471	9808	5479	26761
	Clust.Coef.	0.134	0.0077	0.307	0.109	0.226	0.23	0.223	0.233	0.213	0.287	0.254	0.346
	Diameter	8	10	7	8	20	26	20	25	26	15	19	15
Prostate	**Nodes**	938	4290	704	2277	356	543	322	448	920	4298	702	2287
	Edges	3796	10175	3090	6546	3796	10175	3090	6546	33250	914829	16934	24427
	Clust.Coef.	0.328	0.245	0.29	0.25	0.565	0.607	0.541	0.584	0.655	0.703	0.641	0.711
	Diameter	7	10	6	8	6	9	5	8	8	11	7	12

Supplementary Table 3: Topological properties for co-prediction and ARACNE co-expression networks generated for all 8 datasets.

		Co-prediction				**Co-expression (SE)**				**Co-expression (SN)**			
Dataset	**Cat.**	C_1	C_2	C_3	C_4	$SE(C_1)$	$SE(C_2)$	$SE(C_3)$	$SE(C_4)$	$SN(C_1)$	$SN(C_2)$	$SN(C_3)$	$SN(C_4)$
Leukemia	**Nodes**	421	1480	294	988	1024	1426	1356	1577	422	1480	294	989
	Edges	1529	2294	2154	2646	1529	2294	2154	2646	512	2416	327	1479
	Clust.Coef.	0.712	0.155	0.589	0.330	0.002	0.002	0.002	0.002	0.000	0.002	0.000	0.002
	Diameter	5	6	4	6	17	19	22	19	9	17	11	19
LungH	**Nodes**	429	1419	382	1030	907	1614	1653	2066	429	1419	382	1030
	Edges	1068	2317	2398	3410	1068	2317	2398	3410	435	1924	375	1250
	Clust.Coef.	0.344	0.298	0.430	0.404	0.007	0.006	0.006	0.005	0.013	0.006	0.012	0.007
	Diameter	5	8	5	7	23	16	15	13	14	18	10	18
LungM	**Nodes**	91	919	48	247	143	1321	96	370	91	920	48	247
	Edges	134	1858	78	410	134	1858	78	410	72	1127	34	259
	Clust.Coef.	0.379	0.262	0.418	0.475	0.000	0.002	0.000	0.009	0.000	0.005	0.000	0.014
	Diameter	3	5	3	3	13	17	11	18	11	17	5	11
CNS	**Nodes**	501	4257	494	3538	2002	4509	3581	5342	502	4257	494	3538
	Edges	4302	25069	12769	40840	4302	25069	12769	41661	513	20409	505	12358
	Clust.Coef.	0.743	0.255	0.521	0.302	0.004	0.005	0.006	0.026	0.004	0.005	0.004	0.005
	Diameter	4	7	4	6	12	8	12	7	20	9	20	12
Dlbcl	**Nodes**	201	1699	201	1617	380	1452	1191	2236	201	1699	201	1617
	Edges	848	10471	7351	33170	848	10471	7351	33890	269	14149	269	12903
	Clust.Coef.	0.872	0.574	0.642	0.453	0.136	0.126	0.140	0.176	0.113	0.110	0.113	0.115
	Diameter	3	5	3	5	13	9	11	5	12	8	12	9
GSE2191	**Nodes**	890	4802	846	3561	2574	5226	3846	5027	890	4802	846	3561
	Edges	3290	13424	6469	12074	3290	13424	6469	12076	846	10671	794	5564
	Clust.Coef.	0.488	0.082	0.317	0.291	0.002	0.002	0.002	0.002	0.004	0.002	0.004	0.002
	Diameter	5	9	5	9	19	13	16	13	30	15	30	17
GS3726	**Nodes**	668	2077	524	1170	1362	2167	1546	2279	668	2166	524	1170
	Edges	1761	3255	2051	3502	1761	3255	2051	3502	787	3250	597	1455
	Clust.Coef.	0.134	0.077	0.307	0.109	0.024	0.053	0.029	0.050	0.014	0.053	0.016	0.021
	Diameter	8	10	7	8	20	20	19	18	15	20	13	19
Prostate	**Nodes**	938	4290	704	2277	2760	6805	2268	4575	939	4290	704	2277
	Edges	3796	10175	3090	6546	3796	10175	3090	6546	1300	6095	1017	3102
	Clust.Coef.	0.328	0.245	0.290	0.250	0.005	0.003	0.006	0.003	0.001	0.003	0.002	0.005
	Diameter	7	10	6	8	13	13	15	13	12	13	9	15

Supplementary Table 4: Topological properties for co-prediction and MIC co-expression networks generated for all 8 datasets.

		Co-prediction				Co-expression (SE)				Co-expression (SN)			
Dataset	**Cat.**	C_1	C_2	C_3	C_4	$SE(C_1)$	$SE(C_2)$	$SE(C_3)$	$SE(C_4)$	$SN(C_1)$	$SN(C_2)$	$SN(C_3)$	$SN(C_4)$
Leukemia	**Nodes**	421	1480	294	988	640	807	780	896	421	1480	294	989
	Edges	1529	2294	2154	2646	1529	2294	2155	2647	749	6173	432	3096
	Clust.Coef.	0.712	0.155	0.589	0.330	0.162	0.182	0.180	0.179	0.138	0.180	0.127	0.175
	Diameter	5	6	4	6	18	29	27	29	10	17	8	18
LungH	**Nodes**	429	1419	382	1030	384	685	703	944	429	1419	382	1030
	Edges	1068	2317	2398	3410	1068	2317	2399	3410	1264	5867	1045	3841
	Clust.Coef.	0.344	0.298	0.430	0.404	0.349	0.308	0.305	0.302	0.339	0.282	0.343	0.305
	Diameter	5	8	5	7	9	13	13	17	7	18	9	19
LungM	**Nodes**	91	919	48	247	118	626	79	219	91	919	48	247
	Edges	134	1858	78	410	134	1858	78	410	93	3109	38	484
	Clust.Coef.	0.379	0.262	0.418	0.475	0.212	0.272	0.213	0.306	0.208	0.235	0.153	0.302
	Diameter	3	5	3	3	8	18	7	8	6	14	3	7
CNS	**Nodes**	501	4257	494	3538	1424	3104	2357	3725	501	4257	495	3538
	Edges	4302	25069	12769	40840	4305	25131	12771	40850	704	62208	694	36027
	Clust.Coef.	0.743	0.255	0.521	0.302	0.124	0.154	0.144	0.159	0.089	0.162	0.091	0.161
	Diameter	4	7	4	6	17	11	12	10	11	10	11	10
Dlbcl	**Nodes**	201	1699	201	1617	475	1140	1047	1453	203	1699	203	1617
	Edges	848	10471	7351	33170	848	10471	7362	33172	196	74773	196	59307
	Clust.Coef.	0.872	0.574	0.642	0.453	0.111	0.240	0.219	0.319	0.082	0.381	0.082	0.366
	Diameter	3	5	3	5	21	13	16	15	11	11	11	11
GSE2191	**Nodes**	890	4802	846	3561	1700	4129	2617	3883	890	4803	846	3563
	Edges	3290	13424	6469	12074	3299	13433	6469	12207	1380	17797	1271	10540
	Clust.Coef.	0.488	0.082	0.317	0.291	0.109	0.095	0.098	0.098	0.120	0.095	0.118	0.099
	Diameter	5	9	5	9	22	15	18	16	19	15	21	15
GS3726	**Nodes**	668	2077	524	1170	1271	1921	1357	1996	672	2152	526	1172
	Edges	1761	3255	2051	3502	1890	3261	2056	3524	852	3921	538	1705
	Clust.Coef.	0.134	0.077	0.307	0.109	0.110	0.100	0.104	0.100	0.126	0.099	0.121	0.117
	Diameter	8	10	7	8	23	29	23	28	14	24	14	24
Prostate	**Nodes**	938	4290	704	2277	687	839	667	773	964	4290	712	2277
	Edges	3796	10175	3090	6546	3981	10186	3777	8257	15928	1763794	5254	308709
	Clust.Coef.	0.328	0.245	0.290	0.250	0.167	0.278	0.169	0.265	0.313	0.758	0.218	0.661
	Diameter	7	10	6	8	7	8	7	8	7	8	7	9

When analysing FuNeL networks we observed, as expected, that configurations having feature selection (C_1 and C_3) lead to networks with a smaller number of nodes than when the original set of attributes is used (C_2 and C_4). Furthermore, the second phase of machine learning modeling (C_3 and C_4) tends to reduce the number of nodes as it uses a reduced set of attributes as input (only significant nodes and their neighbours from the first training phase), while increasing both clustering coefficient and number of edges.

When comparing FuNeL and co-expression networks we notice that the ARACNE *SE* counterparts have in general more nodes. The same patter can be found in $SE(C_2$ and $SE(C_4)$ counterparts generated with PCC and MIC, while it's not true for the *SE*-networks based on configurations that use feature selection (C_1 and C_2). Conversely, *SN*-networks differ according to the inference method used. In fact ARACNE generated *SN* counterparts with less edges, while this is true only for $SN(C_1)$ and $SN(C_3)$ inferred with MIC and PCC. The clustering coefficient is constantly lower in ARACNE networks than in FuNeL, this is probably due to the pruning phase operated by the method. A similar trend can be noticed for MIC networks with some exceptions (e.g Prostate $SN(C_2)$ and $SN(C_4)$). A more balanced situation occurs when FuNeL is contrasted with PCC, in fact networks generated with feature selection (C_1 and C_3) have a lower coefficient than their co-expression counterparts. Finally a clear pattern emerge when analysing the diameter of the networks. Co-prediction networks are always more compact than co-expression counterparts having up to 3 time lower diameter for MIC and PCC and up to 7 time lower for ARACNE.

4 Enrichment Score analysis

In this section we report the network average rankings, based on the Enrichment Score, across the 8 datasets for each inferring method. The networks are ranked between 1 and N (where $N = 4$ for FuNeL and $N = 8$ for PCC, ARACNE and MIC: 4 $SE(C_i)$ + 4 $SN(C_i)$). We considered Gene Ontology terms (biological process (BP), molecular function (MF) and cellular component (CC)) and biological pathways. The last row of each table represents the average rank across different biological categories.

Supplementary Table 5: Average network ranks for each method across the 8 datasets (based on ES).

Cat.	**C1**	**C2**	**C3**	**C4**
GO BP	3	4	2	**1**
GO MF	4	2.5	**1**	2.5
GO CC	2	4	**1**	3
Pathways	4	2	3	**1**
Average	3.25	3.125	**1.75**	1.88

(a) FuNeL networks

	PCC (SE)				**PCC (SN)**			
Cat.	C_1	C_2	C_3	C_4	C_1	C_2	C_3	C_4
GO BP	2	4	**1**	3	6	7	5	8
GO MF	8	3.5	5	6	7	3.5	**1**	2
GO CC	2	5	4	6	**1**	8	3	7
Pathways	6	5	4	3	7	**1**	8	2
Average	4.5	4.38	**3.5**	4.5	5.25	4.88	4.25	4.75

(b) PCC networks

	ARACNE (SE)				**ARACNE (SN)**			
Cat.	C_1	C_2	C_3	C_4	C_1	C_2	C_3	C_4
GO BP	4	7	5.5	8	2	5.5	**1**	3
GO MF	4	8	4	7	2	6	**1**	4
GO CC	3	7	5	8	2	6	**1**	4
Pathways	**1**	4	2	6	7	5	8	3
Average	3	6.5	4.13	7.25	3.25	5.63	**2.75**	3.5

(c) ARACNE networks

	MIC (SE)				**MIC (SN)**			
Cat.	C_1	C_2	C_3	C_4	C_1	C_2	C_3	C_4
GO BP	2	6	4	5	3	8	**1**	7
GO MF	**1**	6	4	7	3	8	2	5
GO CC	3	5.5	4	5.5	2	8	**1**	7
Pathways	2.5	4	**1**	5.5	8	5.5	7	2.5
Average	**2.13**	5.38	3.25	5.75	4	7.38	2.75	5.38

(d) MIC networks

We also compared the generated networks against FuNeL. The networks are ranked from 1 to 12: 4 C_i + 4 $SE(C_i)$ + 4 $SN(C_i)$. The ranks in Supplementary Table 6 are averaged across the 8 datasets, for each biological category and for each network. The row-wise rank is given in brackets and the highest ranks are shown with bold font. The following abbreviations were used for GO categories: biological process (BP), molecular function (MF) and cellular component (CC).

Supplementary Table 6: Average network ranks (co-prediction vs. co-expression).

		Co-prediction				**Co-expression (SE)**				**Co-expression (SN)**			
Method	**Cat.**	C_1	C_2	C_3	C_4	$SE(C_1)$	$SE(C_2)$	$SE(C_3)$	$SE(C_4)$	$SN(C_1)$	$SN(C_2)$	$SN(C_3)$	$SN(C_4)$
PCC	**GO BP**	6.06 (6)	7.00 (7.5)	7.00 (7.5)	5.88 (3.5)	5.88 (3.5)	5.88 (3.5)	**5.12 (1)**	5.88 (3.5)	7.06 (9)	7.75 (12)	7.12 (10)	7.38 (11)
	GO MF	7.81 (11)	5.38 (2)	6.19 (5)	5.62 (3)	9.12 (12)	6.50 (7.5)	6.50 (7.5)	7.38 (10)	7.19 (9)	6.25 (6)	**4.31 (1)**	5.75 (4)
	GO CC	4.31 (5)	11.00 (12)	4.19 (4)	9.00 (10)	3.88 (2)	6.25 (6.5)	6.25 (6.5)	8.12 (8)	**3.19 (1)**	9.38 (11)	4.06 (3)	8.38 (9)
	Pathways	8.12 (10.5)	4.75 (2)	8.12 (10.5)	**4.38 (1)**	6.94 (8)	6.50 (6)	6.69 (7)	5.88 (5)	7.62 (9)	5.00 (3)	8.50 (12)	5.50 (4)
ARACNE	**GO BP**	6.69 (7)	6.25 (5.5)	6.94 (10)	**4.75 (1)**	6.25 (5.5)	8.12 (11)	6.88 (8.5)	9.25 (12)	5.19 (3)	6.88 (8.5)	5.06 (2)	5.75 (4)
	GO MF	7.44 (10)	6.50 (8)	6.19 (6.5)	5.69 (3)	6.00 (5)	8.75 (12)	5.62 (2)	8.62 (11)	5.81 (4)	7.00 (9)	**4.19 (1)**	6.19 (6.5)
	GO CC	4.31 (4)	10.75 (12)	3.44 (3)	8.25 (8)	5.50 (5)	9.38 (10)	6.75 (7)	10.12 (11)	2.44 (2)	8.88 (9)	**2.31 (1)**	5.88 (6)
	Pathways	7.88 (10.5)	5.38 (3)	7.88 (10.5)	5.25 (2)	**4.88 (1)**	6.25 (6)	5.50 (4)	7.00 (8)	7.56 (9)	6.50 (7)	8.38 (12)	5.56 (5)
MIC	**GO BP**	7.44 (8.5)	7.88 (11)	7.44 (8.5)	6.38 (5.5)	4.12 (2)	7.00 (7)	6.00 (4)	6.38 (5.5)	4.81 (3)	9.12 (12)	**3.94 (1)**	7.50 (10)
	GO MF	8.06 (10)	8.50 (12)	7.19 (8)	7.75 (9)	**3.62 (1)**	6.50 (5.5)	5.12 (3)	6.75 (7)	5.31 (4)	8.12 (11)	4.56 (2)	6.50 (5.5)
	GO CC	5.19 (6)	11.62 (12)	4.44 (4)	10.12 (10)	4.25 (3)	7.12 (7.5)	5.12 (5)	7.12 (7.5)	3.19 (2)	10.38 (11)	**1.69 (1)**	7.75 (9)
	Pathways	8.75 (12)	**4.75 (1)**	7.50 (10)	5.50 (3)	6.00 (5)	6.50 (6)	5.00 (2)	6.62 (7)	7.56 (11)	7.00 (9)	6.94 (8)	5.88 (4)

5 Disease association analysis

In this section we report the network average rankings across the 8 datasets for every inferring method based on the gene-disease association properties: participation in triangular relationship and proximity. We used two sources for the disease associations: Malacards (Rappaport *et al.*, 2013) (a meta-database of human maladies consolidated from 64 independent sources) and manually curated databases (OMIM (Hamosh *et al.*, 2005), Orphanet (INSERM, 1997), Uniprot (Magrane and Consortium, 2011) and CTD (Davis *et al.*, 2015)). The networks are ranked between 1 and N (where $N = 4$ for FuNeL and $N = 8$ for PCC, ARACNE and MIC: 4 $SE(C_i)$ + 4 $SN(C_i)$). The number of disease-associated genes participating in a triangle is denoted as 1A, 2A and 3A. The last row of each table represents the average rank across different metrics.

Supplementary Table 7: Average network ranks for each method across the 8 datasets (based on disease associations from Malacards).

Cat.	C1	C2	C3	C4
1A	3	**1**	4	2
2A	4	**1**	2	3
3A	3	3	**1**	3
Proximity	3	**1**	4	2
Average	3.25	**1.5**	2.75	2.5

(a) FuNeL networks

	PCC (SE)				PCC (SN)			
Cat.	C_1	C_2	C_3	C_4	C_1	C_2	C_3	C_4
1A	8	5	6	3	4	**1**	2	7
2A	7	4	5	3	8	2	6	**1**
3A	6.5	6.5	6.5	6.5	3	2	4	**1**
Proximity	**1.5**	5	3	**1.5**	7	6	8	4
Average	5.75	5.13	5.13	3.5	5.5	**2.75**	5	3.25

(b) PCC networks

	ARACNE (SE)				ARACNE (SN)			
Cat.	C_1	C_2	C_3	C_4	C_1	C_2	C_3	C_4
1A	4	8	7	5.5	2.5	2.5	5.5	**1**
2A	7	6	8	**1**	4	3	2	5
3A	5.5	**1**	5.5	2	5.5	5.5	5.5	5.5
Proximity	7	3	5	**1**	6	2	8	4
Average	5.88	4.5	6.38	**2.38**	4.5	3.25	5.25	3.88

(c) ARACNE networks

	MIC (SE)				MIC (SN)			
Cat.	C_1	C_2	C_3	C_4	C_1	C_2	C_3	C_4
1A	6	3	7.5	**1**	5	2	7.5	4
2A	8	3	5	2	7	**1**	6	4
3A	7	2	8	6	5	**1**	3	4
Proximity	6	**1**	2	5	7	4	8	3
Average	6.75	2.25	5.63	3.5	6	2	6.13	3.75

(d) MIC networks

Supplementary Table 8: Average network ranks for each method across the 8 datasets (based on disease associations from curated databases).

Cat.	C1	C2	C3	C4
1A	3	**1**	4	2
2A	3	4	**1**	2
3A	**1**	2.5	2.5	4
Proximity	4	**1**	3	2
Average	2.75	**2.13**	2.67	2.5

(a) FuNeL networks

	PCC (SE)				PCC (SN)			
Cat.	C_1	C_2	C_3	C_4	C_1	C_2	C_3	C_4
1A	8	4	6.5	**1.5**	6.5	**1.5**	3	5
2A	2	7	5.5	3	4	5.5	**1**	8
3A	5	7	4	8	**1**	6	3	2
Proximity	4	8	2.5	7	5	2.5	**1**	6
Average	4.5	6.5	4.63	4.88	5.13	3.88	**2**	5.25

(b) PCC networks

	ARACNE (SE)				ARACNE (SN)			
Cat.	C_1	C_2	C_3	C_4	C_1	C_2	C_3	C_4
1A	3	5	8	6	2	4	**1**	7
2A	5	**1**	2	3	7.5	6	7.5	4
3A	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5
Proximity	6	5	4	**1**	7.5	3	7.5	2
Average	4.63	3.88	4.63	**3.63**	5.38	4.38	5.13	4.38

(c) ARACNE networks

	MIC (SE)				MIC (SN)			
Cat.	C_1	C_2	C_3	C_4	C_1	C_2	C_3	C_4
1A	6	2	4	3	8	1	7	5
2A	7	3	8	**1.5**	5.5	**1.5**	5.5	4
3A	4	**1**	8	2	6	5	7	3
Proximity	5.5	**1**	3	4	7	2	8	5.5
Average	5.63	**1.75**	5.75	2.63	6.63	2.38	6.88	4.38

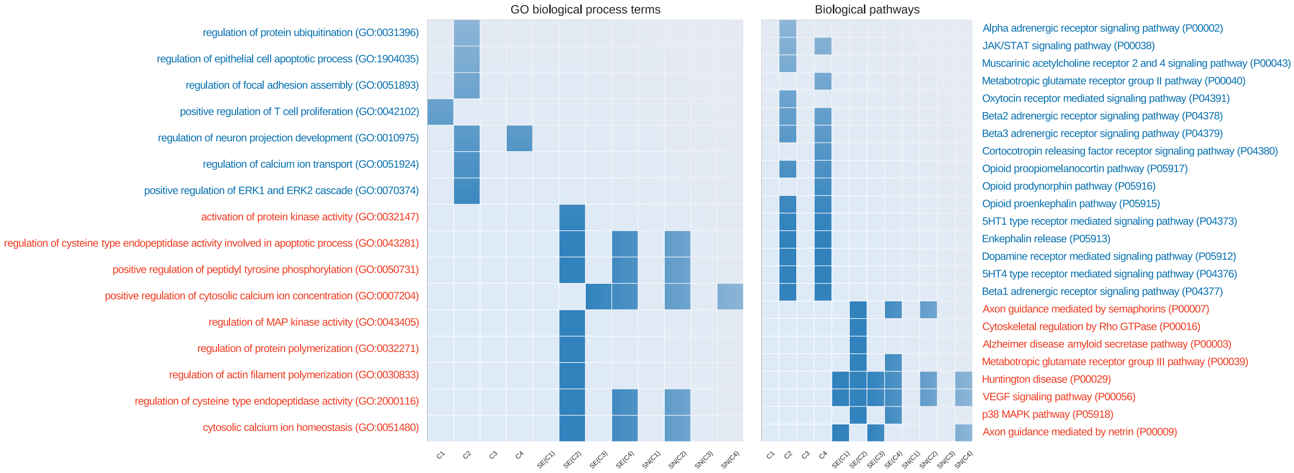
(d) MIC networks

6 Case study: prostate cancer dataset

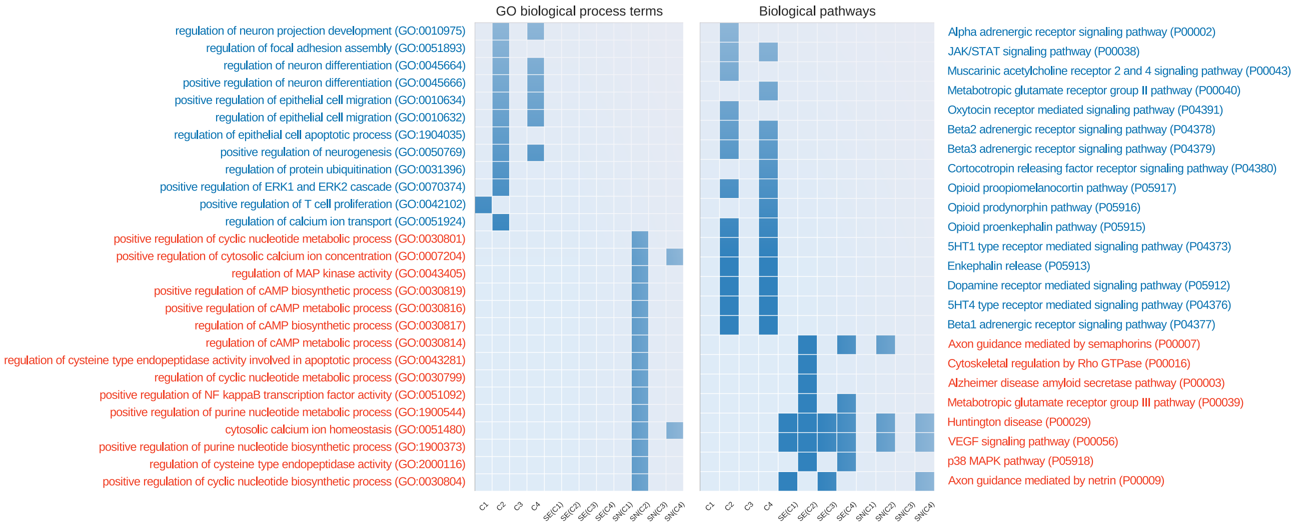
In this section we report the additional results from the analysis performed using the prostate dataset (Singh *et al.*, 2002) as a case study. In particular we show: 1) the overlap of enriched terms between co-prediction and co-expression networks, 2) the overlap between GO terms associated to the hubs of the networks generated with different methods and FuNeL and 3) the average percentages of alteration for key nodes of both co-prediction and co-expression networks in an independent dataset.

6.1 Overlap of the enriched terms

We performed an analysis on the enriched terms of each network to highlight the complementary nature of the co-prediction and the co-expression paradigm. We generated heatmaps showing the unique terms associated only to co-prediction or co-expression networks. The main manuscript includes the comparison between FuNeL and PCC networks, in here we report the analysis performed considering ARACNE (Supplementary Figure 2) and MIC (Supplementary Figure 3) networks. For the sake of readability we filtered out the generic GO terms (with depth < 9 in the GO hierarchical structure).



Supplementary Figure 2: Number of non-common enriched GO terms (biological process) for each network configuration (generated from the prostate cancer dataset). On the x-axis we show the 12 investigated networks. On the y-axis we show the names of enriched terms unique to co-prediction or ARACNE co-expression networks. Red terms are associated with co-expression networks, blue with co-prediction. Empty columns indicate networks with no unique terms.



Supplementary Figure 3: Number of non-common enriched GO terms (biological process) for each network configuration (generated from the prostate cancer dataset). On the x-axis we show the 12 investigated networks. On the y-axis we show the names of enriched terms unique to co-prediction or MIC co-expression networks. Red terms are associated with co-expression networks, blue with co-prediction. Empty columns indicate networks with no unique terms.

When comparing ARACNE and FuNeL, we found 16 unique pathways for co-prediction networks and 8 for co-expression. In terms of unique GO terms, the overlap was more balanced, 7 for co-prediction networks and 9 for co-expression networks. C_2 and C_4, generated without feature selection, had the largest number of unique pathways, while $SE(C_2)$ had the highest number of terms for ARACNE. The comparison of FuNeL with MIC generated many empty columns (Supplementary Figure 3) for the GO terms because several networks resulted having no unique enriched terms. All the 15 unique GO terms related to MIC were associated to $SN(C_2)$ (and with $SN(C_4)$ in two cases), conversely FuNeL had more networks sharing the 12 unique terms. Finally, as noticed for in the ARACNE comparison, FuNeL networks are more enriched in biological pathways: 16 against 8 unique terms for MIC co-expression.

6.2 Overlap of hub-related terms

We also analysed the gene associated to the hubs of each network in order to compare the biological knowledge associated to them. A node v was considered to be a hub if its degree was at least one standard deviation above the mean network degree, that is if:

$$d(v) > \mu_d + \sigma_d \tag{4}$$

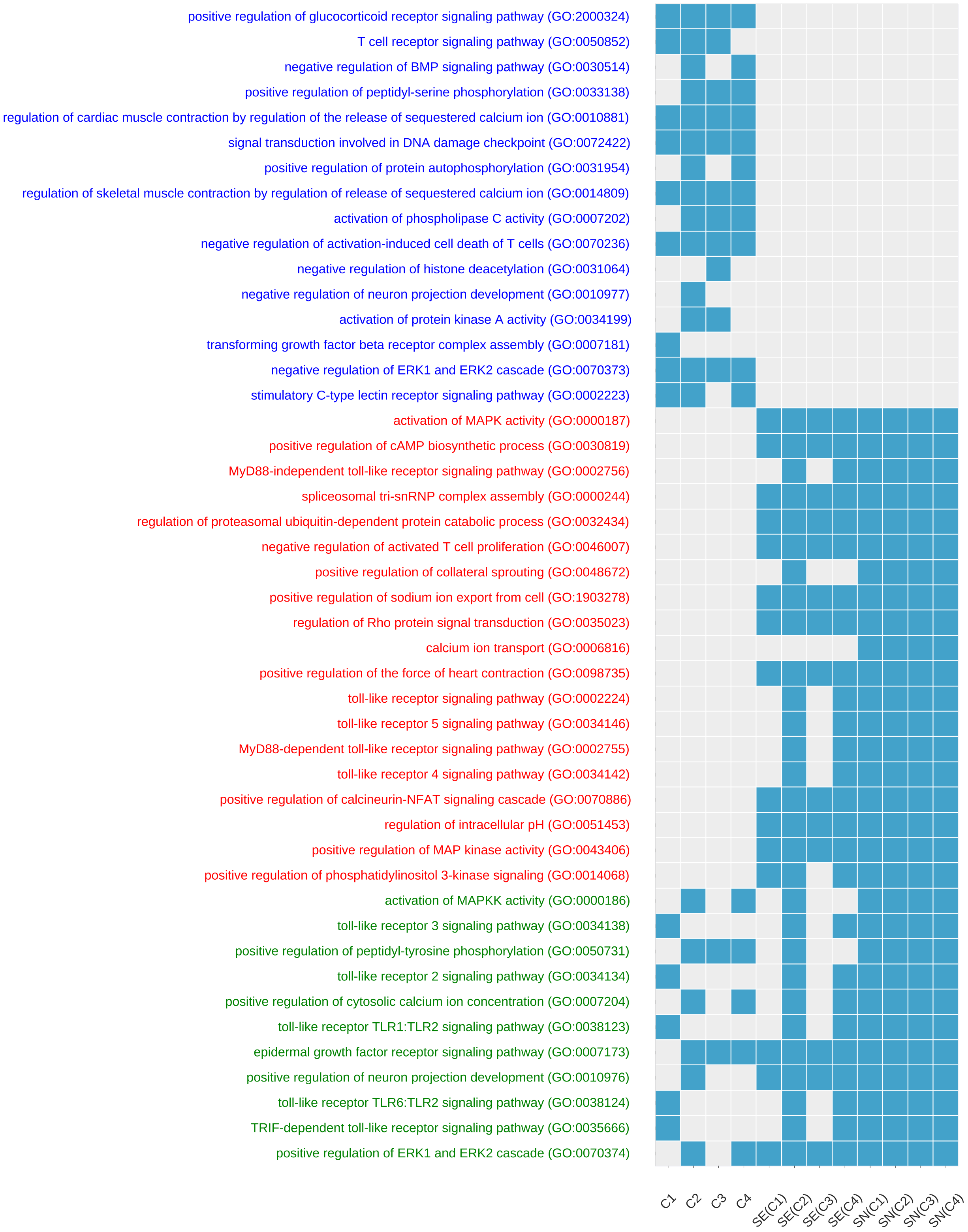
where $d(v)$ is a degree of the node v, and μ_d and σ_d are the mean and standard deviation of a network node degree distribution.

To compare the networks, we used the 10 most frequent GO terms (biological processes) shared among each network's hubs. Supplementary Figures 4 to 6 show the terms-overlap analysis between FuNeL networks and PCC, ARACNE and MIC respectively. To make this analysis more specific we have discarded the most generic / most common terms (which could be be associated with many genes), we considered only the GO terms situated at level 10 of the GO hierarchy or lower.

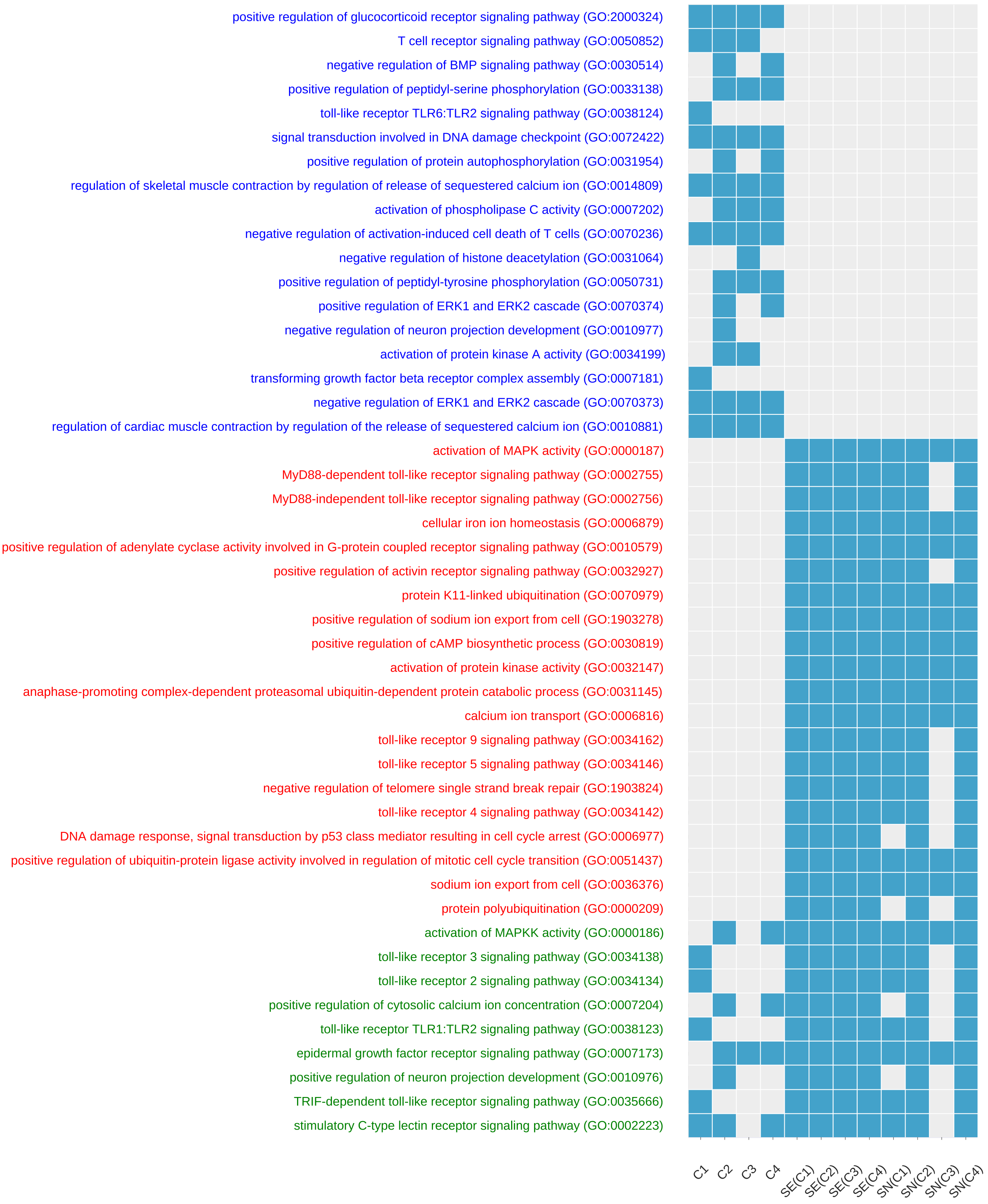
Blue terms were found only in co-prediction networks, red terms were found only in co-expression networks, and green terms were found in both. In Supplementary Table 9 we summarise the number of unique and common terms shared between networks created with different approaches. This analysis further highlights the complementary nature of co-prediction and co-expression approach, the terms that are paradigm-specif always outnumber the common ones.

Supplementary Table 9: Unique and common terms from networks' hubs

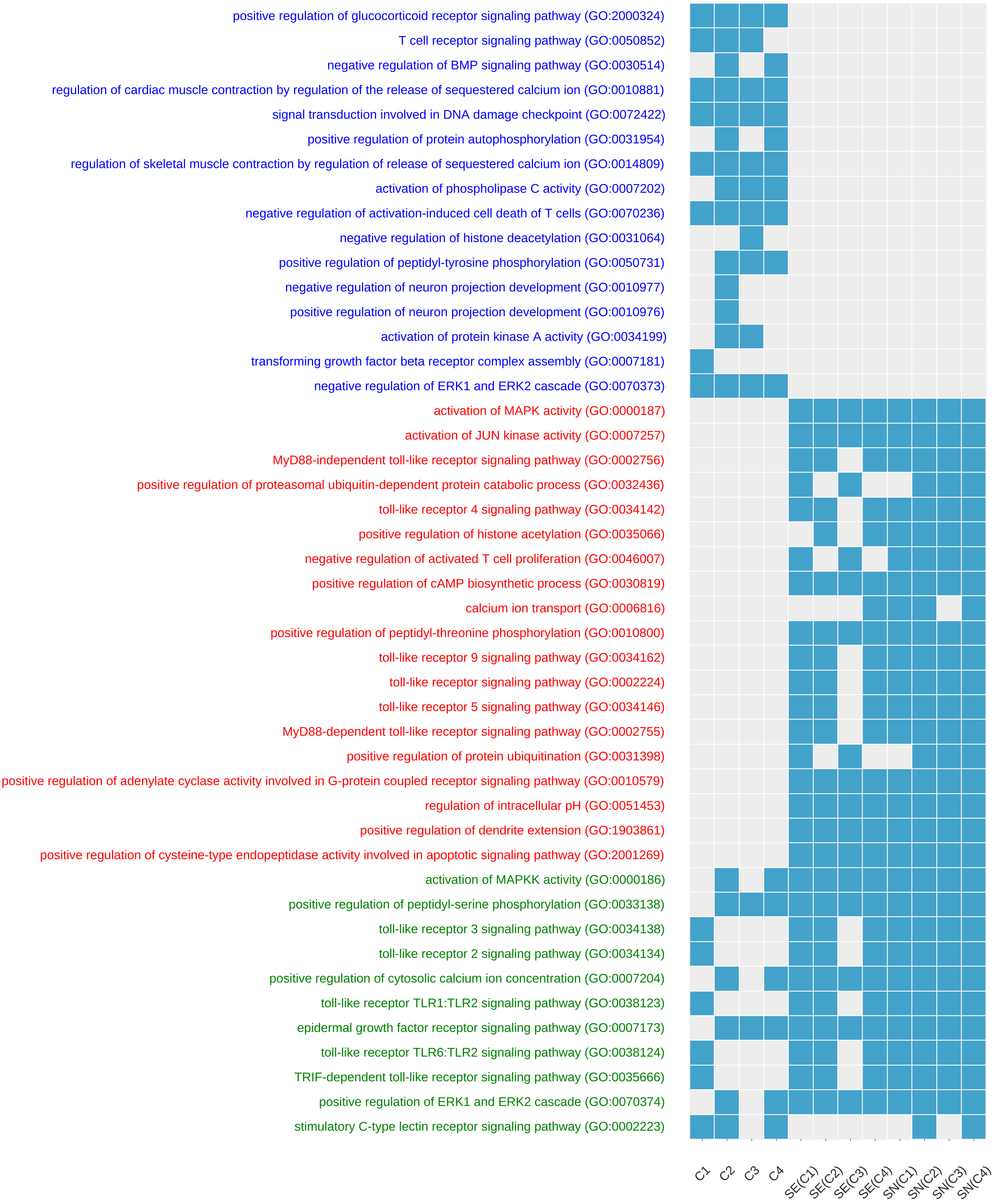
Terms	**PCC**	**ARACNE**	**MIC**
Co-prediction	16	18	16
Co-expression	19	20	19
Common	11	9	11



Supplementary Figure 4: Top 10 most frequent biological processes from Gene Ontology found in the network hubs when comparing FuNeL and PCC co-expression networks.



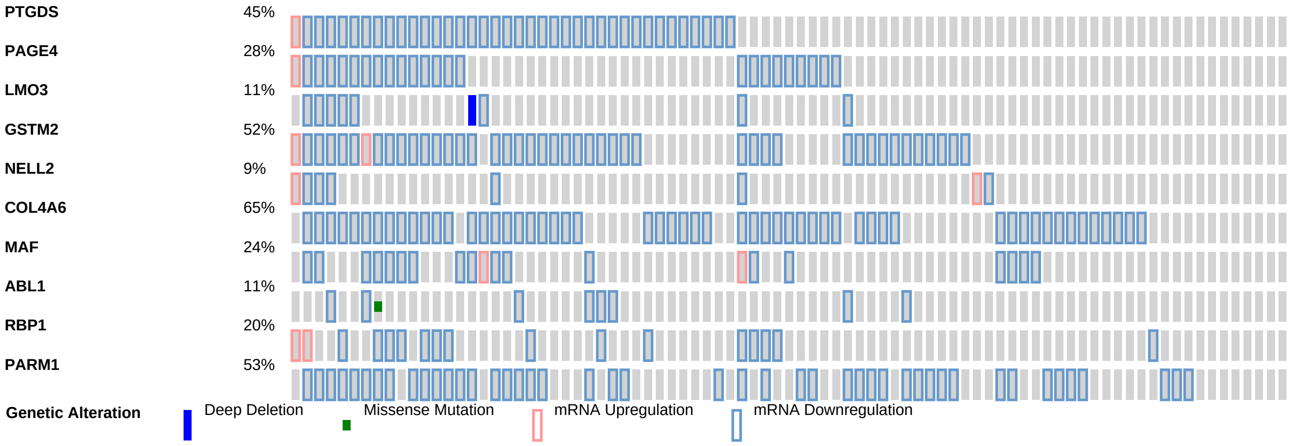
Supplementary Figure 5: Top 10 most frequent biological processes from Gene Ontology found in the network hubs when comparing FuNeL and ARACNE co-expression networks.



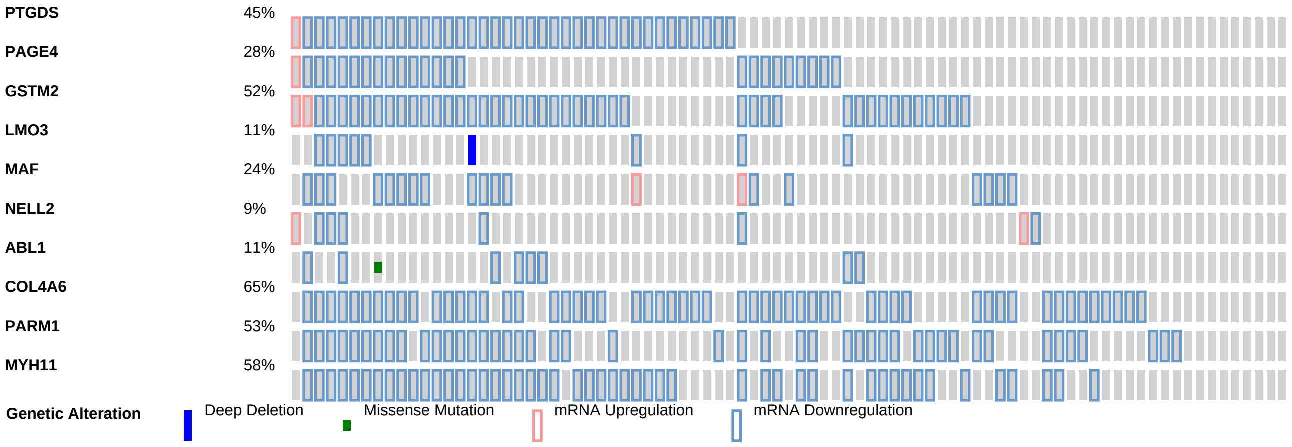
Supplementary Figure 6: Top 10 most frequent biological processes from Gene Ontology found in the network hubs when comparing FuNeL and MIC co-expression networks.

6.3 Validation on independent dataset

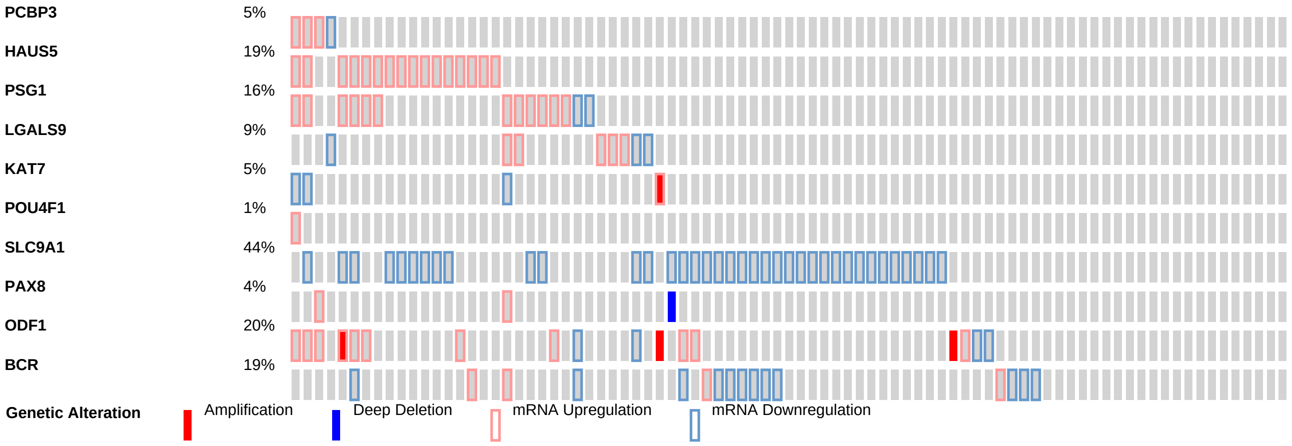
In this section we report additional informations about the analysis performed using data from the independent prostate cancer study (Taylor *et al.*, 2010) available in the cBioPortal for Cancer Genomics (Cerami *et al.*, 2012). In particular we report the full list of alterations for the topologically important genes analysed in the main article. The Supplementary Figure 7–14 show the percentage of altered tumour samples for top 10 hubs (nodes with highest degree) and top 10 central nodes (with highest betweenness centrality) in the best performing networks according to the gene-disease association analysis (using the information from the curated databases). The selected networks are C_2 for FuNeL, $SN(C_3)$ for PCC, $SE(C_4)$ for ARACNE and $SE(C_2)$ for MIC. For all of them we report the alterations for both hubs and central nodes.



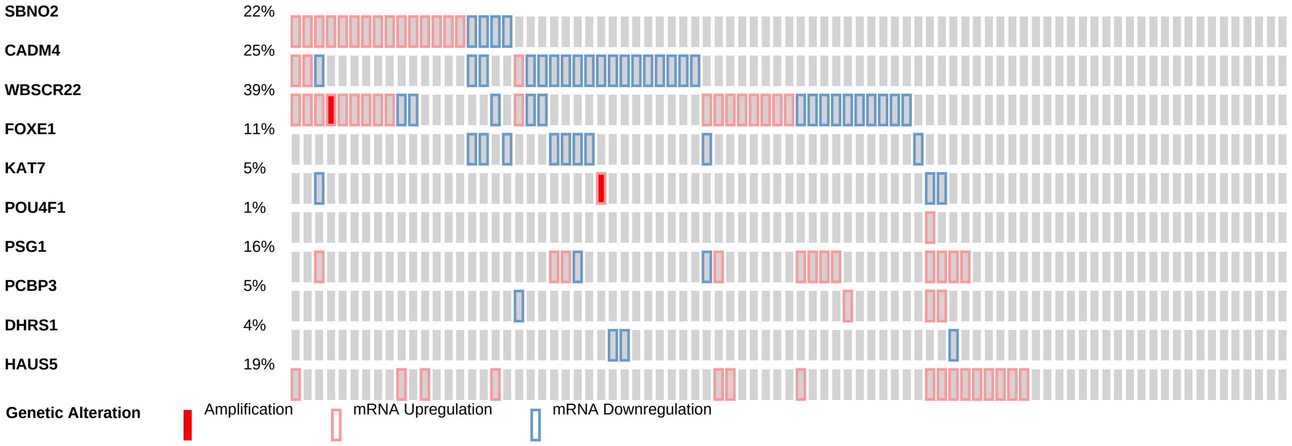
Supplementary Figure 7: Percentage of alterations in tumour samples from an independent cancer genomic study. Genes with **highest degree** (hubs) in C_2 network are shown.



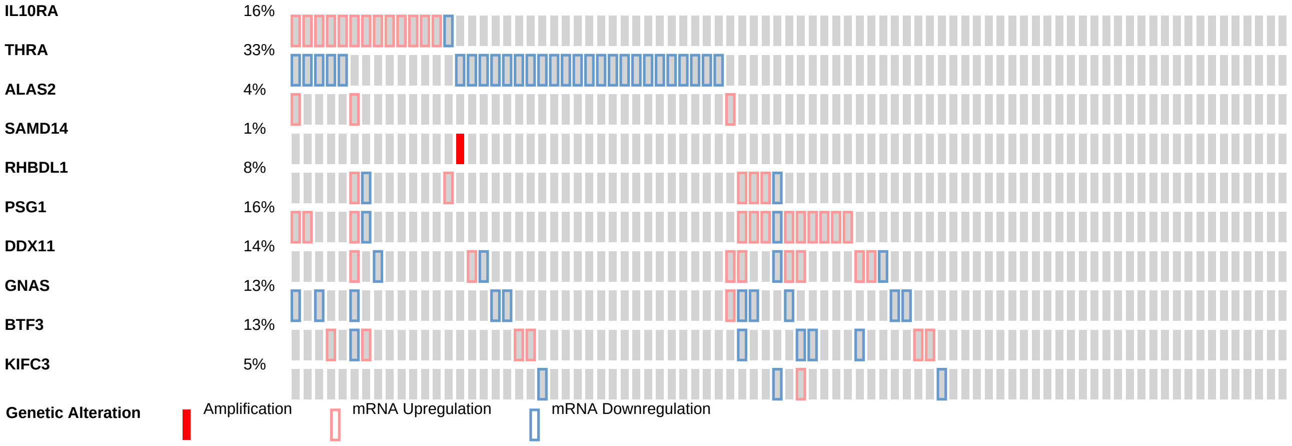
Supplementary Figure 8: Percentage of alterations in tumour samples from an independent cancer genomic study. Genes with **highest betweenness centrality** (central nodes) in C_2 network are shown.



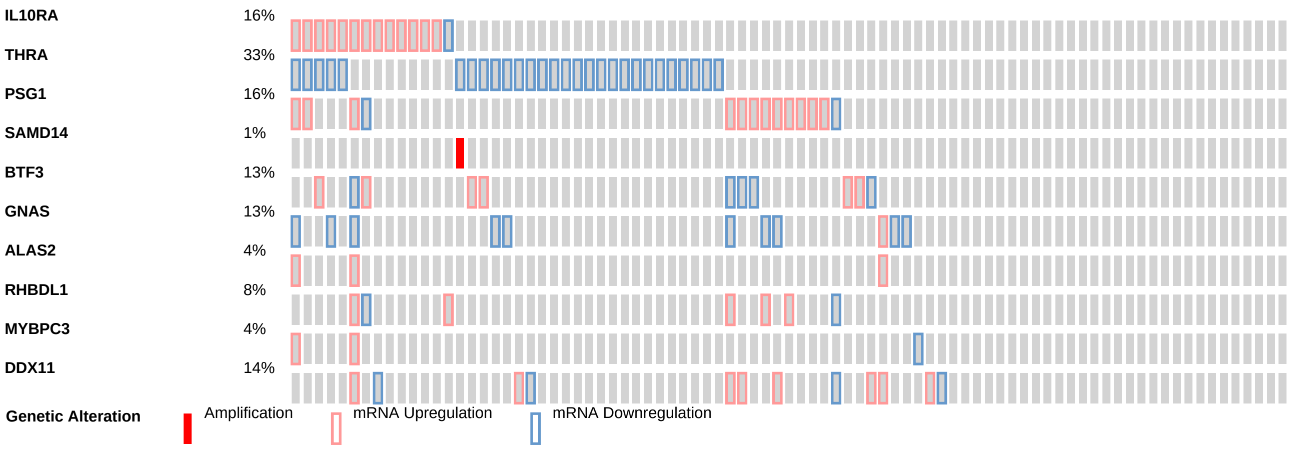
Supplementary Figure 9: Percentage of alterations in tumour samples from an independent cancer genomic study. Genes with **highest degree** (hubs) in PCC $SN(C_3)$ network are shown.



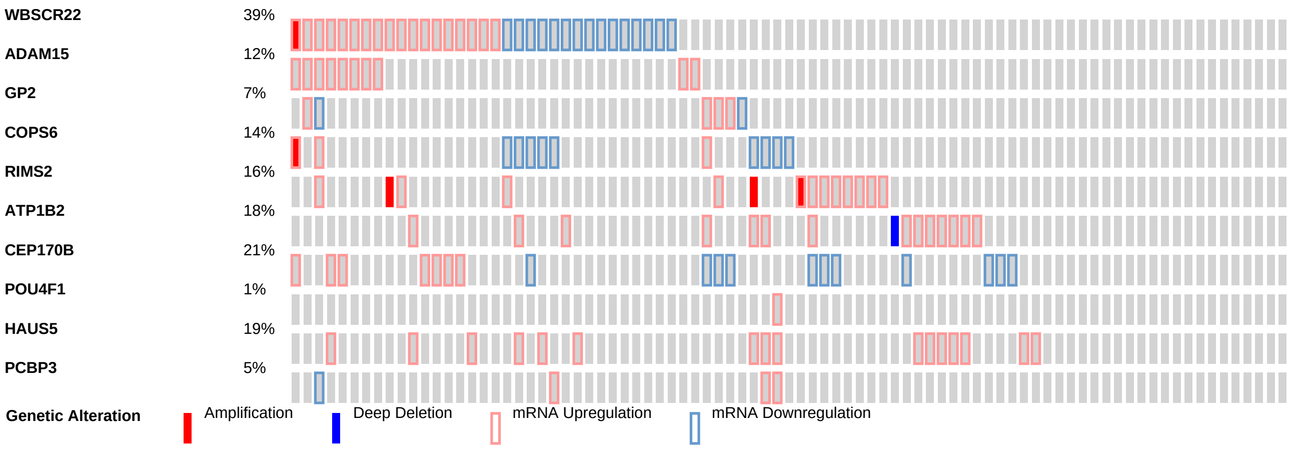
Supplementary Figure 10: Percentage of alterations in tumour samples from an independent cancer genomic study. Genes with **highest betweenness centrality** (central nodes) in PCC $SN(C_3)$ network are shown.



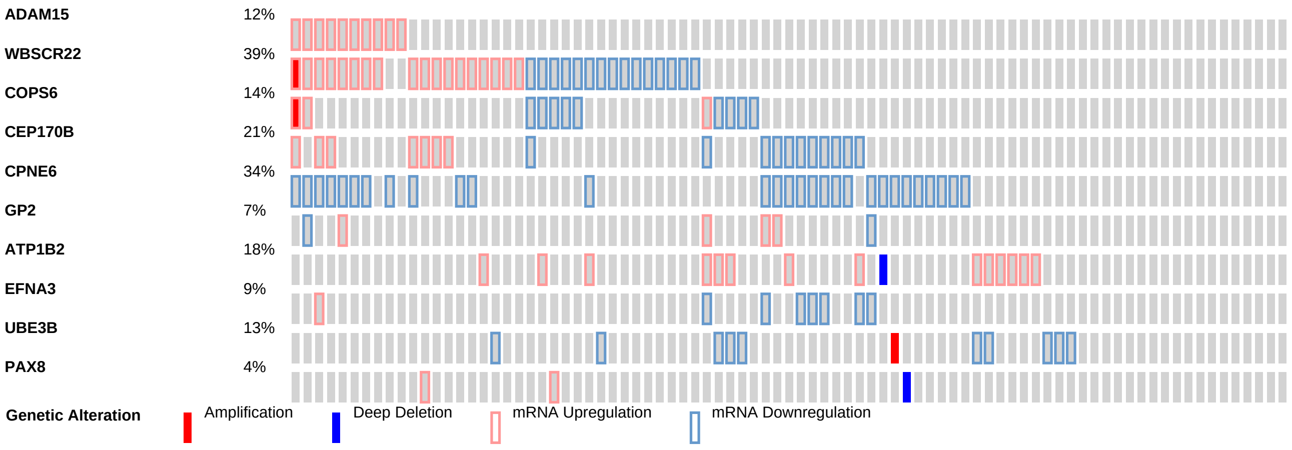
Supplementary Figure 11: Percentage of alterations in tumour samples from an independent cancer genomic study. Genes with **highest degree** (hubs) in ARACNE $SE(C_4)$ network are shown.



Supplementary Figure 12: Percentage of alterations in tumour samples from an independent cancer genomic study. Genes with **highest betweenness centrality** (central nodes) in ARACNE $SE(C_4)$ network are shown.



Supplementary Figure 13: Percentage of alterations in tumour samples from an independent cancer genomic study. Genes with **highest degree** (hubs) in MIC $SE(C_2)$ network are shown.

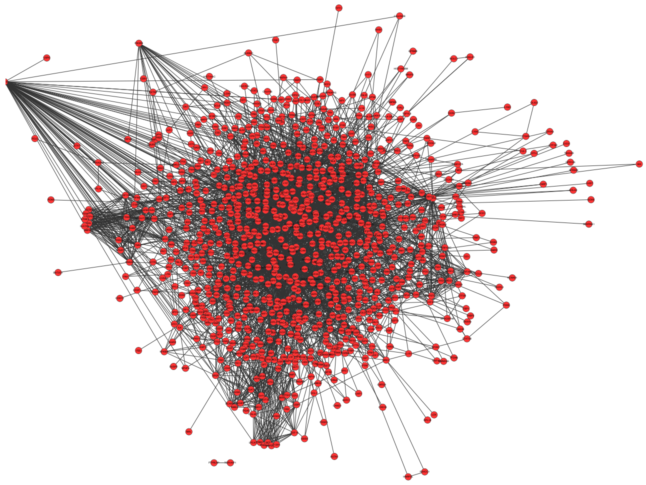


Supplementary Figure 14: Percentage of alterations in tumour samples from an independent cancer genomic study. Genes with **highest betweenness centrality** (central nodes) in MIC $SE(C_2)$ network are shown.

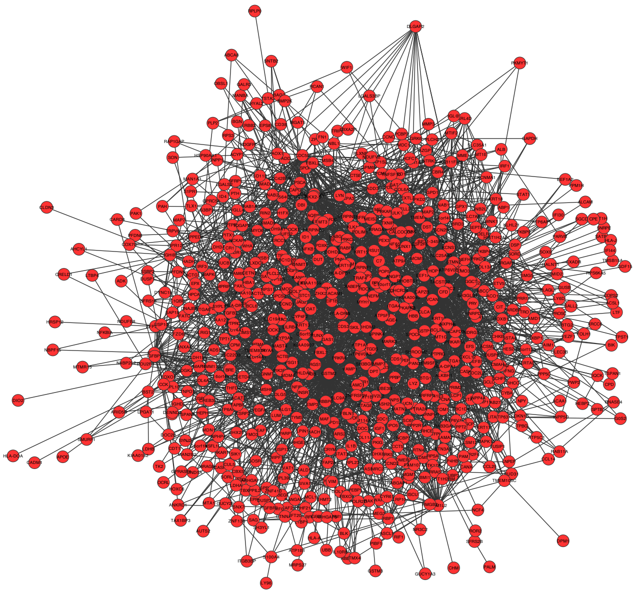
7 Visualisation of the co-prediction and co-expression networks

In this section we include the layouts of co-prediction and co-expression networks for three of the datasets used in the main article: *Prostate* (used in the case study) and *Lung-Michigan* (small networks). Only a few examples are shown, for which the differences in the topology of the networks generated with the two approaches is the most visible. The networks were visualised using the *Organic Layout* in Cytoscape (Shannon *et al.*, 2003).

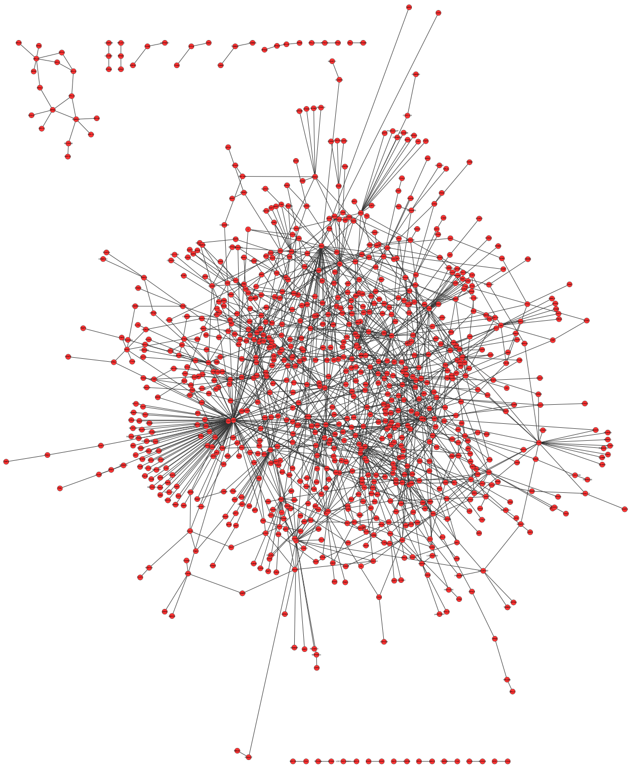
Prostate: FuNeL – C_1



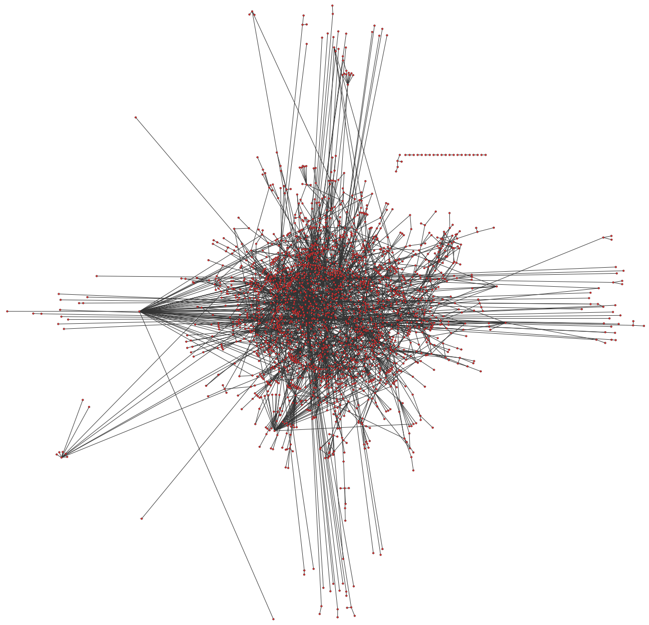
Prostate: FuNeL – C_3



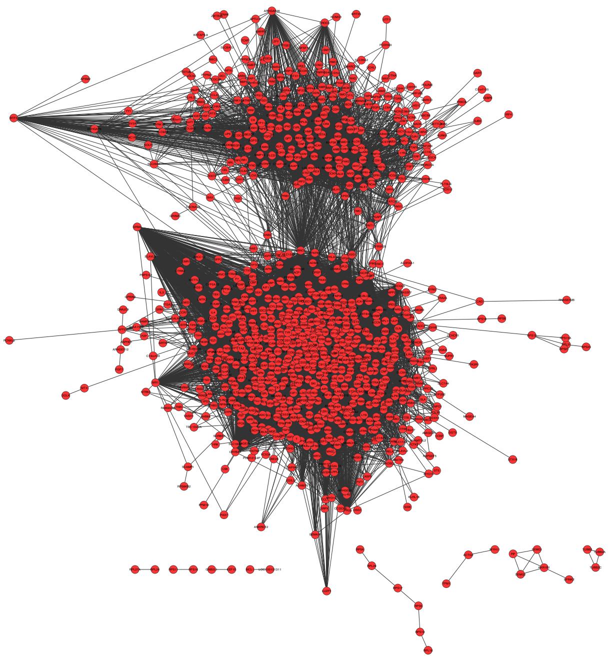
Prostate: ARACNE – $SN(C_1)$



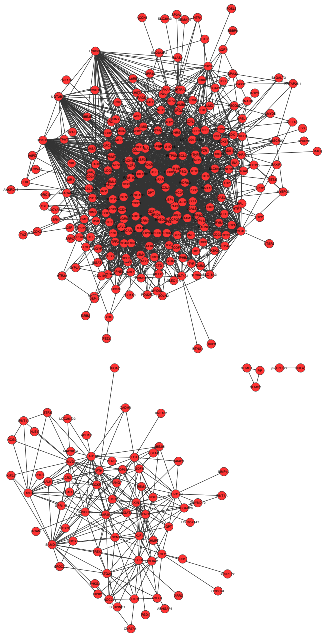
Prostate: ARACNE – $SE(C_3)$



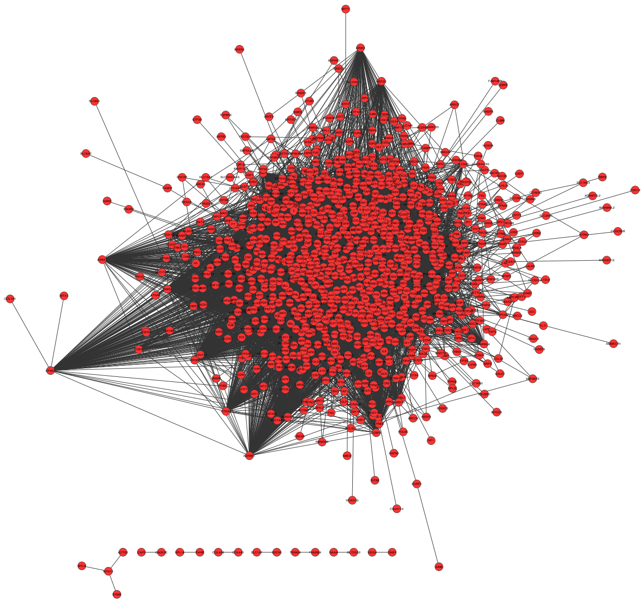
Prostate: PCC – $SN(C_1)$



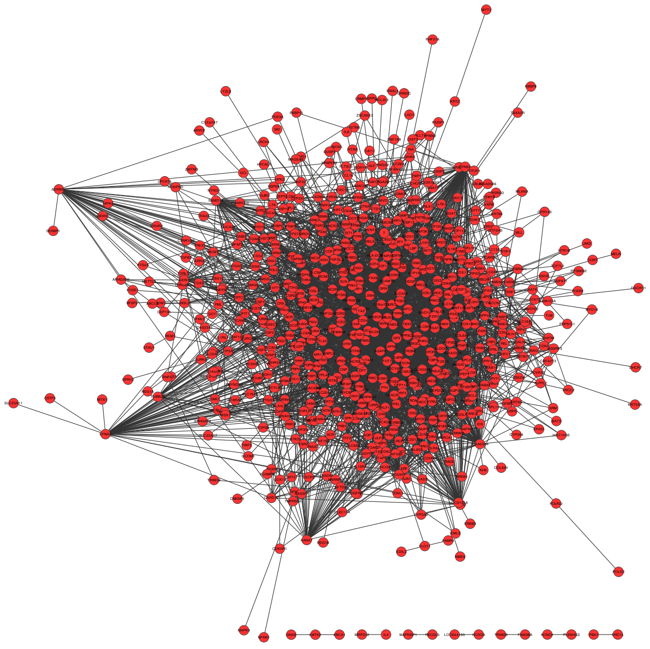
Prostate: PCC – $SE(C_3)$



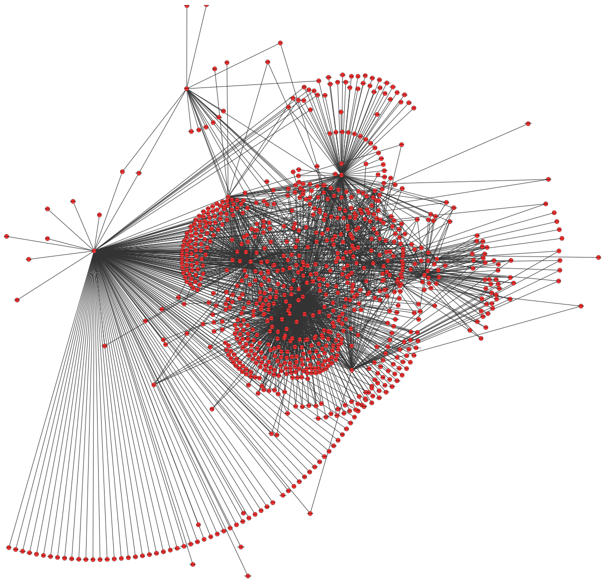
Prostate: MIC – $SN(C_1)$



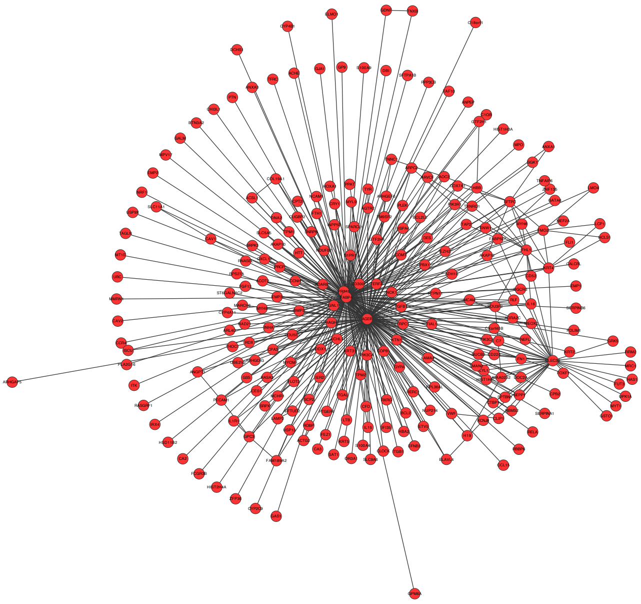
Prostate: MIC – $SE(C_3)$



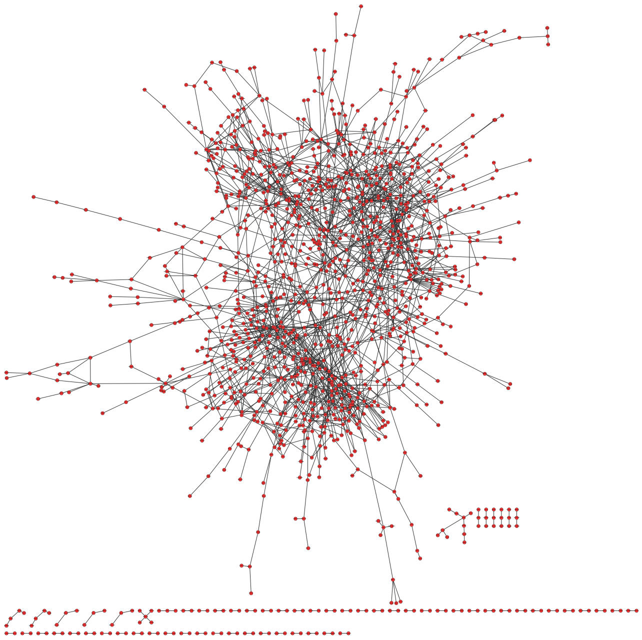
Lung-Michigan: FuNeL – C_2



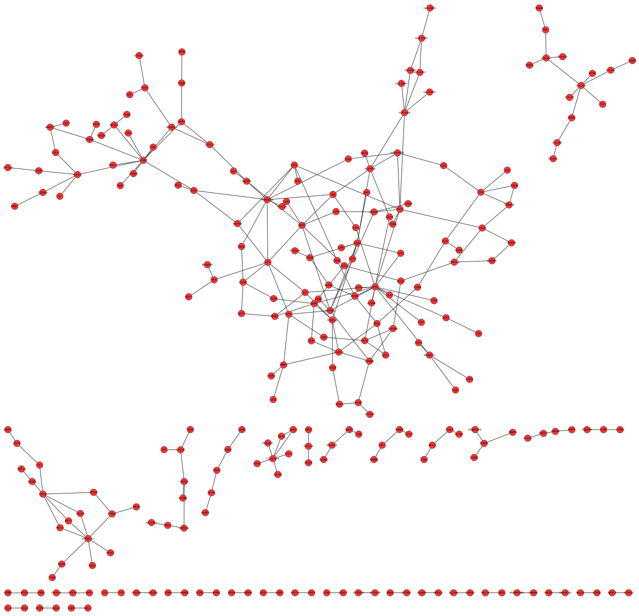
Lung-Michigan: FuNeL – C_4

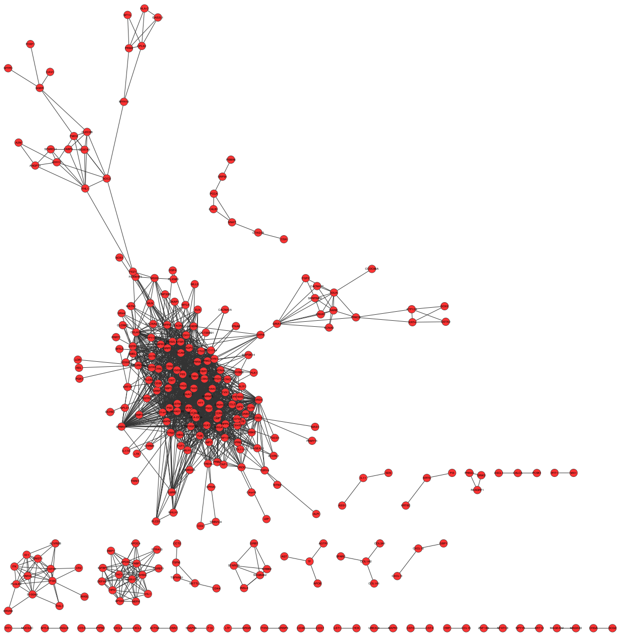


Lung-Michigan: ARACNE – $SE(C_2)$

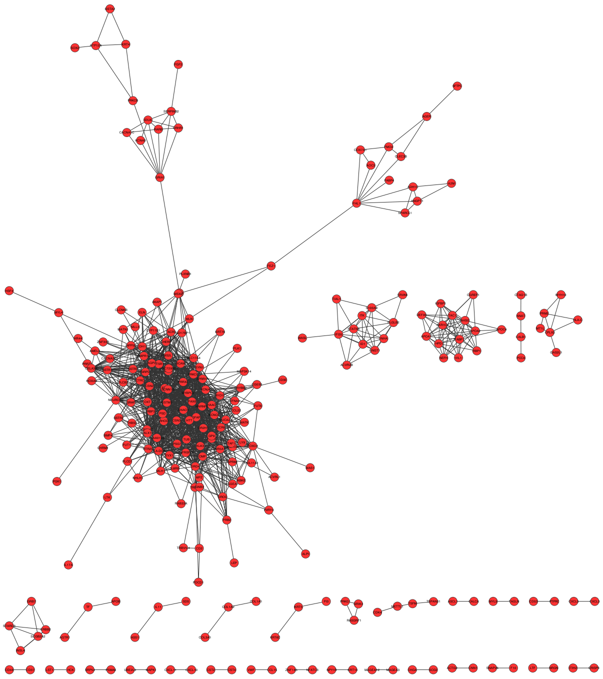


Lung-Michigan: ARACNE – $SN(C_4)$

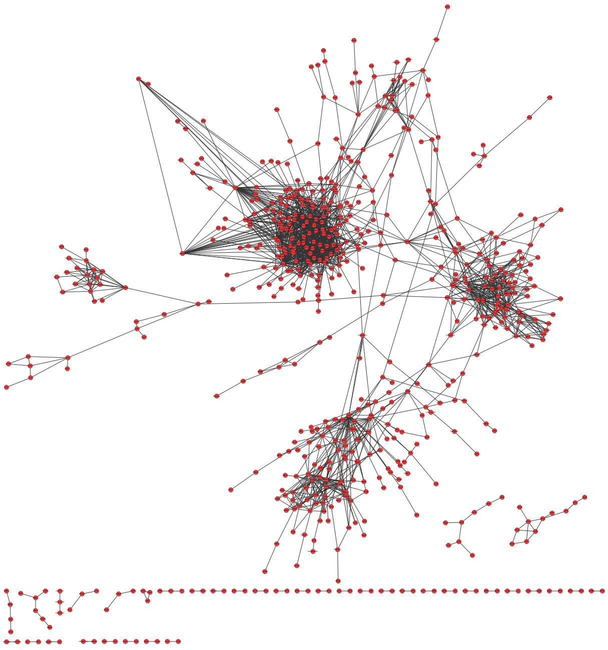




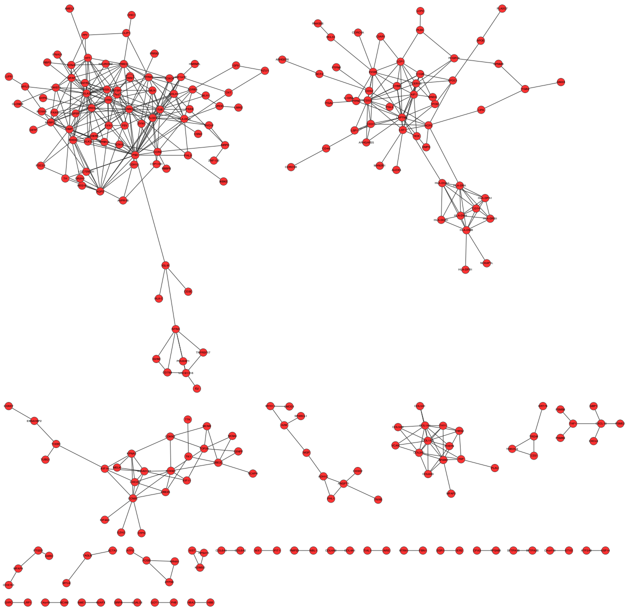
Lung-Michigan: PCC $- \, SE(C_2)$



Lung-Michigan: PCC $- \, SN(C_4)$



Lung-Michigan: MIC $- \, SE(C_2)$



Lung-Michigan: MIC $- \, SN(C_4)$

References

Cerami, E. et al (2012). The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. *Cancer Discovery*, **2**(5), 401–404.

Chowdary, D. et al (2006). Prognostic gene expression signatures can be measured in tissues collected in RNAlater preservative. *The Journal of Molecular Diagnostics*, **8**(1), 31–39.

Davis, A.P. et al (2015). The Comparative Toxicogenomics Database's 10th year anniversary: update 2015. *Nucleic Acids Research*, **43**(D1), D914–D920.

Hamosh, A. et al (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, **33**(suppl 1), D514–D517.

INSERM (1997). Orphanet: an online database of rare diseases and orphan drugs.

Magrane, M. et al (2011). UniProt Knowledgebase: a hub of integrated protein data. *Database*, **2011**.

Rappaport, N. et al (2013). MalaCards: an integrated compendium for diseases and their annotation. *Database*, **2013**.

Shannon, P. et al (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, **13**(11), 2498–2504.

Singh, D. et al (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, **1**(2), 203–209.

Taylor, B.S. et al (2010). Integrative Genomic Profiling of Human Prostate Cancer. *Cancer Cell*, **18**(1), 11–22.