Individualized
PaediatricCure

D4.3

Topological analysis of multi-omics and multi-cancer molecular networks resulting in the definition of molecular mechanisms

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	53 months
Programme	H2020-SC1-DTH-2018-1

Deliverable type	Report
Deliverable reference number	SC1-DTH-07-826121 / D4.3
Work package contributing to the deliverable	WP4
Due date	May 2022 – M41
Actual submission date	31st May 2022

Responsible organisation	CURIE
Editor	Andrei Zinovyev
Dissemination level	PU
Revision	1.0

Abstract	Three types of network-based analysis of gene-gene interaction networks have been suggested and tested on the multi-omics paediatric cancer datasets. User-friendly computational environment for joint application of matrix factorization and network analysis has been implemented.
Keywords	Multi-omics data; Neuroblastoma; networks; modules; survival analysis

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

Editor

Andrei Zinovyev (CURIE)

Jane Merlevede (CURIE)

Contributors (ordered according to beneficiary numbers)

Petrizzelli Marianyela (CURIE)

Davide Cirillo (BSC)

Disclaimer

The information in this document is provided "as is", and no guarantee or warranty is given that the information is fit for any particular purpose. The users thereof use the information at their sole risk and liability.
This document has gone through the consortium's internal review process and is still subject to the review of the European Commission. Updates to the content may be made at a later stage.

Executive Summary

The deliverable D4.3 presents the results of application of network-based analysis of multi-omics data methods to characterize the molecular mechanisms of paediatric cancers. Three distinct approaches have been presented, based on joint application of matrix factorization and analysing network topology. The analysis reveals some proteins and genes known to participate in the tumorigenesis of particular tumour types as well as reveal novel candidates to be investigated for the role of cancer driver genes of potential prognostic or diagnostic biomarkers.

A part of the deliverable is an interested grated user-friendly environment BIODICA which allows users without experience of programming to apply matrix factorization and interpret the results using several types of network-based analyses.

As a part of the deliverable, we curated and harmonized a previously unpublished multi-omics neuroblastoma dataset (kindly provided by AMC partner, Dr. Jan Koster), including 5 levels of omics description and a comprehensive set of clinical features including survival data. This dataset represents a still rare type of multiomics descriptions of paediatric tumours. It can serve in the future either for investigating the multi-omics mechanisms of neuroblastoma progression or developing mutli-omics signatures for better prognosis and diagnosis or benchmarking existing methods. The dataset can be accessed by the iPC consortium via NextCloud platform at https://data.ipc-project.bsc.es/s/PjBPxJFjn5JJQEq (login required).

The software code used to produce this deliverable is available from the following web-sites:

1) BIODICA tool : https://sysbio-curie.github.io/biodica-environment/
2) Code for performing systematic network-based OFTEN analysis in order to identify subnetworks of PPI interactions associated to metagenes extracted with matrix factorization: provided as a Python notebook from iPC github: https://github.com/iPC-project-H2020/wp4-deliverables/tree/main/D4.3
3) Method for constructing a multi-omics gene-gene association network: provided as a Python notebook from iPC github: https://github.com/iPC-project-H2020/wp4-deliverables/tree/main/D4.3

The results of the analyses performed are available at https://data.ipc-project.bsc.es/s/PjBPxJFjn5JJQEq (login required).

Table of Content

List of Figures

List of Tables

Chapter 1 Introduction

Network-based analyses of multi-omics data have shown to be successful in unravelling the molecular mechanisms underlying oncogenesis and providing biomarkers for predicting the success of cancer treatment outcome [Dimitrakopoulos et al, 2018, Colaprico et al, 2020, as examples]. By *networks* in this kind of analysis one can mean:

1) the physical protein-protein interactions networks (and then the task is to find an association of a pattern in a multi-omcs dataset to the fixed topology of PPI network) or

2) the representation of the mutli-omics dataset in a gene-gene association (e.g., correlation) graph, derived from the data itself (and then the task is to interpret the topology of the graph in terms of biological mechanisms).

There is no established golden standard methodology for network-based analysis of multi-omics datasets, especially in the case when the number of samples in the dataset is small and the overlap between omics modalities is not complete, which is frequently the case of paediatric multi-omics datasets.

In this deliverable we suggest several methodologies of network-based analysis of multi-omics datasets, based on joint application of matrix factorization, namely Independent Component Analysis or ICA (a powerful technique for dimensionality reduction of wide omics data) and graph theory-based methods for analysing networks. We investigate three strategies reported in Chapters 2,3 and 4 correspondingly:

1) Crossing the results of application of ICA to multi-omics data (result of deliverable 3.1) with high-confidence human PPI network structure, using OFTEN method inside BIODICA computational pipeline implemented and published as a part of this deliverable.
2) Crossing the results of application of ICA with the network modules inferred from the gene-gene association study by COSIFER (provided as deliverable 4.1 of iPC)
3) Infer multi-level gene-gene association networks directly from the results of ICA application to mutli-omics datasets and analysing the topology of the resulting networks. The suggested methodology has not yet been published.

In Chapter 2 and Chapter 4 we use a concrete example of an unpublished multi-omics neuroblastoma dataset kindly provided to iPC consortium by AMC partner (Dr. Jan Koster). This dataset and the way it was prepared for application of the network-based analysis is described below in section 1.1. In order to apply methods relying on the structure of PPI networks, we had to select a source of high-confidence protein-protein interactions. The PPI network used in our analyses is described in section 1.2 below.

1.1 Preprocessing and harmonizing a new multi-omics neuroblastoma dataset (provided by AMC partner)

AMC partner shared with CURIE and BSC partner a neuroblastoma dataset, not previously analysed in an integrative way, containing multi-omics description of an NB cohort and extensive clinical data associated. The dataset contains five layers of omics of the tumours of the same patient cohort treated. These layers are:

1) Copy number profiling obtained by application of CGH technology (97 samples)
2) Profile of somatic mutations (87 samples)
3) Transcriptomic profiling using Affymetrix exon array HuEx10 (110 samples)
4) Transcriptomic profiling using Affymetrix U133Plus microarray platform (122 samples)
5) Methylation profiling using Illumina 450K platform (59 samples)

Clinical data contained information about collected samples organised in 59 fields containing numerical and categorical values, with a number of missing values.

In order to perform the integrative data analysis, all the values from various levels of omics profiling were mapped to the gene names, using provided annotation files. All samples were mapped to the patient ids from which the sample has originated. 49 patient ids had a complete set of matched omics profiles from 5 different levels (see Figure 1).

The dataset was composed of 5 data matrices of the same dimension, containing genes as rows and samples as columns. Thus, each data matrix contained 49 columns and 15115 rows. The complete data matrices of omics profiles (including those samples that did not have all 5 layers of matching omics profiles) were also stored at the NextCloud platform.

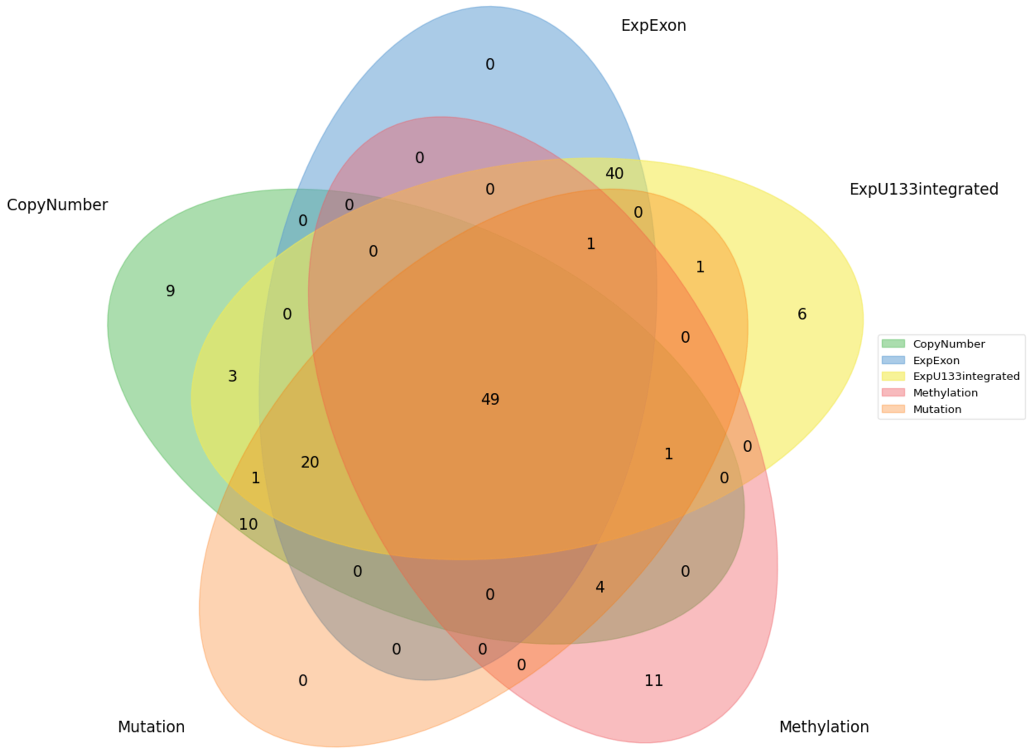


Figure 1: The matching structure of omics profiles obtained at different layers of AMC neuroblastoma multi-omics dataset. 49 patients have omics profiles obtained from 5 different omics platforms.

1.2 Preparing a high-quality reference PPI network for the network-based analysis of ICA metagenes

Several types of network-based analysis of multi-omics data involve the use of PPI networks [Kuperstein et al, 2015]. A PPI network of a particular organism serves as a reference of functional relatedness between proteins without splitting the network into pathways or reference gene sets. The PPI network structure can be superimposed with the results of computing the statistical association between genes or proteins. However, the results and even the applicability of a PPI network-based computational method depends on the concrete PPI network being used. Moreover, one of the critical parameters of PPI network reconstruction for application of computational methods is the network density, i.e. the average number of interactions per protein in the network. The field of systems biology suggests several large-scale reconstructions of PPIs, including STRING, BioGrid, PathwayCommons databases, in which the quality of curation and the confidence of individual interactions can be quite diverse.

From our previous experience, we prefer to use PPI reconstructions of intermediate density, therefore focusing on the most reliable and confident PPIs resulting either from screening literature or well-designed experiments. For this deliverable, we revised three such PPI reconstructions:

1) Human Protein Reference Database (HPRD), https://www.hprd.org/. This PPI network was shown to provide biologically significant results in multiple studies and suits well for our purposes. Nice property of this network is a systematic representation of human protein complexes composition which can be included into the PPI network using specific edge types. However, HPRD has not been updated since 2009.
2) The Human Reference Interactome (HuRi), http://www.interactome-atlas.org/, provided by the Center for Cancer Systems Biology at Dana-Farber Cancer Institute.
3) HINT (High-quality INTeractomes) is a curated compilation of high-quality protein-protein interactions from 8 interactome resources, http://hint.yulab.org/.

After some initial tests we found out that the more recent HuRi and HINT PPI reconstructions suffer from underrepresentation of certain basic biological functions (such as cell cycle), unlike the older HPRD. Also, at the intersection of these three databases we found a small number of common interactions (15 binary interactions) and only 4169 common proteins. Therefore, for our PPI network-based analyses we used a union of HPRD+HuRi+HINT interactions, which we called the *H^3combined* PPI network. We further eliminated self-interactions and multiple interactions between the same pair of proteins which resulted in a network of 16558 proteins connected by 166838 binary interactions. The network is provided as a part of BIODICA software distribution from https://github.com/LabBandSB/BIODICA.

Chapter 2 Integrated environment BIODICA for multi-omics data analysis, based on matrix factorization and network-based analysis

2.1 BIODICA environment description

The recent progress of high throughput omics technologies has made molecular data more accessible and has fostered the development of many computational analyses to exploit the rich information they offer. Such analyses require efficient tools to handle the high dimensionality of these data and reveal the underlying biological processes.

Independent Component Analysis (ICA) is a statistical and computational method which aims to represent observed signals as linear mixtures of independent latent factors. ICA has been successfully applied to omics data with the hypothesis that observed molecular profiles result from linear combinations of unobserved biological and technical processes (Liebermeister, 2002). In particular, it has been shown to extract interpretable and reproducible components and has stood out from other popular methods like Principal Component Analysis (PCA) or Non-negative Matrix Factorization (NMF) (Sompairac et al., 2019).

In order to facilitate the use of ICA for non-experienced in programming users we developed BIODICA, a complete computational environment for a user-friendly application of ICA to omics data. It encompasses a set of tools to extract and interpret reproducible independent components, using methods that already proved to be successful in multiple studies (Biton et al., 2014; Aynaud et al., 2020).

The computational core of BIODICA is the Python package stabilized-ica. It implements a stabilization procedure which addresses the variability of the solutions of ICA algorithms when run multiple times (Himberg and Hyvarinen, 2003). When applied to transcriptomics data, not only did this procedure provide a quantification of the significance of the independent components but it also extracted more reproducible ones than standard ICA (Cantini et al., 2019). Besides, it allowed the development of an approach for selecting the optimal number of independent components to extract from omics data (Kairov et al., 2017), which is also available in BIODICA.

BIODICA provides a unique toolbox to help the biological interpretation of the extracted components, combining different annotation and visualization methods which already proved their usefulness (Teschendorff et al., 2007; Kondratova et al. 2019). Several knowledge-based annotation methods are proposed, such as functional enrichment analysis using ToppFun (Chen et al., 2009), Gene Set Enrichment Analysis (Subramanian et al., 2005).

In relation to WP4 and D4.3, BIODICA includes *three network-based tools for interpreting the results of application of ICA to transcriptomic or other omics datasets*:

1) network-based enrichment analysis using known graphs of protein-protein interactions referenced further as OFTEN (Optimally Functionally Enriched Network)
2) visualisation of computed independent components on top of molecular interaction maps, using NaviCell platform developed in (Bonnet et al., 2015) and re-designed as a part of WP4 (see report on deliverable D4.2)
3) tool, based on application of Mutual Nearest Neighbours (MNN), to match the components extracted from several independent omics data sets. Studying the reproducibility of independent components across multiple data sets may help distinguishing biological signals that are specific to a particular disease/data type or technical biases that are specific to particular conditions (Biton et al., 2014; Cantini et al., 2019).

BIODICA comes with a user-friendly Graphical User Interface called BIODICA Navigator, providing non-experienced users a no-code access to all the BIODICA functionalities. It facilitates communication with biology experts, producing sortable and interactive HTML-based reports. The interface has been designed and validated in several studies, including a study of Ewing sarcoma at single-cell level (Aynaud et al., 2020), as part of iPC project reported in D3.1.

BIODICA tool is available from https://sysbio-curie.github.io/biodica-environment/ and the manuscript describing the tool was published in Bioinformatics Oxford journal (Captier et al, 2022).

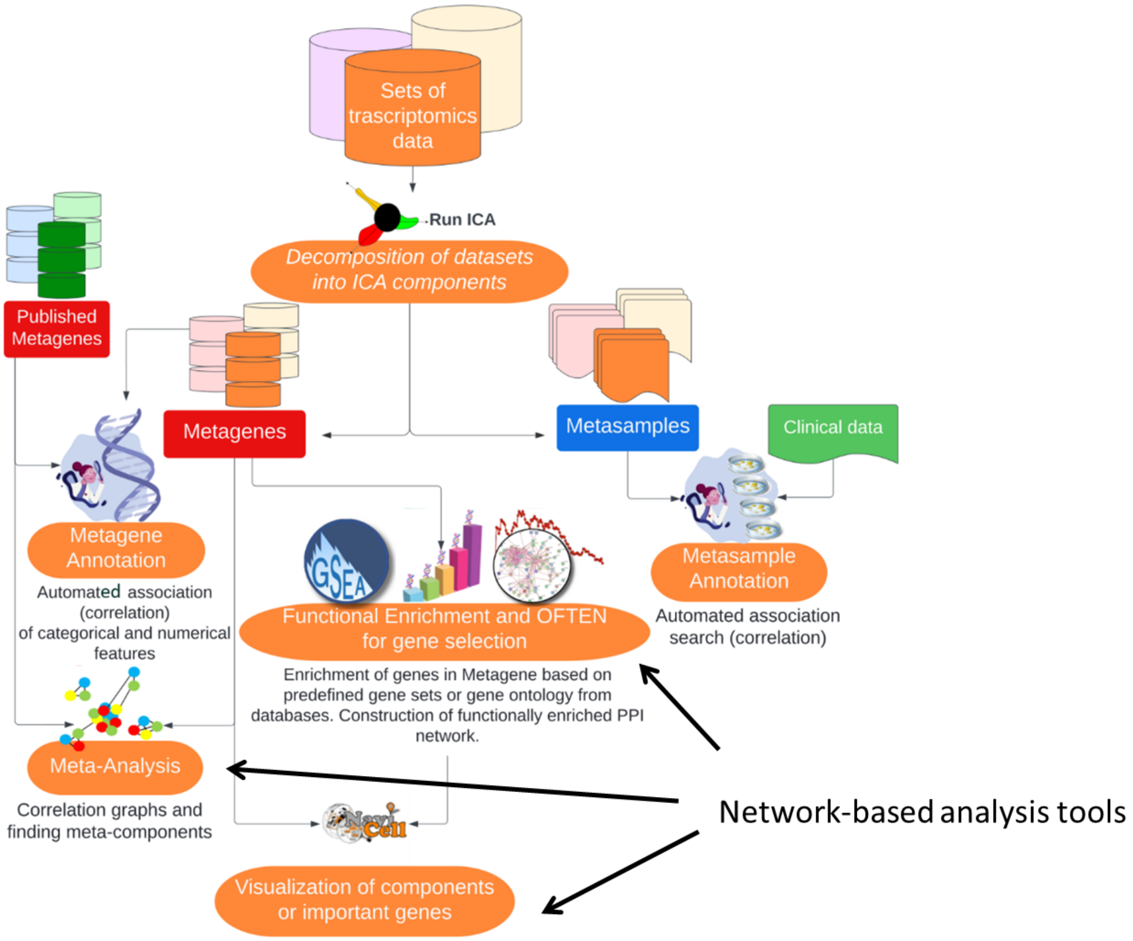

Sets of trascriptomics data
Run ICA
Decomposition of datasets into ICA components
Published Metagenes
Metagenes
Metasamples
Clinical data
Metagene Annotation
Automated association (correlation) of categorical and numerical features
GSEA
Functional Enrichment and OFTEN for gene selection
Enrichment of genes in Metagene based on predefined gene sets or gene ontology from databases. Construction of functionally enriched PPI network.
Metasample Annotation
Automated association search (correlation)
Meta-Analysis
Correlation graphs and finding meta-components
NaviCell
Network-based analysis tools
Visualization of components or important genes

Figure 2: Workflow of BIODICA toolbox. Three network-based analysis tools are indicated by arrows.

2.2 Method OFTEN for associating a Protein-protein interaction network with a ranked list of genes

After ICA decomposition, each metagene could be associated to a subnetwork in a global network of pairwise interactions such as protein-protein interactions (PPI). BIODICA tool can perform OFTEN analysis, a method for selecting a number of genes in a ranked gene list such that this set forms the Optimally Functionally Enriched Network (OFTEN), formed by known physical interactions between genes or their products.

Briefly, OFTEN analysis applied to an IC metagene consists in several steps:

1) For the k top weighted/top ranked genes in the metagene, it maps them on the interaction graph and measure the size C(k’) of the largest component of the subnetwork formed by the k’ genes among the k top genes found in the interaction graph;

2) k’ genes are then randomly sampled and the size of the largest component of the subnetwork they form is measured R(k’) - this step is repeated NumberOfPerms times;

3) Finally a percolation score is computed S(k) = (1/k')*(C(k') - Mean(R(k'))) to assess whether the largest component of the subnetwork formed by the top ranked genes of the metagene is highly non-random or not.

These 3 steps are repeated for a number of top ranked genes k going from Min to Max with a step of size Step. At the end, the largest number k_opt after which the percolation score goes down is selected and the largest component of the connected subnetwork is associated with the IC metagene.

OFTEN analysis included in BIODICA is illustrated with a tutorial available from the BIODICA site : https://sysbio-curie.github.io/biodica-environment/docs/tutorials/tuto_often/. There also exists a possibility to apply OFTEN from command line.

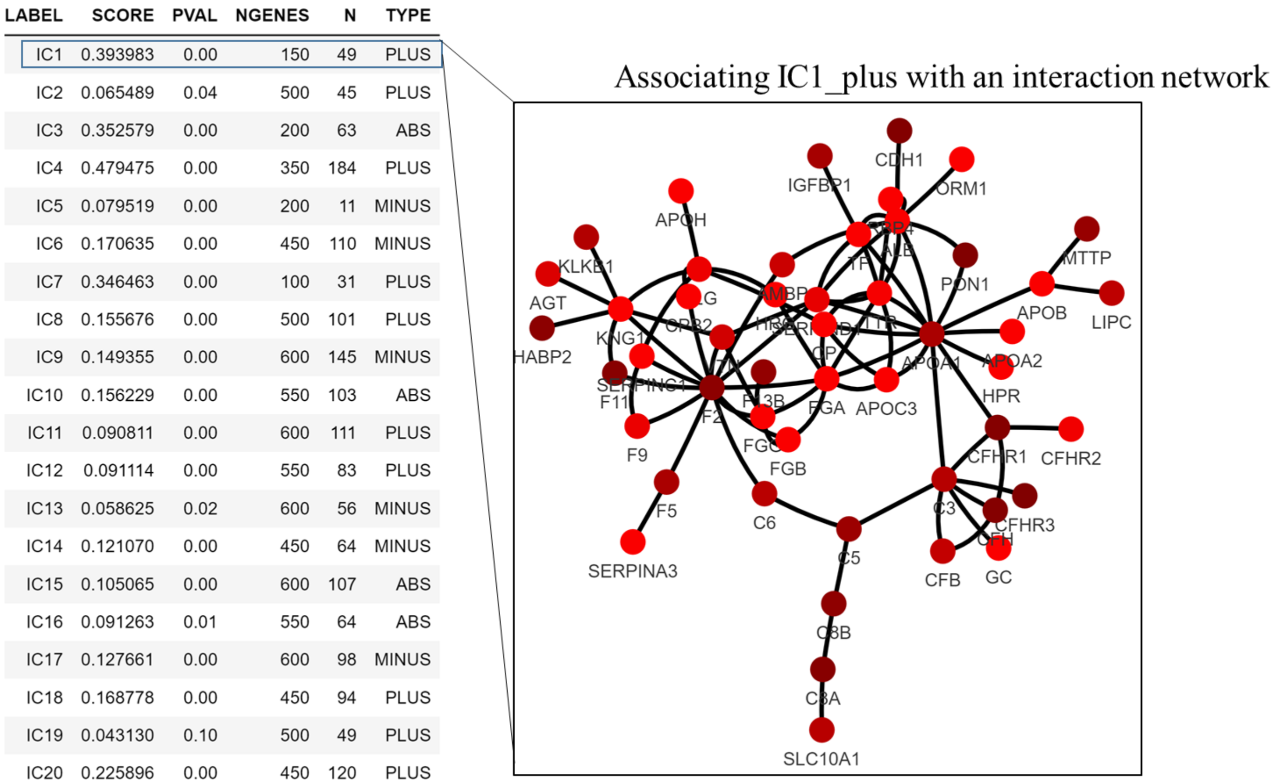
2.3 Application of OFTEN to the results of matrix factorization of paediatric multi-omics datasets

OFTEN algorithm can be used to associate a PPI subnetwork to a metagene or ranked list of genes such as metagenes computed with the use of matrix factorization approaches such as ICA. We systematically applied OFTEN to the results of application of matrix factorization methods in D3.1. In order to achieve this, we developed a simple Python interface to OFTEN using a command line way of launching analyses from within BIODICA. Example notebook illustrating this interface is provided through iPC github in https://github.com/iPC-project-H2020/wp4-deliverables/tree/main/D4.3 .

As an example result of such analysis, we've applied OFTEN to the results of ICA application to the multiomics neuroblastoma dataset, provided by AMC partner and described in Introduction to this deliverable. Application of OFTEN to a particular ICA decomposition results in a table like the one shown in Figure 3.

Subnetworks significantly associated (exceeding a certain threshold in terms of the association score and having a small p-value) with at least one of the independent components can be merged together and visualised as a single PPI subnetwork associated with the ICA decomposition altogether (see Figure 4). Thus for the ICA decomposition of the Affymetrix U133 microarray neuroblastoma dataset, this leads to creation of a network with 1027 nodes and 1671 edges. Remarkably, this network has a modular organisation, with major modules expectedly associated with cell cycle, immune infiltration, extracellular matrix and apolipoproteins. However, some modules emerged from the genes contributing to several independent components (like the one denoted as “axon guidance” in Figure 4).

Interestingly, when we applied OFTEN to the results of ICA decomposition of gene expression profiling by Affymetrix Human Exon array, we identified a network containing 1267 nodes and 2023 edges. The intersection of two networks (411 nodes and 334 edges) is shown in Figure 5. The intersection is highly significant (p-value~10^{-200}) which indicates the reproducibility of the results of ICA+OFTEN analysis with respect to changing the platform of gene expression profiling. This analysis of two modalities reveals expected major network modules related to cell cycle, immune infiltration, extracellular matrix, but some smaller subnetworks can represent interesting targets to be investigated in neuroblastoma. Thus, MEOX2, a homeobox-containing transcription factor that plays an essential role in developing tissues, has recently been shown to play a role in the cancerogenesis of gliomas [Tachon et al, 2021]. Some of the proteins highlighted by this analysis were already suggested as potential biomarkers such as C3 protein [Kim et al, 2014].



LABEL	SCORE	PVAL	NGENES	N	TYPE
IC1	0.393983	0.00	150	49	PLUS
IC2	0.065489	0.04	500	45	PLUS
IC3	0.352579	0.00	200	63	ABS
IC4	0.479475	0.00	350	184	PLUS
IC5	0.079519	0.00	200	11	MINUS
IC6	0.170635	0.00	450	110	MINUS
IC7	0.346463	0.00	100	31	PLUS
IC8	0.155676	0.00	500	101	PLUS
IC9	0.149355	0.00	600	145	MINUS
IC10	0.156229	0.00	550	103	ABS
IC11	0.090811	0.00	600	111	PLUS
IC12	0.091114	0.00	550	83	PLUS
IC13	0.058625	0.02	600	56	MINUS
IC14	0.121070	0.00	450	64	MINUS
IC15	0.105065	0.00	600	107	ABS
IC16	0.091263	0.01	550	64	ABS
IC17	0.127661	0.00	600	98	MINUS
IC18	0.168778	0.00	450	94	PLUS
IC19	0.043130	0.10	500	49	PLUS
IC20	0.225896	0.00	450	120	PLUS

Associating IC1_plus with an interaction network

Figure 3: Result of OFTEN application to the analysis of gene expression independent components.

The table shows the components analysed (IC1-IC20), the PPI network association score (SCORE column), the p-value of this score (PVAL column), number of top-contributing genes taken to extract the subnetwork (NGENES column), number of protein from the top-contributing genes actually found in the PPI network (N column), and the tail of the component where the strongest association with the PPI network is found (TYPE column, PLUS for positive tail, MINUS for negative tail, ABS for the ranking of absolute values of component weights). OFTEN automatically generates a Cytoscape.js interactive representation of the subnetwork associated with each component which can be browsed online (shown on the right for IC1, as an example).

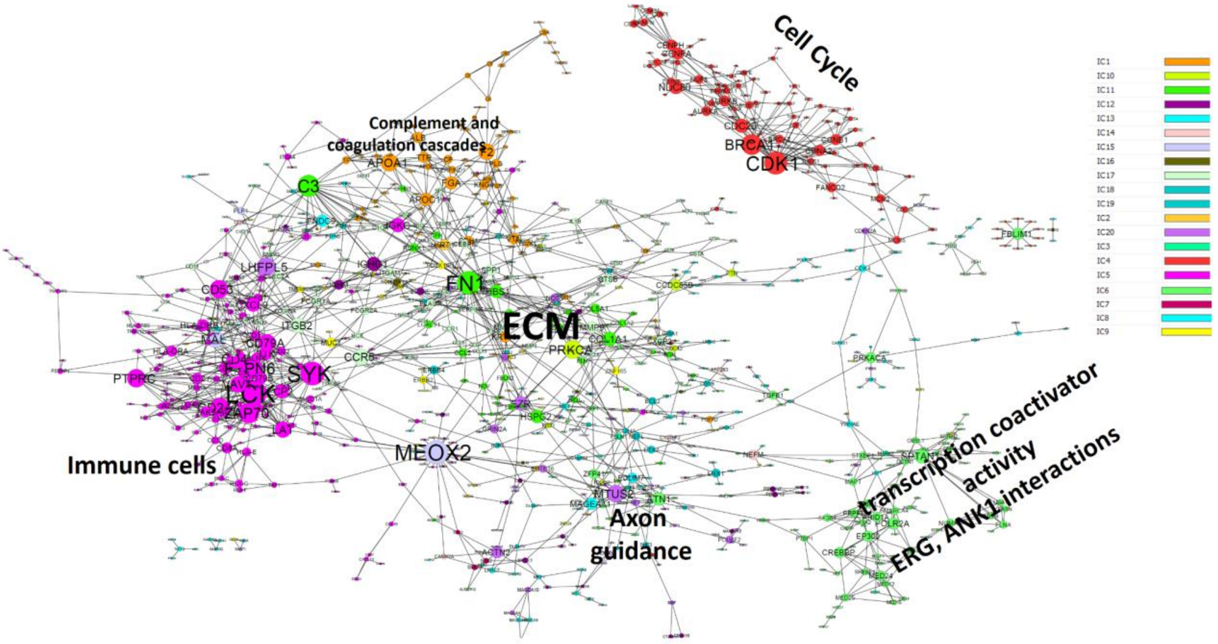


Figure 4: Results of application of OFTEN method to all 20 ICA components of Affymetrix U133 microarray-based expression profiling of neuroblastoma.

All PPI subnetworks significantly associated with independent components are shown simultaneously. The colour of the node indicates the independent component in which the gene has the largest contribution (see the legend). The size of the node reflects its connectivity in the network shown here.

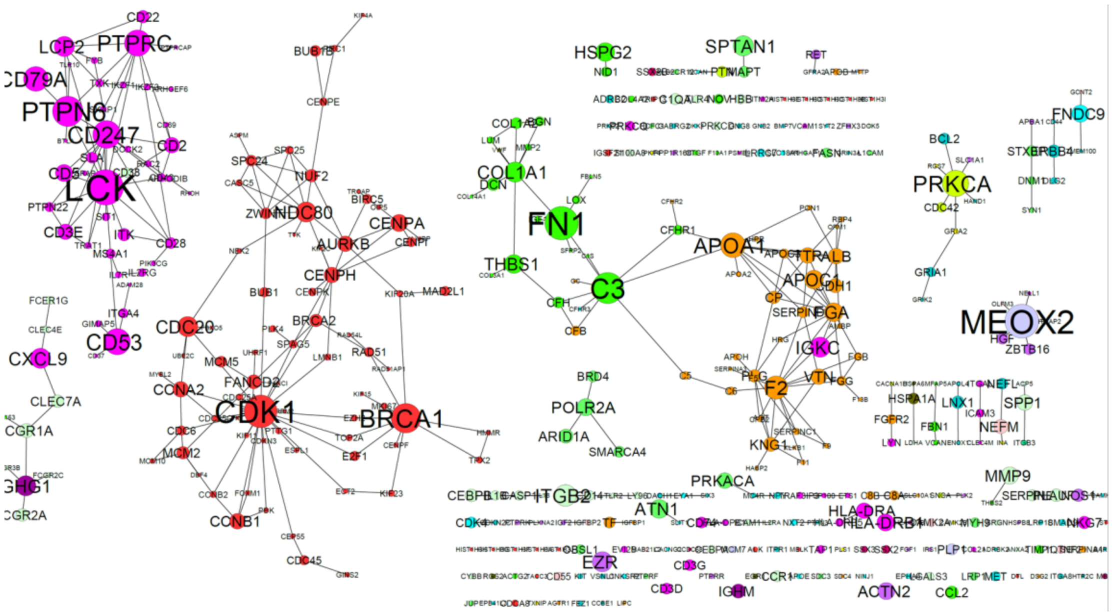


Figure 5: Intersecting the results of application of OFTEN algorithm to two expression datasets decomposed into independent components, one obtained with Affymetrix U133 platform and another one with Affymetrix human exon arrays.

The colors and node sizes are the same as in Figure 4.

Chapter 3 Crossing the results of matrix factorization and network-based analysis of multi-omics data

There are multiple ways to decipher molecular mechanisms active in cancer. Actually, two of the deliverables completed so far are quite similar, having the aim to identify molecular entities from solid paediatric omics data. Indeed, in D4.1, networks of molecular entities were derived from multi-omics datasets of several cancer types. In D3.1, communities of molecular mechanisms were extracted from gene expression of the solid tumour types. The molecular similarity networks deciphered in D4.1 are, in some way, comparable with the communities obtained in D3.1.

In this chapter, we investigated if some communities obtained in D3.1 are related to some modules defined in D4.1. The chapter is organised as follows: we first summarised the work and findings in D3.1 and D4.1. Then, we explained the strategy we used for comparing the findings and selecting the most insightful outputs. Finally, we described and discussed the most strongly related communities and networks.

3.1 Summary of D4.1 and D3.1

3.1.1 Results obtained in D4.1

In D4.1: “Building of cancer type-specific multi-layered molecular and patient similarity networks”, the aim was to provide multi-layered molecular and patient networks. A pipeline was designed to Build Molecular Networks and Patient-Patient Similarity Networks (BMNPPSN). Figure 6, borrowed from deliverable D4.1, illustrates this pipeline. The first step consists in running the COSIFER package, which is designed for: 1. creating the collection of molecular and patient similarity networks, based on the application of multiple computational methods and 2. integrating these networks together. The second step consists in community detection from these networks and the annotation of the extracted communities using enrichment analysis. Other analyses were developed in D4.1 but are not related to our work in D4.3.

An initial corpus of paediatric cancer-specific networks was created, computed from 4 cancer datasets relevant for iPC, encompassing several paediatric cancer types: MB, NB and ES. Table 1 lists these datasets. Three of them are multi-omics datasets.

In D4.3, we started from the molecular networks constructed from this corpus. 9 networks were used in this analysis and processed as described in the text below.

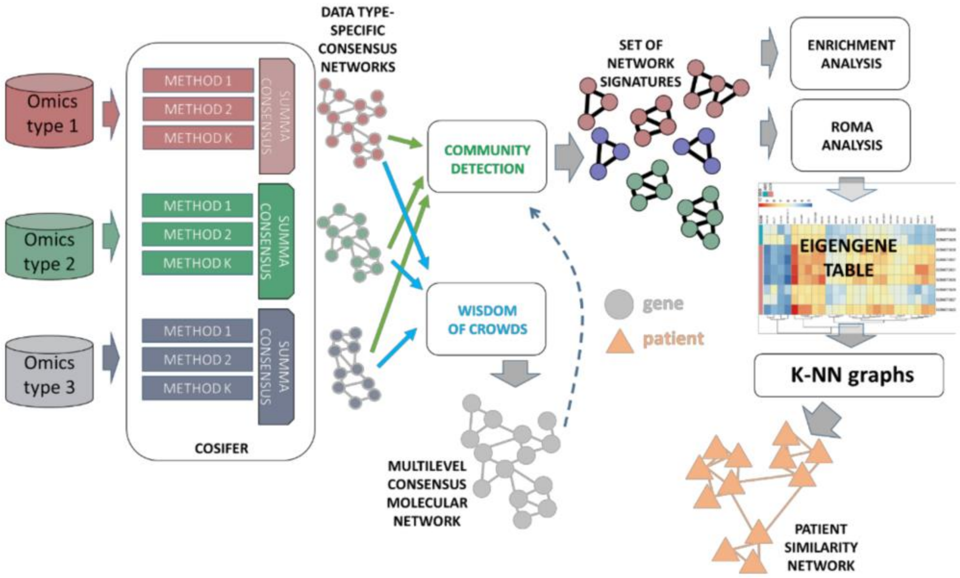


Figure 6: Schema of the WP4 pipeline for constructing paediatric cancer-specific networks from multi-omics data.

dataset name	cancer type	layer	Nb of subnetworks
Cavalli_GE	MB	GE	74
Cavalli_methylation	MB	methylation	117
Forget_GE	MB	GE	38
Forget_methylation	MB	methylation	222
Forget_proteomics	MB	proteomics	123
Forget_phospho_proteomics	MB	phospho_proteomics	112
Henrich_GE	NB	GE	121
Henrich_methylation	NB	methylation	223
Postel-Vinay	ES	GE	32

Table 1: 9 molecular networks used as input to define the subnetworks: 2 multi-layers MB networks, 1 multi-layers NB network, 1 single-layer ES network

We used the results obtained from the molecular networks only, to be comparable with the work done in D3.1. COSIFER includes 10 methodologies for network inference and 3 different consensus strategies to integrate the predictions of individual methods. Again, for the sake of comparison, we used only one method (Spearman correlation) for computing statistical associations between molecular profiles. We used the R package BMNPPSN developed in D4.1 on each data type network from which we detected communities using Markov clustering, allowing the detection of modules. The only variation from what was done in D4.1 is the number of edges used from the input network. Here we kept the 10^4 highest weights, instead of 10^5 as in D4.1.

3.1.2 Results obtained in D3.1

In D3.1: “Identification of important regulatory elements using multi-level matrix factorization approaches”, we performed unsupervised deconvolution using matrix factorization (stabilized ICA - sICA) of gene expression data from each of the 4 solid tumour types of interest in iPC, as described in Figure 7, borrowed from D3.1. We then performed a meta-analysis of the weighted metagenes defined in the 4 analyses. 142 communities were derived. Off note, when using sICA, each extracted component is a vector of gene weights, where the highest and smallest weights show the most contributing genes. Thus, there is a positive tail and a negative tail.

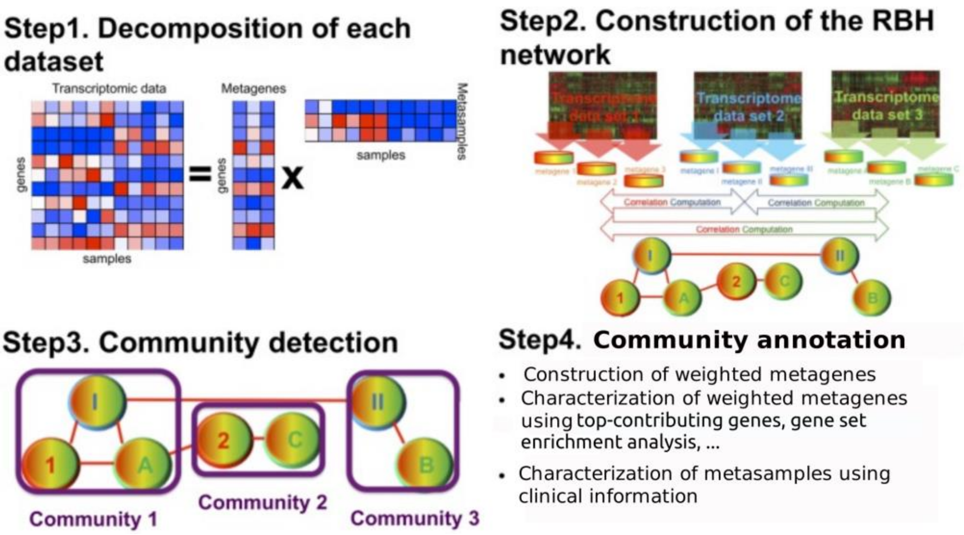

Step1. Decomposition of each dataset
Transcriptomic data
Metagenes
Metasamples
genes
genes
samples
samples
Step2. Construction of the RBH network
Transcriptome data set 1
Transcriptome data set 2
Transcriptome data set 3
Correlation Computation
Correlation Computation
Correlation Computation
Step3. Community detection
Community 1
Community 2
Community 3
Step4. Community annotation
Construction of weighted metagenes
Characterization of weighted metagenes using top-contributing genes, gene set enrichment analysis, ...
Characterization of metasamples using clinical information

Figure 7: Overview of unsupervised deconvolution using matrix factorization (adapted from Cantini L. et al, 2019)

We retrieved general pathways as cell cycle and extracellular matrix. We identified several pathways linked to immunity: innate immune response, adaptive immune response (“neutrophil activation”; “lymphocyte activation”; “T cell activation”). We also observed angiogenesis. In addition, we identified dysregulation of several pathways known to be implicated in several of the four tumour types. For example, Wnt signalling is known to be deficient in the four tumour types. It was retrieved in several communities. Another example is the PI3K-Akt pathway, which also appeared as significantly enriched in several communities.

We used the results of the meta-analysis performed on gene expression of the solid tumour datasets, in particular the meta-weighted metagenes generated in D3.1.

3.2 Identification of regulatory elements reproducible across independent methods

3.2.1 Comparison of regulatory elements obtained using multi-level matrix factorization and multi-layered molecular similarity networks

To identify reliable regulatory elements, *i.e.* molecular entities that are identified using various approaches, we built gene sets from the subnetworks of each omics layer from D4.1. We then performed enrichments into these gene sets for the 142 communities, in the two tails of each community. For each tail, 1062 gene sets were tested. We required at least 5 genes to be part of the enrichment, which led to the exclusion of numerous modules since a lot of them contain a small number of genes.

There are a total of 113 significant enrichments in positive tails and 106 significant enrichments in negative tails, using a threshold of 0.01 for the adjusted p-value. More precisely:

- 100 (22 unique) / 93 (17 unique) significant enrichments “GE” in positive / negative tails.
- 9 (4 unique) / 7 (3 unique) significant enrichments “Methylation” in positive / negative tails.
- 4 (2 unique) / 6 (1 unique) significant enrichments “Proteomics” in positive / negative tails.
- 0 / 0 significant enrichments “Phosphoproteomics” in positive / negative tails.

The majority of the enrichments are in the modules coming from GE networks, as expected since the communities were detected from gene expression data in D3.1. Nevertheless, there are also significant enrichments in subnetworks from methylation and proteomics layers. Few communities show enrichments in several layers (see Table 2). A total of 82 communities are enriched (adjusted p-value: 10^{-2}) in the molecular modules defined in D4.1.

3.2.2 Filter stringency of the significantly enriched communities

We first used a more stringent cutoff for the adjusted p-value (10^{-5}) to keep a community, leading to 56 communities. We then excluded communities related to ribosome (RPL* and RPS* genes) and HLA complexes (HLA* genes), since these genes are often present in gene expression analysis as artefacts. Off note, other families of genes were highlighted and might be artefactual, like C1Q complex, SNORD116 gene cluster, …

The 28 remaining communities are listed in Table 2 15 out of 28 communities are a mixture of bulk and single cell data, showing reproducibility across data types. For each community, the total number of components, the number of components coming from bulk and single-cell datasets used in D3.1 is indicated. The next columns indicate the side of the tail being enriched, the tumour types of the components included (in D3.1) as well as the datasets (in D4.1) in which the community is significantly enriched. More details are provided in the table community_stats.xls in NextCloud (at https://data.ipc-project.bsc.es/s/zbtQpnWLygneRJ4). One of these community is described in detail in the next section.

	Nb comp	Nb bulk	Nb sc	Sig tail	Tumour type D3.1	Enriched modules D4.1
C39	21	9	11	neg	ES,HB,MB,NB	Henrich_GE_31,Henrich_GE_6
C40	21	9	11	neg	ES,HB,MB,NB	Henrich_GE_31,Henrich_GE_6
C42	17	4	13	pos	ES,MB,NB	Cavalli_GE_45
C43	17	5	12	pos	ES,NB	Cavalli_GE_5
C44	17	2	15	neg	ES,HB,MB,NB	Cavalli_GE_1
C45	17	2	15	neg	ES,HB,MB,NB	Cavalli_GE_1
C57	12	2	10	neg	ES,HB,MB,NB	Cavalli_GE_1,Forget_GE_3,Postel-Vinay_GE_5
C60	11	4	7	neg	ES,MB,NB	Henrich_methylation_10
C64	9	2	7	pos	ES,HB,NB	Henrich_methylation_10
C65	9	2	7	pos	ES,HB,NB	Henrich_methylation_10

	Nb comp	Nb bulk	Nb sc	Sig tail	Tumour type D3.1	Enriched modules D4.1
C69	8	5	3	pos	ES,HB,MB,NB	Postel-Vinay_GE_19,Cavalli_GE_5
C70	8	3	5	neg	ES,NB	Cavalli_GE_45
C72	7	0	7	pos	NB	Cavalli_GE_1,Forget_GE_3
C75	6	3	3	pos	ES,MB,NB	Henrich_GE_7
C76	6	0	6	pos	MB,NB	Cavalli_GE_1,Forget_GE_3
C83	6	0	6	neg	MB,NB	Cavalli_GE_1,Forget_GE_3
C84	6	0	6	pos	NB	Forget_GE_1
C89	5	1	4	neg	ES,HB,MB,NB	Cavalli_GE_1,Forget_proteomics_59,Forget_GE_3
C97	4	0	4	pos	ES,MB	Cavalli_GE_11
C98	4	0	4	neg	MB	Cavalli_GE_2,Forget_GE_1
C103	4	0	4	neg	NB	Henrich_methylation_3
C107	4	0	4	neg	ES,NB	Forget_GE_1
C119	3	3	0	neg	ES,MB	Henrich_methylation_16
C120	3	0	3	pos	MB	Forget_GE_3,Cavalli_GE_1,Forget_proteomics_59
C122	3	0	3	pos	MB,NB	Postel-Vinay_GE_15
C129	3	2	1	pos	ES,MB	Cavalli_GE_5,Postel-Vinay_GE_15,Postel-Vinay_GE_5,Postel-Vinay_GE_19
C131	3	3	0	pos	MB	Cavalli_GE_25
C141	2	2	0	pos	HB	Henrich_methylation_10

Table 2: Description of the 28 communities with significant associations with molecular networks established in D4.1

3.3 Annotations of the highlighted communities

We characterised these communities first by looking at the top-contributing genes obtained by application of matrix factorization (table table_top_contributing_genes.txt in NextCloud) and then by looking at the annotations performed in D4.1 and D3.1. The annotations obtained by the 2 approaches are given in the table community_complete_description.xls (in NextCloud). It highlights similar pathways retrieved in different communities and describes in detail the annotations obtained for a community and its associated modules. Table 3summarises the main pathways and an example is provided in Figure 8.

Immune response	C43, C69, C122, C129
Replication	C72, C76, C83, C89, C120
Protein folding/RNA processing	C42, C64, C65, C70
Translation / RNA splicing	C98, C107

Table 3: Summary of the most frequently retrieved molecular entities among the 28 communities

As an example, we selected C122, a community linked to immunity. Its top-contributive genes in the positive tail are highly significantly enriched in GO terms (10^{-22}) and the whole community has a quite high Nominal Enrichment Score (2.7) as reported in the html file in NextCloud. Figure 8 shows its annotations using GO. C122 includes components from MB and NB datasets and is associated with a module derived from ES.

Comments on the highlighted communities:

Combining results from D4.1 and D3.1 led to the identification of 28 communities with adjusted p-value less than 10^{-5}. The description of the communities, the analysis described here and the different results can be browsed in the html report provided in NextCloud "Report_molecular_entities_identified_by_MF_D3.1_network_D4.1.html".

More than half of these 28 communities show enrichments in molecular networks, which are coherent between the annotations of the 2 methods, showing consistencies between these two approaches. The communities span mainly 4 molecular entities, illustrated in Table 3, but other molecular entities are present in single communities, *e.g.* intermediate filament in C119.

While this approach is an interesting way to highlight reproducible (through different datasets, tumour types or methods) pathways, we also identified limitations: for example, we do not retrieve expected processes such as Wnt signalling or PI3K-AKT acting in these four cancer types. While being identified by each method independently, the comparison in the way we ran it led to the exclusion of these pathways, often because the number of genes in the definition of the modules was small and there were not enough genes in common when performing the enrichment analyses.

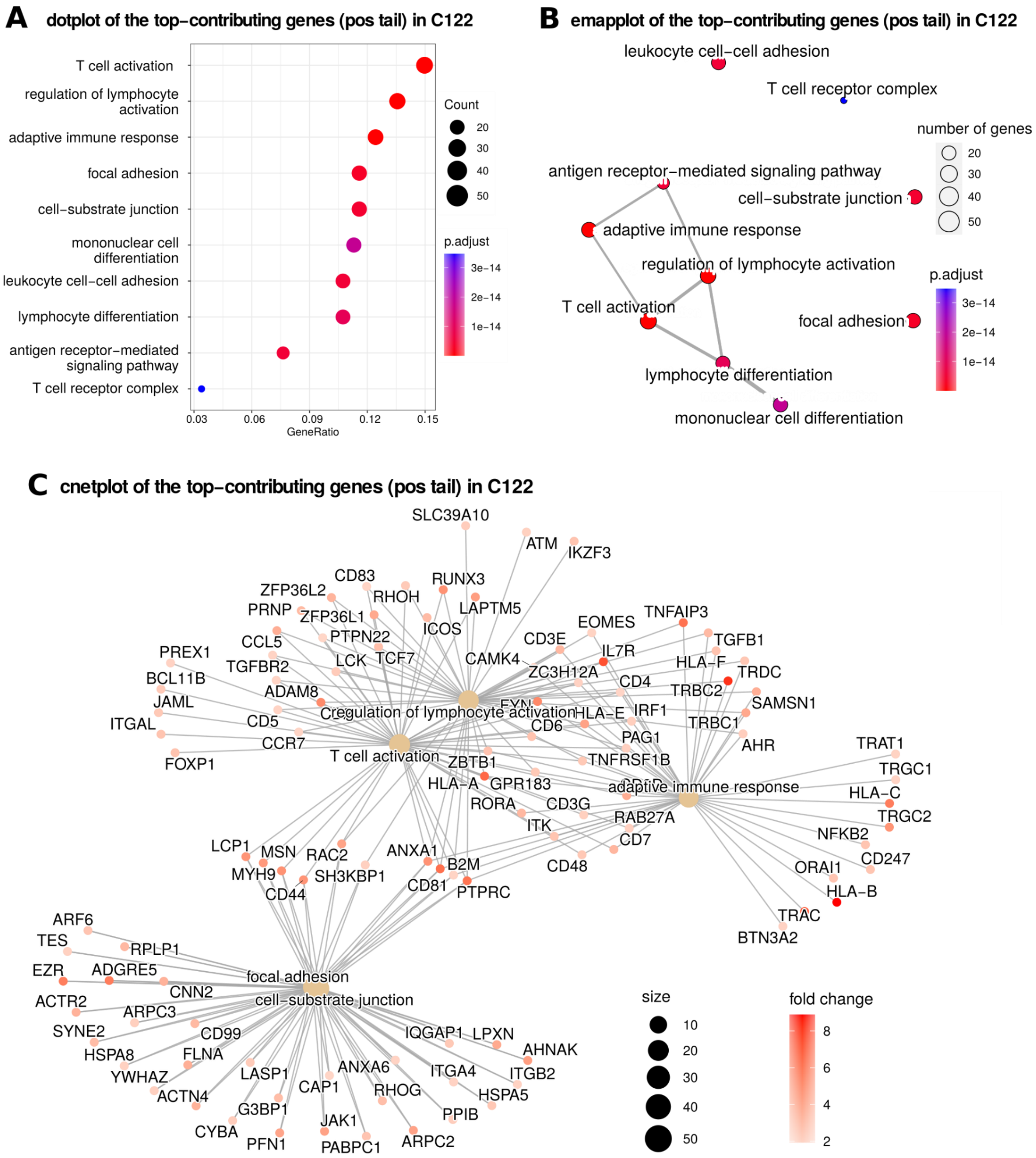


Figure 8: Annotations of the community 122 using GO on the top-contributive genes in the positive tail. A: the 10 most significant categories. B: connections between the 10 most significant categories. C: most contributing genes in each category of the 10 most significant categories.

Chapter 4 Computing network modules from joint analysis of multi-omics datasets. Case study of AMC neuroblastoma dataset

In this Chapter we present a case study of the novel network-based methodology, combining the results of matrix factorization with topological network analysis in order to derive multi-omics signatures and use them for performing the survival analysis as well as to find other associations with clinical information. This methodology can be applied to any multi-omics dataset containing a sufficient number of samples at each level of description (~50 samples minimum, ~100 samples desired). At the same time, the methodology does not necessarily require exact matching of samples for all omics levels: it only assumes the same population (cohort) structure of patients. In this report we present a case study of application of this methodology to the analysis of multi-omics neuroblastoma dataset, provided by the AMC partner and described in section 1.1, which satisfies these conditions.

4.1 Joint analysis of several omics layers representing a cohort of neuroblastoma patients

In order to establish the network of multi-omics gene-gene associations, we jointly analysed the multi-omics dataset, using the following steps:

1) Each omics layer is independently decomposed into latent factors, using matrix factorization. In our case study we used Stabilised Independent Component Analysis as it is implemented in BIODICA software (described in Chapter 2). Alternatively, it is possible to use other matrix factorization methods such as NMF either for analysing each omics level independently or jointly (such as MOFA or tensorial ICA). The resulting latent factors are recapitulated as metagenes (weights provided for each gene).
2) Each gene was characterised by a vector of concatenated weights from each latent factor extracted from each omics level. Therefore, each gene was characterised by a profile of latent factors from all omics levels.
3) Pearson correlation network was computed using the concatenated vectors of weights. An association link between two genes was established if the Pearson correlation exceeded a threshold *p*.
4) Each gene was tagged with a set of tags indicating if it appears in a set of top-contributing genes in one or several of the omics modalities. The gene appears in a set of top-contributing genes if it has a weight exceeding a certain threshold *t* in a metagene describing a latent factor.

In the case of multi-omics neuroblastoma dataset described in section 1.1, four distinct omics modalities have been used: EE - Exon Array (Expression), EM - Expression microarray, MT - Methylation array, CN - Copy number changes. Each of the modality was deconvoluted into 20 independent latent factors, denoted further as IC1_EE, IC2_EE, … , IC20_EE, IC1_EM, IC2_EM, …, IC20_EM, IC1_MT, IC2_MT, …, IC20_MT, IC1_CN, IC2_CN, … , IC20_CN (80 factors in total). Therefore, each gene was characterised by a vector of 80 weights in different factors. An example of significant correlation between two genes (NASP, SFPQ) is shown in Figure 9. In this case one can see that many latent factors from different modalities (CN, EE, EM, MT) contribute to the correlation. Indeed, these two genes are located on the same chromosome 1p arm, which might explain their co-amplification, co-expression and co-methylation. The thresholds used were p = 0.65 and t = 3.0. The NASP gene was tagged as ‘EM’ because it has a weight exceeding 3.0 only in one metagene IC4_EM (cell cycle-related) and is shown in red color in Figure 10. SPFQ gene does not receive any tag because it has a contribution less than 3.0 in all metagenes and is shown in grey in Figure 10.

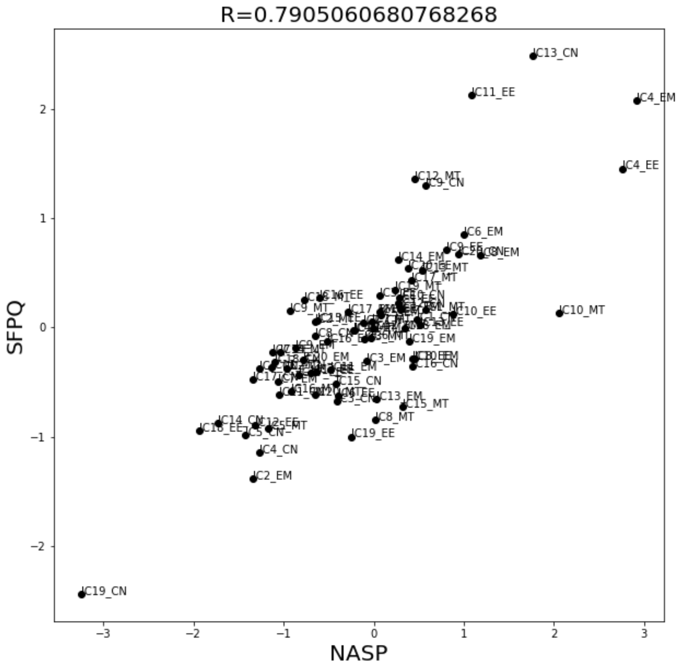


Figure 9: Example of statistical association between two genes across several omics modalities. Each point represents a contribution of a gene to an independent component (IC) computed for a specific omics modality (EE,EM,MT,CN). EE - Exon Array (Expression), EM - Expression microarray, MT - Methylation array, CN - Copy number changes.

The resulting network of multi-omics gene associations manifested an interesting topological organisation as a star-like pattern with a central network module strongly enriched in cell cycle genes (Bonferroni-corrected p-value=10^{-28} with GO:1901990, 'regulation of mitotic cell cycle phase transition') surrounded by modules of different nature, with prevalence of the modules resulting from connections between genes sharing similar copy number variation (CNV) patterns (containing genes shown in green in Figure 10). The clusters of genes (modules) having green colour result from the fact of co-localization of genes in the same or close genomic loci. However, together with CNV-associated modules of genes, the network contained modules resulting from both gene expression and DNA methylation levels, or from gene expression and CNV levels. These modules reflect the role of immune cell infiltration, interferon signalling, presence of other blood cells, role of extracellular matrix (ECM) and cancer-associated fibroblasts (CAFs).

Such network organisation underlines the neuroblastoma cancer type as being characterised by recurrent gene CNVs. Connections of modules associated predominantly to a particular CNV event with the central cell cycle module can reflect the evolutionary advantages provided to tumoral cells in terms of proliferation by each independent CNV event.

Accordingly to the objectives of the D4.3 we performed two types of topological network analysis in order to define and score the molecular mechanisms underlying neuroblastoma progression:

1) We applied clustering of the network presented in Figure 10 in order to identify the precise composition of multi-omics gene modules, and developed a score combining several levels of multi-omics description. This score was used for univariate survival regression analysis which allowed us to rank the modules with respect to their importance in predicting the

patient's treatment outcome.

2) We performed a topological analysis of the whole network presented at Figure 10 in order to define the ‘interface’ genes that connect various gene modules between each other and in particular with the central Cell Cycle module. These genes are possible candidates for neuroblastoma cancer drivers.

The results of two types of the analysis are briefly documented in the following two sections.

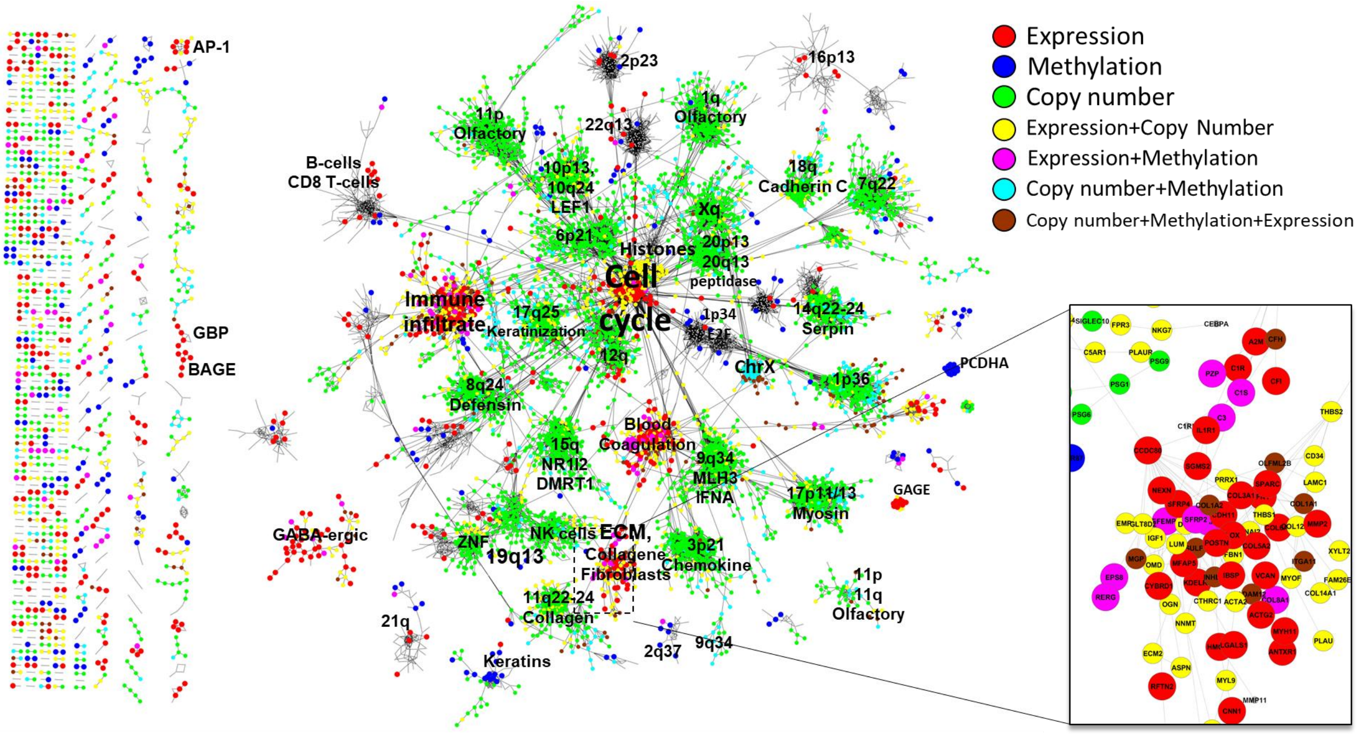


Figure 10: Network-based analysis of the neuroblastoma multiomics dataset provided by AMC partner, resulting in the definition of multi-omics modules (gene signatures), that can be scored using omics data from several modalities (gene expression, DNA methylation, gene copy number).

4.2 Defining multi-omics modules and use them to associate with clinical data

4.2.1 Decomposing the network into multi-omics modules and scoring them from the data

We performed clustering of the network presented in Figure 10, in order to define multi-omics modules, using the following algorithm:

1) We first decomposed the graph of gene-gene associations into connected components. We left only those components which contained at least k=5 genes with tags (coloured genes in Figure 10.
2) Those connected components containing more than 100 genes were further clustered using the Markov Chain Clustering algorithm (MCL).

In application to the graph shown in Figure 10 this resulted in the definition of 93 modules. In order to name them, each module was denoted by the name of the gene having the largest connectivity inside the module, prefixed with ‘CL_’. Thus, the central cell cycle-related module was named as

‘CL_CCNB2’ since the gene CCNB2 had the largest connectivity among all other 135 genes composing the module.

Afterwards, each module was assigned a score in those patients which had the molecular profiles corresponding to all gene tags in the module. For example, if a module contained genes tagged ‘EM’ and ‘CN’ then all patients having expression microarray and copy number change profiles received a score for this module, while other patients received ‘NaN’ score value.

The score itself was computed as a mean value of molecular measurements normalised to z-scores. For each tagged gene in the module those z-scores were taken into the mean computation which corresponded to the tags of the gene. For example, if a gene was tagged with ‘EM’, ‘EE’ and ‘MT’ then the z-scores from expression microarrays, exon arrays, DNA methylation levels were summed up. Importantly, the z-scores from the MT (methylation) level were taken with a negative sign, assuming that hypermethylation of a gene region leads to decrease of it’s activity.

4.2.2 Using multi-omics module scores in survival analysis

The obtained scores were used to compute the univariate survival regression both for overall survival (OVS) and progression-free survival (PFS). 22 out of 93 modules have shown significant association either to overall or progression free survival with p-value<0.05 and FDR=0.05. Some examples of how the multi-omics module scores stratify the patients are shown in Figure 11. This analysis highlighted some of the known survival prediction factors such as cell cycle and CNV of the 17p genomic locus, but also highlighted other possible predictors such as the hypomethylation of the genes EEF2, RPS15, CIRBP, OAZ1, associated with poor prognosis. Of note, these genes are co-localized at the 19p13.3 genomic locus. EEF2 is a major elongation factor and EEF2K kinase was previously reported to be required for neuroblastoma tumours possessing MYCN amplification to adapt to the nutrient deprivation conditions [Delaidelli et al, 2017]. RPS15 and CIRBP proteins are also involved in the translation control and were reported to be associated with various cancer types.

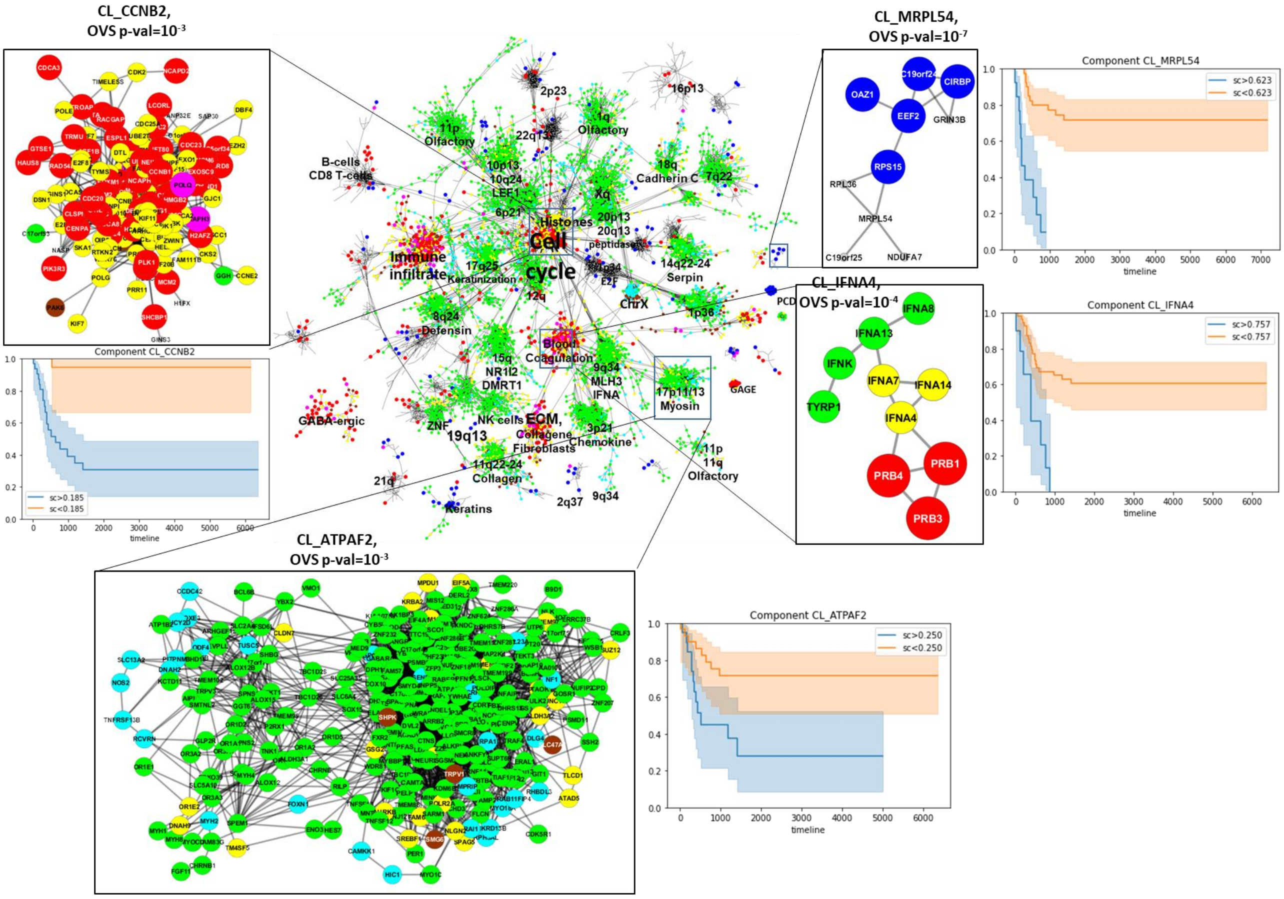


Figure 11: Use of multi-omics modules to stratify patients with respect to their survival. 4 out of 22 multi-omics significantly associated (p-val<0.05, FDR=0.05) with overall survival modules are exemplified by zoom and the corresponding Kaplan-Meier curves (the threshold for the module score, separating the patients into 2 groups, was optimized with respect to the significance of the logrank test).

The colours of the genes are deciphered in the legend of Figure 10.

4.2.3 Using multi-omics module scores to associate the modules with clinical data

The obtained multi-omics module scores were analysed for their associations with clinical variables provided with the neuroblastoma multi-omics datasets. This analysis was performed using BIODICA software, described in Chapter 2. The overview of significant associations (p-val<0.001) is provided in Figure 12 and some examples of significant associations are shown in Figure 13. Overall, this analysis reveals a subset of 15-17 multi-omics modules significantly associated with variables related to patient age, survival data, staging of tumours. Yet another group of modules (including CL_AHRR, where one of the module members is TERT gene known to be dysregulated in neuroblastoma) is associated with genome-related variables such as mycn_amp, loh1p even though this does not translate into an association with patient survival. Some of the modules were specifically associated with selected clinical variables such as gender (for example, this was expected for the module CL_MORF4L2, containing the genes from the X chromosome).

.

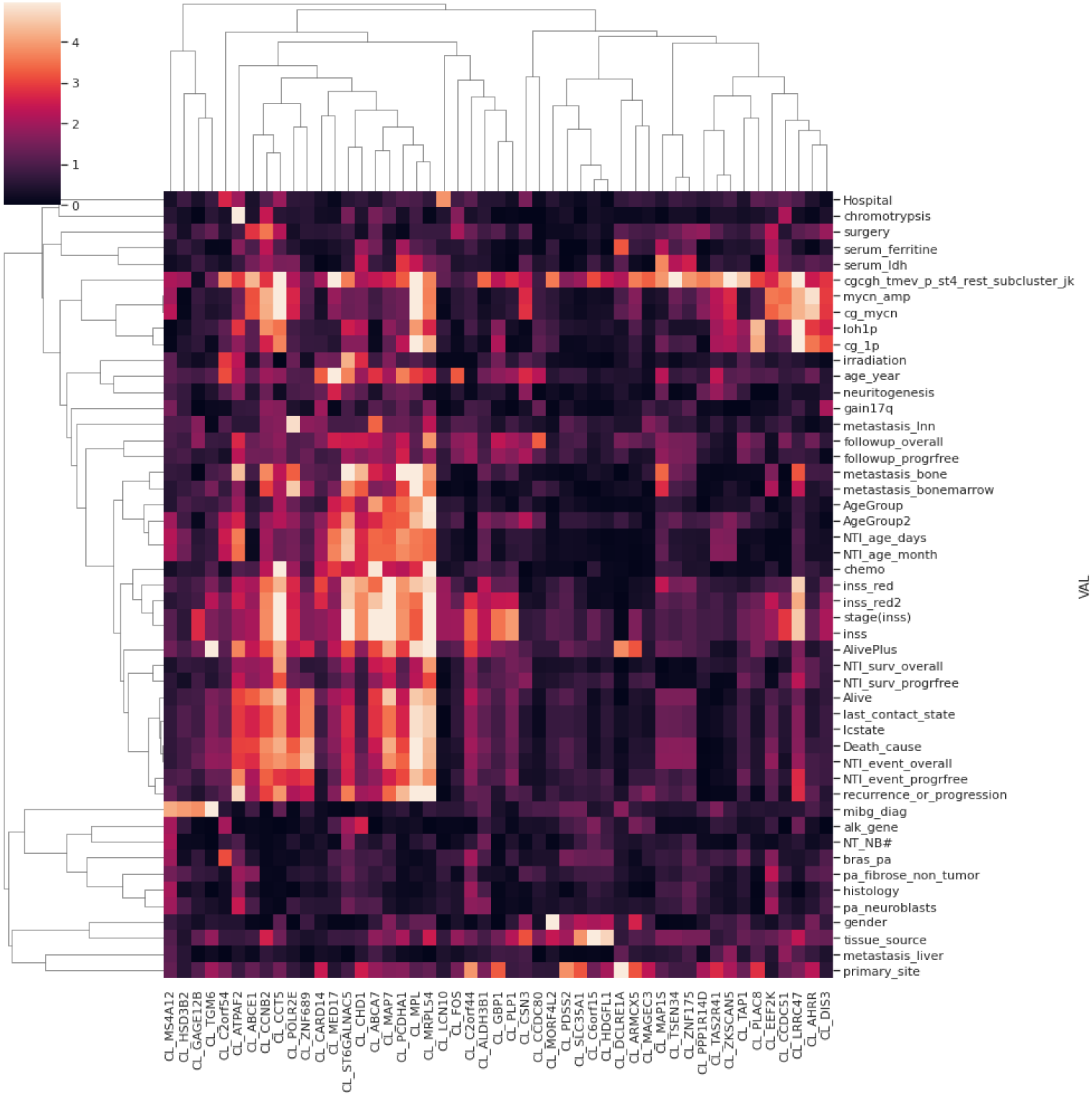


Figure 12: The most significant associations of clinical variables with multi-omics module scores. The values shown in the heatmap are -log10 p-values of the association scores. Some examples of these associations are shown in Figure 13.

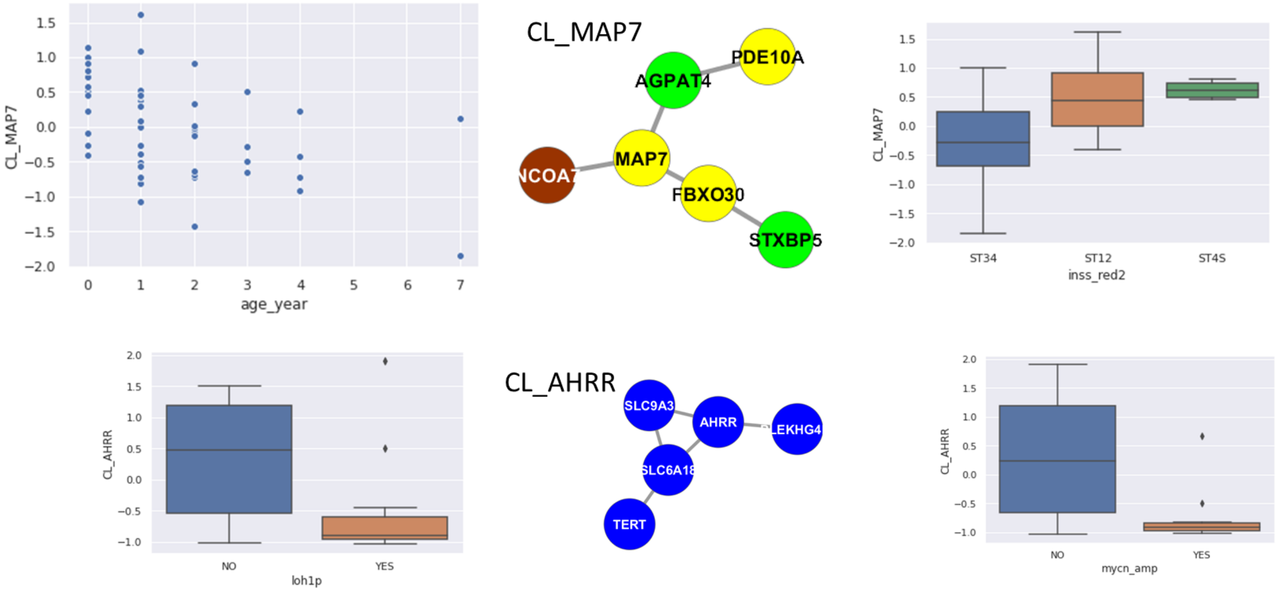


Figure 13: Examples of significant associations between multi-omics module scores and some clinical variables.

4.3 Defining potential cancer drivers as connector genes between the cell cycle module and other modules

We were inspired by the topological organization of the multi-omics network shown in Figure 10, and used this structure in order to identify potential cancer drivers in NB. The idea was that a potential cancer driver gene must serve as a connector between a cell cycle module and another module. In order to quantify this, we computed the in-betweenness score, for each gene in the network. In-betweenness is supposed to identify the connector genes between network communities as those through which many shortest paths in the network pass through.

An example of such a cancer driver is SPAG5 gene (Figure 14) which connects a module, collecting genes with large weights on the locus 17p (metagene CN_IC18 or module CL_ATPAF2), with the cell cycle module. SPAG5 gene has simultaneously large weights in expression-based metagenes EM_IC4 and EE_IC4 and in the 17p gain-associated metagene CN_IC18 (Figure 14, B). Of note, mitotic spindle protein SPAG5 or Sperm-associated Antigen 5 is considered as an emergent oncogene in many cancer types [He et al, 2020].

Other cancer drivers identified using such topological analysis of the multi-omics gene-gene association network can be extracted from Figure 14, A. Interestingly, these drivers are grouped into short connected sequences, such as GSG2→SPAG5→TOP2A, POLD1→WDR62→ASF1B, PHF8→POLA1→CENPI→KIF4A, NCAPG2→ANLN, WDR34→MELK, EXOSC10→E2F2→CDCA8, INTS8→RAD54B→PBK, NUP133→LIN9→NUF2, CASP8AP2→TTK, CEP76→SKA1, NUP107→RAD51AP1, RAF1→FANCD2, PPM1D→BRIP1, USP11→KDM5C→SMC1A→POLA1→CENPI, VRK1→DLGAP5. The direction here indicates approaching the cell cycle module. Many of thus identified cancer drivers were previously reported to be involved in the neuroblastoma genesis.

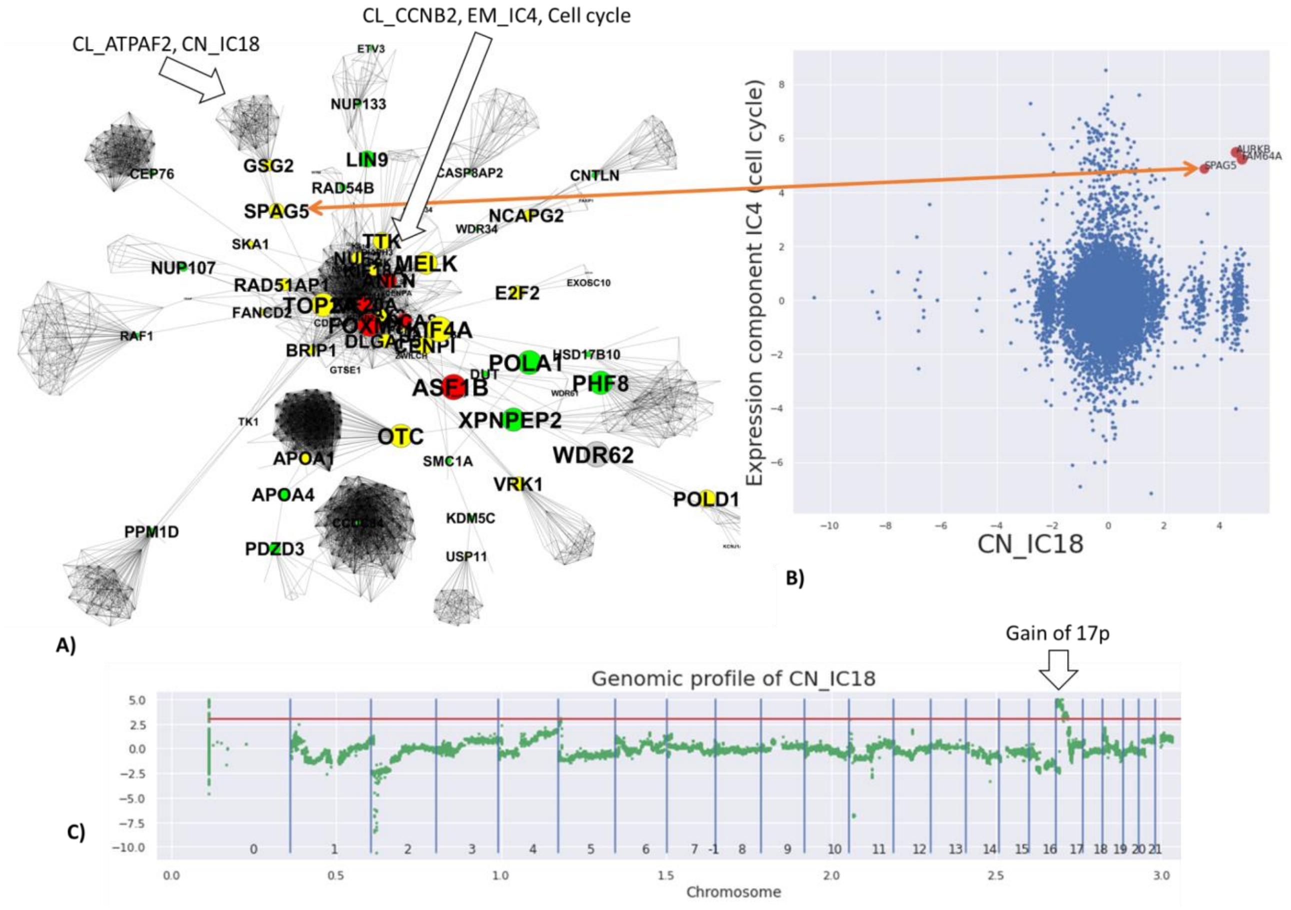


Figure 14: Example of identifying potential cancer drivers in NB.

A) Network extracted from the network shown in Figure 11, by taking the genes with the largest in-betweenness and their nearest neighbours. In-betweenness value is visualised by node and node label size. SPAG5 gene is a connector between two modules in the network, cell cycle multi-omics module (CL_CCNB2 or EM_IC4) and a module associated with CNV of 17p locus (CL_ATPAF2 or CN_IC18). B) Three genes, including SPAG5 simultaneously have large weights in the cell cycle metagene EM_IC4 and in the metagene CN_IC18 associated with CNV of 17p locus. C) genomic profile of the metagene weights of CN_IC18, showing that it is associated with 17p gain. Chromosome 0 here corresponds to the genes with undefined genomic location.

Chapter 5 Summary and Conclusion

This deliverable is devoted to adaptation of several methodologies of network based analysis to interpret the results of application of machine learning techniques to multi-omics datasets for paediatric cancers. The objective of this analysis is to identify molecular mechanisms of tumour progression affecting several levels of omics profiles. Topological analysis of physical PPI networks or networks of statistical associations is performed in order to prioritise certain protein functions or decompose the networks into functional modules that can serve as multi-omics signatures for diagnosis and prognosis.

The obtained results will serve as a basis for discussion with experts in the biology of particular cancer types, as a resource for discovering potential biomarkers and as a collection of multi-omics signatures.

The objectives of the iPC deliverable D4.3 which is a part of Task 4.3 “Identification of molecular mechanisms and network biomarkers based on several data types” are fully achieved.

Chapter 6 List of Abbreviations

Abbreviation	Translation
OFTEN	Optimally Functionally Enriched Network
BIODICA	Biological Data Analysis using ICA
ICA	Independent Component Analysis
PPI	Protein-protein interaction
CNV	Copy Number Variations
OVS	Overall survival
PFS	Progression-free survival
MCL	Markov Chain Clustering algorithm

Chapter 7 Bibliography

[1] Dimitrakopoulos C, Hindupur SK, Häfliger L, Behr J, Montazeri H, Hall MN, Beerenwinkel N. Network-based integration of multi-omics data for prioritizing cancer genes. Bioinformatics. 2018 Jul 15;34(14):2441-2448.

[2] Colaprico A, Olsen C, Bailey MH, Odom GJ, Terkelsen T, Silva TC, Olsen AV, Cantini L, Zinovyev A, Barillot E, Noushmehr H, Bertoli G, Castiglioni I, Cava C, Bontempi G, Chen XS, Papaleo E. Interpreting pathways to discover cancer driver genes with Moonlight. Nat Commun. 2020 Jan 3;11(1):69.

[3] Kuperstein I, Grieco L, Cohen DP, Thieffry D, Zinovyev A, Barillot E. The shortest path is not the one you know: application of biological network resources in precision oncology research. Mutagenesis. 2015 Mar;30(2):191-204.

[4] Sompairac N, Nazarov PV, Czerwinska U, Cantini L, Biton A, Molkenov A, Zhumadilov Z, Barillot E, Radvanyi F, Gorban A, Kairov U, Zinovyev A. Independent Component Analysis for Unraveling the Complexity of Cancer Omics Datasets. Int J Mol Sci. 2019 Sep 7;20(18):4414.

[5] Biton A, Bernard-Pierrot I, Lou Y, Krucker C, Chapeaublanc E, Rubio-Pérez C, López-Bigas N, Kamoun A, Neuzillet Y, Gestraud P, Grieco L, Rebouissou S, de Reyniès A, Benhamou S, Lebret T, Southgate J, Barillot E, Allory Y, Zinovyev A, Radvanyi F. Independent component analysis uncovers the landscape of the bladder tumor transcriptome and reveals insights into luminal and basal subtypes. Cell Rep. 2014 Nov 20;9(4):1235-45.

[6] Aynaud MM, Mirabeau O, Gruel N, Grossetête S, Boeva V, Durand S, Surdez D, Saulnier O, Zaïdi S, Gribkova S, Fouché A, Kairov U, Raynal V, Tirode F, Grünewald TGP, Bohec M, Baulande S, Janoueix-Lerosey I, Vert JP, Barillot E, Delattre O, Zinovyev A. Transcriptional Programs Define Intratumoral Heterogeneity of Ewing Sarcoma at Single-Cell Resolution. Cell Rep. 2020 Feb 11;30(6):1767-1779.e6.

[7] Himberg, J. and Hyvarinen, A. (2003). Icasso: software for investigating the reliability of ica estimates by clustering and visualization. In 2003 IEEE XIII Workshop on Neural Networks for Signal Processing (IEEE Cat. No.03TH8718), pages 259–268.

[8] Kairov U, Cantini L, Greco A, Molkenov A, Czerwinska U, Barillot E, Zinovyev A. Determining the optimal number of independent components for reproducible transcriptomic data analysis. BMC Genomics. 2017 Sep 11;18(1):712.

[9] Teschendorff AE, Journée M, Absil PA, Sepulchre R, Caldas C. Elucidating the altered transcriptional programs in breast cancer using independent component analysis. PLoS Comput Biol. 2007 Aug;3(8):e161. doi: 10.1371/journal.pcbi.0030161.

[10] Kondratova M, Czerwinska U, Sompairac N, Amigorena SD, Soumelis V, Barillot E, Zinovyev A, Kuperstein I. A multiscale signalling network map of innate immune response in cancer reveals cell heterogeneity signatures. Nat Commun. 2019 Oct 22;10(1):4808.

[11] Chen J, Bardes EE, Aronow BJ, Jegga AG. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W305-11.

[12] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005 Oct 25;102(43):15545-50.

[13] Bonnet E, Viara E, Kuperstein I, Calzone L, Cohen DP, Barillot E, Zinovyev A. NaviCell Web Service for network-based data visualization. Nucleic Acids Res. 2015 Jul 1;43(W1):W560-5. doi: 10.1093/nar/gkv450.

[14] Cantini L, Kairov U, de Reyniès A, Barillot E, Radvanyi F, Zinovyev A. Assessing reproducibility of matrix factorization methods in independent transcriptomes. Bioinformatics. 2019 Nov 1;35(21):4307-4313.

[15] Captier N, Merlevede J, Molkenov A, Seisenova A, Zhubanchaliyev A, Nazarov PV, Barillot E, Kairov U, Zinovyev A. BIODICA: a computational environment for Independent Component Analysis of omics data. Bioinformatics. 2022 Apr 6:btac204. doi: 10.1093/bioinformatics/btac204.

[16] Tachon G, Masliantsev K, Rivet P, Desette A, Milin S, Gueret E, Wager M, Karayan-Tapon L, Guichet PO. MEOX2 Transcription Factor Is Involved in Survival and Adhesion of Glioma Stem-like Cells. Cancers (Basel). 2021 Nov 25;13(23):5943.

[17] Kim PY, Tan O, Diakiw SM, Carter D, Sekerye EO, Wasinger VC, Liu T, Kavallaris M, Norris MD, Haber M, Chesler L, Dolnikov A, Trahair TN, Cheung NK, Marshall GM, Cheung BB. Identification of plasma complement C3 as a potential biomarker for neuroblastoma using a quantitative proteomic approach. J Proteomics. 2014 Jan 16;96:1-12.

[18] Delaidelli A, Negri GL, Jan A, Jansonius B, El-Naggar A, Lim JKM, Khan D, Oo HZ, Carnie CJ, Remke M, Maris JM, Leprivier G, Sorensen PH. MYCN amplified neuroblastoma requires the mRNA translation regulator eEF2 kinase to adapt to nutrient deprivation. Cell Death Differ. 2017 Sep;24(9):1564-1576.

[19] He J, Green AR, Li Y, Chan SYT, Liu DX. SPAG5: An Emerging Oncogene. Trends Cancer. 2020 Jul;6(7):543-547.