D3.1

Identification of important regulatory elements using multi-level matrix factorization approaches

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 / D3.1 / 1.0
Work package contributing to the deliverable	WP3
Due date	May, 2021 – M29
Actual submission date	2nd June, 2021

Responsible organisation	CURIE
Editor	Jane Merlevede, Juan Carrillo-Reixach, Alvaro del Rio, Pavel Sumazin, Carolina Armengol, Andrei Zinovyev
Dissemination level	Public
Revision	1.0

Abstract	D3.1 describes the techniques for dimensionality reduction used in iPC and their application to a selection of cohorts (at different omics levels) as well as a meta-analysis of the four solid tumor types of interest. The goal of the deliverable is to provide a list of pathways and biological functions having a key role in multiple paediatric cancers.
Keywords	Unsupervised deconvolution; Matrix factorization; Multiomics data integration; data integration; Meta-analysis

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

Editor

Jane Merlevede (CURIE)

Andrei Zinovyev (CURIE)

Carolina Armengol (IGTP)

Juan Carrillo-Reixach (IGTP)

Alvaro del Rio (IGTP)

Pavel Sumazin (BCM)

Contributors

CURIE

IGTP

BCM

Disclaimer

Executive Summary

Task 3.1: Investigation of cross-cancers molecular similarities

D3.1 reports the application of unsupervised deconvolution using various approaches with the aim of looking for processes acting in the paediatric solid tumors of interest in iPC, that are Ewing sarcoma (ES), Medulloblastoma (MB), Neuroblastoma (NB) and Hepatoblastoma (HB). Our approach is based on matrix factorization, either at a single layer (gene expression) or at multi-omics layers.

D3.1 describes the techniques for dimensionality reduction used in iPC and their application to a selection of cohorts (at different omics levels). The goal of the deliverable is to provide a list of pathways and biological functions having a key role in multiple paediatric cancers. Moreover, for each identified component, we select the most contributing genes, pathways, transcriptional regulators that can be used as biomarkers.

D3.1 provides a set of components / communities annotated using public databases, that describes various processes acting in each cancer type or in a combination of these cancer types. These resources are available at: https://data.ipc-project.bsc.es/s/qAWM2oba4zJEtTq

D3.1 provides a data resource of 112 datasets across the four solid tumor types, that can be used for other deliverables of iPC. These resources are available at: https://data.ipc-project.bsc.es/s/bFC9mPtqPEgfNMn.

D3.1 provides methods to reproduce the analyses discussed here for interested users: https://github.com/merlevede/iPC_WP3_D3.1. More generally, the users interested in applying ICA on their own datasets can use the jupyter notebook developed in our team by Nicolas Captier: https://github.com/ncaptier/Stabilized_ICA.

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

“Unsupervised deconvolution” is an approach that allows to identify processes acting in data, without searching for variation between predefined sample sets.

“Dimensionality reduction” / “matrix factorization” is an algorithm that allows the decomposition of a matrix into the product of two lower dimensionality rectangular matrices. The aim is to summarize the amount of information into principal sources.

Table of Content

List of Figures

List of Tables

Chapter 1 Introduction

The amount of data produced per cancer type has increased dramatically in the last couple of years and enables to address questions that could not be considered before. This is in particular true for the most frequent adult cancers but, to a smaller extent, also for some pediatric tumors. We observe both an increase in the number of tumors sequenced as well as a broader spectrum of tumor characterization using multi omics data, such as genome, transcriptome, methylome, epigenome, proteome, ... The integration of this diversity of data, together with relatively large cohorts allows now to address important questions, as i) what are the mechanisms acting (in each “layer” of) a cancer type?, ii) can the patients be stratified across subtypes based on genomic or epigenomic features?, iii) can we predict survival and therapeutic response using clinical, genomic and epigenomic features?

In this work, we looked at the mechanisms underlying each layer of a cancer type. For the four solid tumors of interest in iPC, namely Ewing sarcoma, Medulloblastoma, Neuroblastoma and Hepatoblastoma, we focused in particular on the transcriptomic level. We also looked at the common and specific mechanisms in these cancer types, searching for common processes between these quite different pediatric cancers. In addition, we also investigated additional layers, proteomics and phosphoproteomics layers, on a Medulloblastoma cohort (Forget, 2018) as well as the CBTTC cohort (The Children’s Brain Tumor Tissue Consortium) from partner CHOP (Children Hospital Of Philadelphia), searching for mechanisms that would be detectable specifically using multiple layers.

The goal of this deliverable is to identify the sources of between sample variance present in multi-omics datasets and annotate them in terms of perturbed molecular mechanisms. High-throughput data can be seen as a mixture of biological factors. In order to deconvolute this mixture, we employed various existing single and multi-level matrix factorization approaches, such as independent component analysis (Comon, 1994), tensor independent component analysis (Teschendorff, 2018) and Multi-Omics Factor Analysis (Argelaguet, 2018). The results of these analyses include a list of pathways and biological functions having key or important roles in these paediatric cancers. They also are general processes expected in most cancer types. Moreover, for each identified factor, we selected the most contributing regulatory genes and transcriptional regulators. Some of them might have the potential to be employed as biomarkers.

In this report, we provide a description of the datasets considered in our approaches in chapter 2 and a short overview of the methods used in the field as well as a description of the method we developed in chapter 3. We performed unsupervised deconvolutions of gene expression data from the solid tumor types of interest in iPC as well as their meta-analysis in chapter 4. We also illustrate the use of multi-omics data integration on Medulloblastoma mainly in chapter 4. In this chapter, we briefly summarize the results of the application of the methods, providing more details for two particular datasets and for our meta-analysis. Finally, in the conclusion chapter, we summarize the findings and discuss the dissemination of our work in the iPC consortium as well as future work.

Our results are accessible on NextCloud. We hope that our biological and clinician partners will be interested in looking at the catalogues of pathways identified by these approaches. We initiated discussion with Carolina Armengol (IGTP) for HB and Olivier Ayraud (CURIE) for Medulloblastoma. Also, the datasets collected can be used for other analyses and are accessible on NextCloud. Finally, the methods are also online. All the useful links are provided in the report.

Chapter 2 Datasets used for unsupervised deconvolutions

We used various datasets to perform unsupervised deconvolutions on single or multi-layer(s) to identify pathways and biological functions having a key role in paediatric cancers. The data can be browsed at https://data.ipc-project.bsc.es/s/bFC9mPtqPEgfNMn.

2.1 Description of the datasets used for matrix factorization

We collected various gene expression data (bulk + single cell ; patient + PDX + cell lines ; comparable adult tumors when available ; controls when available) for each solid tumor type of interest in iPC. The numbers of these datasets and their sample sizes are reported in the table below.

	Nb of bulk datasets (Nb of samples)					**Nb of single cell datasets (Nb of cells)**				
	patient	PDX	cell line	control	adult	patient	PDX	cell line	control	adult
ES	5 (432)						8 (12,338)	4 (7,895)	1 (96)	
MB	7 (1,743)				2 (696)	25 (7,745)	11 (946)			
NB	4 (804)	1 (33)	1 (39)			28 (146,800)			9 (63,776)	
HB	3 (134)				1 (373)					
LK	9 (2158)	1 (90)			1 (173)	23 (36,085)			7 (7,574)	

Table 1 : Description of the datasets collected for each tumor type

Note 1: The number of datasets indicated above is the number of cohorts of patients for bulk data while it is the number of patients for single cell data.

Note 2: For each bulk dataset, the sample size is the number of samples (patients, PDX, cell lines) while for a single cell dataset, the sample size is the number of cells per patient/PDX/cell line/controls.

Note 3: We collected datasets for leukemias (ALL and AML) but in the end focused on the solid tumors. In the following, we do not present results on leukemia.

In addition to these datasets, 2 controls (full embryos, 36,002 cells) were used to study the solid tumors as explained later in the methods chapter.

One of our aims is to compare pediatric tumors to their adult counterpart. As comparable adult tumors, we used GBM and LGG TCGA datasets to be analyzed with Medulloblastoma and Liver Hepatocellular Carcinoma TCGA dataset to be analyzed with Hepatoblastoma.

2.2 Description of the datasets used for multi-omics integration

For the integration of multi-omics layers, we mainly used the Medulloblastoma cohort published in (Forget, 2018) and the Children's Brain Tumor Tissue Consortium (CBTTC) cohort from partner CHOP. The cohort in (Forget, 2018) contains gene expression data from 35 patients, as well as proteomics and phosphoproteomics data for 38 patients. CBTTC gathers a wide range of brain tumors, from which we selected only the four most represented: High and Low Grade Gliomas,

Medulloblastoma and Ependymoma, which gave in total gene expression data for 605 patients, as well as proteomics and phosphoproteomics data for 166 and 165 patients respectively. Patients for whom a layer is missing are represented in grey in the corresponding layer in Figure 1.

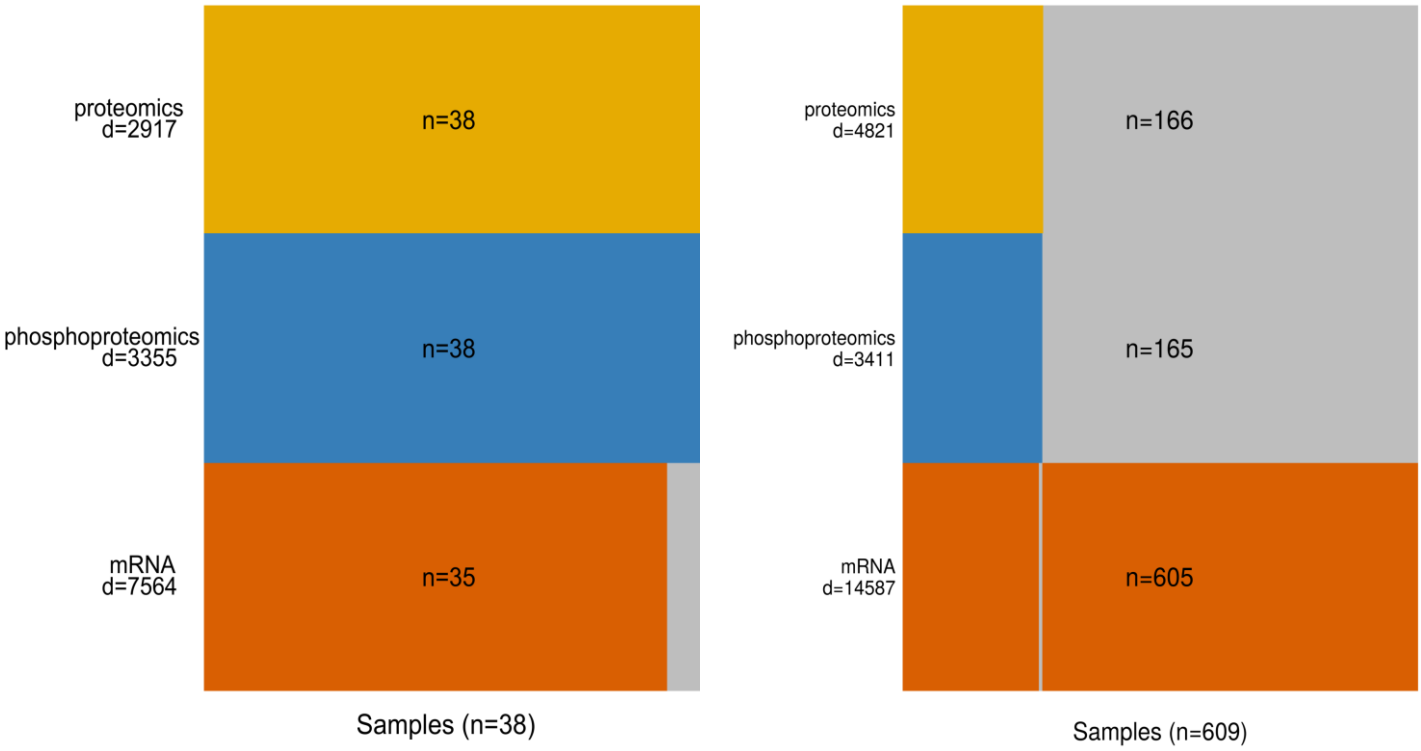


Figure 1: Description of the MB cohort (Forget, 2018) (left) and the CBTTC cohort (from CHOP partner) (right). The number of cases and features used per layer is indicated

Chapter 3 Deconvolutions by matrix factorization and multi-omics data integration

Different approaches were used to explore deconvolution of solid pediatric tumors. First, we performed matrix factorization on gene expression data on multiple tumor types. Then, we addressed cross-cancers molecular similarities by performing a meta-analysis on the results obtained on each tumor type. Then, we performed multi-omics data integration on a Medulloblastoma dataset and the CBTTC cohort. Lastly, we discuss alternatives to matrix factorization to decipher mechanisms acting in tumors.

3.1 Deconvolution and interpretation of gene expression data

To analyse the gene expression data, our approach consists in 4 steps, from decomposing a single dataset to interpreting signals found across many datasets as illustrated in Figure 2. The workflow below summarizes them and few details are given about each step in the following paragraphs. In the next chapter, we show an application of this pipeline to ES and HB data in particular.

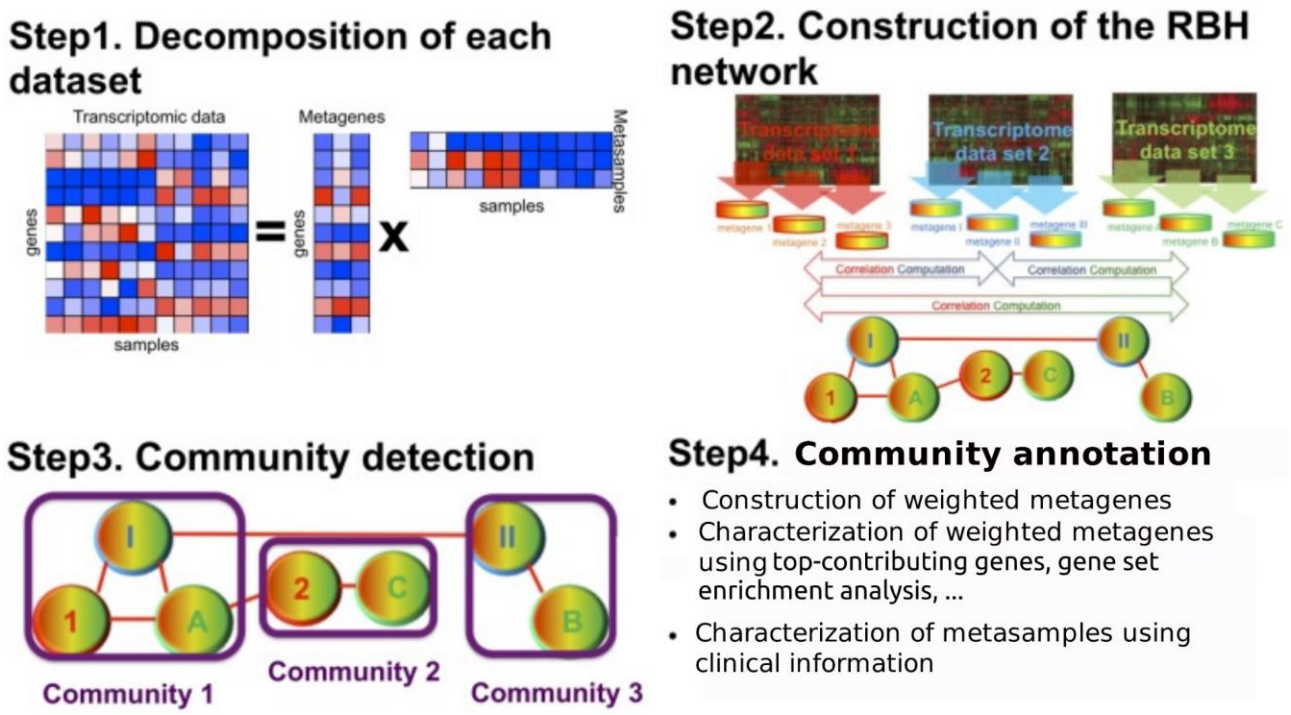


Figure 2 : Overview of unsupervised deconvolution using matrix factorization (Adapted from (Cantini L. K., 2019))

3.1.1 Matrix factorization of gene expression data using sICA

There exist various methods to perform matrix factorization. The most widely known and used are Principal Component Analysis (PCA) and Non-negative Matrix Factorization (NMF). Based on our previous work on analysing gene expression data, we demonstrated the advantage of ICA over these alternatives in (Cantini L. K., 2019). Thus, we selected sICA, an iterative application of ICA that ensures robustness of the identified components (Hyvärinen, 2001). Each dataset is decomposed

into 2 matrices, S “metagenes” and A “metasamples”. S is a matrix with the number of genes in rows and the number of components in columns and A is a matrix with the number of components in rows and the number of samples in columns.
For each tumor type, we systematically applied sICA on each dataset using the following criteria to define the number of components: max(50, nb of samples/3). From our previous work in (Kairov, 2017), we showed that usually 50 composants is sufficient to see most of the relevant biological signals and taking too many components does not alter the quality of the first extracted components. Thus, we get a collection of metagenes and metasamples for each tumor type.

3.1.2 Reciprocal Best Hits

Using different datasets of a tumor type should allow to see a signal or pathway several times. Thus, an important and robust mechanism should be identified, not in every, but in most of or several of the datasets.
The idea through Reciprocal Best Hit (RBH) analysis is to identify components reflecting the same processes across various datasets, as introduced in (Cantini L. K., 2019). This name was chosen in analogy with the namesake common definition of orthology in comparative genomics (Bork, 1998) and (Tatusov, 1997).
We use correlation to compare all the pairs of metagenes between two given datasets. If a pair has a reciprocal best correlation, then this correlation gives a weight to this relationship.

3.1.3 Community detection

Going further, the best hits across various datasets give rise to communities, an ensemble of components sharing biological properties. To detect such communities, we used the Markov Clustering algorithm (Enright, 2002).

3.1.4 Community annotation

Our ultimate goal is to characterize these communities. To do so, we summarize the information of all the metagenes forming a community by constructing a weighted metagene, an average of the different metagenes. Thus, we can then perform annotations of these synthetic weighted metagenes.

3.1.5 Method implementation availability

To perform sICA and RBH analysis, we adapted the python implementation developed in our group at https://github.com/ncaptier/Stabilized_ICA. This notebook was developed during the course of this deliverable. For interested users, this can be easily applied through its python notebook. To perform step 3 and 4, community detection and annotation, we used various R packages through Rmarkdown. The code used for our deliverable is accessible through https://github.com/merlevede/iPC_WP3_D3.1. We gave a subset of ES datasets to make this code usable.

3.2 Meta-analysis of gene expression data of solid tumors using matrix factorization

To identify signals shared across the different pediatric tumor types of interest in iPC, we focused on the four solid tumors ES, MB, NB and HB. Using all their metagenes, computed on each analysis, we performed steps 2-4, as previously. We first looked for reciprocal best hits among the matrices describing the different communities, then performed clustering and finally annotated the identified meta-communities.

3.3 Multi-omics data integration

We first discuss a short overview of existing methods for multi-omics data integration and then briefly describe the two methods we applied.

3.3.1 Overview of multi-omics data integration methods

There exist several benchmarks tackling this problem, *e.g.* (Pierre-Jean, 2020) and (Cantini L. Z., 2021). The methods can be classified according to different criteria, like the integration stage, their assumptions concerning the components and their approach.
First, they can be characterized as late or early integration, *i.e.* combining results obtained on single omics *vs* intermediate integration.
Second, the hidden components can be considered to be shared across the different layers or to be different in each omics layer or the components can be considered as a combination of both.
Third, there are 4 approaches, among which we consider here only the first one:

- Dimensionality Reduction / Matrix Factorization: intNMF, JIVE (joint and individual variation explained), iCluster, iClusterPlus, MoCluster, tICA, scikit-fusion
- Network-based methods / Similarity matrices: SNF (Similarity network fusion), LRACluster (Low-rank clustering), PINSPlus (Perturbation clustering for data INtegration and disease Subtyping), ConsensusClustering, mixKernel. This kind of approach was the topic of D4.1.
- Canonical correlation analysis: RGCCA (regularized generalized canonical correlation analysis), SGCCA (sparse generalized canonical correlation analysis), MCIA (Multiple co-inertia analysis)
- Bayesian methods: MOFA (MF), MSFA (multi-study factor analysis)

The published benchmarks are useful to get guidelines based on what the authors tested. For example, in (Cantini L. Z., 2021), the authors recommend the use of RGCCA, MCIA, MOFA for looking at clinical associations ; the use of tICA and MCIA for deciphering biological processes and pathways association and the use of intNMF for clustering. Depending on the questions to be addressed, one might consider one specific approach.
In the results part, we applied MOFA and tICA, two R packages, on the Medulloblastoma cohort from Forget et al and MOFA on the CBTTC cohort. We briefly describe here the approach of these two matrix-factorization based methods.

3.3.2 MOFA and tICA

Multi-Omics Factor Analysis (MOFA) (Argelaguet, 2018) is characterized by intermediate integration, the components are considered as a combination of shared factors across the different layers and specific factors in each omics layer. It aims at explaining relationships between groups of variables. It extracts the maximum common variance from all variables and puts them into a common score. Each layer is decomposed into a set of weight matrices W and a single matrix of factors Z. A weight matrix for a layer is a set of vectors, which each of them represents a factor / component and defines the importance of each gene in this factor. Z is a set of vectors, each of them represents a factor / component and defines the contribution of each sample in this factor.

Tensorial Independent Component Analysis (tICA) (Teschendorff, 2018) is based on the usage of a tensor. A tensor can be seen as a matrix of N dimensions. In our case, it has 3 dimensions. In the first dimension, there are the samples (like patients), in the second dimension, the features (like genes) and in the third dimension, the data types (like gene expression, proteomics, copy number, …). When applying tICA, one needs to have the same features for the different omics, which can be a limitation when dealing with methylation or phosphoproteomics data. Also, missing values are not accepted.

3.4 Alternatives to matrix factorization

This deliverable makes a part of the Task 3.1: “Investigation of cross-cancers molecular similarities”. (Multi-omics) matrix factorization approaches are among the non-supervised approaches that can be used but are not the unique way to address this question. Few reviews discuss the range of available methods for this challenge, as the one by (Avila Cobos, 2018) (UGHENT partner) and a book chapter by our team (Czerwińska, 2019). The following link points to a list of various non / semi supervised and supervised approaches for expression and methylation data: https://cancer-heterogeneity.github.io/data_challenges_tools.html.

The main alternatives to matrix factorization are supervised or guided approaches. The most commonly used group of methods is ordinary least squares (OLS), and in particular non-negative regression in this context, with non-negative least squares method (NNLS). A second group of methods are support vector regression approaches with linear kernel, including CIBERSORT (Newman, 2015) and ImmuCC (Chen, 2017).

Some successful methods, such as MCPCounter (Becht, 2016) abandon the regression-based problem formulation and rely on more or less simple scoring approaches (such as taking mean expression of marker genes) in order to compute some relative cellular component abundances, which are comparable between tumor samples but not comparable between different components.

Advantages and drawbacks were put in light for both supervised and unsupervised methods and are nicely discussed in the two reviews previously mentioned. One advantage of unsupervised deconvolution is the absence of requirements concerning the expected number of cell types and the cell types themselves. This was an advantage as in this work we considered four different tumor types and we did not know in advance which cell types to expect. In addition, unsupervised deconvolution is not restricted to cell type identification. It also allows the detection of various processes, from very general signals, like cell cycle, to tumor-related signals, like angiogenesis.

Chapter 4 Results

In this chapter, we describe the results of the different analyses, starting with the application of matrix factorization on the solid tumor types. We provide a brief summary of the findings made on each tumor type and an extended description of the findings on ES. We then describe the interpretation of the results on HB tumors provided by partner IGTP. Then, the results of the meta-analysis comparing the various tumors are discussed. Next, we describe the integration of different omics layers on MB and additional brain tumors. Finally, we present a recent work based on unsupervised clustering, and not on matrix factorization, where partner BCM addressed questions related to our topic in hepatocellular carcinomas from young children to young adults.

4.1 Summary of the unsupervised deconvolutions using matrix factorization on gene expression data of ES, MB, NB and HB

The application of the approach described in the previous chapter on each solid tumor type led to the detection of communities acting in these tumor types. From the 20 datasets gathered for ES tumors, we identified 29 communities of at least 3 components. From the 45 datasets gathered for MB tumors, we identified 47 communities of at least 3 components. From the 45 datasets gathered for NB tumors, we identified 75 communities of at least 3 components, with at least 3 components from tumors. From the 6 datasets gathered for HB tumors, we identified 11 communities of at least 2 components, including not only controls. All these communities were annotated and their descriptions are available and browsable by the interested partners at: https://data.ipc-project.bsc.es/s/qAWM2oba4zJEtTq.

4.2 Matrix factorization on gene expression data of ES

When working on our workflow, we wondered if it is possible to analyse similarly bulk and single cell data. One aspect to look at to answer this question, is to see if some of the identified communities gathered these two data types. As illustrated in Table 1 below, it is the case, and some expected signals, like cell cycle in community 1, mixes these data types. In total, 11 out of 29 communities are a mixture of bulk and single cell data, 3 contains only bulk and 15 contains only single cell datasets.

	nb_comp	nb_comp_unique	**nb_bulk**	**nb_sc**	nb_patient	nb_cellline	nb_PDX	nb_control
C1	20	20	**5**	**15**	5	4	8	3
C2	14	12	**3**	**9**	3	0	7	2
C3	13	13	**3**	**10**	3	2	6	2
C4	13	13	**0**	**13**	0	4	7	2
C5	10	10	**3**	**7**	3	4	1	2
C6	10	10	**1**	**9**	1	3	4	2
C7	9	9	**0**	**9**	0	1	6	2
C8	8	8	**0**	**8**	0	3	5	0
C9	7	7	**0**	**7**	0	1	4	2

	nb_comp	nb_comp_unique	**nb_bulk**	**nb_sc**	nb_patient	nb_cellline	nb_PDX	nb_control
C10	7	7	**5**	**2**	5	0	0	2
C11	7	7	**0**	**7**	0	0	7	0
C12	7	7	**5**	**2**	5	0	0	2
C13	6	6	**4**	**2**	4	0	0	2
C14	5	5	**0**	**5**	0	3	0	2
C15	5	5	**0**	**5**	0	2	1	2
C16	5	5	**0**	**5**	0	1	3	1
C17	5	5	**3**	**2**	3	1	1	0
C18	4	4	**0**	**4**	0	2	0	2
C19	4	4	**0**	**4**	0	2	0	2
C20	4	4	**0**	**4**	0	2	0	2
C21	4	4	**4**	**0**	4	0	0	0
C22	3	3	**0**	**3**	0	3	0	0
C23	3	3	**0**	**3**	0	1	0	2
C24	3	3	**0**	**3**	0	0	1	2
C25	3	3	**0**	**3**	0	1	0	2
C26	3	3	**1**	**2**	1	0	0	2
C27	3	3	**3**	**0**	3	0	0	0
C28	3	3	**2**	**1**	2	0	1	0
C29	3	3	**3**	**0**	3	0	0	0

Table 2 : Description of the communities identified in ES tumor datasets

We confirmed our findings by comparing our results with the ones from (Aynaud, 2020). In this paper, the same single cell datasets were analysed. 30 independent components were extracted, including cell cycle, extracellular matrix organization and one component defined as specific signature IC-EWS containing direct targets of EWS-FLI1 from (Aynaud, 2020). Most of (24/30 IC+ and 21/30 IC-) the components identified in (Aynaud, 2020) were retrieved. In particular, cell cycle phases in C1+ (G1/S) and C10+ (G2/M), the specific signature IC-EWS containing direct targets of EWS-FLI1 in C21+, the hallmark of ECM in C15+, C20-, C25+ and C27+ are identified as well. As most of the datasets are shared between the two analyses, it is expected to retrieve similar results, but it proves that adding bulk to single cell data did not perturb the signals.

In addition, other interesting mechanisms were identified like immune response in C4+ (neutrophil activation, neutrophil degranulation), angiogenesis in components C19+ and C26-. Interestingly, both tails of the components 26 are significantly enriched, the negative tail in angiogenesis and the positive tail in RNA splicing.

4.3 Interpretation of the deconvolution of gene expression data of HB

The results of unsupervised deconvolution of HB datasets have been analysed by the group of Carolina Armengol Niell (iPC partner), who had the following conclusions.

First, they report the importance of the 14q32 cluster: the genes of the 14q32 regions (e.g. *SNORD113* and *SNORD114*) were reported as top contributing genes in several communities. This result fits well with their recent paper (Carrillo-Reixach, 2020) in which they reported a strong overexpression of this cluster in HB and its correlation with poor prognosis. The 14q32 genes in the community 10 are not the same as they described in the 14q32 cluster associated with HB. In their opinion, this is a significant result because it appears in the community "HB specific" (community 11) but it is also shared with adult HCC and in other paediatric malignancies as previously reported.

Second, "Glucuronidation" appears as a top contributing pathway in several communities. After a comprehensive evaluation in their gene expression dataset (32 tumors and 32 non-tumors), they found that glucuronidation is strongly downregulated in tumors compared with non-tumors. This profile fits well with the fact that the UDP-glucuronosyltransferases (UGTs) - key regulatory enzymes of this pathway - are upregulated during human hepatic development (Strassburg, 2002). The glucuronidation is a well-recognized phase II metabolic pathway for a variety of chemicals including drugs and endogenous substances and it is considered to be a detoxification process or a defence mechanism that helps humans remove unwanted substances and this agrees with the functions of mature hepatocytes. Overall, it seems that this finding could be a consequence of the degree of immaturity of the tumors but probably not a driver of tumorigenesis. However, they found that UGTs are upregulated by the NFE2L2/NRF2 pathway, which in turn plays an important role in liver protection against DNA damage (Paonessa, 2011). Since mutations in *NFE2L2* have been reported in about 10% of the HBs and specially in aggressive tumors, they investigated whether UGT genes and the Glucuronidation pathway could be upregulated in HBs of poor prognosis. Accordingly, they performed GSEA analysis to investigate their enrichment in tumors classified as high-risk (MRS-3, defined in (Carrillo-Reixach, 2020)) vs. tumors with intermediate or low-risk (MRS2 and MRS1). Despite no significant difference was found, there was a trend to be up-regulated in more aggressive tumors. Because this pathway could have an important role in cell protection against DNA damage, the activation of glucuronidation could have a protective effect of tumor cells against chemotherapy. In conclusion, this pathway might be interesting if it appears in the communities of other paediatric liver tumors and it could be worth studying in detail its association with patient outcome/chemotherapy response.

Third, HB has been associated with deficiencies of imprinting. Some of the top-contributing genes and chromosomal regions that contained these genes (i.e. *IGF2* and *INS-IGF2* at 11p15, *PEG10* at 7q21) are present in the community 11. This might be interesting as one of the hallmarks of HB is the overexpression of imprinted genes. This is also a common event in other pediatric tumors and it would be interesting to see if it appears in the study of other tumor types.

In addition, they observed an activation of Wnt signaling (main HB hallmark, (Armengol, 2011)), PI3K-Akt pathway, cell cycle regulation, TGF-beta signaling and DNA repair processes, all of which have been reported as deregulated in HB. They also identified a down-regulated pathway related to xenobiotic metabolism processes and cytochrome P450 that are down-regulated in tumors as they are specific to mature hepatocytes. Other significant pathways/genes are also related to metabolic functions (i.e. ascorbate and aldarate metabolism in community 11) and are probably related to mature hepatocyte functions.

Finally, they observed an enrichment of transcription factors (TFs) related to liver development (e.g. HNF and FOX family genes) in different communities. These TFs appeared in 8 communities. For example, they identified several TFs related to inflammation: the community 8 has a remarkable enrichment of TFs (e.g. STAT3, STAT4, CEBPB, SP1) related to the expression of inflammatory genes and involved in IFN gamma production, Th1 response, among others. Interestingly, STAT3

inhibition prevents the establishment of a pro-metastatic niche and inhibits liver metastasis (Lee JW et al, 2019). All this data corroborates the results presented in Hirsch, TZ et al., 2021, in which they identified a HB subgroup that strongly expressed TFs involved in hepatic differentiation (HNF1A, HNF4A) that includes tumors with different degrees of immune infiltration.

4.4 Meta Analysis of gene expression data of solid tumors using matrix factorization

To identify signals shared across the different pediatric tumor types, we used the metagenes extracted from the four solid tumors. In total, we had 112 matrices of metagenes, which contain between 8 and 50 metagenes. We first looked for reciprocal best hits among their metagenes describing the different communities. The RBH network included in total 4,748 nodes, *i.e.* metagenes, and 53,460 reciprocal best hits. To keep only meaningful associations, we removed RBH with a correlation below 0.2. These generated isolated nodes that we filtered out. Thus, the network was composed of 2,305 nodes and 25,394 edges. On this new network, we performed Markov clustering and the resulting network consisted of 2,305 nodes and 22,007 edges. 255 communities composed of at least 2 components were identified, as shown in Figure 3. We then required one of the following criteria on these communities: communities should consist in i. >=3 components coming from any combination of pediatric tumors, ii. >=2 components coming from any combination of pediatric HB tumors or adult liver cancers, iii. >=2 components coming from any combination of adult tumors (GBM, LGG, HCC). Using these criteria, 142 communities remained. Finally, we annotated these 142 communities as previously. These results are available at: https://data.ipc-project.bsc.es/s/qAWM2oba4zJEtTq.

In these 142 communities, the number of components per community is between 2 and 80. 70% of the (99 out of 142) communities are a mixture of bulk and single cell data, 5 contain only bulk and 38 contain only single cell datasets. 67 out of 142 (47%) communities do not contain any control. If the reader is in particular interested in tumor related signals exclusively, he can focus on communities containing no component from the control datasets. If one wants to investigate the mechanisms found only in children, we can exclude communities containing components coming from adult datasets and focus on communities without such components; there are 98 out of 142 (69%) such communities.

Detailed investigation of these communities is ongoing with experts. First, we retrieved general pathways as cell cycle (C1-, ...) and extracellular matrix (C25+, C26+, C31+, C39+, C40+, C47+, C59-, C70-, C71+, C74+, C80+, C83+). We also observed angiogenesis (C7+, C62-, C70+), a usual pathway expected in cancer studies.

Second, we identified several pathways linked to immunity: innate immune response (C99+, C106-), adaptive immune response (C101+, “neutrophil activation”; C110+, “lymphocyte activation”; C122+, “T cell activation”; C129+, “T cell activation”). Several pathways annotated as “humoral immune response mediated by circulating immunoglobulin” were observed (C13+, C14+, C15+, C38-). Nevertheless, these communities need to be further studied as the immunoglobulin genes might be captured in a specific and biased manner during sequencing, as for ribosomes. Several additional communities were linked to immunoglobulins, in particular annotated as “immunoglobulin complex” with often very small p values.

Third, we also identified dysregulation of several pathways known to be implicated in several of the four tumor types. For example, Wnt signaling is known to be deficient in the four tumor types. Community 93+ reports several pathways, including “Wnt signaling pathway” (p value adjusted 10^{-5}). Community 95- reports “negative regulation of Wnt signaling pathway” as its most significantly enriched pathway (p value adjusted 10^{-3}). Also, community 128+ reports “Wnt signaling pathway” ex-aequo with various pathways (p value adjusted 10^{-4}).
Another example is the PI3K-Akt pathway, which also appeared as significantly enriched in several communities (C11+, C12+, C17+, C18+, C25+, C26+, C31+, C33+, C34+, …) with p values adjusted often around 10^{-5}.

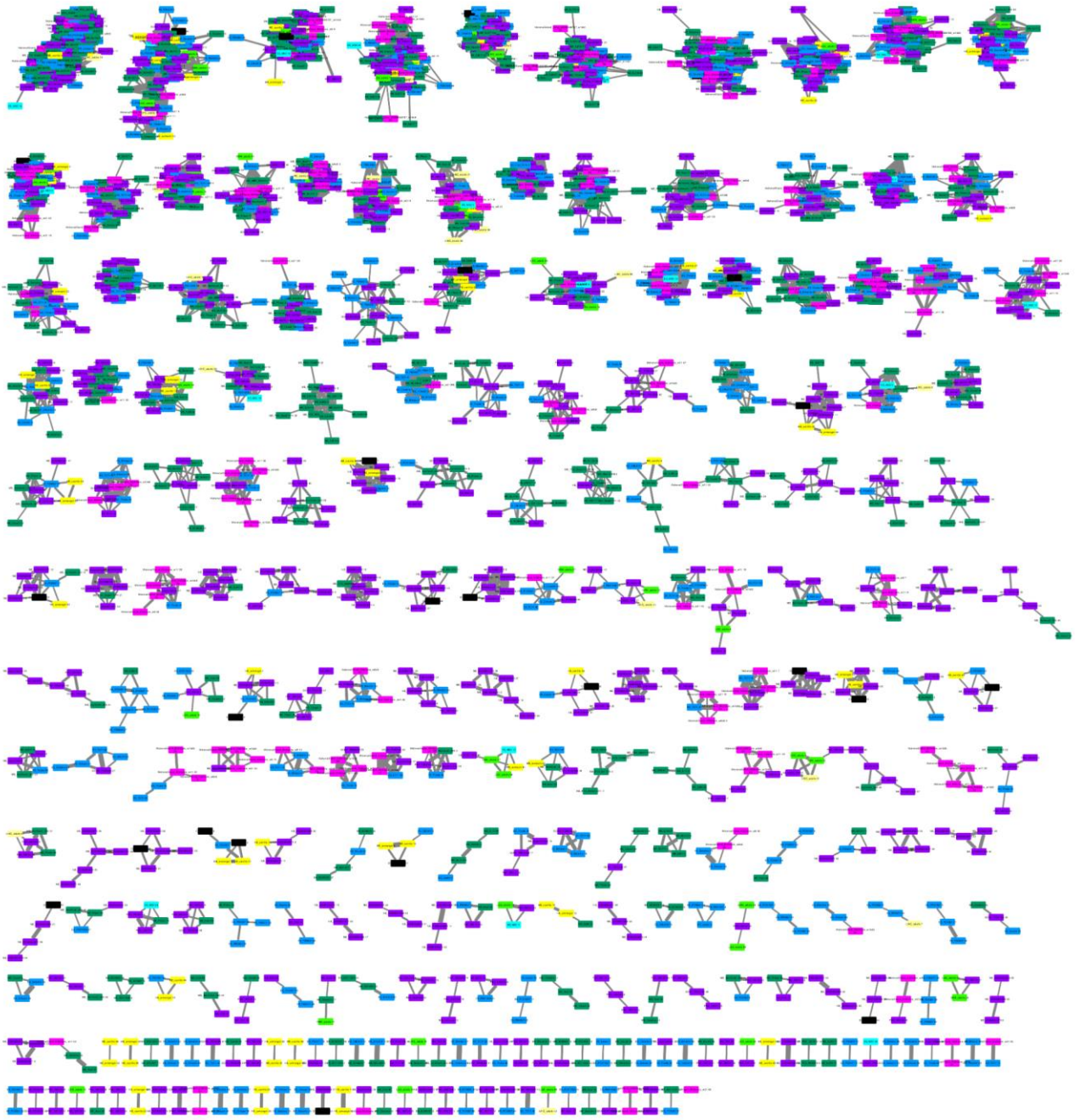
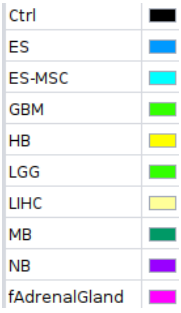


Figure 3 : Graph obtained after performing markov clustering on the RBH graph of the weighted metagenes of the 4 solid tumor types for weights above 0.2

The legend for the different pediatric tumors (ES, MB, NB and ES), the different adult tumors (GBM, LGG and LIHC), the controls coming from full embryos (Ctrl) as well as control cells for ES, mesenchymal cells (ES-MSC) and for NB fetal adrenal glands (fAdrenalGland) is shown here.



4.5 Multi-omics data integration of MB data

To look at a tumor in a broader view, and not be limited to gene expression data, we can perform multi-omics integration when multi omics are available. In this report, we first illustrate the use of MOFA on MB. Running MOFA on the small cohort of MB patients with 35 samples with gene expression and 38 with proteomics and phosphoproteomics, the model extracted 12 factors, with the requirement that a factor should explain at least 2% of the variance. About 60% of the variance of each layer is explained by these 12 factors as shown in Figure 4. The first factor is explained by the 3 layers, but dominated by proteomics and phosphoproteomics, while the second factor, also explained by the 3 layers, is dominated by gene expression data. Factor 5 is only explained by proteomics and phosphoproteomics.

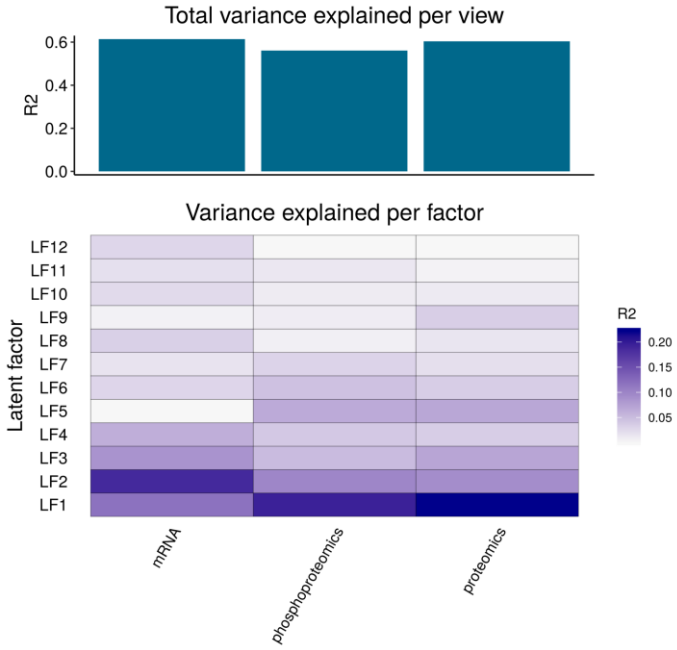


Figure 4 : Variance explained by each layer using MOFA on the MB cohort from (Forget, 2018)

Let us now have a look at the matrix of factors.
This matrix shows the factors in row and the samples in columns, so it is possible to look at the representation of the samples using their coordinates in any factor. Below we plotted the coordinates of the samples according to factors 1 and 2 and then 1 and 3 in Figure 5. Factor 1 extracted the G4 subgroup, while factor 2 distinguished SHH subgroups from the remaining subgroups. Finally, the third factor allowed to separate the WNT subgroup. These 3 first identified factors allow to separate the 4 MB subgroups of patients in this cohort, which was not fully possible using a single layer, whatever the layer (gene expression; proteomics, phosphoproteomics, methylation) (Forget, 2018).

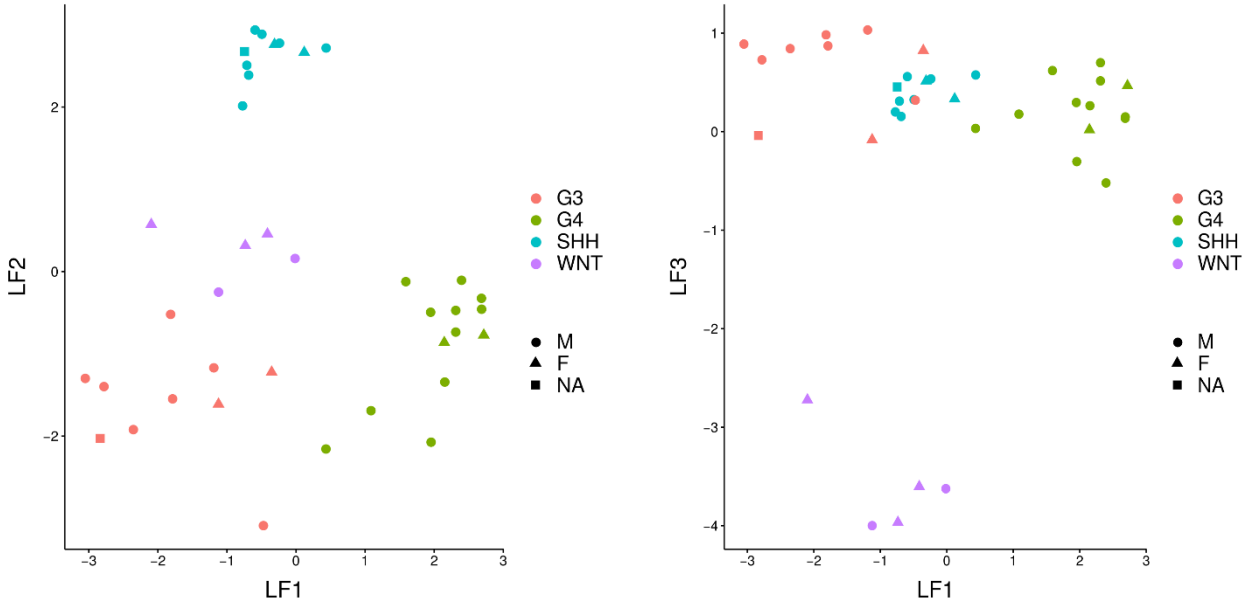


Figure 5 : Projection of the samples onto the first two latent factors (left) and onto latent factors 1 and 3 (right)

Let us now have a look at the weight matrices. We can look at the features having the highest weights in the different factors for each layer. In addition, we can perform gene set enrichment analysis based

on the weights attributed to each gene. Table 3 shows the most significantly enriched pathways for each of the 12 factors, for the layers gene expression and proteomics using Reactome database. Interestingly, the first 3 factors have significant enrichment in pathways representing the characteristics of patient subgroups. Other factors have enrichment in expected pathways like cell cycle. We also retrieved pathways that can make sense in the context of brain tumors, like "phototransduction cascade" in factor 4.

Factor	GSEA mRNA	GSEA Proteomics
1		rRNA processing
2	Hedgehog "on" state	
3	Negative regulation of WNT	Extracellular matrix organization
4	phototransduction cascade	
5		cell cycle
7	Interleukin-10 signaling	Semaphorin interactions
9	Integrin cell surface interactions, ECM proteoglycans	Respiratory electron transport
11		base excision repair
12	cell cycle	
6, 8, 10		

Table 3: Pathway analysis performed on the results obtained using MOFA on gene expression and proteomics layers of MB tumors (Forget, 2018)

Then, we tried tICA on this cohort. Before the decomposition, the subspace dimension is estimated based on random matrix theory algorithm. Using this approach, 7 components were chosen. After the decomposition, it is quite straightforward to correlate the inferred components with associated phenotypes from the R package. As shown in Figure 6, all components are correlated with one (or more) clinical features. Then, we can look at the weights of the samples in a given factor. For example, groups 3 and 4 have the most different weight distributions in component 6, while the weight distributions of SHH and WNT are in between and very close from each other, see Figure 7.

Concerning the gene space, we can extract the most contributing genes in each component. There is no implementation to run enrichment analysis, but this can be done using other R packages.

The results of the multi-omics data analysis on this small MB cohort were described in a poster presented in a data challenge on "Matrix factorization and deconvolution methods to quantify tumor heterogeneity in cancer research" in which we participated in 2018 and 2019: https://data-institute.univ-grenoble-alpes.fr/research/data-science-for-life-sciences/health-data-challenge-2nd-edition-matrix-factorization-and-deconvolution-methods-to-quantify-tumor-heterogeneity-in-cancer-research-801962.htm. This was also discussed in a blog post on the iPC website: https://ipc-project.eu/health-data-challenge-matrix-factorization-and-deconvolution-methods-to-quantify-tumor-heterogeneity-in-cancer-research/

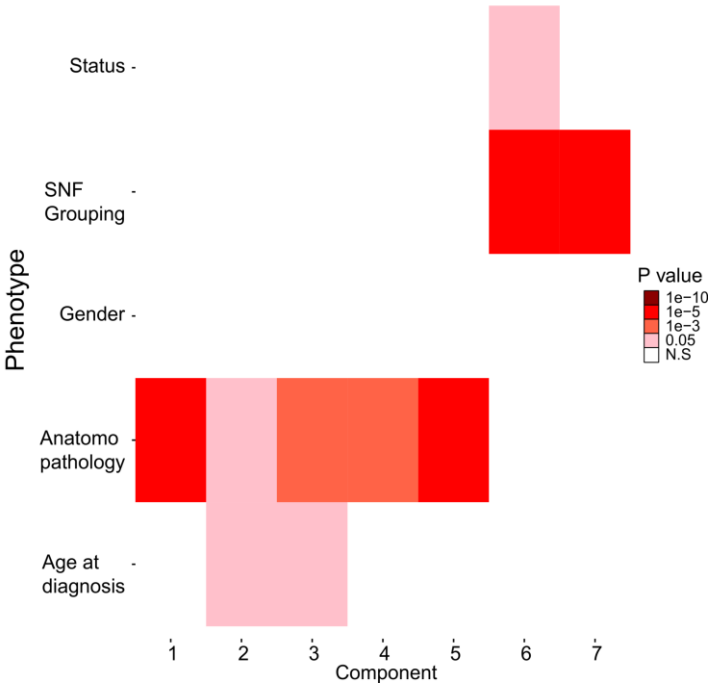


Figure 6: Associations between the identified components and clinical informations

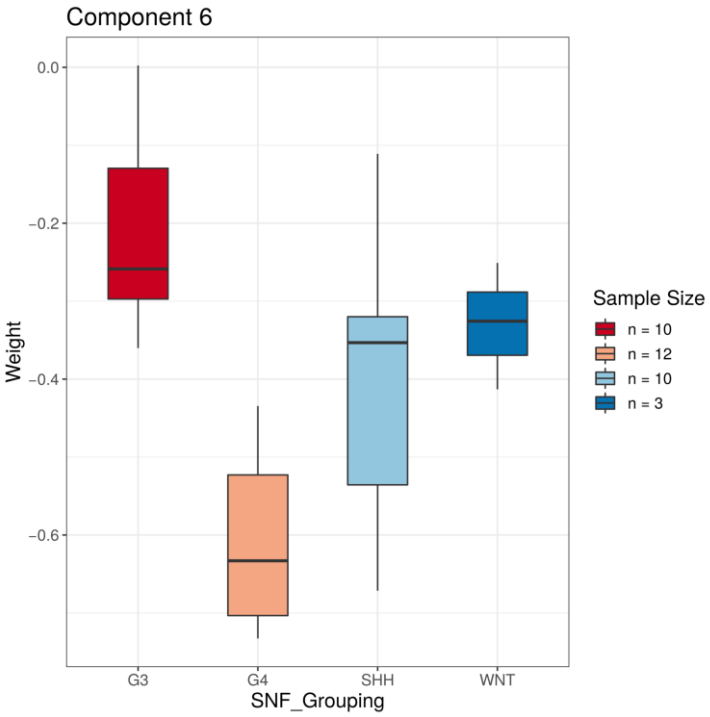


Figure 7: Weights given by tensorICA to each sample. Distributions are shown according to the MB subgroups

Next, we looked at the CBTTC cohort, which gathers a wide range of brain tumors, from which we selected only the four most represented: High and Low Grade Gliomas, Medulloblastoma and Ependymoma, which gave in total gene expression data from 605 patients, as well as proteomics and phosphoproteomics data for 166 and 165 patients respectively (see Figure 1).

Running MOFA on this cohort, the model extracted 9 factors, with the requirement that a factor should explain at least 2% of the variance. Between 50% and 60% of the variance of each layer is explained by these factors as shown in Figure 8. The first factor is explained by the 3 layers, slightly dominated by proteomics and phosphoproteomics, while the second factor, also explained by the 3 layers, is largely dominated by phosphoproteomics data and to a smaller extent by proteomics. Factors 3 and 4 are explained by the 3 layers and factor 5 is explained by gene expression data only.

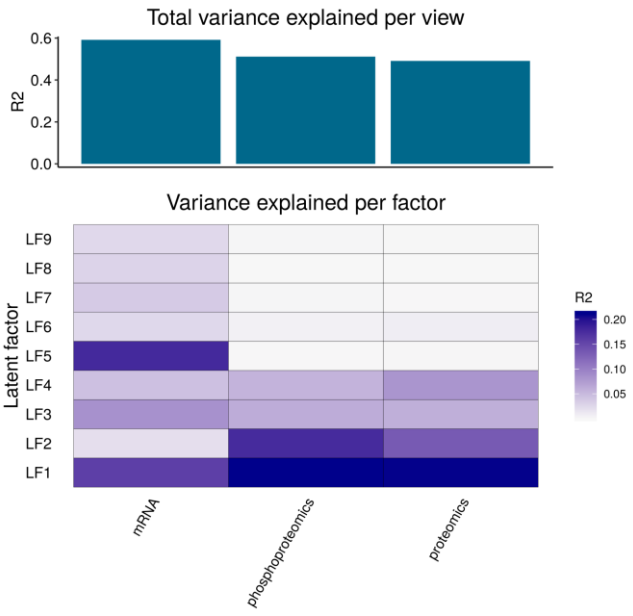


Figure 8:Variance explained by each layer using MOFA on the CBTTC cohort from CHOP

Let us now have a look at the matrix of factors. Factors 1, 3 and 6 allowed to extract, to some extent, MB or MB subtypes from the other tumor types, so we focused on them here. Below we plotted the coordinates of the samples according to factors 1 and 2 ; 1 and 3 ; 1 and 4 and 1 and 6 in Figure 9, Figure 10, Figure 11 and Figure 12. Factor 1 creates a continuum of the four tumor types, from MB, then Ependymoma, followed by HGAT and finally LGAT, even if the tumor types are not perfectly separated. Using factor 3 in addition to factor 1 allows to extract pretty well MB from the other tumor types, while the combination of factor 1 and 6 allows to separate, not perfectly, Group 3 and Group 4 from SHH and WNT subgroups. To illustrate other factors, we show that factor 4 extracts most of the Ependymoma tumors.

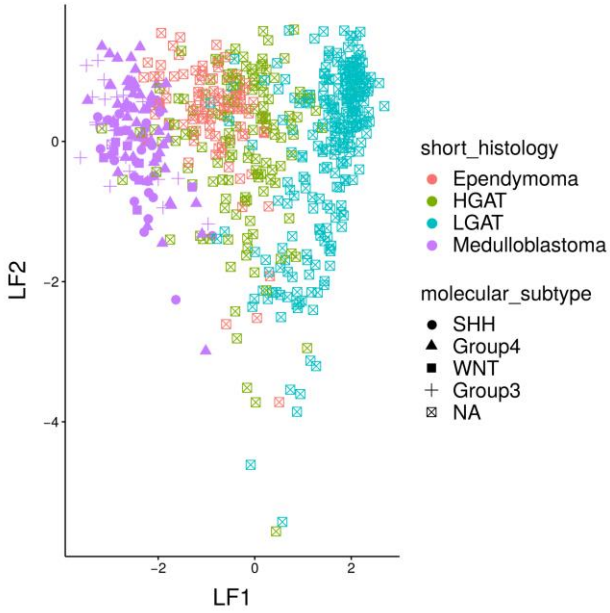


Figure 9 : Projection of the CBTTC samples onto the first two latent factors

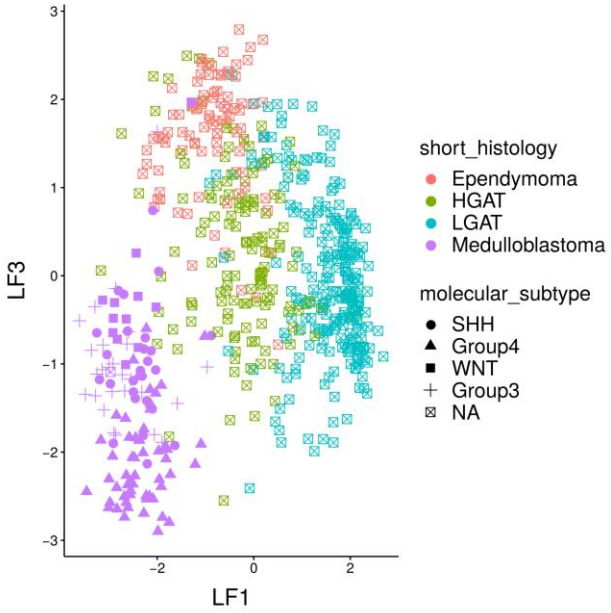


Figure 10 : Projection of the CBTTC samples onto the latent factors 1 and 3

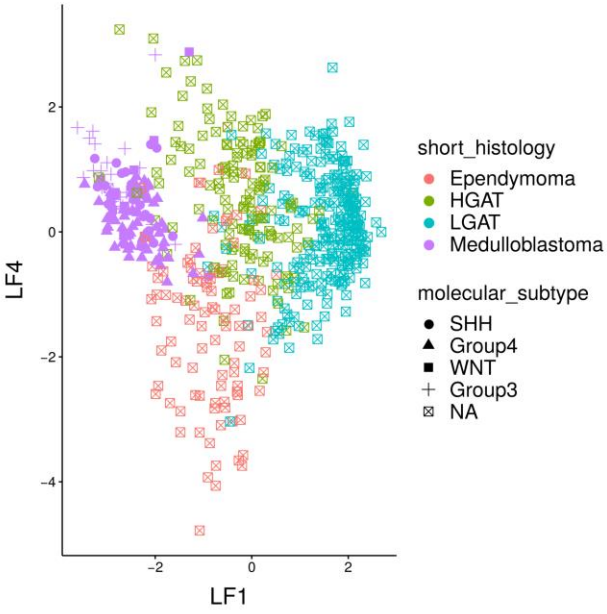


Figure 11: Projection of the CBTTC samples onto the latent factors 1 and 4

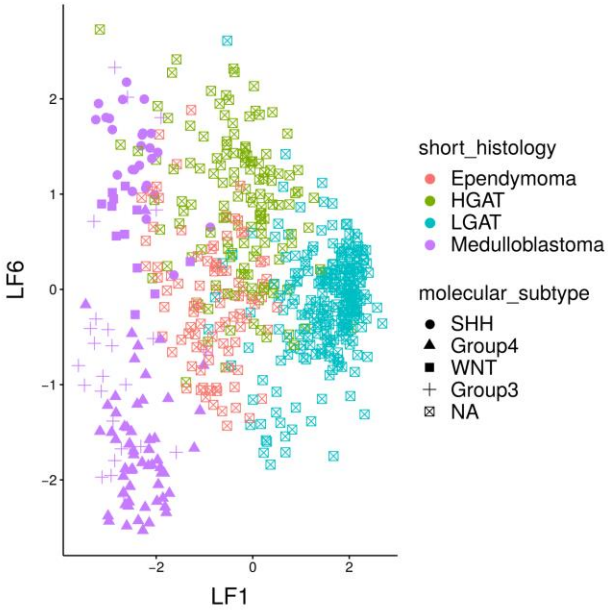


Figure 12: Projection of the CBTTC samples onto the latent factors 1 and 6

Let us now have a look at the weight matrices. We can again look at the features having the highest weights in the different factors but also perform gene set enrichment analysis based on the weights attributed to each gene. Table 4 lists the most significantly enriched pathways for selected factors, for the layers gene expression and proteomics, using Reactome database. Interestingly, the first factor has significant enrichment in pathways representing DNA methylation which was demonstrated to be a powerful approach to classify pediatric brain tumors (Kumar, 2018). Factor 3, which contributes to extract MB from the other tumors, is enriched in ECM organization and RNA processing. Finally factor 6, which separates Groups 3 and 4 from WNT and SHH subgroups, is enriched in translation, which is a key aspect in these subgroups (Forget, 2018).

Factor	GSEA mRNA	GSEA Proteomics
1	DNA methylation, chromatin modifications	TCA cycle, respiratory electron transport
2	Neuronal Synapses	Neurotransmitter Release Cycle
3	ECM organization and RNA processing	Extracellular matrix organization
4	ECM organization and RNA processing	mRNA processing
6	Translation	transcription, translation

Table 4 : Pathway analysis performed on the results obtained using MOFA on gene expression and proteomics layers of selected brain tumors from the CBTTC cohort

4.6 Unsupervised clustering for deciphering intra-tumor heterogeneity in hepatocellular cancers

Hepatoblastoma is one of the tumor types of interest in iPC and we started discussions with partners of iPC working on HB. The group of Carolina Armengol (IGTP) interpreted the results we got from the application of our pipeline as previously described. Pavel Sumazin (BCM) shared with us an ongoing project that we would like to mention as an illustration of deciphering intra-tumor heterogeneity using other approaches than matrix factorization.
Malignant hepatocellular cancers are the most common primary liver malignancies in children and hepatoblastomas (HBs) account for more than two-thirds of these cases. Most HBs respond to chemotherapy and have favorable clinical outcomes, but high-risk HBs have a 3-year overall-survival rate below 50% and guidelines for their classification and particularly the identification of patients that would benefit from aggressive treatments, are still evolving. HB risk-stratification efforts using clinical, histological, and molecular parameters have been reported to be successful in retrospective studies and are being tested in clinical trials. However, risk assessment is particularly challenging for cancers with mixed histologies, including tumors in the recently proposed provisional hepatocellular neoplasm not otherwise specified (HCN NOS) category, which exhibit either intermediate or combined HB and hepatocellular carcinoma (HCC) histological features.
Pediatric hepatocellular neoplasms with HCN NOS histological features were first observed over a decade ago, but only a handful of HCN NOSs have been characterized and little is known about their underlying biology or associated clinical features, including outcomes. BCM molecularly characterized a series of clinically annotated HCN NOSs that demonstrated either intermediate HB/HCC histology (equivocal HCN NOS) or distinct coexisting areas with either HB or HCC histological features (biphasic HCN NOS) in the same tumor. In addition, omics profiling of HBs with focal pleomorphism or anaplasia suggested underlying molecular features previously observed in specific HCC subtypes.

BCM profiled gene expression from 33 tumor samples, including 8 equivocal HCN NOS, 11 biphasic HCN NOS, 10 HBs with focal pleomorphism or anaplasia and 4 HBs of older patients with usual HB histology. Unsupervised clustering of these profiles - together with control profiles of 7 low-risk HBs, 7 non-cancer pediatric liver samples and 4 pediatric non-fibrolamellar HCCs with upregulated WNT-signaling pathway genes and no virus detection - revealed marked differences between the profiles of cancer and non-cancer samples and similarities between the expression profiles of HBC and of low-risk HBs as well as of HCCs. As shown in Figure 13, all non-cancer profiles clustered together, as did all low-risk HBs and all HCCs. Identified clusters were enriched with dysregulated genes from at least 4 cancer pathways. Tumor profiles showed upregulation of WNT-signaling pathway genes. The cluster containing low-risk HBs included cancers with upregulated PI3K-AKT signaling-pathway and cell-cycle genes, while the cluster containing HCCs included cancers with upregulated NF-kB signaling pathway genes. Interestingly the third cluster, composed of profiles of HBCs and older HBs that did not cluster with low-risk HBs or HCCs showed intermediate expression levels of these pathways; see Figure 14. Consequently, the four Figure 13 clusters were named HCC-like, HB-like, Intermediate (the third cluster) and Normal.

In summary, this study suggested that HCN NOS and other high-risk hepatoblastomas share common features with HBs and HCCs, and fill an information gap between the molecular profiles of HBs and HCCs, thus demonstrating that there is a spectrum of disease ranging from HBs to HCCs.

Finally, this kind of data is a nice opportunity to apply our pipeline. In addition, we could add this dataset to the one we already used for unsupervised deconvolution of HB tumors, as soon as this cohort is published.

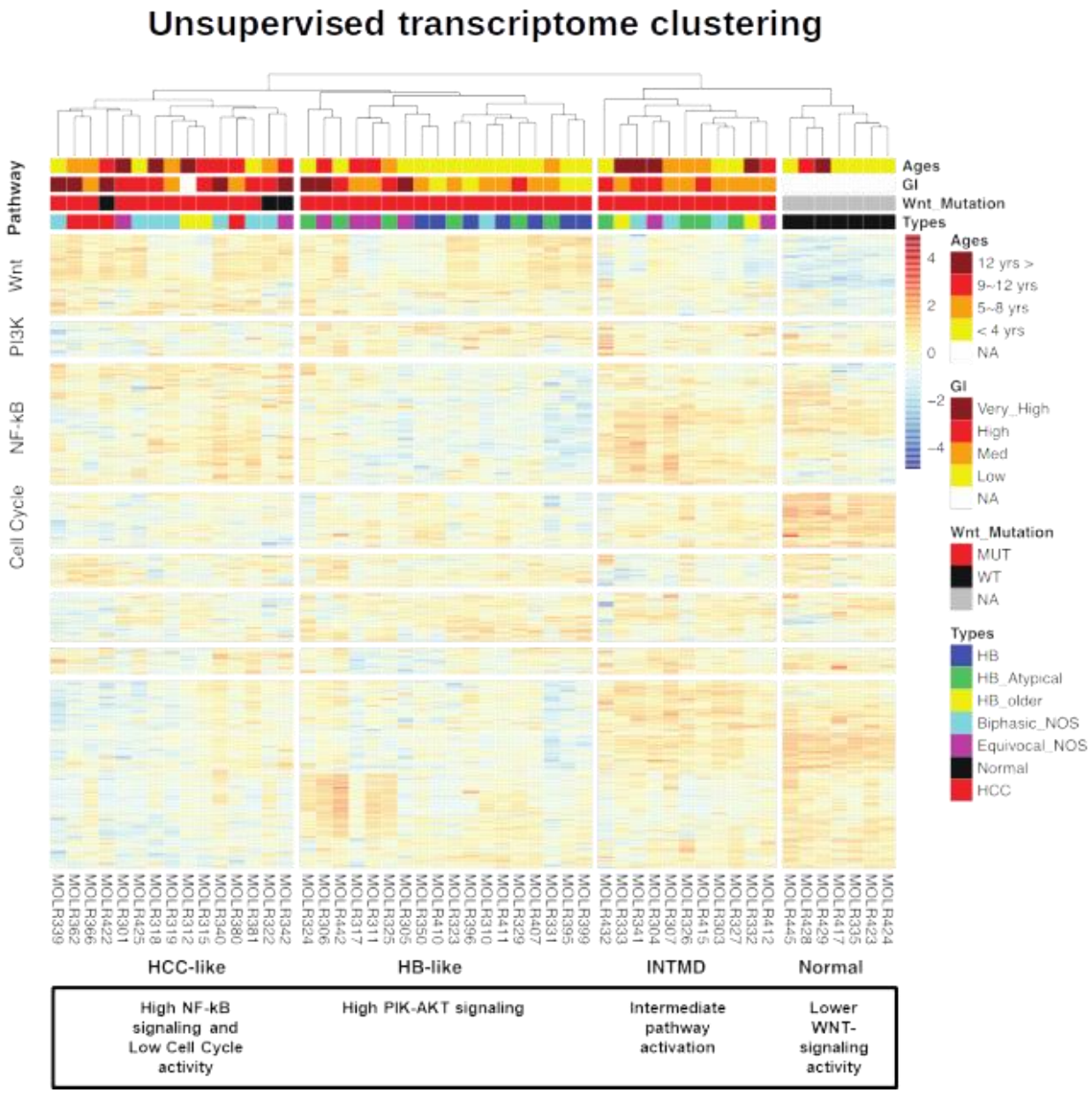


Figure 13: Unsupervised clustering of the expression profiles of non-cancer liver samples, HCCs, low-risk HBs, HCN NOSs and HBs with focal pleomorphism or anaplasia

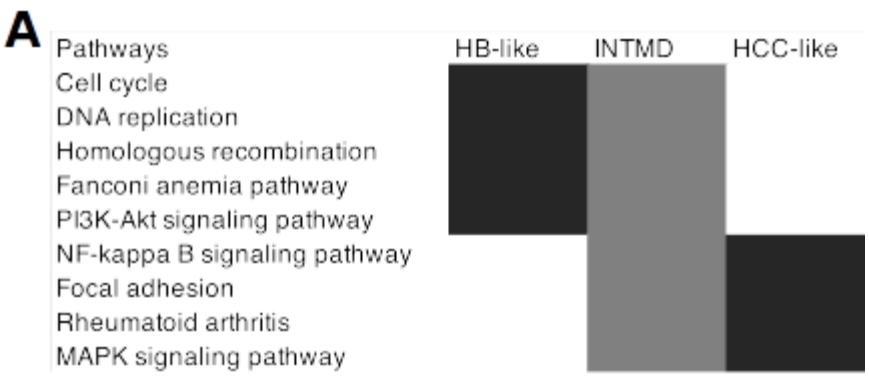


Figure 14: Activated pathways in the identified clusters called HB-like, HCC-like and intermediate

Chapter 5 Conclusions and future work

In this report, we tackled the challenge of the identification of pathways and regulatory elements acting in solid pediatric tumors, namely Ewing sarcoma, Medulloblastoma, Neuroblastoma and Hepatoblastoma. We addressed this fundamental question using various approaches, all based on unsupervised deconvolution. Leukemia is also part of the tumors of interest in iPC. Our main goal was to perform a meta-analysis of different pediatric tumors, but we suspect that mixing solid tumors and leukemia would dilute the findings, this is why we discarded leukemia in a first place. We prioritized the analysis of solid tumors at first, with an idea to analyze available leukemia datasets at the second stage, which has been launched.

First, we performed single layer analysis on each tumor type at gene expression level, allowing to identify some processes acting in these tumors. Then, we took advantage of these results to perform, in a single layer, a meta-analysis across these tumor types at the gene expression level. This led to the identification of common processes acting in these tumors. Next, we performed multi-omics analysis on a specific tumor type, Medulloblastoma, where using only 3 components, it was possible to distinguish the subtypes, in the cohort we studied. Finally, we analyzed multi-omics layers across four pediatric brain tumor types and concentrated on the factors allowing to extract MB or MB subtypes.

These results are available to the consortium through NextCloud at https://data.ipc-project.bsc.es/s/qAWM2oba4zJEtTq. Also, the implementation of our pipeline to perform unsupervised deconvolution of gene expression data is available at: https://github.com/merlevede/iPC_WP3_D3.1. More generally, the users interested in applying ICA on their own datasets can use the jupyter notebook developed in our team by Nicolas Captier: https://github.com/ncaptier/Stabilized_ICA. Finally, the data used in this work comprising 112 datasets is available at: https://data.ipc-project.bsc.es/s/bFC9mPtqPEgfNMn.

Taken together, these results and the methodological developments around them allowed us to fulfil the requirements for this deliverable. We now expected feedback from biologists and clinical partners that are involved in each of these tumor types. We already initiated discussions with Olivier Ayraud (CURIE) about Medulloblastoma and with Carolina Armengol (IGTP) about Hepatoblastoma. This work was a nice opportunity to start new collaborations.

As future work, we would like to investigate in more detail the communities detected in particular in the meta-analysis. For example, we did not yet look specifically into communities shared across pediatric and adult cases or communities specific to one of them. We probably suffer from the small number of adult cohorts used here (only 3). It was difficult to have more as Ewing sarcoma and Neuroblastoma do not have adult counterparts. Another aspect we would like to study more is the association of clinical information with the identified factors, as we mainly focused in this work on the interpretation of metagenes/the gene weights attributed in the identified components.

Finally, we would have liked to have more time to address in depth the question of the presence of immune cells by the use of reference profiles. Indeed, we can perform unsupervised deconvolution on such reference profiles and then use them in a meta-analysis to see to which communities they belong. Nevertheless, it is not straightforward to “name” each component obtained from the deconvolution of reference profiles.

Chapter 6 References

Argelaguet, R. V. (2018). Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular systems biology*.

Armengol, C. C. (2011). Wnt signaling and hepatocarcinogenesis: the hepatoblastoma model. *The international journal of biochemistry & cell biology*.

Avila Cobos, F. V. (2018). Computational deconvolution of transcriptomics data from mixed cell populations. *Bioinformatics*.

Aynaud, M. M. (2020). Transcriptional programs define intratumoral heterogeneity of Ewing sarcoma at single-cell resolution. *Cell reports*.

Becht, E. G. (2016). Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome biology*.

Bork, P. D.-L. (1998). Predicting function: from genes to genomes and back. *Journal of molecular biology*.

Cantini, L. K. (2019). Assessing reproducibility of matrix factorization methods in independent transcriptomes. . *Bioinformatics*.

Cantini, L. Z. (2021). Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. *Nature communications*.

Carrillo-Reixach, J. T.-C.-S.-F. (2020). Epigenetic footprint enables molecular risk stratification of hepatoblastoma with clinical implications. *Journal of hepatology*.

Chen, Z. H. (2017). Inference of immune cell composition on the expression profiles of mouse tissue. *Scientific reports*.

Comon, P. (1994). Independent component analysis, a new concept? *Signal processing*.

Czerwińska, U. K. (2019). *Computational Systems Biology Approaches in Cancer Research.*

Enright, A. J. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic acids research*.

Forget, A. M. (2018). Aberrant ERBB4-SRC signaling as a hallmark of group 4 medulloblastoma revealed by integrative phosphoproteomic profiling. *Cancer cell*.

Hyvärinen, A. K. (2001). Independent component analysis, adaptive and learning systems for signal processing, communications, and control. *John Wiley & Sons*.

Kairov, U. C. (2017). Determining the optimal number of independent components for reproducible transcriptomic data analysis. *BMC genomics*.

Kumar, R. L. (2018). Advances in the classification of pediatric brain tumors through DNA methylation profiling: from research tool to frontline diagnostic. *Cancer*.

Newman, A. M. (2015). Robust enumeration of cell subsets from tissue expression profiles. *Nature methods*.

Paonessa, J. D. (2011). Identification of an unintended consequence of Nrf2-directed cytoprotection against a key tobacco carcinogen plus a counteracting chemopreventive intervention. *Cancer research*.

Pierre-Jean, M. D. (2020). Clustering and variable selection evaluation of 13 unsupervised methods for multi-omics data integration. *Briefings in bioinformatics*.

Strassburg, C. P. (2002). Developmental aspects of human hepatic drug glucuronidation in young children and adults. *Gut*.

Tatusov, R. L. (1997). A genomic perspective on protein families. *Science*.

Teschendorff, A. E. (2018). Tensorial blind source separation for improved analysis of multi-omic data. *Genome biology*.