D7.3

Identification of cell subpopulations in each tumour type, their association with response to therapy, and prediction of effective alternative therapies

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 / D7.3 / V1.0
Work package contributing to the deliverable	WP7
Due date	30th September 2022 – M45
Actual submission date	30th September 2022

Responsible organisation	BCM
Editor	Pavel Sumazin
Dissemination level	PU
Revision	V1.0

Abstract	Tumour decomposition into cells and subtypes and inference about the effects of treatments and perturbations on each tumour component (cell or tumor subclone).
Keywords	Public, inference methods, personalized medicine

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

Editor

Pavel Sumazin (BCM)

Contributors (ordered according to beneficiary numbers)

All iPC partners contributed to this deliverable, with major contributions by Ghent, BCM, IGTP, IBM, and UZH.

Disclaimer

Executive Summary

The primary goal of iPC is to inform the selection of therapies for paediatric cancer patients based on personalized patient data, including demographic, clinical, and molecular data. Towards this aim, iPC has collected data and designed computational models to associate data with disease states, therapeutic responses, and outcomes. Here, we describe efforts to decompose cancers into tumor subclones that convey higher risks for therapeutic resistance and poor outcomes. We describe progress in delineating high-risk subclones in hepatoblastoma, paediatric AML, and neuroblastoma. We also describe iPC-developed technologies that were used to accomplish these tasks.

Table of Content

Executive Summary II

Chapter 1 Introduction 1

Chapter 2 Hepatoblastoma predictive subcolnes 2

Chapter 3 Single-cell to bulk RNA-Seq Deconvolution 6

Chapter 4 Chemoresistant AML subclones 13

Chapter 5 Outcomes-predictive neuroblastoma (NB) subclones 15

Chapter 6 Summary and Conclusion 16

Chapter 7 References 17

List of Figures

Figure 1: Preliminary proposal for a diagnostic algorithm for HBCs. 2

Figure 2: Glypican-3 staining of a section of a high-risk HBC revealed areas enriched for HCC- and HB-like cells forming tumor areas with fetal and embryonal features, and areas with macrotrabecular patterns. Non-tumor cells are Glypican-3 negative. 3

Figure 3: CNAs and histology of a biphasic HBC. 3

Figure 4: Representative histopathology of HBCs with (A) areas of low-risk HB-like cells that are adjacent to high-risk HCC-like cells (4X), and (B) more homogeneous patterns of low-risk HB cells that are co-localized with other HBC cells with high nuclear-cytoplasmic ratios, pleomorphism, and increased mitotic count (10X). 4

Figure 5: Tumor subclones are associated with collections of somatic alterations, each with a unique copy number. 4

Figure 6: Inferred HBC etiology by Chimæra, with a focus on CNAs. 5

Figure 7: HBCs that contains a HB, HBC, and HCC cells. 5

Figure 8: PCA and snRNA profiles of 2 same-HBC (MOLR314) biopsies and derived PDXs that respond differentially to treatment by Cisplatin. 5

Figure 9: Mixture design. 7

Figure 10: Performance comparisons suggested that DWLS deconvolution is the most accurate on our cell mixture assays (top) and on datasets with concurrent bulk RNA-Seq and scRNA-Seq or snRNA-Seq profiles (bottom). 9

Figure 11: Mixture deconvolution with transformed RNA-seq data.. 12

Figure 12: Analysis of 6 paired diagnosis-relapse scRNA-Seq Paediatric AML samples. 14

Figure 13: (A) Analysis of concurrent RNA-Seq and scRNA-Seq profiles of 14 NB identified 15 NB subclones including NB-s1 and NB-S2 (B) whose abundance was predictive of outcomes in TARGET samples that were profiled by RNA-Seq. (C) Significantly upregulated genes in cells included semaphores, HRAS (NB-s1), and MYC targets. 15

Chapter 1 Introduction

iPC researchers have made progress on studying tumor subclones in each of our proposed cancer types. Here, we focus on work in hepatoblastoma (HB), paediatric AML, and neuroblastoma (NB).

We recently proposed a new HB subtype, HB and hepatocellular carcinoma (HCC) histological features (HBC) in order to reclassify some high-risk paediatric liver tumors and to invite therapies that are focused on this subtype (Sumazin et al., 2022). In Figure 1 we describe a diagnostic protocol to identify HBCs in the molecular pathology lab and distinguish them from HBs and HCCs. We used a combination of technologies, including iPC-developed multi-WES strategies (Manica et al., 2020), and single-cell and single-nuclei DNA and RNA sequencing to study HB, HBC, and HCC subclones. Our findings include the identification of chemoresistant subclones and biomarkers that can be used to identify these subclones at diagnosis.

We collected scRNA-Seq and snRNA-Seq samples for each tumor type, with the aim to use these samples to define high-risk subclones and characterize their transcriptomes. Our preliminary tests identified the top-performing methods to deconvolve RNA-Seq profiled samples using the transcriptomes of high-risk subclones, however, the precision of the top methods was poor in some data sets. Consequently, as described in Chapter 3, we developed a method with dramatically improved deconvolution accuracy. In Chapters 4 and 5 we showed that this method can now be applied to test cases in paediatric cancer, identifying regulatory network modules that are predictive of therapeutic resistance and outcomes in paediatric AML and NB. Our work in D7.3 paves the way to developing paediatric cancer atlases that can be used to predict outcomes for patients based on RNA-Seq profiles of their tumors at diagnosis.

Chapter 2 Hepatoblastoma predictive subcolnes

Hepatoblastoma (HB) is one of the most genetically stable types of cancer (Gröbner et al., 2018). However, higher risk HBs and HBs in older patients contain genetic mutations (Sumazin et al., 2017) and display pathway dysregulation that have been more commonly observed in hepatocellular carcinoma (HCC). We recently proposed a liver-cancer subtype (HBCs) that is associated with higher risk and can be described as an intermediate state between HBs and HCC. Our proposed diagnostic protocol for identifying HBCs is depicted in Figure 1 (Sumazin et al., 2022). The algorithm reclassifies some hepatocellular neoplasms into HBCs, and 3 types of HBCs are considered. Biphasic HCN NOS HBCs have discernible tumor areas with distinct HB or HCC features, while equivocal UCN NOS HBCs are characterized by cytological features and growth patterns that were intermediate between HB and HCC. HB FPAs present focal atypical histological features including anaplasia, areas with significantly increased pleomorphism, or prominent macrotrabecular pattern. We note that *Hepatocellular neoplasm not otherwise specified* (HCN NOS) is a recently proposed provisional category of liver cancers (López-Terrada et al., 2014) with diagnostic and treatment protocols that remain evolving. Here, we differentiate between 3 HBC subtypes, including HB FPAs and the 2 HCN NOS subtypes, biphasic HCN NOS and equivocal HCN NOS.

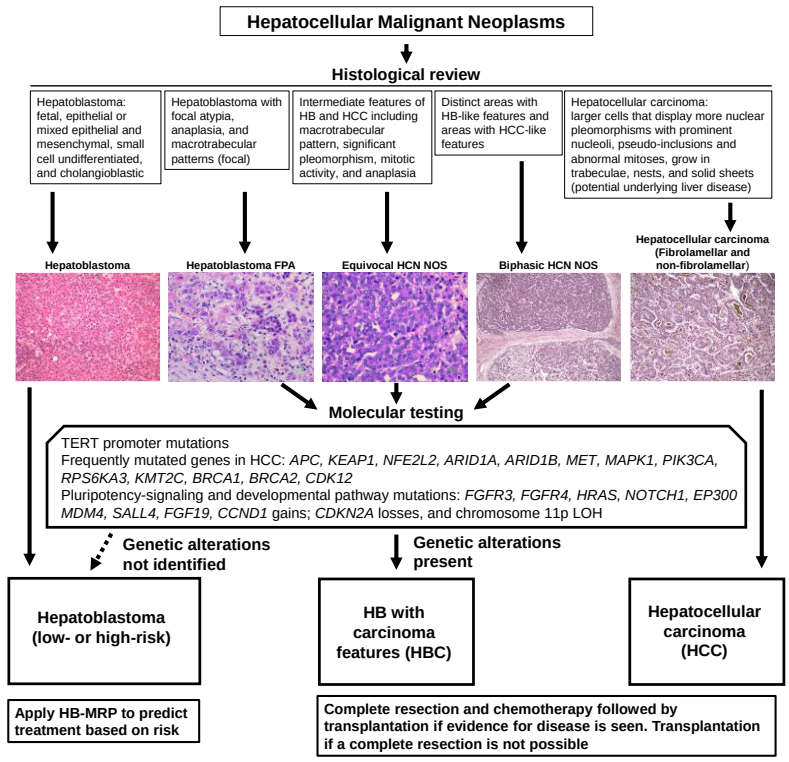

Figure 1: Preliminary proposal for a diagnostic algorithm for HBCs. Hepatocellular malignant neoplasms are categorized by histological review to specific HB subtypes, HB FPAs, equivocal and biphasic HCN NOSs, and HCCs. Fetal HBs are well-differentiated, uniform in size with round nuclei and have minimal mitotic activity. Embryonal HBs have high nucleus to cytoplasm ratios, angulated nuclei, primitive tubules. Small-cell undifferentiated HBs have no architectural pattern, minimal pale amphophilic cytoplasms, and round to oval nuclei with fine chromatin. Cholangioblastic HBs are bile duct cells, usually at the periphery of epithelial islands. Fibrolamellar and non-fibrolamellar HCCs are characterized by larger cells that display more nuclear pleomorphisms with prominent nucleoli, pseudo-inclusions and abnormal mitoses, grow in trabeculae, nests, and solid sheets. HB FPAs have focal atypia, anaplasia, and macrotrabecular pattern. HCN NOSs include cells with features that are commonly associated with both HBs and HCCs (equivocal), or distinct HB-like and HCC-like regions (biphasic HCN NOSs). While the HB and HCC subtypes mentioned can be diagnosed by histology alone, we identified characteristic genetic features that could be used to confirm HBC diagnoses.

The creation of the HBC liver cancer category is intended to promote HBC-specific treatment that would lead to improved outcomes for these patients and to highlight the biology of these transitional cancers. We hypothesize that HBCs are intermediate childhood liver tumors that are partially chemosensitive and may be curable even if the cancer appears to be difficult to fully resect at diagnosis. However, our preliminary work suggests that improved outcomes for HBCs require

aggressive intervention and that HBC patients may benefit from targeted HBC-personalized therapies. The biology of HBCs is of import because they offer a unique snapshot into the transformation of aggressive HBs into HCCs, and we hypothesize that studying HBCs will help map out the diversity of HBs and the evolution of HCCs from HB precursors.

HBCs are composed of multiple tumor subclones

Our recent work suggested that hepatocellular neoplasms that display a combination of HB and HCC histological features (HBC) either across large tumor areas or even focally are associated with poor outcomes. We showed that these cancers are genetically heterogeneous and contain both tumor subclones that molecularly resemble chemosensitive fetal or embryonal HBs and other subclones that are chemoresistant and better resemble HCCs. In our study, aggressive therapies—high-dosage chemotherapy followed by complete resection or transplantation at early treatment stages—significantly (hazard ratio 2.5X-4.3X at 95% confidence) improved outcomes for HBC patients (Sumazin et al., 2022).

This is especially the case for tumors that are visibly multiclonal with either well-defined regions (Figures 2, 3, 4A) or co-localized cells with mixed histologies (Figure 4B). To identify genetically distinct tumor subclones in low- and high-risk tumors, we expanded our preliminary studies using iPC technologies and methods, including Chimæra (Manica et al., 2020). Chimæra predicts tumor phylogenies by comparing the genetic composition of regions across space and time (Figure 6). It deconvolves CNAs, mutation frequencies, mutation-to-subclone associations, and to predict the clonal composition of profiled biopsies from assays that average profiles across cellular ensembles—e.g., RNA-Seq or NanoString profiling of heterogeneous tumors (Figure 3). Chimæra improves on the accuracy of mutation frequency and copy number deconvolution by profiling multiple biopsies from the same tumor. It is designed for tumors with high genomic instability and includes an optimization step to improve mutation-frequency estimates in each biopsy. The advantages of Chimæra analysis over single-cell CNA-profiling are the inclusion of analysis of both mutations and CNAs from the same samples, our ability to profile FFPE with higher accuracy and select multiple tumor regions based on histology, and its (ten-fold) lower cost. When used in tandem, the two approaches are complementary.

To maximize Chimera's benefit, we focused on cancers with histologically distinct regions. Our results revealed that these cancers undergo complex genetic development, from pre-malignant states to HB-like, HBC-like, and sometimes HCC-like states. As an example, we show the inferred etiology of an HBC, as observed by Chimera based on 6 tumor regions and at 2 time points (Figure 6). All tumor cells in these samples were derived from cancer cells with amplifications in chromosomal arms 1q and 2q. Some HB-like cancers

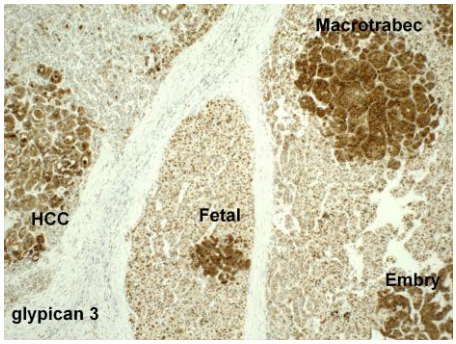

Figure 2: Glypican-3 staining of a section of a high-risk HBC revealed areas enriched for HCC- and HB-like cells forming tumor areas with fetal and embryonal features, and areas with macrotrabecular patterns. Non-tumor cells are Glypican-3

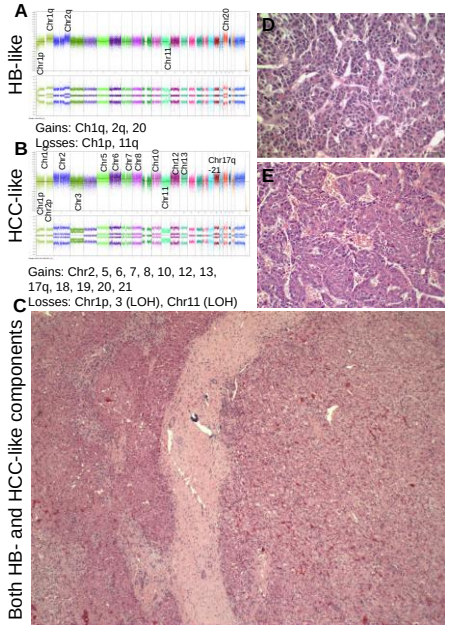

Figure 3: CNAs and histology of a biphasic HBC. CNAs of an (A) HB-like region and (B) an HCC-like region were estimated by Affymetrix OncoScan and suggested a wide-range of chromosomal gains and losses, especially in the HCC-like region. (C) Histology of the HB-like region on the left (also D) and HCC-like region on the right (also E) that has a nodule with frank macrotrabecular pattern and other HCC features.

acquired amplifications in chromosome 8, and these cells transformed to HBCs after acquiring additional mutations, including amplifications of chromosome 17. chr17(+) HBC cells were a small minority at the time of diagnosis, but this cancer population expanded dramatically at the time of transplant and in the metastatic lesion (Figure 6).

In addition, we used single nuclei copy number alteration (CNA) profiles to demonstrate HBC's genetic clonality and hypothesized that genetic clones have differential treatment responses. Indeed, analysis of molecular profiles of patients before and after therapy and same-patient PDXs with divergent responses to treatment suggested the presence of treatment-response predictive expression and genetic alterations. CNA profiles, including in single-cell resolution, and profiles of multiple regions per tumor suggested the presence of high-risk HBC subclones with clonal CNAs.

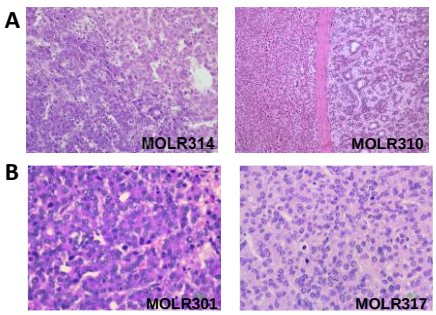

Figure 4: Representative histopathology of HBCs with (A) areas of low-risk HB-like cells that are adjacent to high-risk HCC-like cells (4X), and (B) more homogeneous patterns of low-risk HB cells that are co-localized with other HBC cells with high nuclear-cytoplasmic ratios, pleomorphism, and increased mitotic count (10X).

Single-cell RNA Sequencing identifies differential risk

We relied on a combination of scRNA-Seq and scCNA-Seq assays to identify tumor subclones that are associated with differential risk. In addition, we concurrently evaluated

primary samples and PDXs using scRNA-Seq before and after treatment by the standard of care chemotherapy. The results pointed to the presence of chemoresistant and risk-associated tumor subclones with elevated expression of cMYC, YAP1, and IGF2 pathways. We are currently targeting these pathways in mammalian models to biochemically investigate their effects on chemosensitivity, vascular invasion, and metastasis.

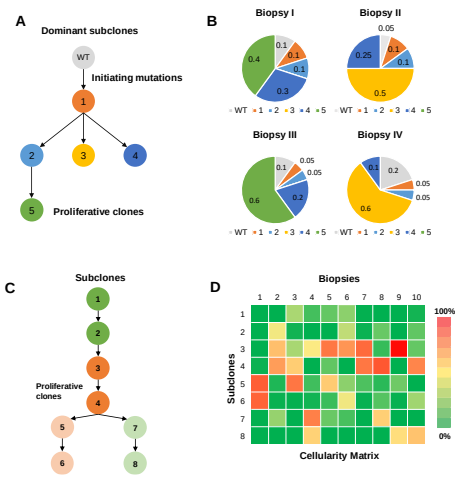

Figure 5: Tumor subclones are associated with collections of somatic alterations, each with a unique copy number. (A) Chimæra infers clonal phylogeny based on (B) WES profiles of multiple biopsies from the same tumor. As an example, (C) Inferred phylogeny from 10 biopsies (regions) was used to estimate (D) the clonal composition of each biopsy.

Our analyses of HBCs suggested that they are high-risk cancers that contain a mixture of both HB and HCC features. We used single-cell resolution profiling to investigate the cellular composition of these tumors and discover whether they are composed of a combination of HB and HCC cells or whether these tumors have HBC cells. Our results revealed that some tumors include HB and HCC cells, as well as cells that cannot be classified as either HB or HCC and contain features of both tumor classes (Figure 7). We concluded that HBC may contain any mixture of these 3 cell types in various proportions. Our results demonstrated that HBCs should be treated as a separate high-risk paediatric liver cancer subtype.

Summary

In total, we studied tumor subclones in samples from 48 HBC, 12 HCC, and 73 HB pediatric patients. We used single-cell resolution assays to study subclones of genetically heterogenous cancers, and matched PDXs and primary tumor samples, including PDX responses to treatments (Figure 8). Our results revealed new tumor biology with implications for liver cancer diagnosis and treatment. Namely, we defined a new subtype of liver cancer and showed that this high-risk subtype is composed of a variety of cancer cells, including high- and low-risk cancer cells. We showed that high-risk cancer cells may be derived from lower-risk cells and that the increase of risk is associated with the acquisition of genetic alterations,

including mutations and CNAs. We showed that HBCs require aggressive treatment and that the outcomes of resected chemo-treated patients is very poor. Instead, the outcomes of HBC patients that were transplanted was 3x better. We identified biomarkers of risk that significantly predict patient outcomes and are differentially expressed across tumor subtypes. In addition, we assembled a panel of biomarkers for diagnosis in the molecular pathology lab. Our investigation results were followed up in PDXs, including PDXs that showed differential responses to chemotherapy (Figure 8). To reveal biomarkers of chemosensitivity and test targeted therapeutics that could improve outcomes for patients with high-risk chemoresistant tumor subclones.

Our study of pediatric liver cancers revealed the potential of technologies using multiple biopsies per patient, as well as single-cell resolution profiling in predicting risk, understanding the evolution of cancers, and identifying the source of resistance to treatment. However, the high cost of profiling and the sample requirements of scRNA-Seq, snRNA-Seq, and scCNA-Seq prevent the use of these technologies on a large scale. Instead, we proposed to comprehensively catalogue cell types for each cancer type that can later be used to deconvolve patient samples and predict cell type compositions based on bulk RNA-Seq profiles of cancer samples.

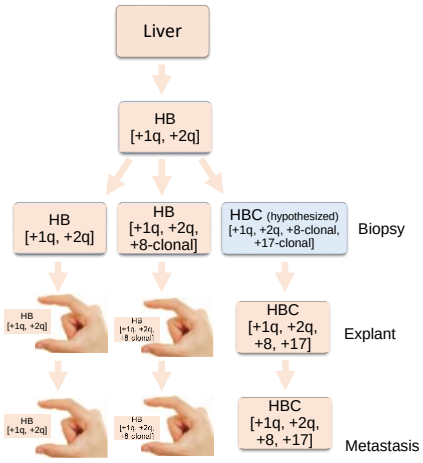

Figure 6: Inferred HBC etiology by Chimæra, with a focus on CNAs.

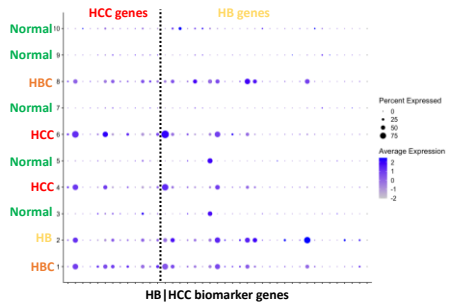

Figure 7: HBCs that contains a HB, HBC, and HCC cells.

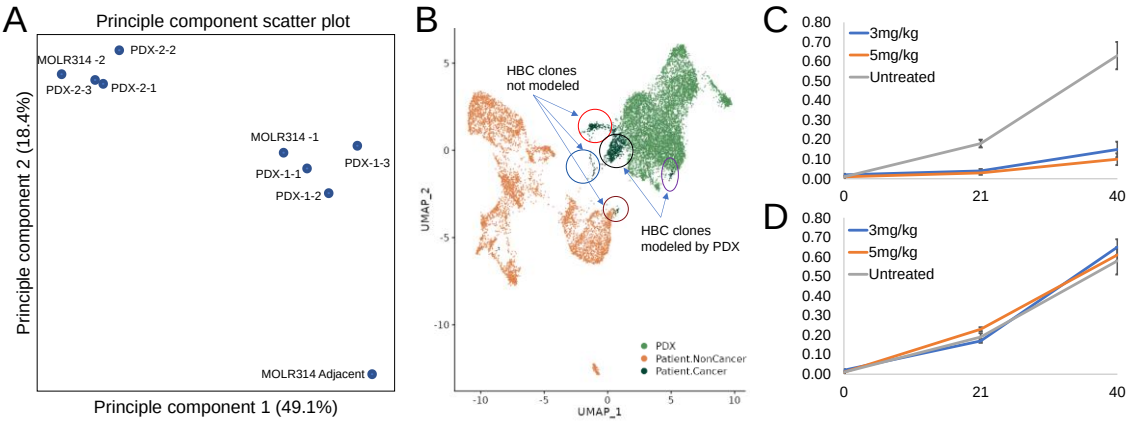

Figure 8: PCA and snRNA profiles of 2 same-HBC (MOLR314) biopsies and derived PDXs that respond differentially to treatment by Cisplatin.

(A) PCA of RNA-Seq profiles of 2 biopsies and PDXs derived from them, and a normal adjacent sample, suggested gene expression differences between the biopsies and their derived PDX clusters (designated as -1 and -2). (B) Single-cell RNA profiles of a patient biopsy and PDX. 80% of the patient cells were normal, but almost 2,000 cancer cells were profiled, revealing multiple HBC clones, including clones that were and weren't modelled by the PDX. (C) Growth of the orthotopic same-HBC PDXs after Cisplatin treatment in 10-day cycles suggested the presence of both chemosensitive and (D) resistant HBC subclones.

Chapter 3 Single-cell to bulk RNA-Seq Deconvolution

Single-cell and single-nuclei RNA-sequencing (scRNA-Seq and snRNA-Seq) technologies have revolutionized our ability to accurately quantify various cell types and states in healthy or diseased tissue biopsies. Based on cell-type specific gene expression signatures, individual cells can be labeled and enumerated, allowing comparisons of cell fractions between tissue samples and associated changes in cell fractions with relevant biological or clinical parameters. For example, scRNA-seq analyses of tumor biopsies have revealed differences in various immune cell types and states between patients that respond to immune checkpoint inhibition therapies and patients that do not. Such findings could reveal novel therapeutic applications or provide us with biomarkers to improve therapy response prediction.

While these methods provide cell-type specific information at unprecedented resolution, they also come with several challenges that prevent efficient implementation in a clinical setting. One major challenge is the high per sample cost for library preparation and sequencing when relying on commercially available solutions. In addition, there are stringent requirements associated with sample collection, processing, and storage as to avoid technical biases in the gene expression profiles. For example, the temperature at which tissues are dissociated in single cell suspensions can impact stress response in the cells, and subsequent changes in gene expression. Moreover, cryopreservation of dissociated cells has been shown to result in a loss of epithelial cell types (Denisenko et al., 2020).

Bulk RNA-Seq methods are less impacted by these factors, yet they only provide an averaged gene expression profile of the entire tissue sample. Over the past years, numerous computational methods, collectively referred to as deconvolution methods, have been developed to infer cell type proportions from bulk RNA-Seq profiles using a reference matrix composed of cell-type specific gene expression signatures (Avila Cobos et al., 2020; Decamps et al., 2021). While some of these methods only rely on the reference matrix to enumerate cell type proportions, others make use of scRNA-seq data from the same tissue type as the bulk RNA-Seq profiles to perform deconvolution. In various benchmarking efforts, we and others have shown that different factors impact the performance of deconvolution methods, including data transformation, data normalization and the composition of the reference matrix. While these studies have been informative to understand which factors impact performance, they have failed to quantify the absolute performance of deconvolution methods compared to scRNA-Seq, which is considered the gold standard.

We envision using single-cell profiling assays of each tumor type to categorize cell types for these tumors and create a tumor-specific catalog with clinical predictions associated with cell types. This catalog could then be used to predict sample composition using deconvolution of its bulk RNA-Seq profile. The composition of samples will help select therapies to target all cancer cell types present in each tumor. In order to achieve this goal, we needed to evaluate deconvolution methods. In total, we evaluated 6 deconvolution methods in 8 different datasets, each composed of samples that were profiled by both bulk RNA-Seq and scRNA-seq or snRNA-seq. One of these datasets was composed of artificial cell-line mixtures where expression and abundance are well characterized, and where both relative and absolute quality of deconvolution efforts can be evaluated. By comparing observed cell fractions (derived from deconvolution of bulk RNA-Seq profiles) to expected cell fractions (derived from scRNA-seq or snRNA-Seq profiles of the same samples), we were able to accurately quantify the absolute performance of each deconvolution method. Our results highlight consistent performance differences between methods across datasets and reveal a dramatic performance improvement when transforming bulk RNA-Seq profiles to scRNA-seq spaces before applying deconvolution. Our conclusions led to the development of a new deconvolution method, which produced improved results for all our tests.

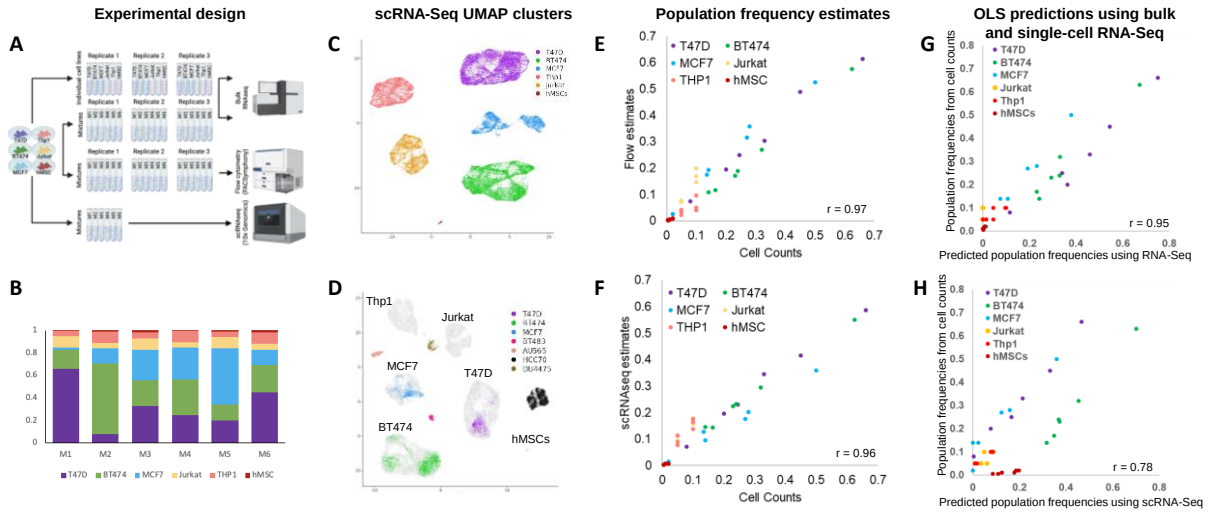

Figure 9: Mixture design.

(A) Breast cancer cells (BT474, T47D, and MCF7), leukemia cells (THP1 and Jurkat), and human mesenchymal stem cells (hMSCs) were cultured and used to generate 6 mixed cell populations. Each cell line was profiled individually by bulk RNA-Seq in triplicates, and each mixture was profiled by bulk RNA-Seq and flow cytometry in triplicates as well as by a 10x genomics Chromium controller to generate 6 scRNA-Seq datasets. (B) Mixtures were composed of varying proportions of cancer cells, with leukemia cells accounting for 15% and hMSC accounting for 0.5% (M1 and M4) to 2% (M3 and M6) of cells. (C) Clusters of cells in mixtures profiled by scRNA-Seq data corresponded to composing cell types, as identified based on biomarker expression profiles. (D) The integration of scRNA-Seq profiles of our mixtures (in gray) and scRNA-Seq profiles of BT474, T47D, MCF7, BT483, AU565, HCC70, and DU4475 (Gambardella et. al., 2022) revealed a significant overlap between profiles of BT474, T47D, and MCF7 cells but not BT483, AU565, HCC70, and DU4475 cells and our inferred BT474, T47D, and MCF7 clusters, respectively. (E) Cell counts at the time of mixture generation were significantly correlated with cellular compositions, as estimated by flow cytometry (r=0.97) and (F) by scRNA-Seq clusters (r=0.96). However, the correlation between cellular composition estimates by flow cytometry and scRNA-Seq clusters was significantly lower (r=0.92, p<0.05 by Fisher's transformation). (G) Ordinary least squares regression (OLS) deconvolved the composition of our mixtures with high accuracy (r=0.95). (H) OLS also deconvolved our mixtures using cell-type expression estimates from scRNA-cluster profiles with significant accuracy (r=.72, p<1E-4), but because of inaccuracies in scRNA-Seq derived expression profiles, OLS accuracy was significantly (p<1E-5) lower than when using bulk RNA-Seq profiles of each cell type.

Cell mixtures used to evaluate scRNA-Seq accuracy

While concurrent bulk RNA-seq and scRNA-seq assays can be used to evaluate deconvolution accuracy, they lack controls for both true composition and cell-type expression estimates. Namely, divergent estimates from the two assays cannot be resolved, and technical analysis errors may not be identified due to missing information. Consequently, accurate and fully resolved deconvolution-strategy evaluations require fully characterized datasets, where the expression profiles and composition of each cell type are known with high degrees of accuracy. Towards this aim, we developed a solid tumor model that includes multiple solid-tumor cell types, immune cells, and lower-abundance stem cells. Following this model, we established *in vitro* cell mixtures that are composed of varying proportions of cells from 3 breast cancer lines (T47D, BT474, MCF7), monocytes (Thp1), lymphocytes (Jurkat), and stem cells (hMSC).

Mixture composition was estimated using input cell counts. Cells from each cell line and each mixture were profiled by bulk RNA-seq in triplicates. Mixtures were profiled by flow cytometry to

independently evaluate their composition and by scRNA-seq (Figure 9A, Supplemental Table 1). The proportions of breast cancer cell lines varied across mixtures, with some mixtures composed predominantly of one cell type (e.g., 66% of Mixture 1 were T47D cells) and others having a balanced composition (e.g., Mixture 4). Monocytes and lymphocytes accounted for 15% of the mixtures, and hMSCs abundance varied from 0.5% to 2% (Figure 9B).

UMAP analysis of mixture scRNA-seq profiles verified the existence of 6 clusters with biomarkers that correspond to their 6 composing cell types (Figure 9C). We confirmed breast-cancer cell type identities by integrating 7 scRNA-seq profiles of breast-cancer cell samples(Gambardella et al., 2022) including T47D, BT474, MCF7 and 4 cell lines that were not used in our mixtures (BT483, AU565, HCC70, DU4475); see Figure 2D. Cellular composition estimates based on absolute cell counts that were determined when assembling the mixtures showed high correlations with composition estimates by flow cytometry and scRNA-seq (r=0.97 and r=0.96 respectively; Figure 10E,F). However, the correlation between estimates by flow cytometry and scRNA-Seq clusters were significantly lower (r=0.92, p<0.05 by Fisher's transformation). This suggested composition estimates by cell counts are the most accurate, and that flow cytometry and scRNA-Seq introduce partially independent errors to composition estimates. Overall, however, these results confirmed the mixture composition as estimated by cell counting and demonstrate that it is reflected in scRNA-seq data with good accuracy.

Unlike scRNA-Seq based mixture composition estimates, which had high correlation with composition estimates by cell counts (r=0.96, Figure 9F), the correlation between scRNA-Seq based expression profiles of individual cell types (clusters) and bulk expression profiles of these cells were not as high. Pearson correlation of the profiles of T47D, BT474, and MCF7 cells and their respective bulk RNA-Seq profiles were r=0.53, r=0.53, and r=0.55, respectively; Jurkat and Thp1 had correlations of r=0.66 and r=0.63, respectively; hMSCs, which were the least abundant cells in each mixture, were correlated at r=0.16 with their bulk RNA-Seq profiles. Restricting comparisons to the top expressed genes did not improve these correlations.

To evaluate the effect of expression-estimate inaccuracies on the quality of deconvolution, we tested the predictive accuracy of ordinary least squares regression (OLS) in predicting mixture composition from its bulk profiles and using either scRNA-Seq or bulk-derived expression profile estimates for each cell type. OLS is used to solve a simple set of linear equations that seeks to find the optimal composition P of a set of mixtures with bulk RNA profiles Z to minimize the difference between the observed bulk RNA-Seq profiles and the abundance-weighted sums of the expression profiles of composing cell types X. Namely, given bulk expression profiles $\{z_i \in Z\}$ of each mixture, and expression estimates for each cell type j $\{x_j \in X\}$, we see to identify $p_{ij} \in P$ across all mixtures i and cell types j to minimize Equation 1.

$$\min_P \sum_i |z_i - \sum_j p_{ij} x_j|^2 \text{ for mixture } i \text{ and cell type } j \qquad \textbf{Equation 1}$$

Our results suggested that OLS can estimate mixture composition with high accuracy when input expression profile estimates are accurate. Namely, using bulk RNA-Seq profiles of each cell type, OLS composition predictions had Pearson correlations of r=0.95 with mixture composition estimates by cell counts (Figure 9G). However, when using scRNA-Seq based expression profile estimates of the cell types, this correlation declined to r=0.78 (Figure 9H). Note that the correlation r=0.78 is significant at p<1E-5, suggesting that, overall, OLS can accurately predict composition in our mixtures using scRNA-Seq based expression estimates. However, as expected, hMSC composition estimates were the least accurate (Figure 9H).

Having confirmed the quality and validity of the in vitro cell mixtures and associated data, we applied our deconvolution benchmarking framework to the bulk RNA-Seq and scRNA-seq data of the 6 mixtures (Figure 3A). We observed substantial differences in performance (i.e. observed versus expected fractions) between deconvolution methods, with DWLS outperforming the other 5 methods, irrespective of the bulk RNA-seq and scRNA-seq normalization strategy. Overall, normalization of the bulk RNA-seq data with TPM resulted in better performance compared to TMM, LogNormalize or no normalization. Normalization of the scRNA-seq derived reference matrix did not impact performance. All methods performed poorly on the hMSC cells, which were present at fractions of 1% or lower in each of the mixtures. All methods also underestimated the fraction of Jurkat cells in several mixtures, but this was most pronounced for CIBERSORT, FARDEEP, RLR and NNLS. MuSiC also underestimated the fraction of THP1 cells. Together, these observations demonstrate

that, in an ideal setting, with concordant scRNA-seq and bulk RNA-seq, deconvolution with DWLS leads to the most accurate cell type proportions estimates, as depicted in Figure 10. However, DWLS's accuracy on some tissues, and especially on cryopreserved tissues was poor, producing composition estimates that are too inaccurate to be used in practice. To overcome this, we set out to develop alternative methods that integrate strategies across multiple past deconvolution efforts.

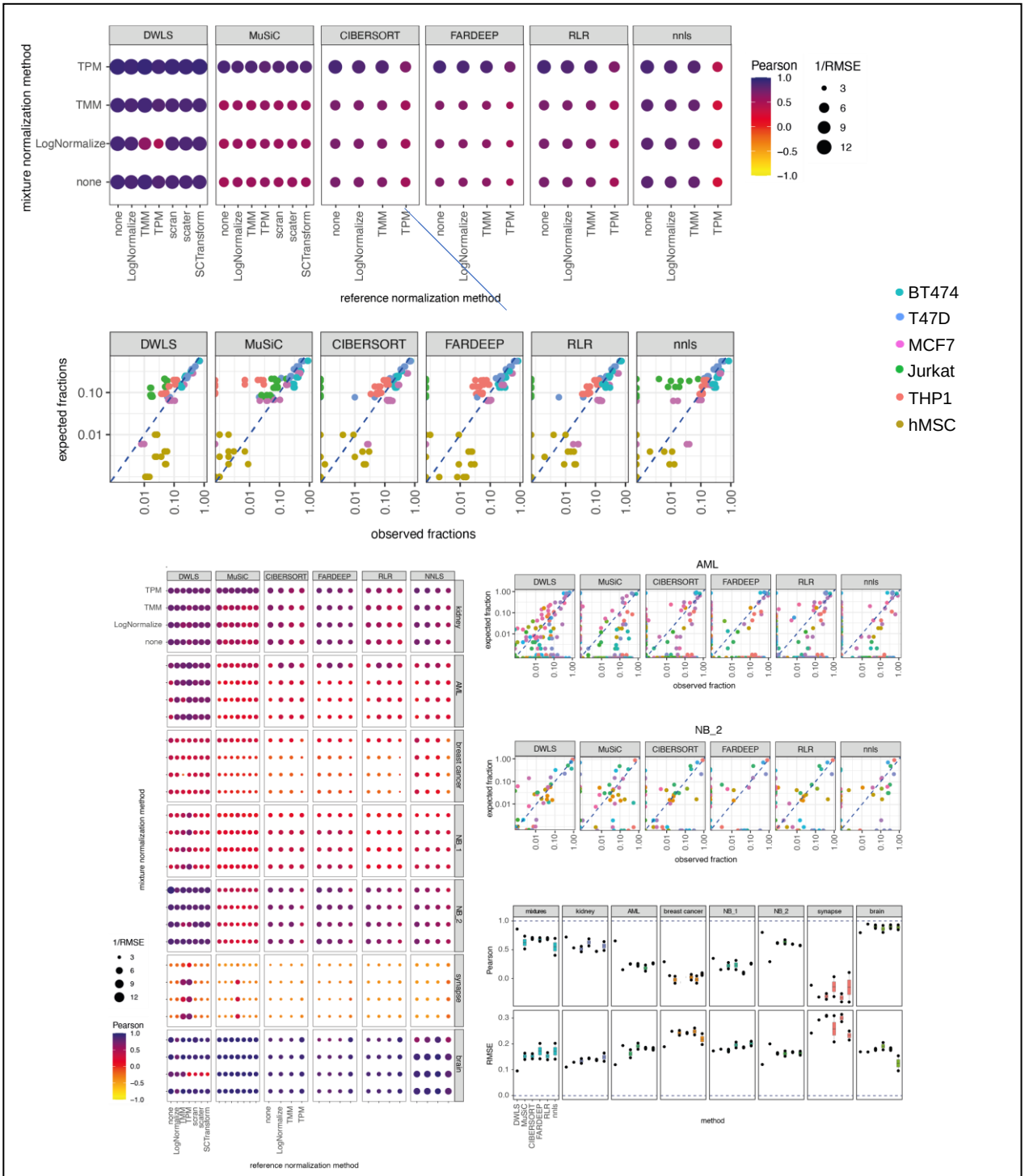

Figure 10: Performance comparisons suggested that DWLS deconvolution is the most accurate on our cell mixture assays (top) and on datasets with concurrent bulk RNA-Seq and scRNA-Seq or snRNA-Seq profiles (bottom). However, even when outperforming other methods DWLS's performance was poor on some datasets.

Hybrid bulk RNA-Seq transformation and optimization with Janus

Janus is a conversion-dampened weighted least squares strategy to transform and deconvolve bulk RNA-seq data into scnRNA-seq vector spaces. Similar to Bisque (Jew et al., 2020), Janus learns a simple linear transformation from concurrent bulk RNA-seq and scnRNA-seq profiles.

Namely, given bulk RNA-seq profiles Z and concurrent pseudobulk scnRNA-seq derived profiles $\hat{Z}$ of the samples in the dataset, then the bulk RNA-seq expression profile of each gene g that is expressed in both the bulk and scnRNA-Seq profiles is mapped to its pseudobulk profile according to Equation 2; $\hat{z}_{g,i}$ and $z_{g,i}$ are the pseudobulk and bulk profiles of gene g in sample i, respectively, and the coefficient a_g and constant b_g form the linear transformation for each gene g.

$$\underset{a,b}{\operatorname{argmin}} \sum_i \left(\hat{z}_{g,i} - \left(a_g z_{g,i} + b_g\right)\right)^2 \qquad \textbf{Equation 2}$$

This linear transformation is applied to all other bulk RNA-seq profiles to transform them to scnRNA-seq space. The transformation minimizes the deviation between a sample's pseudobulk and bulk RNA-seq profiles by mapping the bulk RNA-seq expression profile of each gene to the magnitude and deviation of pseudobulk scnRNA-seq values. Equation 2 applies naturally when converting bulk RNA-seq profiles with no concurrent scnRNA-seq profiles. However, when testing deconvolution on our datasets, which included concurrent bulk RNA-seq and scnRNA-seq profiles for each sample, we used a leave-one-out strategy. Here, the linear transformation was optimized using all but one sample and was then used to transform the bulk RNA-seq profile of the remaining sample. This transformed profile was then used to predict the composition of the sample with a dampened weighted least squares strategy like DWLS (Tsoucas et al., 2019). Deconvolution performance was determined using cell counts for our cell mixtures and estimates from single-cell profiles for patient samples with concurrent bulk and scnRNA-seq profiles. Cell counts are the most accurate and unbiased estimates for our cell mixtures, and single-cell estimates are our only estimates for the true composition of patient samples.

We note that the proposed simple linear transformation in Equation 2 is one of many. Indeed, Bisque proposed an alternative transformation that could be used more generally. We tested the following formulation, and it performed equivalently to that of Equation 2. Let $\overline{\hat{z}_g}$ denote the average expression estimate of gene g in pseudobulk profiles $\hat{Z}$ and $\overline{z_g}$ the average expression of this gene in all bulk RNA-seq profiles—including the matching and other bulk RNA-seq and profiles Z, and let $\widehat{\sigma_g}$ and σ_g denote their respective standard deviations. Then the transformed profile for gene g in sample i ($\vec{z}_{g,i}$) is given in Equation 3. This formulation does not require concurrent RNA-seq and scnRNA-seq profiling.

$$\vec{z}_{g,i} = \overline{\hat{z}_g} + \frac{\widehat{\sigma_g}}{\sigma_g}\left(z_{g,i} - \overline{z_g}\right) \qquad \textbf{Equation 3}$$

Following transformation using Equations 2 or 3, Janus adopts a simplification of DWLS's strategy to deconvolve transformed bulk profiles. Janus doesn't require signature gene selections, and instead uses all genes with nonzero expression in both the transformed bulk and scnRNA-seq profiles. The objective function is identical to the one employed by OLS (Equation 1), however, here, the Janus process seeks to identify $\tilde{p}_{ij} \in \tilde{P}$ that minimizes the discrepancy between transformed bulk RNA-seq profiles $\vec{Z}$ and the abundance-weighted sums of the expression profiles of composing cell types X. Consequently, following the iterative process proposed by Tsoucas et al., Janus minimizes this dampened weighted discrepancy until convergence is reached at iteration l, so that $\left\|\tilde{P}^{(l)} - \tilde{P}^{(l-1)}\right\| \leq 0.01$.(Tsoucas et al., 2019).

Performance evaluation of deconvolution methods

DWLS consistently outperformed other deconvolution methods in our tests. However, its accuracy was poor in several datasets, limiting its potential applications. Note that lower accuracy may be due to method-independent factors, including physically different cellular compositions between scnRNA-Seq and bulk RNA-Seq samples, and technical differences in sample processing that results in diverging estimates. Most importantly, deconvolution accuracy is dependent on accurate gene expression estimates, and—as is the case for our cell mixtures—scnRNA-Seq-derived gene expression profiles may be imprecise. We showed that OLS-based deconvolution using bulk RNA-Seq profiles of each cell type (Figure 9G) produced more accurate results than deconvolution using scRNA-seq-derived profiles (Figure 9H). Similarly, deconvolution with DWLS using bulk RNA-Seq profiles of each cell type was in excellent agreement with mixture composition as estimated by cell counts (Figure 11A), and its performance declined when using scRNA-seq-derived profiles (Figure 11B). However, in both cases, deconvolution with DWLS was more accurate than OLS: r=0.98 vs. r=0.95 when using bulk RNA-Seq profiles, and r=0.93 vs. r=0.78 when using scRNA-seq-derived profiles, respectively. However, deconvolution with Janus—employing linear bulk RNA-Seq transformation followed by dampened weighted least squares—further improved deconvolution accuracy (r=0.95, Figure 11C).

To systematically test the benefit of bulk transformation and deconvolution with Janus, we compared the performance of Janus, DWLS, and OLS on our cell mixture set, paediatric AML, FAC2004 NB (NB1), Jansky NB (NB2), ROSEMAP brain, breast cancer, and kidney profiles using a leave-one-out cross-validation strategy. Namely, iteratively, concurrent RNA-Seq and scnRNA-Seq profiles of all but one of the samples were used to predict the composition of the remaining sample based on its bulk RNA-seq profile (Figure 11D). We compared composition estimates by the three deconvolution prediction methods and estimates based on cell counts (for cell mixtures) or from scnRNA-Seq (all other samples), denoting both their Pearson correlation and RMSE. Our results demonstrated consistently and significantly improved prediction accuracy with Janus, resulting in improved correlations and reduced error for each sample tested.

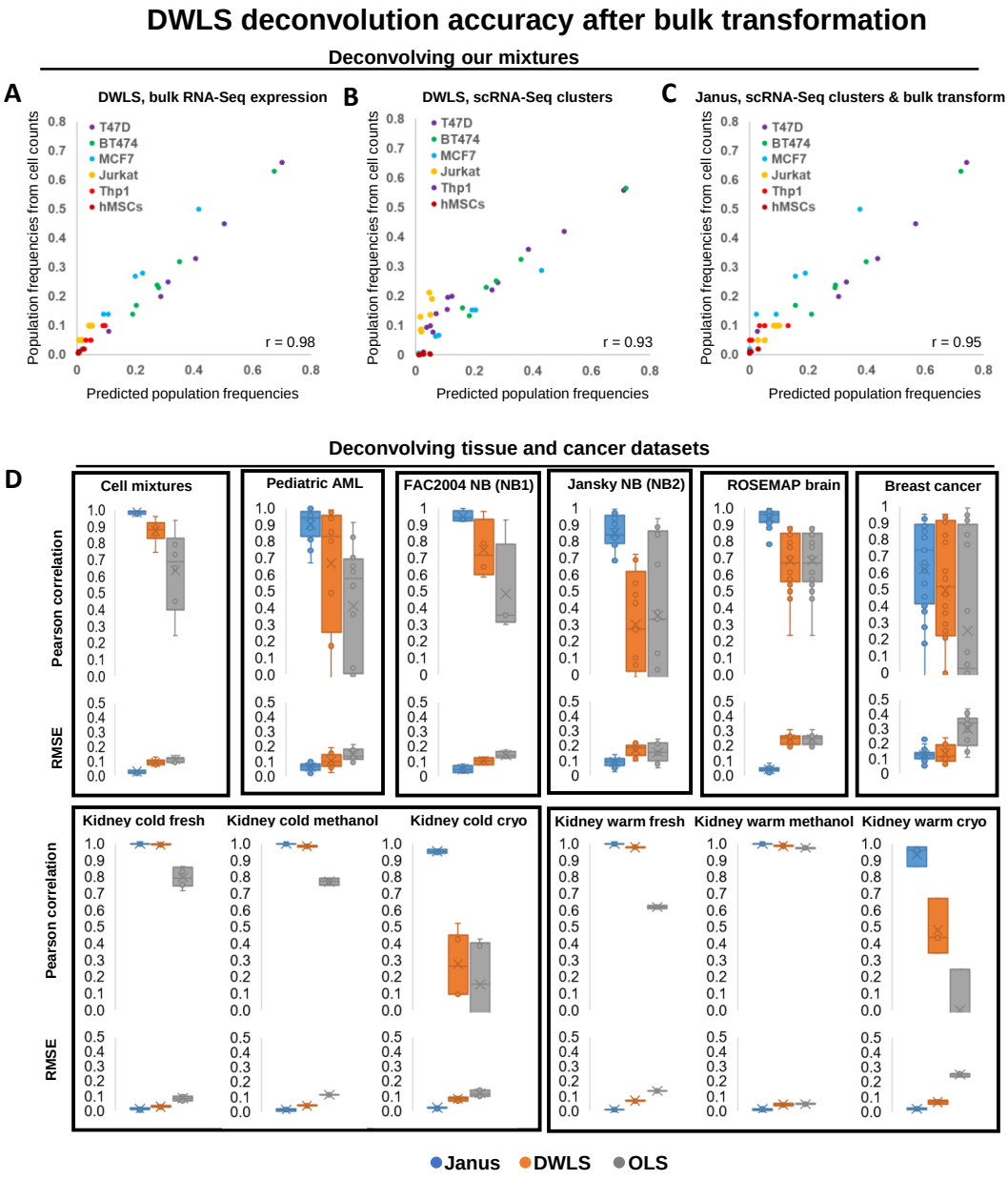

Figure 11: Mixture deconvolution with transformed RNA-seq data. (A) DWLS deconvolved the composition of our mixtures with near-perfect accuracy when given the bulk RNA-seq expression profiles of each cell type (r=0.98), and (B) with high accuracy when using cell-type expression estimates from scRNA-cluster profiles (r=0.93). (C) Bulk RNA-seq transformation significantly improved DWLS deconvolution accuracy when using cell-type expression estimates from scRNA-cluster profiles (r=0.95, p<2E-4). (D) We compared the accuracy of predicted composition estimates by OLS, DWLS, and DWLS following bulk RNA-seq transformation on datasets with concurrent bulk and scRNA or snRNA expression profiling, using Pearson correlation and RMSE.

Chapter 4 Chemoresistant AML subclones

Children with acute myeloid leukemia (AML) are not benefiting from advances in personalized treatment and immunotherapy because there are few known disease-driving mutations and AML-specific antigens for targeting. Most children with AML are still treated with traditional chemotherapy and often stem cell transplant which carries a large risk of toxicity, and still 40% of children will fail this approach and die of relapse (Gamis et al., 2014; Rubnitz et al., 2010). The relapsed disease is often attributed to a subset of cells that withstand chemotherapy and eventually resurge. Clinically, this phenomenon manifests as “measurable residual disease” (MRD), which refers to residual AML cells detected in the bone marrow after chemotherapy. About 30% of paediatric AML patients have MRD detected by flow cytometry after one cycle of treatment, and this finding is strongly associated with future relapse (Loken et al., 2012).

Intrinsic and extrinsic survival mechanisms contribute to MRD. For example, AML cells that survive cytarabine treatment demonstrate distinct intrinsic metabolic dependencies (Farge, Boyd, Konopleva), as well as environment-induced anti-apoptosis gene and protein expression profiles. Equally important to surviving drug toxicity is escaping immune surveillance. Several mechanisms by which AML cells evade immune cell eradication have been described, including checkpoint protein expression and immunosuppressive cytokine secretion. T cells, NK cells and monocytes in the AML bone marrow niche all demonstrate a loss of anti-AML reactivity. Understanding the intrinsic and extrinsic mechanisms that allow an AML cell to survive chemotherapy and immune eradication will pave the way for new rational therapies to counter these mechanisms.

We postulated that individual AML cases contain subpopulations with enhanced chemoresistance and immune evasion capacity and that these cells can often be detected as MRD after the first cycle of chemotherapy. We have constructed regulatory interaction graphs based on public RNA-Seq data, and have used several high-dimensional approaches, including CITE-seq and single-cell RNA-seq, to compare AML subpopulations between diagnosis and relapse. We have identified surface markers and gene expression programs that are enriched in subpopulations that expand between diagnosis and relapse, suggesting they represent potential targets for eradicating chemoresistant AML. We applied our biological and computational expertise to analyze end-induction bone marrow samples by CITE-seq. Distinguishing a rare population of malignant myeloid cells from recovering normal stem and progenitor cells at the single-cell level can be challenging, but if successful it would be immensely informative.

In total, we profiled paired diagnostics samples before treatment and at relapse for 6 patients; pre-treatment AML samples are enriched for chemosensitive cancer cells, while relapse AML samples are enriched for chemoresistant cancer cells(Rasche et al., 2018). We then used cell types and expression profiles from these scRNA-Seq data to deconvolve diagnostic samples from a total of 181 paediatric AML (Bolouri et al., 2018) patients that were profiled by the TARGET consortium. Paired diagnostic-relapse paediatric AML samples were collected with the objective to identify chemoresistant tumor subclones. After integration and clustering (Figure 12A), we sought to identify AML subclones (clusters) that are present before treatment and expand at relapse. One AML cluster was the only predicted AML subclone that included diagnostic and relapse cells from at least half of the patients and expanded at relapse. We refer to this subclone as the AML expanding subclone, or AML-X for short (Figure 12B). We used Janus, DWLS, and CIBERSORT to predict the composition of TARGET AML samples, and then used the predicted abundance of AML-X cells in each diagnostic sample for survival analysis. Abundance estimates by DWLS and CIBERSORT were not predictive of outcomes, however, estimates by Janus suggested that diagnostic samples whose composition included at least 5.8% AML-X cells had significantly worse outcomes (p=1.90E-3, Figure 12C). Janus composition estimates were also the only ones that were predictive of survival by Cox regression (p=6E-4, compared to p=0.07 using DWLS). Note that in the 3 scRNA-Seq profiled samples with AML-X cells, 10%-11% of AML cells were identified as AML-X cells at diagnosis. The most upregulated genes in AML-X were MALAT1, NEAT1, and ZEB2 (Figure 12D). The long noncoding

RNAs NEAT1 and MALAT1 colocalize in Chr11Q13.1, are co-expressed in AML, and are predicted to transcriptionally co-inhibit hundreds of genes(Chiu et al., 2018; Lorenzi et al., 2021). Their common targets were significantly downregulated in AML-X (Figure 12E). Moreover, NEAT1 has been previously implicated with chemoresistance in cancer (Adriaens et al., 2016), and both NEAT1 and MALAT1 have been associated with poor prognosis in childhood leukemia (Pouyanrad et al., 2019). In addition, MALAT1 has been shown to post-transcriptionally upregulate ZEB2 in cancer (Cheng et al., 2021; Xiao et al., 2015); ZEB2 was the third most upregulated gene in AML-X. In its totality, this evidence suggests that lncRNAs, including NEAT1 and MALAT1 may play key roles in regulating chemoresistance in paediatric AML.

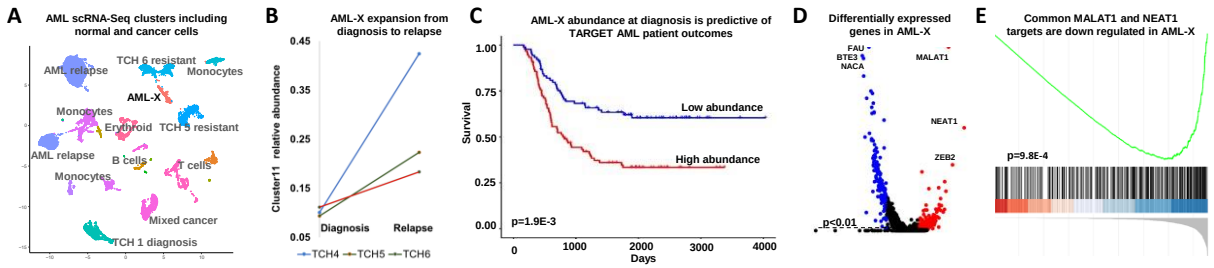

Figure 12: Analysis of 6 paired diagnosis-relapse scRNA-Seq Paediatric AML samples. (A) The integration of the 12 samples revealed a combination of AML and normal cell types, including AML-X, which contained AML cells from both the diagnostic and relapse samples of AML patients TCH4, TCH5, and TCH6. (B) The proportion of AML-X cancer cells was higher in the relapse samples of TCH4, TCH5, and TCH6 than in the corresponding diagnostic samples. (C) Janus predicted AML-X cell abundance in RNA-Seq profiles of diagnostic TARGET AML samples was predictive of outcomes; note that abundance estimates by other methods were not predictive. (D) Upregulated genes in AML-X cells included NEAT1, MALAT1, and ZEB2. (E) BigHorn-predicted targets of NEAT1 and MALAT1 were significantly enriched for downregulated genes in AML-X cells.

Chapter 5 Outcomes-predictive neuroblastoma (NB) subclones

We produced concurrent bulk RNA-Seq and scRNA-Seq NB profiles to evaluate deconvolution methods and to identify NB subclones that are predictive of outcomes. The NB dataset included scRNA-Seq profiles of 14 NB samples (Jansky et al., 2021). We used cell types and expression profiles from these scRNA-Seq data to deconvolve diagnostic samples from a total of 161 NB (Pugh et al., 2013) patients that were profiled by the TARGET consortium. In total, 15 NB subclones were identified (Figure 13A), and each one was tested for outcomes predictions based on abundance estimates by Janus, DWLS, CIBERSORT using TARGET RNA-Seq data. In total, two subclones were identified to be predictive of outcomes using Janus abundance estimates (Figure 13, NB-s1 and NB-s2 at p=1.4E-3 and p=1.0E-2, respectively). No cluster was predictive of outcomes using DWLS or CIBERSORT. Interestingly both subclones showed enrichment for upregulated MYC and RAS signalling pathway genes, as well as upregulation of semaphore genes. MYC is a key driver in NB, and the RAS pathway (Eleveld et al., 2015; Pugh et al., 2013) and semaphore genes (Delloye-Bourgeois et al., 2017) have been previously implicated in poor prognosis for NB patients. In addition, MALAT1 was identified as upregulated in NB-s1 cells, and its targets were significantly inhibited, similarly to what was observed for chemoresistant AML.

In total, our analysis suggests that bulk RNA-Seq deconvolution with Janus can help identify drivers of poor prognosis in AML. Our work suggested that a predictive panel based on MYC and RAS signalling pathways could help classify NB patients at diagnosis.

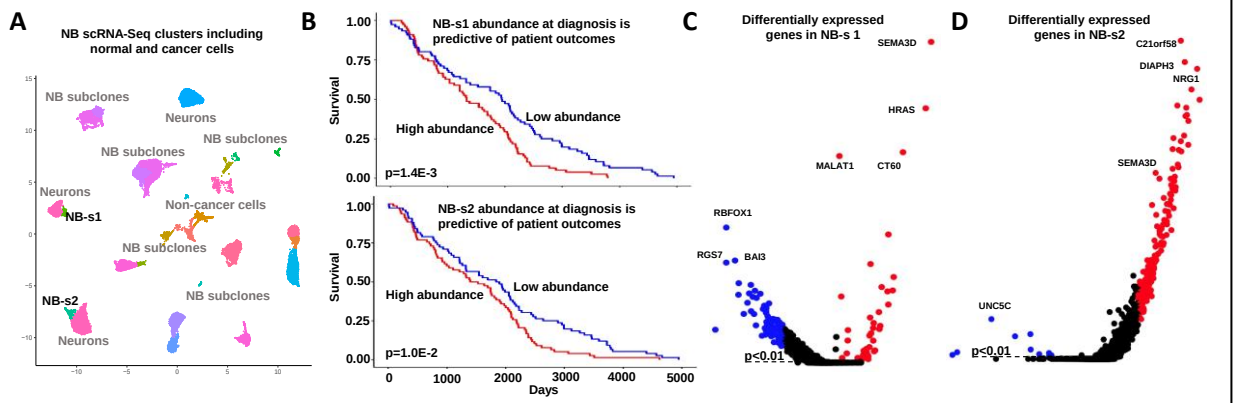

Figure 13: (A) Analysis of concurrent RNA-Seq and scRNA-Seq profiles of 14 NB identified 15 NB subclones including NB-s1 and NB-S2 (B) whose abundance was predictive of outcomes in TARGET samples that were profiled by RNA-Seq. (C) Significantly upregulated genes in cells included semaphores, HRAS (NB-s1), and MYC targets.

Chapter 6 Summary and Conclusion

We set out to develop technologies to identify cell subpopulations in each tumour type, associate them with responses to therapies, and use these data to predict therapeutic responses. During the past 3 years, iPC has developed technologies to address all these challenges. Here we report on the use of these technologies to identify chemoresistant cancer cells that are associated with relapse and metastases in HB, paediatric AML, and NB. These success stories produce a framework to predict and evaluate therapies for paediatric cancer based on the composition of tumors at diagnosis. We envision that this framework could be expanded to these and other cancer types in the immediate future.

Chapter 7 References

[1] Adriaens, C., Standaert, L., Barra, J., Latil, M., Verfaillie, A., Kalev, P., Boeckx, B., Wijnhoven, P. W., Radaelli, E., and Vermi, W. (2016). p53 induces formation of NEAT1 lncRNA-containing paraspeckles that modulate replication stress response and chemosensitivity. Nature medicine 22, 861-868.

[2] Avila Cobos, F., Alquicira-Hernandez, J., Powell, J. E., Mestdagh, P., and De Preter, K. (2020). Benchmarking of cell type deconvolution pipelines for transcriptomics data. Nature communications *11*, 1-14.

[3] Bolouri, H., Farrar, J. E., Triche, T., Ries, R. E., Lim, E. L., Alonzo, T. A., Ma, Y., Moore, R., Mungall, A. J., and Marra, M. A. (2018). The molecular landscape of pediatric acute myeloid leukemia reveals recurrent structural alterations and age-specific mutational interactions. Nature medicine 24, 103-112.

[4] Cheng, H., Zhao, H., Xiao, X., Huang, Q., Zeng, W., Tian, B., Ma, T., Lu, D., Jin, Y., and Li, Y. (2021). Long non-coding RNA MALAT1 upregulates ZEB2 expression to promote malignant progression of glioma by attenuating miR-124. Molecular Neurobiology 58, 1006-1016.

[5] Chiu, H.-S., Somvanshi, S., Patel, E., Chen, T.-W., Singh, V. P., Zorman, B., Patil, S. L., Pan, Y., Chatterjee, S. S., TCGA, Sood, A. K., Gunaratne, P. H., and Sumazin, P. (2018). Pan-cancer analysis of lncRNA regulation supports their targeting of cancer genes in each tumor context. Cell reports 23, 297-312. PMC5906131

[6] Decamps, C., Arnaud, A., Petitprez, F., Ayadi, M., Baurès, A., Armenoult, L., Escalera, S., Guyon, I., Nicolle, R., and Tomasini, R. (2021). DECONbench: a benchmarking platform dedicated to deconvolution methods for tumor heterogeneity quantification. BMC bioinformatics 22, 1-17.

[7] Delloye-Bourgeois, C., Bertin, L., Thoinet, K., Jarrosson, L., Kindbeiter, K., Buffet, T., Tauszig-Delamasure, S., Bozon, M., Marabelle, A., and Combaret, V. (2017). Microenvironment-driven shift of cohesion/detachment balance within tumors induces a switch toward metastasis in neuroblastoma. Cancer Cell *32*, 427-443. e428. Bibliography

[8] Denisenko, E., Guo, B. B., Jones, M., Hou, R., de Kock, L., Lassmann, T., Poppe, D., Clément, O., Simmons, R. K., Lister, R., and Forrest, A. R. R. (2020). Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. Genome Biol *21*, 130. PMCPMC7265231

[9] Eleveld, T. F., Oldridge, D. A., Bernard, V., Koster, J., Daage, L. C., Diskin, S. J., Schild, L., Bentahar, N. B., Bellini, A., and Chicard, M. (2015). Relapsed neuroblastomas show frequent RAS-MAPK pathway mutations. Nature genetics 47, 864-871.

[10] Gambardella, G., Viscido, G., Tumaini, B., Isacchi, A., Bosotti, R., and di Bernardo, D. (2022). A single-cell analysis of breast cancer cell lines to study tumour heterogeneity and drug response. Nature communications *13*, 1-12.

[11] Gamis, A. S., Alonzo, T. A., Meshinchi, S., Sung, L., Gerbing, R. B., Raimondi, S. C., Hirsch, B. A., Kahwash, S. B., Heerema-McKenney, A., Winter, L., Glick, K., Davies, S. M., Byron, P., Smith, F. O., and Aplenc, R. (2014). Gemtuzumab ozogamicin in children and adolescents with de novo acute myeloid leukemia improves event-free survival by reducing relapse risk: results from the randomized phase III Children's Oncology Group trial AAML0531. J Clin Oncol 32, 3021-3032. PMCPMC4162498

[12] Gröbner, S. N., Worst, B. C., Weischenfeldt, J., Buchhalter, I., Kleinheinz, K., Rudneva, V. A., Johann, P. D., Balasubramanian, G. P., Segura-Wang, M., and Brabetz, S. (2018). The landscape of genomic alterations across childhood cancers. Nature 555, 321-327.

[13] Jansky, S., Sharma, A. K., Körber, V., Quintero, A., Toprak, U. H., Wecht, E. M., Gartlgruber, M., Greco, A., Chomsky, E., and Grünewald, T. G. (2021). Single-cell transcriptomic analyses provide insights into the developmental origins of neuroblastoma. Nature genetics *53*, 683-693.

[14] Jew, B., Alvarez, M., Rahmani, E., Miao, Z., Ko, A., Garske, K. M., Sul, J. H., Pietiläinen, K. H., Pajukanta, P., and Halperin, E. (2020). Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. Nature communications 11, 1-11.

[15] Loken, M. R., Alonzo, T. A., Pardo, L., Gerbing, R. B., Raimondi, S. C., Hirsch, B. A., Ho, P. A., Franklin, J., Cooper, T. M., Gamis, A. S., and Meshinchi, S. (2012). Residual disease detected by multidimensional flow cytometry signifies high relapse risk in patients with de novo acute myeloid leukemia: a report from Children's Oncology Group. Blood 120, 1581-1588.

[16] López-Terrada, D., Alaggio, R., de Dávila, M. T., Czauderna, P., Hiyama, E., Katzenstein, H., Leuschner, I., Malogolowkin, M., Meyers, R., and Ranganathan, S. (2014). Towards an international pediatric liver tumor consensus classification: proceedings of the Los Angeles COG liver tumors symposium. Modern Pathology 27, 472-491.

[17] Lorenzi, L., Chiu, H.-S., Avila Cobos, F., Gross, S., Volders, P.-J., Cannoodt, R., Nuytens, J., Vanderheyden, K., Anckaert, J., and Lefever, S. (2021). The RNA Atlas expands the catalog of human non-coding RNAs. Nature biotechnology 39, 1453-1465.

[18] Manica, M., Chouvarine, P., Mathis, R., Wagner, U., Oehl, K., Saba, K., Roditi, L. D. V., Pati, A. N., Martínez, M. R., Wild, P. J., and Sumazin, P. (2020). Inferring clonal composition from multiple biopsies of the same tumor. NPJ Syst Biol Appl 6, 27.

[19] Pouyanrad, S., Rahgozar, S., and Ghodousi, E. S. (2019). Dysregulation of miR-335-3p, targeted by NEAT1 and MALAT1 long non-coding RNAs, is associated with poor prognosis in childhood acute lymphoblastic leukemia. Gene 692, 35-43.

[20] Pugh, T. J., Morozova, O., Attiyeh, E. F., Asgharzadeh, S., Wei, J. S., Auclair, D., Carter, S. L., Cibulskis, K., Hanna, M., Kiezun, A., Kim, J., Lawrence, M. S., Lichenstein, L., McKenna, A., Pedamallu, C. S., Ramos, A. H., Shefler, E., Sivachenko, A., Sougnez, C., Stewart, C., Ally, A., Birol, I., Chiu, R., Corbett, R. D., Hirst, M., Jackman, S. D., Kamoh, B., Khodabakshi, A. H., Krzywinski, M., Lo, A., Moore, R. A., Mungall, K. L., Qian, J., Tam, A., Thiessen, N., Zhao, Y., Cole, K. A., Diamond, M., Diskin, S. J., Mosse, Y. P., Wood, A. C., Ji, L., Sposto, R., Badgett, T., London, W. B., Moyer, Y., Gastier-Foster, J. M., Smith, M. A., Guidry Auvil, J. M., Gerhard, D. S., Hogarty, M. D., Jones, S. J., Lander, E. S., Gabriel, S. B., Getz, G., Seeger, R. C., Khan, J., Marra, M. A., Meyerson, M., and Maris, J. M. (2013). The genetic landscape of high-risk neuroblastoma. Nat Genet 45, 279-284. PMC3682833

[21] Rasche, M., Zimmermann, M., Borschel, L., Bourquin, J.-P., Dworzak, M., Klingebiel, T., Lehrnbecher, T., Creutzig, U., Klusmann, J.-H., and Reinhardt, D. (2018). Successes and challenges in the treatment of pediatric acute myeloid leukemia: a retrospective analysis of the AML-BFM trials from 1987 to 2012. Leukemia 32, 2167-2177.

[22] Rubnitz, J. E., Inaba, H., Dahl, G., Ribeiro, R. C., Bowman, W. P., Taub, J., Pounds, S., Razzouk, B. I., Lacayo, N. J., Cao, X., Meshinchi, S., Degar, B., Airewele, G., Raimondi, S. C., Onciu, M., Coustan-Smith, E., Downing, J. R., Leung, W., Pui, C. H., and Campana, D. (2010). Minimal residual disease-directed therapy for childhood acute myeloid leukaemia: results of the AML02 multicentre trial. Lancet Oncol 11, 543-552.

[23] Sumazin, P., Chen, Y., Treviño, L. R., Sarabia, S. F., Hampton, O. A., Patel, K., Mistretta, T. A., Zorman, B., Thompson, P., and Heczey, A. (2017). Genomic analysis of hepatoblastoma identifies distinct molecular and prognostic subgroups. Hepatology 65, 104-121.

[24] Sumazin, P., Peters, T. L., Sarabia, S. F., Kim, H. R., Urbicain, M., Hollingsworth, E. F., Alvarez, K. R., Perez, C. R., Pozza, A., Najaf Panah, M. J., Epps, J. L., Scorsone, K., Zorman, B., Katzenstein, H., O'Neill, A. F., Meyers, R., Tiao, G., Geller, J., Ranganathan, S., Rangaswami, A. A., Woodfield, S. E., Goss, J. A., Vasudevan, S. A., Heczey, A., Roy, A., Fisher, K. E., Alaggio, R., Patel, K. R., Finegold, M. J., and López-Terrada, D. H. (2022). Hepatoblastomas with carcinoma features

represent a biological spectrum of aggressive neoplasms in children and young adults. Journal of Hepatology 77, 1026-1037.

[25] Tsoucas, D., Dong, R., Chen, H., Zhu, Q., Guo, G., and Yuan, G.-C. (2019). Accurate estimation of cell-type composition from gene expression data. Nature communications 10, 1-9.

[26] Xiao, H., Tang, K., Liu, P., Chen, K., Hu, J., Zeng, J., Xiao, W., Yu, G., Yao, W., and Zhou, H. (2015). LncRNA MALAT1 functions as a competing endogenous RNA to regulate ZEB2 expression by sponging miR-200s in clear cell kidney carcinoma. Oncotarget 6, 38005.