D4.1

Building of cancer type-specific multi-layered molecular and patient similarity networks

Project number	826121
Project acronym	iPC
Project title	individualizedPaediatricCure: Cloud-based virtual-patient models for precision paediatric oncology
Start date of the project	1st January, 2019
Duration	48 months
Programme	H2020-SC1-DTH-2018-1

Deliverable type	Report
Deliverable reference number	D4.1
Work package contributing to the deliverable	WP4
Due date	31th December, 2020
Actual submission date	15th January, 2021

Responsible organisation	IBM
Editor	Matteo Manica, Joris Cadow, Marianyela Petrizzelli, Andrei Zinovyev, Davide Cirillo
Dissemination level	Public
Revision	V1.0

Abstract	Molecular and patient networks for specific cancer-type enables an unbiased understanding of disease mechanisms. We report the analysis performed on multiple omic levels for a collection of paediatric cancer types using single as well as consensus network inference methods.
Keywords	Molecular networks, patient networks

The project iPC has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826121.

Editor

Matteo Manica (IBM)

Joris Cadow (IBM)

Marianyela Petrizzelli (CURIE)

Andrei Zinovyev (CURIE)

Davide Cirillo (BSC)

Contributors (ordered according to beneficiary numbers)

IBM, CURIE, BSC

Disclaimer

Executive Summary

D4.1 report describes the techniques for network inference used in iPC and their application to a selection of paediatric patient cohorts at different omic levels. The goal of the deliverable is to generate the networks that will be used in the downstream WP4 tasks as well as the other project activities involving the usage of networks.

D4.1 provides a set of computational tools for generating molecular and patient networks, assembled in a computational pipeline

D4.1 provides a computational data resource containing already computed molecular and patient networks constructed for several paediatric cancer types analysed in the project.

“Data” or “Omic data” is defined as experimental measurement of a set of molecular entities of interest, e.g., gene expression (transcriptomic) data, methylation data, etc.

“Network” is defined as a set of connected entities, be it patients or molecular entities (e.g., genes, proteins, etc.), in a graph structure composed by a set of nodes and a set of edges.

“Network inference” is defined as a methodology aimed to reconstruct a network from data.

Table of Content

List of Figures

List of Tables

Chapter 1 Introduction

High-throughput technologies enabled the measurement of thousands of molecular entities and allowed an unprecedented resolution in the investigation of the internal regulatory apparatus of the cell.

Such technological advances translated into the need for appropriate methodologies, named network inference methods, that can be employed to analyze the data produced and transform the readouts from biological samples into organized maps of interaction.

These methods can play a dual role: on one hand, they allow us to build graph of interactions between the measured entities, enabling a deeper understanding of the molecular mechanisms underlying a disease; on the other hand, they can be used to construct sample, i.e., patient, similarity maps, allowing to define complex subtypes and subpopulations.

In this report, we describe three main results achieved during the work on this deliverable:

1) Implementing and validating COSIFER: a tool for computing networks of statistical associations between molecular profiles, based on winner methods from the DREAM challenge on biological network inference (Manica et al., 2020)

2) Building a pipeline for creating the collection of molecular and patient similarity networks, based on application of multiple computational methods, collected in the COSIFER package as well as the methods for integrating these networks together, using multi-omics datasets as input

3) Creating an initial corpus of paediatric cancer-specific networks computed on four cancer datasets relevant for the iPC, three of which are multi-omics datasets

The deliverable report is structured as follows. In the Chapter 2, we provide a description of the datasets considered for creating the initial set of paediatric cancer-specific networks. In Chapter 3, we describe COSIFER and the methods adopted for the analysis of the cohorts included in the study. In Chapter 4, we summarize the results of the application of the methods to selected paediatric cancer datasets, providing more details for one particular dataset. Finally, in the conclusion chapter, we summarize the achieved results and lay out a plan definition for the exploitation of the generated networks in upcoming project deliverables.

Chapter 2 Datasets used to produce the initial set of paediatric cancer-specific networks

We considered 4 multi-omics datasets from three paediatric tumors, namely one from Ewing sarcoma, two medulloblastoma and one neuroblastoma. The Ewing sarcoma data consists of gene expression profiles from 117 patients (Postel-Vinay et al., 2012). The medulloblastoma datasets come from two different cohorts of patients. The first, described in detail in (Forget et al., 2018) consists of gene expression, methylation, proteomic and phospho-proteomic profiles from 38 patients (with missing values); the second, described in detail in (Cavalli et al., 2017), consists of gene expression and methylation profiles from a cohort of 763 patients. The neuroblastoma dataset (Henrich et al., 2016) consists of microarray-based comparative genomic hybridization (aCGH), gene expression and methylation profiles from a cohort of 105 patients.

All datasets have been homogenized in terms of gene identifiers. All datasets but the medulloblastoma from (Forget et al., 2018) were downloaded from the R2 repository (https://hgserver1.amc.nl/cgi-bin/r2/main.cgi).

Table 1. Datasets used in the work on D4.1

Tumor type	Omics data	Num. genes	Samples	Reference
Medulloblastoma	Phospho-proteomics	4,476 (gene x sites)	35	(Forget et al., 2018)
	Proteomics	3,892		
	Methylation	23,348		
	Gene expression	30,257		
Medulloblastoma	Methylation	18,144	763	(Cavalli et al., 2017)
	Gene expression	18,479		
Ewing sarcoma	Gene expression	17,789	117	(Postel-Vinay et al., 2012)
Neuroblastoma	aCGH	112	105	(Henrich et al., 2016)
	Methylation	18,421		
	Gene expression	19,320		

Chapter 3 Pipeline for computing and analysis of paediatric cancer-specific networks

3.1 COSIFER package for computing networks of statistical associations between molecular profiles

COSIFER, is a package and companion web-based platform to infer molecular networks from expression data using state-of-the-art consensus approaches. COSIFER includes a selection of state-of-the-art methodologies for network inference and different consensus strategies to integrate the predictions of individual methods and generate robust networks.

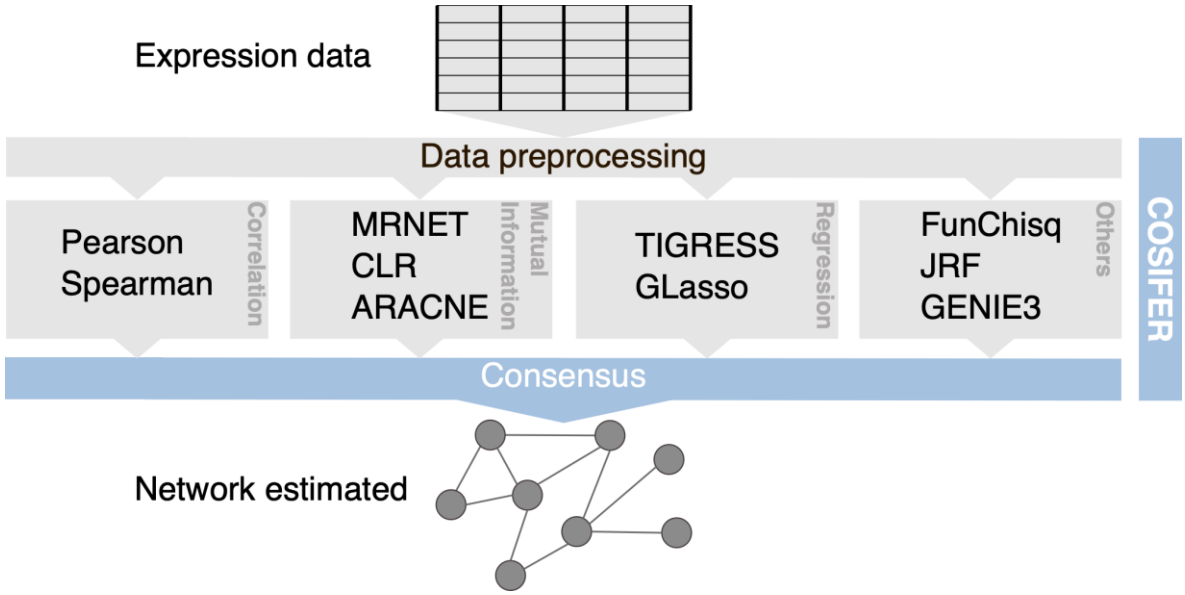

Figure 1. COSIFER workflow.

COSIFER implements 10 different unsupervised network inference methods and 3 integration strategies to produce a high-confidence consensus network. In addition, COSIFER implements several utilities to preprocess data including: on-the-fly decompression based on file extension, data standardisation, mean imputation (imputation of missing values using the mean of the measured data) and the possibility of inferring pathway-specific networks if the user provides a .gmt file. A user can select different preprocessing approaches, inference methods and consensus strategy to process expression data, and COSIFER returns pairwise interaction networks between the measured molecular entities. The resulting consensus network as well as single-method inferred networks are stored as edge lists in gzipped .csv format. The edge list is composed of triplets, listing the interacting entities and the interaction intensity, which is a real value $\in [0,1]$ associated with the strength of the predicted interaction. No threshold is applied to the output.

3.1.1 Inference methods

COSIFER implements 10 unsupervised inference methods. Algorithms have been chosen based on their inference performance, as reported in the literature (Iyer et al., 2017; Maetschke et al., 2014; Marbach et al., 2012), as well as their distinctive theoretical foundations, a key aspect for robust performance of consensus networks (Dietterich, 2000). Thus, COSIFER comprises methods based

on correlation, mutual information, regression, tree ensembles and functional χ2-test based methods.

An extensive list of the methods is reported in Table 2, we refer to the original publications for the details of each method or to the supplementary of COSIFER available online at Bioinformatics (here).

Table 2. Methods for computing statistical associations between molecular profiles implemented in COSIFER.

Method	Source
Correlation	
Pearson's correlation coefficient	(Butte & Kohane, 1999; Pearson, 1895),
Spearman's correlation coefficient	(Butte & Kohane, 1999; Spearman, 1987)
Mutual Information	
ARACNE	(Margolin et al., 2006; Meyer et al., 2008)
CLR	(Faith et al., 2007; Meyer et al., 2008)
MRNET	(Meyer et al., 2007, 2008)
Regression	
GLasso	(Friedman et al., 2008)
TIGRESS	(Haury et al., 2012)
Other Approaches	
JRF	(Petralia et al., 2016)
FunChisq	(Zhang & Song, 2013)
GENIE3	(Huynh-Thu et al., 2010)

3.1.2 Consensus methods

In a seminal work (Marbach et al., 2012), it has been shown that, assuming independence between individual methods, specifically between the ranks of the scores associated to edge predictions, allows to exploit the central limit theorem resulting in an average rank distribution that approaches a Gaussian distribution whose variance shrinks with the number of predictions, i.e., methods considered. This guarantees that a consensus approach can approximate accurately and robustly the latent edge distribution given a set of independent methods.

In COSIFER we take advantage of this notion by providing different consensus methodologies alongside the single inference methods presented above. The consensus methods implemented acts on the output of the single methods in the form of weighted adjacency matrices, i.e., real-valued matrices, where the elements indicate the strength of an interaction. In consensus mode, the matrices are first scaled between in a unitary interval using min-max scaling, to ensure they are comparable, and then combined using one of the available methodologies.

The methods implemented are: i) the Wisdom of the Crowds (WOC) (Marbach et al., 2012), where, by averaging the rank of the interactions predicted by each method and assigning a zero rank to those interactions not predicted by a method, we implement a voting scheme that is used to build the consensus network; ii) WOC (hard), a variant WOC that is computing averages based on the number of methods that actually predicted an interaction, hence explicitly accounting for sparsity of the networks inferred by the individual methods; iii) Similarity Network Fusion (SNF) (Wang et al., 2014), a method that uses a sparse kernel approximation of the similarity matrices that represent the interactions in an iterative process that converges to the resulting aggregated network and iv) SUMMA (Unsupervised Evaluation and Weighted Aggregation of Ranked Predictions) (Ahsen et al., 2018), an ensemble learning algorithm, completely unsupervised, that estimates the AUROC (Area under the Receiver Operating Characteristic) of each single inference method and associates a weight to each method proportional to its estimated AUROC. The consensus network edge intensities/score are the ones directly predicted from the different methods.

All methods are reported in Table 3 with associated references.

Table 3. Methods for computing consensus networks implemented in COSIFER

Method	Source
WOC	(Marbach et al., 2012)
WOC (hard)	(Marbach et al., 2012) (averaging over existing predictions)
SNF	(Wang et al., 2014)
SUMMA	(Ahsen et al., 2018)

3.1.3 COSIFER availability

COSIFER is available as a pip-installable Python package distributed under MIT license conditions via GitHub. The library implements utilities, functions and classes to handle different types of expression data for network inference and consensus network prediction in multiple formats. COSIFER follows object-oriented programming principles allowing an easy extension and addition of new inference methods. Upon installation the command `cosifer` is made available. The command takes as input a symbol-separated file (e.g. .csv, .tsv, etc.) containing omic data, and generates an output folder where the single/consensus inference methods results are stored as an edge list in a compressed .csv format. This ensures compatibility with most popular graph processing and visualization libraries.

The command allow the user to tweak different aspects of the data processing and inference pipeline: data standardization control; orientation of the data matrix (samples on rows or columns); inference method selection as well as the consensus method of choice; and optionally, provide a .gmt file to perform inference on all the gene sets of interest. For more details on COSIFER module and the script usage, we include in the deliverable the user guide, that is also available online (here).

To broaden its adoption, COSIFER is also available as a docker image on Docker Hub: https://hub.docker.com/r/tsenit/cosifer. The docker image is used as the basis for a Binder-hosted notebook, that shows an example how to parallelize inference workflows using COSIFER.

To ease its usage for experimental scientists, COSIFER is also available as a login-free service on IBM Cloud at https://ibm.biz/cosifer-aas. The web application integrates basic functionalities, such as data upload/processing and choice of inferences as well as consensus methods. Given the hosting server computational limitations, only WOC, WOC (hard) and SUMMA are available on the web GUI. The GUI allows selection of molecular entities of interest and supports interactive network visualization. A tutorial on how to use COSIFER web GUI is available here.

The manuscript describing the functionality of COSIFER has been published in Bioinformatics journal, with acknowledgements from iPC project, and available under open access: https://doi.org/10.1093/bioinformatics/btaa942.

3.2. Pipeline for network-based analysis of multi-omics datasets

3.2.1. General description of the pipeline

Figure 2 represents a diagrammatic view of the workflow developed for D4.1 to construct both molecular and patient similarity networks from a set of multi-omics profiles. The workflow uses several tools:

1) COSIFER package for computing networks of statistical associations between molecular profiles, using multiple methods.
2) Standard community detection algorithm using Markov Clustering (MCL).
3) Standard gene enrichment tool used in order to define the biological meaning of the communities and make a selection of them.
4) ROMA analysis which quantifies a table of eigengene scores from identified communities, which is used by COSIFER in order to construct the patient similarity network.

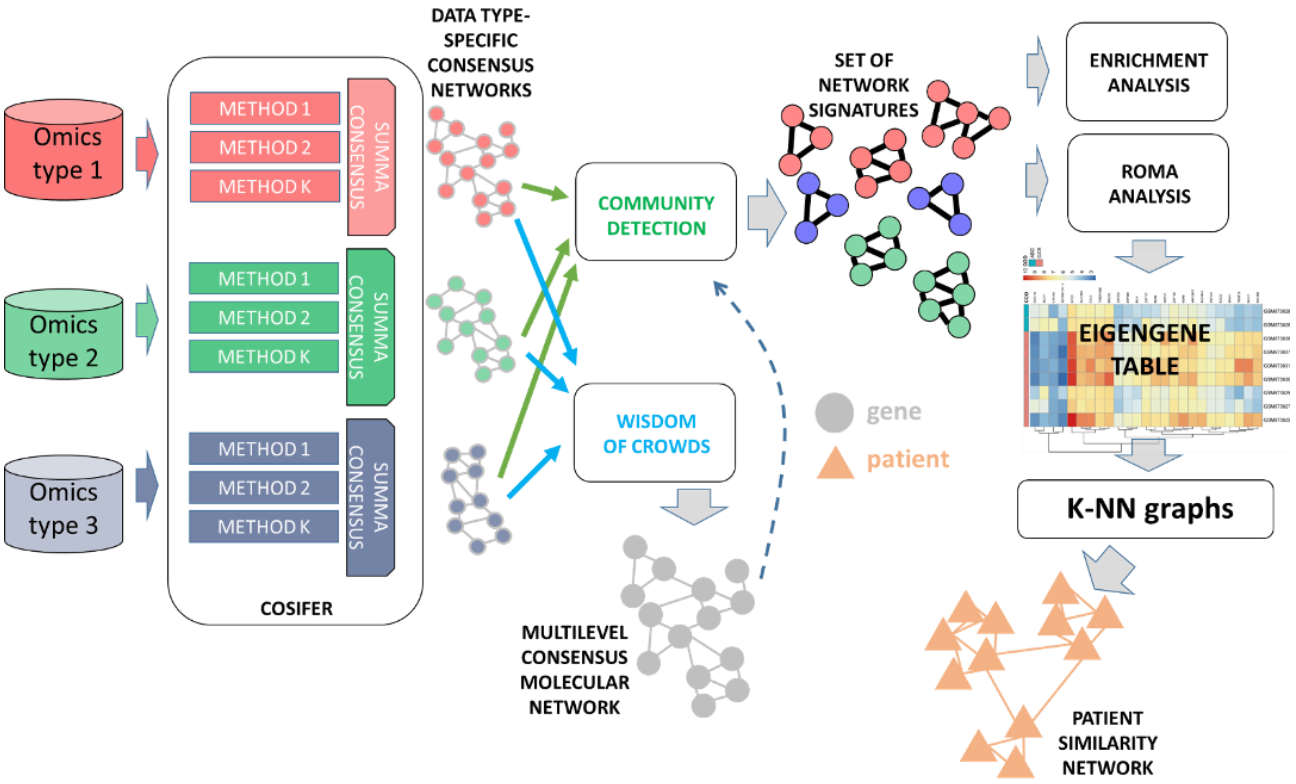

Figure 2. Schema of the WP4 pipeline for constructing paediatric cancer-specific networks from multi-omics data.

3.2.2. Construction and analysis of molecular networks

At the first step of the application of the pipeline, COSIFER package is systematically applied to all levels of a multi-omics dataset independently. A consensus across all methods for finding statistical associations network is constructed using SUMMA approach, one per each data type, which leads to creation of data type-specific consensus networks. All the networks use genes as nodes, and the strategy of mapping the molecular profile onto genes can be different for different data types. Thus, for methylation profiles, the methylation signal is summarized on the gene level before application of COSIFER, while for phospho-proteomics data, COSIFER is applied first at the level of individual protein phospho-forms, and then the edges identified for phospho-forms of the same protein are merged together, taking the maximum weight as the new edge weight.

In order to construct the multi-level consensus network, combining the information from many data types, we applied the wisdom of crowds approach (Marbach et al., 2012).

In order to extract key features from the networks inferred using COSIFER (which are very dense since they contain all statistically significant associations), we set up an *a priori* edge-weight threshold based on the feasibility of the further network analysis. We therefore created sub-networks selecting for the most confident ten thousand gene-gene statistical associations from the initially computed COSIFER network. Within the sub-network, we analyze the disconnected components separately and we select the largest connected component for further analysis. We analyze the statistical properties of these networks, i.e., degree distribution, community structure and the biological community function.

Network communities representing biological modules in the inferred network have been inferred using the mcl package in R. Willing to apply the ROMA (Martignetti et al., 2016) method for the construction of eigengene profiles, we tuned the inflation parameter of the MCL to have as many as possible clusters with more than ten genes. For finding the network communities, two approaches can be utilized. First, they can be detected from data type-specific consensus networks, having in

mind that for some data types it is almost impossible to define the community structure in the corresponding network (see example in further). Second approach consists in defining communities from the multi-level consensus network, after application of the wisdom of crowds approach (this approach is shown by a dashed line in Figure 2).

Gene enrichment analysis of the resulting communities has been performed using the gprofiler2 package in R.

3.2.3. Construction and analysis of patient similarity networks

In order to construct patient similarity networks from multi-omics data, it is necessary to define a metric (distance) between a set of molecular profiles of different types representing the same patient/tumor (Pai & Bader, 2018). This task is highly non-trivial and prone to the influence of non-biological factors affecting the variance of molecular profiles and the differences in the statistical properties of molecular profiles of different kinds. Therefore, simple merging of various omics measurements appears to be a suboptimal choice for defining the distance.

In order to define a meaningful metric, in D4.1 we decided to build a new set of features based on the analysis of networks representing different levels of multi-omics data, as described in the previous section. Each new feature represents a network community extracted from network analysis, containing at least 10 genes. In Figure 2, we call network signatures a set of such features. For computing the distance between molecular profiles, the user can either use all such communities or make an expert-based selection of those communities representing relevant biological functions, as indicated by the gene enrichment analysis performed during network analysis.

The genes composing each community have distinct profiles in each data type. In order to summarize their profiles in one score, we applied a well-established approach based on quantifying so-called eigengenes, using Representation and Quantification of Module Activity (ROMA) method (Martignetti et al., 2016). In brief, ROMA activity quantification is based on the simplest uni-factor linear model of gene regulation that approximates the expression data of a gene set by its first principal component. As a result of such quantification, a set of omics datasets for each data type and a set of network signatures defined per each data type, is summarized in the eigengene table, where rows are patients/samples, and the columns are new features (network signatures). The eigengene table contains scores of the eigengenes, quantified with the use of ROMA.

All the analysis scripts used for the deliverable that can be used to reproduce the analysis presented in the report are available here.

Chapter 4 Collection of paediatric cancer-specific networks

4.1. Creating the initial version of the computational resource as a collection of networks for paediatric cancer-

We've applied the constructed pipeline defined in the previous chapter, to four paediatric cancer datasets, described in Chapter 2. For each dataset, we created the following items:

1) Set of data type-specific consensus networks computed with COSIFER, one per each data type, containing all statistically significant associations. These networks are generally dense and require a significant amount of disk space to store them.
2) Set of reduced data type-specific consensus networks, containing only top weighted edges. These networks are amenable for the further network-based analyses and require much less disk space to store them.
3) Multi-level consensus molecular network, one per the whole multi-omics dataset, resulting from application of the wisdom of crowds approach.
4) Set of network signatures resulting from application of community detection algorithm, for each data type-specific consensus network.
5) Results of gene enrichment analyses for each network community of sufficient size (>10 genes).
6) Table of eigengene scores, keeping the results of quantification of network signatures using ROMA.
7) Patient similarity network, computed by applying k-NN to the table of eigengenes.

All these items for each of the four analyzed paediatric cancer datasets, were stored in the iPC data repository: https://data.ipc-project.bsc.es/s/wQHByo32o3JHXwS.

For the purpose of reporting, in the next section, we describe in detail the results of the application of the developed pipeline for the network-based analysis of the medulloblastoma dataset from (Forget et al., 2018).

4.2. Example network-based analysis of the multi-omics data for medulloblastoma

The multi-omics dataset from (Forget et al., 2018) is a relatively small (35 samples) paediatric cancer dataset but it has the advantage of containing four levels of data types, amenable for the network-based analysis (gene expression, protein expression, DNA methylation and phospho-proteomics levels). In the further we provide details of D4.1 pipeline application to this dataset at the individual data type levels for the analysis of molecular networks and construction of patient-patient similarity networks.

4.2.1. Level of protein expression

The complete network, computed by COSIFER, consisted of 3,892 nodes and 7,211,426 edges, which was too dense to apply network-based analysis. The subgraph obtained by considering the 10^5 strongest statistical associations contained 3,316 nodes. Within the subgraph there are 2

connected components, the largest one consisting of 3,314 nodes. The distribution of weights for both complete and reduced networks is shown in Figure 3.

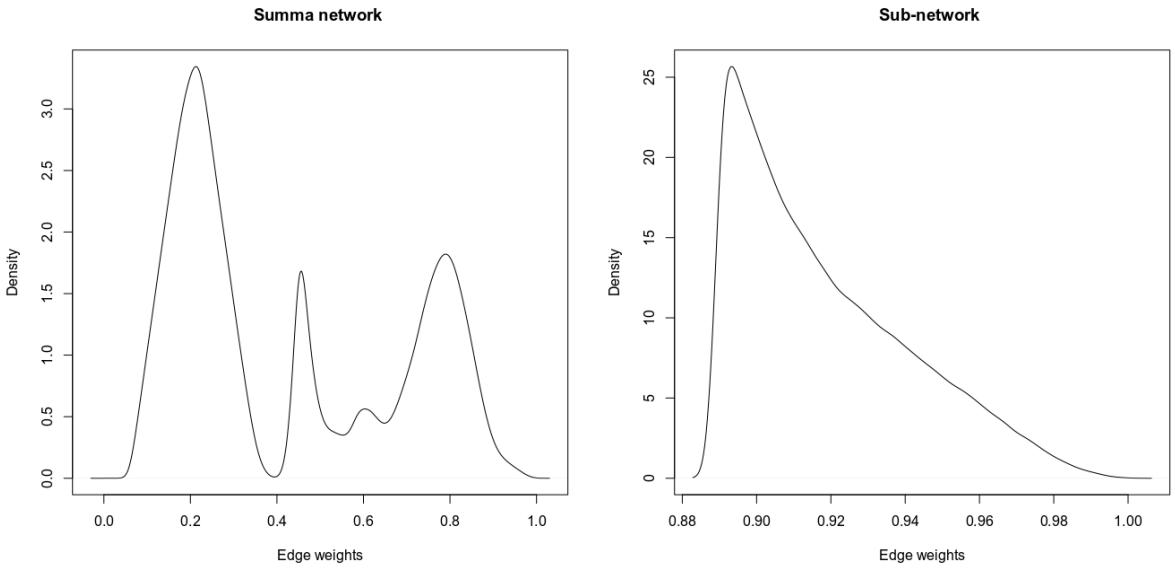

Figure 3. Distribution of edge weights for SUMMA inference network and the reduced subnetwork for the protein expression level of the medulloblastoma dataset.

We applied the standard MCL algorithm in order to identify network communities in the reduced network. The main parameter in the application of this algorithm is inflation, which defines the granularity of the resulting communities. Therefore, we need to optimize this parameter and in order to do this we used the following criterion: the number of communities containing more than 10 genes. Optimizing this criterion allows us to achieve better interpretability of the resulting communities. For example, MCL clustering with inflation parameter equal to 1.8 allows us to identify 45 clusters (network communities) with more than 10 genes. With this criterion we obtain 399 communities among which 136 were composed by just 2 genes. The distribution of the number of genes found in the clusters is shown in Figure 4.

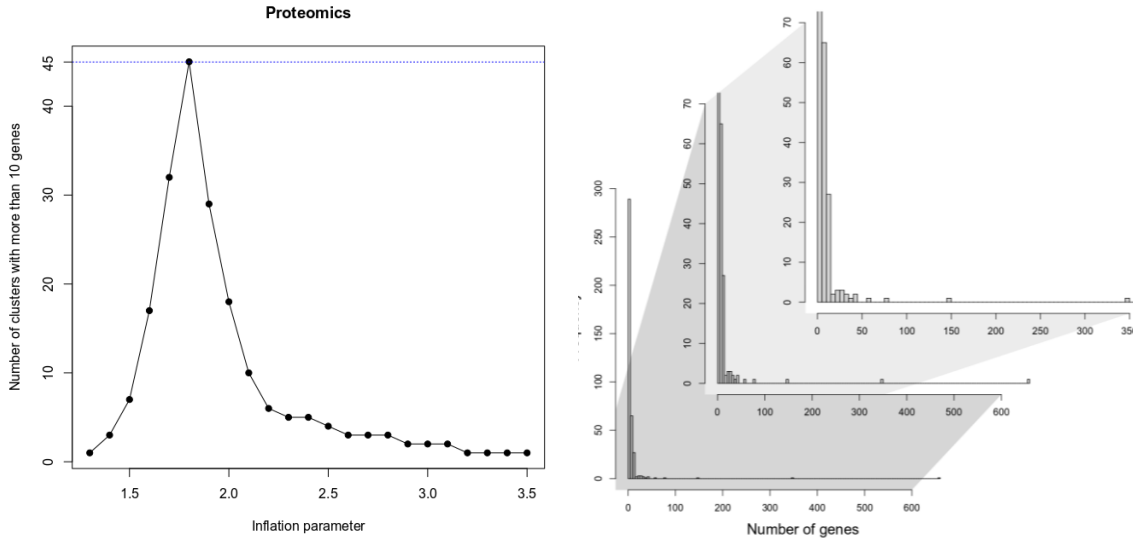

Figure 4. A) Inflation parameter versus the number of clusters with more than 10 genes in the medulloblastoma network, computed for protein expression data. B) Histogram representing the frequency of clusters for a given number of genes.

The largest community consists of 656 nodes and is enriched in proteins linked to the spliceosome, exocyst Sec6/8 complex, Exocyst complex, SF3b complex and THO complex. The second largest community, composed of 348 nodes, was found to be enriched in proteins involved in EIF3 complex, cellular component organization or biogenesis, ribonucleoprotein complex biogenesis and viral process. ROMA was applied with the 45 network communities (clusters) with more than 10 genes. The resulting eigenGene matrix consists of 39 eigengenes and 35 samples. Visualization of these communities annotated with gene enrichment analysis results is provided in Figure 5.

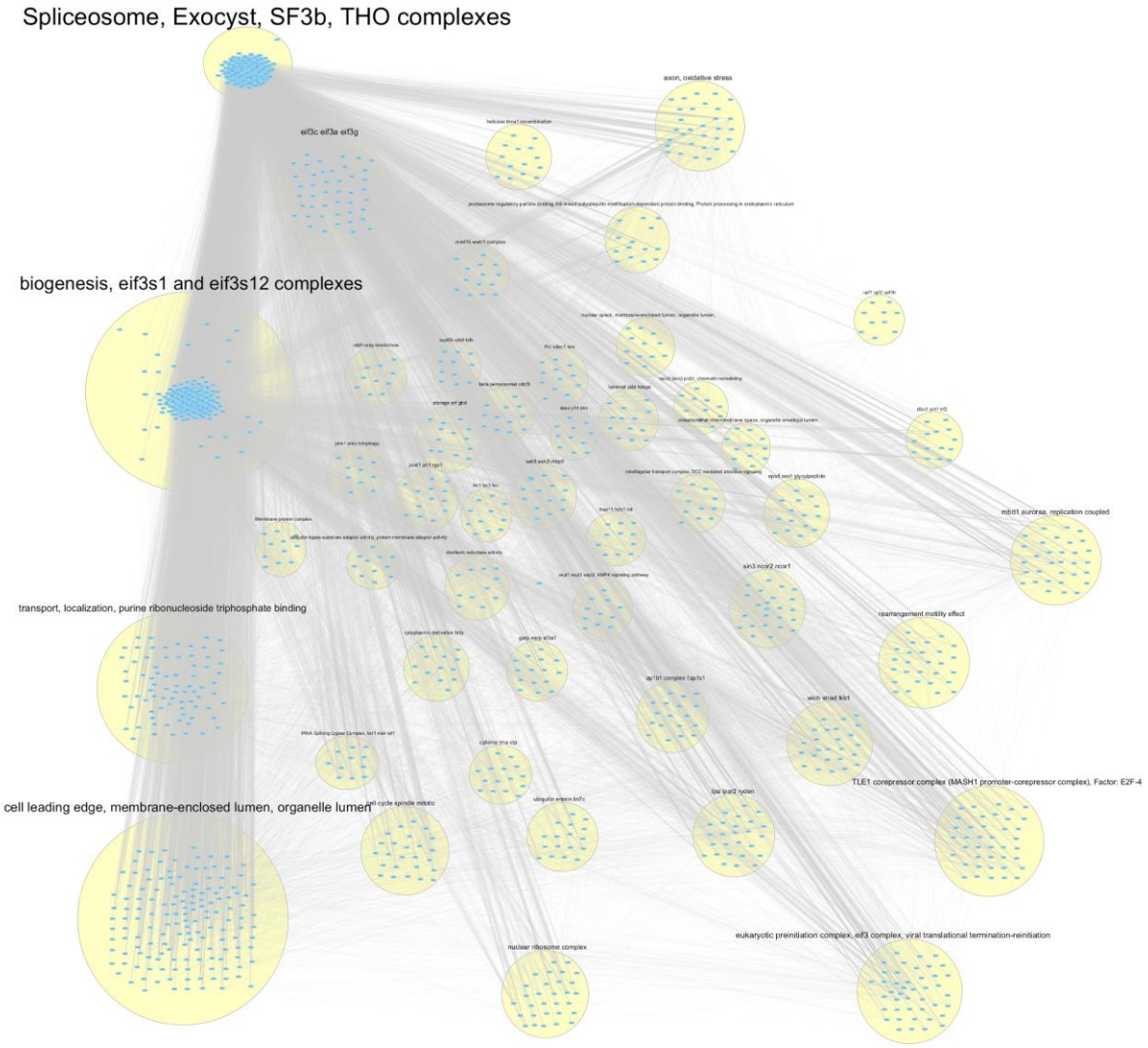

Figure 5. Network communities identified at the level of protein expression in the analysis of medulloblastoma dataset.

In Figure 5 each circle represents a community, annotated by the results of gene enrichment analysis. The connections between genes belonging to different communities are also shown.

4.2.2. Level of DNA methylation

The complete network computed by COFISER consisted of 23,348 nodes and 215,023,696 edges. The subgraph obtained by considering the 10^5 strongest interactions consisted of 3,704 nodes.

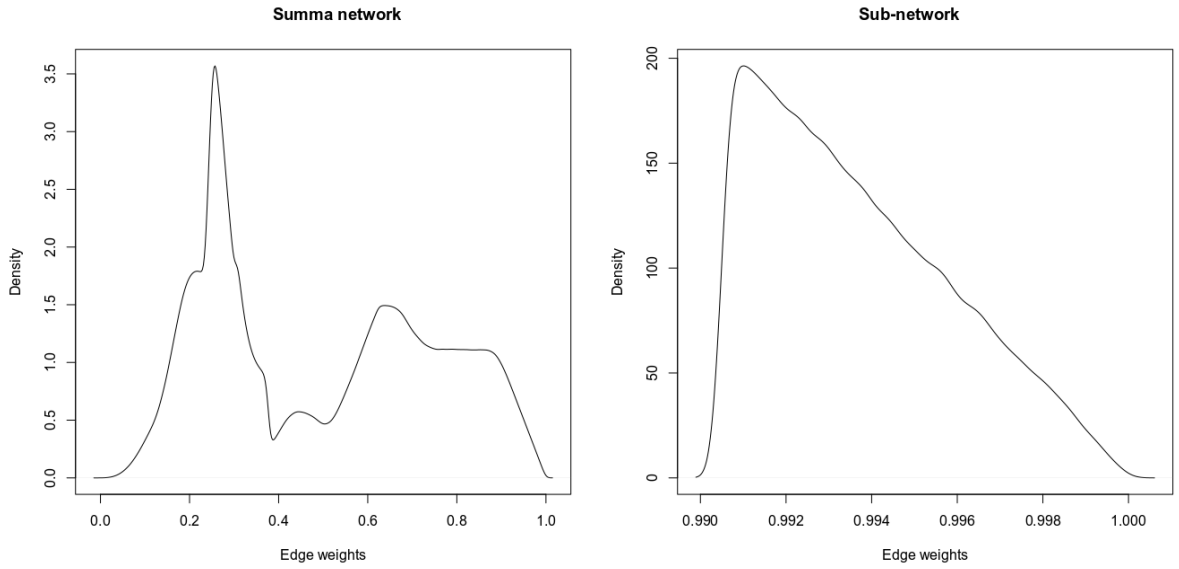

Figure 6. Distribution of edge weights for SUMMA inference network and the reduced subnetwork for the DNA methylation level of the medulloblastoma dataset.

Within the subgraph, there were no disconnected components. MCL clustering with inflation parameter equal to 1.8 leads to 29 clusters with more than 10 nodes (Figure 7A). The distribution of cluster cardinality is shown in Figure 7B. It identified 344 communities. The largest one consists of 1,641 genes, whereas 150 communities consisted of only two genes.

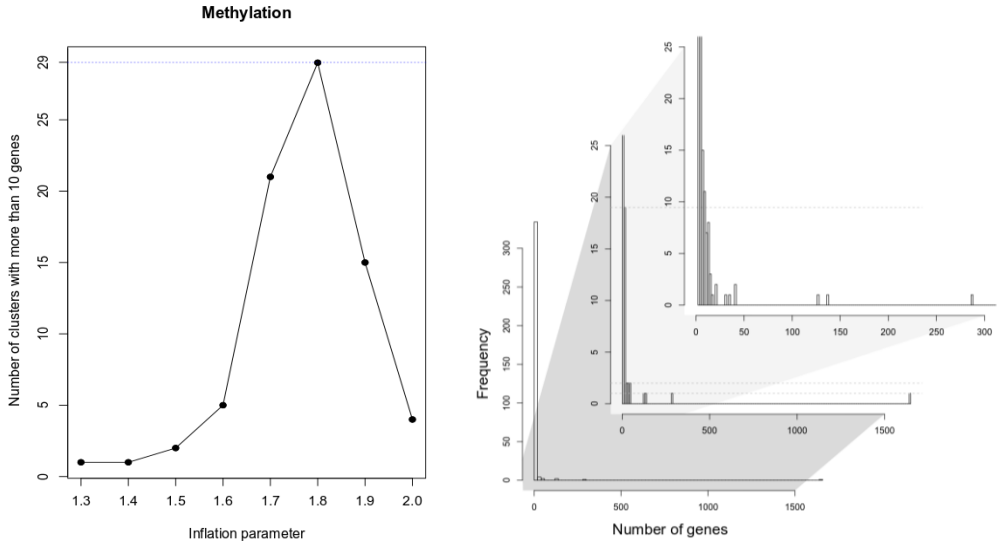

Figure 7. A) Inflation parameter versus the number of clusters with more than 10 genes in the medulloblastoma network, computed for DNA methylation data. B) Histogram representing the frequency of clusters for a given number of genes.

The largest component was associated with localization, transport and the nervous system. The second largest connected component consisted of 287 nodes and consisted of genes associated with the RUNX1-CBF-beta-DNA complex, Factor: ZF5. Overall communities were associated with specific gene complexes and functions. Among the inferred communities with more than 10 nodes only one could not be associated with a biological function. Enrichment results are summarized in Figure 8.

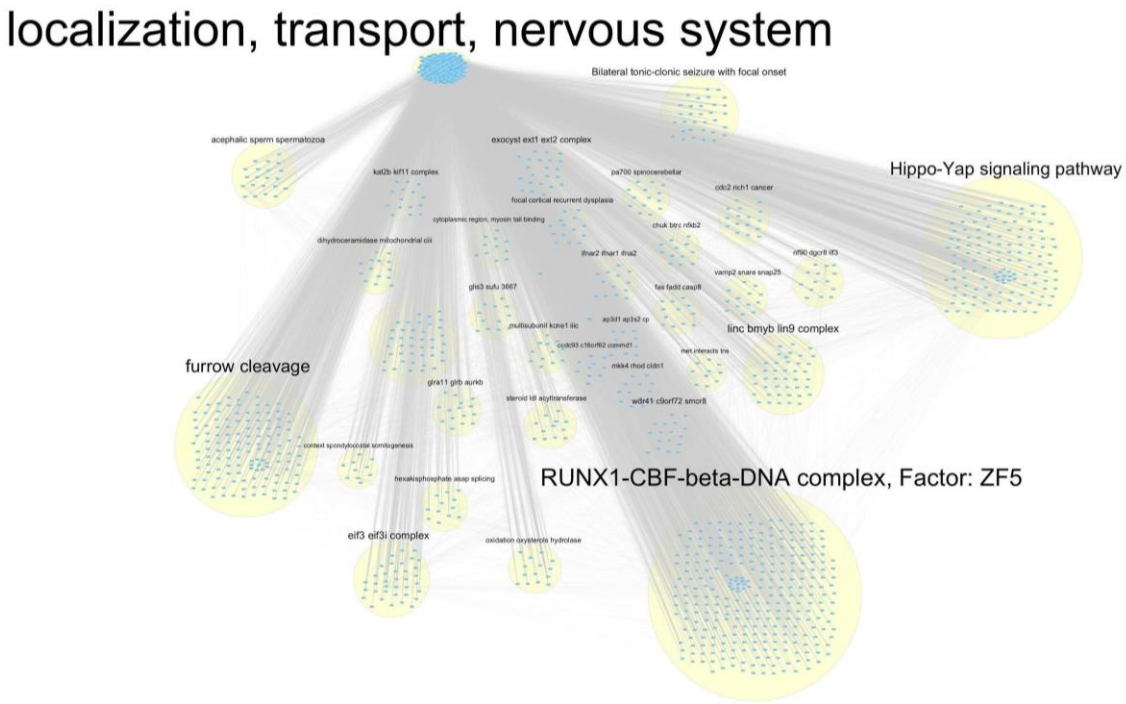

Figure 8. Network communities identified at the level of DNA methylation in the analysis of medulloblastoma dataset.

In Figure 8 each circle represents a community, annotated by the results of gene enrichment analysis. The connections between genes belonging to different communities are also shown.

ROMA was thus applied on the 29 clusters with more than 10 genes. The resulting eigenGene matrix consists of 25 eigenGenes and 35 samples.

4.2.3. Level of gene expression

For gene expression data, the complete network computed by COFISER consisted of 30,257 nodes and more than 4*10^8 edges. The subgraph obtained by considering the 10^5 strongest interactions was connected and consisted of 4,889 nodes.

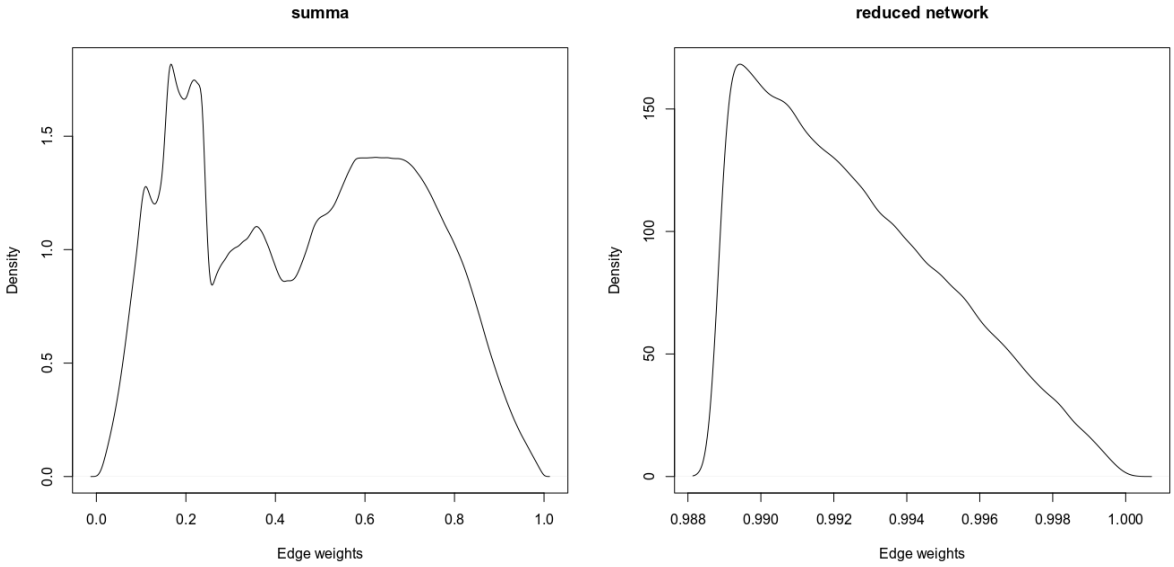

Figure 9. Distribution of edge weights for SUMMA inference network and the reduced subnetwork for the gene expression level of the medulloblastoma dataset.

MCL clustering with inflation parameter equal to 1.7 leads to 88 clusters with more than 10 nodes (Figure 10A). It identified 644 communities, the largest one consisting of 837 genes, whereas 186 communities consisted of only two genes. The distribution of cluster cardinality is shown in Figure 10B.

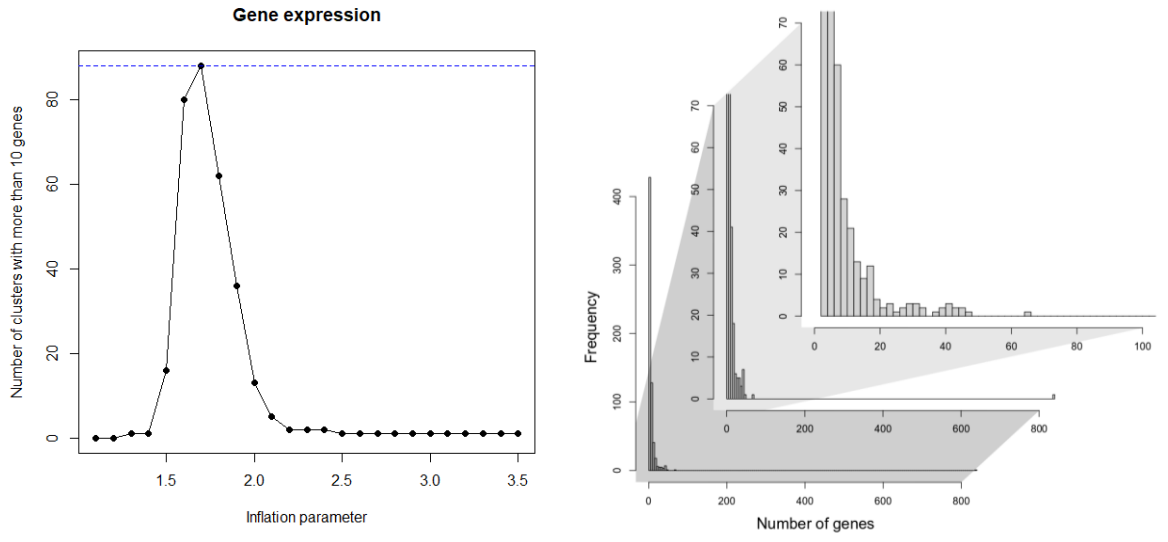

Figure 10. A) Inflation parameter versus the number of clusters with more than 10 genes in the medulloblastoma network, computed for gene expression data. B) Histogram representing the frequency of clusters for a given number of genes.

The largest community was enriched in genes associated with chemical homeostasis, intrinsic component of plasma membrane, plasma membrane region, integral component of plasma membrane, ion channel complex. The second largest community consists of 66 genes and is associated with 3'-5' DNA helicase activity, four-way junction helicase activity and lymph node function.

EigenGenes were calculated using 88 clusters, 6 of which were filtered out after outliers detection. The resulting matrix consisted of 82 eigen genes and 35 samples.

4.2.4. Level of phospho-proteomics

The complete network computed by COFISER consisted of 4,476 nodes and 9,703,387 edges. The subgraph obtained by considering the 10^5 strongest interactions consisted of 3,220 nodes.

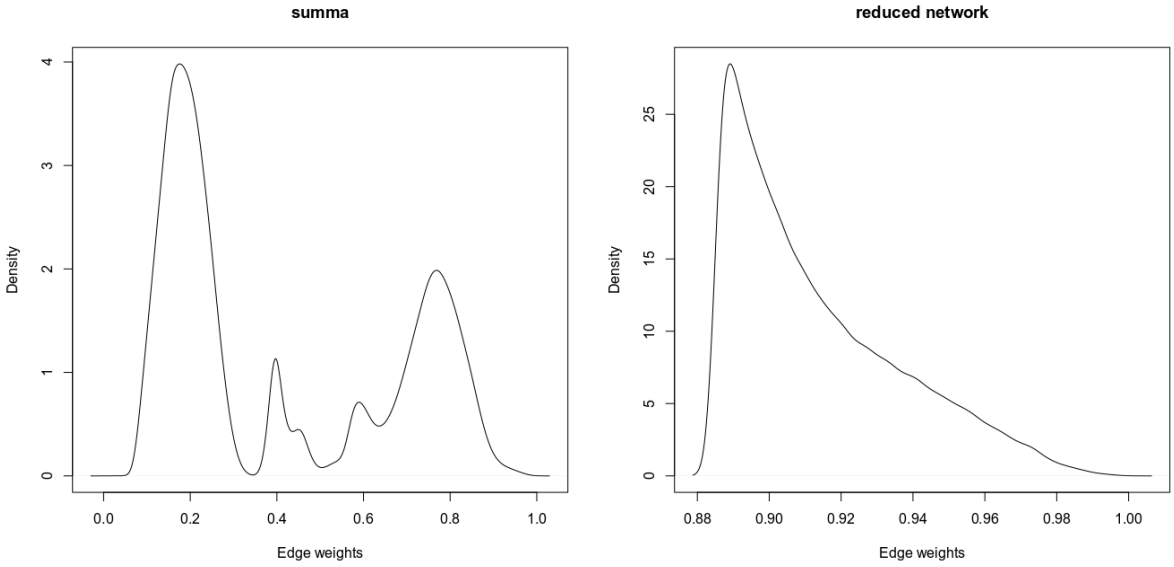

Figure 11. Distribution of edge weights for SUMMA inference network and the reduced subnetwork for the phospho-proteomics level of the medulloblastoma dataset.

MCL clustering with inflation parameter equal to 1.8 leads to 54 clusters with more than 10 nodes (Figure 10A). It identified 173 communities, the largest one consisting of 438 genes, whereas 109 communities consisted of only two genes. The distribution of cluster cardinality is shown in Figure 10B.

EigenGenes were calculated using 54 clusters, 2 of which were filtered out after outliers detection. The resulting matrix consisted of 52 eigen genes and 35 samples.

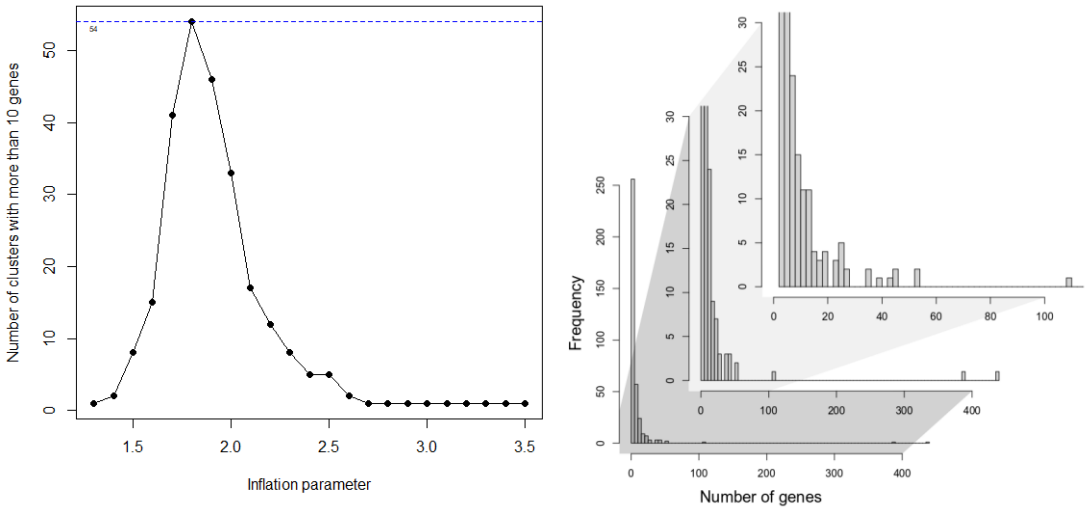

Figure 12. A) Inflation parameter versus the number of clusters with more than 10 genes in the medulloblastoma network, computed for phospho-proteomic data. B) Histogram representing the frequency of clusters for a given number of genes.

4.2.5. Patient-Patient similarity network

Patient-patient similarity network has been constructed using as an underlying truth the eigenGene matrices produced at the individual level. Indeed, a preliminary PCA analysis of the eigen gene matrices shows that the 4 different medulloblastoma are characterized at each individual level by a combination of eigengenes (Figure 13).

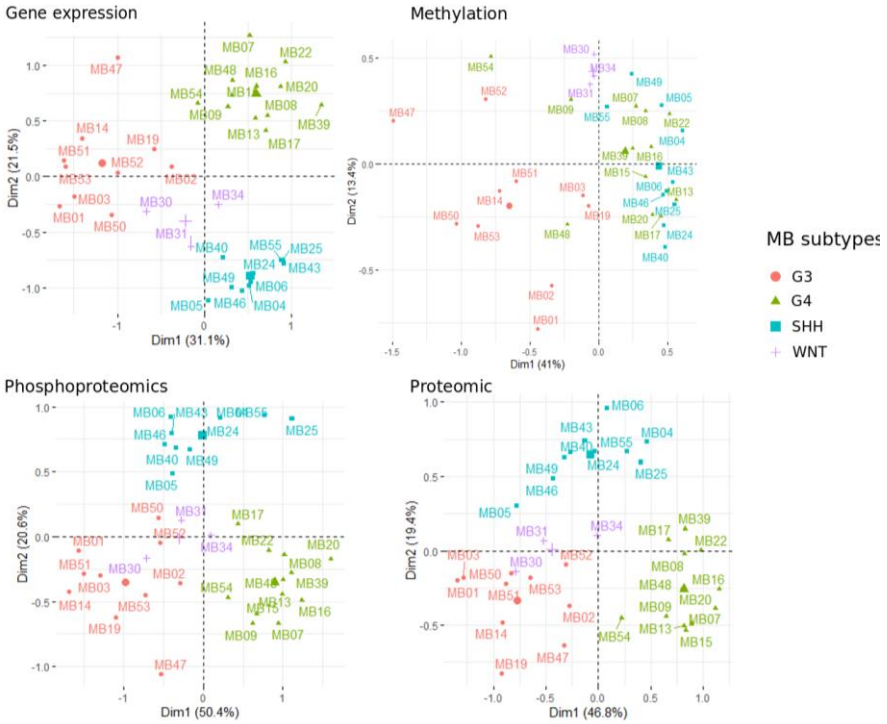

Figure 13. Principal component analysis for the samples x eigengenes matrices retrieved using ROMA at each individual level. Samples are colored according to the medulloblastoma subtype they belong to.

Specifically, eigenGene matrices are concatenated to produce a global description of the samples, that resulted in a matrix of 35 samples and 198 eigenGenes (39 from proteomic level, 52 from phospho-proteomic, 25 from methylation and 82 from gene expression level).

Using as a metric the Euclidean distance between samples, we computed the K-nearest neighbor graphs for different values of k (Figure 14). We see that the k-NN network allows to cluster samples from the 4 well known medulloblastoma subgroups namely WNT, SHH, G3 and G4.

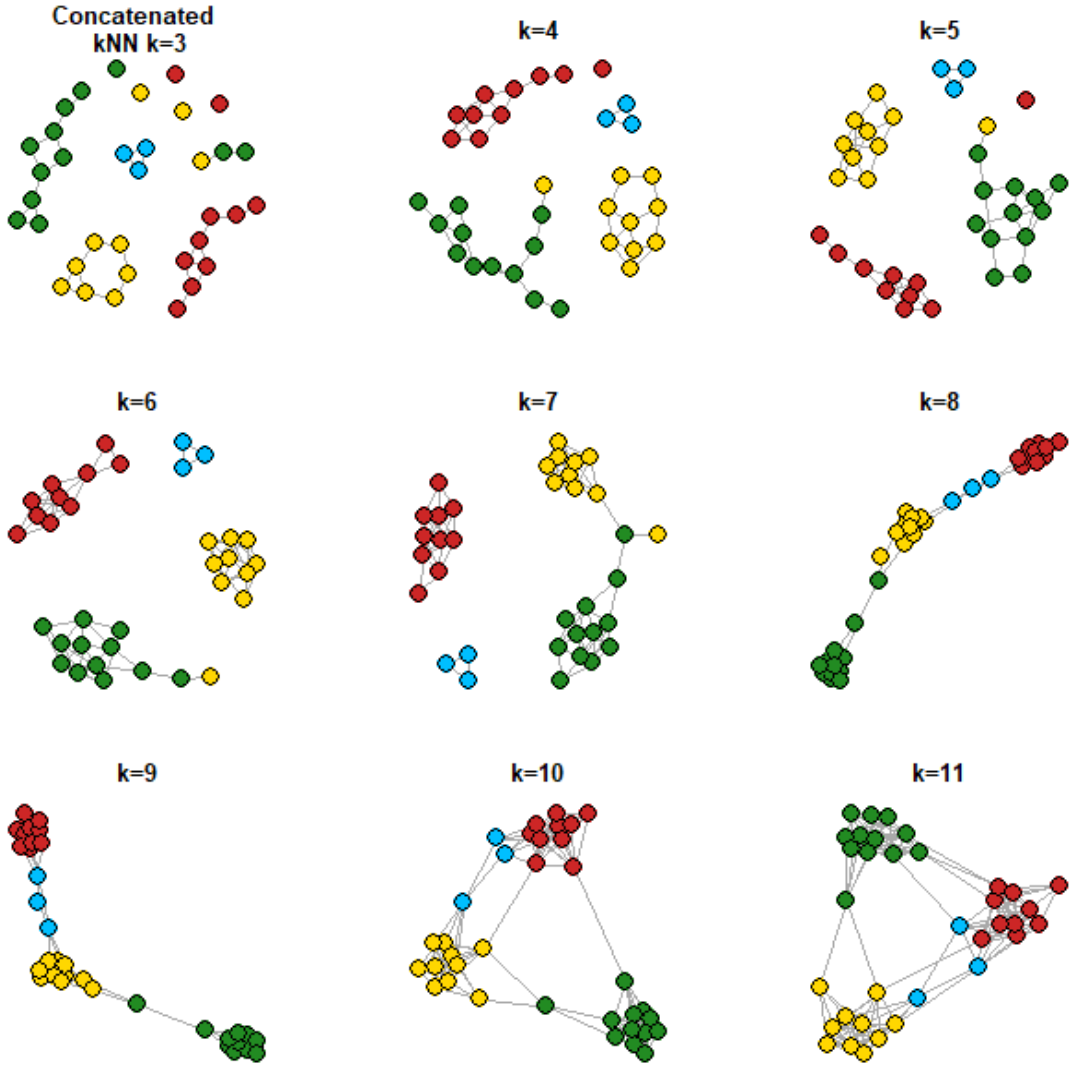

Figure 14. Patient similarity networks obtained using k-NN the concatenated eigen gene matrix using as a metric the Euclidean distance between samples, for k = 3, …, 11. Color codes: WNT blue, SHH red, G3 yellow, G4 green.

Chapter 5 Conclusions and future work

The work on this deliverable produced the required methodology for the network-based analysis of paediatric cancer multi-omics datasets, and the first version of paediatric cancer-specific network data. Therefore, the objectives of D4.1 have been fully accomplished.

In the future the initiated collection of network resources will be extended with the analysis of other paediatric cancer datasets, from the same types considered in D4.1 or other types (in particular, leukemia and hepatoblastoma). A particular interest is applying the network analysis pipeline developed for bulk datasets, to single cell data which are becoming increasingly available for the specified paediatric datasets. The extended collection of networks will be used to define molecular mechanisms driving the paediatric cancer progression, as a part of the work on D4.3, and for refining the definition of molecular subtypes of paediatric cancers as a part of the work on D4.4.

In the framework of iPC, network-based analysis will be compared with the results of application of matrix factorization methods (part of WP3). The constructed networks should be a valuable resource for the mathematical modeling efforts within iPC project.

The network communities computed and stored in the iPC project database will be exposed to the final user for interactive online browsing with the use of NaviCell network visualization platform, with a possibility to visualize various sources of omics data on top of the network visualization. This will be accomplished as a part of D4.2.

To demonstrate the analytical advantage of network representation for paediatric cancer-specific datasets, we anticipate here initial results of the application of a new methodology for the functional analysis of the medulloblastoma proteogenomic dataset described in the previous sections (Forget et al., 2018). In particular, we have developed a methodology for dimensionality reduction, based on multilayer network community detection (Didier et al., 2015), aiming to facilitate the molecular characterization of patient stratification. This graph-based approach allows to identify the minimal set of genes that characterize disease subgroups based on their persistent association in the multilayer network at different levels of community resolution. By applying this method, we have achieved a highly accurate reconstruction of the known medulloblastoma subtypes (accuracy > 94%) while offering a clear characterization of the associated gene functions, with downstream implications for diagnosis and therapeutic interventions. More details about this work, which is currently under review by a scientific publisher, will be provided in the next deliverable D4.2.

The applicability and benefits of network-based approaches to multi-omics data integration are demonstrated by the development of efficient computational tools for the analysis and interpretation of large volumes of heterogeneous biomedical data. Nevertheless, graph representation of big data entails challenges and obstacles which are mainly related to the processing and analysis of highly dense networks. In this regard, recent algorithmic implementations, such as methods for graph sparsification, allow overcoming the difficulties associated with these aspects. Graph sparsification aims to approximate an arbitrary graph, such as a large and densely connected graph, by a graph with fewer edges or vertices while maintaining essential structural properties. Graph sparsification methods with direct applications in the biomedical area, such as the disparity filter algorithm (Serrano et al., 2009) and the distance closure algorithm (Simas & Rocha, 2015), provide a robust frameworks for studying highly dense networks which can be conveniently used in the context of iPC network-based analytical workflows as a part of D4.2 or further deliverables.

Bibliography

Ahseny, M. E., Vogelzy, R. M., & Stolovitzkyzy, G. A. (2018). Unsupervised evaluation and weighted aggregation of ranked predictions. *ArXiv*, *20*(166), 1–40. http://jmlr.org/papers/v20/18-094.html.

Butte, A. J., & Kohane, I. S. (1999). Unsupervised knowledge discovery in medical databases using relevance networks. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, 711–715. /pmc/articles/PMC2232846/?report=abstract

Cavalli, F. M. G., Remke, M., Rampasek, L., Peacock, J., Shih, D. J. H., Luu, B., Garzia, L., Torchia, J., Nor, C., Morrissy, A. S., Agnihotri, S., Thompson, Y. Y., Kuzan-Fischer, C. M., Farooq, H., Isaev, K., Daniels, C., Cho, B. K., Kim, S. K., Wang, K. C., … Taylor, M. D. (2017). Intertumoral Heterogeneity within Medulloblastoma Subgroups. *Cancer Cell*, *31*(6), 737-754.e6. https://doi.org/10.1016/j.ccell.2017.05.005

Didier, G., Brun, C., & Baudot, A. (2015). Identifying communities from multiplex biological networks. *PeerJ*, *2015*(12). https://doi.org/10.7717/peerj.1525

Dietterich, T. G. (2000). Ensemble methods in machine learning. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *1857 LNCS*, 1–15. https://doi.org/10.1007/3-540-45014-9_1

Faith, J. J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J. J., & Gardner, T. S. (2007). Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biology*, *5*(1), 0054–0066. https://doi.org/10.1371/journal.pbio.0050008

Forget, A., Martignetti, L., Puget, S., Calzone, L., Brabetz, S., Picard, D., Montagud, A., Liva, S., Sta, A., Dingli, F., Arras, G., Rivera, J., Loew, D., Besnard, A., Lacombe, J., Pagès, M., Varlet, P., Dufour, C., Yu, H., … Ayrault, O. (2018). Aberrant ERBB4-SRC Signaling as a Hallmark of Group 4 Medulloblastoma Revealed by Integrative Phosphoproteomic Profiling. *Cancer Cell*, *34*(3), 379-395.e7. https://doi.org/10.1016/j.ccell.2018.08.002

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, *9*(3), 432–441. https://doi.org/10.1093/biostatistics/kxm045

Haury, A. C., Mordelet, F., Vera-Licona, P., & Vert, J. P. (2012). TIGRESS: Trustful Inference of Gene REgulation using Stability Selection. *BMC Systems Biology*, *6*(1), 145. https://doi.org/10.1186/1752-0509-6-145

Henrich, K. O., Bender, S., Saadati, M., Dreidax, D., Gartlgruber, M., Shao, C., Herrmann, C., Wiesenfarth, M., Parzonka, M., Wehrmann, L., Fischer, M., Duffy, D. J., Bell, E., Torkov, A., Schmezer, P., Plass, C., Höfer, T., Benner, A., Pfister, S. M., & Westermann, F. (2016). Integrative genome-scale analysis identifies epigenetic mechanisms of transcriptional deregulation in unfavorable neuroblastomas. *Cancer Research*, *76*(18), 5523–5537. https://doi.org/10.1158/0008-5472.CAN-15-2507

Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., & Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE*, *5*(9), e12776. https://doi.org/10.1371/journal.pone.0012776

Iyer, A. S., Osmanbeyoglu, H. U., & Leslie, C. S. (2017). Computational methods to dissect gene regulatory networks in cancer. In *Current Opinion in Systems Biology* (Vol. 2, pp. 115–122). Elsevier Ltd. https://doi.org/10.1016/j.coisb.2017.04.004

Maetschke, S. R., Madhamshettiwar, P. B., Davis, M. J., & Ragan, M. A. (2014). Supervised, semi-supervised and unsupervised inference of gene regulatory networks. *Briefings in Bioinformatics*, *15*(2), 195–211. https://doi.org/10.1093/bib/bbt034

Manica, M., Bunne, C., Mathis, R., Cadow, J., Ahsen, M. E., Stolovitzky, G. A., & Martínez, M. R.

(2020). COSIFER: a Python package for the consensus inference of molecular interaction networks. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btaa942

Marbach, D., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., Allison, K. R., Kellis, M., Collins, J. J., Aderhold, A., Stolovitzky, G., Bonneau, R., Chen, Y., Cordero, F., Crane, M., Dondelinger, F., Drton, M., Esposito, R., Foygel, R., … Zimmer, R. (2012). Wisdom of crowds for robust gene network inference. *Nature Methods*, *9*(8), 796–804. https://doi.org/10.1038/nmeth.2016

Margolin, A. A., Wang, K., Lim, W. K., Kustagi, M., Nemenman, I., & Califano, A. (2006). Reverse engineering cellular networks. *Nature Protocols*, *1*(2), 662–671. https://doi.org/10.1038/nprot.2006.106

Martignetti, L., Calzone, L., Bonnet, E., Barillot, E., & Zinovyev, A. (2016). ROMA: Representation and quantification of module activity from target expression data. *Frontiers in Genetics*, *7*(FEB). https://doi.org/10.3389/fgene.2016.00018

Meyer, P. E., Kontos, K., Lafitte, F., & Bontempi, G. (2007). Information-theoretic inference of large transcriptional regulatory networks. *Eurasip Journal on Bioinformatics and Systems Biology*, *2007*(1), 1–9. https://doi.org/10.1155/2007/79879

Meyer, P. E., Lafitte, F., & Bontempi, G. (2008). Minet: A r/bioconductor package for inferring large transcriptional networks using mutual information. *BMC Bioinformatics*, *9*(1), 1–10. https://doi.org/10.1186/1471-2105-9-461

Pai, S., & Bader, G. D. (2018). Patient Similarity Networks for Precision Medicine. In *Journal of Molecular Biology* (Vol. 430, Issue 18, pp. 2924–2938). Academic Press. https://doi.org/10.1016/j.jmb.2018.05.037

Pearson, K. (1895). VII. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, *58*(347–352), 240–242. https://doi.org/10.1098/rspl.1895.0041

Petralia, F., Song, W. M., Tu, Z., & Wang, P. (2016). New Method for Joint Network Analysis Reveals Common and Different Coexpression Patterns among Genes and Proteins in Breast Cancer. *Journal of Proteome Research*, *15*(3), 743–754. https://doi.org/10.1021/acs.jproteome.5b00925

Postel-Vinay, S., Véron, A. S., Tirode, F., Pierron, G., Reynaud, S., Kovar, H., Oberlin, O., Lapouble, E., Ballet, S., Lucchesi, C., Kontny, U., González-Neira, A., Picci, P., Alonso, J., Patino-Garcia, A., De Paillerets, B. B., Laud, K., Dina, C., Froguel, P., … Delattre, O. (2012). Common variants near TARDBP and EGR2 are associated with susceptibility to Ewing sarcoma. *Nature Genetics*, *44*(3), 323–327. https://doi.org/10.1038/ng.1085

Serrano, M. Á., Boguñá, M., & Vespignani, A. (2009). Extracting the multiscale backbone of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(16), 6483–6488. https://doi.org/10.1073/pnas.0808904106

Simas, T., & Rocha, L. M. (2015). Distance closures on complex networks. *Network Science*, *3*(2), 227–268. https://doi.org/10.1017/nws.2015.11

Spearman, C. (1987). The proof and measurement of association between two things. By C. Spearman, 1904. *The American Journal of Psychology*, *100*(3–4), 441–471. https://doi.org/10.2307/1422689

Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., & Goldenberg, A. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods*, *11*(3), 333–337. https://doi.org/10.1038/nmeth.2810

Zhang, Y., & Song, M. (2013). *Deciphering Interactions in Causal Networks without Parametric Assumptions*. http://arxiv.org/abs/1311.2707