D4.4

Consensus multi-omics subtypes of paediatric cancers

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 / D3.2 / 1.0
Work package contributing to the deliverable	WP4
Due date	31st May, 2022
Actual submission date	30th May, 2022

Responsible organisation	BSC
Editor	Davide Cirillo
Dissemination level	PU
Revision	V1.0

Abstract	We report on the implementation of a method for multilayer community trajectory analysis and its applications, including a published study on medulloblastoma, a study on congenital myasthenic syndromes, and a study on the functional characterization of commonalities among a selection of paediatric tumours.
Keywords	Multilayer networks, functional analysis, paediatric cancers

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

Editor

Davide Cirillo (BSC)

Contributors (ordered according to beneficiary numbers)

Iker Núñez-Carpintero (BSC)

Salvador Capella-Gutierrez (BSC)

Jane Merlevede (CURIE)

Marianyela Petrizzelli (CURIE)

Anrei Zinovyev (CURIE)

Disclaimer

Executive Summary

In this deliverable, we introduce the concept of multilayer community trajectories and its applications in paediatric cancers for the identification of a functional consensus among them. In Chapter 1, an introduction to the challenges and opportunities of the functional characterization of paediatric cancers and the need of identifying commonalities (consensus) among different paediatric cancers are discussed. In Chapter 2, community detection in multilayer networks and the CmmD (Continual Multiplex network Module Detector) method, which computes multilayer community trajectories in a given multilayer network, are presented. The section reports on the application of multilayer community trajectories for functional characterization in two scenarios. In the first scenario, multilayer community trajectories are used in an optimization process to find the minimal number of genes that recapitulate known subgroups of medulloblastoma, a paediatric cancer (see Section 2.4). This work, authored by BSC and CURIE, has been recently published in iScience (Cell Press) [1]. In the second scenario, we identified modules of genes participating in the same multilayer community trajectories that distinguish between severe and non-severe phenotypes of congenital myasthenic syndromes, a rare disease (see Section 2.5). This work will be soon submitted to a high-impact journal for peer-reviewing. In Chapter 3, we calculate multilayer community trajectories of high-quality gene signatures of a number of paediatric cancers, and identify a functional consensus among them. In Chapter 4, we conclude the deliverable and outline future work in this area. All the files used in the analyses presented here are available at the iPC Nextcloud repository: https://data.ipc-project.bsc.es/s/GXWppiddJAZ6HHm. The code to reproduce the results in Chapter 3 is available at an iPC GitHub repository: https://github.com/iPC-project-H2020/D4.4

Table of Content

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Translation
CmmD	Continual Multiplex network Module Detector
ALL	Acute lymphoid leukaemia
HB	Hepatoblastoma
NB	Neuroblastoma
ES	Ewing's sarcoma
MB	Medulloblastoma
MYCN	MYCN amplified
MNA-LR	MYCN non-amplified low risk
MNA-HR	MYCN non-amplified high risk
MES	Mesenchymal
WNT	Wint
SH	Sonic hedgehog
G3	Group 3
G4	Group 4
CMS	Congenital Myasthenic Syndromes
NMJ	Neuromuscular junction
AChR	Acetylcholine receptor
GO	Gene Ontology

Chapter 1 Introduction

Over the last decades, major advances have been made in understanding the pathogenesis and treatment of paediatric cancers as well as improving survival rates. In particular, the next generation of highly parallel sequencing platforms have enabled the thorough analysis of genetic and epigenetic abnormalities in tumour cells leading to the discovery of several genes involved in childhood cancers [2]. Despite such remarkable improvements, paediatric oncologists are facing significant challenges spanning several aspects that require further investigation and efforts, in particular the identification of specific *molecular subtypes of disease* aimed to implement targeted therapies that could substantially help enhance cure rates and reduce long-term sequelae improving the overall quality of life of patients. The main difficulties in the discovery of molecular subtypes in childhood oncology reside in countless factors, including the marked heterogeneity within paediatric cancer types, the relationship between survival rates and disease stage and age at clinical presentation, the intractable drug resistance in individuals with specific genetic alterations, the little number of clinical trials and the challenges related to their effective design, among many others.

The project iPC is focusing on five paediatric cancers, namely medulloblastoma, Ewing's sarcoma, neuroblastoma, hepatoblastoma, and acute lymphoid leukaemia. Molecular subtypes of disease are not known for all these paediatric cancers (**Table 1**). In the cases of neuroblastoma, medulloblastoma, and hepatoblastoma, molecular signatures supported by multi-omics evidence have been revealed, composed of a varied number of genes (from tens of candidate genes to thousands). Surface proteins of B-cells and T-cells are used to classify different types of acute lymphoid leukaemia. Finally, sets of genes with multiple genomic alterations have been identified in Ewing's sarcoma. All these collections of putatively critical genes for childhood cancer makes the development of novel and more insightful tools for gene functional analysis an open challenge and an opportunity in this area. In particular, graph-based approaches demonstrated a great potential in identifying functional characteristics of paediatric rare diseases [1] and hold the promise to effectively incorporate the molecular characterization of paediatric cancers into clinical classifications and actionable interventions.

In this deliverable, we leverage a novel approach for functional characterization of candidate genes in paediatric cancer that BSC and CURIE published in iScience (Cell Press) in 2021 [1]. We report on the application of this methodology based on multilayer networks to analyse individual paediatric cancers (medulloblastoma) and rare diseases (congenital myasthenic syndromes) as well as multiple of them (neuroblastoma, medulloblastoma, Ewing's sarcoma). The objective of the deliverable is to uncover a functional consensus among distinct paediatric cancers that were selected based on the availability of sufficiently populated candidate gene sets.

Paediatric cancer	Molecular subtypes	Reference
Acute lymphoid leukaemia (ALL)	ALL subtypes can be classified by the unique set of proteins, known as "immunophenotypes", found on the surface of leukaemia cells: ● B-cell ALL ○ with recurrent genetic abnormalities ○ not otherwise specified ○ provisional entities* ● T-cell ALL ○ provisional entities*	[3]

	• Mixed phenotype acute leukemias (MPAL)	
Hepatoblastoma (HB)	HB subtypes can be classified by a transcriptomic and epigenomic 16-gene signature (C1, C2) or the increased expression of epithelial-mesenchymal transition markers (C2B).	[4–6]
Neuroblastoma (NB)	NB subtypes can be classified using a 1476-gene signature supported by multi-omic evidence (chromatin conformation, epigenetics, transcriptomics): • MYCN (MYCN amplified) • MNA-LR(MYCN non-amplified low risk) • MNA-HR (MYCN non-amplified high risk) • MES (mesenchymal)	[7]
Ewing's sarcoma (ES)	No molecular subtypes of ES are known. However, STAG2 and CDKN2A genetic alterations have been observed to be mutually exclusive, the first being associated with highly aggressive tumours.	[8]
Medulloblastoma (MB)	MB subtypes can be classified using a 3868-gene signature supported by multi-omic evidence (proteomics, phosphoproteomics, epigenetics, transcriptomics): • WNT • SHH • G3 • G4	[9]

Table 1: Molecular subtypes described for the paediatric cancers under study in the iPC project.

* A provisional entity is a disease subtype for which there is not enough supporting evidence.

Chapter 2 Multilayer community trajectories

2.1 Multilayer networks

A *multilayer network* is a network organised into multiple layers representing different types of nodes and edges (**Figure 1**). Formally, a multilayer network is defined as a quadruplet $M = (V_M, E_M, V, L)$, where V denotes the set of nodes in the multilayer network, L denotes the set of layers , $V_M \subseteq V \times L$ denotes the sets of nodes $v \in V$ contained in each layer, and $E_M \subseteq V_M \times V_M$ denotes the sets of edges connecting tuples of nodes and layers (v, l), $(v', l') \in V_M$ [10]. In a multilayer network, an edge can be an *intralayer edge*, i.e. it connects nodes in the same layer ($l = l'$), or *interlayer edge*, i.e. it connects nodes from different layers ($l \neq l'$). If the nodes connected by the interlayer edges are the same (i.e., they represent the same entity in different layers), interlayer edges are also called identity or coupling edges.

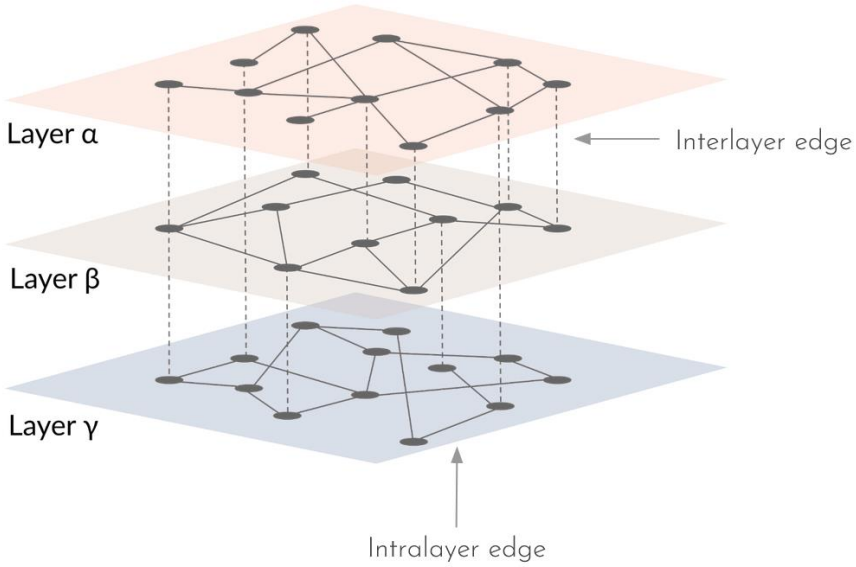

Figure 1: A multilayer network composed of three layers (α, β, γ).

Nodes are connected within each layer through intralayer edges and among different layers through interlayer edges. In the represented multilayer network, interlayer edges only connect the same nodes in each layer.

Despite offering the means to achieve a comprehensive view of human diseases by accounting for the complexity of accumulated biomedical data, multilayer networks exhibit a range of research challenges that still requires substantial investigation. Among them, community detection in multilayer networks is an area of investigation particularly promising for biomedicine, facilitating the evaluation of relevant associations among genes and the identification of candidate targets for drug development and repurposing [11].

2.2 Community detection in multilayer networks

Community detection in multilayer networks can be achieved with several methods, the *Louvain algorithm* being among the most efficient. The Louvain algorithm for community detection consists of two recursive steps. In the first step, nodes are assigned to communities and then moved to others until no increase in modularity is observed. In the second step, the identified communities are

aggregated so that a new graph is created and the entire process starts again and proceeds until convergence. A community (c) is defined as groups of densely connected nodes in the different layers $l \in L$. The algorithm is parametrized to the resolution parameter γ: the higher the value of γ, the smaller the size of the detected multilayer communities. The software MolTi [12,13] implements the Louvain algorithm for community detection in multilayer networks and defines the *modularity of a multilayer network* X as

$$Multilayer\ modularity = \sum_{l} \frac{w^{(l)}}{2m^{(l)}} \sum_{\substack{\{i,j\} \\ i \neq j}} \left(X_{i,j}^{(l)} - \gamma \frac{S_i^{(l)} S_j^{(l)}}{2m^{(l)}} \right) \delta_{c_i,c_j}$$

where the first sum runs over all layers of the multilayer network and the second over all edges $\{i, j\}$ of each layer l. $X^{(l)}_{i,j}$ is the weight of the edge $\{i, j\}$ in a layer l; $S(l)_i$ is the sum of the weights of all the edges involving vertex i in that layer; $m(l)$ is the sum of the weights of all the edges of that layer; $\delta_{ci,cj}$ is equal to 1 if i and j belong to the same community ($ci = cj$) and to 0 otherwise; γ is the resolution parameter; $w^{(l)}$ is the user-defined weight associated to the layer l. In our calculations, $w^{(l)}$ and $X^{(l)}_{i,j}$ are both equal to 1, so that $m^{(l)}$ represents the total number of edges in l and $S^{(l)}_i$ and $S^{(l)}j$ represent the degree of nodes i and j, respectively.

2.3 Multilayer community trajectories and the CmmD tool

BSC in collaboration with iPC partner CURIE published a research article on the application of community detection in multilayer networks to paediatric cancers, specifically medulloblastoma, in the journal iScience (Cell Press) in March 2021 [1]. The publication, titled "*The multilayer community structure of medulloblastoma*", introduces a methodology for the analysis of the community structures that coexist in a multilayer network at different values of modularity resolution and exploit this structural property for several purposes, specifically dimensionality reduction and functional characterization of medulloblastoma subtypes.

We first constructed a multilayer network with identity interlayer edges composed of five layers of gene associations retrieved from reputable knowledge bases (molecular interactions, targeting drugs, variant-associated diseases, pathways, common reaction metabolites) [1]. We then collected the different community structures found in a range of modularity resolution (γ), identifying an endpoint of interest for this range as the value where the average community size, as a function of the number of communities, establishes a plateau, i.e. where the first derivative equals zero with 0.05 margin of error (**Figure 2**). The endpoint was found at γ=12 (964 multilayer communities).

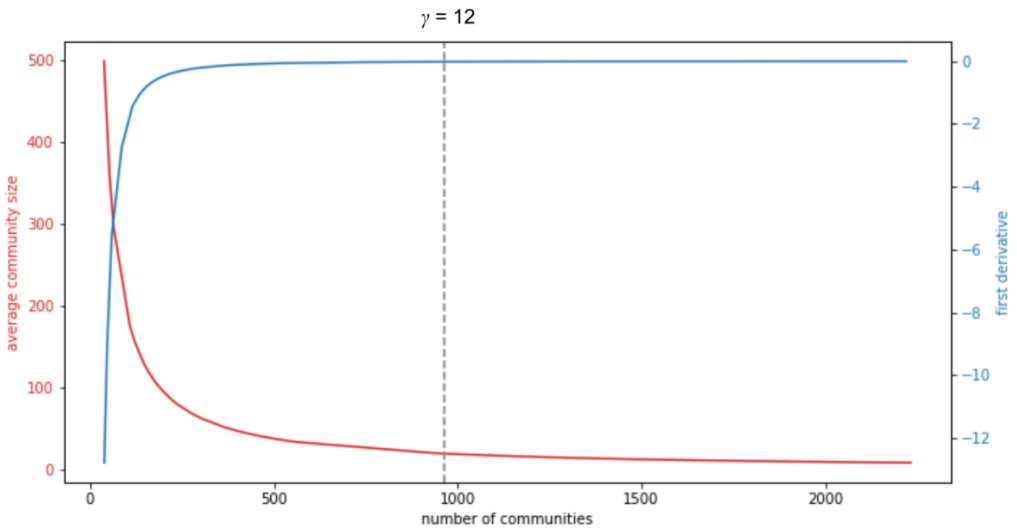

Figure 2: Identification of the resolution range of interest.

The modularity resolution parameter (γ) determines the number of communities and their size. The most dramatic changes in both size and number of communities occur in an initial range of resolution, which enables detection of genes that are strongly associated. We identified the endpoint of this range (γ=12) as the value where the average community size, as a function of the number of communities, establishes a plateau (i.e. its first derivative equals zero with 0.05 margin of error).

The gene composition of the communities may vary depending on the resolution. Some genes tend to share the same communities at all resolutions, while others have different trajectories while progressively increasing the resolution. To compare the trajectories of each gene along the communities, we computed the pairwise Hamming distance among the vectors of communities visited by each gene in the resolution range. We refer to these vectors as *multilayer community trajectories*. The higher the distance, the more times two genes belong to different communities within this range (**Figure 3**).

The obtained distance matrix can be used for several purposes. In the publication, we apply it in an optimization process to find the minimal number of genes that recapitulate known medulloblastoma subgroups (see Section 2.4). In other applications, we used such distance matrices to identify modules of genes that consistently share the same communities at any resolution (see Section 2.5) or clusters of trajectories that share similar functions among different paediatric cancers (see Chapter 3). The *R package CmmD*, which automates these operations, is available at https://github.com/ikernunezca/CmmD.

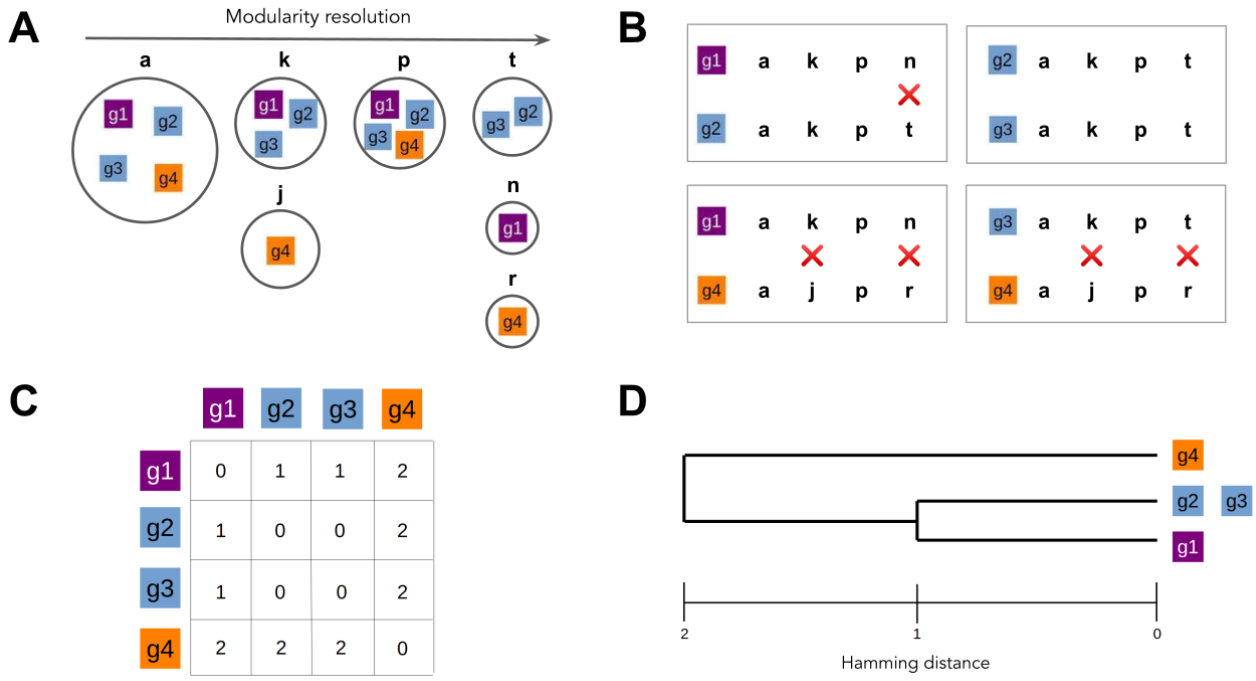

Figure 3: Identification of multilayer community trajectories.

For a given set of genes, we identified the multilayer communities to which they belong in a range of modularity resolution (A). We then computed the pairwise Hamming distances of the trajectories of communities visited by each gene (B). The corresponding distance matrix (C) was represented in the form of a dendrogram (D) used for clustering analysis.

2.4 Application to medulloblastoma

Medulloblastoma (MB) is a malignant and fast-growing primary central nervous system tumour, which originates from embryonic cells of the brain or spinal cord with no known causes and a preferential manifestation in children (one to nine years old). Despite being rare, medulloblastoma is

the most common cancerous brain tumour in children. Four molecular disease subtypes of paediatric medulloblastomas with distinct clinicopathological features have been identified: WNT, SHH, Group 3 (G3), and Group 4 (G4) [14]. WNT is associated with the most favourable prognosis, while SHH and G4 with intermediate-level prognosis and G3 with the worst outcome.

The biomedical goal of the study is to identify the minimal number of genes, among those revealed by proteogenomic data, that recapitulate the four biomedically relevant medulloblastoma subtypes (WNT, SHH, G3, and G4) (Forget et al. 2018). Identifying a minimal set of genes is crucial for both the definition of molecular signatures and the research on disease mechanisms. To achieve this goal, we performed a series of hierarchical clustering analyses (Ward's linkage method) where the similarity between two patients (A and B) was measured as the Jaccard index (J) of sets of altered genes selected using two parameters, θ and λ:

$$J(A_{\theta,\lambda}, B_{\theta,\lambda}) = \frac{A_{\theta,\lambda} \cap B_{\theta,\lambda}}{A_{\theta,\lambda} \cup B_{\theta,\lambda}}$$

The parameter θ defines the maximum Hamming distance allowed to include genes in the analysis, while the parameter λ defines the maximum number of them that must co-occur in the same communities along their trajectories. For dimensionality reduction purposes, small values of θ and λ guarantee a selection of genes with similar trajectories and in minimal numbers. Hence, we formulated an optimization procedure to systematically evaluate values of θ and λ to identify the ones that maximise the accuracy of the corresponding optimal clusters, found using the partitioning around medoids (PAM) algorithm, that recapitulate patient stratification into the four medulloblastoma subtypes (WNT, SHH, G3, G4).

We achieved the highest accuracy (94.94%) and MCC (87%) with 5 clusters (WNT, SHH, G4, G3, and G3-G4) (**Figure 4**), by selecting for each patient those genes that are represented in the communities in sets of, at most, 6 (λ = 6), and that are always part of the same communities along their trajectories (θ = 0). Strikingly, such high accuracy corresponds to a strict selection of genes, indicating that only a small portion of the genes altered in a patient is sufficient to accomplish an accurate patient segregation. This observation implies that the selected genes are tightly associated and never leave the communities they belong to along their trajectories.

Additional analyses, including robustness and sensitivity analyses, and layer-specific enrichments of the minimal set of genes identified are reported in the publication [1].

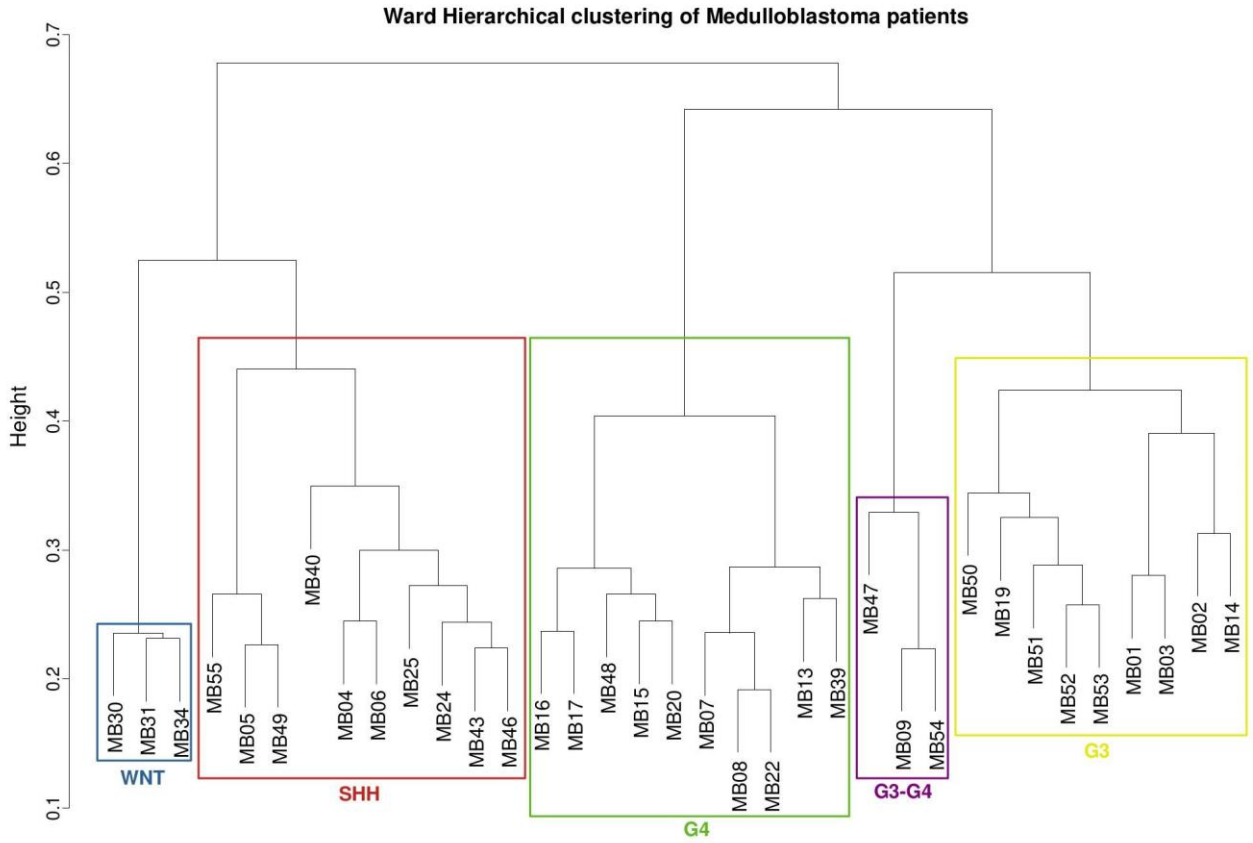

Figure 4: Ward's linkage hierarchical clustering obtained at λ = 6 and θ = 0.

The rectangles indicate the 5 clusters suggested by PAM (partitioning around medoids) criteria. The colour of each cluster indicates the original patient stratification into the four medulloblastoma subtypes via network fusion (Forget et al. 2018): WNT group (blue), SHH group (red), Group 4 (G4, green), Group 3 (G3, yellow). A fifth cluster is depicted in purple, including 3 patients originally assigned to groups G3 (MB47) and G4 (MB09 and MB54).

2.5 Other applications (congenital ,myasthenic syndromes)

We have recently applied CmmD and the multilayer community trajectory analysis to the study of rare diseases other than paediatric cancers. Congenital Myasthenic Syndromes (CMS) are a group of diverse rare diseases characterised by neuromuscular junction (NMJ) dysfunctions resulting in muscle weakness. While causative genes have been previously described for these conditions, a molecular explanation for the observed differences in phenotypic severity remains unclear. In collaboration with research institutions in Spain, The Netherlands and Canada, we undertook an in-depth analysis of a cohort of 20 CMS patients from an isolated population bearing the same homozygous mutation of the *CHRNE* gene, which encodes a subunit of the acetylcholine receptor (AChR) (**Figure 5**). By identifying genes that are found in the same communities in the entire range of resolution parameter and segregate between severe and non-severe cases, our results show that CMS severity can be ascribed to the personalised impairment of specific classes of NMJ proteins, namely extracellular matrix components (proteoglycans, tenascins, chromogranins) and postsynaptic modulators of AChR clustering. By coupling multilayer community trajectory analysis and omics information, we identified and experimentally validated in *D. rerio* the modifying effect of a gene that was previously unknown to be a NMJ interactor. This work is about to be submitted to a high-impact journal for peer-reviewing.

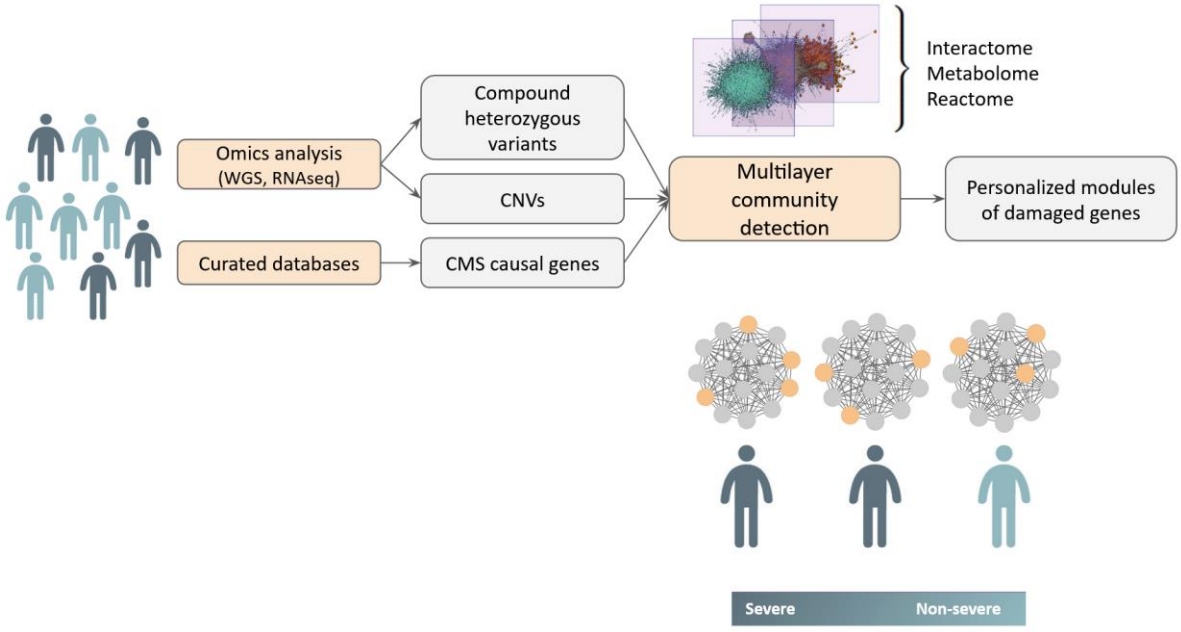

Figure 5: Analytical workflow employed to address the severity of a cohort of patients affected by Congenital Myasthenic Syndromes (CMS).

A multi-scale functional analysis approach, based on multilayer networks, was used to identify the relationships between an association of functionally related alterations obtained from omics data (Whole Genome Sequencing, WGS; RNA-sequencing, RNAseq) with known CMS causal genes. Modules of CMS linked genes that emerged from this analysis were characterised at the level of single individuals.

Chapter 3 A functional consensus of paediatric cancers based on multilayer community trajectories

3.1 Clusters of multilayer community trajectories

In this analysis we leverage the multilayer network used in a recent publication [1]. This multilayer network is composed of five layers (molecular interactions, targeting drugs, variant-associated diseases, pathways, common reaction metabolites). The layers contain a total of 18,948 genes (Entrez identifiers) for which we computed pairwise Hamming distances based on their multilayer community trajectories in a defined range of modularity resolution (see Section 2.3 in Chapter 2). We focused our analysis on cancer types with sufficiently large gene sets that are significantly altered or representing molecular signatures supported by multi-omics evidence (medulloblastoma, neuroblastoma, Ewing's sarcoma; see **Table 1**). Then, we identified clusters of genes based on similarity in their multilayer community trajectories comparing a *relaxed clustering criterion* (similar trajectories) and a *strict clustering criterion* (identical trajectories). Finally, we computed Gene Ontology enrichments in each cluster of each paediatric tumour and evaluated the functional consensus among them.

3.1.1 Relaxed clustering criterion (similar trajectories)

While the distance metric is defined (Hamming distance), the optimal linkage method to hierarchically cluster the multilayer community trajectories was chosen based on the *cophenetic correlation coefficient*. If the clustering is valid, the linking of objects in the cluster tree should have a strong correlation with the distances between objects in the distance vector. The optimal number of clusters is the one corresponding to the maximum *Calinski-Harabasz score*.

3.1.2 Strict clustering criterion (identical trajectories)

As shown in the case of medulloblastoma and congenital myasthenic syndromes (see Sections 2.4 and 2.5 in Chapter 2), the identification of genes that exhibit exactly the same trajectories is a strong indication of functional relatedness among them. The partition of genes with identical trajectories corresponds to those with a Hamming distance of zero.

3.2 Neuroblastoma

To functionally characterise neuroblastoma, we used the signature of 1476 genes targeted by the super enhancers identified and assigned to four disease subtypes (MYCN, MNA-LR, MNA-HR, MES) by Gartlgruber and collaborators [7]. After converting gene symbols to Entrez identifiers, 1153 genes of this signature were found in the multilayer network. The relaxed clustering criterion (similar trajectories) led to an optimal number of 743 clusters of which 139 contained more than one gene and were populated by an average of 3.97 genes (**Table 2**; **Figures Figure** 6**-Figure** 7). The strict clustering criterion (identical trajectories) identified 124 clusters with more than one gene (3.87 genes per cluster on average).

linkage method	cophenetic correlation coefficient
average	0.886239
weighted	0.855589
complete	0.815867
centroid	0.693149
median	0.520462
single	0.436420
ward	0.341389

Table 2: Cophenetic correlation coefficient for several linkage methods.

The optimal hierarchical clustering method for the neuroblastoma gene signature is ‘average’.

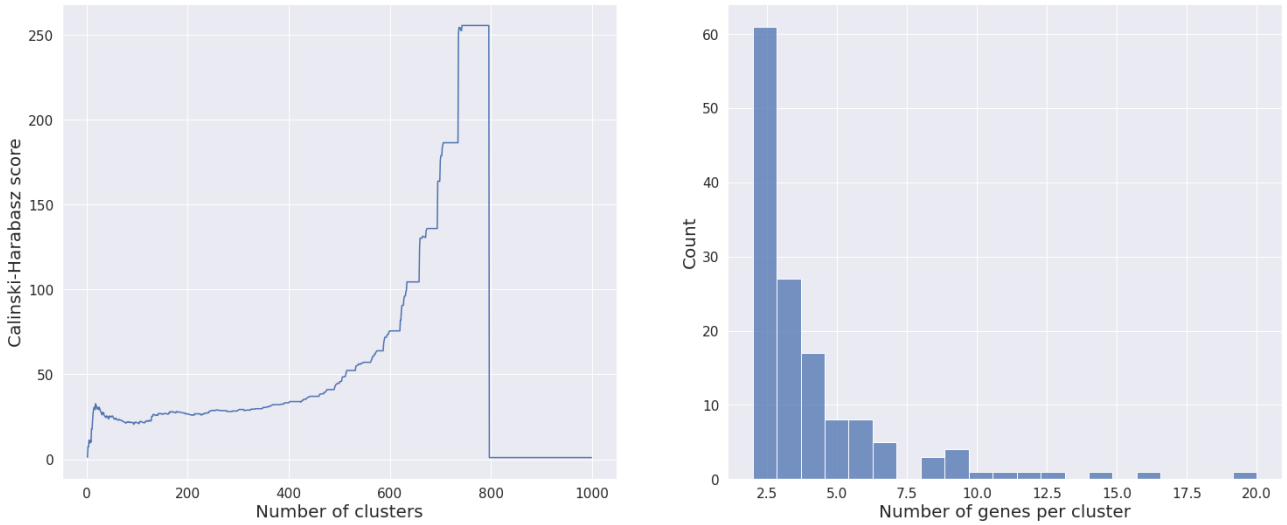

Figure 6: Optimal number of multilayer community trajectory clusters in neuroblastoma.

Number of clusters as a function of the Calinski-Harabasz score; the optimal number of clusters is 743 (left panel). Counts of the number of genes in clusters populated by more than one gene (average 3.97; right panel).

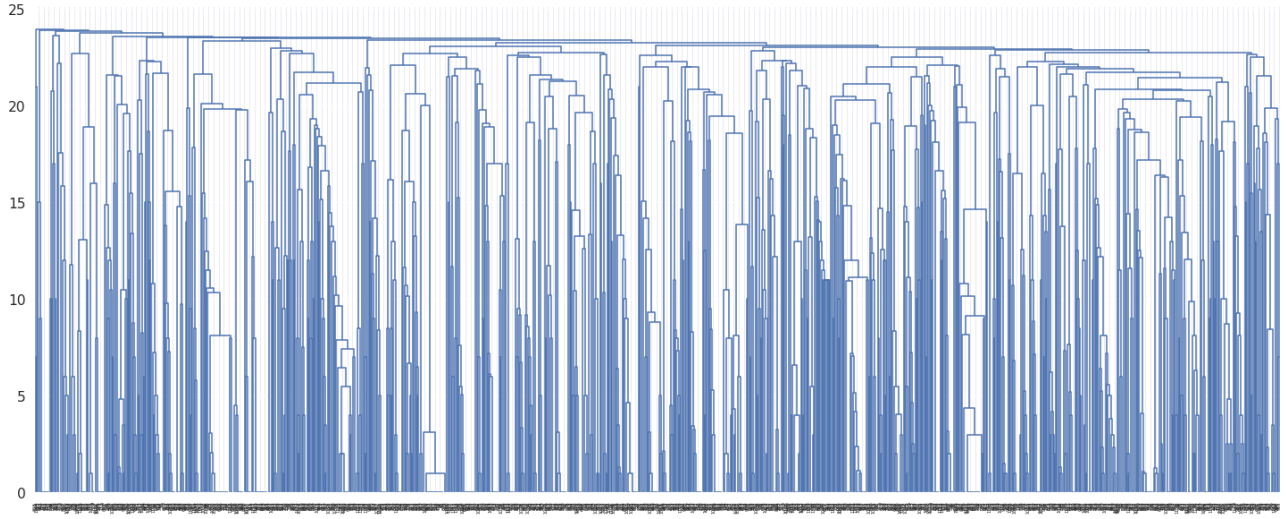

Figure 7: Dendrogram showing the hierarchical clustering of neuroblastoma signature genes based on multilayer community trajectories.

3.3 Medulloblastoma

In the case of medulloblastoma, we used the 3838 proteogenomic signature quantified by Forget and collaborators [9]. After converting gene symbols to Entrez identifiers and filtering out the genes that were not present in the multilayer network, we obtained a list of 3677 genes. After clustering the genes by multilayer community trajectories, we identified 5 clusters with more than one gene, with an average number of 735.4 genes per cluster (**Table 3**; **Figures Figure** 8**-Figure** 9). The strict clustering criterion (identical trajectories) identified 274 clusters with more than one gene (6.89 genes per cluster on average).

linkage method	cophenetic correlation coefficient
average	0.925982
weighted	0.878883
complete	0.830810
centroid	0.766416
median	0.558795
single	0.475755
ward	0.316795

Table 3: Cophenetic correlation coefficient for several linkage methods.

The optimal hierarchical clustering method for the medulloblastoma gene signature is ‘average’.

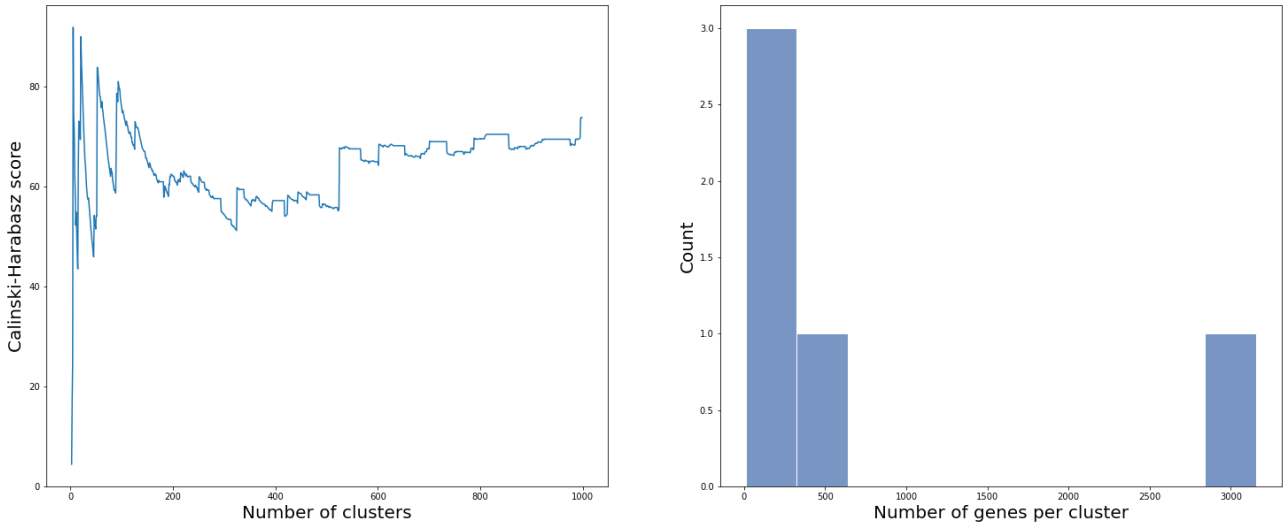

Figure 8: Optimal number of multilayer community trajectory clusters in medulloblastoma.

Number of clusters as a function of the Calinski-Harabasz score; the optimal number of clusters is 5 (left panel). Counts of the number of genes in clusters populated by more than one gene (average 735.4; right panel).

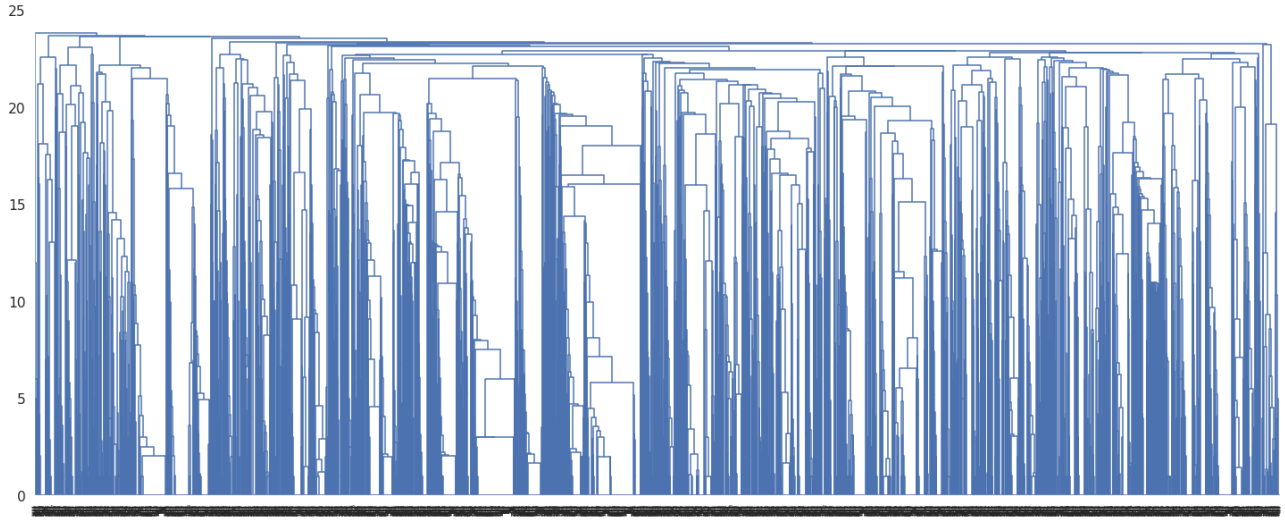

Figure 9: Dendrogram showing the optimal hierarchical clustering of medulloblastoma signature genes based on multilayer community trajectories.

3.4 Ewing's sarcoma

To functionally characterise Ewing's sarcoma, we used the set of 50 significantly mutated genes identified by Tirode and collaborators [8]. After converting gene symbols to Entrez identifiers, 46 genes of this set were found in the multilayer network. The hierarchical clustering of the multilayer community trajectories of these genes detected 39 optimal clusters of which 4 contain more than one gene (2.75 genes per cluster on average) (**Table 4; Figures Figure** 10-**Figure** 11). Similarly, the strict clustering criterion (identical trajectories) identified 3 clusters with more than one gene (2.33 genes per cluster on average).

linkage method	cophenetic correlation coefficient
average	0.992301
weighted	0.991969
complete	0.988707
single	0.969303
centroid	0.928789
median	0.871606
ward	0.598645

Table 4: Cophenetic correlation coefficient for several linkage methods.

The optimal hierarchical clustering method for the Ewing's gene signature is 'average'.

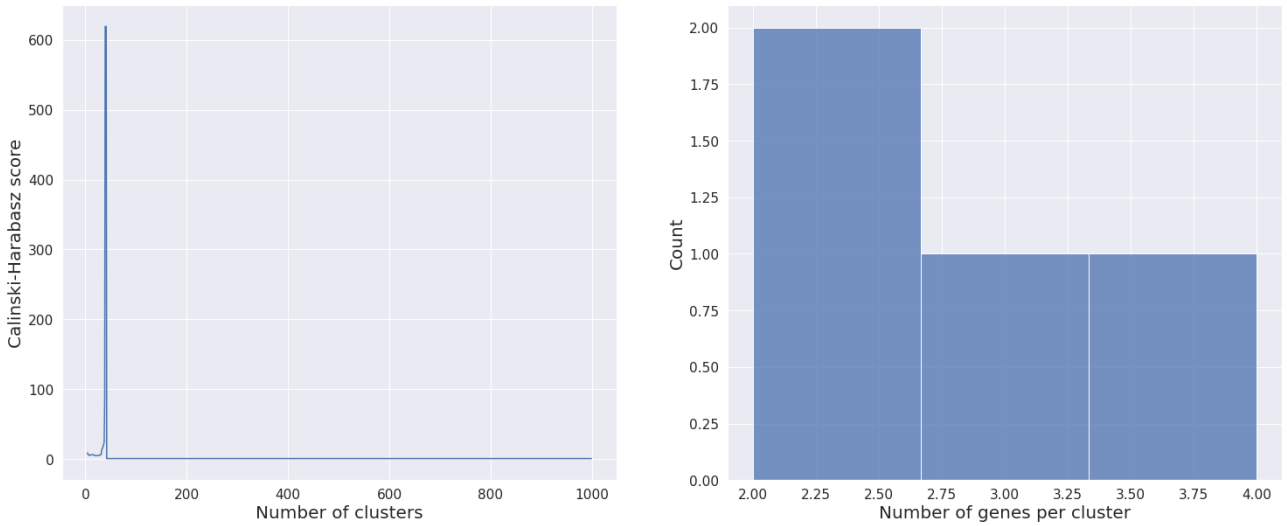

Figure 10: Optimal number of multilayer community trajectory clusters in Ewing's sarcoma.

Number of clusters as a function of the Calinski-Harabasz score; the optimal number of clusters is 4 (left panel). Counts of the number of genes in clusters populated by more than one gene (average 2.75; right panel).

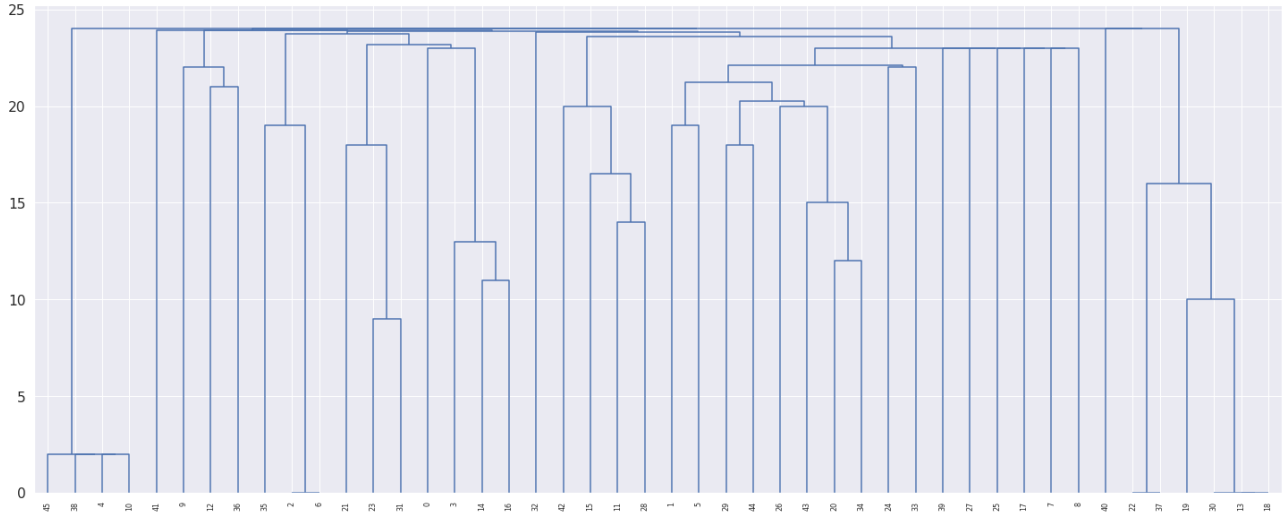

Figure 11: Dendrogram showing the optimal hierarchical clustering of Ewing's sarcoma signature genes based on multilayer community trajectories.

3.5 Functional consensus

By comparing the results of the relaxed and strict clustering criteria (**Table 5**), it is apparent how the strict clustering criterion is able to capture a more diverse landscape of gene groups in medulloblastoma then the relaxed criterion, while resulting in similar clusters in neuroblastoma and Ewing's sarcoma. For this reason, we computed Gene Ontology (GO) functional enrichments of the genes found in the clusters identified with the strict clustering criterion. In particular, we focused on the 12 common clusters that are represented in the three paediatric cancers (**Figure 12; Table 6**), and visually rendered the SimRel semantic similarity of the corresponding 359 enriched GO terms

using the Tree Map visualisation tool of Revigo (http://revigo.irb.hr/) [15] (**Figures Figure** 13**-Figure** 15). All GO enrichments has been computed using g:Profiler[1] [16].

Among the many functional enrichments, the 25 genes belonging to cluster 9 (see Table 6) are enriched in functions related to telomere dynamics, microtubules, and binding to PP2CA, a major phosphatase for microtubule-associated proteins [17]. Interestingly, three genes (*PDCL3*, *STK24*, *STRIP1*) are common between neuroblastoma and medulloblastoma, and one (*STRN4*) is common between medulloblastoma and Ewing's sarcoma. All the other genes are uniquely found in the three paediatric cancers gene signatures although they share exactly the same multilayer community trajectory, thus a tight functional relatedness among them. Few other genes from this analysis are in common between signatures. *TBCD* (cluster 3, functionally related to cytoskeleton), *HPCAL1* (cluster 5, functionally related to extracellular exosome), is found in neuroblastoma and medulloblastoma, while *ANK2* and *NUP155* (cluster 1, functionally related to ion channel activities) are found in medulloblastoma and Ewing's sarcoma. In the majority of the case, the common clusters are composed of genes from distinct signatures. For instance, the 11 genes belonging to cluster 7 are enriched in hormonal regulation and response but none of them is found in multiple cancers. These genes belong to the same communities but are specific to each paediatric cancer.

Paediatric cancer	Relaxed clustering criterion (similar trajectories)		Strict clustering criterion (identical trajectories)	
	Number of clusters with >1 genes	Average number of genes per cluster	Number of clusters with >1 genes	Average number of genes per cluster
Neuroblastoma	139	3.97	124	3.87
Medulloblastoma	5	735.4	274	6.89
Ewing's sarcoma	4	2.75	3	2.33

Table 5: Number of clusters and average number of genes per cluster recovered by two clustering criteria.

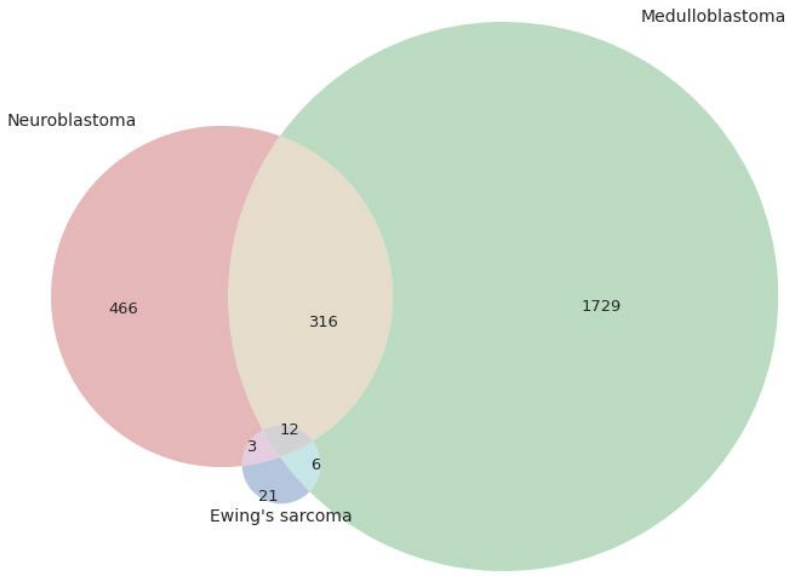

Figure 12: Venn diagrams showing the intersection of clusters of multilayer community trajectories identified in neuroblastoma, medulloblastoma and Ewing's sarcoma.

[1] Main parameters: annotated domain scope, Bonferroni adjusted p-values with 0.05 significant threshold, experimental evidence codes, intersection size >3.

cluster ID	Neuroblastoma	Medulloblastoma	Ewing's sarcoma
0	ZNF77;ZNF703;ZNF704;ZFP30;VWCE;ZNF169;ZNF664;ZNF605;ZNF662	ZNF24;ZKSCAN1;TRIM28;ZNF316	ZNF721
1	MYL4;SCN5A;CACNA1C;DPP6;ANK2;RYR3;KCNH2;RYR2	AKAP9;GPD1L;NUP155;GNAI2;ANK2	KCNIP1;NUP155;DPP10
2	SIGLEC10;CD1D	CXADR	TREML1
3	AGBL4;TBCD	TUBB;TUBB3;EML2;ATAT1;TUBB2A;TUBB6;TBCC;TUBB4B;TBCA;TUBB2B;TUBA1C;EML4;ARL2;TUBB4A;TTC5;TBCD	AGBL1
4	GNAZ;NPY;GPR18;GAL;GRM8;APLNR;CXCR4	HEBP1;GPSM1	GRM4
5	DAAM2;CD164;HPCAL1;SPRY1;SPRY2;KRT19;CREB5	KRT18;HPCAL1;GLRX3;DSG2;CAPNS1;KRT15;KRT8	LCE1C;KRT38
6	ARHGAP28;SOX11;TCF7L2;PAX4;IRS2	PTPN1;HMGA1;APPL1;IRS1;IGF2BP2	PCDHB10
7	PRLHR;HRH1	GNA11;GNAQ;NLN	OXTR
8	TRIM67;DLC1;CCND1;ODC1	CTNNB1;BAX;PTPN12;SRC	TP53
9	STRIP1;RASSF3;CDCA4;PDCL3;STK24;ASTN2;KLHDC8A	MOB4;CCT7;STRN;CCT3;STRN3;TTC27;STRIP1;PDCL3;STK24;CCT5;STRN4;TCP1;CCT4;CCT6A;PDCD10;CTTNBP2NL;STK26	STRN4
10	SORL1	TGFBI	APCS
11	NR6A1;RORA	NR2C2;NR2C2AP;NRBP1	NR0B1

Table 6: Genes that belong to the common clusters of multilayer community trajectories found in neuroblastoma, medulloblastoma, and Ewing's sarcoma using the strict clustering criterion.

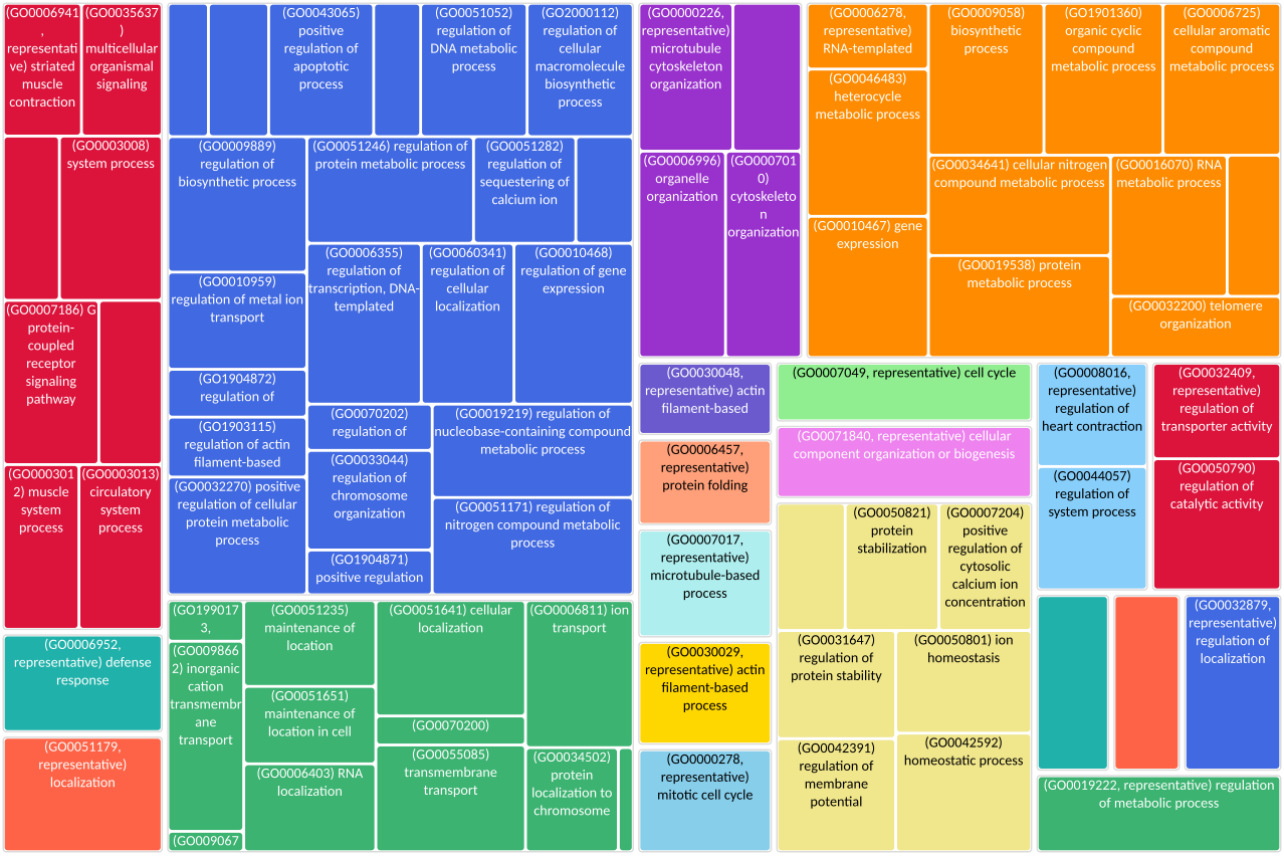

Figure 13: Revigo Tree Map visualisation of pairwise SimRel semantic similarities of common enriched GO terms (Biological Processes) in the common clusters of multilayer community trajectories of neuroblastoma, medulloblastoma and Ewing's sarcoma.

Figure 14: Revigo Tree Map visualisation of pairwise SimRel semantic similarities of common enriched GO terms (Molecular Functions) in the common clusters of multilayer community trajectories of neuroblastoma, medulloblastoma and Ewing's sarcoma.

Figure 15: Revigo Tree Map visualisation of pairwise SimRel semantic similarities of common enriched GO terms (Cellular Components) in the common clusters of multilayer community trajectories of neuroblastoma, medulloblastoma and Ewing's sarcoma.

Chapter 4 Summary and Conclusions

Molecular disease subtyping is a fundamental tool to achieve an effective patient stratification for clinical trials, preventive and therapeutic interventions. In some cancers, such as breast cancer and blood cancers, subtyping has been very successful thanks to the statistical power brought by cohorts composed of large numbers of patients. Rare diseases, such as paediatric cancers, represent a more challenging situation since, by definition, they affect a small number of patients and studies that in most cases are in the order of tens of subjects. In our view, a meaningful molecular subtyping of rare diseases can be achieved by leveraging the wealth of biomedical information that is available in public knowledge bases and that can be integrated in the form of a multilayer network. In particular, achieving patient stratification by means of structural features (multilayer community trajectories) extracted from a general-purpose multilayer network represents a way to both identify the minimal set of genes that characterise the subgroups and, most importantly, to obtain information about the types of relations that define the associations of such genes (e.g. targeting drugs, pathways, physical interactions).

In this deliverable we leveraged a novel multilayer network community analysis framework that was originally applied to a study on medulloblastoma and published in a collaborative effort between iPC partners BSC and iPC [1]. We reported on that work and further applications of this methodology to other rare diseases (congenital myasthenic syndromes). Moreover, we explored the ability of multilayer community trajectories to enable the identification of a functional consensus among diverse paediatric cancers, namely medulloblastoma, neuroblastoma, and Ewing's sarcoma. We tested different strategies to determine gene clusters based on multilayer community trajectories and computed functional enrichments that revealed specific processes that involve genes belonging to the same communities but acting distinctively in the three cancers.

This report demonstrates the ability of multilayer community trajectory analysis to uncover functional communalities among paediatric cancers. The main limitation of this work consists in the availability of candidate gene sets to employ in the analysis. We used available gene signatures for three paediatric cancers that are sufficiently populated. Nevertheless, the type of study would benefit from a set of *bona fide* gene candidates supported by multi-omics evidence and involved in key cancer processes, such as the meta-gene identified by the work developed in the context of D4.3 or the paediatric cancer genes mentioned in the literature such as those detected in D3.2.

Chapter 5 Bibliography

[1] Núñez-Carpintero I, Petrizzelli M, Zinovyev A, Cirillo D, Valencia A. The multilayer community structure of medulloblastoma. iScience 2021;24:102365.

[2] Pui C-H, Gajjar AJ, Kane JR, Qaddoumi IA, Pappo AS. Challenging issues in pediatric oncology. Nat Rev Clin Oncol 2011;8:540–9.

[3] Leukemia & Lymphoma Society. Acute Lymphoblastic Leukemia (ALL) in Children and Teens 2021. https://www.lls.org/booklet/acute-lymphoblastic-leukemia-all-children-and-teens (accessed May 14, 2022).

[4] Cairo S, Armengol C, De Reyniès A, Wei Y, Thomas E, Renard C-A, et al. Hepatic stem-like phenotype and interplay of Wnt/beta-catenin and Myc signaling in aggressive childhood liver cancer. Cancer Cell 2008;14:471–84.

[5] Hooks KB, Audoux J, Fazli H, Lesjean S, Ernault T, Dugot-Senant N, et al. New insights into diagnosis and therapeutic options for proliferative hepatoblastoma. Hepatology 2018;68:89–102.

[6] Carrillo-Reixach J, Torrens L, Simon-Coma M, Royo L, Domingo-Sàbat M, Abril-Fornaguera J, et al. Epigenetic footprint enables molecular risk stratification of hepatoblastoma with clinical implications. J Hepatol 2020;73:328–41.

[7] Gartlgruber M, Sharma AK, Quintero A, Dreidax D, Jansky S, Park Y-G, et al. Super enhancers define regulatory subtypes and cell identity in neuroblastoma. Nat Cancer 2021;2:114–28.

[8] Tirode F, Surdez D, Ma X, Parker M, Le Deley MC, Bahrami A, et al. Genomic landscape of Ewing sarcoma defines an aggressive subtype with co-association of STAG2 and TP53 mutations. Cancer Discov 2014;4:1342–53.

[9] Forget A, Martignetti L, Puget S, Calzone L, Brabetz S, Picard D, et al. Aberrant ERBB4-SRC Signaling as a Hallmark of Group 4 Medulloblastoma Revealed by Integrative Phosphoproteomic Profiling. Cancer Cell 2018;34:379–95.e7.

[10] Kivelä M, Arenas A, Barthelemy M, Gleeson JP, Moreno Y, Porter MA. Multilayer networks. J Complex Netw 2014;2:203–71.

[11] Halu A, De Domenico M, Arenas A, Sharma A. The multiplex network of human diseases. NPJ Syst Biol Appl 2019;5:15.

[12] Didier G, Valdeolivas A, Baudot A. Identifying communities from multiplex biological networks by randomized optimization of modularity. F1000Res 2018;7:1042.

[13] Didier G, Brun C, Baudot A. Identifying communities from multiplex biological networks. PeerJ 2015;3:e1525.

[14] Taylor MD, Northcott PA, Korshunov A, Remke M, Cho Y-J, Clifford SC, et al. Molecular subgroups of medulloblastoma: the current consensus. Acta Neuropathol 2012;123:465–72.

[15] Supek F, Bošnjak M, Škunca N, Šmuc T. REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS One 2011;6:e21800.

[16] Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res 2019;47:W191–8.

[17] Watkins GR, Wang N, Mazalouskas MD, Gomez RJ, Guthrie CR, Kraemer BC, et al. Monoubiquitination promotes calpain cleavage of the protein phosphatase 2A (PP2A) regulatory subunit α4, altering PP2A stability and microtubule-associated protein phosphorylation. J Biol Chem 2012;287:24207–15.