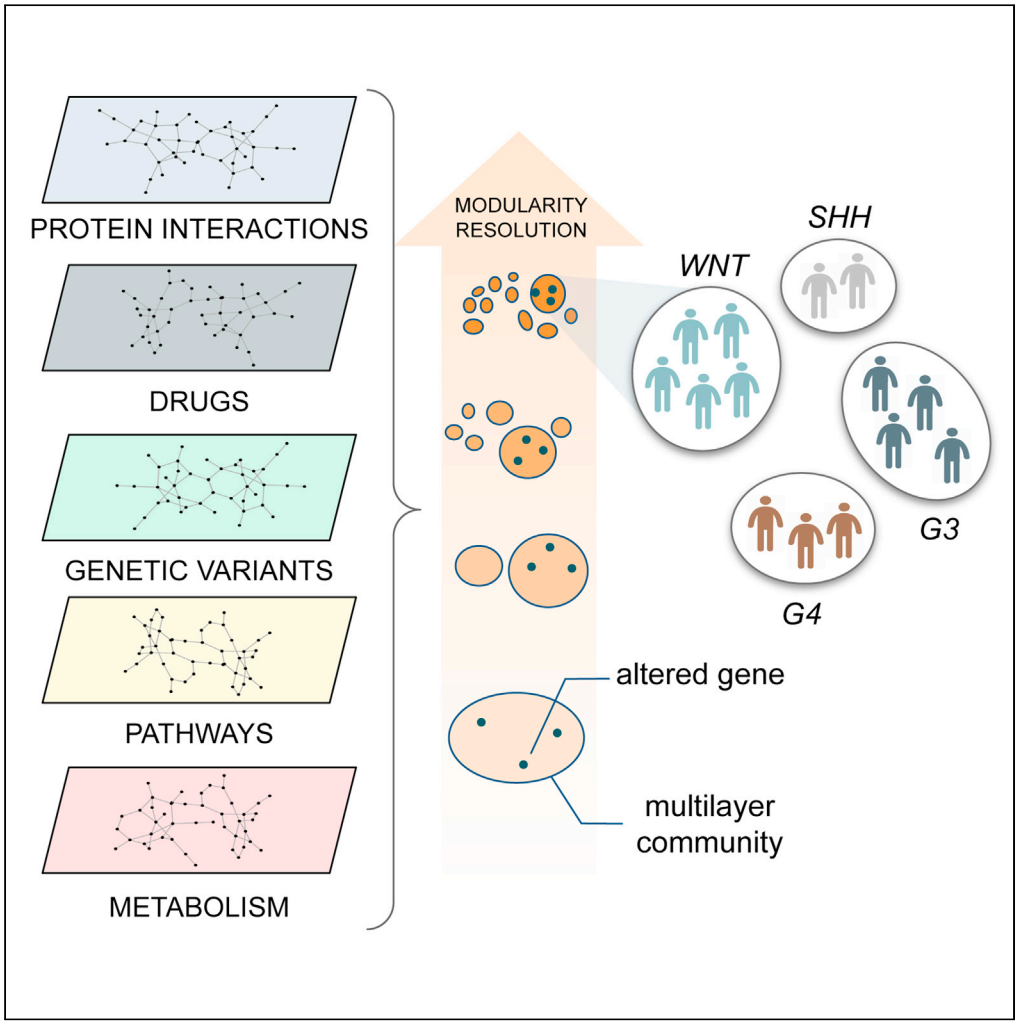
iScience

CellPress
OPEN ACCESS

Article

The multilayer community structure of medulloblastoma



Iker Núñez-Carpintero, Marianyela Petrizzelli, Andrei Zinovyev, Davide Cirillo, Alfonso Valencia

davide.cirillo@bsc.es

Highlights

The molecular interpretation of rare diseases is a challenging task

Multilayer networks allow patient stratification and explainability

We identify subgroup-specific genes and multilayer associations in medulloblastoma

Multilayer community analysis enables the molecular interpretation of rare diseases

Núñez-Carpintero et al., iScience 24, 102365 April 23, 2021
https://doi.org/10.1016/j.isci.2021.102365

iScience

CellPress
OPEN ACCESS

Article

The multilayer community structure of medulloblastoma

Iker Núñez-Carpintero,[1] Marianyela Petrizzelli,[2,3,4] Andrei Zinovyev,[2,3,4,5] Davide Cirillo,[1,7,*] and Alfonso Valencia[1,6]

SUMMARY

Multilayer networks allow interpreting the molecular basis of diseases, which is particularly challenging in rare diseases where the number of cases is small compared with the size of the associated multi-omics datasets. In this work, we develop a dimensionality reduction methodology to identify the minimal set of genes that characterize disease subgroups based on their persistent association in multilayer network communities. We use this approach to the study of medulloblastoma, a childhood brain tumor, using proteogenomic data. Our approach is able to recapitulate known medulloblastoma subgroups (accuracy >94%) and provide a clear characterization of gene associations, with the downstream implications for diagnosis and therapeutic interventions. We verified the general applicability of our method on an independent medulloblastoma dataset (accuracy >98%). This approach opens the door to a new generation of multilayer network-based methods able to overcome the specific dimensionality limitations of rare disease datasets.

INTRODUCTION

To improve our understanding of complex systems, it is crucial to take into account the multiple types of relationships that inherently define natural systems. The study of the so-called multilayer networks (alternatively multiplex networks) has recently become one of the most important directions in network science (Kivela et al., 2014; Aleta and Moreno 2019). A multilayer network is a network organized into multiple layers representing different types of nodes and edges (Figure S1). Despite offering the means to achieve a comprehensive view of human diseases by accounting for the complexity of accumulated biomedical data, biological multilayer networks exhibit a range of research challenges that still require substantial investigation (Kristensen et al., 2014). Among them, community detection in multilayer networks is an area of investigation that is particularly promising for biomedicine, facilitating the evaluation of relevant associations among genes and the identification of candidate targets for drug development and repurposing (Halu et al., 2019; Valdeolivas et al., 2019).

Popular strategies for community detection in networks include the Louvain algorithm (Blondel et al., 2008), a greedy optimization technique, to maximize a network structural metric that is called modularity (Newman and Girvan 2004). Modularity is defined as the fraction of edges within a group of nodes that is significantly enriched when compared with a random model. It measures the strength of a given partition of the network (Reichardt and Bornholdt 2006). The Louvain algorithm is one of the most widely used meta-heuristics for community detection in large networks. It outperforms other community detection algorithms in accuracy, scalability, and computing time (Yang et al., 2016). Moreover, the algorithm is implemented in a number of network analysis software, and it has been recently adapted to multilayer networks (Didier et al. 2015; Didier et al. 2018).

Nevertheless, community structure determination in networks remains an open problem to such an extent that the preferred formulation of communities is often domain specific (Porter et al. 2009). One major conundrum of modularity-based approaches to community detection is the intrinsic limit of resolution, by which it is *a priori* impossible to rule out that a community defined at a certain level of resolution may be composed of a cluster of smaller communities (Fortunato and Barthélemy 2007; Lancichinetti and Fortunato 2011). In other words, multiple topological descriptions, each one with its own importance, coexist at different scales that are detected at alternative values of resolution (Arenas et al. 2008). As a

[1]Barcelona Supercomputing Center (BSC), C/ Jordi Girona 29, 08034, Barcelona, Spain

[2]Institut Curie, PSL Research University, 75005 Paris, France

[3]INSERM, U900, 75005 Paris, France

[4]MINES ParisTech, PSL Research University, CBIO-Centre for Computational Biology, 75006 Paris, France

[5]Lobachevsky University, 603000 Nizhny Novgorod, Russia

[6]ICREA - Institució Catalana de Recerca i Estudis Avançats, Pg. Lluís Companys 23, 08010, Barcelona, Spain

[7]Lead contact

*Correspondence: davide.cirillo@bsc.es

https://doi.org/10.1016/j.isci.2021.102365

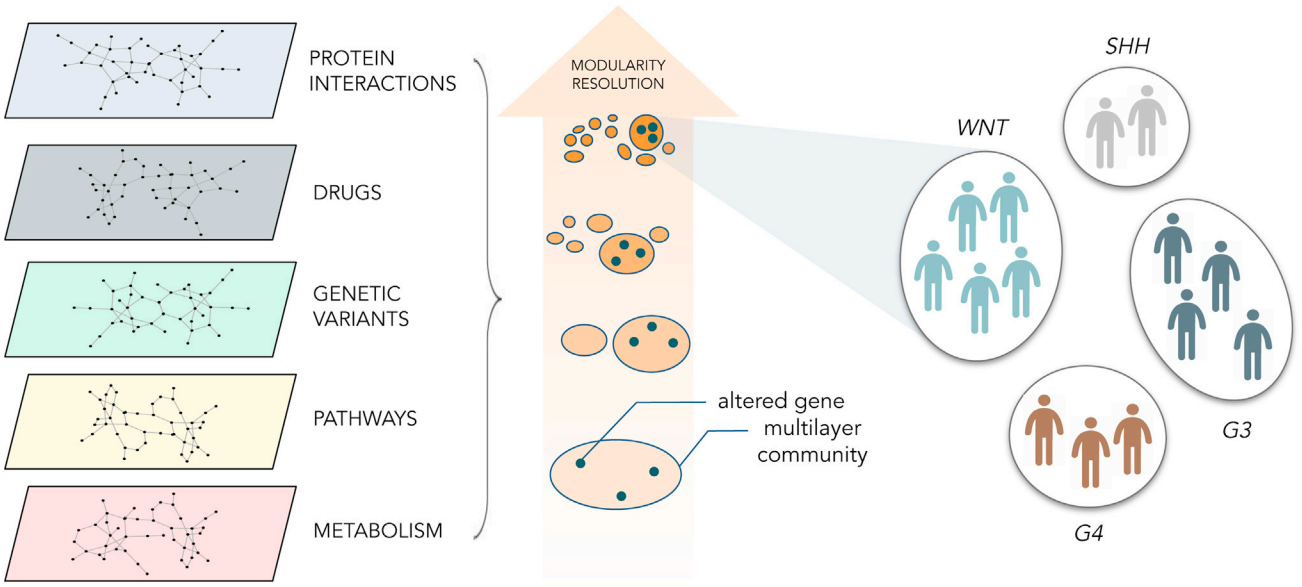


Figure 1. Multilayer community structure analysis of medulloblastoma subgroups

Using multilayer community structure analysis on a network describing gene-gene associations based on protein interactions, drug targets, genetic variants, pathways, and metabolic reactions, we identified the minimum sets of altered genes that optimally cluster the patients with medulloblastoma into previously described subgroups. See also Figures S1 and S2.

consequence, the identification of meaningful network communities, such as groups of genes of interest that robustly express strong associations, heavily depends on the choice of the resolution value to be used. This limitation can be overcome through the identification of stable partitions at different values of resolution. Indeed, the detection of persistent partitions when changing the resolution is indicative of strong modular structures (Arenas et al. 2008).

Here we implemented a methodology to identify groups of genes that are systematically found to belong to the same communities across a range of different resolution values. In this view, two or more genes of interest that are consistently found in the same communities at different values of resolution will be deemed strongly associated based on the multiple biological evidence from the multilayer network. We applied this concept to the analysis of the multilayer community structure of genes altered in a cohort of patients with medulloblastoma (MB) who were previously stratified based on proteogenomic data (Forget et al., 2018) (Figure 1). To this aim, we implemented a dimensionality reduction methodology based on the persistent association of genes in the multilayer network communities (see methods: "multilayer community structure analysis" and Figure 2).

MB is a malignant and fast-growing primary central nervous system tumor, which originates from embryonic cells of the brain or spinal cord with no known causes and a preferential manifestation in children (aged 1–9 years). Despite being rare, MB is the most common cancerous brain tumor in children. Four molecular disease subgroups of pediatric MB with distinct clinicopathological features have been identified: WNT, SHH, Group 3 (G3), and Group 4 (G4) (Taylor et al., 2012; Northcott et al., 2011). WNT is associated with the most favorable prognosis, whereas SHH and G4 are associated with intermediate-level prognosis and G3 with the worst outcome. Seven genes exhibit recurrent genetic alterations in the four subgroups (*SHH* in SHH group, *CTNNB1* in WNT group, *MYC* and MYCN in G3 and G4, *ERBB4*, *SRC*, and *CDK6* in G4 (Kool et al., 2012; Ramaswamy et al., 2016; Taylor et al., 2012; Northcott et al., 2014; Robinson et al., 2012; Northcott et al., 2017; Kool et al., 2014; Clifford et al., 2006; Forget et al., 2018). Each subgroup presents substantial biological heterogeneity and survival differences (Jones et al., 2012) so much so that the identification of more than four subgroups has been recently proposed, in particular as concerns the heterogeneity of G3 and G4 (Schwalbe et al., 2017).

Our results show that our multilayer community structure analysis is able to recapitulate the four MB subgroups (accuracy 94.94%), as well as better characterize them by identifying distinct minimal sets of genes

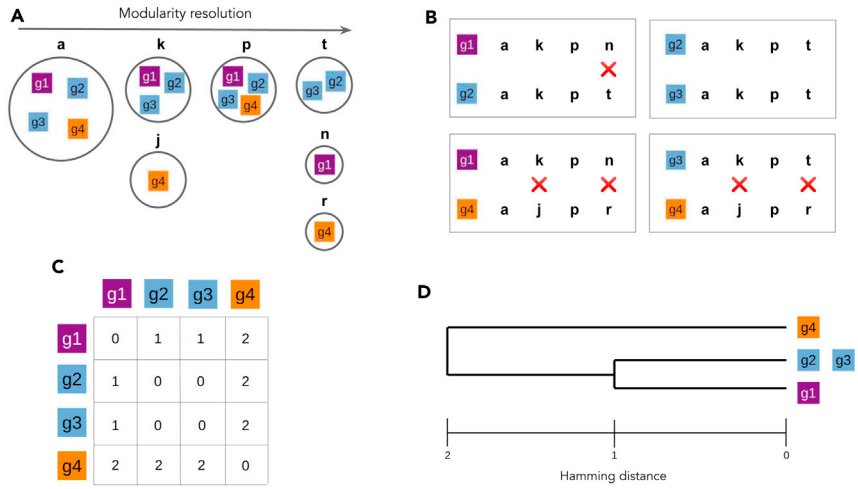


Figure 2. Identification of multilayer community trajectories

(A–D) 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. See also Figure S3.

with strong associations based on multiple layers of evidence (Figure S2). We further verify the applicability of our method using an independent MB multi-omics dataset, achieving a very high performance also in this case (accuracy 98.29%). This work represents an important step forward not only in the characterization of MB subgroups but also, in general, in rare tumor research, where the absence of large patient sample cohorts makes the identification of supporting evidence for candidate genes an extremely challenging task.

RESULTS

Multilayer community trajectories

To implement a way to monitor the behavior of multilayer communities containing MB genes upon changes of the modularity resolution, we initially sought to take into account gene mentions in abstracts of scientific publications about MB (see methods: "data sources of medulloblastoma genes"). By interrogating PubTator Central (PTC) (Wei et al., 2019), we retrieved a list of 1,941 multi-species genes, consisting of 1,475 human genes (76%), 389 murine genes (20%), and 77 genes of other species (4%). We identified the multilayer communities to which the human genes (1,387 out of 1,475, represented in the multilayer network) belong in a range of modularity resolution (see methods: "multilayer community structure analysis" and Figure S3). We conceived this particular analysis as a proof of concept for the multilayer community structure analysis.

As shown in Figure 3, there are plain differences in the trajectories of the communities that are visited by each gene. Interestingly, the trajectories of seven genes, whose recurrent genetic alterations are well-known hallmark features of the four molecular disease subgroups (see introduction), branch off from well-separated communities, with the exception of *SRC* and *CTNNB1*, which are physical interactors (IntAct: EBI-15951997).

The landscape of these multilayer community trajectories can be further explored to investigate the so-called operations on dynamic communities (Cazabet et al. 2017), such as birth (a new community appears), death (a community vanishes), and resurgence (a community disappears and appears again later on). Along the explored range of modularity resolution, the 2,186 unique multilayer communities of the text-mined MB genes experience a total of 2,517 death events and 673 resurgence events (Figure S4), indicating not only a high level of instability (all communities disappear at least once) but also a high level of commutability (some communities reappear several times with the same exact composition). These observations led us to realize that each gene is characterized by its own journey throughout the communities found at different levels of resolution. For this reason, we further tested the hypothesis that tracing such trajectories for a set

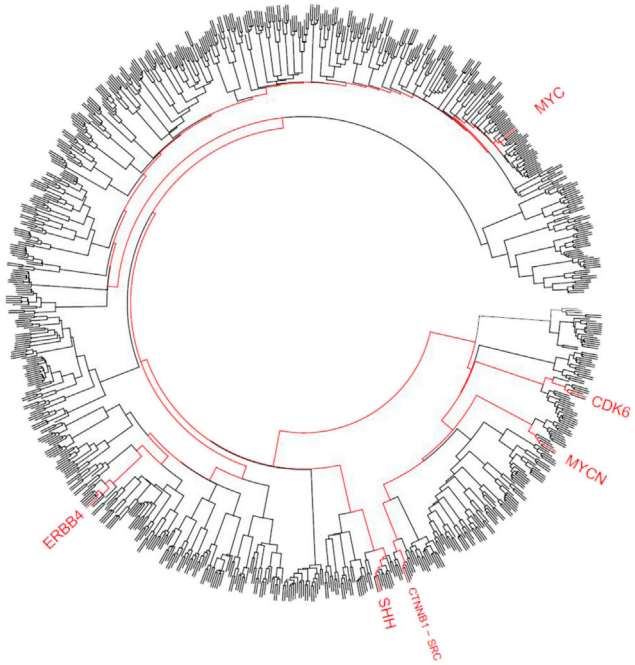


Figure 3. Dendrogram of multilayer community trajectories

The dendrogram represents the Hamming distance among the trajectories of the communities visited by each gene associated to medulloblastoma by text mining in a range of modularity resolution (see methods: "multilayer community structure analysis"). Trajectories of seven genes that are known to characterize medulloblastoma subgroups (see introduction) are highlighted in red. See also Figure S4.

of disease-related genes could be exploited for patient clustering purposes (see methods: "identification of the minimal set of genes that define medulloblastoma subgroups").

Medulloblastoma patient stratification through multilayer structure analysis

We sought to use the trajectories of the multilayer communities visited by the genes altered in MB to achieve patient stratification. Our reference (ground truth) consists of the four classical subgroups (WNT, SHH, G3, G4), which represent a standard categorization of MB despite substantial heterogeneity and the possibility of a more granular stratification have been reported (see introduction). The four subgroups have been recently investigated via network fusion using a cohort of patients with proteogenomic information (Forget et al., 2018). We reanalyzed this cohort to optimally recapitulate the four subgroups, while aiming to reduce the number of critical genes required for this stratification.

We retrieved lists of genes altered in 35 patients who display complete datasets (DNA methylation, RNA sequencing, proteomics, and phosphoproteomics) (see methods: "data sources of medulloblastoma genes"). Partial datasets are available for three additional patients (MB10, MB21, and MB33) that we retained as a validation set (see results: "sensitivity analyses"). We performed a hierarchical clustering based on the multilayer community trajectories of an optimal selection of minimal sets of genes. Optimality means that the features of these selected genes, in terms of their representation in the multilayer communities (parameter λ) and the similarity of their trajectories (parameter θ), allow clustering patients with the maximum accuracy and Matthews correlation coefficient (MCC) to the four subgroups of reference (see methods: "identification of the minimal set of genes that define medulloblastoma subgroups").

We achieved the highest accuracy (94.94%) and MCC (87%) with five clusters (WNT, SHH, G4, G3, and G3-G4), by selecting for each patient those genes that are represented in the communities in sets of at most 6 ($\lambda = 6$) and that are always part of the same communities along their trajectories ($\theta = 0$) (Figures 4 and 5, Tables S1–S3). 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

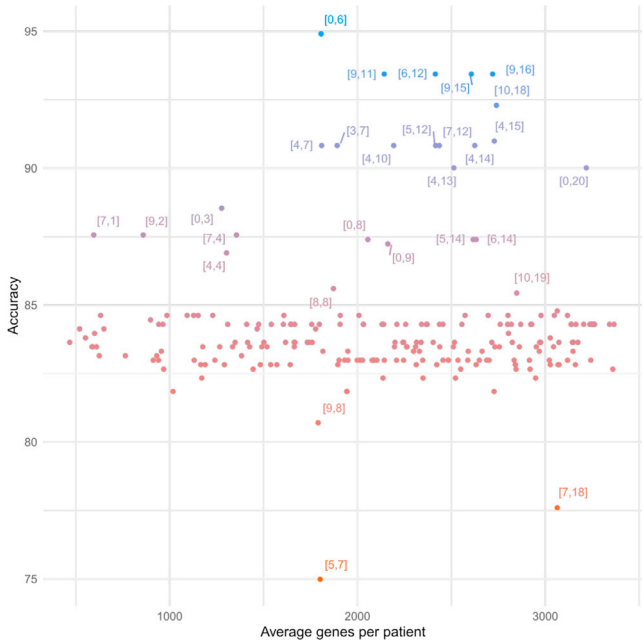


Figure 4. Parameters optimization

Scatterplot comparing the average genes per patient obtained by each iteration of the optimization procedure (see methods: "identification of the minimal set of genes that define medulloblastoma subgroups") and its corresponding accuracy. Values next to each point highlight the corresponding [θ,λ] combination. See also Figures S5–S7 and Tables S1–S3.

segregation. This observation implies that the selected genes are tightly associated and never leave the communities they belong to along their trajectories. An important aspect of this result is that, despite our reference being of four subgroups, we identified five clusters, indicating that only few patients escape the classical categorization and subtler stratas may exist, as suggested in recent studies (Schwalbe et al., 2017; Archer et al., 2018).

Classification of patients with partial molecular information

As the datasets of three patients consist of partial molecular information (see methods: "data sources of medulloblastoma genes"), we excluded these samples from the parameter optimization procedure and used them as a validation set. The three patients belong to subgroups G4 (patient MB10) and WNT (patients MB21 and MB33) (Forget et al., 2018). We assigned each one of the three patients to the cluster of the most similar among the remaining 35 patients based on the Jaccard Index (J) parametrized by the optimal θ and λ (see methods: "identification of the minimal set of genes that define medulloblastoma subgroups"). Patient MB10 shows the highest similarity to patient MB22 (J = 0.263), who belongs to G4 subgroup likewise eight patients in the following ranking positions (Table S4). Patient MB21 shows the highest similarity to three patients of the WNT subgroup (MB31 J = 0.2653; MB34 J = 0.2631; MB30 J = 0.2601). Finally, patient MB33 shows high similarity to two patients of WNT subgroup (MB30 J = 0.2168; MB34 J = 0.2106). Of note, patient MB31 of the WNT subgroup is the fourth most similar patient to MB33 (J = 0.2080), MB16 of the G4 subgroup being the third (J = 0.2081). These results show that the parameters for gene selection optimized based on patients with complete molecular information allow classifying the patients who have only partial molecular information with high accuracy (all three patients are correctly classified).

Robustness analyses

The identified values of θ and λ, optimized on 35 patients, correspond to an average of 1,812.74 genes per patient (SD = 106.97) (i.e., an average dimensionality reduction of 87.56% (SD = 0.44) per patient) (Table S5). Moreover, some of these genes are uniquely found among all patients of distinct clusters (148 genes in G3

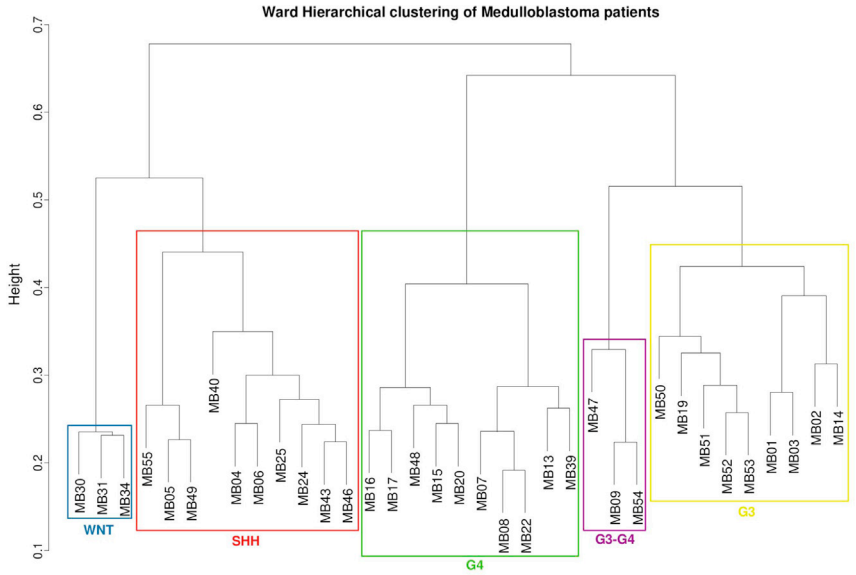


Figure 5. Clustering of medulloblastoma patients

Ward's linkage hierarchical clustering obtained at $\lambda = 6$ and $\theta = 0$. The rectangles indicate the five clusters suggested by PAM (partitioning around medoids) criteria. The color of each cluster indicates the original patient stratification into the four medulloblastoma subgroups (Forget et al., 2018): WNT (blue), SHH (red), Group 4 (G4, green), Group 3 (G3, yellow). A fifth cluster is depicted in purple, including three patients originally assigned to subgroups G3 (MB47) and G4 (MB09 and MB54). See also Figures S8 and S9 and Tables S4, S5, and S6.

patients; 83 genes in SHH patients; 115 genes in G4 patients; 46 genes in G3-G4 patients; 260 genes in WNT patients).

We evaluated the robustness of our results with two types of robustness analyses. In the first analysis, we shuffled the altered genes across the cohort 10,000 times, maintaining the same number of genes for each patient as in the original data. This procedure yielded an average accuracy of 54.76% (SD = 0.11) with $\theta = 0$ and $\lambda = 6$ (Figure S5). The distribution of the average optimization accuracies of the randomized sets shows dramatically lower values than those of the original data, indicating that our optimization procedure, when based on a meaningful clinical stratification, is able to identify non-random and very specific gene-subgroup associations (Figure S6).

In the second analysis, we recursively performed the optimization procedure after excluding the identified minimal set of genes at each iteration. We observed a progressive decrease in accuracy and, as expected, higher values of optimal θ and λ in later iterations, indicating less effective gene selection and dimensionality reduction (Figure S7). Overall, we observed that this decay in accuracy upon iterative removal of selected genes can be divided into three phases: a short initial phase (accuracies between 94.94 and 88.57) in which large sets of genes are removed at each iteration (1027.72 on average), a long intermediate phase (accuracies between 79.76 and 69.96) in which less genes are removed (23.31 on average), and a short final phase (between 57.06 and 31.43) in which an average of 1.08 gene is removed at each iteration before the accuracy drops to 0. At the end of this procedure, the cumulative number of removed genes is 5,950.63 (average per patient; 38 patients). These results show the effectiveness of the greedy nature of our optimization algorithm, which is able to achieve high accuracies even when the pool of genes it operates upon is largely reduced.

Sensitivity analyses

To test if our clusters are a good representation of the similarities among patients, we performed a sensitivity analysis with two approaches for clustering significance assessment. The first, based on multiscale bootstrap resampling (Suzuki and Shimodaira 2006), assigns a confidence value, known as approximately unbiased probability value (pvAU), to each cluster. High pvAU indicates high confidence in the clusters. The second, based on a Monte Carlo procedure (Kimes et al., 2017), assigns an empirical p value and a Gaussian

approximate p value to each cluster. An important difference between the two approaches is that the multi-scale bootstrap resampling approach tends to be less conservative than the Monte Carlo-based procedure, which outperformed the first with simulated and real-world data (Kimes et al., 2017).

At the root node, WNT and SHH subgroups are significantly separated from G3 and G4 subgroups with empirical p value of 1.08e-02 (Gaussian approximate p value of 2.59e-03) (Figures S8 and S9). Such two large partitions are poorly supported by the data (pvAU 49.23% and 63.05%, respectively), indicating the possibility of a finer subdivision. Indeed, WNT subgroup significantly separates from SSH subgroup with empirical p value of 5.45e-02 (Gaussian approximate p value of 3.55e-02), whereas G4 subgroup significantly separates from G3 subgroup with empirical p value of 1.85e-02 (Gaussian approximate p value of 6.75e-03).

Unlike the three main subgroups WNT (pvAU 100%), G4 (pvAU 99.97%), and G3 (pvAU 92.79%), SHH appears to be poorly supported by the data as a unique cluster (pvAU = 38.41%), whereas two SHH sub-clusters might exist (pvAU 99.88% and pvAU 99.55%, respectively), although their separation is not statistically significant (empirical p value 1.02e-01; Gaussian approximate p-value 9.74e-02). Of note, a finer partition of SHH subgroup into multiple sub-clusters has been reported by recent studies (Schwalbe et al., 2017; Archer et al., 2018).

The fifth cluster (G3-G4), despite being composed of two patients previously described as G4 (MB09 and MB54), and one as G3 (MB47), is supported by the data (pvAU 83.98%), but its separation from the G3 subgroup is not statistically significant (empirical p value 1.01e-01; Gaussian approximate p value 9.18e-02). Interestingly, patients of this cluster were all assigned to G4 via network fusion and to G3 only using methylation data (Forget et al., 2018). Indeed, an overlap of genetic features between G3 and G4 has also been reported by a study on risk stratification (Schwalbe et al., 2017).

Overall, these sensitivity analyses indicate that (1) 4 of 5 clusters found in our optimization procedure are statistically significant based on a Monte Carlo approach (Kimes et al., 2017) and recapitulate the classical MB molecular subgroups and (2) the small fifth cluster (G3-G4) shares similarities with G3 whose heterogeneity was previously observed (Schwalbe et al., 2017; Forget et al., 2018).

Provenance analysis of the identified gene communities

By performing a network enrichment analysis test (Signorelli et al. 2016), we identified the most significantly overrepresented intra-layer edges among the genes of the minimal sets identified for each patient in each cluster (see methods: ''multilayer network enrichment analysis''). In the following, we analyze those associations that are unique of the five clusters and enriched in all patients of each cluster (Table S6). Beside this strict requirement, several other enriched associations are shared among clusters and can be further explored (see resource availability: ''data and code availability''). Overall, we found that the minimal set of genes found in all patients of WNT, SHH, and G4 clusters are uniquely enriched in very specific associations in each layer, whereas G3-G4 and G3 clusters tend to display less specific enrichments (i.e., either several or no enriched associations). This reduction of enrichment specificity from WNT to G3 suggests an interesting parallel with the prognosis spectrum of the four classical subgroups, from best (WNT) to worst (G3) outcomes.

- **Molecular associations.** As for the molecular interaction layer, WNT cluster presents enrichment in four proteins: ACVR2A, a receptor involved in the activin signaling pathway (Chen et al., 2006), which is also enriched in this cluster in the pathways layer; ATP4A, a subunit of the ATPase H^+/K^+, a membrane transporter that is target of the Hedgehog signaling pathway, whose low levels of β1 subunit have been related to cell proliferation in MB models (Lee et al., 2015); POU2F2, which has been recently found to play a role in spinal cord development in a mouse model (Masgutova et al., 2019) and suspected to be regulated by miRNAs in MB (Venkataraman et al., 2013); and RBM48, a protein found to be amplified across several cancer tissues and cell lines and that may have a role in apoptotic processes (Hart et al., 2015). SHH cluster is uniquely enriched in molecular interactions of various gene products, including two proto-oncogenes (*ETS1* [Cao et al., 2015] and *JUND* [Elliott et al., 2019]), a calcium voltage-gated channel (CACNA1A) significantly downregulated in MB and other brain tumors (Phan et al., 2017), and interestingly a long noncoding RNA (*LINC00461*), expressed predominantly in the brain and involved in tumorigenesis (Yang et al., 2017). G4 cluster

only presents enrichment in the interactions of ARID4A, a member of the ARID family such as ARID1B, a repressor of Wnt/β-catenin signaling (Vasileiou et al., 2015). G3 cluster is enriched in interactions of the ABC transporter, ABCA3, suspected to be involved in chemoresistance in brain tumor progression (Hadjipanayis and Van Meir 2009); the dystrophin-glycoprotein SGCB; the SUMO ligase PIAS1, which increases the activity of Gli proteins on the Hedgehog pathway (Niewiadomski et al., 2019); and the heat shock protein DNAJB5, which regulates histone deacetylase (HDAC) nuclear shuttling, whose inhibition is considered to be a promising therapy in MB (Becher 2019).

- **Drug-target associations.** As for the drug layer, G3 cluster is the only one showing a unique enrichment in all patients, namely, in lubeluzole, an inhibitor of nitric oxide (NO) synthesis (Maiese et al. 1997). This observation points toward the role of oxidative stress in MB under the light of results from NO synthesis inhibition in experimental models (Haag et al., 2012) and clinical trials in G3 subgroup (Bakhshinyan et al., 2019).
- **Variant-disease associations.** As for the disease layer, the enriched associations may indicate overlapping features between MB and molecular processes underlying other pathologies. WNT cluster is uniquely associated with alveolar rhabdomyosarcoma, a common soft tissue sarcoma in children (Barr 2011), and familial prostate carcinoma. The implication of the overactivation of the Hedgehog signaling pathway in both MB and rhabdomyosarcoma (Azatyan et al., 2019) as well as in prostate cancer (Amakye et al. 2013; Ng and Curran 2011) has been extensively reported. SHH cluster is uniquely associated with macular degeneration and syndromic craniosynostosis, also characterized by ocular abnormalities, suggesting a link with the ophthalmic complications of MB, which occur as a result of the disease and its treatments (Cassidy et al., 2000). Polydactyly (Crane et al., 2018) appears to be uniquely enriched in G4 cluster, MB being a feature of several disorders of infants often characterized by akin skeletal abnormality such as Gorlin syndrome (Lo Muzio 2008) and others (Osterling et al., 2011). Cluster G3-G4 shows many enriched diseases, including several cancers, forms of hypogonadism, and interestingly Dravet syndrome, a genetic disorder that causes severe epilepsy in infants. Of note, MB is among the most frequent tumors of cerebellum presenting with seizures (5%) (Sánchez Fernández and Loddenkemper, 2017). G3 does not display unique enrichments in the disease layer.
- **Pathway associations.** As for the pathway layer, the WNT cluster is uniquely enriched in cell differentiation in early embryogenesis, such as Nodal (Brown et al., 2011) and Activin (Chen et al., 2006) signaling; immune response, such as Dendritic cell-associated C-type lectin-2 (Dectin-2) carbohydrates receptor activity (Graham and Brown 2009); protein metabolism, such as insulin-like growth factor (IGF) regulation (Holly and Perks 2006); and defects in the mismatch repair (MMR) system (Chao and Lipkin 2006). SHH cluster is uniquely enriched in potassium channels of the neuronal system, such as the Kir channel (Radeke et al. 1999), and signal transduction, such as calcitonin (Sexton et al. 1999) and Hedgehog (Briscoe and Thérond 2013) signaling. G4 is associated with fusion events in the *FGFR1* gene (Braun and Shannon 2004) and neuronal system transmission, such as excitatory synaptic transmission by glutamate receptors (Kessels and Malinow 2009). The great majority of these pathways have been directly or indirectly related to MB in the literature, such as the interplay between the embryonic morphogens Nodal and Hedgehog in brain development (Rohr et al., 2001), the activation of Activin signaling in a subset of G3 subgroup (Morabito et al., 2019), the role of *FGFR1* in gliomas (Egbivwie et al., 2019), and the importance of carbohydrate antigen recognition in MB (Read et al., 2009). Clusters G3-G4 and G3 show a varied landscape of enriched pathways.
- **Metabolic reaction associations.** As for the metabolome layer, uniquely enriched metabolites of the WNT cluster are ferricytochrome C (part of mitochondrial respiratory electron transport chain), nicotinamide nucleotide (a derivative of niacin, a form of vitamin B3), superoxide anion (a reactive oxygen species), and ribose 5-phosphate (a precursor to many biomolecules, including DNA and RNA). SHH shows unique enrichments for nicotinate D-ribonucleotide (part of cofactor biosynthesis) and pantothenate (vitamin B5), whereas G4 is uniquely enriched in sulfate (the major sulfur source in humans). Cluster G3 does not present uniquely enriched metabolites, whereas G3-G4 shows several.

Method verification on an independent cohort

To further verify the applicability of our methodology, we performed the same analytical procedure on an independent, non-overlapping, multi-omics MB cohort (Archer et al., 2018) (see methods: "data sources of medulloblastoma genes"). This cohort study collects proteogenomics data from 45 patients and proposes

a finer categorization of SHH and G3 subgroups (SHHa, SHHb, G3a, G3b). A total of 39 patients display complete multi-omics information, whereas 6 lack RNA sequencing, including all 3 patients of the WNT subgroup.

In a first analysis, we were able to recapitulate the 5 clusters (SHHa, SHHb, G3a, G3b, G4) of the 39 patients with complete multi-omics information, achieving the highest accuracy (98.29%, MCC = 0.95) with optimized parameters $\lambda = 3$ and $\theta = 0$ (Figure S10), which corresponds to an average of 842.2 genes per patient (SD = 145.12) and average dimensionality reduction of 92.83% (SD = 0.578). All patients are correctly assigned to their subgroups, whereas only MB136, labeled as a SHHb member, clusters with the SHHa subgroup.

As for the previous analysis, the patients with incomplete multi-omics information were used as validation set and assigned individually to subgroups based on the Jaccard Index (J) (see methods: "identification of the minimal set of genes that define medulloblastoma subgroups"). Patients MB037, MB018, and MB282 are correctly classified as SHHa, G3a, and G4, the most similar patients being MB239 (J = 0.177), MB226 (J = 0.136), and MB091 (J = 0.166), respectively.

In a second analysis, we included all 45 patients achieving the highest accuracy (95.56%, MCC = 0.85) with 7 clusters and $\lambda = 5$ and $\theta = 1$ (Figure S11), which corresponds to an average of 1,073.58 genes per patient (SD = 161.94) and average dimensionality reduction of 90.59% (SD = 1.06). The performance reduction suggests that the addition of patients presenting missing data in the parameters optimization procedure can decrease its performance.

DISCUSSION

Molecular disease subtyping is a fundamental tool to achieve an effective patient stratification for clinical trials and 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 represent a more challenging setting because, by definition, they affect a small number of patients with studies that, in most cases, are in the order of tens of subjects. MB, such as other pediatric cancers, is an illustrative example, two MB subgroups being very well distinguishable (SHH and WNT) and two others being far less characterized (G3 and G4).

In our vision, 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 multilayer networks. 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 characterize 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, molecular interactions). This way of accomplishing two objectives with one action constitutes the main achievement of our methodology.

In this regard, this work is additionally motivated by the relevance and urgency of implementing computational solutions based on biological multilayer networks. Borrowing from social network science, we use multiplexity as a way to evaluate intimacy of gene associations in MB: the more tightly a group of genes is connected through multiple types of features, the more clearly defined and explainable that community will be (Dickison et al. 2016).

Our results show that we can accurately recapitulate the four established MB subgroups using proteogenomic data and correctly classify the patients with partial molecular profiles. The approach enables an effective dimensionality reduction leading to the identification of a minimal set of altered genes that are sufficient to define MB subgroups. Moreover, the use of a multilayer network in this context allows the retrieval and analysis of the multiple associations among the identified genes, enabling a high level of interpretation of the patient subgroups and the spectrum of prognosis that characterize them, from best (WNT) to worst (G3) outcomes. Analyzing the provenance of the associations that determine the detected communities is extremely beneficial to better characterize the molecular determinants of the patient subgroups and, in turn, achieve a high level of explainability, a matter of considerable debate in computational biology lately (Adadi and Berrada 2020).

An additional important aspect that emerges from our results is that the precise clinical stratification of patients and the completeness of multi-omics information can lead to a better optimization and finer molecular characterization. Indeed, the overall performances of our optimization approach, in terms of both clustering accuracy and dimensionality reduction, are higher using a patient stratification of reference of six subgroups (Archer et al., 2018) compared with the traditional four subgroups (Forget et al., 2018). This indicates that precise clinical hypotheses can lead to precise molecular characterization of patient subgroups, making multilayer networks a powerful and unique tool especially for the study of rare diseases.

Limitations of the study

The main limitations of the study include (1) the scope of the multilayer network, (2) the reliability of the patient stratification of reference, and (3) the suitability of modularity as a quality function for community detection. As for the multilayer network, we distilled high-quality information from reputable and widely used knowledge bases (see methods: "data sources for the construction of the multilayer network"). Our multilayer network encapsulates a comprehensive view of fundamental aspects of human biology, but it can be further expanded to layers with a different content. As for the patient stratification of reference, the categorization of the cohort under study, based on network fusion (Forget et al., 2018), is one of the most recent and highly accurate attempts to cluster patients with MB using multi-omics information. Our analysis can be repurposed for different MB cohorts, available at data sharing platforms such as R2 (http://r2.amc.nl) and Cavatica (www.cavatica.org), among others. As for modularity, it is one of the most well-known quality functions for community detection (Chen et al., 2018). Moreover, the Louvain algorithm has been adapted for multilayer networks (Didier et al. 2018; Didier et al. 2015). Nevertheless, our approach can be applied to other quality functions (e.g., Hamiltonians, partition density) and more recent algorithms, such as the Leiden algorithm (Traag et al. 2019), which, to our knowledge, has currently not been adapted to multilayer networks.

Resource availability

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Davide Cirillo (davide.cirillo@bsc.es).

Materials availability

This study did not generate reagents, cell lines, or any biological material.

Data and code availability

The data and code generated during this study is available at dedicated GitHub repositories. The developed CmmD package is available at https://github.com/ikernunezca/CmmD. The code to reproduce all the figures and tables is available at https://github.com/ikernunezca/Medulloblastoma, where the complete lists of network enrichments and the processing of MB gene lists from the cohorts under study are also available. The text mining process is automated in the workflow available at https://github.com/cirillodavide/ipc_textmining. The procedure to generate the multilayer network used in this work is available at https://github.com/cirillodavide/gene_multilayer_network.

METHODS

All methods can be found in the accompanying transparent methods supplemental file.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2021.102365.

ACKNOWLEDGMENTS

This work was supported by the Agence Nationale de la Recherche in the program Investissements d'Avenir (project No. ANR-19-P3IA-0001; PRAIRIE 3IA Institute), the European Commission's Horizon 2020 Program, H2020-SC1-DTH-2018-1, "iPC - individualizedPaediatricCure" (ref. 826121), and the Ministry of Science and Higher Education of the Russian Federation (Project No. 14.Y26.31.0022).

The authors would like to thank Anaïs Baudot and Léo Pio-Lopez (Marseille Medical Genetics, Inserm) for advising about multilayer community structure analysis, María Rodríguez Martínez and Matteo Manica (IBM Research, Zurich) for text mining support, and François Serra and Miguel Ponce de León (Barcelona Supercomputing Center) for the insightful discussions.

AUTHOR CONTRIBUTIONS

A.V. and D.C. conceived the study; I.N.C. designed and implemented the computational analyses in consultation with D.C.; M.P. and A.Z. processed the medulloblastoma proteogenomic data. A.V. supervised the project. All the authors contributed to the writing of the article.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: October 13, 2020
Revised: March 17, 2021
Accepted: March 24, 2021
Published: April 23, 2021

REFERENCES

Adadi, A., and Berrada, M. (2020). Explainable AI for healthcare: from black box to interpretable models. In Embedded Systems and Artificial Intelligence, *1076*, V. Bhateja, S. Satapathy, and H. Satori, eds (Singapore: Springer), pp. 327–337.

Aleta, A., and Moreno, Y. (2019). Multilayer networks in a nutshell. Annu. Rev. Condens. Matter Phys. https://doi.org/10.1146/annurev-conmatphys-031218-013259.

Amakye, D., Jagani, Z., and Dorsch, M. (2013). Unraveling the therapeutic potential of the hedgehog pathway in cancer. Nat. Med. *19*, 1410–1422.

Archer, T.C., Ehrenberger, T., Mundt, F., Gold, M.P., Krug, K., Mah, C.K., Mahoney, E.L., Daniel, C.J., LeNail, A., Ramamoorthy, D., et al. (2018). Proteomics, post-translational modifications, and integrative analyses reveal molecular heterogeneity within medulloblastoma subgroups. Cancer Cell *34*, 396–410.e8.

Arenas, A., Fernández, A., and Gómez, S. (2008). Analysis of the structure of complex networks at different resolution levels. New J. Phys. https://doi.org/10.1088/1367-2630/10/5/053039.

Azatyan, A., Gallo-Oller, G., Diao, Y., Selivanova, G., Johnsen, J.I., and Zaphiropoulos, P.G. (2019). RITA downregulates hedgehog-GLI in medulloblastoma and rhabdomyosarcoma via JNK-dependent but p53-independent mechanism. Cancer Lett. *442*, 341–350.

Bakhshinyan, D., Adile, A., Venugopal, C., Singh, M., Qazi, M., Kameda-Smith, M., and Singh, S. (2019). MEDU-25. genes preserving stem cell state in group 3 MB BTICs contribute to therapy evasion and relapse. Neuro-Oncology *21*, ii108.

Barr, F.G. (2011). Soft tissue tumors: alveolar rhabdomyosarcoma. Atlas Genet. Cytogenet. Oncol. Haematol. *12*, https://doi.org/10.4267/2042/44650.

Becher, O.J. (2019). HDAC inhibitors to the rescue in sonic hedgehog medulloblastoma. Neuro Oncol. https://doi.org/10.1093/neuonc/noz115.

Blondel, V.D., Guillaume, J.L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. J. Stat. Mech. Theor. Exp. https://doi.org/10.1088/1742-5468/2008/10/p10008.

Braun, B.S., and Shannon, K. (2004). The sum is greater than the FGFR1 partner. Cancer Cell *5*, 203–204.

Briscoe, J., and Thérond, P.P. (2013). The mechanisms of hedgehog signalling and its roles in development and disease. Nat. Rev. *14*, 416–429.

Brown, S., Teo, A., Pauklin, S., Hannan, N., Cho, C.H.-H., Lim, B., Vardy, L., Dunn, N.R., Trotter, M., Pedersen, R., et al. (2011). Activin/nodal signaling controls divergent transcriptional networks in human embryonic stem cells and in endoderm progenitors. Stem Cells *29*, 1176–1185.

Cao, P., Fan, F., Dong, G., Yu, C., Feng, S., Song, E., Shi, G., Liang, Y., and Liang, G. (2015). Estrogen receptor α enhances the transcriptional activity of ETS-1 and promotes the proliferation, migration and invasion of neuroblastoma cell in a ligand dependent manner. BMC Cancer *15*, 491.

Cassidy, L., Stirling, R., May, K., Picton, S., and Doran, R. (2000). Ophthalmic complications of childhood medulloblastoma. Med. Pediatr. Oncol. *34*, 43–47.

Cazabet, R., Rossetti, G., and Amblard, F. (2017). Dynamic community detection. In Encyclopedia of Social Network Analysis and Mining, *2*, R. Alhajj and J. Rokne, eds. (Springer), pp. 1–10.

Chao, E.C., and Lipkin, S.M. (2006). Molecular models for the tissue specificity of DNA mismatch repair-deficient carcinogenesis. Nucleic Acids Res. *34*, 840–852.

Chen, S., Wang, Z.Z., Bao, M.H., Tang, L., Zhou, J., Xiang, J., Li, J.M., and Yi, C.H. (2018). Adaptive multi-resolution modularity for detecting communities in networks. Physica A Stat. Mech. Appl. *491*, 591–603.

Chen, Y.G., Wang, Q., Lin, S.L., Donald Chang, C., Chuang, J., and Ying, S.Y. (2006). Activin signaling and its role in regulation of cell proliferation, apoptosis, and carcinogenesis. Exp. Biol. Med. *231*, 534–544.

Clifford, S.C., Lusher, M.E., Lindsey, J.C., Langdon, J.A., Gilbertson, R.J., Straughton, D., and Ellison, D.W. (2006). Wnt/Wingless Pathway Activation and Chromosome 6 Loss Characterize a Distinct Molecular Sub-Group of Medulloblastomas Associated with a Favorable Prognosis. Cell Cycle *5*, 2666–2670.

Crane, J., Chang, V., Lee, H., Yong, W., Salamon, N., Kianmahd, J., Dorrani, N., Martinez-Agosto, J., and Davidson, T. (2018). PATH-23. germline gnas mutation in an 18-month-old with medulloblastoma. Neuro Oncol. *20*, vi163.

Dickison, M.E., Magnani, M., and Rossi, L. (2016). Multilayer Social Networks (Cambridge University Press).

Didier, G., Brun, C., and Baudot, A. (2015). Identifying communities from multiplex biological networks. PeerJ *3*, e1525.

Didier, G., Valdeolivas, A., and Baudot, A. (2018). Identifying communities from multiplex biological networks by randomized optimization of modularity. F1000Res. *7*, 1042.

Egbiwie, N., Cockle, J.V., Humphries, M., Ismail, A., Esteves, F., Taylor, C., Karakoula, K., Morton, R., Warr, T., Short, S.C., and Brüning-Richardson, A. (2019). FGFR1 expression and role in migration in low and high grade pediatric gliomas. Front. Oncol. *9*, 103.

Elliott, B., Millena, A.C., Matyunina, L., Zhang, M., Zou, J., Wang, G., Zhang, Q., Bowen, N., Eaton, V., Webb, G., et al. (2019). Essential role of JunD in cell proliferation is mediated via MYC signaling in prostate cancer cells. Cancer Lett. *448*, 155–167.

Forget, A., Martignetti, L., Puget, S., Calzone, L., Brabetz, S., Picard, D., Montagud, A., Liva, S., Sta, A., Dingli, F., et al. (2018). Aberrant ERBB4-SRC signaling as a hallmark of group 4 medulloblastoma revealed by integrative phosphoproteomic profiling. Cancer Cell *34*, 379–395.e7.

Fortunato, S., and Barthélemy, M. (2007). Resolution limit in community detection. Proc. Natl. Acad. Sci. U S A *104*, 36–41.

Graham, L.M., and Brown, G.D. (2009). The dectin-2 family of C-type lectins in immunity and homeostasis. Cytokine *48*, 148–155.

Haag, D., Zipper, P., Westrich, V., Karra, D., Pfleger, K., Toedt, G., Blond, F., Delhomme, N., Hahn, M., Reifenberger, J., et al. (2012). Nos2 inactivation promotes the development of medulloblastoma in Ptch1(+/-) mice by deregulation of gap43-dependent granule cell precursor migration. PLoS Genet. *8*, e1002572.

Hadjipanayis, C.G., and Van Meir, E.G. (2009). Brain cancer propagating cells: biology, genetics and targeted therapies. Trends Mol. Med. *15*, 519–530.

Halu, A., De Domenico, M., Arenas, A., and Sharma, A. (2019). The multiplex network of human diseases. NPJ Syst Biol Appl *5*, 15.

Hart, T., Chandrashekhar, M., Aregger, M., Steinhart, Z., Brown, K.R., MacLeod, G., Mis, M., Zimmermann, M., Fradet-Turcotte, A., Sun, S., et al. (2015). High-resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities. Cell *163*, 1515–1526.

Holly, J., and Perks, C. (2006). The role of insulin-like Growth factor binding proteins. Neuroendocrinology *83*, 154–160.

Jones, D.T.W., Jäger, N., Kool, M., Zichner, T., Hutter, B., Sultan, M., Cho, Y.J., Pugh, T.J., Hovestadt, V., Stütz, A.M., et al. (2012). Dissecting the genomic complexity underlying medulloblastoma. Nature *488*, 100–105.

Kessels, H.W., and Malinow, R. (2009). Synaptic AMPA receptor plasticity and behavior. Neuron *61*, 340–350.

Kimes, P.K., Liu, Y., Hayes, D.N., and Marron, J.S. (2017). Statistical significance for hierarchical clustering. Biometrics *73*, 811–821.

Kivela, M., Arenas, A., Barthelemy, M., Gleeson, J.P., Moreno, Y., and Porter, M.A. (2014). Multilayer networks. J. Complex Netw. *2*, 203–271.

Kool, M., Jones, D.T.W., Jäger, N., Northcott, P.A., Pugh, T.J., Hovestadt, V., Piro, R.M., Esparza, L.A., Markant, S.L., Remke, M., et al. (2014). Genome sequencing of SHH medulloblastoma predicts genotype-related response to smoothened inhibition. Cancer Cell *25*, 393–405.

Kool, M., Korshunov, A., Remke, M., David, T., Jones, W., Schlanstein, M., Northcott, P.A., Cho, Y.J., Koster, J., Schouten-van Meeteren, A., van Vuurden, D., et al. (2012). Molecular subgroups of medulloblastoma: an international meta-analysis of transcriptome, genetic aberrations, and clinical data of WNT, SHH, group 3, and group 4 medulloblastomas. Acta Neuropathol. *123*, 473–484.

Kristensen, V.N., Lingjærde, O.C., Russnes, H.G., Vollan, H.K.M., Frigessi, A., and Børresen-Dale, A.L. (2014). Principles and methods of integrative genomic analyses in cancer. Nat. Rev. Cancer *14*, 299–313.

Lancichinetti, A., and Fortunato, S. (2011). Limits of modularity maximization in community detection. Phys. Rev. E *84*, 066122.

Lee, S.J., Litan, A., Li, Z., Graves, B., Lindsey, S., Barwe, S.P., and Langhans, S.A. (2015). Na,K-ATPase β1-subunit is a target of sonic hedgehog signaling and enhances medulloblastoma tumorigenicity. Mol. Cancer *14*, 159.

Lo Muzio, L. (2008). Nevoid basal cell carcinoma syndrome (Gorlin syndrome). Orphanet J. Rare Dis. *3*, 32.

Maiese, K., TenBroeke, M., and Kue, I. (1997). Neuroprotection of lubeluzole is mediated through the signal transduction pathways of nitric oxide. J. Neurochem. *68*, 710–714.

Masgutova, G., Harris, A., Jacob, B., Corcoran, L.M., and Clotman, F. (2019). Pou2f2 regulates the distribution of dorsal interneurons in the mouse developing spinal cord. Front. Mol. Neurosci. *12*, 263.

Morabito, M., Larcher, M., Cavalli, F.M., Foray, C., Antoine, F., Mirabal-Ortega, L., Andrianteranagna, M., Druillennec, S., Garancher, A., Masliah-Planchon, J., et al. (2019). An autocrine ActivinB mechanism drives TGFβ/activin signaling in group 3 medulloblastoma. EMBO Mol. Med. *11*, e9830.

Newman, M.E.J., and Girvan, M. (2004). "Finding and evaluating community structure in networks. Phys. Rev. E *69*, 026113.

Ng, J.M., and Curran, T. (2011). "The hedgehog's tale: developing strategies for targeting cancer. Nat. Rev. *11*, 493–501.

Niewiadomski, P., Niedziółka, S.M., Markiewicz, Ł., Uśpieński, T., Baran, B., and Chojnowska, K. (2019). Gli proteins: regulation in development and cancer. Cells *8*, https://doi.org/10.3390/cells8020147.

Northcott, P.A., Buchhalter, I., Morrissy, A.S., Hovestadt, V., Weischenfeldt, J., Ehrenberger, T., Gröbner, S., Segura-Wang, M., Zichner, T., Rudneva, V.A., et al. (2017). The whole-genome landscape of medulloblastoma subtypes. Nature *547*, 311–317.

Northcott, P.A., Korshunov, A., Witt, H., Hielscher, T., Eberhart, C.G., Mack, S., Bouffet, E., Clifford, S.C., Hawkins, C.E., French, P., et al. (2011). Medulloblastoma comprises four distinct molecular variants. J. Clin. Oncol. *29*, 1408–1414.

Northcott, P.A., Lee, C., Zichner, T., Stütz, A.M., Erkek, S., Kawauchi, D., David, J., Shih, H., Hovestadt, V., Zapatka, M., Sturm, D., et al. (2014). Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. Nature *511*, 428–434.

Osterling, W.L., Boyer, R.S., Hedlund, G.L., and Bale, J.F., Jr. (2011). MPPH syndrome: two new cases. Pediatr. Neurol. *44*, 370–373.

Phan, N.N., Wang, C.Y., Chen, C.F., Sun, Z., Lai, M.-D., and Lin, Y.-C. (2017). Voltage-gated calcium channels: novel targets for cancer therapy. Oncol. Lett. *14*, 2059–2074.

Porter, M.A., Onnela, J.P., and Mucha, P.J. (2009). Communities in networks. arXiv. http://arxiv.org/abs/0902.3788.

Radeke, C.M., Conti, L.R., and Vandenberg, C.A. (1999). Inward rectifier potassium channel Kir 2.3 is inhibited by internal sulfhydryl modification. Neuroreport *10*, 3277–3282.

Ramaswamy, V., Remke, M., Bouffet, E., Bailey, S., Steven, C.C., Doz, F., Kool, M., Dufour, C., Vassal V, Milde, T., et al. (2016). Risk stratification of childhood medulloblastoma in the molecular era: the current consensus. Acta Neuropathol. *131*, 821–831.

Read, T.-A., Fogarty, M.P., Markant, S.L., McLendon, R.E., Wei, Z., Ellison, D.W., Febbo, P.G., and Wechsler-Reya, R.J. (2009). Identification of CD15 as a marker for tumor-propagating cells in a mouse model of medulloblastoma. Cancer Cell *15*, 135–147.

Reichardt, J., and Bornholdt, S. (2006). Statistical mechanics of community detection. Phys. Rev. E *74*, 016110.

Robinson, G., Parker, M., Kranenburg, T.A., Lu, C., Chen, X., Ding, L., Phoenix, T.N., Hedlund, E., Wei, L., Zhu, X., et al. (2012). Novel mutations target distinct subgroups of medulloblastoma. Nature *488*, 43–48.

Rohr, K.B., Anukampa Barth, K., Varga, Z.M., and Wilson, S.W. (2001). The nodal pathway acts upstream of hedgehog signaling to specify ventral telencephalic identity. Neuron *29*, 341–351.

Sánchez Fernández, I., and Loddenkemper, T. (2017). Seizures caused by brain tumors in children. Seizure *44*, 98–107.

Schwalbe, E.C., Lindsey, J.C., Nakjang, S., Crosier, S., Smith, A.J., Hicks, D., Rafiee, G., Hill, R.M., Iliasova, A., Stone, T., et al. (2017). Novel molecular subgroups for clinical classification and outcome prediction in childhood medulloblastoma: a cohort study. Lancet Oncol. *18*, 958–971.

Sexton, P.M., Findlay, D.M., and Martin, T.J. (1999). Calcitonin. Curr. Med. Chem. *6*, 1067–1093.

Signorelli, M., Vinciotti, V., and Wit, E.C. (2016). NEAT: an efficient network enrichment analysis test. BMC Bioinformatics *17*, 352.

Suzuki, R., and Shimodaira, H. (2006). Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics *22*, 1540–1542.

Taylor, M.D., Northcott, P.A., Korshunov, A., Remke, M., Cho, Y.-J., Clifford, S.C., Eberhart, C.G., Parsons, D.W., Rutkowski, S., Gajjar, A., et al. (2012). Molecular subgroups of medulloblastoma: the current consensus. Acta Neuropathol. *123*, 465–472.

Traag, V.A., Waltman, L., and van Eck, N.J. (2019). From Louvain to leiden: guaranteeing well-connected communities. Sci. Rep. *9*, 5233.

Valdeolivas, A., Tichit, L., Navarro, C., Perrin, S., Odelin, G., Levy, N., Cau, P., Remy, E., and Baudot, A. (2019). Random walk with restart on multiplex and heterogeneous biological networks. Bioinformatics *35*, 497–505.

Vasileiou, G., Ekici, A.B., Uebe, S., Zweier, C., Hoyer, J., Engels, H., Behrens, J., Reis, A., and Hadjihannas, M.V. (2015). Chromatin-remodeling-factor ARID1B represses wnt/β-catenin signaling. Am. J. Hum. Genet. *97*, 445–456.

Venkataraman, S., Birks, D.K., Balakrishnan, I., Alimova, I., Harris, P.S., Patel, P.R., Handler, M.H., Dubuc, A., Taylor, M.D., Foreman, N.K., et al. (2013). MicroRNA 218 acts as a tumor suppressor by targeting multiple cancer phenotype-associated genes in medulloblastoma. J. Biol. Chem. *288*, 1918–1928.

Wei, C.-H., Allot, A., Leaman, R., and Lu, Z. (2019). PubTator central: automated concept annotation for biomedical full text articles. Nucleic Acids Res. *47*, W587–W593.

Yang, Y., Ren, M., Song, C., Li, D., Hussain Soomro, S., Xiong, Y., Zhang, H., and Fu, H. (2017). LINC00461, a long non-coding RNA, is important for the proliferation and migration of glioma cells. Oncotarget *8*, 84123–84139.

Yang, Z., Algesheimer, R., and Claudio, J.T. (2016). A comparative analysis of community detection algorithms on artificial networks. Sci. Rep. *6*, 1–18.

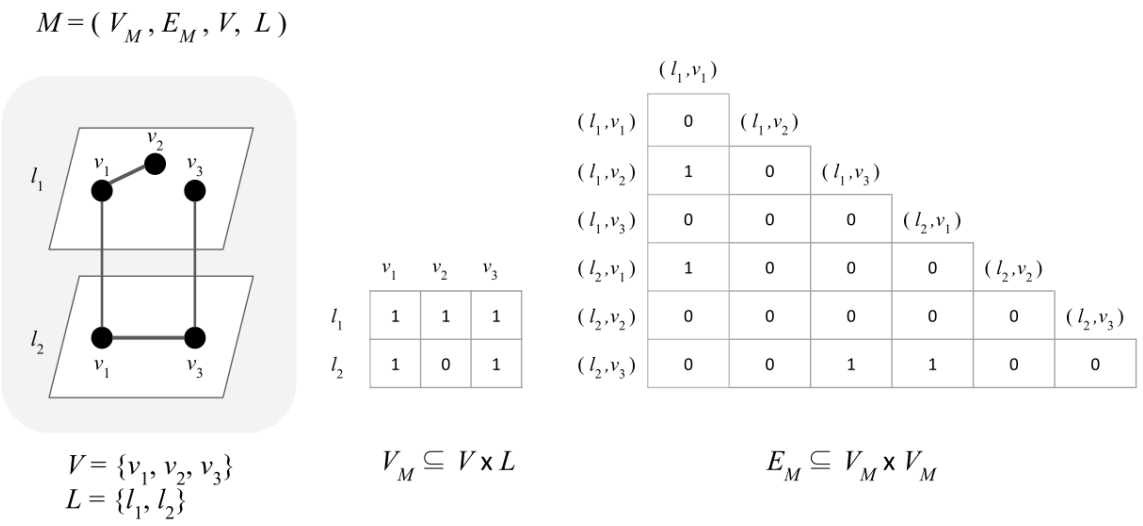
iScience, Volume *24*

Supplemental information

The multilayer community structure of medulloblastoma

Iker Núñez-Carpintero, Marianyela Petrizzelli, Andrei Zinovyev, Davide Cirillo, and Alfonso Valencia

Supplemental Figures



	v_1	v_2	v_3
l_1	1	1	1
l_2	1	0	1

	(l_1,v_1)	(l_1,v_2)	(l_1,v_3)	(l_2,v_1)	(l_2,v_2)	(l_2,v_3)
(l_1,v_1)	0					
(l_1,v_2)	1	0				
(l_1,v_3)	0	0	0			
(l_2,v_1)	1	0	0	0		
(l_2,v_2)	0	0	0	0	0	
(l_2,v_3)	0	0	1	1	0	0

Figure S1. Multilayer network definition, Related to Figure 1 and Figure 2. A multilayer network M, such as the one represented inside the grey area, is defined as a quadruplet of four elements (V_M, E_M, V, and L). V and L are the sets of nodes and layers of M, respectively. V_M and E_M are the sets of nodes contained in each layer and edges connecting them within (intra-layer) and between (inter-layer) layers, respectively. As the one represented here, we build a multilayer network where inter-layer edges only connect the same nodes in each layer.

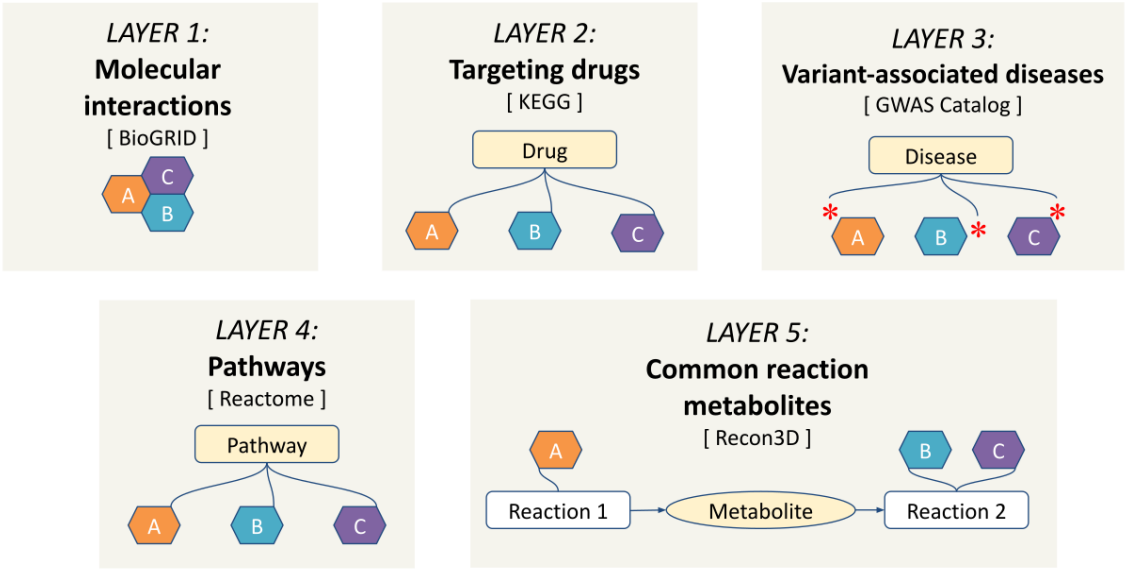


Figure S2. Gene-gene association represented in the fiver layers of the multilayer network, Related to Figure 1 and Figure 2. Gene entities are represented as hexagons. Associations retrieved from the databases in squared parentheses are represented as curved lines. Red asterisks indicate mutations.

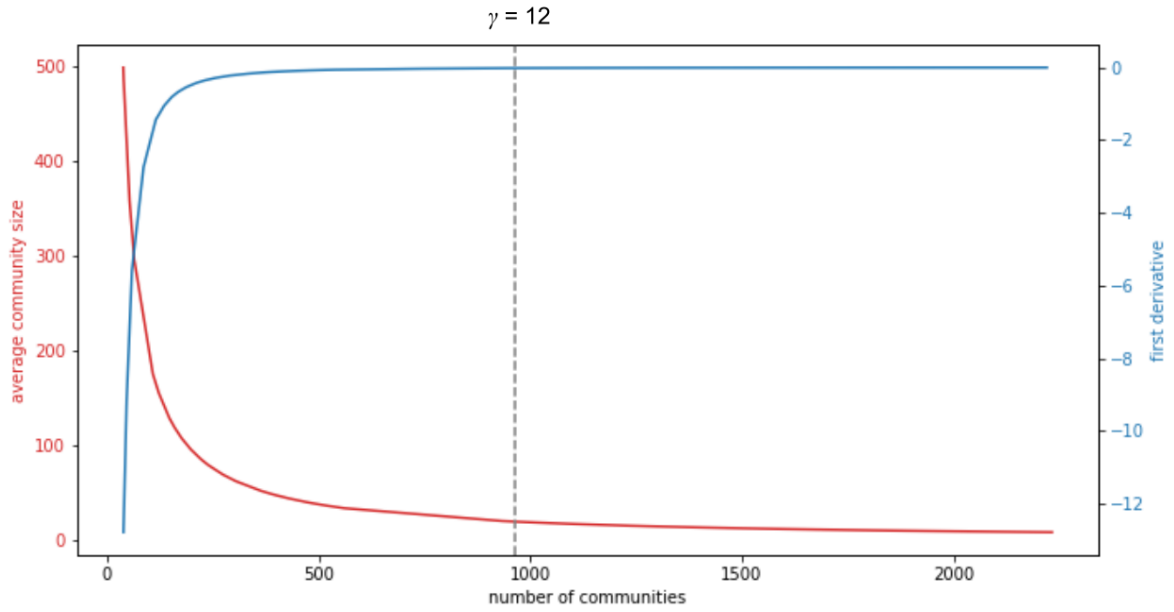


Figure S3. Identification of the resolution range of interest, Related to Figure 2. The modularity resolution parameter (γ) determines the number of communities and their size. The most dramatic changes in both size and number of communities occur in an initial range of resolution, which enables to detect 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).

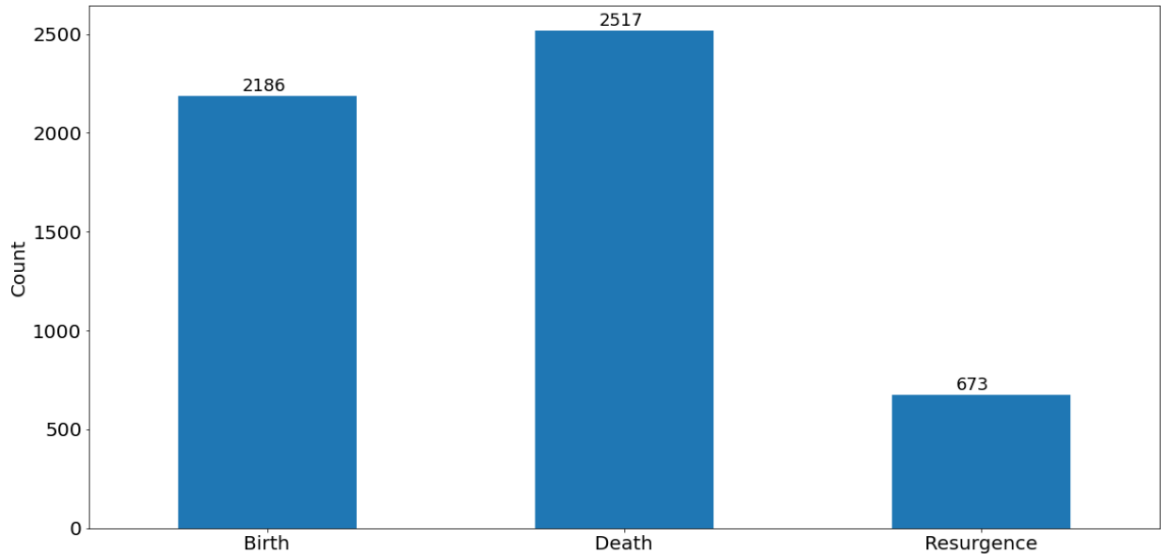


Figure S4. Operations on dynamic communities, Related to Figure 3. Count of dynamic events (birth, death, and resurgence) in the multilayer communities that contain text-mined medulloblastoma genes.

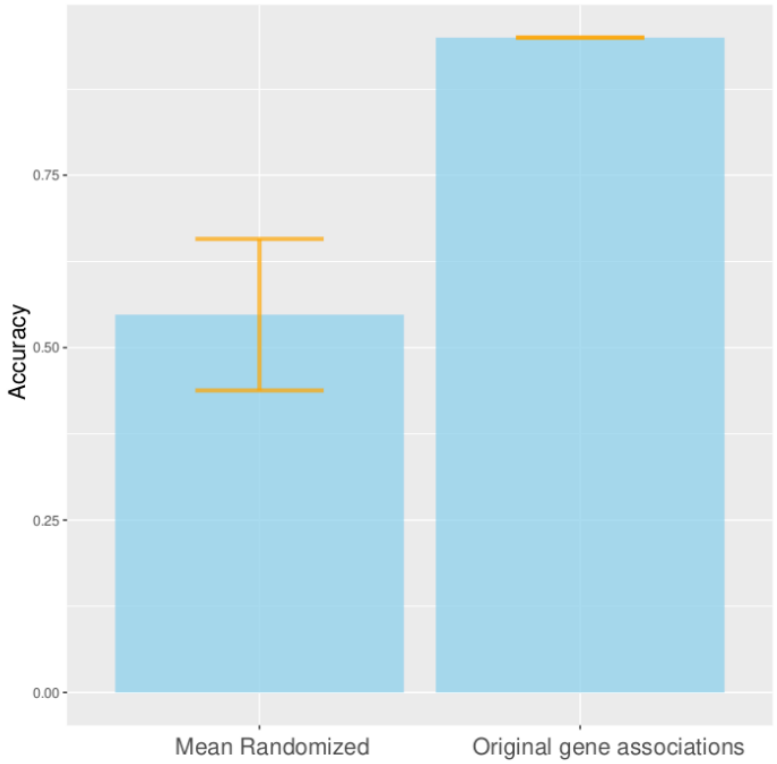


Figure S5. Gene shuffling test, Related to Figure 4 and Figure S6. The bar plots show the comparison between the highest accuracy achieved with the optimization procedure (94.94%, “Original gene associations”) and the average accuracy achieved by shuffling the genes in the cohort 10,000 times (54.76%, SD = 0.11, “Mean Randomized”), maintaining the same number of genes for each patient as in the original data and using the optimal parameters $\theta = 0$ and $\lambda = 6$.

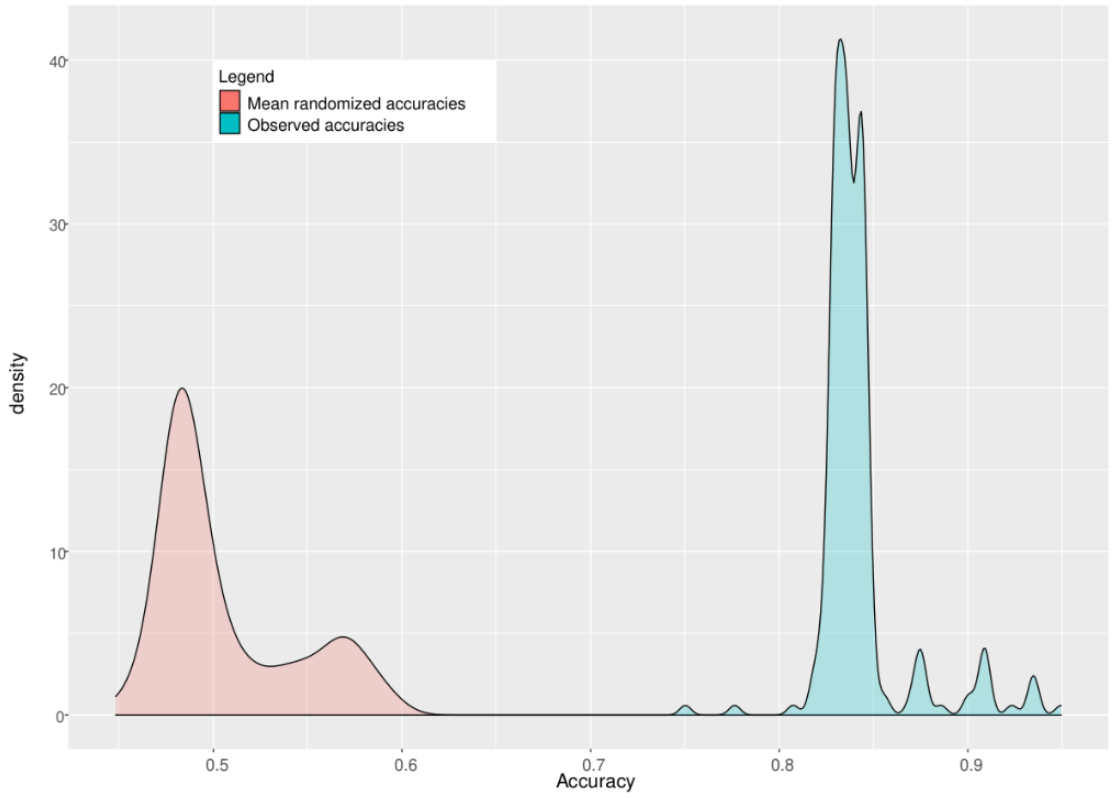


Figure S6. Distributions of optimization accuracies, Related to Figure 4 and Figure S4. The distribution of the optimization accuracies in the original data is reported in green, and the distribution of the average optimization accuracies after shuffling the altered genes across the cohort 10,000 times, maintaining the same number of genes for each patient is reported in red.

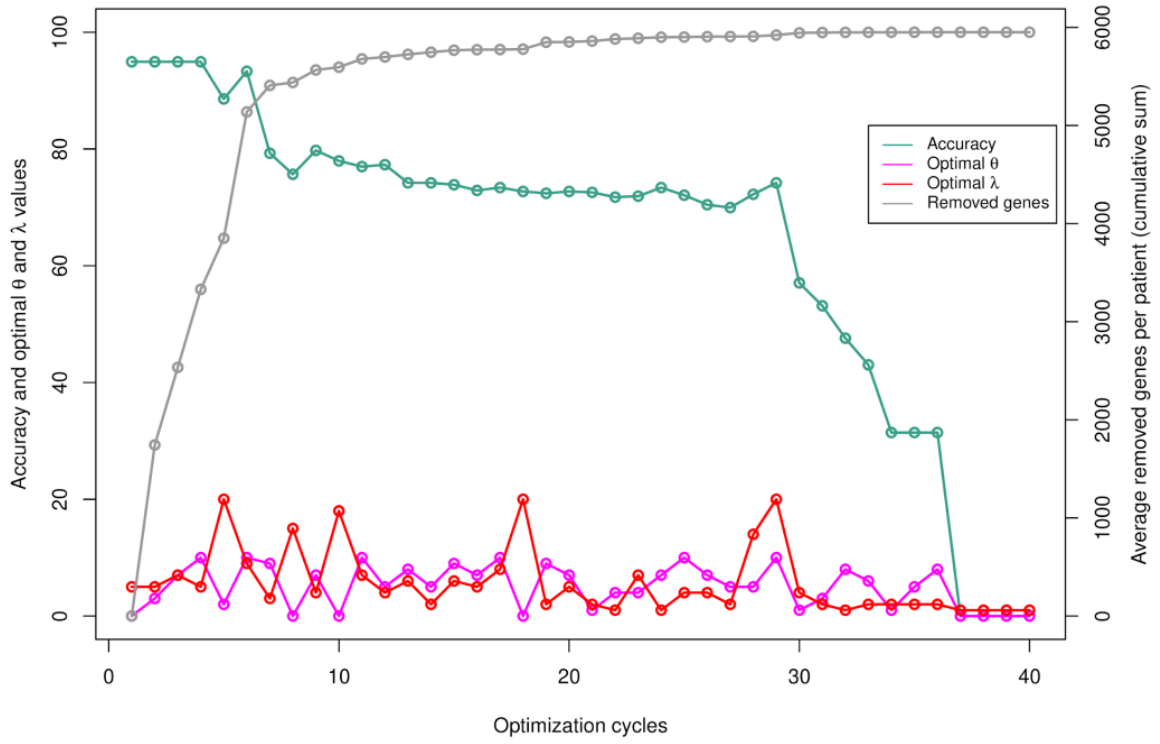


Figure S7. Recursive exclusion test, Related to Figure 4. The plot shows the iterative removal of selected genes in the cohort of 38 medulloblastoma patients. At every iteration, the minimal set of genes, found at optimal values of θ (purple line) and λ (red line) corresponding to highest accuracy (green line), is removed and the optimization procedure is repeated. The cumulative average number of genes per patient that are removed at every iteration is reported (grey line).

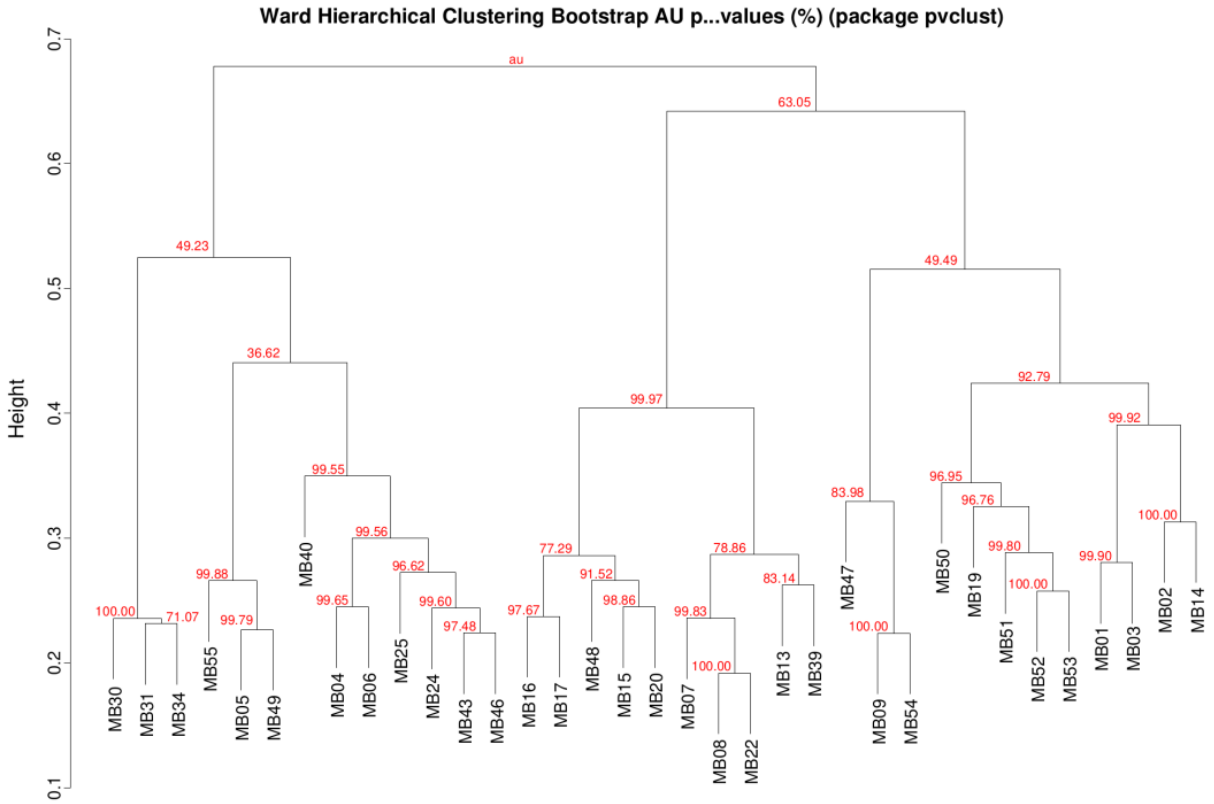


Figure S8. Clustering significance, Related to Figure 5. Significance assessment of hierarchical clustering (Ward method) of medulloblastoma patients using multiscale bootstrap resampling (Suzuki and Shimodaira, 2006). AU p values (%), or approximately unbiased probability value (pvAU), is reported in red on top of each cluster.

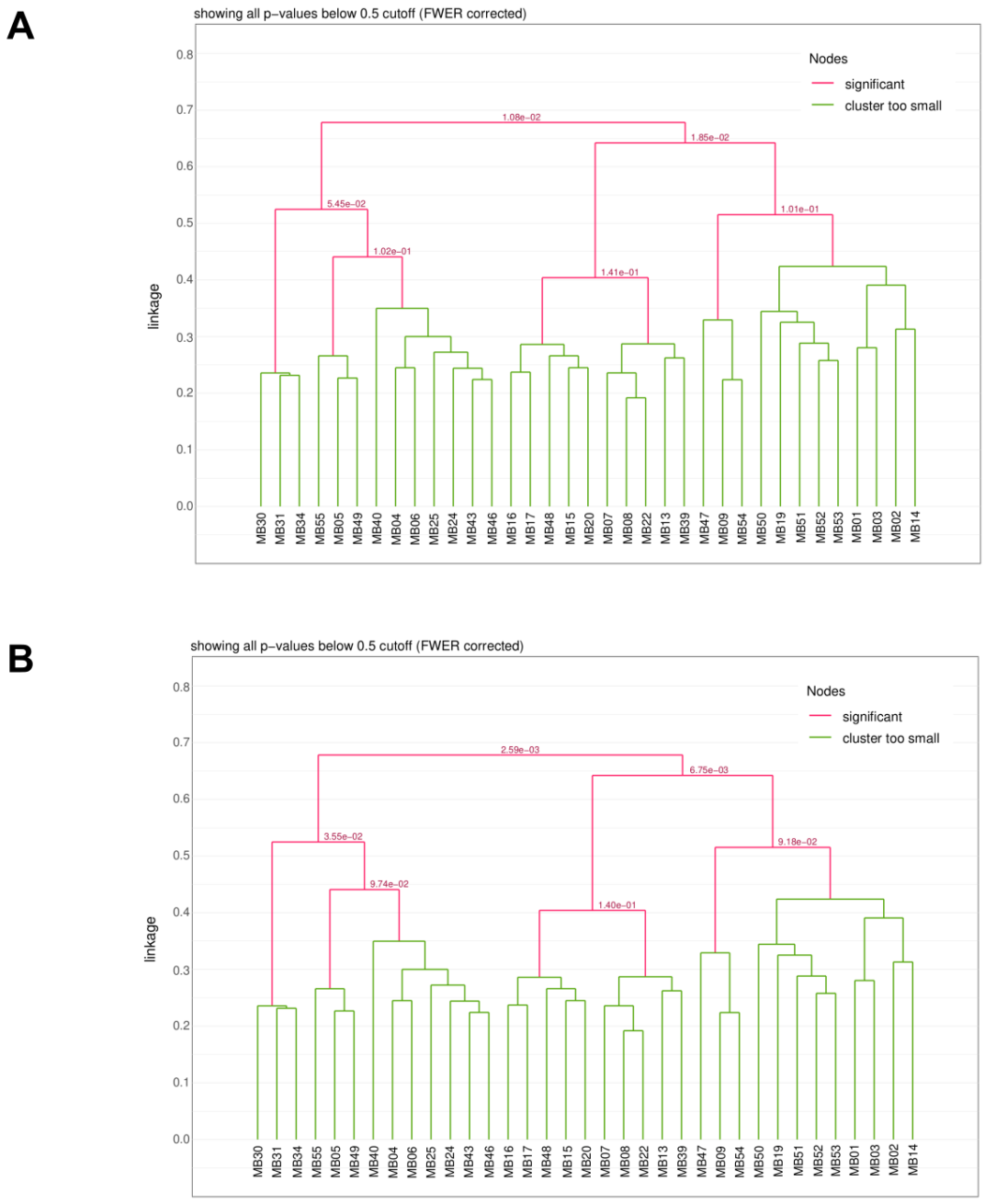


Figure S9. Clustering significance, Related to Figure 5. Significant assessment of hierarchical clustering (Ward method) of medulloblastoma patients using a Monte Carlo procedure (Kimes *et al.*, 2017). (A) empirical p-value and (B) Gaussian approximate p-value are reported in red on top of each cluster.

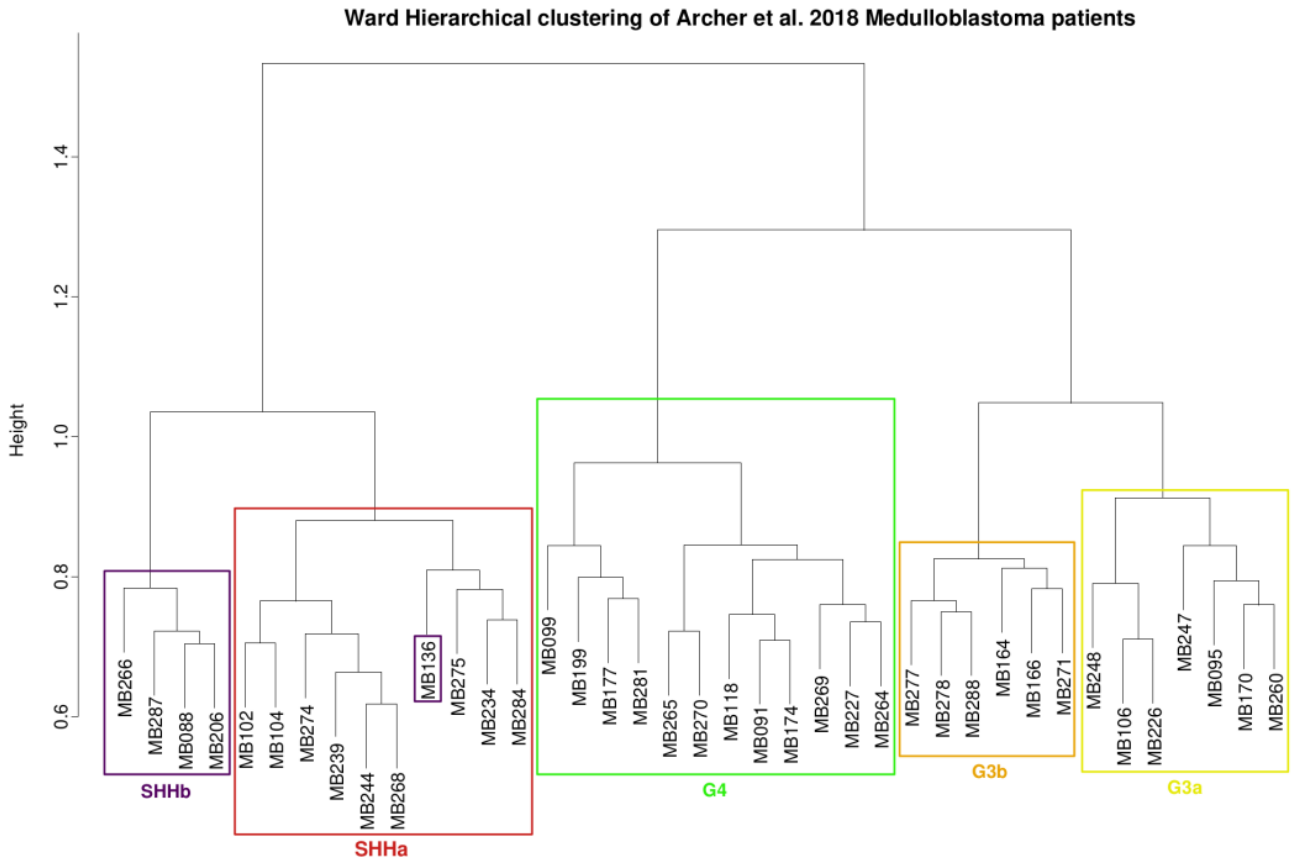


Figure S10. Hierarchical clustering of medulloblastoma patients from Archer et al. 2018, Related to Figure 2 and Figure S11. Ward's linkage hierarchical clustering obtained at λ = 3 and θ = 0 for patients with complete multi-omics data (Archer *et al.*, 2018). Rectangles indicate the 5 clusters suggested by PAM (partitioning around medoids) criteria. The color of each cluster indicates the original patient stratification into the five medulloblastoma subgroups: SHHa (red), SHHb (purple), Group 4 (G4, green), Group 3a (G3, yellow), Group 3b (G3b, orange). Patient MB136, originally labeled as SHHb subgroup and highlighted with a purple lower level rectangle, clusters within the SHHa subgroup.

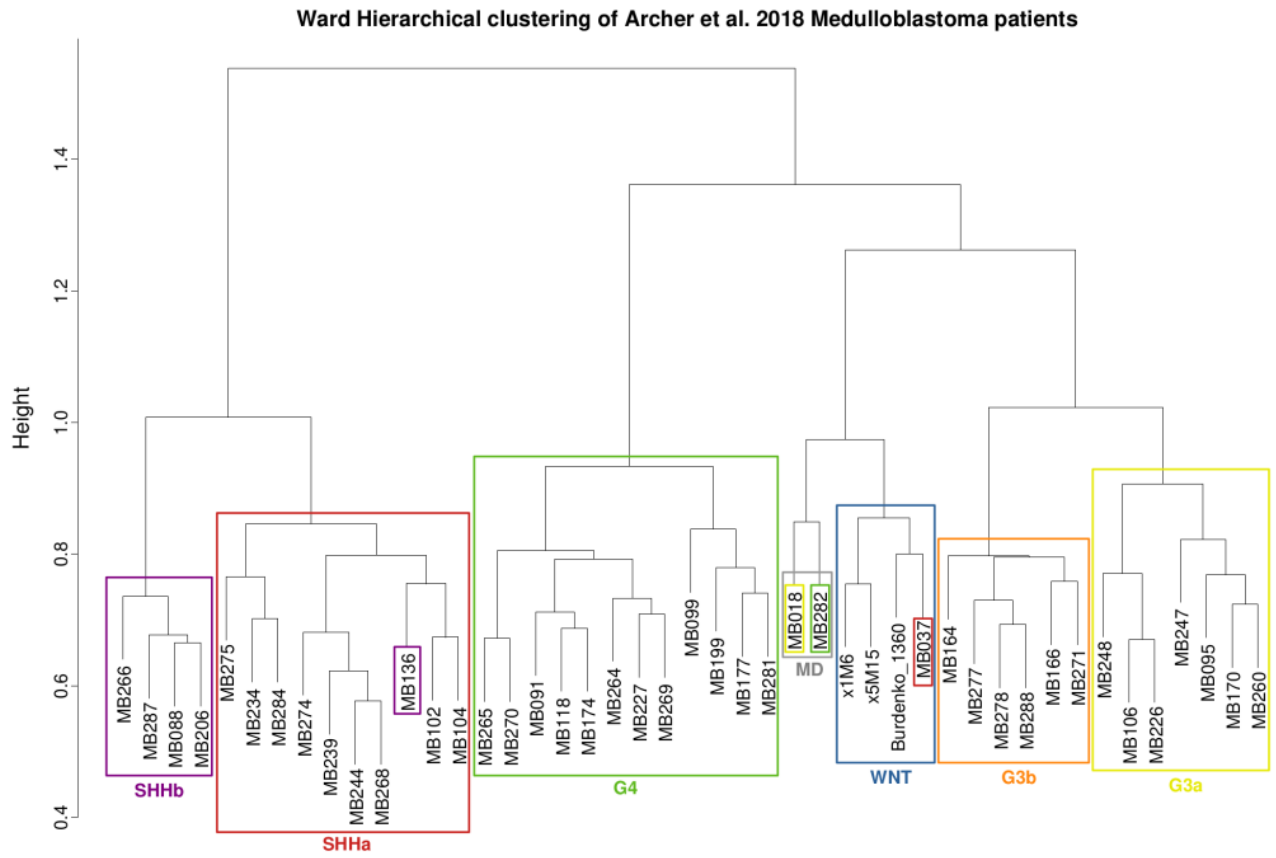


Figure S11. Hierarchical clustering of medulloblastoma patients from Archer et al. 2018, Related to Figure 2 and Figure S10. Ward's linkage hierarchical clustering obtained at λ = 5 and θ = 1 for patients with complete and incomplete multi-omics data (Archer *et al.*, 2018). Rectangles indicate the 7 clusters suggested by PAM (partitioning around medoids) criteria. The color of each cluster indicates the original patient stratification into the six medulloblastoma subgroups: WNT (blue), SHHa (red), SHHb (purple), Group 4 (G4, green), Group 3a (G3, yellow), Group 3b (G3b, orange). Patients with missing data cluster together (MD, Missing Data). Misclassified patients are highlighted with lower level rectangles indicating their original subgroup.

Supplemental Tables

Table S1. Optimal number of clusters, Related to Figure 4. The matrix shows the optimal number of clusters, based on the partitioning around medoids (PAM) algorithm, for combinations of parameters θ (rows) and λ (columns).

θ \ λ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0	9	10	7	9	9	5	10	6	8	8	10	8	10	9	8	8	9	8	9	6
1	8	10	10	10	9	8	10	8	8	9	8	8	8	8	8	8	9	9	10	8
2	8	8	8	10	8	9	8	10	9	10	8	10	9	9	8	8	8	9	8	10
3	10	10	9	5	9	9	5	10	10	9	10	8	8	10	8	10	10	8	8	8
4	10	10	10	4	8	9	5	10	10	5	9	9	5	5	5	10	8	8	8	8
5	10	8	9	7	8	8	5	10	10	10	9	5	8	6	10	10	8	9	8	8
6	8	9	10	7	9	8	8	10	10	5	9	4	9	6	10	9	9	9	8	8
7	7	9	7	7	7	7	9	8	10	8	9	5	10	10	5	7	9	5	8	8
8	8	8	8	8	8	8	9	4	8	8	8	8	10	10	10	10	9	9	8	8
9	8	7	8	8	8	8	9	4	8	8	4	5	8	10	4	4	9	9	10	10
10	8	9	8	8	9	9	9	8	8	9	10	10	9	9	9	10	9	5	8	10

Table S2. Optimization accuracies, Related to Figure 4. The matrix shows the accuracies of the optimization procedure (see Methods: “Identification of the minimal set of genes that define medulloblastoma subgroups”) for combinations of parameters θ (rows) and λ (columns). The maximum accuracy achieved is highlighted in bold.

θ \ λ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0	0.842	0.819	0.886	0.829	0.837	**0.949**	0.819	0.874	0.873	0.835	0.824	0.835	0.827	0.833	0.843	0.843	0.837	0.829	0.837	0.9
1	0.84	0.827	0.83	0.827	0.837	0.843	0.83	0.843	0.843	0.837	0.837	0.843	0.847	0.847	0.843	0.843	0.848	0.837	0.83	0.843
2	0.835	0.843	0.847	0.835	0.843	0.837	0.847	0.83	0.83	0.83	0.843	0.83	0.833	0.837	0.843	0.843	0.847	0.833	0.843	0.827
3	0.835	0.83	0.829	0.832	0.829	0.837	0.909	0.83	0.824	0.837	0.83	0.843	0.843	0.83	0.847	0.83	0.83	0.843	0.843	0.843
4	0.835	0.83	0.83	0.869	0.843	0.829	0.909	0.83	0.83	0.909	0.833	0.837	0.9	0.909	0.91	0.83	0.835	0.835	0.843	0.843
5	0.832	0.843	0.829	0.835	0.835	0.843	0.75	0.819	0.83	0.83	0.829	0.909	0.835	0.874	0.819	0.827	0.847	0.837	0.843	0.843
6	0.847	0.833	0.824	0.837	0.837	0.843	0.843	0.83	0.83	0.835	0.837	0.935	0.837	0.874	0.835	0.829	0.833	0.829	0.837	0.843
7	0.876	0.832	0.835	0.876	0.835	0.835	0.833	0.833	0.83	0.843	0.833	0.909	0.83	0.83	0.835	0.835	0.837	0.776	0.829	0.843
8	0.838	0.845	0.847	0.843	0.842	0.847	0.837	0.856	0.847	0.843	0.843	0.843	0.83	0.83	0.83	0.84	0.827	0.838	0.847	0.843
9	0.842	0.876	0.847	0.835	0.843	0.843	0.837	0.807	0.843	0.843	0.935	0.835	0.843	0.83	0.935	0.935	0.837	0.843	0.829	0.83
10	0.837	0.832	0.847	0.847	0.829	0.837	0.829	0.843	0.842	0.829	0.83	0.83	0.837	0.835	0.829	0.824	0.829	0.923	0.855	0.824

Table S3. Optimization MCC, Related to Figure 4. The matrix shows the Matthews Correlation Coefficient (MCC) of the optimization procedure (see Methods: "Identification of the minimal set of genes that define medulloblastoma subgroups") for combinations of parameters θ (rows) and λ (columns). The maximum MCC value achieved is highlighted in bold.

θ \ λ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0	0.603	0.536	0.719	0.564	0.590	**0.876**	0.536	0.682	0.685	0.582	0.554	0.582	0.563	0.581	0.608	0.608	0.590	0.564	0.590	0.754
1	0.596	0.563	0.572	0.563	0.590	0.608	0.572	0.608	0.608	0.590	0.590	0.608	0.617	0.617	0.608	0.608	0.624	0.590	0.572	0.608
2	0.582	0.608	0.617	0.586	0.608	0.590	0.617	0.572	0.572	0.572	0.608	0.572	0.578	0.590	0.608	0.608	0.617	0.581	0.608	0.563
3	0.589	0.572	0.564	0.577	0.564	0.590	0.771	0.572	0.554	0.590	0.572	0.608	0.608	0.572	0.617	0.572	0.572	0.608	0.608	0.608
4	0.589	0.572	0.572	0.678	0.608	0.564	0.771	0.572	0.572	0.771	0.578	0.590	0.751	0.771	0.776	0.572	0.582	0.582	0.608	0.608
5	0.577	0.608	0.564	0.575	0.582	0.608	0.382	0.536	0.572	0.572	0.564	0.771	0.582	0.682	0.536	0.563	0.617	0.590	0.608	0.608
6	0.617	0.581	0.554	0.579	0.590	0.608	0.608	0.572	0.572	0.582	0.590	0.841	0.590	0.682	0.589	0.564	0.581	0.564	0.590	0.608
7	0.694	0.577	0.575	0.694	0.575	0.575	0.581	0.578	0.572	0.608	0.581	0.771	0.572	0.576	0.582	0.575	0.590	0.421	0.564	0.608
8	0.595	0.612	0.617	0.608	0.603	0.617	0.590	0.643	0.617	0.608	0.608	0.608	0.572	0.572	0.572	0.603	0.563	0.595	0.617	0.608
9	0.603	0.694	0.617	0.582	0.608	0.608	0.590	0.521	0.608	0.608	0.841	0.582	0.608	0.572	0.841	0.841	0.590	0.611	0.567	0.572
10	0.587	0.577	0.617	0.617	0.564	0.590	0.564	0.608	0.603	0.564	0.572	0.572	0.590	0.586	0.564	0.554	0.564	0.810	0.642	0.554

Table S4. Classification of patients with partial datasets, Related to Figure 5.
The table reports the values of the Jaccard Index (J), parametrized by the optimal θ and λ, between the 3 patients with partial datasets (MB10, MB21, MB33) and the 35 patients with complete datasets (see Methods: “Data sources of medulloblastoma genes”).

	"MB10"	"MB21"	"MB33"
"MB01"	0.20855106888361	0.219810040705563	0.199903194578896
"MB02"	0.226933830382106	0.232198142414861	0.202247191011236
"MB03"	0.230385487528345	0.246086956521739	0.203089504770559
"MB04"	0.236396890717878	0.247598253275109	0.193577566711895
"MB05"	0.224057602710716	0.236489232019504	0.185480486781368
"MB06"	0.2255299954894	0.239740820734341	0.190839694656489
"MB07"	0.247404063205418	0.241379310344828	0.191169977924945
"MB08"	0.255896751223854	0.245812395309883	0.191443388072602
"MB09"	0.234858387799564	0.238391376451078	0.19559585492228
"MB10"	1	0.387596899224806	0.417813765182186
"MB13"	0.252108716026242	0.240667545015371	0.196420376319413
"MB14"	0.224178962398858	0.229138475417231	0.200670498084291
"MB15"	0.245346062052506	0.238565022421525	0.199434229137199
"MB16"	0.260057471264368	0.244523915958873	0.208097928436912
"MB17"	0.256641366223909	0.238726790450928	0.205223880597015
"MB19"	0.226726057906459	0.23021582733813	0.196706720071206
"MB20"	0.245344506517691	0.237636761487965	0.19560238204306
"MB21"	0.387596899224806	1	0.47027027027027
"MB22"	0.263229308005427	0.251486830926083	0.197016235190873
"MB24"	0.228245363766049	0.235772357723577	0.198476915754403
"MB25"	0.219874100719424	0.225834046193328	0.190647482014389
"MB30"	0.225081890500702	0.260118235561619	0.216873212583413
"MB31"	0.219325842696629	0.265342163355408	0.208029197080292
"MB33"	0.417813765182186	0.47027027027027	1
"MB34"	0.231185218566922	0.263134851138354	0.210621879255561
"MB39"	0.243792325056433	0.246463780540077	0.193433895297249
"MB40"	0.210699202252464	0.221179624664879	0.191943127962085
"MB43"	0.214088397790055	0.220472440944882	0.188539741219963
"MB46"	0.232285312060066	0.232372505543237	0.200093720712277
"MB47"	0.208278291501541	0.229626485568761	0.183826778612461
"MB48"	0.236533957845433	0.244097995545657	0.194895591647332
"MB49"	0.216193656093489	0.234702093397746	0.172910662824208
"MB50"	0.208942390369733	0.227743271221532	0.176949443016281
"MB51"	0.22202565236621	0.242437153813379	0.195921985815603
"MB52"	0.232150678931231	0.253062948880439	0.194782608695652
"MB53"	0.236637734125171	0.251093613298338	0.203644646924829
"MB54"	0.208281573498965	0.230861723446894	0.176980198019802
"MB55"	0.201853344077357	0.224043715846995	0.166733306677329

Table S5. Minimal set of genes, Related to Figure 5 (attached dataset). Minimal sets of altered genes associated with each one of the 38 medulloblastoma patients from (Forget *et al.*, 2018). The labels of the original subgroups (clusters) and the ones assigned after the optimization procedure are reported.

Table S6. Multilayer network enrichment analysis, Related to Figure 5 (attached dataset). The table reports those associations (edges) among the minimal sets of genes that are enriched in all the patients of a cluster and unique of each cluster (WNT, SHH, G3, G4, G3-G4) for a specific layer of the multilayer network (see Methods: “Multilayer network enrichment analysis”). Association IDs are grounded in databases (see Methods: “Data sources for the construction of the multilayer network”).

Transparent Methods

Multilayer network definition

A network (i.e. a graph or a *monoplex*) is defined as a tuple $G = (V, E)$, where V denotes the set of nodes (or vertices) in the network and $E \subseteq V \times V$ denotes the set of edges (or links) connecting them (Bollobás 1998). A graph composed of multiple networks, called layers, is referred to as a multilayer network. 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 $l \in L$, $V_M \subseteq V \times L$ 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 $(v, l), (v', l') \in V_M$ (Kivela et al. 2014) (**Figure S1**). In a multilayer network, an edge can be intra-layer, i.e. it connects nodes in the same layer ($l = l'$), or inter-layer, i.e. it connects nodes from different layers ($l \neq l'$). We built a multilayer network consisting of 5 layers and inter-layer edges imposed only between the same nodes, if any, on different layers.

Multilayer community detection

Communities in the multilayer network have been detected using MolTi software (Didier, Valdeolivas, and Baudot 2018; Didier, Brun, and Baudot 2015), which is available at https://github.com/gilles-didier/MolTi-DREAM. MolTi adapts the Louvain clustering algorithm with modularity maximization to multilayer networks. 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 a group 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. In MolTi, modularity of a multilayer network X is defined 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 (c_i = c_j) 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.

Data sources for the construction of the multilayer network

We created a multilayer network consisting of five layers in which nodes represent genes (Entrez identifiers), intra-layer edges represent different types of associations retrieved from publicly available knowledge bases and inter-layer edges exist between the same nodes in the different layers (**Figure S2**). All the data was downloaded on October 19, 2019, and it is available at https://github.com/cirillodavide/gene_multilayer_network.

Molecular associations. In this layer, two genes are connected if a physical or genetic association exists. Molecular associations between human genes were obtained from BioGRID, release 3.5.177. BioGRID (Oughtred et al. 2019) is a multi-species database of interactions, curated from high-throughput datasets and individual studies. Among other prominent primary databases, BioGRID shows the highest coverage for both interactions and proteins (Bajpai et al. 2019).

Drug-target associations. In this layer, two genes are connected if they are both targets of the same drug. Drug-target associations between human genes were obtained from KEGG BRITE “Target-based Classification of Compounds”, release br08310. KEGG BRITE (Kanehisa et al. 2019) is a manually curated database of functional hierarchies of various biological objects, such as Drug classifications. The Target-based Classification of Compounds consists of six categories

(Protein-coupled receptors, Nuclear receptors, Ion channels, Transportes, Enzymes, Others). One-to-one and unclassified gene-target associations were excluded.

Variant-disease associations. In this layer, two genes are connected if they are both reported to be associated with the same disease in genome-wide association studies (GWAS). Variant-disease associations between human genes were obtained from Monarch Disease Ontology (MonDO), released 2019-09-30. MonDO (Mungall et al. 2017) is a multi-species ontology generated by merging and harmonizing multiple disease resources (ORDO/Orphanet, DO, OMIM, MESH, etc.). In MonDO, gene-disease associations are inferred by integrating gene variants (SNPs, SNVs, QTLs, CNVs, among others) from significant GWAS hits. We retrieved MonDO entries with associated OMIM identifiers from the OWL file, filtering for evidence code ECO:0000220 (sequencing assay evidence) through the Monarch Solr search service.

Pathway associations. In this layer, two genes are connected if they are both annotated to the same pathway. Pathway associations between human genes were obtained from Reactome, release 70. Reactome (Fabregat et al. 2018) is a manually curated pathway database. Associations were retrieved from the lowest level pathway diagram of Reactome hierarchy. We found that all annotations are associated with IEA (inferred from electronic annotations) and TAS (traceable author statement) evidence codes.

Metabolic reaction associations. In this layer, two genes are connected if they are involved in metabolic reactions where product metabolites of one reaction are reactant metabolites of the other one. Metabolic reaction associations between human genes were obtained from Recon3D (Brunk et al. 2018) through BiGG Models (http://bigg.ucsd.edu), released 2019-09-12. Recon3D is the largest human metabolic network model. Superconnected metabolites (e.g. ATP, CO_2, H_2O) (Croes et al. 2006) were excluded.

Data sources of medulloblastoma genes

We aim to study the community structures of a multilayer network that contains medulloblastoma-associated genes. We selected genes for our study from two sources: (1) genes mentioned in scientific publications about medulloblastoma identified via text mining; (2) genes that are altered in medulloblastoma patients

based on two recent proteogenomic studies (Forget et al. 2018; Archer et al. 2018). The text mined data have been used as a proof-of-concept for the multilayer community structure analysis. The proteogenomic datasets have been used to identify the minimal sets of genes that characterize the medulloblastoma subgroups.

Text mined medulloblastoma genes. PubTator Central (PTC) (Wei et al. 2019) was used to retrieve gene mentions in abstracts of scientific publications indexed in PubMed with the MeSH term “medulloblastoma” (D008527) in February 2020 (see Resource Availability: “Data and Code Availability”).

Medulloblastoma genes from proteogenomic data. Subgroups of 38 medulloblastoma patients (WNT, SHH, G3, G4) were retrieved from (Forget et al. 2018). While 35 patients present DNA methylation, RNA sequencing, proteomic and phosphoproteomic profiles, 3 patients (MB10, MB21, MB33) present only partial molecular information (the three lack RNA sequencing) and were used for validation. Gene methylation levels were mapped from CpG sites using the biomaRt package in R. When multiple CpG sites fell on a gene position, the median value was considered; when it fell on a region that is not annotated, the nearest gene was considered. Based on these pre-processed datasets (Forget et al. 2018), lists of genes, henceforth called “altered genes”, were obtained by selecting the top 30% of the distribution of each data type. All the items of such lists were converted to Entrez identifiers, resulting in a total of 14039.6 altered genes per patient on average (see Resource Availability: “Data and Code Availability”).

Subgroups of 45 medulloblastoma patients (WNT, SHHa, SHHb, G3a, G3b and G4) were retrieved from (Archer et al. 2018). While 39 patients present DNA acetylation, RNA sequencing, proteomics and phosphoproteomics profiles, 6 patients lack RNA sequencing information, including all 3 patients of the WNT subgroup. When multiple DNA acetylation measurements were linked to the same gene, the median value was considered. Altered genes were obtained with the same criterion as previously described and gene symbols converted to Entrez identifiers, resulting in a total of 11608.6 (SD= 2264.524) altered genes per patient on average.

Multilayer community structure analysis

We analyzed how the multilayer community structure varies within a range of modularity resolution (γ) where the most dramatic changes in size and composition

of the-communities are observed before both reach a plateau. We identified the endpoint of 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 S3**). The endpoint was found at γ=12 (964 multilayer communities), indicating that $\gamma \in (0,12]$ is the range of interest for our study.

To compare the trajectories of each gene along the communities, we computed the pairwise Hamming distance (Hamming 1950) among the vectors of communities visited by each gene in the range $\gamma \in (0,12]$ with an interval of 0.5. 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 2**).

Identification of the minimal set of genes that define medulloblastoma subgroups

The biomedical goal of the study is to identify the minimal number of genes that recapitulate the four biomedically relevant medulloblastoma subgroups (WNT, SHH, G3, and G4) (Forget et al. 2018). Identifying a minimal set of genes is crucial for both the definition of diagnostic 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 purpose, small values of θ and λ guarantee a selection of genes with similar trajectories and in minimal numbers. For instance, with θ = 2 and λ = 4, patient similarity is computed using sets of at most four genes that did not belong to

the same communities at most twice along their trajectories. For each of these clustering analyses, we identified the optimal number of clusters using the partitioning around medoids (PAM) algorithm (Kaufman and Rousseeuw 1987) (**Table S1**).

Based on this approach, we formulated an optimization procedure to systematically evaluate values of θ and λ to identify the ones that maximize the accuracy of recapitualiting patient stratification into the four medulloblastoma subgroups (WNT, SHH, Group 3, and Group 4). We defined accuracy as

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

where true positives (TP) are patients of the same subgroup who are clustered together, true negatives (TN) are patients of different subgroups who are not clustered together, false positives (FP) are patients of different subgroups who are clustered together, and false negatives (FN) are patients of the same subgroup who are not clustered together. The same optimization procedure can also be formulated to maximize the Matthews Correlation Coefficient (MCC), which is defined as

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

In both cases, the optimal parameters found are θ = 0 and λ = 6, corresponding to an accuracy of 94.94% (**Figure 4** and **Table S2**) and an MCC 87% (**Table S3**). The optimal number of clusters based on PAM is 5, suggesting the existence of subtle differences in few patients (see Results: “Medulloblastoma patient stratification through multilayer structure analysis”).

Multilayer network enrichment analysis

To detect overrepresented features (drugs, pathways, etc.) that characterize each cluster, we performed a network enrichment analysis test (NEAT) (Signorelli, Vinciotti, and Wit 2016) in each layer of the multilayer network. NEAT tests whether

the number of edges between two groups of nodes is significantly higher (over-enriched) than by chance, assuming a hypergeometric null distribution. In our analyses, the two groups of nodes are (a) the minimal set of genes of a patient that are present in a layer, and (b) the genes annotated to a certain feature of that layer (e.g., the genes annotated to a specific drug in the drug layer). In the specific case of the molecular interaction layer, the annotation feature consists of the neighborhood of each gene of the minimal set of a patient. Once we identify significant hits for each patient using a p-value cutoff of 0.01 (Benjamini-Hochberg correction for multiple testing), we select those features that are enriched in all the patients of a cluster and unique to each cluster (**Table S6**).

Computational resources

All calculations were performed using the R statistical environment, in particular the packages stats (hierarchical clustering), fpc (k-medoids clustering), pvclust (clustering significance by multiscale bootstrap resampling), sigclust2 (clustering significance by Monte Carlo procedure), and neat (network enrichment analysis). To ease the detection and analysis of the multilayer community trajectories, we developed the R package CmmD, which is openly available at https://github.com/ikernunezca/CmmD.

Supplemental References

Archer, Tenley C., Tobias Ehrenberger, Filip Mundt, Maxwell P. Gold, Karsten Krug, Clarence K. Mah, Elizabeth L. Mahoney, et al. 2018. “Proteomics, Post-Translational Modifications, and Integrative Analyses Reveal Molecular Heterogeneity within Medulloblastoma Subgroups.” *Cancer Cell* 34 (3): 396–410.e8.

Bajpai, Akhilesh Kumar, Sravanthi Davuluri, Kriti Tiwary, Sithalechumi Narayanan, Sailaja Oguru, Kavyashree Basavaraju, Deena Dayalan, Kavitha Thirumurugan, and Kshitish K. Acharya. 2019. “How Helpful Are the Protein-Protein Interaction Databases and Which Ones?” *Cold Spring Harbor Laboratory*. https://doi.org/10.1101/566372.

Bollobás, Béla. 1998. “Ramsey Theory.” *Modern Graph Theory*. https://doi.org/10.1007/978-1-4612-0619-4_6.

Brunk, Elizabeth, Swagatika Sahoo, Daniel C. Zielinski, Ali Altunkaya, Andreas Dräger, Nathan Mih, Francesco Gatto, et al. 2018. “Recon3D Enables a Three-Dimensional View of Gene Variation in Human Metabolism.” *Nature Biotechnology* 36 (3): 272–81.

Croes, Didier, Fabian Couche, Shoshana J. Wodak, and Jacques van Helden. 2006. “Inferring Meaningful Pathways in Weighted Metabolic Networks.” *Journal of Molecular Biology* 356 (1): 222–36.

Didier, Gilles, Christine Brun, and Anaïs Baudot. 2015. “Identifying Communities from Multiplex Biological Networks.” *PeerJ* 3 (December): e1525.

Didier, Gilles, Alberto Valdeolivas, and Anaïs Baudot. 2018. “Identifying Communities from Multiplex Biological Networks by Randomized Optimization of Modularity.” *F1000Research* 7 (July): 1042.

Fabregat, Antonio, Steven Jupe, Lisa Matthews, Konstantinos Sidiropoulos, Marc Gillespie, Phani Garapati, Robin Haw, et al. 2018. “The Reactome Pathway Knowledgebase.” *Nucleic Acids Research* 46 (D1): D649–55.

Forget, Antoine, Loredana Martignetti, Stéphanie Puget, Laurence Calzone, Sebastian Brabetz, Daniel Picard, Arnau Montagud, et al. 2018. “Aberrant ERBB4-SRC Signaling as a Hallmark of Group 4 Medulloblastoma Revealed by Integrative Phosphoproteomic Profiling.” *Cancer Cell* 34 (3): 379–95.e7.

Hamming, R. W. 1950. “Error Detecting and Error Correcting Codes.” *Bell System Technical Journal*. https://doi.org/10.1002/j.1538-7305.1950.tb00463.x.

Kanehisa, Minoru, Yoko Sato, Miho Furumichi, Kanae Morishima, and Mao Tanabe. 2019. “New Approach for Understanding Genome Variations in KEGG.” *Nucleic*

Acids Research 47 (D1): D590–95.

Kaufman, Leonard, and Peter Rousseeuw. 1987. *Clustering by Means of Medoids*. https://wis.kuleuven.be/stat/robust/papers/publications-1987/kaufmanrousseeuw-clusteringbymedoids-l1norm-1987.pdf

Kimes, Patrick K., Yufeng Liu, David Neil Hayes, and James Stephen Marron. 2017. “Statistical Significance for Hierarchical Clustering.” *Biometrics* 73 (3): 811–21.

Kivela, M., A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter. 2014. “Multilayer Networks.” *Journal of Complex Networks* 2 (3): 203–71.

Mungall, Christopher J., Julie A. McMurry, Sebastian Köhler, James P. Balhoff, Charles Borromeo, Matthew Brush, Seth Carbon, et al. 2017. “The Monarch Initiative: An Integrative Data and Analytic Platform Connecting Phenotypes to Genotypes across Species.” *Nucleic Acids Research* 45 (D1): D712–22.

Oughtred, Rose, Chris Stark, Bobby-Joe Breitkreutz, Jennifer Rust, Lorrie Boucher, Christie Chang, Nadine Kolas, et al. 2019. “The BioGRID Interaction Database: 2019 Update.” *Nucleic Acids Research* 47 (D1): D529–41.

Wei, Chih-Hsuan, Alexis Allot, Robert Leaman, and Zhiyong Lu. 2019. “PubTator Central: Automated Concept Annotation for Biomedical Full Text Articles.” *Nucleic Acids Research* 47 (W1): W587–93.

Signorelli, Mirko, Veronica Vinciotti, and Ernst C. Wit. 2016. “NEAT: An Efficient Network Enrichment Analysis Test.” *BMC Bioinformatics* 17 (1): 352.

Suzuki, Ryota, and Hidetoshi Shimodaira. 2006. “Pvclust: An R Package for Assessing the Uncertainty in Hierarchical Clustering.” *Bioinformatics* 22 (12): 1540–42.