Inaugural dissertation

for

obtaining the doctoral degree

of the

Combined Faculty of Mathematics, Engineering and Natural Sciences

of the

Ruprecht – Karls – University

Heidelberg

Presented by

Enrique Blanco Carmona, M.Sc.

born in: Valencia, Spain

Oral examination: October 22nd, 2024

Characterizing tumor heterogeneity in ATRT and IDH-mutant glioma tumors using single-cell multi-omics analyses

Referees: Prof. Dr. Benedikt Brors

Dr. Şevin Turcan

Abstract

Over the course of the last decade, the world health organization (WHO) classification of tumors of the central nervous system (CNS) has started to incorporate different molecular insights as decision criteria for the categorization of different tumor types, which have promoted the surge of novel tumor types and subtypes. While methodology advances have enabled for an easier and more accurate diagnosis of the tumor cases, the underlying biology and tumor heterogeneity between tumor types and subtypes remain to be fully elucidated. This is evidenced by the dismal prognosis some tumor subtypes possess, underscoring the need for more effective and subtype-targeted tumor therapies.

Developing in parallel to the new CNS tumor classification, single-cell technologies have emerged as very powerful approaches to perform comparative analysis of tumor subtypes at different Omics layers, including transcriptomics and chromatin accessibility. Within this context, the two research projects I have worked on during my PhD focused on understanding the tumor heterogeneity depicted by the various IDH-mutant glioma or atypical teratoid/rhabdoid tumor (ATRT) subtypes.

In both cases, single-cell analyses identified a novel tumor cell subpopulation. In the case of IDH-mutant gliomas, this was a non-cycling, ribosomal-enriched tumor cell population harboring a stemness phenotype and exhibiting expression of elongation factors and oncogenes (annotated as RE). For ATRTs, a “rhabdoid ground-state” tumor cell population was identified and characterized across all SMARCB1-deficient ATRT subtypes, which presented high stemness activity, together with an expression profile resembling that of neuroblasts with cycling activity (annotated as IPC-like). Both these tumor cell populations in IDH-mutant gliomas and ATRTs, upon validation in external datasets, hold promise for the development of subtype-specific therapies, albeit further research is still needed.

Further analyses on the IDH-mutant glioma cohort revealed a differential composition of tumor-associated macrophages (TAM) across subtypes, with an increased prevalence of pro-inflammatory TAM states in astrocytomas, for which immunohistochemistry (IHC) staining revealed elevated p-STAT1 expression, suggesting the promotion of a pro-inflammatory microenvironment in astrocytomas. Longitudinal analyses on paired primary-recurrent

astrocytomas sample pairs demonstrated that the composition of tumor cell types across patients at tumor recurrence remained consistent, emphasizing their therapeutic potential.

Subsequent analyses on the ATRT cohort are still being carried out. These include the examination of the single-cell chromatin accessibility data, and the characterization of the crosstalk between both tumor cell populations and the tumor and its microenvironment. Additional experiments encompass the validation, in ATRT organoid models, of druggable targets designed to push tumor cells into differentiated cell states within ATRT subtype-specific tumor cell lineages. Other analyses include the validation and spatial distribution of the various tumor and TME cell pupations in ATRTs of all three subtypes using spatial transcriptomics.

Finally, in order to address the increasing need of streamlined alternatives to generate high-quality, publication-ready data visualizations of single-cell transcriptomics data, I developed a software package for R, `SCpubr`. The software tool provides data visualization one-liner functions, the scope of which range from simpler visualization tasks such as inspecting dimensional reduction embeddings, displaying cell type composition, or assessing the expression or enrichment of selected genes, to inspecting the output of more complex analyses, such as copy number variant analysis or gene set enrichment analysis. Altogether, the scientific community has successfully adopted `SCpubr` for visualizing single-cell transcriptomic data, as evidenced by its growing number of citations.

Zusammenfassung

Im Laufe des letzten Jahrzehnts hat die Weltgesundheitsorganisation (WHO) bei der Klassifizierung von Tumoren des Zentralnervensystems (ZNS) damit begonnen, verschiedene molekulare Erkenntnisse als Entscheidungskriterien für die Kategorisierung verschiedener Tumortypen zu berücksichtigen, was eine Welle von neuartigen Tumortypen und -subtypen hervorgerufen hat. Während Fortschritte in der Methodik eine einfachere und genauere Diagnose der Tumorfälle ermöglicht haben, müssen die zugrunde liegende Biologie und Tumorheterogenität zwischen Tumortypen und -subtypen noch vollständig aufgeklärt werden. Dies wird durch die schlechte Prognose einiger Tumorsubtypen belegt und unterstreicht die Notwendigkeit wirksamerer und mehr auf Subtypen spezialisierter Tumortherapien.

Parallel zur neuen Klassifizierung von ZNS-Tumoren entwickelten sich Einzelzelltechnologien als sehr einflussreiche Ansätze für die vergleichende Analyse von Tumorsubtypen auf verschiedenen Omics-Ebenen, einschließlich den Ebenen der Transkriptomik und der Chromatinzugänglichkeit. In diesem Zusammenhang konzentrierten sich die beiden Forschungsprojekte, an denen ich während meiner Doktorarbeit gearbeitet habe, auf das Verstehen der Tumorheterogenität, welche sich in den verschiedenen Subtypen von IDH-Gliomen oder atypischen teratoiden/rhabdoiden Tumoren (ATRT) widerspiegelt.

In beiden Fällen identifizierten Einzelzellanalysen eine neuartige Tumorzellsubpopulation. Im Fall der IDH-Gliome handelte es sich um eine nicht proliferierende, ribosomal angereicherte Tumorzellpopulation, die einen stammzellartigen Phänotyp aufwies und sich durch die Expression von Elongationsfaktoren und Onkogenen auszeichnete (annotiert als RE). Für ATRTs wurde eine Tumorzellpopulation mit „rhabdoidem Grundzustand“ über alle SMARCB1-defizienten ATRT-Subtypen hinweg identifiziert und charakterisiert, welche eine hohe stammzellartige Aktivität sowie ein Expressionsprofil aufwies, das dem von Neuroblasten mit proliferierender Aktivität ähnelte (als IPC-ähnlich bezeichnet). Sowohl die genannte Tumorzellpopulation der IDH-Gliome und als auch die der ATRTs scheinen nach Validierung in externen Datensätzen vielversprechend für die Entwicklung subtypspezifischer Therapien, auch wenn hierfür noch weitere Forschung erforderlich sein wird.

Weitere Analysen der IDH-Gliomkohorte ergaben eine unterschiedliche Zusammensetzung von tumorassoziierter Makrophagen (TAM) über die Subtypen hinweg, mit einer erhöhten Prävalenz proinflammatorischer TAM-Zustände bei Astrozytomen, bei welchen die immunhistochemische (IHC) Färbung eine erhöhte p-STAT1-Expression aufzeigte, was auf die Förderung einer proinflammatorischen Mikroumgebung in Astrozytomen schließen lässt. Längsschnittanalysen an Probenpaaren aus primären und rezidivierenden Astrozytomen zeigten, dass die Zusammensetzung der Tumorzelltypen über alle Patienten hinweg beim Tumorrezidiv konstant blieb, was ihr therapeutisches Potenzial unterstreicht.

Nachfolgende Analysen der ATRT-Kohorte werden noch durchgeführt. Dazu gehören die Untersuchung der Daten zur Chromatinzugänglichkeit auf Einzelzellebene und die Charakterisierung der Kommunikation sowohl zwischen Tumorzellpopulationen als auch zwischen dem Tumor und seiner Mikroumgebung. Weitere Experimente umfassen die Validierung von Zielstrukturen, die einer Medikamentenbehandlung zugänglich sind und die darauf ausgelegt sind, Tumorzellen in differenzierte Zellzustände innerhalb ATRT-Subtyp-spezifischer Abstammungslinien zu bringen, in ATRT-Organoidmodellen. Weitere Analysen umfassen die Validierung und Untersuchung der räumlichen Verteilung der verschiedenen Tumor- und TME-Zellpopationen in ATRTs aller drei Subtypen mithilfe räumlicher Transkriptomik.

Zu guter Letzt, um dem zunehmenden Bedarf an zielgerichteten Alternativen zur Generierung hochwertiger, publikationsbereiter Datenvisualisierungen von Einzelzell-Transkriptomdaten gerecht zu werden, habe ich ein Softwarepaket für R, `SCpubr`, entwickelt. Das Softwaretool bietet einzeilige Funktionen zur Datenvisualisierung, deren Umfang von einfacheren Visualisierungsaufgaben wie der Überprüfung von Dimensionsreduktionseinbettungen, der Beschreibung der Zelltypzusammensetzung oder der Beurteilung der Expression oder Anreicherung ausgewählter Gene, bis hin zur Überprüfung der Ergebnisse komplexerer Analysen reicht, wie beispielsweise der Analyse der Kopiezahlvarianten oder Analysen zur Anreicherung von bestimmten Gengruppen. Insgesamt hat die wissenschaftliche Gemeinschaft `SCpubr` erfolgreich zur Visualisierung transkriptomischer Einzelzelldaten angenommen, was durch die wachsende Zahl von Zitaten demonstriert wird.

List of abbreviations

ALT	alternative lengthening of telomeres
AS	astrocytoma
ATAC	assay for transposase-accessible chromatin
ATRT	atypical teratoid/rhabdoid tumor
ATRX	alpha thalassemia/mental retardation
BAM	border-associated macrophage
BMD	bone-marrow-derived macrophage
CBTRUS	central brain tumor registry of the United States
CCA	canonical correlation analysis
CGGA	Chinese glioma genome atlas
CNS	central nervous system
CNV	copy number variant
CollecTRI	collection of transcriptional regulatory interactions
CP	choroid plexus
CRAN	comprehensive R archive network
CTCF	CCCTC-binding factor
D-2HG	D-2-hydroxyglutarate
DNA	deoxyribonucleic acid
ecMRT	extra-cranial MRT
ESC	embryonic stem cell
FDR	false discovery rate
G-CIMP	glioma-associated CpG island methylator phenotype
GLASS	glioma longitudinal analysis
GO	gene ontology
GRCh38	genome reference consortium human build 38
GSEA	gene set enrichment analysis
GSVA	gene set variation analysis
GTF	gene transfer file
HIF1α	hypoxia-inducible factor 1 α
HVG	highly variable gene

IDH	isocitrate dehydrogenase
IHC	immunohistochemistry
IPC	intermediate precursor cell
JmjC	jumonji-C
KEGG	Kyoto encyclopedia of genes and genomes
MGMT	O(6)-methylguanine-DNA methyltransferase
mRNA	messenger RNA
MRT	malignant rhabdoid tumor
MSigDB	molecular signatures database
MVP	mean-variance plot
NMF	non-negative matrix factorization
NPC	neuronal precursor cell
NSPC	neuronal stem/precursor cell
OD	oligodendroglioma
OPC	oligodendrocyte precursor cell
PC	principal component
PCA	principal component analysis
PDGFRA	platelet-derived growth factor receptor alpha
PHATE	potential of heat-diffusion for affinity-based trajectory embedding
PROGENy	pathway responsive genes for activity inference
PSC	pluripotent stem cell
QC	quality control
RG	radial glia
RNA	ribonucleic acid
SCpubr	single-cell publication ready
scRNAseq	single-cell RNA sequencing
SHH	sonic hedgehog
snATACseq	single-nucleus ATAC sequencing
snRNAseq	single-nuclei RNA sequencing
SVD	singular value decomposition
SWI/SNF complex	switch/sucrose non-fermentable

t-SNE	t-distributed stochastic neighbor embedding
TAM	tumor-associated macrophages
TCGA	the cancer genome atlas
TERT	telomerase reverse transcriptase
TET	ten-eleven translocation
TF	transcription factor
TF-IDF	term frequency-inverse document frequency
TME	tumor microenvironment
TMZ	temozolomide
TSS	transcription start site
TYR	tyrosinase
UMAP	uniform manifold approximation and projection
UMI	unique molecular identifier
VST	variance stabilizing transformation
WGS	whole genome sequencing
WHO	world health organization
α-KG	α-ketoglutarate

Table of Contents

Prologue

The following thesis has been structured into several, independent chapters. First, I will place IDH-mutant gliomas and atypical teratoid/rhabdoid tumors (ATRT) within the context of central nervous system (CNS) tumors, briefly delving into the different major groups of CNS tumors and their overall incidence in the general population.

Then, I will provide a general introduction to single-cell genomics covering the different available methods and analyses that can be utilized in single-cell transcriptomics and chromatin accessibility data, narrowing them down to the ones relevant to the different projects I have worked on.

Next, I will expand on my contribution to single-cell methodology development with the publication of `SCpubr`, my R package, that generates high-quality data visualizations of single-cell transcriptomics data ready for publication in scientific journals, which I have employed throughout both of my research projects.

Following that, I will extensively illustrate the results of my research in both ATRT and IDH-mutant glioma tumors. Each project will encompass its own chapter, following a traditional structure: starting with an introduction summarizing the state of the art, followed by the methodology associated to the research project, continuing with a report of the results obtained, and finalizing with their discussion.

Afterwards, I will move on to the final remarks in an epilogue chapter, where I will elaborate on the current status of single-cell technologies in the context of tumor research and how the advent of spatial transcriptomics can reshape the way research is performed.

Finally, I will conclude my thesis by outlining the various research publications I am part of, both ongoing and published.

Introduction to CNS tumors

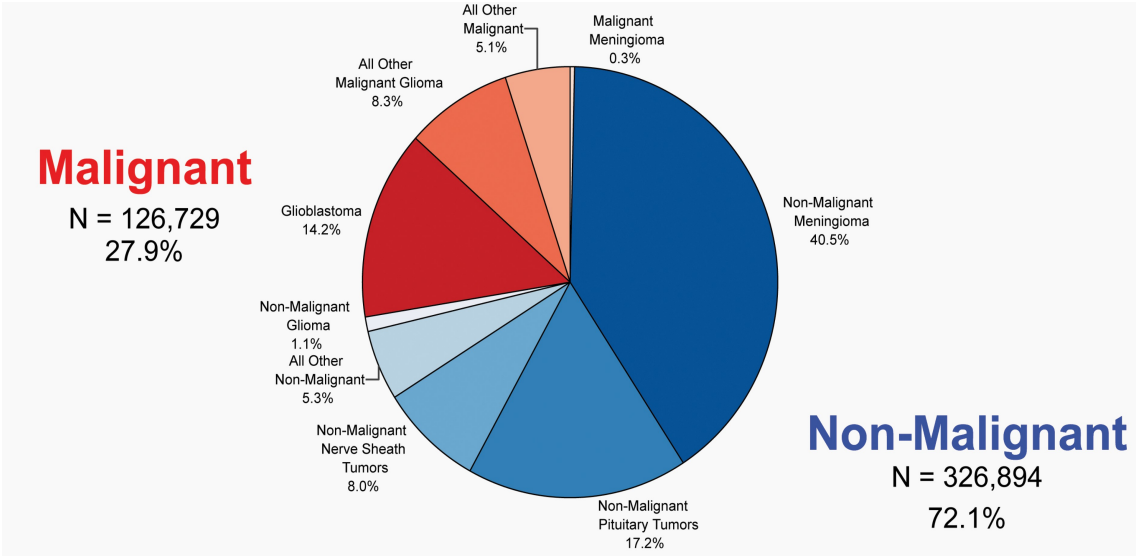
Central nervous system (CNS) tumors encompass a wide variety of different tumor types and subtypes, each with distinct molecular and clinical characteristics. This complexity has led to global efforts to accurately classify and diagnose these tumors. Traditionally, classification heavily relied on histology, but the latest versions of the world health organization (WHO) classification of CNS tumors have started to also include molecular features in the diagnostic criteria. In particular, DNA methylation profiling has contributed significantly to a more accurate diagnosis of CNS tumors and has led to the identification of many new tumor types and subtypes that were added to the 5th edition of the WHO classification of CNS tumors[1]. In this thesis chapter, I aim to provide a general overview of the major CNS tumor types, highlighting the placement of IDH-mutant gliomas and atypical teratoid/rhabdoid tumors (ATRTs), the specific tumor types I have researched. While I wrote the original text of this chapter, I used ChatGPT to enhance its readability.

Based on the 5th edition of the WHO classification of CNS tumors, these tumors can be broadly categorized into 12 major groups: gliomas, glioneural tumors and neuronal tumors; choroid plexus tumors; embryonal tumors; pineal tumors, cranial and paraspinal nerve tumors; meningiomas; mesenchymal, non-meningothelial tumors involving the CNS; melanocytic tumors; hematolymphoid tumors involving the CNS; germ cell tumors; tumors of the sellar region; and metastases to the CNS. Furthermore, the classification includes a separate chapter encompassing various genetic tumor syndromes involving the CNS. Examples within this group include, for instance, Li-Fraumeni syndrome, a familial disposition syndrome associated with germ-line mutations in *TP53*[2], which can develop different pediatric and adult tumor types, such as medulloblastoma[3], among others. Each major group comprises various tumor types, defined not only by histological traits but also by molecular features. For instance, IDH-mutant gliomas are characterized by recurrent mutations in the *IDH1* or *IDH2* genes[4] and belong to the larger group of gliomas, which includes many other tumor types that are molecularly and clinically distinct. ATRTs are part of the embryonal tumors, exhibiting biallelic inactivation of either the *SMARCB1*[5] or *SMARCA4*[6] genes. Other embryonal tumor types include medulloblastoma, cribriform neuroepithelial tumor (CRINET), embryonal tumor with multilayered rosettes (ETMRs), CNS neuroblastoma with *FOXR2* activation, and CNS embryonal

tumors with *BCOR* alterations. ATRTs are distinct from these other embryonal tumors based on DNA methylation, transcriptomics, and their mutational landscape. Subsequently, molecular features are crucial for accurate diagnosis, including specific DNA methylation profiles or characteristic genetic aberrations such as gene fusions or (in)activation of specific oncogenic drivers[7]. Furthermore, recent research has introduced machine learning-based methylation classifiers that, when combined with current diagnostic standards, assist in diagnosing the most challenging cases[8]. Methylation data is particularly valuable as it indicates the cell of origin[9,10] and can help infer the primary site of metastatic cancers[11].

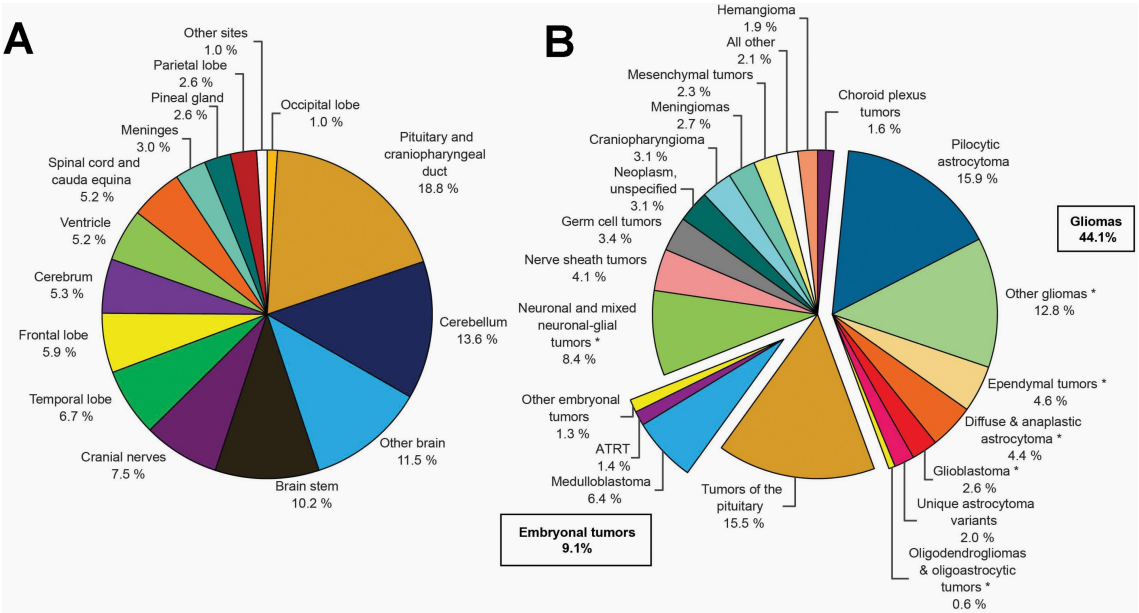
Furthermore, not all major CNS tumor groups occur equally in the population. According to the latest CBTRUS statistical reports[12,13], across all age groups, the majority of CNS tumor cases are benign (72.1%, 17.88 cases per 100.000 population[12]), with malignant cases exhibiting a lower incidence (27.9%, 6.94 cases per 100.000 population[12]), with the most common CNS tumor type being non-malignant meningioma (40.5%) and glioblastoma for malignant cases (14.2%) (Figure 1). However, in children, malignant cases have a higher incidence rate (3.55 cases per 100.000 population[13]) than non-malignant cases (2.67 cases per 100.000 population[13]), and therefore the majority of CNS tumors in children between 0-14 years are classified as malignant (65.8%)[14]. IDH-mutant gliomas are also malignant CNS tumors, most common in older children and younger adults, and can present as astrocytoma, IDH-mutant (incidence rate of 0.45 cases per 100.000 population[12]) or oligodendroglioma, IDH-mutant and 1p/19q codeleted (incidence rate of 0.29 cases per 100.000 population[12]). ATRTs are rare embryonal CNS tumors, mostly occurring in very young children (incidence rate of 0.09 cases per 100.000 population[13]), and are extremely rare in adults.

Focusing on pediatric CNS tumor cases only, tumors can arise throughout the CNS, but most commonly affected sites are the pituitary and craniopharyngeal duct (18.8%), followed by the cerebellum (13.6%) (Figure 2A). Among the different tumor types, gliomas are the most common (44.1%), which include low-grade gliomas such as pilocytic astrocytomas, and high-grade gliomas like ependymal tumors, astrocytomas, oligodendrogliomas, and glioblastomas. Furthermore, embryonal tumors account for a total of 9.1% of pediatric cases, with ATRTs representing 1.4% of all pediatric CNS cases (Figure 2B).



a. Percentages may not add up to 100% due to rounding.
b. Includes histopathologies with ICD−O−3 behavior code of /3 from choroid plexus tumors, neuronal and mixed neuronal−glial tumors, tumors of the pineal region, embryonal tumors, nerve sheath tumors, mesenchymal tumors, primary melanocytic lesions, lymphoma, other hematopoietic neoplasms, germ cell tumors, tumors of the pituitary, craniopharyngioma, hemangioma, neoplasm unspecified, and all other.
c. Includes histopathologies with ICD−O−3 behavior code of /0 or /1 from neuronal and mixed neuronal−glial tumors, tumors of the pineal region, embryonal tumors, other tumors of cranial and spinal nerves, mesenchymal tumors, primary melanocytic lesions, other hematopoietic neoplasms, germ cell tumors, craniopharyngioma, hemangioma, neoplasm unspecified, and all other.
Abbreviations: CBTRUS, Central Brain Tumor Registry of the United States; US, United States; NPCR, National Program of Cancer Registries; SEER, Surveillance, Epidemiology, and End Results.

Figure 1: "Distribution[a] of Primary Brain and Other Central Nervous System Tumors by Behavior (Five-Year Total=453,623; Annual Average Cases=90,725), CBTRUS Statistical Report: US Cancer Statistics—NPCR and SEER, 2016-2020". ICD-O-3 codes represent an international coding of tumor types where the first four digits designate the histology term, with the fifth digit (separated by a "/"), referring to the behavior code (whether a tumor is malignant, benign, in situ or uncertain)[15]. Reprinted with permission from Ostrom, *et al*.[12]



*All or some of this histopathology is included in the CBTRUS definition of gliomas including ICO-O-3 histopathology codes 9360-9384 and 9391-9480 (Table 2).
a. Percentages may not add up to 100% due to rounding.
Abbreviations: CBTRUS, Central Brain Tumor Registry of the United States; NPCR, National Program of Cancer Registries; SEER, Surveillance, Epidemiology, and End Results Program; ATRT, Atypical Teratoid/Rhabdoid Tumor.

Figure 2: "Distribution[a] in Children and Adolescents (Ages 0-19 Years) of Primary Brain and Other Central Nervous System Tumors (Five-Year Total=24,999; Annual Average Cases=5,000) by A) Site and B) Histopathology Subtypes, CBTRUS Statistical Report: US Cancer Statistics—NPCR and SEER, 2016-2020" ICD-O-3 codes represent an international coding of tumor types where the first four digits designate the histology term, with the fifth digit (separated by a "/"), referring to the behavior code (whether a tumor is malignant, benign, in situ or uncertain)[15]. Reprinted with permission from Ostrom, *et al*.[12]

Overall, tumor types and subtypes are molecularly and clinically different, exhibiting varying prevalence across different age groups and tumor types. This highlights an inherent tumor heterogeneity that still needs to be explored. Further research is essential to understand the biological differences driving these tumor types and subtypes. Single-cell technologies are proving to be key in this effort, as they can identify novel tumor cell types with therapeutic potential.

Introduction to single-cell genomics

In this introductory chapter, I aim to provide a comprehensive summary of the typical steps involved in analyzing single-cell transcriptomics and chromatin accessibility sequencing data, restricting it to the methods applied throughout my research. The structure of this chapter will be based on a recent review, which outlines current best practices in the field of single-cell genomics[16]. While I wrote the original text of this chapter, I used ChatGPT to enhance its readability.

One of the major breakthroughs in molecular biology over the past decade has been the transition from bulk sequencing methods to single-cell techniques, and more recently, spatial transcriptomics. From targeting the transcriptome to encompassing other Omics layers such as chromatin accessibility[17], T cell receptor (TCR)/B cell receptor (BCR) repertoires[18], surface proteins[19] and spatial location[20], single cell technologies have become essential tools in the study of tumor types[21]. This shift has led to the development of numerous software tools, with over a thousand dedicated to single-cell transcriptomic analysis alone[22]. These tools range from popular computational frameworks like `Seurat`[23] or `Scanpy`[24] to specialized method benchmarks and best practices workflows, all aimed at ensuring the correct data analysis and visualization of single-cell transcriptomics data[16,25–29].

Single-cell transcriptomics

Expanding on the principles of bulk transcriptomics, single-cell RNA sequencing (scRNAseq, fresh tissue) / single-nuclei RNA sequencing (snRNAseq, frozen tissue) measures the quantity of mRNA molecules per cell. To accomplish this, tissue is extracted and digested to isolate the cells. Generally, scRNAseq technologies can be divided into plate-based and droplet-based protocols, the latter rising in popularity[30]. In both protocol types, after sequencing, raw reads are mapped to a reference genome and matched to their respective originating cell. Each sequencing method employs a proprietary approach to tackle this task, typically involving the assignment of a barcode to each mRNA molecule prior to sequencing. These barcodes are then tracked during reference mapping to generate a count matrix. Individual mRNA molecules originating from a single cell are commonly denoted as unique molecular identifiers (UMI).

Data pre-processing

Typically, the initial step of any scRNAseq analysis involves performing quality control (QC), normalization, feature selection and dimensionality reduction (`Figure 3A`). In essence, the objective is to eliminate any potential sources of bias that could affect the analysis, including technical artifacts derived from the sequencing steps or potential biological confounders such as cell cycle or apoptosis[16,31].

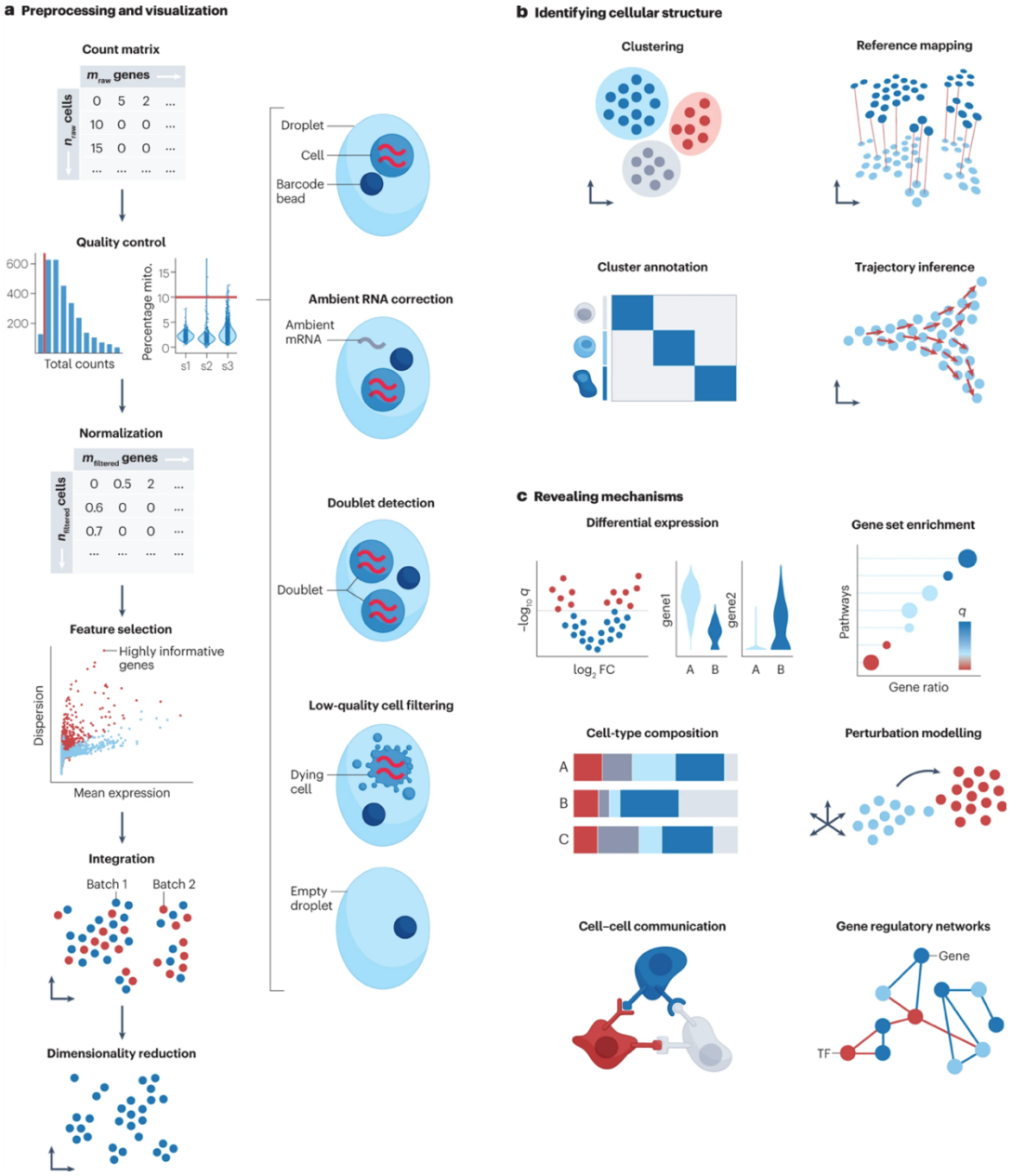


Figure 3: "Overview of unimodal analysis steps for scRNA-seq. a, Count matrices of cells by genes are obtained from raw data processing pipelines. To ensure that only high-quality cells are captured, count matrices are corrected for cell-free ambient RNA and filtered for doublets and low-quality or dying cells. The latter is done by removing outliers with respect to quality control metrics (the number of counts per barcode, called count depth or library size, the number of genes per barcode and the fraction of counts from mitochondrial genes per barcode

(percentage mito.)). All counts represent successful capture, reverse transcription and sequencing of an mRNA molecule. These steps vary across cells, and therefore count depths for identical cells can differ. Hence, when comparing gene expression between cells, differences may originate solely from sampling effects. This is addressed by normalization to obtain correct relative gene abundances between cells. Single-cell RNA sequencing (scRNA-seq) data sets can contain counts for up to 30,000 genes for humans. However, most genes are not informative, with many genes having no observed expression. Therefore, the most variably expressed genes are selected. Different batches of data are integrated to obtain a corrected data matrix across samples. To ease computational burden and to reduce noise, dimensionality reduction techniques are commonly applied. This further allows for the low-dimensional embedding of the transcriptomics data for visualization purposes. **b**, The corrected space can then be organized into clusters, which represent groups of cells with similar gene expression profiles, annotated by labels of interest such as cell type. The annotation can be conducted manually using prior knowledge or with automatic annotation approaches. Continuous processes, such as transitions between cell identities during differentiation or reprogramming, can be inferred to describe cellular diversity that does not fit into discrete classes. **c**, Depending on the question of interest and experimental set-up, conditions in the data set can be tested for upregulated or downregulated genes (differential expression analysis), effects on pathways (gene set enrichment) and changes in cell-type composition. Perturbation modelling enables the assessment of the effect of induced perturbations and the prediction of unmeasured perturbations. Expression patterns of ligands and receptors can reveal altered cell–cell communication. Transcriptomics data further enable the recovery of gene regulatory networks. q, q value." Reprinted with permission from Heumos, *et al*.[16]

Quality control (QC)

Quality control of scRNAseq data primarily involves filtering out low-quality cells, ambient DNA or instances where multiple cells might have been captured together (doublets) or none at all (empty droplets). Typically, three metrics play a key role in detecting low-quality cells: the number of UMIs per cell (count depth), the number of genes per cell and the fraction of mitochondrial genes. A high fraction of mitochondrial genes often indicates dying cells[32,33]. Examining the distribution of these covariates independently for each sample allows for the establishment of individualized thresholds tailored to each sample. While these metrics can vary significantly across samples and sequencing techniques, community-based guidelines have been developed to provide some ground-basis of what are commonly used thresholds.[25,27,34].

Another significant confounding factor is cell-free RNA, often termed as ambient RNA. This is RNA that does not belong to a captured cell, but that actually comes as a contamination from other lysed cells in the solution, which can ultimately lead to the detection of markers identifying different cell populations within the same group of cells, therefore adding noise and mixing two genuine populations together[35]. Various software tools have been developed to estimate and eliminate ambient RNA, such as `SoupX`[35], `CellBender`[36] and `Decontx`[37].

Ensuring that each droplet contains a single cell is a crucial final step of the quality control. Empty droplets are relatively straightforward to identify as they exhibit minimal to no expression. By applying the aforementioned QC cutoffs, empty droplets can be effectively removed from the dataset. However, identifying doublets presents a more complex challenge, as they come in different nature: homotypic and heterotypic. Homotypic doublets occur when two cells of the same cell type are captured in a single droplet, while heterotypic doublets contain cells from different cell types. Heterotypic doublets result in a barcode with expression profiles from two distinct cell types, while homotypic doublets manifest as a barcode with high expression of marker genes of a particular cell population.

The consensus approach is to target homotypic doublets by applying QC cutoffs on the number of UMIs and genes per cells. For heterotypic doublets, a diverse array of tools and methods have been developed, leading to benchmarking reviews[38,39] that recommend the use of `scDblFinder`[38,40] and `DoubletFinder`[39,41]. Additionally, meta-analysis tools such as `Demuxafy`[42], which consolidate the output of several doublet detection tools, are beginning to emerge.

Normalization

Within a single dataset, cells may exhibit a different number of transcripts. Consequently, it becomes necessary to transform raw counts to ensure that the expression profiles across cells are comparable. This transformation is followed by a variance stabilization step to mitigate the influence of outliers on the structure of the data[43] and scaling, which standardizes counts to have zero mean and unit variance. Numerous normalization methods have been developed over time to address this need, each with its own strengths and weaknesses. Benchmarking reviews have demonstrated that simpler normalization methods perform comparably to more sophisticated ones[44]. Methods like the shifted logarithm[44] are better-suited for cases where optimal dimensional reduction is desired. Others, such as the approach proposed by `scran`[45], are most effective for datasets requiring strong batch correction. Additionally, methods based on analytical Pearson residuals excel at feature selection and identification of rare cell types[46]. One widely adopted normalization method, available in `Seurat`, is SCtransform[47], which has recently been updated[48]. Therefore, the choice of normalization method depends on the specific goals of downstream analysis.

Feature selection

After normalization, the next step involves identifying the most variable genes through feature selection, which serves as the basis for dimensionality reduction. Various feature selection methods are available, broadly categorized into empirical-distribution-based methods, generative-model-based methods, and distribution-free methods, with the first group being the most widely used[49]. Empirical-distribution-based methods rank genes based on summary statistics, typically gene expression variance or dispersion, computed over the empirical distribution of genes across cells[49]. Genes are ranked accordingly, with the top-ranking genes selected as highly variable genes (HVGs).

However, the selection of the number of HVGs can impact downstream analysis. A meta-analysis of 153 studies revealed that the selected number of HVGs ranged from 1.000 to 3.000 in 65% of studies[49]. Popular empirical-distribution-based methods implemented in `Seurat` and `Scanpy` are dispersion (DISP), mean-variance plot (MVP) and variance stabilizing transformation (VST).

Dimensionality reduction

The subsequent step in data pre-processing involves dimensionality reduction. Starting from the basis that the expression values of each gene act as independent variables in the analysis, a single-cell transcriptomics experiment has as many degrees of freedom as genes. The selection of HVGs serves as an initial step towards reducing the number of degrees of freedom in the analysis, while dimensionality reduction methods aim to either summarize the underlying topology of the data or visualize it[16]. Principal component analysis (PCA) is a popular method used to summarize the data, while techniques like uniform manifold approximation and projection (UMAP)[50], t-distributed stochastic neighbor embedding (t-SNE)[51] and potential of heat-diffusion for affinity-based trajectory embedding (PHATE)[52] are employed for data visualization[16,27,52].

While dimensionality reduction methods are valuable for analyzing and interpreting single cell-transcriptomics data, reviews suggest not to solely rely on the topology of two-dimensional dimensional reduction embeddings but rather using them as a complementary tool to results derived from quantitative analyses[16,53].

Data integration and batch correction

Following normalization and dimensional reduction, eliminating confounding sources of variation is a critical step for accurate downstream analyses. These sources can be of two types: technical and biological. An example of technical variation could be a dataset comprising several samples collected from different laboratories or under varying experimental conditions (batch), thereby introducing batch effects. To address this, data integration tools have been developed over the last years. A recent benchmark review evaluating up to 16 different methods[54] recommended the use of canonical correlation analysis (CCA)[55] or `Harmony`[56] for cases where the batch effect is rather simple. For datasets with more complex batch effects, tools such as `scANVI`[57], `scVI`[58], `scGen`[59] or `Scanorama`[60] are preferred.

Examples of biological variation include the cell cycle effect, where observed differences in cells stem from their differential phases in the cell cycle during sequencing rather than a genuine biological effect. Recent benchmarks have recommended the use of the cell cycle regression methods offered by tools such as `Scanpy`[24] and `Seurat`[47]. These methods compare the mean expression values of the cell cycle genes to reference cell cycle-based gene signatures. Subsequently, the use of `Tricycle`[61], a software tool that maps the dataset to a cell cycle-derived dimensional reduction embedding, is suggested, particularly for highly heterogeneous datasets[62]. Ultimately, deciding whether to remove this batch effect depends on whether significant cell cycle differences are observed in the dataset, and doing so risks losing the identification of cycling cells in the dataset.

Clustering and cell type annotation

After pre-processing the data, biological questions can be addressed. The initial step involves generating groups of cells with similar expression profiles, known as cell clusters. These clusters can be then analyzed to determine their identity through cell type annotation (`Figure 3B`).

Clustering

A diverse range of clustering algorithms can be applied to address this question, leading to benchmarking reviews that rank the optimal clustering algorithms for single-cell transcriptomics[63,64]. The findings from such studies suggested the use of Louvain algorithm, which serves as default implementation in `Seurat`. However, further research has highlighted

that the Louvain algorithm may produce poorly connected communities. This insight led to the development of Leiden algorithm[65], which is currently recommended as the algorithm of choice in reviews of single-cell transcriptomics[16,27].

Cell type annotation

Cell type annotation is the process of assigning biological meaning to clusters. This process typically involves assessing how transcriptionally similar a given cluster is to other known cell types. Current best practices suggest approaching cell type annotation in three consecutive steps: automatic annotation, manual annotation and expert validation[16,66]. Automatic annotation involves the use of pre-trained classifier models or reference mappers. Reference mappers such as `scArches`[67], `Symphony`[68] or `Azimuth`[23] are software tools that utilize annotated external datasets or atlases to compute a co-embedding with the query datasets, followed by label transfer. The second step involves manual annotation, which can be achieved by examining the expression of selected marker genes on given clusters, the combined expression of gene signatures across clusters, or the computation and analysis of differentially expressed genes between each cluster. Finally, expert validation is advisable, especially in cases where prior reference datasets are lacking[66].

Downstream analyses

Once annotations have been assigned to the different clusters in the dataset, different downstream analyses become feasible. These include but are not limited to: differential expression analysis, gene set enrichment analysis, copy-number variant analysis, ligand-receptor analysis, deconvolution analysis, and non-negative matrix factorization (`Figure 3C`).

Differential expression analysis

Differential expression analysis aims to identify genes that are either up- or downregulated across conditions. Methods for testing differential expression can be broadly classified into pseudobulk-based and cell-level-based approaches[16]. Pseudobulk-based methods involve aggregating count data across all cells within a biological replicate, resulting in a count matrix resembling that of bulk transcriptomics. Differential expression testing is then performed by widely adopted tools such as `edgeR`[69], `DESeq2`[70] or `limma`[71]. On the other hand, cell-level-based methods utilize generalized mixed models[16]. Benchmarking reviews[72,73] have shown minimal

overlap in results across differential expression methods, often favoring the use of bulk transcriptomics methods[16,74,75]. `Scanpy` defaults to using t-test for differential expression analysis, while `Seurat` resorts to Wilcoxon rank-sum test.

Gene set enrichment analysis

Gene set enrichment analysis aims to move beyond individual genes to explore terms that represent specific biological functions or phenotypes, such as pathways, transcription factors and its downstream targets (regulon), gene ontology terms, among others. This is facilitated by prior knowledge databases containing such terms and their associated genes, including Gene Ontology[76], KEGG[77], Reactome[78] or MSigDB[79], as well as weighted gene set databases like PROGENy[80] for pathways, or DoRothEA[81] and CollecTRI[82] for regulons. Enrichment analysis is typically performed using methods such as hypergeometric tests, GSEA[83] or GSVA[84], among others[16]. Due to its widespread use, gene set enrichment analysis has promoted the development of framework tools like `decoupleR`[85], which integrate various prior knowledge networks and enrichment methods. An important aspect to take into consideration when performing gene set enrichment analysis is the choice of prior knowledge network, as it significantly influences the range of terms the method can identify as enriched[16,86].

Cell-cell communication

An annotated dataset enables the examination of interactions between different cell populations by assessing the joint expression of ligand-receptor pairs, a process commonly referred to as cell-cell communication. Numerous tools have been developed to quantify this effect, including `CellChat`[87], `CellPhoneDB`[88] and `SingleCellSignalR`[89]. These tools typically utilize databases of ligand-receptor interactions to infer crosstalk between clusters. Research has demonstrated than the choice of tool significantly influences the observed results[90]. To address this variability, framework tools such as `LIANA` (for R)[90] and `LIANA+` (for python)[91] have been developed.

These tools allow for the computation of ligand-receptor interactions using multiple methods and rank the results to generate a consensus scoring across tools[90]. Additionally, inferring interactions across all possible cluster pairs is not recommended, as the results might be overwhelming to interpret due to the sheer number of significant interactions returned. Therefore, narrowing the scope of the analysis to selected clusters is advisable, coupled with

expert validation to extract biologically significant insights out of the statistically significant interactions.

Copy number variant (CNV) analysis

Copy number variant (CNV) analysis, in the context of single-cell transcriptomics, involves inferring of copy-number aberrations by comparing expression profiles between reference and target clusters. This analysis is particularly relevant when studying tumor datasets, as the presence or absence of specific CNV events can help classify cells as healthy or tumor cells. Numerous tools have been developed to tackle CNV analysis in single-cell transcriptomics, including `XClone`[92], `sciCNV`[93], `copyKAT`[94] or `inferCNV`[95]. However, there are currently no benchmarking reviews comparing different tools, nor are there any framework tools that generate a consensus CNV calling out of the results from multiple tools. Regarding `inferCNV`, the tool was originally designed for SMARTseq2 data but can also be applied to 10X datasets. Since overall counts in 10X datasets are lower than in SMARTseq2 datasets, increasing the sensitivity of the analysis by generating metacells (artificial cells designed by aggregating the raw counts of several cells within a given cluster) may be advisable (see `Copy Number Variant analysis`).

Cell deconvolution analysis

Over the decades, a multitude of bulk transcriptomics datasets have been generated due to scientific research and routine protocols for tumor diagnosis. These datasets represent heterogeneous mixtures of the various cell types present in the original tissue. Cell deconvolution methods aim to infer cell proportions from bulk transcriptomics data by using a reference gene expression profile from annotated cell types, such as pseudobulked annotated single-cell transcriptomics data[96]. This approach is especially useful for validating the empirical proportions from a given single-cell dataset against reference datasets and for repurposing extensive existing data, such as the cancer genome atlas program (TCGA), for novel research utilizing single-cell datasets[97]. Many tools have been developed over the years to perform cell deconvolution. A benchmarking review evaluating 20 different cell deconvolution tools recommended the use of regression-based tools like `CIBERSORT`[98] when working with bulk transcriptomics datasets. It also suggested pairing such tools with those designed for single-cell transcriptomics data, such as `MuSiC`[99], emphasizing the importance of curating a reference

matrix that accurately represents all cell types present on the query datasets[100]. However, cell deconvolution methods are continuously evolving alongside advancements in the field of single-cell transcriptomics. For instance, there has been the development of cell deconvolution tools tailored for spatial transcriptomics, such as SPOTlight[101].

Non-negative matrix factorization (NMF)

Tumor cells often exhibit a continuum of differentiation, ranging from the most stem-like tumor cells to more differentiated cells, following specific tumor lineages. In single-cell transcriptomics datasets where the differentiation lineages are unknown, unsupervised analysis like non-negative matrix factorization (NMF) can uncover complex biological processes within the data.

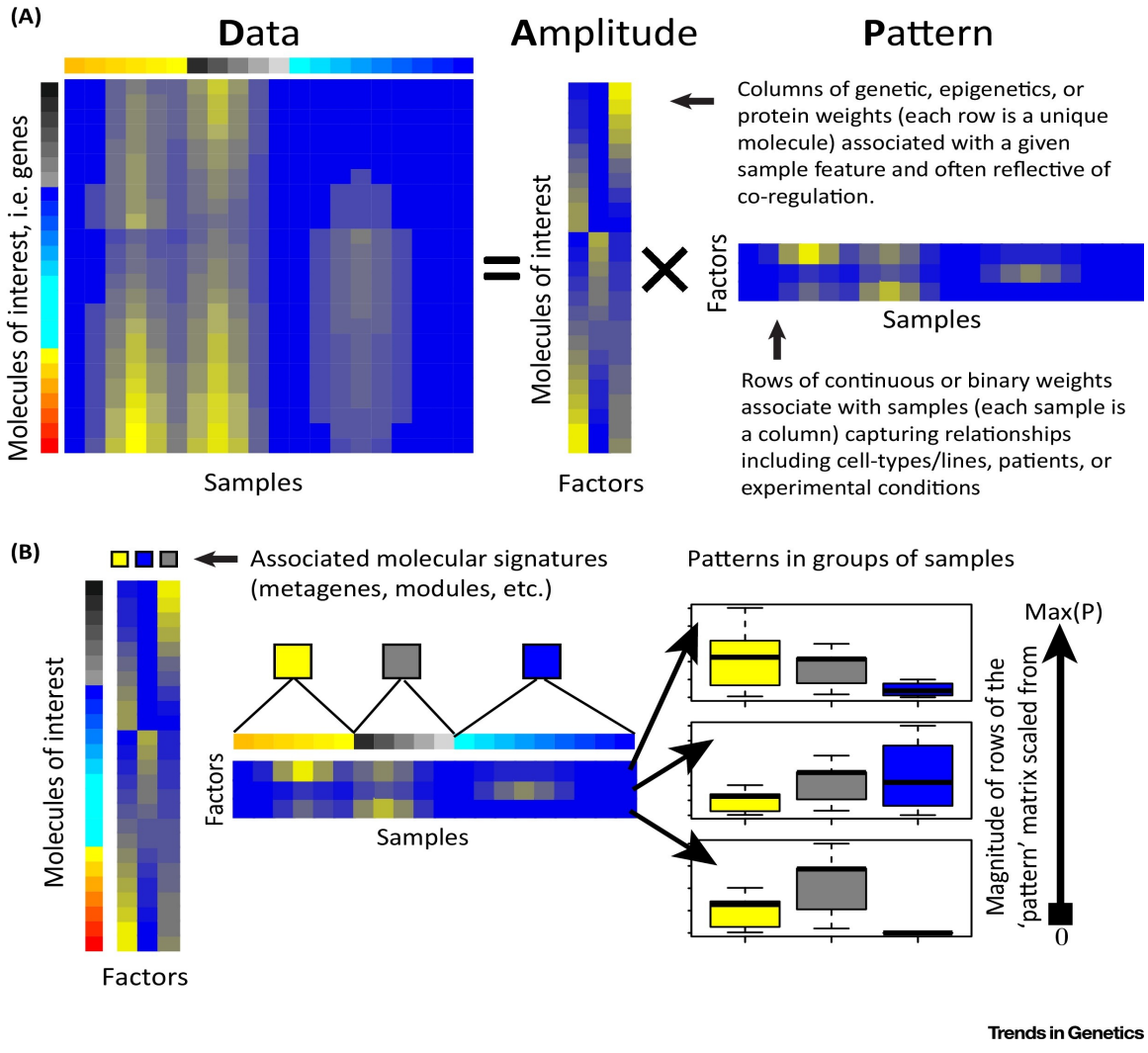


Figure 4: "The Matrix Product of the Amplitude and Pattern Matrices Approximates the Preprocessed Input Data Matrix. (A) The number of columns of the amplitude matrix equals the number of rows in the pattern matrix, and represents the number of dimensions in the low-dimensional representation of the data. Ideally, a pair of one column in the amplitude matrix and the corresponding row of the pattern matrix represents a distinct source of

biological, experimental, and technical variation in each sample (called complex biological processes, CBPs). **(B)** The values in the column of the amplitude matrix then represent the relative weights of each molecule in the CBP, and the values in the row of the pattern matrix represent its relative role in each sample. Plotting of the values of each pattern for a pre-determined sample grouping (here indicated by yellow, grey, and blue) in a boxplot as an example of a visualization technique for the pattern matrix. Abbreviation: Max(P), maximum value of each row of the pattern matrix." Reprinted with permission from Stein-O'Brien, *et al.*[102]

NMF is a matrix factorization technique applied on non-negative matrices, such as count data, which returns two independent matrices with the property of reconstructing the original when multiplied. These matrices are commonly referred as amplitude and pattern matrices (Figure 4A)[102]. The amplitude matrix retains the genes from the original matrix, while the pattern matrix retains the cells, both completed by new rows or columns termed NMF factors[102]. Both matrices offer meaningful insights into the data: in the amplitude matrix, each NMF factor represents a molecular signature or complex biological process, while in the pattern matrix, each NMF factor represents the contribution of a given molecular signature to a given cell (Figure 4B)[102]. When applied to tumor cells on a per-patient basis, NMF yields a set of signatures present within the tumor cells as a whole. These signatures, when correlated, can define molecular patterns present across tumor cases, termed NMF metaprograms[103]. Gene set enrichment analysis applied on the top scoring genes for each NMF metaprogram can reveal their biological nature. Overall, NMF is a powerful tool for uncovering biological patterns across tumor datasets, as evidenced by its use in pan-cancer studies aiming to identify recurrent tumor subpopulations across tumors[104].

Single-cell chromatin accessibility

Chromatin accessibility throughout the whole genome can be assessed through the assay for transposase accessible chromatin with high-throughput sequencing, commonly referred as ATAC-seq[105]. Similar as bulk transcriptomics, this method has evolved to the single-cell resolution, allowing for the investigation of chromatin accessibility in individual cells. In ATAC-seq, chromatin accessibility is assessed by sequencing DNA fragments resulting from the activity of Tn5 transposase enzyme[106]. Tn5 transposase is used to fragment DNA in open regions of the chromatin and add sequencing adapters to the resulting fragments. In regions of the chromatin presenting a closed state, the 3D structure prevents efficient tagging and fragmentation by the enzyme, therefore remaining largely unaffected[105].

Despite being a broadly used technique, there are no standards that define the features of the analysis. The reads are typically summarized into cell-by-peak, cell-by-bin, cell-by-gene and cell-by-TF count matrices, with evidence suggesting the latter two perform less effectively compared to the former[107]. Cell-by-peak count matrices capture chromatin openness across the genome, indicated by the enrichment of Tn5 transposition events in open chromatin regions compared to closed chromatin regions, generally referred as peaks. These matrices tend to be sparse[107], requiring a sufficient number of cells to correctly identify rare cell types[108]. Conversely, cell-by-bin matrices utilize uniformly sized windows (bins) across the genome, measuring the amount of Tn5 transposition events occurring between specified genomic coordinates defined by each bin, making them particularly useful for clustering purposes[108]. `Signac`, a widely adopted tool for the analysis of single-cell chromatin accessibility data[109], primarily utilizes cell-by-peak matrices.

Data pre-processing

From the fragment files, which contain the DNA fragments resulting from Tn5 transposition events[16], various QC metrics can be derived. These metrics encompass the total number of fragments per cell, serving as an indicator of sequencing depth; the enrichment of fragments in transcription start sites (TSS), illustrating the degree of enrichment in open chromatin regions compared to the background[16]; the nucleosome signal, determined by the ratio of long to short fragments, being low ratio indicative of high quality[110]; and the ratio of reads mapping in ENCODE's blacklist regions[111] can be computed, with a higher ratio indicating lower quality[16].

Examining these metrics individually for each dataset is essential for accurately defining QC thresholds to eliminate low-quality cells[16]. To identify and remove doublets from the analysis, best practices recommend utilizing two doublet scoring methods, such as `scDblFinder`[40] and `AMULET`[112], and combining their outputs to filter out doublets[16,40].

Normalization, dimensional reduction and clustering

To normalize the features, recent studies lean towards peak binarization[108,113,114], although alternative methods like modelling the counts have demonstrated better preservation of biological information[115]. For dimensional reduction, benchmarking reviews recommend the use of latent semantic indexing (LSI), facilitated by tools such as `ArchR`[113] or `Signac`[109]; latent

Dirichlet allocation via cisTopic[114]; or spectral embedding through snapATAC[108]. Subsequently, batch correction can be performed using LIGER[54], followed by cell clustering using Leiden algorithm.

Cell type annotation

To annotate the different cell clusters, several approaches are available. Gene activity estimation can be derived from the peaks based on the number of ATAC-seq counts mapping on the gene and two kilobases upstream. Gene activity effectively acts as RNA-seq counts, and enrichment scoring methods querying gene sets for different cell populations can be employed to annotate the cell clusters. Based on the cell clustering, differentially accessible peaks can be inferred similarly to single-cell transcriptomics data, utilizing tools such as edgeR or DESeq2[116]. Differentially accessible peaks can serve as a proxy for TF enrichment analysis using hypergeometric tests[113], with tools like chromVAR[117]. Analysis of top enriched TFs can aid in determining the cell identity of the clusters.

Alternatively, when matching single-cell transcriptomics data is available, cell annotation labels can be transferred from the RNA to the ATAC modality since the cells share matching identifiers. When matching IDs are not available, label transfer methods can be applied to project the labelling from RNA towards the ATAC modality[118]. In short and simplified, this process involves computing dimensional reduction through canonical correlation analysis (CCA) to project both query and reference datasets into a shared dimensional reduction space, from which “anchors” are retrieved. Anchors represent pairs of cells (one for the reference and one for the query dataset) deemed to originate from the same cell type[118]. Therefore, gene activity is correlated with expression, and a prediction score for each cell in the ATAC dataset is returned based on the correlation scores[118].

SCpubr: publication-ready plots for single-cell transcriptomics

Introduction

In recent years, there has been a significant shift towards the use of single-cell technologies in research, with research projects relying on single-cell data across various Omics layers, such as transcriptomics and chromatin accessibility. This shift has created an environment that fostered the development of software tools designed to address the unique analysis challenges posed by single-cell datasets.

With a wide range of specialized software tools available for different types of analyses, efforts have been made to consolidate the most widely used analyses into unified framework tools. This led to the creation of community-favorite tools like `Seurat`[23] for R, and `Scanpy`[24] for python. Moreover, as these framework tools are increasingly adopted by the scientific community, more software tools are developed using them as a foundation. This creates an ecosystem of interdependencies that broadens and simplifies the analysis of single-cell data.

Subsequently, the need for data analysis tools has expanded to the field of data visualization, as effectively displaying results has become increasingly important. Both `Seurat` and `Scanpy` cover basic types of data visualization, leaving further style enhancements to the user. However, these fine-grained style modifications can be time-consuming and technically challenging. Therefore, there is a need for software solutions that offer easy customization of data visualizations for single-cell data.

As a result, several tools have been published in recent years to facilitate the visualization of single-cell data, including `scCustomize`[119], `dittoSeq`[120], `iSEE`[121] shiny app, `LotOfCells`[122], and `plot1cell`[123]. While these tools streamline the generation of figures, users still need to make additional modifications if desired, which can be technically challenging. Proficiency in plotting software like `ggplot2`[124] is often required to make these changes, creating a barrier for users who are not familiar with these tools.

Here, I present `SCpubr` (**S**ingle-**C**ell **pub**lication-**r**eady), a user-friendly R package designed to generate high-quality, publication-ready data visualizations of single-cell transcriptomics datasets (`Figure 5A-B`). Based on the use of `Seurat` objects, `SCpubr` offers a wide variety of

functions, each tailored to a specific data visualization relevant to single-cell transcriptomics. With a syntax designed to match that of `Seurat`, `SCpubr` aims to facilitate seamless integration between the two packages. Additionally, `SCpubr` produces publication-ready plots with a minimalistic yet aesthetic appearance, while still allowing for full customization if desired.

Figure 5: SCpubr logo and banner. (A) SCpubr logo, emphasizing the use of eye-catching color palettes available in the package. The background grid symbolizes cells forming different clusters. The logo was designed by the artist Keryan[125] based on my indications. **(B)** SCpubr banner, created by modifying the X and Y coordinates of the UMAP embedding of a Seurat object to arrange cells into letter formations. Cells are colored based on their position along the UMAP2 component. UMAP1, x-axis; UMAP2, y-axis.

Methods

This R package has been developed independently, being all the design choices and functionalities of the package decided by me. However, I would like to thank Prof. Dr. Marcel Kool for promoting a curiosity-driven working environment that made this project possible. While I wrote the original text of this chapter, I used ChatGPT to enhance its readability.

Datasets

For development and testing purposes, a single-cell transcriptomics dataset comprising 10.000 peripheral blood mononuclear cells (PBMCs) was utilized[126]. This dataset underwent standard quality control, considering several metrics: the number of unique molecular identifiers (UMI) per cell (`nCount_RNA`), the number of genes per cell (`nFeature_RNA`) and the percentage of mitochondrial RNA per cell (`percent.mt`). Cells were filtered out if `nCount_RNA < 1000`, `nFeature_RNA < 500` or `percent.mt > 20%`.

Normalization of the count data was performed using regularized negative binomial regression (RNBR) via the `Seurat::SCTransform()` function. Normalized counts served as the basis for dimensional reduction, first by principal component analysis (PCA), and then by uniform manifold approximation and projection (UMAP) using the top 30 principal components. Cell clusters were identified via the Louvain algorithm[127], using `Seurat::FindNeighbors()` and `Seurat::FindClusters()` functions. This dataset provided the foundation for function development and testing, and is used throughout the user manual website. However, for the purposes of this thesis chapter, the datasets from the ATRT project will be used instead, with the figures being slightly modified in Affinity Designer to achieve a better layout of the panels within figures.

Function development and implementation

The entire `SCpubr` package is coded in R (v4.2.0). Each function within `SCpubr` focuses on a specific type of data visualization, ranging from displaying data distributions to visualizing dimensional reduction embeddings and summarizing the results of specific analyses. Shared functionalities between functions are stored in private functions, promoting a modular and less redundant coding design. The parameter syntax and function names were chosen to match

those of `Seurat`, thus allowing for an easy adoption of `SCpubr` by any current `Seurat` user. That includes parameters such as `group.by`, `split.by` or `features`, among others.

Package development

The various functions were built into an R package following established guidelines[128]. This process included proper code and function documentation using `roxygen2`[129], unit testing with `testthat`[130], and adding the associated metadata and GPLv3 license to the package. `SCpubr` was tested for compliance with CRAN policies using `devtools`[131]. After passing all checks, it was deposited in the CRAN repository, where it is now available for download.

Website development

All the code was stored in a GitHub repository, and a user manual website was generated using `Quarto`[132]. Tutorials for each function were created and compiled with `Quarto`, selecting HTML book as the output format. The resulting files were stored in a separate GitHub repository. GitHub Actions were enabled to publish the book as a website through GitHub Pages.

Manuscript preparation

A manuscript[133] was written and deposited in bioRxiv to allow `SCpubr` to be cited as a research paper rather than as a website link. Further efforts are ongoing to prepare a manuscript for submission to a peer-reviewed journal.

Data availability

The source code is available in the respective GitHub repositories:

- Source code: https://github.com/enblacar/SCpubr
- User manual: https://github.com/enblacar/SCpubr-book

Use cases

The various functions included in SCpubr can be broadly characterized into four main groups, each covering different aspects of data visualization for single-cell transcriptomics datasets: visualization of dimensional reduction embeddings, inspection of data distributions, summarization of results from specific methods, and miscellaneous functions aimed at improving user experience. While this section provides an overview of the major functionalities in SCpubr, not every data visualization type is covered here. Subsequently, an in-depth user manual covering all the plot types and customization options is available online[134].

Visualizing dimensional reduction embeddings is a critical aspect of single-cell transcriptomics analysis, providing a streamlined and direct way to assess the results at various stages of the analysis. For instance, the PCA embedding can be visualized to determine the major sources of variability within a dataset, or the results of data integration can be assessed by displaying the integrated embedding colored by the sources of variation that were regressed out, among many other applications (Figure 6).

For these purposes, SCpubr allows for the visualization of any dimensional reduction embeddings stored in the Seurat object. Categorical data can be projected onto these embeddings (Figure 6A), and SCpubr also supports facetted visualizations where the dimensional reduction silhouette is drawn for clarity (Figure 6B). Additionally, facetted plots where the faceting and color encoding are mapped to two different categorical variables are also possible (Figure 6C). Continuous variables can also be projected onto the dimensional reduction embeddings (Figure 6D), with the color encoding range kept identical across facets to facilitate side-by-side comparisons (Figure 6E).

Next, the inspection of data distributions can be particularly relevant, as it can unveil the nature of data variables such as the enrichment of cells in a given gene marker set and the cell type composition across individual samples, among others (Figure 7). The distribution of data variables can be inspected through various data visualization types, many of which are included in SCpubr. For instance, a pictogram of the distribution of a categorical variable in the dataset can be generated as a waffle plot (Figure 7A), where each tile corresponds to 1% of the cells.

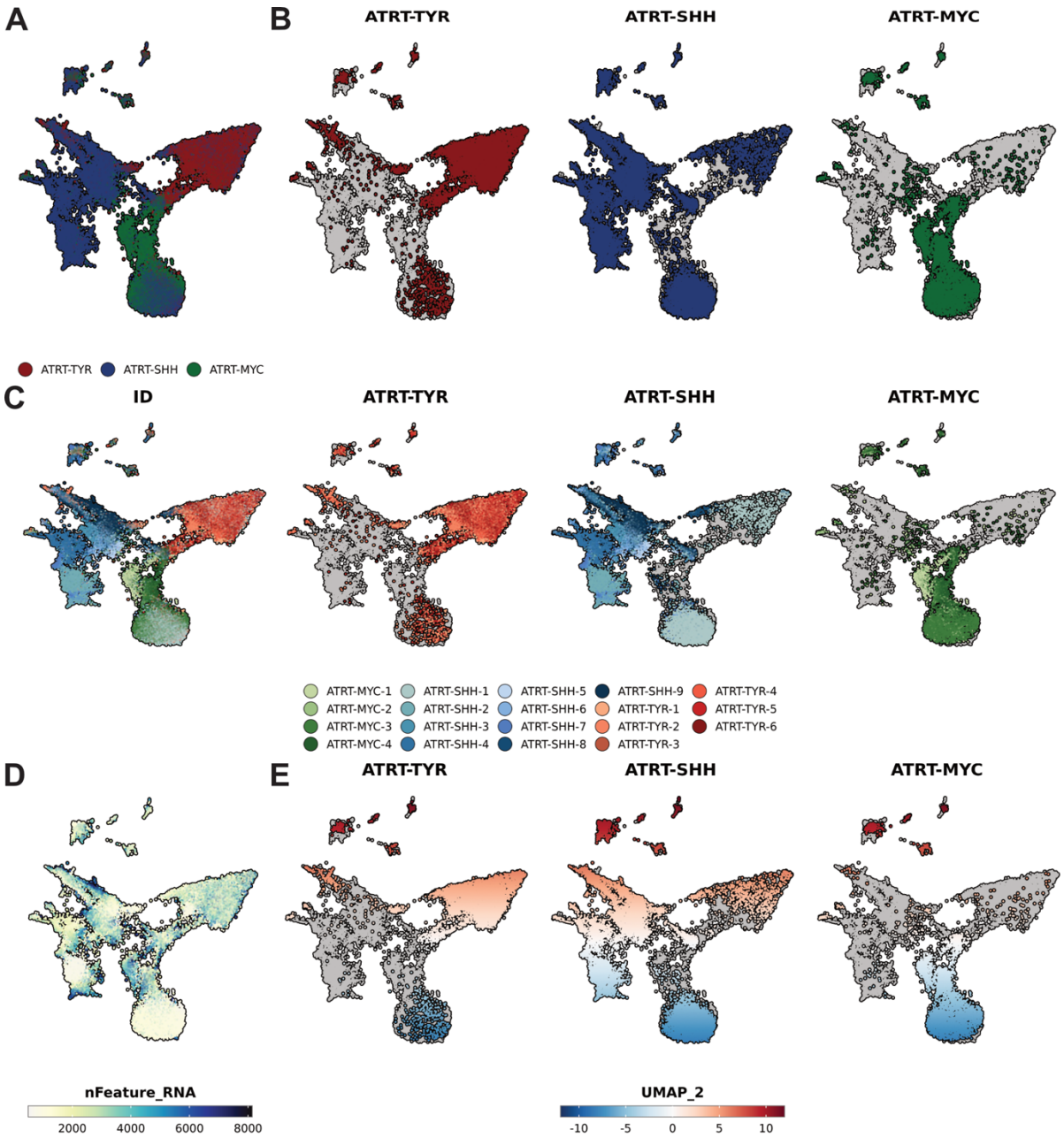


Figure 6: Visualizing dimensional reduction embeddings. UMAP representation of ATRT single-cell transcriptomics dataset, colored based on different scenarios. UMAP1, x-axis; UMAP2, y-axis. **(A)** Colored by ATRT subtype. **(B)** Split by each ATRT subtype, with cells of a given subtype colored and the remaining cells greyed out. **(C)** Split by each ATRT subtype, with cells colored based on the different patients within each subtype, and the remaining cells greyed out. **(D)** Colored based on the number of genes (nFeature_RNA) per cell. **(E)** Split by each ATRT subtype, with cells colored based on their position within UMAP2, and the remaining cells greyed out.

Therefore, waffle plots serve as a visual guide to determine whether there is an overrepresentation of a given group within a data variable. Categorical variables can also be summarized using bar plots (Figure 7B), where the proportion of the different groups is displayed as stacked bars, accounting for up to 100%.

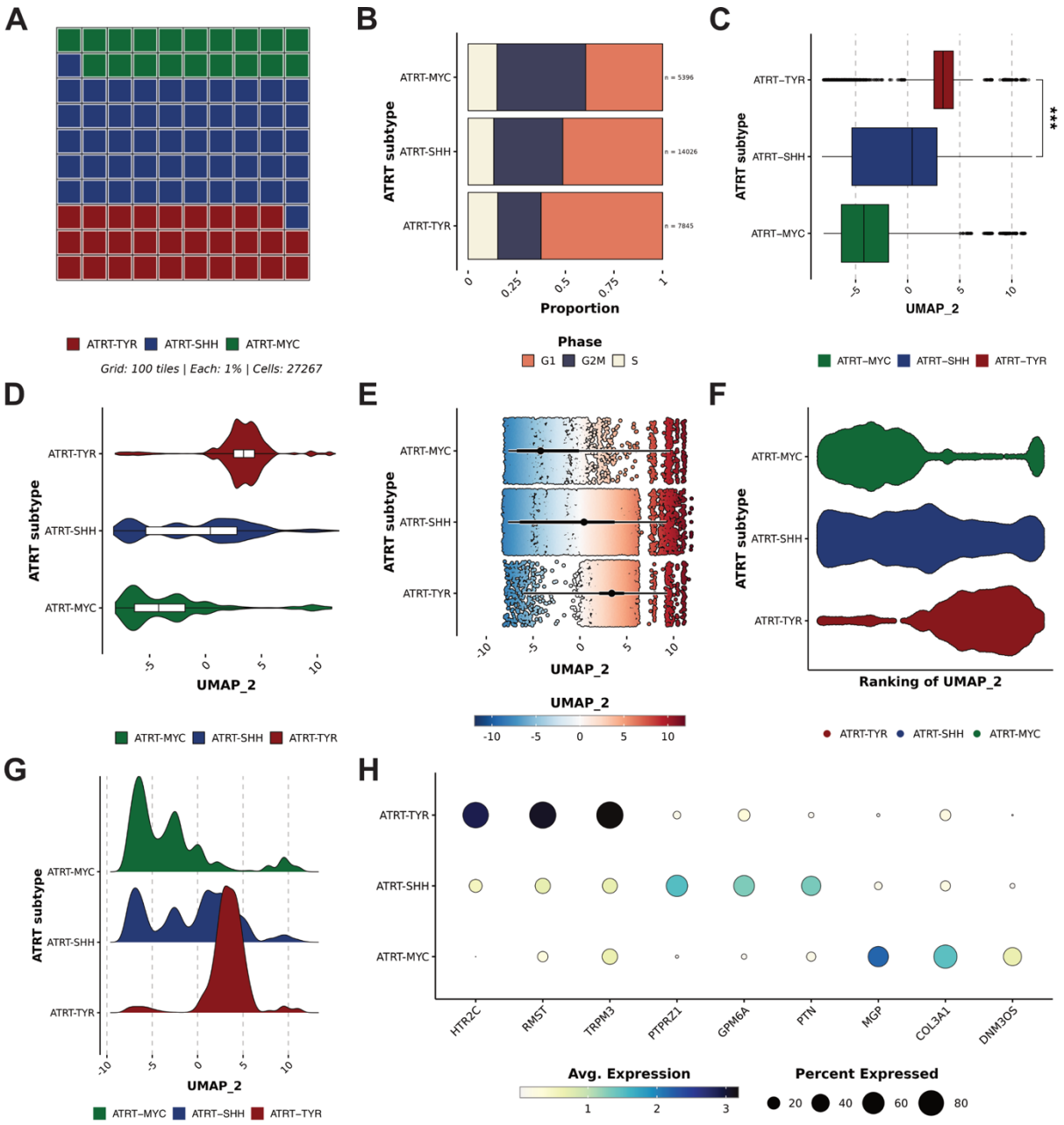


Figure 7: Inspecting distributions of the data. (A) Waffle plot depicting the distribution of ATRT subtypes in the dataset, with each tile representing 1% of the data. **(B)** Stacked bar plot showcasing the cell cycle phase assignment across ATRT subtypes. **(C-G)** Various visualization types inspecting the position of cells along the UMAP2 component: **(C)** box plots showing the differential position cells. Statistical significance between ATRT-TYR and ATRT-SHH cells is tested (Wilcoxon test, *** = 0.001); **(D)** violin plots, displaying the distribution of the data and including a box plot representation; **(E)** geyser plots, categorical scatter plots with jittered dots to identify outliers; **(F)** beeswarm plot, where cells are ranked by their UMAP2 position and dot jittering is defined by density, allowing for higher dispersion in denser regions; **(G)** ridge plots, displaying the distribution of a numerical variable across groups by density. **(H)** Dot plot depicting gene expression, with color encoding the average expression across a group and size representing the fraction of cells expressing the gene.

Furthermore, canonical data visualization types such as box plots (`Figure 7C`), which can be used to test for statistical differences between groups, or violin plots (`Figure 7D`), which help determine the distribution nature of a data variable, are included in `SCpubr`. Additionally, more

niche representations are available. This includes geyser plots (Figure 7E), where a continuous variable is displayed as a categorical jittered scatter plot, allowing for the identification of distribution outliers; beeswarm plots (Figure 7F), where a continuous variable is ranked and displayed across groups, with increased jittering proportional to the density of ranked values; and ridge plots (Figure 7G), where a continuous variable is displayed across groups as a function of density. Finally, continuous variables such as the expression of selected genes can be visualized as a dot plot (Figure 7H), where dot color encodes expression levels and size represents the fraction of cells within a group expressing the gene.

Additionally, following the inspection of dimensional reduction embeddings and data distributions, the summarization of results from different analyses is another crucial aspect of single-cell transcriptomics data analysis. For this, SCpubr provides functions for the most popular analyses, including: visualizing the density of expression of selected genes (Figure 8A); assessing the activity of pathways and regulons across categorical groups (Figure 8B-C); displaying top differentially enriched genes between two conditions as a volcano plot (Figure 8D); assessing the expression and enrichment of selected genes and gene sets across categorical groups (Figure 8E-F); generating correlation matrices based on Jaccard similarity (Figure 8G); classifying cells into three distinct cellular cell states based on gene sets as per Tirosh, *et al.*[135] (Figure 8H); and visualizing top enriched GO terms for a given gene set (Figure 8I).

Additional analyses covered by SCpubr, though not included in this chapter, include the visualization of top ligand-receptor pairs across categorical groups; copy number variant analysis scores across categorical groups, classifying cells into four distinct cellular states based on gene sets as per Neftel, *et al.*[136], and visualizing metadata as a categorical heatmap, among others. Some of these data visualization types are included in the following chapters of this thesis.

Finally, SCpubr offers a set of functions designed to enhance the overall user experience. This include a wrapper that saves plots to disk with specified proportions, resolution, and format, and an installation checker that ensures all necessary dependencies are installed, generating a report on which functions can be used with the current user installation.

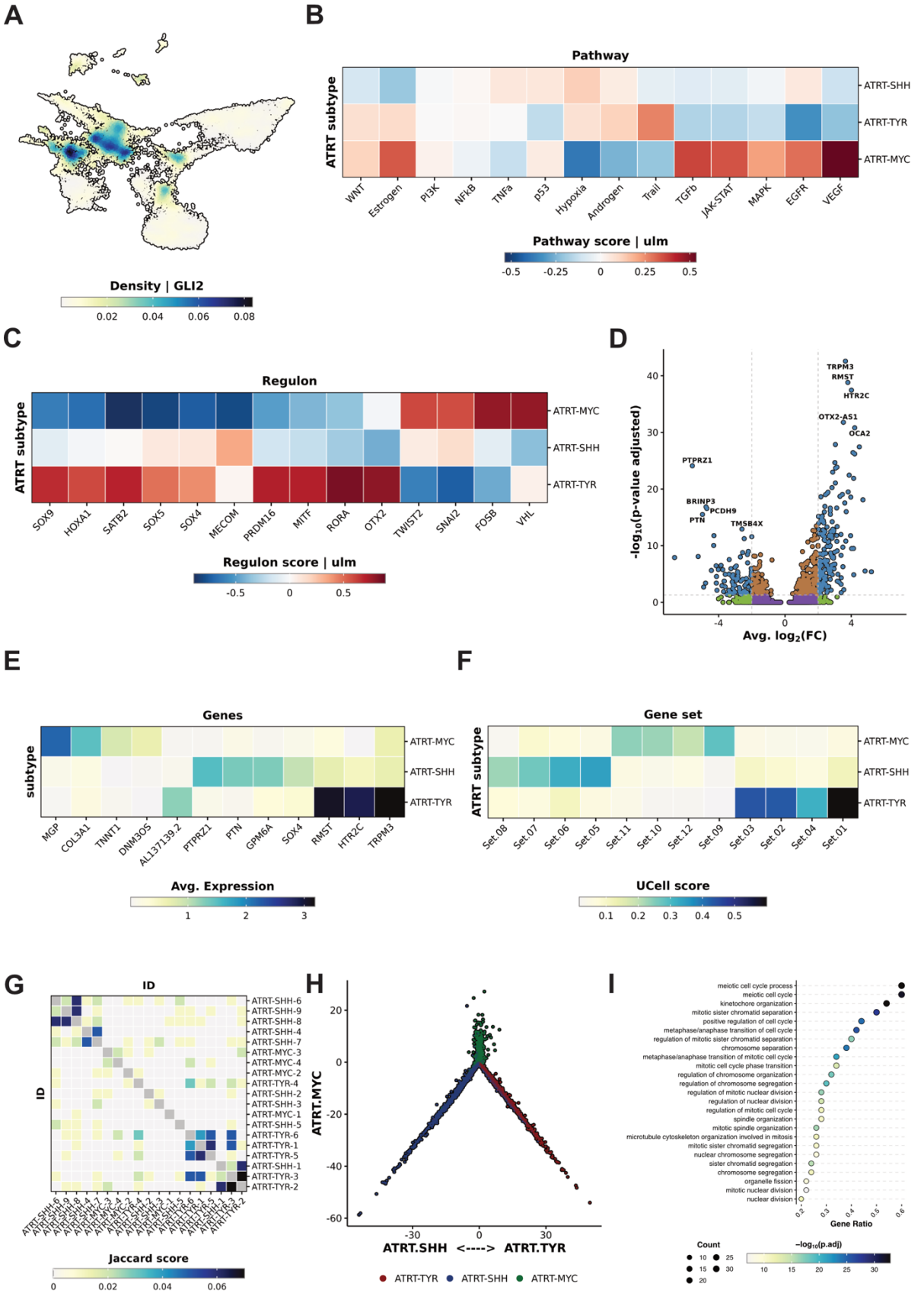


Figure 8: Summarizing experiments. (A) UMAP representation colored by the density of *GLI2* expression. UMAP1, x-axis; UMAP2, y-axis. **(B-C)** Activity scores computed over a set of 500 randomly selected cells for gene pathways **(B)** and regulons **(C)** across ATRT subtypes. ATRT subtypes, columns; pathway or regulon, rows. **(D)** Volcano plot of genes differentially expressed between ATRT-TYR (right) and ATRT-SHH (left) cells, with a

significance threshold of 0.05 and fold change cutoff is set to 2. Genes meeting these criteria are colored in blue. **(E-F)** Average expression values **(E)** and enrichment scores **(F)** of selected genes and gene sets across ATRT subtypes. ATRT subtypes, columns; genes or gene sets, rows. **(G)** Correlation matrix based on Jaccard similarity of the top 100 differentially expressed genes per patients in the ATRT single-cell transcriptomics cohort. **(H)** Cellular states plot adapted from Tirosh, *et al.*[135], discriminating ATRT cells based on the top 100 differentially expressed genes across ATRT subtypes. The Y-axis separates between ATRT-MYC from ATRT-TYR/ATRT-SHH cells, while the X-axis separates between ATRT-TYR from ATRT-SHH cells. **(I)** Dot plot showcasing the top 25 enriched GO terms based on the G2M cell cycle phase gene set available in Seurat, with color encoding for significance and size representing the ratio of genes supporting the term.

Additionally, `SCpubr` allows for the generation of color palettes based on a chosen color, ensuring all colors within the palette maintain the same contrast and brightness values while varying only in hue (`Figure 9A`). Users can also select specific combination of colors based on color theory, including opposite, adjacent, triadic, split complementary, tetradic or square color combinations (`Figure 9B`).

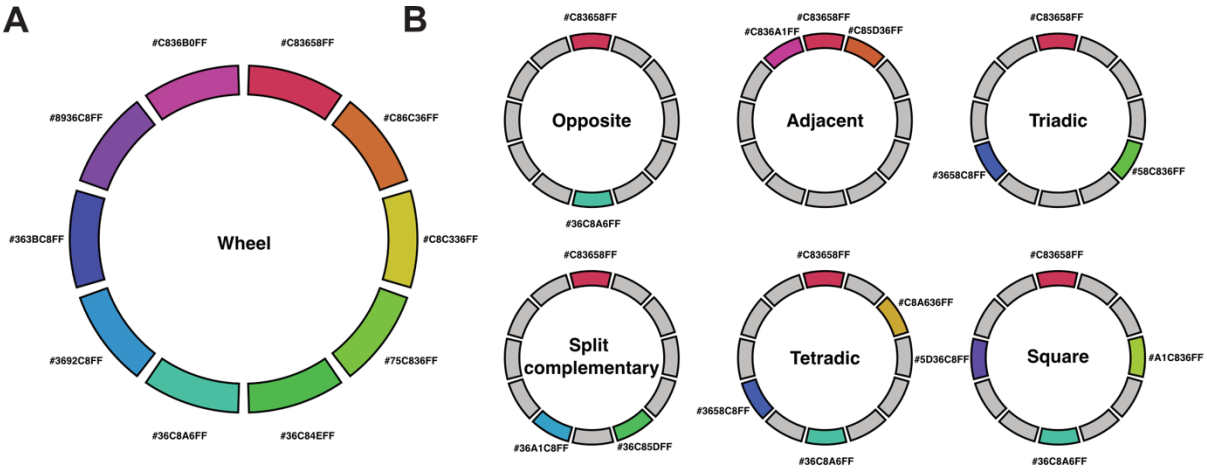


Figure 9: Generating color palettes. (A) A color palette comprising ten different colors generated across the same hue, with constant contrast and brightness values. **(B)** Various color combinations based on color theory derived from the original color palette including opposite, adjacent, triadic, split complementary, tetradic and square combinations.

Outlook

With the recent popularity of single-cell transcriptomic technologies, the repertoire of analysis and software tools available for its analysis is ever increasing. However, most of the available software tools do not offer a wide range of options for customizing data visualizations, often returning plots with default theming, requiring users to apply further style changes if desired. This can be time-consuming and often presents a technical barrier, preventing inexperienced users from performing such modifications.

Aiming to address this issue, I developed `SCpubr` in parallel with my research projects. This led to a gradual development, where new functions were introduced to meet the different needs arising in the projects. As a result, `SCpubr` provides the tools to generate most of the plots required for a publication containing single-cell transcriptomics datasets in a user-friendly manner. While allowing full customization of the resulting plots, using `SCpubr` with default parameters generates data visualizations that are minimalistic and aesthetically pleasing, suitable for publication in scientific journals.

Since its public release on CRAN in February 2022, along with its corresponding manuscript deposited in bioRxiv[133], `SCpubr` has successfully captured the attention of the single-cell community. Currently, `SCpubr` has over 12.000 downloads from CRAN, and its GitHub repository has been starred 132 times. Moreover, the manuscript has been cited 25 times, with a significant proportion of citations coming from high-impact journals. These metrics confirm the need that `SCpubr` addresses within the community. Future plans include further developing `SCpubr` and preparing a manuscript for publication in a scientific journal.

Multi-omics sequencing of atypical teratoid rhabdoid tumors unveils a rhabdoid ground-state population promoting subgroup-specific differentiation trajectories

Introduction

Malignant rhabdoid tumors (MRT) are a relatively rare but highly malignant fraction of pediatric tumors. They are typically characterized by the presence of undifferentiated cells alongside rhabdoid cells, which have the potential to differentiate into various cell lineages, including neuroepithelial, epithelial and mesenchymal cells. MRTs can arise in several anatomical regions. In the central nervous system (CNS), they are referred to as atypical teratoid/rhabdoid tumors (ATRT). MRTs can also occur in the kidney, where they are known as malignant rhabdoid tumors of the kidney, or in soft tissues, collectively termed extra-cranial malignant rhabdoid tumors (ecMRT)[137].

MRTs are characterized by biallelic inactivation of the *SMARCB1* gene in 95% of the cases. In rare cases (5%)[138,139], the *SMARCA4* gene is inactivated[6]. Whole genome sequencing (WGS) studies[140–142] have confirmed that ATRTs exhibit a relatively simple cancer genome, characterized by mutation rates of 0.19 mutations per megabase (Mb) and that the biallelic inactivation of either *SMARCB1* or *SMARCA4* gene is the sole recurrent event across patients[139,141,143,144]. *SMARCB1* is a tumor-suppressor gene that encodes a protein member of the switch/sucrose non-fermentable (SWI/SNF) complex (also known as *INI1*, *SNF5* or *BAF47*)[145], which is involved in chromatin remodeling, cell differentiation and lineage specification[146–148]. The SWI/SNF complexes, also known as BRG1/BRM-associated (BAF) complexes[148], use the energy from ATP hydrolysis to slide nucleosomes[149]. In mammals, SWI/SNF complexes are classified into three main families: canonical BAF (cBAF), polybromo-associated BAF (PBAF), and non-canonical BAF (ncBAF). While all families share a common set of core subunits, they include variable subunits encoded by multi-gene families, adding heterogeneity and slight variation in function[148,150]. On an epigenetic level, the SWI/SNF complex recruits transcription factors and chromatin-remodeling enzymes, promoting cell proliferation and differentiation[151].

In the context of tumorigenesis, up to nine SWI/SNF subunits are known to be linked to different cancer types when mutated[152]. Specifically, the inactivation of either *SMARCB1* or *SMARCA4* gene is present across all rhabdoid tumors, including ATRTs[6,153] (Figure 10). Epigenetically, the loss of SMARCB1 protein results in a global depletion of H3K27Ac (epigenetic mark associated with active gene transcription) and H3K27me3 (epigenetic mark associated with gene expression and enhancer silencing) in ATRTs[154]. However, the loss of SWI/SNF activity and the loss of H3K27ac at promoters and enhancers of genes involved in differentiation programs leads to an increase of H3K27me3 at these sites, as activity of the Polycomb-repressing complex 2 (PRC2), mediated by EZH2, a methylase subunit of the complex[155–157], is no longer inhibited at these sites by the SWI/SNF complex[147,154,158]. However, SWI/SNF activity is not completely lost in ATRTs, as there is still residual activity present at active (super)enhancers and promoters of genes involved in cell cycle regulation and oncogenesis[154]. The exact tumorigenic mechanisms driven by mutations in the SWI/SNF complex are, however, not yet fully characterized. This includes understanding the precise interplay between the SWI/SNF complex and the PCR2 complex, and how exactly *SMARCB1* inactivation affects the function of EZH2[151]. Current hypotheses suggest that these mutations might lead to either dysfunctional transcriptional regulation impacting lineage specification or defects in DNA damage repair mechanisms[148]. Further research in this area could pave the way for improved therapeutic treatments.

Notably, despite ATRTs having a relatively simple caner genome, there is a substantial clinical and molecular heterogeneity between tumor cases. Based on DNA methylation profiling and transcriptome analyses, three main molecular subtypes of ATRTs have been identified, each with distinct molecular and clinical characteristics: ATRT-TYR, ATRT-SHH and ATRT-MYC[139,144,159] (Figure 11). Furthermore, recent methylation studies have demonstrated that SMARCA4-deficient ATRTs constitute a separate subtype from the rest and should be considered as an infrequent fourth ATRT subtype (ATRT-SMARCA4)[160].

The ATRT-TYR subtype, accounting for around 34% of the cases[160], presents at a median age of 12 months[160] and predominantly locates in infratentorial regions[139]. This subtype is characterized by the enrichment of tyrosinase (*TYR*) expression, a feature not observed in other subtypes.

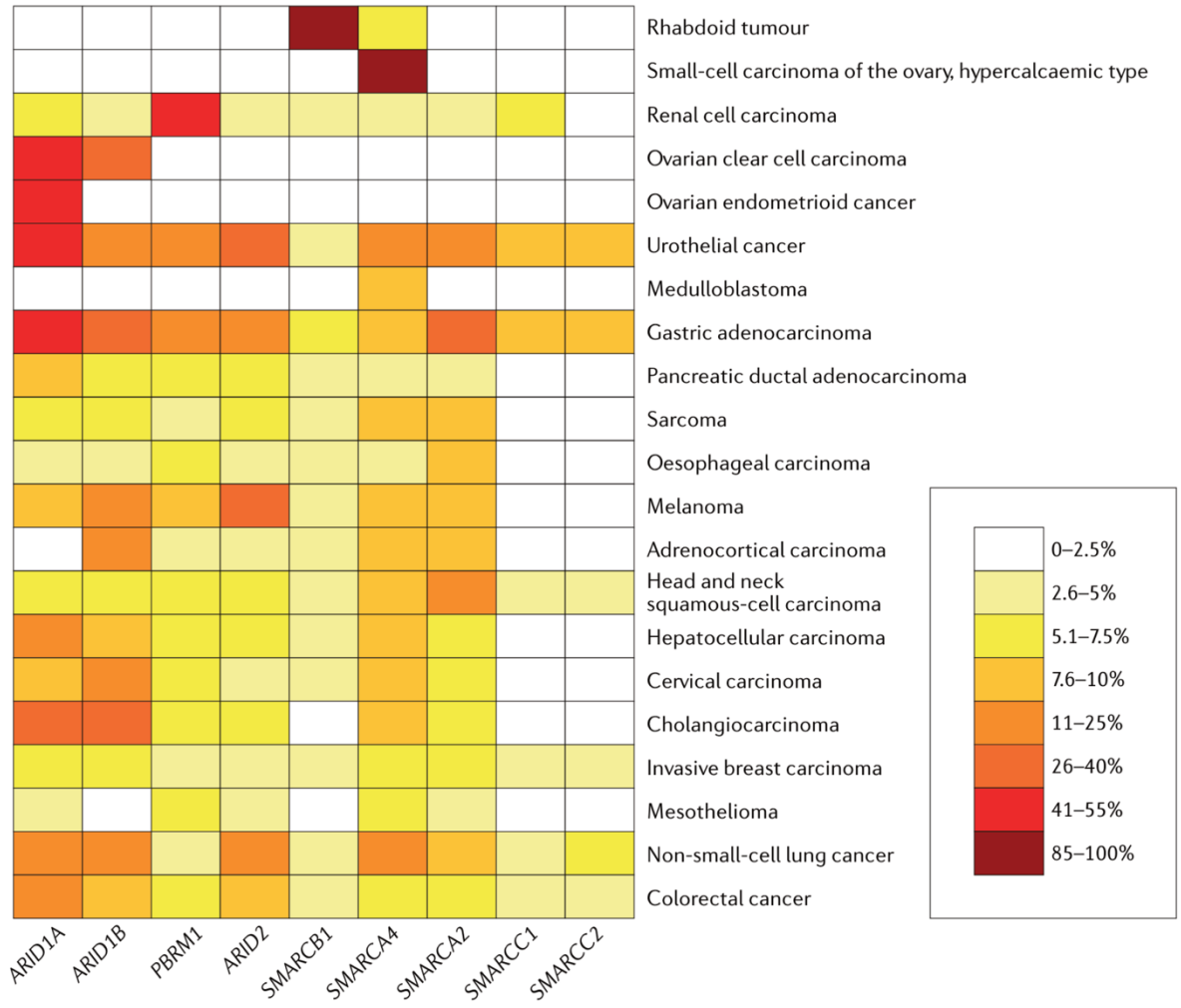


Figure 10: "Frequency and pattern of SWI/SNF subunit mutations across human cancers. The heatmap depicts the frequency of non-synonymous mutations and deletions in select genes encoding components of SWI/SNF complexes across cancer types. Overall, the figure depicts the high prevalence of mutations affecting nine SWI/SNF subunits and the context-specificity of these mutations, with most being highly enriched in certain pediatric and adult malignancies. ARID1A is the most frequently mutated SWI/SNF complex gene, followed by SMARCA4 and PBRM1." Reprinted with permission from Mittal, *et al*.[148]

While TYR expression can be used diagnostically, its role in ATRT tumorigenesis remains unclear[139]. However, its co-expression with *TYRP* and *MITF* suggest a neuroectodermal origin[161]. The methylation profiles of ATRT-TYR are highly similar to cribriform neuroectodermal tumors (CRINETs), indicating a potential shared cell of origin[162]. CRINETs are associated with better outcomes, suggesting that comparative studies of single-cell data from ATRT-TYR and CRINET tumors could provide valuable insight. Biallelic inactivation of SMARCB1 in this subtype typically occurs through a total or partial loss of chromosome 22 in one allele combined with a point mutation in the other allele. Additionally, ATRT-TYR exhibits a higher degree of open chromatin[163].

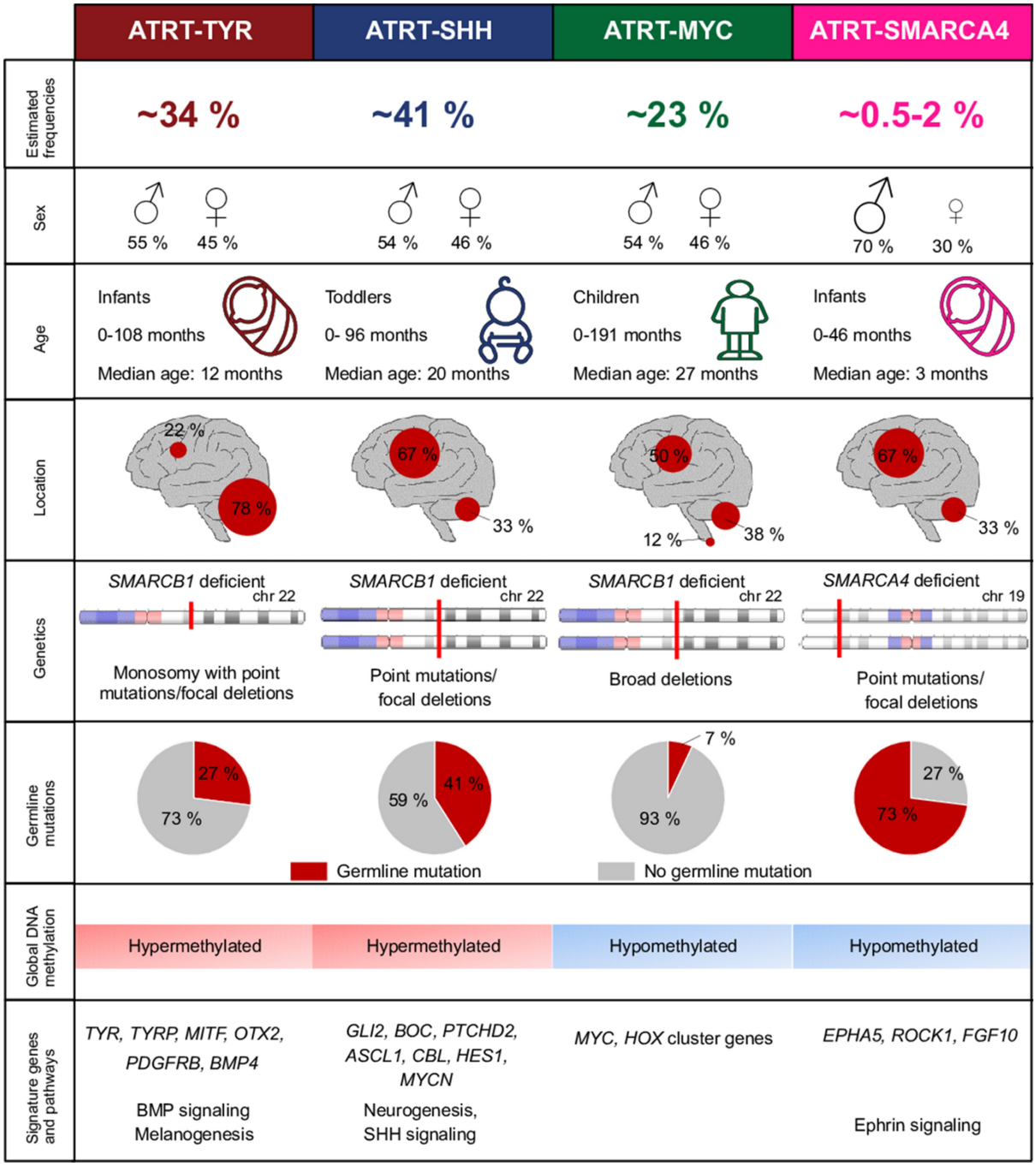


Figure 11: "Proposed model for including ATRT-SMARCA4 in the subgroup classification of ATRTs. Note that frequencies for each subgroup are based on published datasets and represent only a rough estimation. Estimated frequencies of SMARCB1-deficient ATRT subgroups (n=321), their male to female ratios (n=82 for ATRT-TYR, n=105 for ATRT-SHH, n=56 for ATRT-MYC), age (n=62 for ATRT-TYR, n=72 for ATRT-SHH, n=43 for ATRT-MYC), and locations (n=68 for ATRT-TYR, n=91 for ATRT-SHH, n=48 for ATRT-MYC) as well as information regarding genetics, signature genes and pathways is based on the study by Ho *et al.*[139] Frequencies of germline mutations was taken from Frühwald *et al.*[164] Frequencies of ATRT-SMARCA4 is estimated based on studies published by Johann *et al.*[144] and Frühwald *et al.*[164] Information concerning the sex ratio (n=19), age (n=19), location (n=19), and germline mutations (n=10) of ATRT-SMARCA4 are taken from the study presented here and published reports[6,165].Genetics, global DNA methylation levels as well as signature genes and pathways of ATRT-SMARCA4 are based on the here generated results. Design is inspired by the model proposed by Ho *et al.*[139]" Reprinted with permission from Holdhof, *et al*[160].

The ATRT-SHH subtype, which accounts for around 41% of the cases[160], presents at a median age of 20 months[160]. This subtype is characterized by the overexpression of members of the SHH and NOTCH pathways, including *GLI2*, *PTCH1* and *BOC* for the SHH pathway and *ASCL1*, *HES1* and *DTX1* for the NOTCH pathway[139]. Gene set enrichment analysis (GSEA) suggests a neuronal cell of origin for this subtype[139]. ATRT-SHH exhibits inactivation of *SMARCB1* gene through point mutations and focal deletions[139,160]. Methylation profiling reveals two major subtypes within ATRT-SHH: ATRT-SHH-1 and ATRT-SHH-2, with recent studies further dividing ATRT-SHH-1 into ATRT-SHH-1A and ATRT-SHH-1B[166]. ATRT-SHH-1A and ATRT-SHH-1B mainly localize supratentorially, while ATRT-SHH-2 is found infratentorially, with potential extension into the pineal region[139,166]. While all three ATRT-SHH subtypes exhibit overexpression of the SHH and NOTCH pathways, there are notable differences: ATRT-SHH-1B is enriched for expression of the proneural marker *ASCL1*, whereas ATRT-SHH-2 lacks expression of glial markers such as *OLIG2* and *GFAP*[166]. In terms of survival, older patients (> 3 years) with the ATRT-SHH-1B subtype tend to have more favorable outcomes compared to the other two ATRT-SHH subtypes[166]. Despite these findings, further analyses are needed to fully characterize the differences between the three ATRT-SHH subtypes. It has been shown that inactivation of *SMARCB1* can activate the SHH pathway[167], but it remains unclear why this activation appears to be restricted to the ATRT-SHH subtype[139]. A possible explanation could be the potentially different cells of origin of the ATRT subtypes.

The ATRT-MYC subtype, accounting for around 23% of the cases[160], presents at a median age of 27 months[160] and predominantly localizes in supratentorial regions, but can also be found in the spine[139,160]. This subtype is distinguished by *MYC* overexpression. Additionally, ATRT-MYC tumors exhibit overexpression of HOXC cluster genes[144], indicating a likely mesenchymal cell of origin[163]. These tumors are characterized by the inactivation of *SMARCB1* via focal deletions, which can span several hundred kilobases[144,163], while point mutations as an inactivation mechanism are mostly absent. The methylation profiles of ATRT-MYC tumors show similarities to extra-cranial malignant rhabdoid tumors (ecMRTs), suggesting a potential shared cell of origin[159]. Along with the ATRT-SMARCA4 subtype, the ATRT-MYC subtype exhibits an overall hypomethylated state[139,159,160], which is known to be indicative of poor prognosis[168–170].

The ATRT-SMARCA4 subtype, accounting for around 0.5-2% of the cases[160], presents at a median age of three months[160] and predominantly locates supratentorially[160]. This subtype is characterized by either homozygous nonsense or missense inactivation of *SMARCA4* and an overall hypomethylated phenotype[160]. This loss may lead to tumor development and progression through the activation of proto-oncogenes and cancer-germline genes[171,172], possibly explaining why ATRT-SMARCA4 subtype is more aggressive than the SMARCB1-derived subtypes[160]. At the transcriptional level, ATRT-SMARCA4 exhibits upregulation of *EPHA5*, *ROCK1* and *FGF10*, alongside downregulation of *DMRT2*. This results in the enrichment of the *Ephrin forward signaling* pathway, which is crucial for CNS development and is commonly altered in various cancer types[173]. Potential treatment avenues specifically for ATRT-SMARCA4 subtype might involve targeting EPHA5, as has been tested for other pathologies[174–176].

Overall, the cell of origin of ATRTs remain largely elusive, with only hypotheses available. Furthermore, instances of SMARCB1-deficient ATRTs emerging in the context of other tumors such as ependymoma, high-grade glioma and low-grade glial tumors have been described, suggesting the potential for ATRTs to progress from other tumor entities[177–180]. Treatment strategies for ATRTs depend on factors such as tumor location, initial staging, and patient age, with varying outcomes across subtypes. Current therapeutic approaches typically involve multimodal methods combining surgery, radiotherapy and chemotherapy. Nonetheless, fact remains that the overall survival rate for patients with ATRT tumors averages around 17 months[181–183], highlighting the urgent need for more effective and subtype-specific treatments.

In light of current knowledge, the need to delve deeper into the inherent heterogeneity underlying ATRT subtypes, including those within ATRT-SHH, becomes evident. Crucial tasks such as delineating the cell of origin of these subtypes, determining whether it is shared or distinct among them, linking various cell populations to tumor developmental hierarchies, characterizing the enriched pathways and transcriptional factors in different tumor cell lineages, and exploring the interplay between the tumor microenvironment-derived cell populations and the distinct tumor cell types, all require further research. Such studies hold promise for uncovering novel therapeutic targets aimed at improving survival rates of ATRT patients.

To unveil the answers to these questions, I have been granted access to an extensive collection of ATRT single-cell datasets derived from both in-house sequencing efforts and international collaborations. As a discovery cohort, I possess 19 datasets: six ATRT-TYR, ten ATRT-SHH and four ATRT-MYC, all derived from single-nuclei (frozen tissue) using 10X v3 3' technology (n = 12), as well as 10X multiome (RNA + ATAC, n = 7), resulting from a collaboration with Dr. Jarno Drost at the Princess Maxima Center in Utrecht, the Netherlands. Additionally, in collaboration with Sam Behjati at the Wellcome Sanger institute in the UK, I possess one ATRT-TYR single-cell (fresh tissue) 10X v3 3' dataset, ideal for trajectory analysis given that current state-of-the-art tools for such analyses are designed for whole cells rather than just nuclei[184]. For validation purposes, in collaboration with Mariella Filbin at the Dana Farber Cancer Institute in Boston, I have obtained eight SMARTseq2 datasets, comprising three single-cell (ATRT-TYR = 1, Not known = 2) and five single-nuclei datasets (ATRT-TYR = 1, ATRT-SHH = 1, ATRT-MYC = 3). Lastly, for functional validation, I have at my disposal five ATRT cell-line single-nuclei 10X 3'datasets.

Materials and Methods

This section provides an overview of the datasets and methods employed in the project. Each subsection will contain a final statement of the people involved in the experiments. All bioinformatic analyses, unless specified otherwise at the end of each subsection, were conducted by me. Throughout all analyses and data interpretation, I received supervision from Prof. Dr. Marcel Kool, initially also from Dr. Natalie Jäger and Dr. med. Pascal Johann, and later in the project also from Dr. Jarno Drost at the Princess Máxima Center in Utrecht, who is one of the main collaborators on the project. While I wrote the original text of this chapter, I used ChatGPT to enhance its readability.

Ethical statement

Informed consent was obtained in written form from all patients or their respective legal guardians. Patient samples were acquired under the ethical approval of the ethics committee of the DKFZ or Princess Máxima Center, respectively. Approval for Máxima samples and clinical data within the scope of this study was obtained by the Máxima biobank and data access committee (biobank request nr. PMCLAB2018.005).

Sample processing and nuclei isolation

For single-nucleus RNA-seq (snRNA-seq, n = 12), sequencing was conducted by Aniello Federico and Monika Mauermann following established protocols[185]. Briefly, each sample was thawed, sectioned into smaller pieces (1-2mm^3), and minced thoroughly in ice-cold “CHAPS, with salts and Tris” (CS) digestion buffer for five minutes. The homogenized tissues were then filtered twice through 40 µm cell strainers and washed three times with “salt-Tris” (ST) detergent buffer. Subsequently, the samples were centrifuged for five minutes at 500g at 4 °C and the nuclei pellets were resuspended in PBS + 0.05% BSA and counted using the Luna Automated cell counter (Logos Biosystems). If visible agglomerates and/or cell debris were present, samples underwent an additional filtration step using Flowmi 40 µm cell strainers. Approximately 20.000 nuclei per sample were loaded into the Chromium single-cell 3’chip (10X Genomics), where each cell and its respective transcriptome were individually partitioned and barcoded. The steps for cDNA amplification and gene expression library preparation were carried out using the Chromium Next GEM Single Cell 3’ Reagent Kits v3.1 (10X Genomics),

following the manufacturer's instructions. The concentration and quality of cDNA and libraries were assessed using the Qubit dsDNA HS Assay Kit (ThermoFisher Scientific) and TapeStation (Agilent). Equimolar pooled libraries (multiplexes) were sequenced on a NovaSeq 6000 (Illumina) sequencer, according to the manufacturer's instructions.

For single-nucleus multiome samples (n = 7), tissues were processed using manufacturer's standard procedures (Chromium v3, 10x Genomics protocol (CG000338 Rev D)) by Jarno Drost's team at the Princess Máxima Center. This involved mincing and homogenizing tissues using a dounce tissue grinder. After cell lysis, the samples underwent filtration using a 70 µM filter followed by a 40 µM filter. Intact nuclei were sorted based on 7AAD positivity and size of the nuclei.

Single-nucleus RNA-seq analysis

Read alignment and count matrix generation

Sequencing reads were utilized to generate a count matrix using `cellranger count` (v7.0.0)[186]. This version of cellranger inherently considers introns from the reference transcriptome (GRCh38). The count matrices obtained through this method were imported into `R` (v4.2.0) and further processed using `Seurat` (4.9.9.9045)[23,187].

Quality control of snRNAseq data

To ensure the exclusion of low-quality cells and droplets, stringent cutoffs were applied according to standard guidelines[25]. The following metrics were taken into consideration: total number of unique molecular identifiers (UMI) per cell (`nCount_RNA`), total number of genes per cell (`nFeature_RNA`) and percentage of mitochondrial RNA per cell (`percent.mt`). Cells were filtered out if they exhibited `nCount_RNA < 100`, `nFeature_RNA < 500` and `percent.mt > 5%`. Additionally, cells falling outside the mean plus three standard deviations of the distribution of UMI and genes were also excluded, resulting in a dataset containing only high-quality cells.

Doublet removal

To identify and eliminate doublets from the data, `scrublet`[188] (v0.2.1) was employed. As this is a python-based package, it was imported into R using `reticulate` via the

`scrublet$Scrublet()` function. The output of `scrublet` exhibited a bimodal distribution, based on which the threshold to classify a cell as doublet was determined as the local minimum between the two modes of the distribution. This process was conducted independently for each patient, and the cells identified as doublets were subsequently removed from the analysis.

Normalization

To normalize the raw counts, the `Seurat::NormalizeData()` function was applied using default parameters, conducted on a patient-by-patient basis. Subsequently, datasets originating from different patients were merged into a single dataset, and highly variable genes were selected using the `Seurat::FindVariableFeatures()` function. The normalized counts were then scaled using the `Seurat::ScaleData()` function while regressing out the effect of number of UMIs (`nCount_RNA`), genes (`nFeature_RNA`) and percentage of mitochondrial RNA (`percent.mt`) per cell.

Dimensional reduction

Dimensionality reduction was performed by first computing principal component analysis (PCA) via the `Seurat::RunPCA()` function, with the normalized data as the basis for the method and default parameters. A total of 50 principal components (PC) were generated as the output of the method, out of which the top 25 were selected for downstream analysis. This selection process involved inspecting the amount of standard deviation represented by each principal component using the `Seurat::ElbowPlot()` function. Following PCA, uniform manifold approximation and projection[50] (UMAP) was computed using the `Seurat::RunUMAP()` function, also under default parameters.

Data integration

Data integration was performed using the `Harmony`[56] package, via the function `harmony::RunHarmony()`. This was based on the PCA reduction embedding, and the method generated a new embedding known as Harmony reduction. In this new embedding, the effect of cells originating from different patients and sequencing technologies was mitigated. The method was run with default parameters, except for setting `theta` as `1` for patients and `2` for sequencing technologies.

Clustering and cell type annotation of the Tumor Microenvironment

To identify cells and clusters that were not malignant, the composition of patients per cluster was examined using `SCpubr`[133] package via the function `SCpubr::do_BarPlot()`. Clusters composed of cells originating from the majority, if not all, patients were designated as tumor microenvironment (TME). Conversely, clusters predominantly consisting of cells from a single patient were annotated as malignant cells. Additionally, chromosome 22 loss was assessed using copy number variant (CNV) analysis via the `inferCNV`[95] package, with TME clusters serving as reference. The following parameters were applied: `cutoff = 0.1`, `min_cells_per_gene = 3`, `HMM = TRUE`, `HMM_type = "i6"` and `window_length = 201`. To enhance the sensibility of the analysis, metacells were computed, following the method outlined in previous publications[189]. Clusters presenting a chromosome 22 loss were classified as malignant cells, in line with published data that chromosome 22 loss is most common in ATRT-TYR tumors[139].

For annotating the TME clusters, enrichment scores were calculated using the `UCell`[190] package. Various TME cell populations commonly found in tumors were selected, and their corresponding marker gene sets were retrieved from PanglaoDB[191]. Enrichment scores were then visualized as feature plots using `SCpubr::do_FeaturePlot()` and also as enrichment heatmaps using `SCpubr::do_EnrichmentHeatmap()`.

Annotation of tumor cells

A three-step workflow was devised to characterize the different tumor subpopulations within the tumor cells. Initially, supervised annotation was performed based on selected gene sets sourced from publications. The gene sets comprised differentially expressed genes from the cell populations identified in the human fetal brain atlas[192], pan-cancer recurrent cell states[104] derived from non-negative matrix factorization (NMF), and choroid plexus markers obtained from PanglaoDB. Enrichment scores for these gene sets were computed using `UCell`, and transformed into an Assay object compatible with `Seurat`, serving as input for the previously described dimensional reduction workflow. The enrichment scores were then visualized in the context of the new UMAP embedding by plotting them as feature plots with `SCpubr::do_FeaturePlot()` and as heatmaps using `SCpubr::do_EnrichmentHeatmap()`. Clusters exhibiting unique enrichment for a particular gene set were subsequently annotated.

Cells not annotated in the previous step underwent unsupervised annotation using NMF, following previous publications[189]. In summary, NMF was applied independently for each patient, yielding NMF programs that were correlated and clustered using Pearson's correlation. Highly correlated NMF programs were grouped into NMF metaprograms, from which the top 30 scoring genes were extracted and used for annotation. Consequently, clusters were annotated based on the NMF metaprogram in which they were uniquely enriched.

Finally, the remaining cells underwent a new round of clustering, and marker genes for each of the newly computed clusters were identified using `COSGR`[193]. The top differentially enriched genes were analyzed for functional annotation using `Metascape`[194]. In cases where no strong association with a specific phenotype was evident, clusters were annotated based on the ATRT subtype to which they belonged.

Stemness activity analysis

To characterize tumor cell populations based on their activity in different stem cell gene marker sets, `decoupleR`[85] was used. This tool calculates a cell-wise activity score based on prior knowledge networks, which can be generated based on specific gene sets or retrieved from public sources. In this case, marker gene sets for pluripotent stem cell (PSC), embryonic stem cell (ESC) and neuronal stem/precursor cell (NSPC), originating from PanglaoDB Extended 2021 database hosted in `EnrichR`[195] were retrieved and used to build the prior knowledge network. The mode of regulation of the custom network was set to `1`, as all the genes were considered marker genes for their respective cell types. The inferred activities were visualized as a heatmap using `SCpubr::do_AffinityAnalysisPlot()`.

Panel design for spatial transcriptomics using Xenium

In addition to the genes included in the commercially available Xenium Multi-Tissue and Cancer panel (377 genes)[196], a custom panel for spatial transcriptomics was created by incorporating 100 additional genes. To select these additional genes, differential expression analysis across all cell populations was performed using `COSGR`[193], enforcing a 25% of representation of the gene in the cell type. Mitochondrial, ribosomal, and long non-coding genes, as well as alternative spliced variants, were filtered out. The top four DE genes per cell population were then retrieved, resulting in a total of 76 genes. Additional genes representative of different

immune fractions and TFs were included by Jarno Drost's team in the Princess Máxima Center to reach a total of 100 add-on genes. In total, the designed panel comprised 477 genes.

Results

To elucidate distinct cell states within ATRT tumors across each of the ATRT subtypes, single nucleus RNA sequencing (snRNAseq, Marcel Kool's laboratory) and multiome (RNA + ATAC, Jarno Drost's laboratory) on primary patient ATRT tissues (Figure 12) was conducted. My role in this project primarily involved analyzing the RNA data, while Irene Paasen and Jiayou He are currently analyzing the ATAC data and performing wet-lab validations.

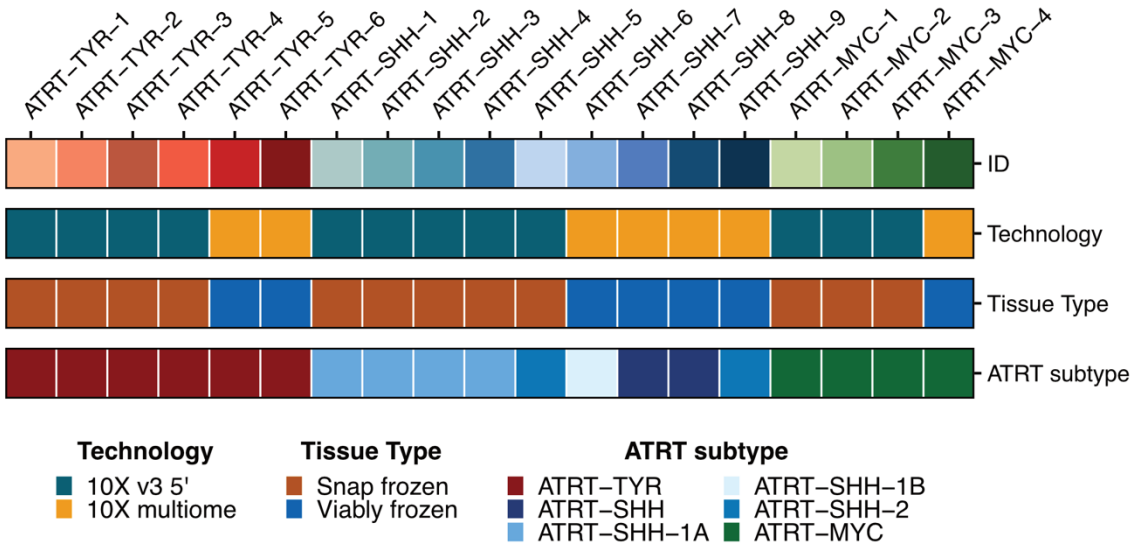


Figure 12: Metadata ATRT discovery cohort. Molecular characteristics of the ATRT discovery cohort are presented. Not all ATRT-SHH patients received an ATRT-SHH subtype diagnosis, therefore called as "ATRT-SHH".

Initially, my focus was on comprehending the nature of the dataset by examining the metadata associated with the patients. The cohort comprised of 19 distinct tumor samples spanning the three ATRT subtypes (ATRT-TYR = 6, ATRT-SHH = 9, ATRT-MYC = 4), utilizing various sequencing techniques (10X 5' v3 = 12, 10X multiome = 7) and originating from different tissue types (Viably frozen = 12, snap frozen = 7). This dataset represents a significant contribution to the field of pediatric neurooncology, particularly given the absence, at the time of writing, of comprehensive single-cell transcriptomics cohorts for the different ATRT subtypes.

Identifying tumor cells apart from tumor microenvironment

To distinguish between tumor and tumor microenvironment cells, stringent quality control was enforced to filter out low quality cells, following recommendations from previous studies[25]. Subsequently, total of 36.601 nuclei passed quality control and underwent normalization according to standard guidelines[34]. Upon dimensional reduction through PCA followed by

UMAP and clustering, a total of 21 distinct clusters were identified. To initiate the cluster labelling process, I opted to inspect the clusters to determine the ones belonging to the tumor microenvironment (TME) and which ones were tumor cells. Various approaches can be employed for this purpose. Typically, TME cells tend to aggregate into a single cluster before integration, whereas the tumor fraction predominantly remains as a patient-specific cluster. To verify this pattern, I examined the composition of patients per cluster (Figure 13A-B). The results suggested that clusters 13, 16, 19 and 20 exhibited a multi-patient composition, while the remaining clusters were predominantly comprised of cells from a single patient.

In addition, copy number variant (CNV) profiles of the tumor cells can be inferred using the TME cell clusters as a reference. Through this analysis, I anticipated identifying a chromosome 22 loss in tumor cells belonging to ATRT-TYR subtype (Figure 14A-C), consistent with previous studies[139,197]. Additionally, projecting chromosome 22 CNV scores onto the merged UMAP revealed a distinct pattern specific to tumor cells of the ATRT-TYR subtype, corresponding to tumor cells (Figure 14C). However, distinguishing tumor cell populations from ATRT-SHH or ATRT-MYC subtype through CNV analyses is not straightforward, as tumors in these subtypes exhibit no CNVs[139].

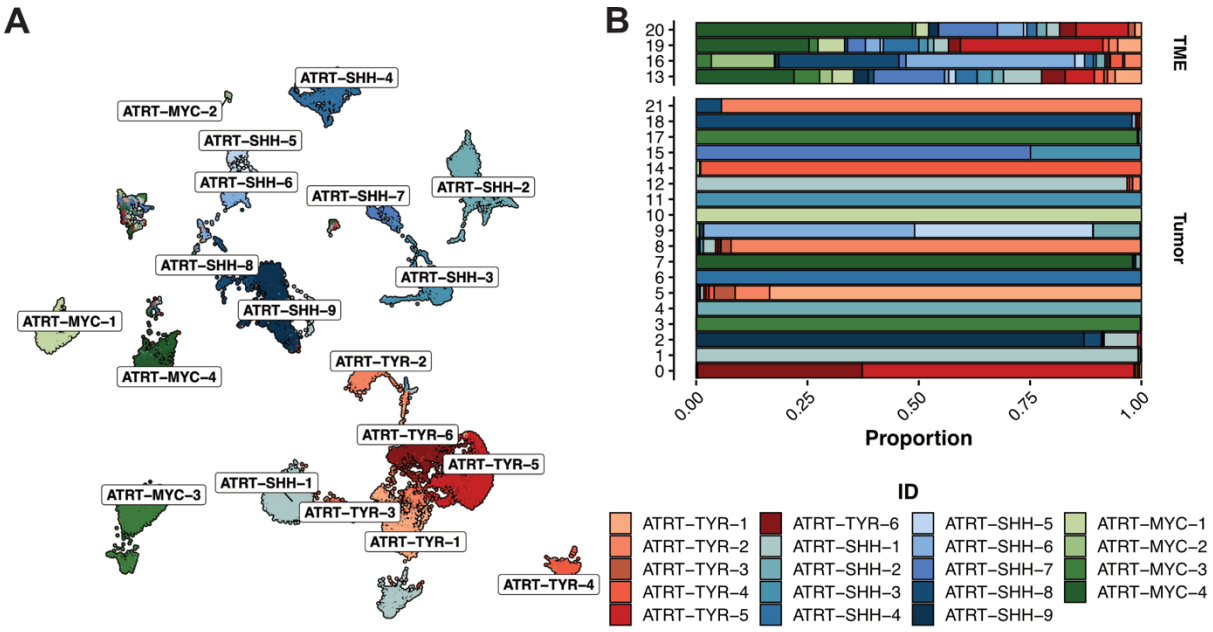


Figure 13: Identifying cell clusters shared across patients. (A) UMAP representation colored by patient. UMAP1, x-axis; UMAP2, y-axis. **(B)** Patient composition per each cluster identified prior to data integration. Clusters located at the top belong to tumor microenvironment and clusters in the bottom group are tumor cells.

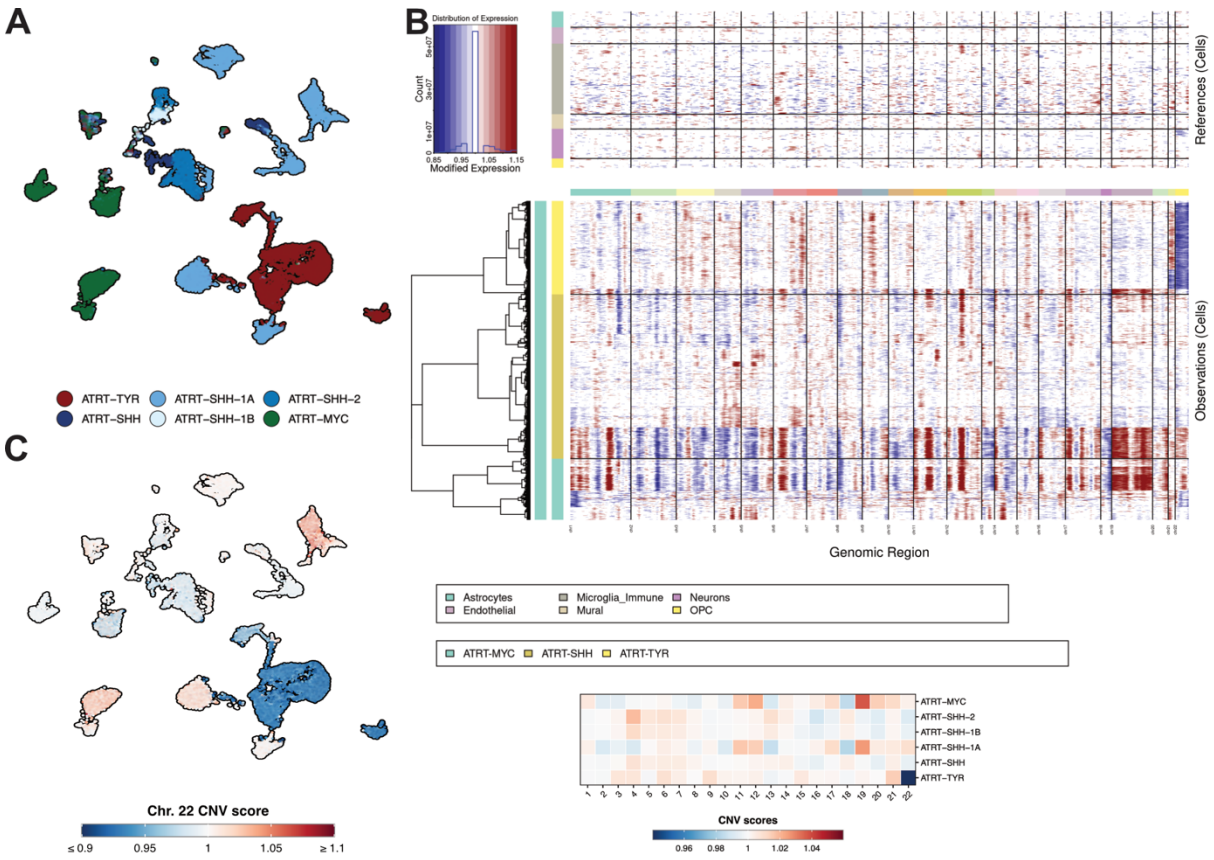


Figure 14: Chromosome 22 loss identifies tumor cells in ATRT-TYR. (A) UMAP representation of ATRT single-cell cohort prior to integration colored by ATRT subtype. UMAP1, x-axis; UMAP2, y-axis. **(B)** Inferred copy number variant profile using cells from tumor microenvironment as a reference (top). CNV scores are further aggregated by ATRT subtype and chromosome and displayed as a heatmap (bottom). Scores around 1 mean no CNV, being higher associated with a chromosome gain and lower with a loss. **(C)** UMAP representation of ATRT single-cell cohort where chromosome 22 scores (see B) have been mapped onto. Chromosome 22 loss can be observed in the regions of the UMAP where ATRT-TYR cells are located (see A). UMAP1, x-axis; UMAP2, y-axis.

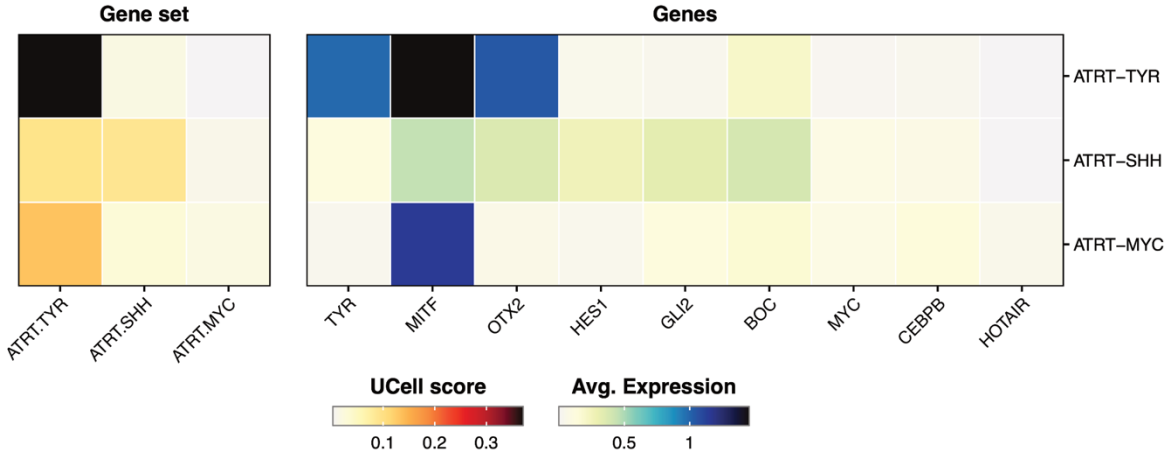


Figure 15: Querying ATRT bulk RNA-seq literature markers on single-cell data. Enrichment (left) and expression (right) heatmap of literature ATRT subtype marker gene sets[139,198].

Following this, I hypothesized that another method to delineate the tumor cells within each ATRT subtype could be computing enrichment scores based on gene marker sets previously identified from bulk RNA sequencing studies on ATRTs[139,198]. Analyzing the enrichment trends of clusters in marker gene sets specific of specific ATRT subtypes (Figure 15) revealed a clear expression pattern for ATRT-TYR-specific marker genes. In contrast, ATRT-SHH and ATRT-MYC markers exhibited more diffuse expression profiles within their respective subtypes. These findings underscore a general lack of specificity of literature-based marker sets derived from ATRT bulk transcriptomics datasets in my single-cell data.

Next, I aimed to gain insights into the biological characteristics of the TME clusters. To achieve this, I retrieved various sets of marker genes for different TME cell populations from PanglaoDB[191] and generated enrichment scores. This analysis revealed a total of six distinct TME populations within my datasets: astrocytes, endothelial cells, microglia and immune cells, neurons, oligodendrocyte precursor cells (OPC) and pericytes (Figure 16A-B).

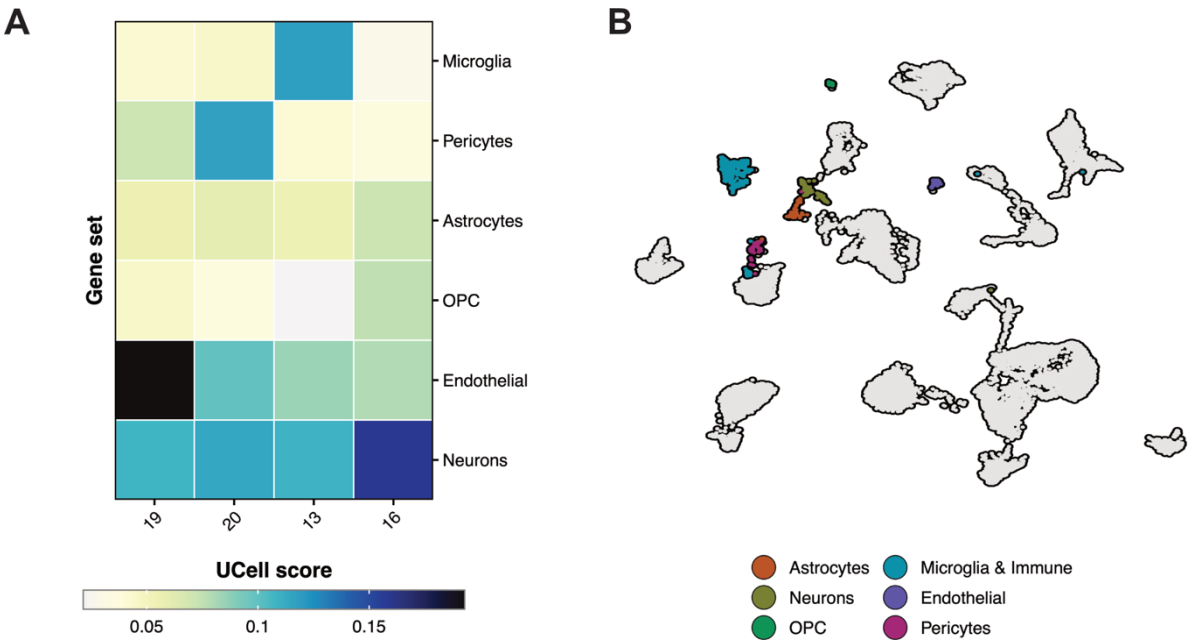


Figure 16: Annotation of TME clusters based on reference marker sets. (A) Enrichment heatmap of literature marker gene sets[191] for tumor microenvironment cell types for a selection of cell clusters shared across patients. **(B)** UMAP representation colored by tumor microenvironment clusters. Grey cells are tumor cells. UMAP1, x-axis; UMAP2, y-axis.

Finally, to exclude batch effects from impacting dimensional reduction, I integrated the dataset to regress out the effects of patient identity and sequencing batch. This resulted in a UMAP reduction where clustering was based on ATRT subtypes while maintaining TME clusters together. The UMAP revealed three main branches, each representing one of the ATRT subtypes, supporting the presence of subtype-specific expression signatures (Figure 17A-B). However, in one of the branches, I observed a mixture of ATRT-MYC and ATRT-SHH cells. Additionally, at the center of the UMAP, there was a region where the three ATRT subtypes converged, and subtype-specific signatures appear to overlap.

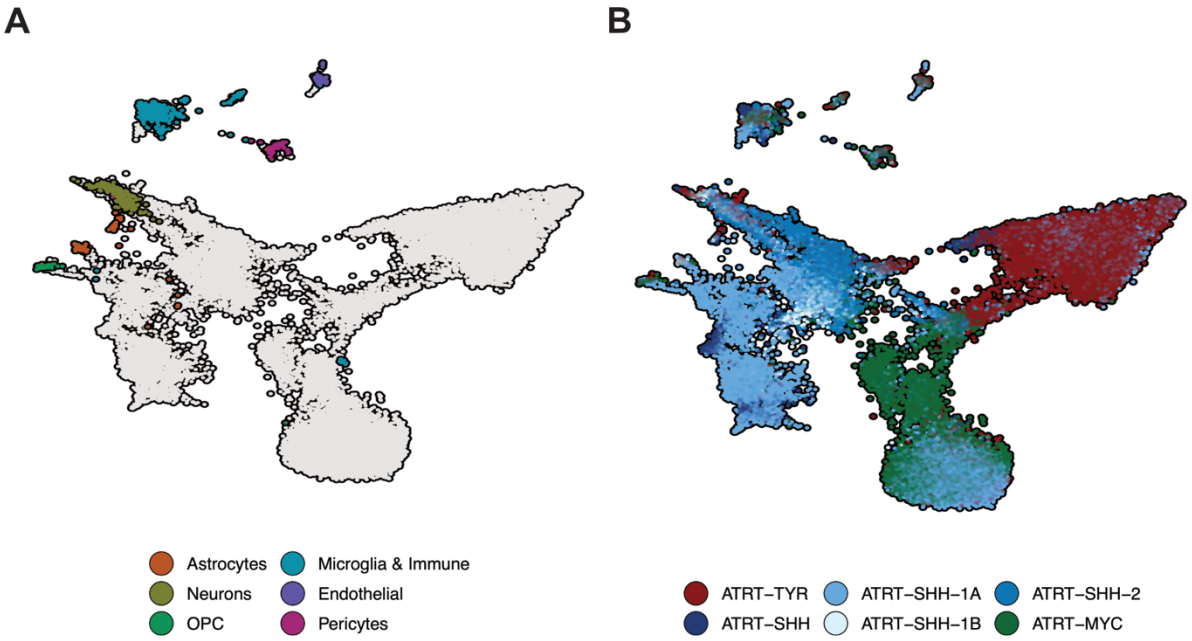


Figure 17: Integration with Harmony. (A) Integrated UMAP representation colored by tumor microenvironment clusters. Grey cells are tumor cells. UMAP1, x-axis; UMAP2, y-axis. **(B)** Integrated UMAP representation colored by ATRT subtype. UMAP1, x-axis; UMAP2, y-axis.

Characterization of tumor lineages: a supervised and unsupervised pipeline

Next, I explored whether the expression patterns of tumor cells resembled those of normal cell types. To achieve this, annotated TME cell types were removed from the analysis, and tumor cells were annotated based on an annotation pipeline I developed (Figure 18). This pipeline involved several stages. Initially, supervised annotation on the tumor cells was conducted based on a collection of literature-derived marker gene sets. The remaining tumor cells underwent supervised annotation through first the computation of non-negative matrix factorization (NMF), resulting in the identification of NMF metaprograms.

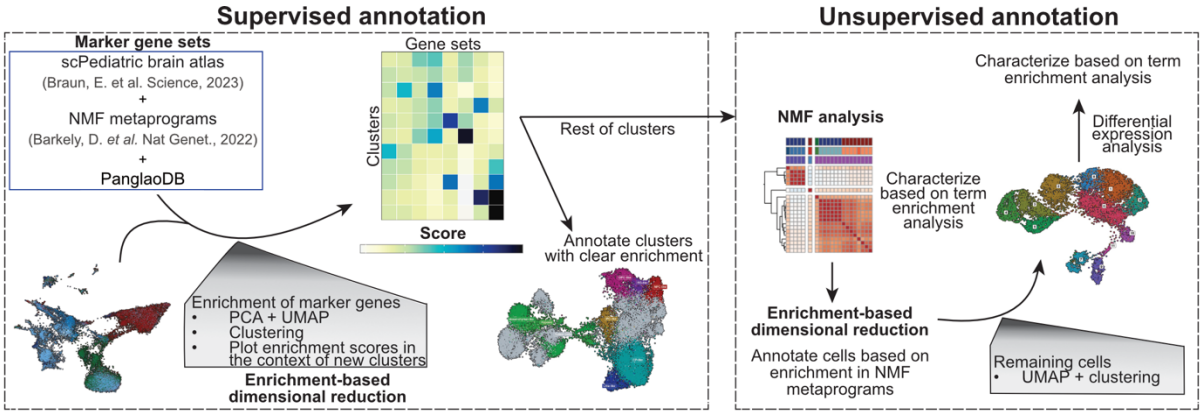


Figure 18: Tumor annotation workflow. Schematic overview of tumor cell annotation workflow. Supervised annotation based on literature marker gene sets (left) following by unsupervised annotation via NMF and differential expression analysis of the remaining, not annotated cells (right). Illustration made by Irene Paassen and modified by me.

Subsequently, unannotated cells were then reclustered, and differentially expressed genes were identified for each cluster. In both variants of unsupervised annotation, I attempted to assign a phenotype to each gene marker set based on Gene Ontology (GO) enrichment analysis. In the supervised annotation step (`Figure 18-left`), I curated literature-derived gene sets that encompassed phenotypes related to fetal development[191,192] and recurrent tumor programs across different cancer types[104]. These gene sets were included as they offered potential phenotypes that might also be observed in ATRTs. My rationale was that the enrichment on these gene signatures could serve as a foundation for clustering and dimensional reduction, leading to clusters with distinct enrichment patterns.

Using this approach, I annotated the clusters as “alike” due to their distinctive enrichment in one or several gene signatures (`Figure 19A-C`). This method unveiled different developmental trajectories across ATRT subtypes: ATRT-TYR cells exhibited choroid plexus-like and cilia-like tumor populations; ATRT-SHH presented radial glia (RG)-like cells, branching into neuronal precursor cell (NPC)-like and oligodendrocyte precursor cell (OPC)-like tumor populations; and ATRT-MYC showcased a mesenchymal-like tumor population (`Figure 19D`). Moreover, the cilia-like population, although present in some ATRT-SHH patients, was entirely absent in ATRT-MYC cases (`Figure 19E`).

Notably, I identified a tumor population shared across all ATRT subtypes with expression signatures resembling those of neuronal intermediate precursor cells (defined as neuroblast-like cells with cycling activity), which I termed IPC-like. Additionally, I identified a cluster of hypoxic cells, primarily mapping to a single patient. Previously, it has been shown that ATRT-

SHH can be further subclassified into ATRT-SHH-1A, ATRT-SHH-1B and ATRT-SHH-2[166]. Interestingly, the majority of NPC-like cells belonged to the ATRT-SHH-2 subtype, while the OPC-like population was predominantly composed of cells from ATRT-SHH-1A (Figure 19F). ATRT-SHH-1B cells were not represented in the NPC-like nor OPC-like populations, but instead were primarily identified as a distinct NMF metaprogram (see below).

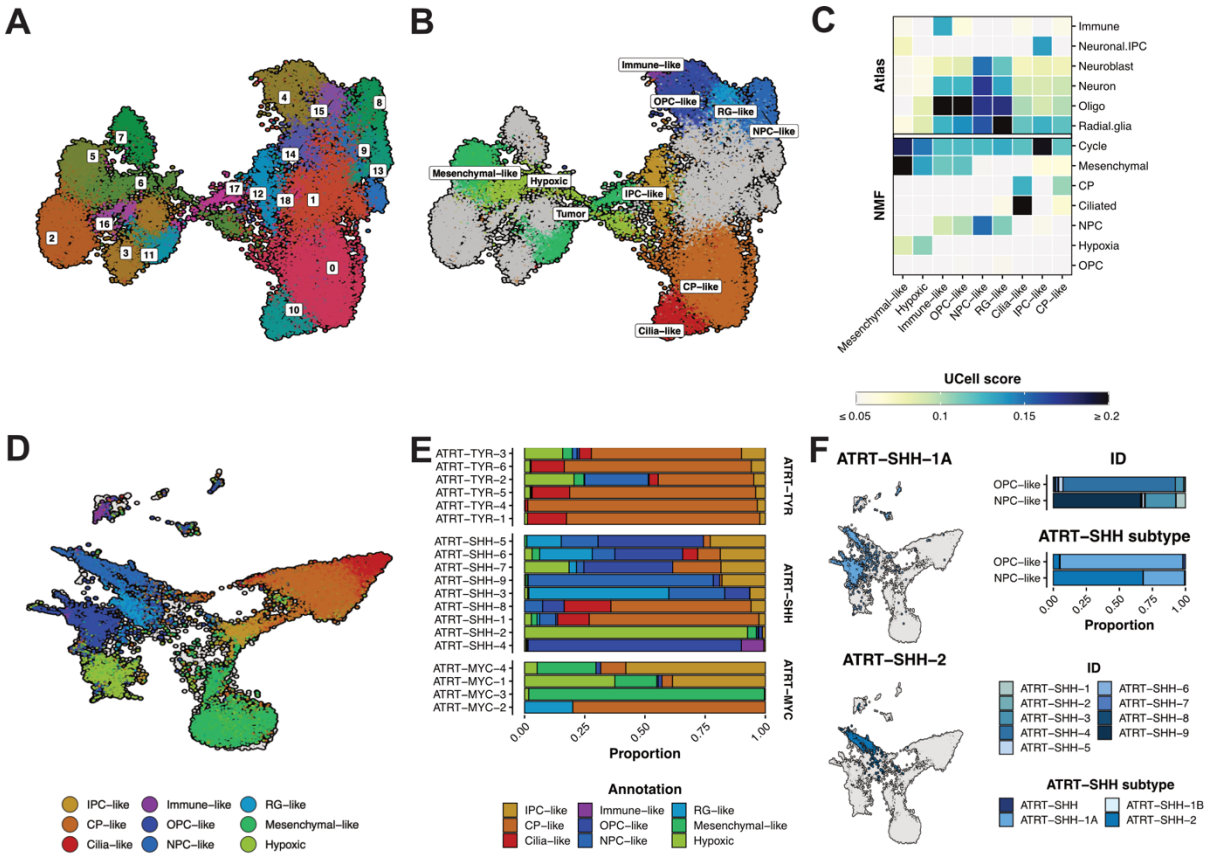


Figure 19: Supervised annotation of tumor cells. (A and B) UMAP representation computed based on enrichment scores in the supervised annotation gene sets, colored by original integration clusters **(A)** and inferred annotation derived from the supervised step in the annotation workflow **(B)**. UMAP1, x-axis; UMAP2, y-axis. **(C)** Heatmap depicting enrichment scores for a representative subset of gene sets used in the supervised annotation step. Choroid plexus markers (CP) were retrieved from PanglaoDB database[191]. Rows: tumor cell populations in **(B)**, columns: gene set. Color gradient is subset to comprise [0.05, 0.2]. OPC: oligodendrocyte precursor cell; NPC: neuronal precursor cell; RG: radial glia; IPC: neuronal intermediate precursor cell; CP: choroid plexus. **(D)** Integrated UMAP representation highlighting tumor cell types. UMAP1, x-axis; UMAP2, y-axis. OPC: oligodendrocyte precursor cell; NPC: neuronal precursor cell; RG: radial glia; IPC: neuronal intermediate precursor cell; CP: choroid plexus. **(E)** Tumor population composition per patient from the cell types resulting from the supervised annotation step. Bars are grouped by ATRT subtype and ordered based on decreasing proportion of IPC-like cells. OPC: oligodendrocyte precursor cell; NPC: neuronal precursor cell; RG: radial glia; IPC: neuronal intermediate precursor cell; CP: choroid plexus. **(F)** Integrated UMAP representation of ATRT-SHH-1A and ATRT-SHH-2 across NPC-like and OPC-like populations (left) and the ATRT-SHH subtype and patient composition of these tumor populations (right).

The fraction of tumor cells that remained unannotated underwent unsupervised annotation (Figure 18-right). This process began with non-negative matrix factorization (NMF) (Figure 20A-D), and for any cells still unannotated, continued through re-clustering followed by differential expression analysis (Figure 20E). Correlation analysis on the NMF programs revealed a total of eight highly correlated NMF metaprograms. NMF programs that were not correlated, or correlated groups mostly originating from a single patient, were excluded (Figure 20C). I performed GO enrichment analysis on both NMF metaprograms and clusters, assigning annotations to those showing clear enrichment, while labelling the remaining cells as "Unannotated" (Figure 20F). NMF-based cell populations were named based on the order of retrieval of NMF metaprograms, while subsequent clusters were named based on the predominant ATRT subtype.

A ground-state rhabdoid tumor cell population at the basis of the developmental hierarchies of each ATRT subtype

Next, I investigated whether the IPC-like population, given its presence across ATRT subtypes and patients (Figure 21A), truly exhibited a cycling nature. To do this, I conducted differential expression analysis comparing the IPC-like population to the other cell populations. I focused on the top 100 differentially expressed genes and subjected them to GO enrichment analysis.

Remarkably, the top 10 enriched GO terms were clearly associated with a cell cycle phenotype (Figure 21B). This observation was consistent with the inferred cell cycle phase for each tumor population: IPC-like cells exhibited a phase assignment predominantly in the S and G2M phases, indicating increased cell proliferation compared to most other tumor cell populations, which were majorly on the G1 phase[199] (Figure 21C). The only other tumor cell population showing a high fraction of cells in the G2/M was the mesenchymal-like, although also exhibiting a small fraction of cells in the G1 phase, unlike the IPC-like cells.

Given that the IPC-like population resembles neuroblasts, I assessed its similarity to different stem cell types using literature marker gene sets from PanglaoDB Augmented 2021[191,195]. These gene sets served as a prior knowledge network, and activity scores were computed across tumor cell populations[85].

The results showed that activity scores for pluripotent stem cell marker genes were highest in IPC-like population, suggesting that this tumor cell population also harbors a stem-like phenotype (Figure 21D-E).

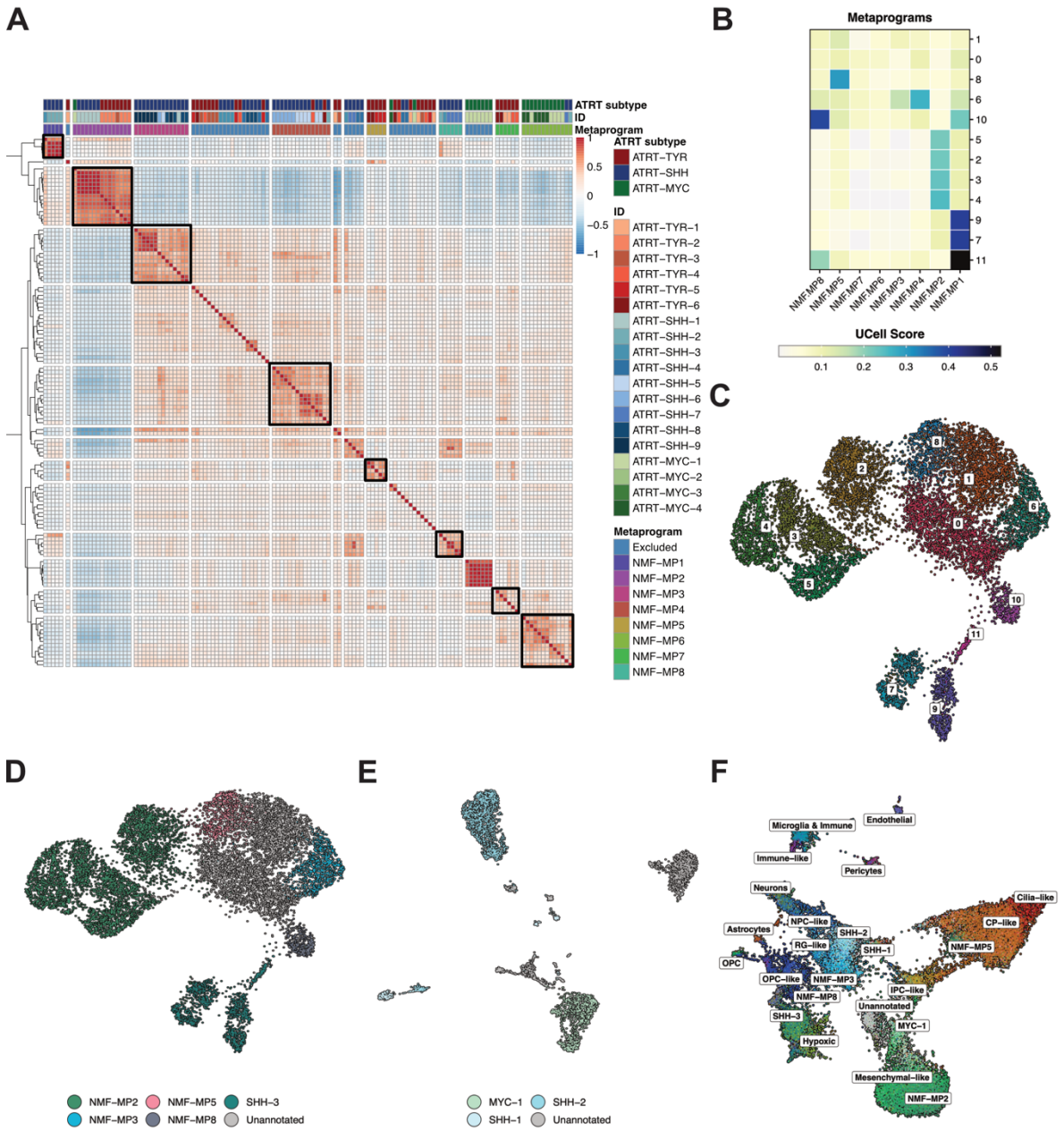


Figure 20: Unsupervised annotation of tumor cells. (A) Heatmap depicting Pearson's correlation score between each pair of NMF programs. Groups of highly correlated programs shared across patients become NMF metaprograms. Non-correlated programs or highly correlated programs unique of single patients are excluded. **(B)** Heatmap depicting enrichment scores for each metaprogram in a re-clustering using only the unannotated cells from the supervised annotation step. **(C, D)** UMAP representation based on enrichment scores in the NMF metaprograms, colored by cluster **(C)** and derived annotation from the unsupervised annotation step based on NMF **(D)**. UMAP1, x-axis; UMAP2, y-axis. **(E)** UMAP representation of the remaining, unannotated, cells after the unsupervised annotation step based on NMF, colored by inferred clusters with slight GO term enrichment, that predominantly associated with a single ATRT subtype. Unannotated cells are a mixture of cells that were not labelled in any of the previous steps. UMAP1, x-axis; UMAP2, y-axis. **(F)** Integrated UMAP representation colored by the inferred cell populations (TME + tumor cells). UMAP1, x-axis; UMAP2, y-axis.

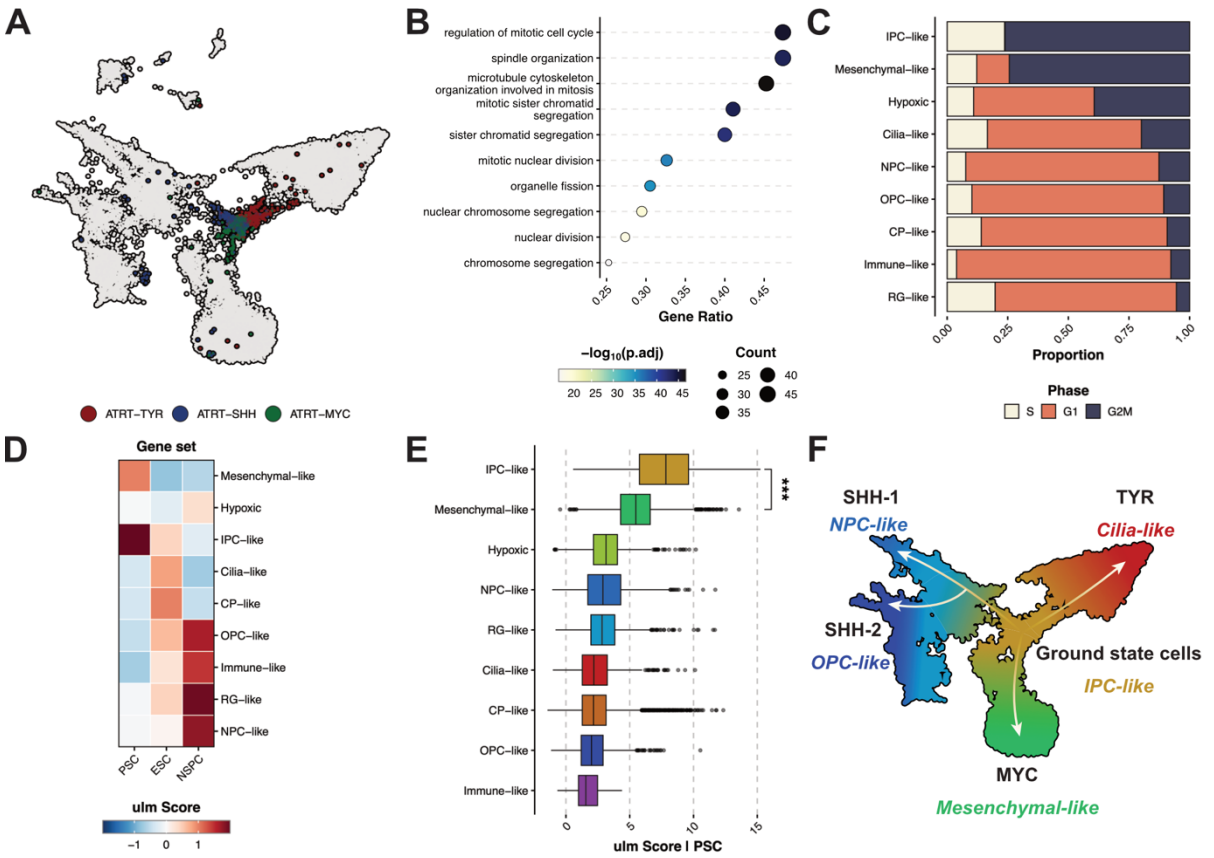


Figure 21: IPC-like population is highly proliferative and displays features of pluripotent stem cells. (A) Integrated UMAP representation highlighting IPC-like cells colored by ATRT subtype. UMAP1, x-axis; UMAP2, y-axis. **(B)** Dot plot depicting the enriched GO terms based on top 100 marker genes for the IPC-like population against the rest of cell identities in the dataset. Top ten GO terms are shown. Color encodes for adjusted p-value and size indicates the number of marker genes supporting each term. **(C)** Cell cycle phase proportion across tumor cell populations. **(D)** Heatmap of activity scores for different literature gene sets associated with stemness (PSC = pluripotent-stem cells, ESC = embryonic stem-cells, NSPC = neuronal stem/precursor cells)[191]. Scores are scaled and centered, thus comparison across columns should be avoided. **(E)** Box plot depicting non-scaled and not centered PSC activity scores across tumor cell populations. Activity scores for IPC-like population are significantly higher than those of the second scoring population, Mesenchymal-like (Wilcoxon test: *** < 0.001). **(F)** Schematic illustration based on the integrated UMAP representation where different ATRT subtype-specific trajectories have been showcased. Arrows indicate the possible direction of differentiation within the tumor cell population. Illustration made by Irene Paassen.

While previous studies support the notion of a differential cell of origin for each ATRT subtype[139], I hypothesize that the loss of SMARCB1 may alter the transcriptome in these distinct cells of origin, causing them to resemble neuronal IPC-like progenitor cells.

These cells may serve as the basis for the various differentiation lineages emerging in each ATRT subtype (Figure 21F). Therefore, these cells are clustered together in the UMAP due to their shared cycling and stem-like properties.

Consequently, due to the overarching cycling nature of the IPC-like tumor cell population, I aimed to ascertain whether their presence was not due to biological batch effect in my analysis. For this, I conducted an experiment to determine whether regressing out the effect of cell cycle genes prior to dimensional reduction and supervised annotation would affect the identification of the IPC-like tumor cell population.

After performing this process, I identified a new IPC-like cluster, the cells of which exhibited a high degree of overlap with those previously annotated as IPC-like (Figure 22). This finding suggested that the cell cycle was not a confounding artifact in my analyses.

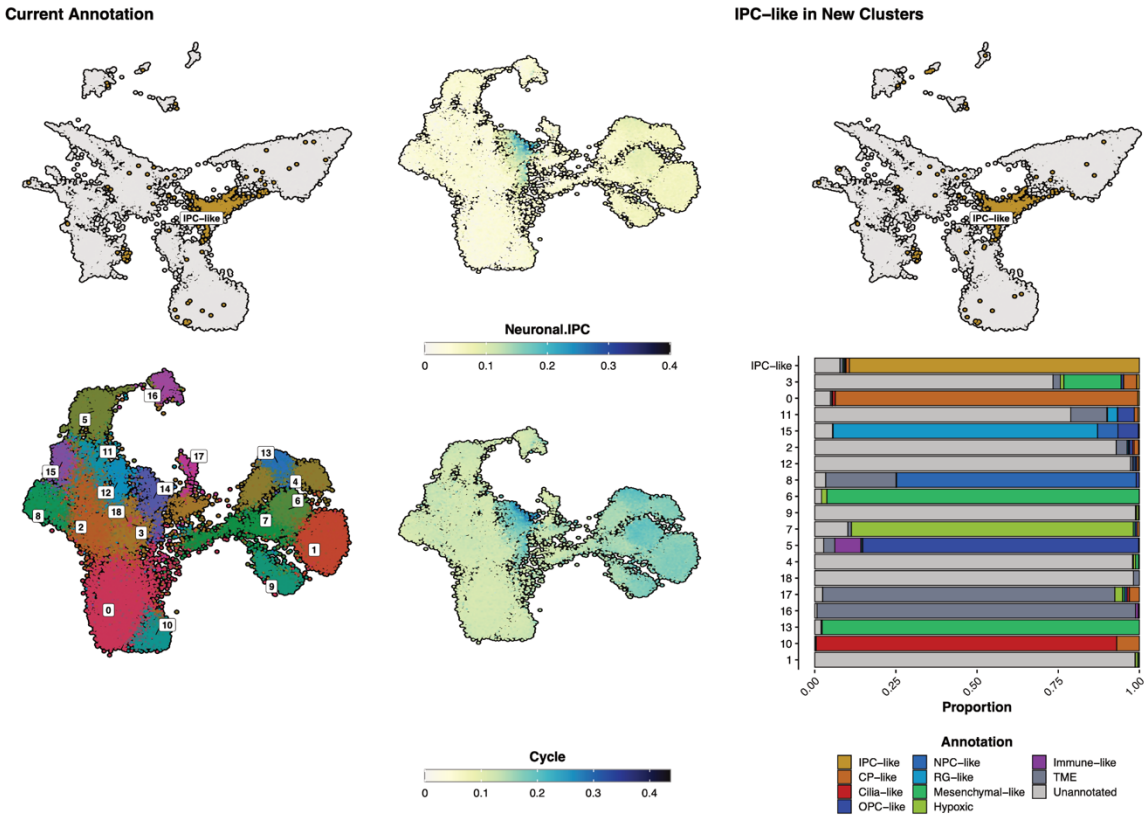


Figure 22: IPC-like bias. Integrated UMAP representation highlighting the IPC-like population (top-left) prior to regressing out the effect of cell cycle. Cell cycle-regressed cells underwent supervised annotation step. UMAP representation based on enrichment scores (bottom-left) and colored by the enrichment is Neuronal IPC and Cycle marker sets are displayed. New IPC-like population is defined and projected onto the original integrated UMAP representation (top-right). Cell type proportion across the new clusters and IPC-like population is displayed as a bar plot (bottom-right), where most of the new IPC-like population is formed by the old one, thus reassuring that I did not find this population as a bias from not regressing out cell cycle effect. UMAP1, x-axis; UMAP2, y-axis.

Inferring the presence of IPC-like cells in additional datasets

Next, I aimed to determine whether the identification of an IPC-like tumor cell population was unique to my datasets or if this population could also be detected in additional datasets. To explore this, I utilized various validation datasets, including single-nuclei (n = 5) and single-cell (n = 3) SMARTseq2 datasets, single-nuclei ATRT cell line 10X v3 3'datasets (n = 5), and fresh-tissue 10x v3 3' single-cell (n = 1) datasets. After conducting quality control, normalization, clustering and dimensional reduction, I proceeded with supervised annotation using the same gene sets as previously described. The findings indicated that, across the various datasets examined, there were cell clusters showing enrichment for both neuronal IPC and cycling gene sets (Figure 23A-D), thus confirming the presence of IPC-like cells as a distinct tumor cell population within ATRT tumors.

Designing a gene panel for Xenium spatial transcriptomics

To examine the spatial arrangement of the previously characterized tumor and tumor microenvironment cell populations, I devised a gene panel for Xenium spatial transcriptomics. To assess the specificity of the selected genes for each cell population, I evaluated their enrichment and expression across all cell populations. The findings revealed that each cell population could be distinguished by, at most, a combination of two gene sets (Figure 24A). Moreover, the expression of individual genes was predominantly restricted to individual cell populations, except for closely related cell types within specific ATRT subtype-specific tumor lineages, such as CP-like and cilia-like tumor cell populations (Figure 24B).

With a tailored gene panel crafted to encompass the heterogeneity within my datasets, the experiment design for the Xenium datasets was then established. This involved selecting representative sample pairs across ATRT subtypes showcasing contrasting levels of immune infiltration, for which spatial transcriptomics data is currently being generated. While my work related to this PhD thesis concluded at this stage, the results stemming from this analysis could yield crucial insights into the interplay between ATRT tumor cells and their microenvironment, offering potential avenues for identifying therapeutical targets.

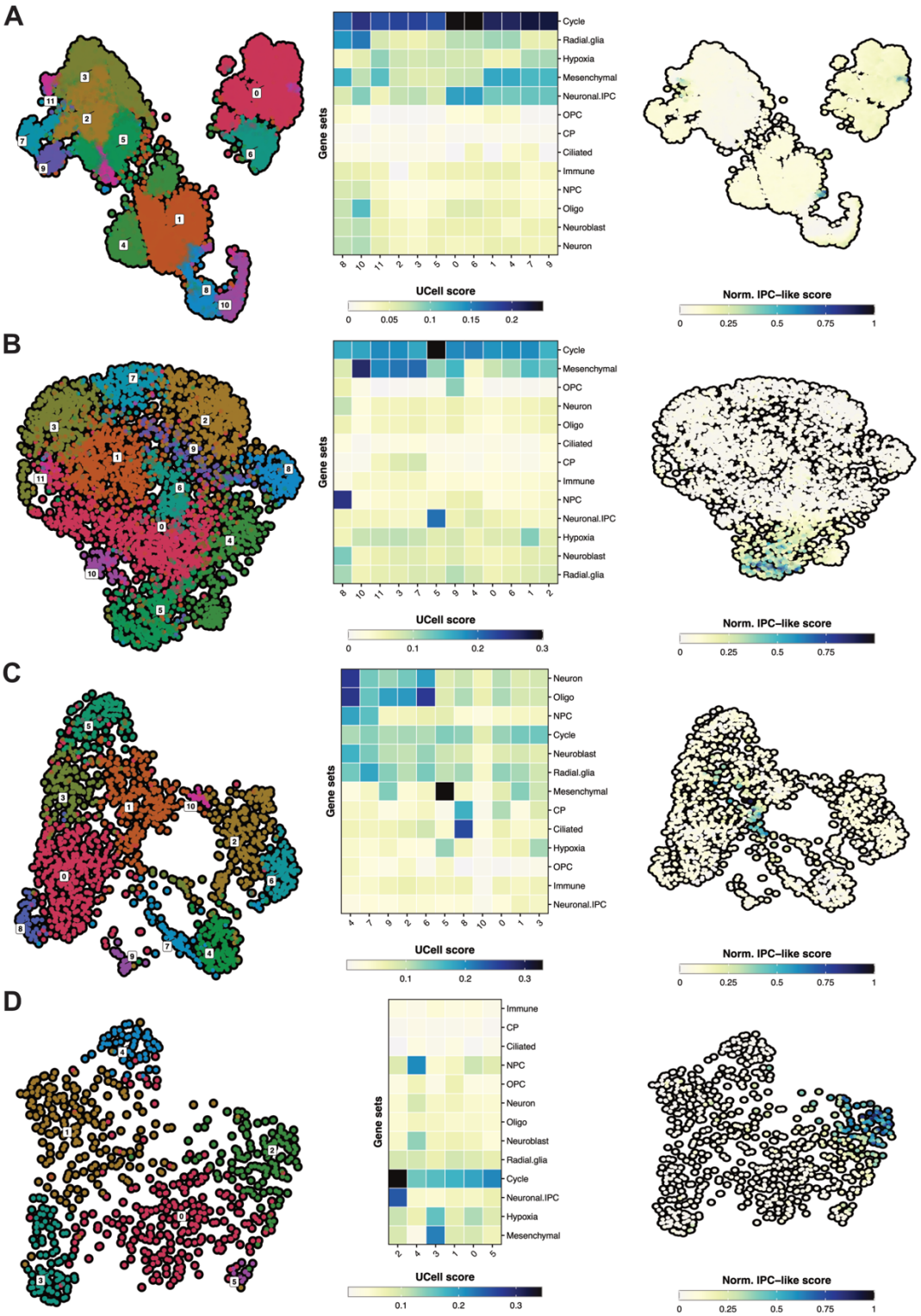


Figure 23: Inferring IPC-like tumor cell population in validation datasets. Four validation datasets are used in order to retrieve, among others, the IPC-like population: 5x 10x v3 5' snRNAseq cell line data **(A)**, 1x 10x v3 5' scRNAseq ATRT-TYR **(B)**, 5x SMARTseq2 snRNAseq **(C)** and 3x SMARTseq2 scRNAseq **(D)**. For each dataset, enrichment scores for the supervised annotation set are computed. UMAP representation and clustering based on enrichment scores are computed (left) and scores are visualized as a heatmap (middle). Scores for Neuronal IPC and Cycle gene sets are normalized to range from 0 to 1 and a linear combination of both is calculated to form a combined score, projected onto the UMAP visualization (right). Cells with highest scoring for both gene sets will have a high combined score, and therefore will be highlighted in the UMAP. Across datasets, a specific region of the UMAP is highlighted, suggesting the presence of IPC-like cells. UMAP1, x-axis; UMAP2, y-axis.

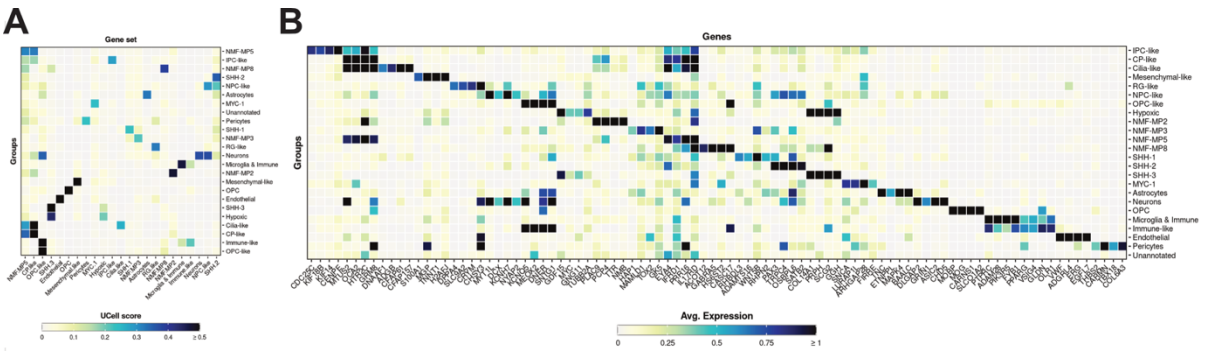


Figure 24: Panel generation for Xenium. Enrichment scores **(A)** and expression levels **(B)** for the different genes and associated gene sets per each cell population. Gene set or individual genes, x-axis; cell clusters, y-axis. Unannotated cells belong to the subset of tumor cells that could not be annotated by neither supervised nor unsupervised annotation methods. Color scale for enrichment scores restricted to a maximum of 0.5 and to a maximum of 1 for average expression level.

Discussion

ATRTs are notorious for their poor prognosis within pediatric brain tumors. Over the past decade, numerous studies have explored the heterogeneity across ATRT subtypes[139,163,200,201]. However, no comparative analysis has been conducted at the single-cell level, nor has the cell of origin of each ATRT subtype been fully characterized, leaving only hypothesis available. In this project, I aimed to systematically compare the inherent heterogeneity present at transcriptional and chromatin accessibility levels within and between SMARCB1-deficient ATRT subtypes at a single-cell level. I devised a comprehensive tumor cell type annotation strategy and characterized the different tumor lineages that ATRT tumor cells commit to.

The results of the tumor cell type annotation revealed ATRT subtype-dependent overarching tumor lineages, consistent with previous studies[139,144]. Specifically, ATRT-TYR cells exhibited a transition from choroid plexus-like cells towards cilia-like cells, while ATRT-SHH cells showed a transition from radial glia-like cells towards either an oligodendrocytic or neuronal/astrocytic lineage. Additionally, ATRT-MYC cells harbored a shift towards a mesenchymal-like lineage. Since ATRT-SHH can be further classified into three distinct subtypes[166], the distinct lineages observed in the data may be linked to specific ATRT-SHH subtypes. Tumor cell type composition analysis suggested a polarization of ATRT-SHH-1A subtype towards the oligodendrocytic lineage and ATRT-SHH-2 towards the neuronal/astrocytic lineage. However, further research on a larger ATRT-SHH cohort is needed to validate these observations.

Despite previous studies suggesting different cells of origin for each ATRT subtype[139,198,202], tumor cells within my datasets converged in an integrated embedding into a tumor cell population shared across patients and ATRT subtypes. These cells exhibited expression signatures resembling those of neuroblasts with cycling activity, which I termed IPC-like cells. Even after accounting for cell cycle genes as a potential confounding factor, the same population persisted, indicating their integration into a single cell cluster based on additional phenotypes beyond cycling nature. Gene set enrichment analysis focusing on stemness-associated gene sets revealed that IPC-like highly scored for pluripotent stem cell gene markers. These findings position IPC-like cells at the root of the differentiation lineages of each ATRT subtype, being integrated together based on their dominant overarching cycling and stem phenotype while still retaining expression signatures characteristic of their subtype of

origin. Consequently, IPC-like cells may serve as a “rhabdoid ground-state” cell population within each patient, potentially crucial for promoting tumorigenesis. Furthermore, IPC-like cells appear not to be a unique finding from my datasets, with their presence being confirmed in the various validation datasets. Therefore, targeting this tumor population to induce differentiation towards the end of the aforementioned tumor lineages could be pivotal in developing ATRT subtype-targeted therapies.

Having identified the differential tumor lineages across ATRT subtypes at a single-cell level and characterized the presence of a ground-state tumor cell population with therapeutic potential, the next crucial step is validation. Functional validation through gene set enrichment analysis at the pathway and transcription factor levels will provide insights into the roles of each tumor cell population in ATRT biology. Additionally, ligand-receptor analysis can shed light on the crosstalk between tumor populations and the microenvironment. Furthermore, contrasting these results at the chromatin accessibility level may be essential to determine whether other epigenetic factors contribute to the observed heterogeneity in ATRTs.

Assessing the role of the IPC-like population as a “rhabdoid ground-state” tumor cell population may involve performing drug testing analysis on organoid models, where a combination of drugs selected to promote differentiation can be tested by comparing viability readouts against a control setting. Furthermore, exploring the spatial disposition of the different tumor cell populations can provide key insights on the interplay between tumor cell populations and tumor microenvironment. By designing experiments where sample pairs per ATRT subtype are selected with varying degrees of immune infiltration, a comparative analysis of tumor and tumor microenvironment composition in relation to immune infiltration can be conducted. Altogether, these results and follow-up experiments hold promise for further research and serve as a foundation for the development of potential ATRT subtype-specific therapies.

Tumor heterogeneity and tumor-microglia interactions in primary and recurrent IDH1-mutant gliomas

Introduction

Diffuse gliomas are prevalent malignant brain tumors in adults[203]. The latest WHO classification of CNS tumors, updated in 2021, categorizes adult-type diffuse gliomas into three main supergroups: astrocytoma, isocitrate dehydrogenase (IDH)-mutant, oligodendroglioma, IDH-mutant and 1p/19q-codeleted and glioblastoma, IDH-wildtype[7]. With the exception of glioblastoma, IDH wildtype, these tumors are collectively referred as IDH-mutant gliomas.

IDH-mutant gliomas are diffusely infiltrating tumors characterized by a hallmark mutation in either the *IDH1* or *IDH2* gene, encoding homologous proteins. The most common *IDH1* mutations, occurring on amino acid 132, includes substitutions such as arginine-to-histidine (*IDH1* R132H, 83-91% of cases)[204]; arginine-to-cysteine (*IDH1* R132C, 3.6-4.6% of cases)[204]; arginine-to-glycine (*IDH1* R132G, 0.6-3.8% of cases)[204]; arginine-to-serine (*IDH1* R132S, 0.8-2.5% of cases)[204]; or arginine-to-leucine (*IDH1*, R132L, 0.5-4.4% of cases)[4,204,205]. In the case of *IDH2* mutations, which occur on amino acid 172, the arginine-to-lysine substitution (*IDH2*, R172K) is the most frequent[206].

These mutations lead to a gain of function that disrupts the conversion of isocitrate to α-ketoglutarate (α-KG), shifting it to a production of D-2-hydroxyglutarate (D-2HG) (Figure 25A), an oncometabolite[207] that decreases the activity of hypoxia-inducible factor 1 α (HIF1α) protein[208], known to be a suppressor of gliomagenesis[209] and also antagonizes via competitive inhibition with other tumor-suppressors part of the α-KG-dependent dioxygenase family such as ten-eleven translocation (TET) DNA modifying enzymes and jumonji-C domain-containing (JmjC) histone demethylases[210,211] (Figure 25B-C). This interference results in DNA hypermethylation, known as the glioma-associated CpG island methylator phenotype (G-CIMP)[212], leading to a differentiation block[213] and the exhibition of tumor cell populations close to that of stem cells in both oligodendrogliomas an astrocytomas[135,214]. G-CIMP can be characterized as “G-CIMP-high” and “G-CIMP-low” (state seems to alter alongside tumor progression)[215], with the latter associated with worse prognosis[216].

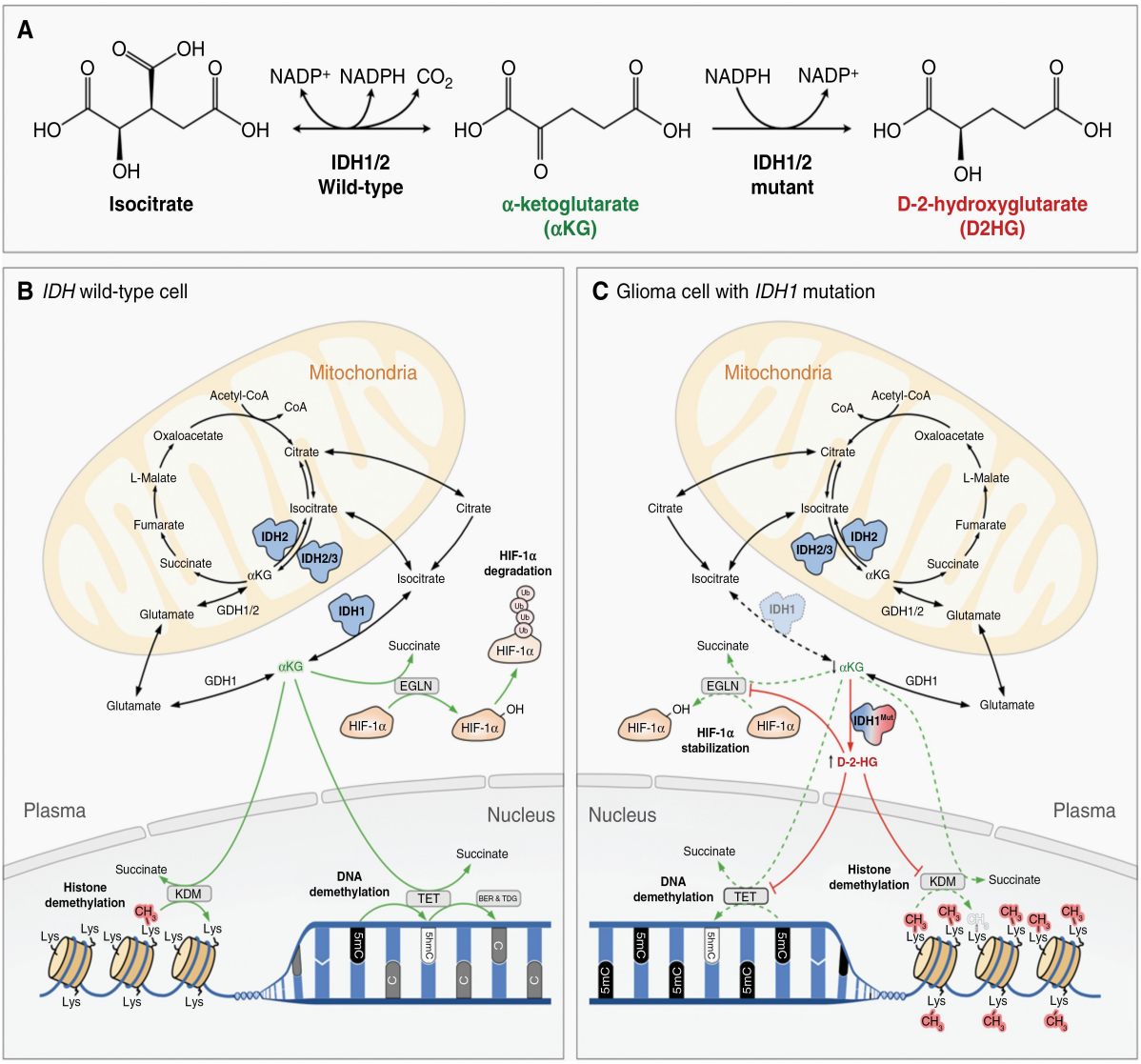


Figure 25: "Functions of normal and mutated IDH enzymes. (A) Normal IDH1 and IDH2 proteins use NADP+ as an electron acceptor to catalyze the oxidative decarboxylation of isocitrate, producing α-ketoglutarate (αKG) and CO2. However, mutant IDH1 and IDH2 produce D-2-hydroxyglutarate (D-2-HG) from α-KG using NADPH as an electron donor. **(B)** The normal isoforms IDH1 and IDH2 catalyze αKG production in the cytoplasm and mitochondria, respectively. As a critical intermediate in the Krebs cycle, αKG is involved in many biological metabolic processes. A superfamily of enzymes called αKG-dependent dioxygenases (αKGDs), including TET, KDM, and EglN, can decarboxylate αKG to succinate while hydroxylating different substrates for various further changes, such as DNA demethylation, histone demethylation, and ubiquitination of transcription factor HIF-1α. **(C)** Mutant IDH1 produces high level D-2-HG. As a structural analog of αKG, excessive D-2-HG competitively inhibits the catalytic efficiency of the TETs and KDMs while paradoxically stimulating EglN activity. Decreased TET and KDM activity causes DNA and histone hypermethylation, respectively, while increased EglN activity lowers HIF. Collectively, these changes affect gene expression, cell division and differentiation." Reprinted with permission from Miller, *et al*.[217]

Studies also show that DNA hypermethylation at cohesion and CCCTC-binding factor (CTCF) binding sites affects chromosomal topology, promoting new chromosomal interactions that induces expression of glioma oncogenes such as platelet-derived growth factor receptor alpha (*PDGFRA*)[218].

Additionally, MGMT promoter methylation is a recurrent event in IDH-mutant gliomas[212,219]. MGMT promoter encodes for a DNA repair protein responsible for removing alkyl groups from the O6 position of guanine in DNA, affecting the efficiency of therapy treatments with alkylating agents such as temozolomide (TMZ)[220]. Hypermethylation at MGMT promoter silences the gene, therefore depleting its protein levels and thus allowing for TMZ to have an increased effect.

Based on histologic and molecular and methylation data, IDH-mutant gliomas can thus be classified into two major subtypes: astrocytomas and oligodendrogliomas[217]. Astrocytomas can range from WHO grade 2-4 and is typically presented at a median age of 38 years for WHO grade 2-3 cases[221], which is slightly higher than for WHO grade 4 cases[222]. Molecularly, astrocytomas often harbor loss-of-function mutations in *TP53* and alpha thalassemia/mental retardation (*ATRX*, 70% cases)[221] genes. Mutations in the *ATRX* gene lead to an abnormal telomere maintenance mechanism commonly referred to as alternative lengthening of telomeres (ALT)[223,224], contributing to genomic instability[225]. This instability can manifest as copy-number abnormalities, including amplifications of oncogenes such as *MYC* or *CCND2*[221]. These molecular features contribute to the pathogenesis and progression of astrocytomas.

Oligodendrogliomas are characterized by a specific genetic alteration being the codeletion of chromosome arms 1p and 19q, resulting from an unbalanced translocation between these chromosome arms[226]. This event is a defining feature for the classification and diagnosis of oligodendrogliomas. Oligodendrogliomas can range from WHO grade 2-3 and are typically diagnosed at a median age of 43 years for CNS grade 2 cases and of 50 years for CNS grade 3 cases[227]. Oligodendrogliomas very frequently exhibit *IDH1* R132H substitution, which is present in approximately 90% of cases. However, some cases may have non-canonical substitutions in either *IDH1* or *IDH2* gene.[205]

Additionally, mutations in telomerase reverse transcriptase (*TERT*) promoter (most of the cases)[228,229], *CIC* (70% of cases)[221,230] or *FUBP1* (20-30% of cases)[230] are commonly observed in oligodendrogliomas. *TERT* promoter mutations, such as C228T or C250T substitutions, lead to transcriptionally upregulation of *TERT*, promoting cellular immortalization and proliferation and telomere stabilization[231], and are considered to be a clonal event in oligodendroglioma formation[232,233]. *CIC*, a repressor of the MAPK pathway[234], is involved in controlling cellular

growth, development and metabolism[235]. Lower levels of *CIC* resulting from mutations are linked to gliomagenesis by promoting the proliferation of neural stem cells[236] and upregulating D-2HG[237]. *FUBP1* is essential for the maintenance and self-renewal of neural stem cells[238] and promotes alternative splicing of oncogenes and tumor suppressor genes[239]. The combined loss of *CIC* and *FUBP1* are linked to shorter tumor recurrence time[240].

Both astrocytomas and oligodendrogliomas can originate in any part of the CNS, with a preference for the supratentorial region[241,242]. They share a common origin from glial progenitor cells such as neural precursor cells, oligodendrocyte precursor cell or astrocytes[243]. These progenitor cells play a crucial role in promoting developmental hierarchies, consisting of astrocyte-like and oligodendrocyte-like lineages[135,214].

In the light of the current knowledge, several questions remain unanswered regarding the transcriptional and epigenetic differences between astrocytomas and oligodendrogliomas at a single-cell resolution. Additionally, it is unclear whether specific subpopulations of tumor-associated microglia/macrophages (TAMs) are associated with particular glioma subtypes, despite evidence supporting distinct TAM activation states in IDH-mutant gliomas[244,245]. The interplay between tumor cell populations and TAMs, potential variations in both tumor and TAM populations between oligodendrogliomas and astrocytomas, and whether tumor and TAM populations vary as a factor of tumor grade and tumor recurrence remain unexplored areas of research.

To address these questions, single-nucleus RNA sequencing data (n = 14, six astrocytomas, IDH-mutant and eight oligodendrogliomas, IDH-mutant. From here onwards, IDH-mutant gliomas will be referred as astrocytomas and oligodendrogliomas) along with matching single-nucleus ATAC sequencing data (n = 11, five astrocytomas and eight oligodendrogliomas) were generated in the laboratory of Dr. Şevin Turcan and made accessible to me for bioinformatics analyses. Additionally, single-nucleus RNA sequencing of another cohort of matching primary-recurrent astrocytoma tumor pairs (n = 12, six tumor pairs) was generated to broaden and validate the findings. This data was analyzed in collaboration with the laboratory of Dr. Holger Heyn, in the Centre for Genomic Regulation/Centro Nacional de Análisis Genómico, Barcelona.

Materials and Methods

This section provides an overview of the dataset generation and methods used in the project. Unless otherwise noted at the end of each subsection, I conducted all bioinformatic analyses. I also generated all the data visualizations for the published article[189]. Throughout the analyses, I received supervision from Prof. Dr. Matthias Schlesner, Dr. Ashwin Narayanan and Dr. Şevin Turcan. While I wrote the original text of this chapter, I used ChatGPT to enhance its readability.

Patient samples

The following paragraph is extracted from Blanco-Carmona, *et al*.[189] and adapted to match the figure references to those in this thesis:

"A total of 14 fresh frozen archival IDH-mutant glioma samples (8 oligodendroglioma and 6 astrocytoma) were used for snRNA-seq and snATAC-seq. Additionally, 6 fresh frozen paired primary and recurrent samples (IDH-mutant astrocytoma) were used for snRNAseq. Patient information and tumor characteristics are provided in Figure 26 *and* Figure 48A*. Fresh frozen samples were collected by the Division of Experimental Neurosurgery, Department of Neurosurgery, University Hospital Heidelberg, and the Department of Neurosurgery, Acıbadem Mehmet Ali Aydınlar University, School of Medicine, Istanbul. Samples used as validation cohort for immunohistochemistry were retrieved from the institutional databases of the Department of Pathology, Spedali Civili of Brescia, and the Department of Neuropathology, Heidelberg University Hospital. All patients provided written informed consent, in accordance with the Declaration of Helsinki. All samples were reviewed and received approval from the respective Institutional Review Boards and local authorities at the institutions where samples were originally collected. Specifically, the samples were approved by the Ethics Committee of Spedali Civili of Brescia, Institutional Review Board at the Medical Faculty of Acıbadem Mehmet Ali Aydınlar University, and the Ethics Committee of Heidelberg University*."

Sample preparation and library construction for sequencing

The following paragraph is extracted from Blanco-Carmona, *et al*.[189]:

"Isolation of nuclei for snRNA-seq and snATAC-seq was performed as previously described. Briefly, fresh frozen tissue samples were cut into small pieces and homogenized using a Dounce

homogenizer in EZ lysis buffer (Sigma Aldrich). Tissue was homogenized 20 times with pestle A and 20 times with pestle B. This was followed by centrifugation, filtration, and buffer-mediated gradient centrifugation to obtain pure single nuclei, which were then used for snRNA-seq and snATAC-seq. Nuclei were counted using a hemocytometer, and their concentration adjusted as needed to meet the optimal range for loading on the 10x Chromium chip. The nuclei were then loaded into the 10x Chromium system using the Single Cell 30 Reagent Kit v3 or v3.1 (for snRNA-seq) and Chromium Next GEM Single Cell ATAC Library & Gel Bead Kit v1.1 (for snATAC-seq) according to the manufacturer's protocol. We aimed to load ~20,000 nuclei for each snRNA-seq run and ~10,000 nuclei for each snATAC-seq run. Following library construction, libraries were sequenced on the Illumina NovaSeq 6000 system."

Single-nucleus RNA-seq analysis

Read alignment and count matrix generation

Raw reads from the sequencer were used to generate a count matrix using the `count` module of `cellranger` (version 3.1.0)[186]. Since reads were from frozen tissue, a modified version of the GRCh38 transcriptome accounting for introns was used[246]. This transcriptome, commonly referred as "pre-mRNA" reference, was generated using the `mkfastq` module of `cellranger`. A custom GTF annotation, where all introns are accounted for as exons, was provided and incorporated during the count matrix generation step. Subsequent analyses were carried out using `Seurat`[118,247] (version 3.1.0 – 4.3.0).

Quality control of snRNAseq data

To remove low-quality cells, including droplets, the following metrics were considered: the total number of unique molecular identifiers (UMI) per cell (`nCount_RNA`), the total number of genes per cell (`nFeature_RNA`) and the percentage of mitochondrial RNA per cell (`percent.mt`). Cells with `nCount_RNA < 1000`, `nFeature_RNA < 500` and `percent.mt > 5%` were filtered out, according to best practices[25]. Additionally, to remove extreme outliers, cells falling outside the mean plus three standard deviations of the distribution of UMI and genes were also filtered out, leaving only high-quality cells.

Doublet removal

Despite cells being of high quality, cells can still exhibit a doublet nature, which is a side-effect of the sequencing technology and needs to be removed. To detect and remove doublets, `scrublet`[188] (version 0.2.1) was used. Following standard guidelines[248], the package was imported into R using `reticulate` with `reticulate::import()` and doublet scores were computed using `scrublet$Scrublet()`. This process returned a list of doublet scores that displayed a bimodal distribution. Cells with doublet scores falling on the second peak (high doublet score) were removed. The threshold to determine whether a cell was a doublet or not was sample-dependent and was set manually upon individual inspection.

Normalization

Cell expression data was normalized using `Seurat::SCTransform()`[48] with default parameters, which performs regularized negative binomial regression on the count data. Following the developer's guidelines[249], individual datasets from each patient were pooled together into a merged dataset for normalization.

Dimensional reduction

To perform dimensionality reduction using principal component analysis (PCA) a total of 3.000 highly variable genes were identified and used as basis for the method. PCA was computed using `Seurat::RunPCA()`, retrieving a total of 50 principal components (PC). Upon manual inspection of the variability captured by each PC using `Seurat::ElbowPlot()`, the top 15 PCs were selected. Subsequently, further dimensional reduction was applied using the selected PCs with `Seurat::RunUMAP()`, which computes uniform manifold approximation and projection (UMAP)[50], resulting in the dimensional reduction embedding that serves as the basis for the analysis.

Data integration

To remove biases arising from each individual patient, integration was applied to the data using `Harmony`[56] with default parameters. Based on the PCA dimensional reduction, a new dimensional reduction was generated where the effect of cells originating from different patients was corrected. This corrected dimensional reduction was then used to generate a new UMAP embedding, commonly referred as the "integrated UMAP".

Clustering and cell type annotation of the Tumor Microenvironment

To cluster the cells based on their transcriptional similarity, Louvain algorithm[127] was applied either on the PCA (merged data) or Harmony embedding (integrated data). This was achieved by using `Seurat::FindNeighbors()` followed by `Seurat::FindClusters()`. The resulting clusters served as the foundation for the subsequent cell type annotation step. Initially, gene sets representing various cell types were retrieved from `panglaoDB`[191], including microglia, oligodendrocytes, neurons, astrocytes, endothelial cells, pericytes and T cells. Enrichment scores were then calculated for each gene signature and cell using `Seurat::AddModuleScore()`. Subsequently, to assess whether a particular cluster correlated with any gene signature, the enrichment scores were visualized as feature plot with `SCpubr::do_FeaturePlot()` and as heatmaps with `SCpubr::do_EnrichmentHeatmap()`.

Copy Number Variant analysis

To identify malignant cells, two distinct methods can be employed. Firstly, in the context of a merged UMAP, cells from the TME typically form clusters regardless of the patient-specific biases. Therefore, malignancy can be inferred by identifying clusters where the majority or entirety of the cells originate from a single patient. In the case of oligodendrogliomas, an additional confirmation step involves detecting the codeletion of chromosome arms 1p and 19q. Therefore, copy number variant (CNV) events were identified from expression data using `inferCNV`[95], using microglia and oligodendrocytes cells as reference. The analysis was carried out with the following parameters: `cutoff = 0.1`, `min_cells_per_gene = 3`, `HMM = TRUE`, `HMM_type = "i6"`, `window_length = 201`. To further enhance the sensitivity of the analysis, raw counts of up to five cells from the same cell cluster were aggregated into "metacells". These metacells were then used in the copy number variant analysis, effectively enhancing the signal while minimizing the inherent noise arising from using 10X datasets when utilizing `inferCNV`.

Non-Negative Matrix Factorization

To reveal the predominant cell states within tumor cells across both oligodendrogliomas and astrocytomas, non-negative matrix factorization (NMF) was used, following established methodology[103]. This involved, for each patient independently, filtering out cells from the TME followed by scaling and centering the normalized count matrix. NMF was conducted using the

`NMF` R package[250] via `NMF::nmf()`function. Key parameters included: `rank = X`, being `X` a number varying from two to ten, `seed = 777` and `method = "snmf/r"`. The rank parameter determined the number of independent NMF programs that are retrieved from each patient, from which mitochondrial genes were excluded, and the top 30 scoring genes were selected for further analysis.

To assess the enrichment level of each cell for a given NMF program, a scoring system developed by Dr. Volker Hovestadt was implemented. For every gene within the NMF program, a control set comprising the top 100 genes with the most similar expression profiles was defined. Expression similarity was determined by subtracting the expression value of the selected gene from the average expression value across all cells. The resulting values were sorted in ascending order, and the top 100 genes (excluding the initially selected gene) constituted the control set. Subsequently, the expression difference between the selected gene and the control set was computed for each cell. This process was repeated for every gene in the NMF program. The output of this scoring method yielded a matrix where each column represented a cell and each row represented a gene in the NMF program. The values of the matrix were then averaged across genes, resulting in a single score for each cell pertaining to the given NMF program. This score effectively quantifies the extent to which a particular cell is enriched in the NMF program.

To identify similarities among the scoring of different NMF programs across cells, a correlation matrix using Pearson's correlation coefficient was computed. Highly correlated NMF programs formed NMF metaprograms. This process was iterated for each value of the NMF rank, selecting the value that produced the most distinct and highest number of NMF metaprograms. To extract the genes driving each NMF metaprogram, I applied a modified version of the previously described scoring method. However, in this iteration, the matrix containing scores was averaged across all cells. This resulted in a single scoring value for each gene, illustrating its relevance across all cells and thus within the metaprogram. The top 30 scoring genes were then chosen as the drivers of each NMF metaprogram.

While the NMF method and initial scoring method were provided by Dr. Volker Hovestadt, the adaptation of the method to identify the driver genes of NMF metaprograms was the outcome of collaborative work between Dr. Christina Blume and me.

Permutation-testing as a method to assign metaprograms to cells

After obtaining the top 30 genes for each NMF metaprogram, I developed a method to statistically identify cells enriched for each NMF metaprogram. This approach involved conducting permutation testing to compare an empirical distribution against a null distribution. The empirical distribution comprised enrichment scores for all cells associated with a given NMF metaprogram. In contrast, the null distribution was generated by permuting the expression values across all cells for each gene belonging to the NMF metaprogram being analyzed. This permutation ensured a complete disruption of any enrichment patterns stemming from the specific combination of genes. Enrichment scores were computed using `Seurat::AddModuleScore()` function with default parameters. Subsequently, new enrichment scores were calculated for all cells based on the permuted expression matrix, thereby representing the null distribution. This permutation process was repeated until a total of one million permuted values were generated.

The determination of whether a particular cell exhibited statistically significant enrichment for a given NMF metaprogram involved computing p-values based on the fraction of cells in the null distribution where the enrichment score surpassed that of the queried cell. To prevent infinite values, a value of plus one (`+1`) was added to both components of the fraction. The number of permuted values generated determined the lowest achievable p-value by the method. Therefore, with a total of one million permuted values, the method allowed for a minimum p-value of up to `1e-6`. This aspect was particularly crucial as p-values required correction for multiple testing.

The correction of p-value was necessary to address two main factors: multiple testing and multiple comparisons. Firstly, multiple testing occurred when the same cells were utilized to shuffle the expression scores. Secondly, multiple comparisons arose when several NMF metaprograms were queried and compared with each other for the same cell. To correct for these issues, the Benjamini-Hochberg method for p-value adjustment was employed[251] using the `stats::p.adjust()` function with parameters `method = "BH"`. Cells were deemed significant in a given comparison if the adjusted p-value was below 0.05 (false discovery rate (FDR) = 5%). When multiple comparisons were conducted, the FDR threshold was further

adjusted by dividing it by the total number of comparisons. This adjustment helped control against the inflation of the alpha error, commonly known as type I error.

This method was developed thanks to the collaborative effort involving Prof. Dr. Matthias Schlesner, Dr. Martin Sill and myself. While I undertook the implementation and coding to test significance of enrichment scores, the process was overseen by Prof. Dr. Matthias Schlesner, and Dr. Martin Sill provided guidance on the template code and ensured the statistical rigor of the method.

Pathway and transcription factor enrichment analysis

To identify which pathways and transcription factors (and downstream targets) are up- or downregulated in the cells, `decoupleR`[85] was used. `DecoupleR` computes activity scores based on prior knowledge networks. For pathway analysis, the network used was `PROGENy`[80], and for the transcription factor enrichment analysis, `DoRothEA`[81]. The resulting activity scores were calculated for each cell and then aggregated by cell population using `SCpubr::do_PathwayActivityPlot()` and `SCpubr::do_TFActivityPlot()`. In the case of the scores based on `DoRothEA`, the top 30 scoring regulons were selected for plotting.

Ligand-Receptor analysis

Inference of ligand-receptor pairs between tumor and microglia subpopulations was performed using `liana` (v.0.1.10)[90]. Aggregated consensus rank was used as filtering metric, excluding interactions for which `rank < 0.05`. From the statistically significant interactions, a biologically relevant subset was selected and then visualized as a dot plot.

This analysis was a collaborative effort. Dr. Marc Elosua-Bayes contributed the code, Dr. Inmaculada Hernández conducted the analysis and Dr. Juan C. Nieto provided the biological expertise to select the relevant interactions. Data visualizations were generated by me.

Diffusion analysis

To characterize tumor subpopulations based on their stemness activity, I gathered various stemness-related marker sets from PanglaoDB[191] and previous publications[252]. Tumor microenvironment cell populations were filtered out before the analysis. Diffusion maps were computed using the `destiny`[253] R package, with the function `destiny::DiffusionMap()`. For this purpose, the `Seurat` objects were converted into Single Cell Experiment (SCE) objects, for

which the function `Seurat::as.SingleCellExperiment()` was used. The results were visualized in two ways: as a dimensional reduction plot with `SCpubr::do_DimPlot()`, and as a bar plot where scaled and centered scores were plotted over the ranking of cell positions along a given diffusion component with `SCpubr::do_DiffusionMapPlot()`.

Deconvolution analysis

To validate the presence of the tumor subpopulations in external datasets, a deconvolution analysis was performed on the cancer genome atlas program (TCGA) and the Chinese glioma genome atlas (CGGA) IDH-mutant glioma bulk transcriptomics datasets using `SPOTlight`[101] (v1.0.3). Initially, the model was trained on a pseudobulked version of the datasets to evaluate its accuracy in predicting the proportion of each cell type within the same datasets. Marker genes for each cell population were retrieved using `Seurat::AddModuleScore()` with the following parameters: `min.pct = 0.5`, `test.use = "MAST"`, `logfc.threshold = 0.5` and `only.pos = TRUE`. The `avg_log2FC` scores obtained were used as weights to initialize the NMF method in `SPOTlight`. The correlation between true and estimated proportions per cell type was calculated patient-wise and displayed as a dot plot, along with the respective correlation coefficients. After confirming a sufficient correlation between the proportions, the datasets coming from TCGA and CGGA were processed. The resulting estimated proportions were displayed as a stacked bar plot.

In this section, the method generation and implementation were carried out by Dr. Marc Elosua-Bayes and Dr. Inmaculada Hernández, while I generated the figures for the analysis.

Generation of custom prior knowledge networks for activity inference analysis

To generate a custom prior knowledge network for use with `decoupleR`, I created a data frame containing that included the pathway name, the genes involved in the pathway and their mode of regulation was generated. Since all the genes were used as marker genes for the analysis, the mode of regulation, which can range between `-1` to `1`, was set to `1` for all genes to indicate that they positively influence the associated phenotype.

Single-nucleus ATAC-seq analysis

The analysis of snATACseq data was conducted using `Signac`[109] R package. Dr. Inmaculada Hernández performed all bioinformatic analyses, while I was responsible for generating the figures.

Quality control and Normalization of count data

To filter out low-quality cells in the snATACseq data, the following variables were defined: the number of fragments per cell (`peak_region_fragments`), the percentage of reads in peaks (`percentage_reads_in_peaks`), the number of reads located in ENCODE's blacklist regions[111] (`blacklist_ratio`), the chromosome binding pattern (`nucleosome_signal`, computed with `Signac::NucleosomeSignal()`), and the enrichment of the peaks transcription start site regions (`TSS.enrichment` , computed with `Signac::TSSEnrichment()`). Cells not meeting the following criteria were filtered out: `3000 < peak_region_fragments < 20000`, `percentage_reads_in_peaks > 15%`, `blacklist_ratio < 0.05`, `TSS.enrichment > 2`. Peak annotation for downstream analyses was conducted using `Ens.Db.Hsapients.v86`[254] R package.

Dimensional reduction and clustering

Dimensional reduction was performed by computing term frequency-inverse document frequency (TF-IDF) using `Signac::RunTFIDF()`, followed by selecting all peaks in the dataset with `Signac::FindTopFeatures()`, using the parameter: `min.cutoff = "q0"`. Singular value decomposition (SVD) was then applied to the TF-IDF matrix, yielding a total of 30 components. The first component was removed as it correlated with sequencing depth. Non-linear dimensionality reduction was subsequently carried out using `Seurat::RunUMAP()`, and cells were clustered using `Seurat::FindNeighbors()` and `Seurat::FindClusters()`.

Computing gene activity, label transfer and integration of the data

Gene activity can be estimated based on the accessibility associated with each gene. This was done using `Signac::GeneActivity()`, which created a new "ACTIVITY" assay representing raw counts. These raw counts were further normalized using `Seurat::SCTransform()`. To generate a co-embedding from the RNA and ATAC (ACTIVITY) assays, shared anchors were identified using `Signac::FindTransferAnchors()` following `Signac::TransferData()`,

corresponding to expression trends shared between the RNA and the ATAC assay[109]. The RNA assay was used as a reference since it was already annotated.

This resulting co-embedding was integrated with harmony using `harmony::RunHarmony()` to remove patient-specific effects, and a new UMAP was generated using the same number of LSI components as previously mentioned.

Transcription factor enrichment analysis

To infer gain or loss of accessibility on peaks within TF motifs, `Signac::AddMotifs()` was used, which is a wrapper for `chromVAR` (version 1.14.0)[117]. Motif activity scores across tumor cell populations were compared using `Seurat::FindAllMarkers()` with default parameters. Significant motifs were annotated to their closest gene with `Signac::ClosestFeature()`, and the top motifs enriched across populations were displayed as a heatmap.

Immunohistochemistry

Immunohistochemistry was performed by J.T., M.C., and P.L.P. as described in Blanco-Carmona*, et al*.[189] The following paragraph is extracted from the publication:

"2 µm sections were cut from formalin-fixed paraffin-embedded (FFPE) tissue samples provided by the Pathological Department of Spedali Civili of Brescia. Sections were de-waxed and rehydrated. Endogenous peroxidase activity was blocked with 0.3% H_2O_2 in methanol for 20 min. Antigen retrieval was performed using a microwave oven or a thermostatic bath in 1.0 mM EDTA buffer (pH 8.0) or in 1.0 mM Citrate buffer (pH 6.0). Sections were then washed in tris-buffered saline (TBS, pH 7.4) and incubated for 1 h with the specific primary antibody diluted in TBS 1% bovine serum albumin. The reaction was revealed by using Dako EnVision System-HRP Labeled Polymer anti-mouse or anti rabbit (Dako) or Novolink Polymer Detection System (Novocastra) followed by diaminobenzydine (DAB) as chromogen and hematoxylin as counterstain. For double immunohistochemistry, after completing the first immune reaction, the second one was revealed by using MACH4 Universal AP Polymer kit (Biocare Medical) followed by Ferangi Blue Chromogen kit (Biocare Medical) and nuclei were counterstained with hematoxylin. Images were acquired with a Nikon DS-Ri2 camera (4908 x 3264 full-pixel) mounted on a Nikon Eclipse 50i microscope equipped with Nikon Plan lenses (x10/0.25; x20/0.40; x40/0.65; x100/1.25) using NIS-Elements 4.3 imaging software (Nikon Corporation).

The following primary antibodies were used: anti-Iba1 rabbit polyclonal (1:300, Wako), anti-CD74 mouse monoclonal (clone LN2) (1:100, BioLegend), anti-CD163 mouse monoclonal (clone 10D6) (1:50, ThermoFisher Scientific), anti-STAT1 (pY701) mouse monoclonal (clone 14/P-STAT1) (1:500, BD Biosciences), anti-EEF2 rabbit monoclonal (EP880Y) (1:250, Abcam), and anti-EEF1A1 rabbit monoclonal (EPR9470) (1:50, Abcam).

For IHC staining of human astrocytomas, 5µm FFPE sections were obtained from the Department of Neuropathology, Heidelberg University Hospital, and stained with the following antibodies: mouse monoclonal CSF1 antibody (1:25, clone 2D10, Sigma-Aldrich #MABF 191), rabbit polyclonal CSF1R antibody (1:50, Proteintech #25949-1-AP), rat monoclonal human PIGF antibody (1:25, clone #358905, R&D #MAB 2642), mouse monoclonal NRP1 antibody (1:50, clone 2H3F6, Proteintech #60067-1-Ig). For CSF1, a pre-treatment steamer with Tris buffer pH 9.0 was performed followed by overnight incubation at room temperature. Detection was performed with ImmPress HRP, universal antibody (horse, anti-mouse) polymer detection kit, peroxidase (Vector # MP 7500). For CSF1R, we performed the same incubation and antigen retrieval as for CSF1 and detected with Dako REAL Detection System, Alkaline Phosphatase/RED, Rabbit/Mouse (Biocompare #K5005). For PIGF, pre-treatment with citrate buffer pH 6.0 was performed, followed by overnight incubation at room temperature. Detection was performed using the secondary antibody - biotinylated anti-rat IgG (Vector # BA-4001, 1:200) for 300 at 37°C, followed by HRP Streptavidin (1:200, Vector #SA-5004), for 30' at 37°C, and DAB (brown color) was used as a chromogen (ImmPress HRP Universal Antibody (Horse Anti-Mouse/Rabbit IgG) Polymer detection Kit, Peroxidase, (Vector # MP 7500). For NRP1 IHC, samples were pretreated with citrate buffer pH 6.0 and incubated for 2 h at 37°C, and detected using the Dako REAL Detection System, Alkaline Phosphatase/RED, rabbit/mouse, (Biocompare #K5005). Samples were imaged at 40x magnification with an Olympus VS.200 slide scanner (Olympus Corporation)."

Results

To gain deeper insights into the tumor heterogeneity between IDH-mutant gliomas subtypes (oligodendrogliomas and astrocytomas), 14 snap-frozen 10X v3 3' primary tissue datasets were processed and sequenced in the lab of Dr. Şevin Turcan. These datasets were entrusted to me for further analyses. My initial step was to get a general overview of the datasets, for which I examined the metadata associated with each patient sample (Figure 26). In summary, the datasets provided a balanced representation across various dimensions: IDH-mutant glioma subtypes (OD = 8, AS = 6), tumor grades (Grade 2 | OD = 3, Grade 3 | OD = 5, Grade 2 | AS = 2, Grade 3 | AS = 4) and sequencing methodologies (snRNAseq | OD = 8, snRNAseq | AS = 6, snATACseq | OD = 4, snATACseq | AS = 5). Furthermore, patients exhibited diversity in MGMT methylation status (Methylated = 6, Not available = 8), TERT status (WT = 3, C228T = 2, C250T = 1, Not available = 8) and gender distribution (Female = 5, Male = 9). Collectively, these datasets present a promising avenue for delving deeper into the biology of IDH-mutant gliomas at a single cell resolution across both oligodendrogliomas and astrocytomas, thus constituting a valuable resource for the scientific community.

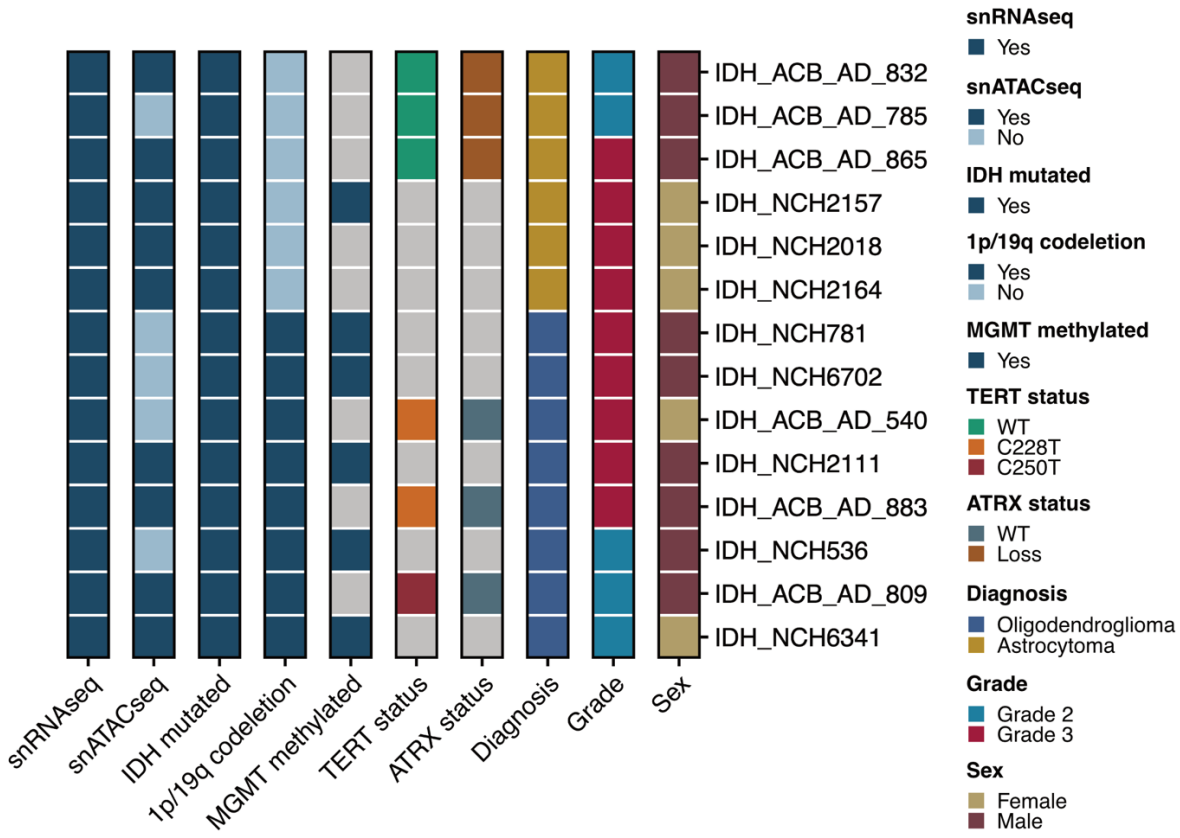


Figure 26: "Metadata of primary samples. Clinical and molecular characteristics of the IDH-mutant glioma cohort for single-nuclei sequencing." Reprinted from Blanco-Carmona, *et al.*[189]

Identifying tumor cells apart from tumor micro-environment

Supervised annotation based on literature gene marker sets

To comprehend the biological nature of the different cell clusters within oligodendroglioma and astrocytomas, first I needed to understand whether the cells belonged to the tumor microenvironment (TME) or were tumor cells. To accomplish this, the cells underwent quality control and dimensional reduction. Subsequently, I analyzed the cells in the context of a merged UMAP. Normally, cells that belong to the TME usually typically cluster together prior to integration, whereas the tumor cells tend to segregate into distinct clusters, each comprising cells predominantly originating from a single patient (Figure 27A-B).

After identifying the clusters that belonged to the TME, my next step was to determine their identity. To achieve this, I conducted enrichment analysis using literature-derived gene marker sets sourced from PanglaoDB[191] (Figure 27C-D). The results showcased cell clusters that were uniquely enriched in specific gene sets, therefore confirming their identity. This allowed for the characterization of: microglia, oligodendrocytes, astrocytes, neurons and pericytes (Figure 27E-F). Subsequently, I integrated the datasets together using Harmony, removing the effect originating from the individual patients. This integration yielded a new UMAP embedding where TME cell populations formed single clusters, while the tumor cells would also aggregate together into a large cluster (Figure 28A-B). At this stage of the analysis, I opted to exclude from the oligodendroglioma dataset a cell cluster comprised solely of cells from a single patient, whose data exhibited low quality. This cluster integrated with microglia cells (Figure 27E, Figure 28A), with subsequent CNV analysis also revealing chromosome arms 1p and 19q codeletion.

Inferring copy number variant events in IDH-mutant gliomas

In the context of oligodendrogliomas, a defining characteristic is the presence of chromosome 1p/19q codeletion, a copy number variant event crucial to their classification. Thus, I anticipated observing this codeletion across the tumor clusters, while being absent in TME clusters. To infer CNV scores, I employed inferCNV, originally developed for inferring CNV events on SMARTseq data[95], and designed an optimization experiment to enhance its efficiency in 10X datasets.

In this experiment, CNV scores were inferred under different scenarios, and efficiency was benchmarked based on its ability to successfully identify chromosome 1p deletion in oligodendrogliomas. Initially, endothelial cells served as a reference, which identified chromosome 1p loss across some of the clusters, therefore deeming clusters that lacked this deletion as TME (Figure 29A). However, the analysis output exhibited a notable amount of noise, with regions classified as gains or losses where no such events should occur. To minimize this noise, I iteratively refined the method by adjusting the clusters utilized as reference (Figure 29B) and also treating them as a single unique reference cluster (Figure 29C).

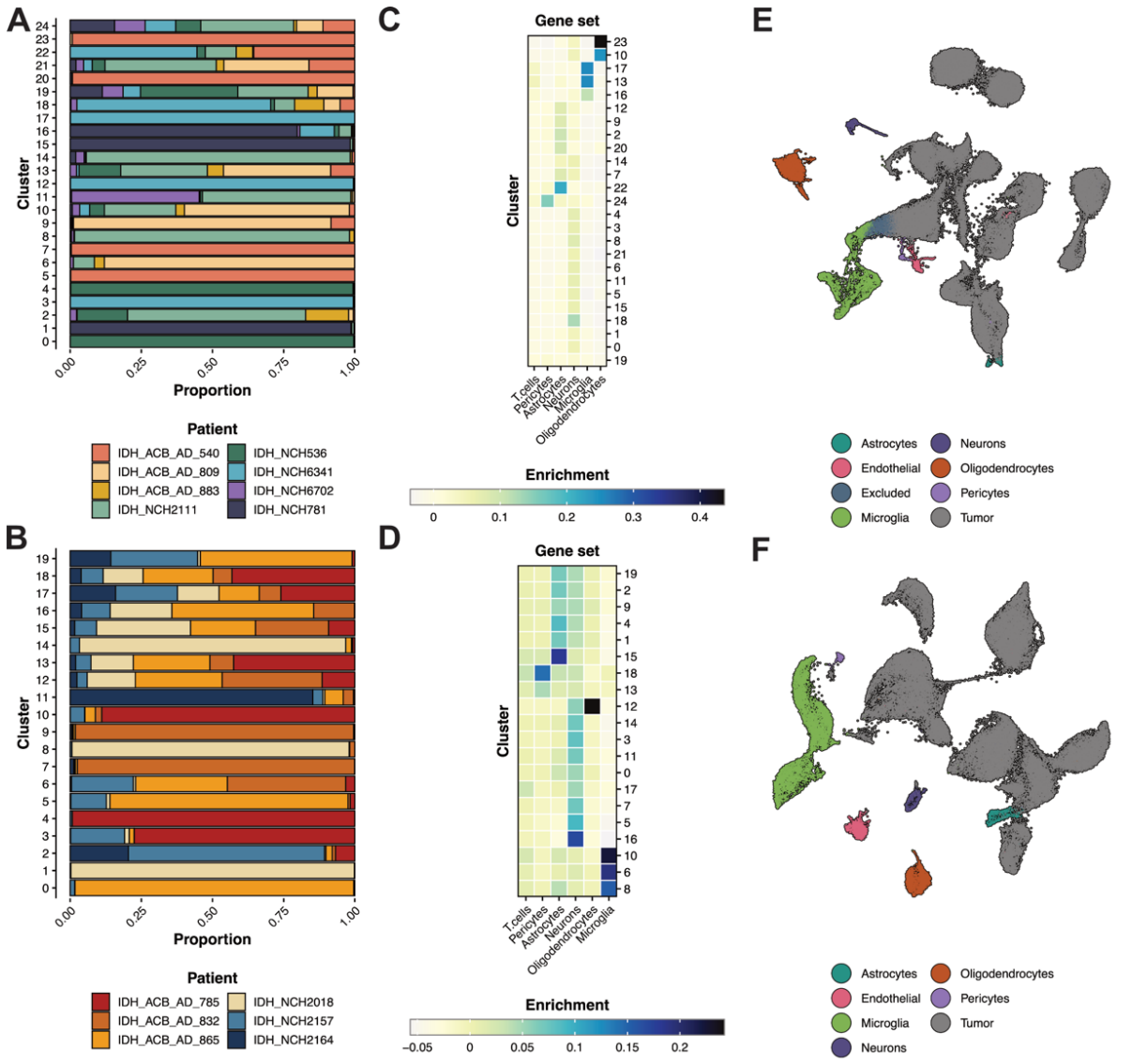


Figure 27: TME annotation. Patient composition across cell clusters for primary oligodendrogliomas **(A)** and astrocytomas **(B)**. Enrichment scores for selected TME cell type gene marker sets for primary oligodendrogliomas **(C)** and astrocytomas **(D)**. UMAP representation for primary oligodendrogliomas **(E)** and astrocytomas **(F)** showcasing TME annotation. UMAP1, x-axis; UMAP2, y-axis.

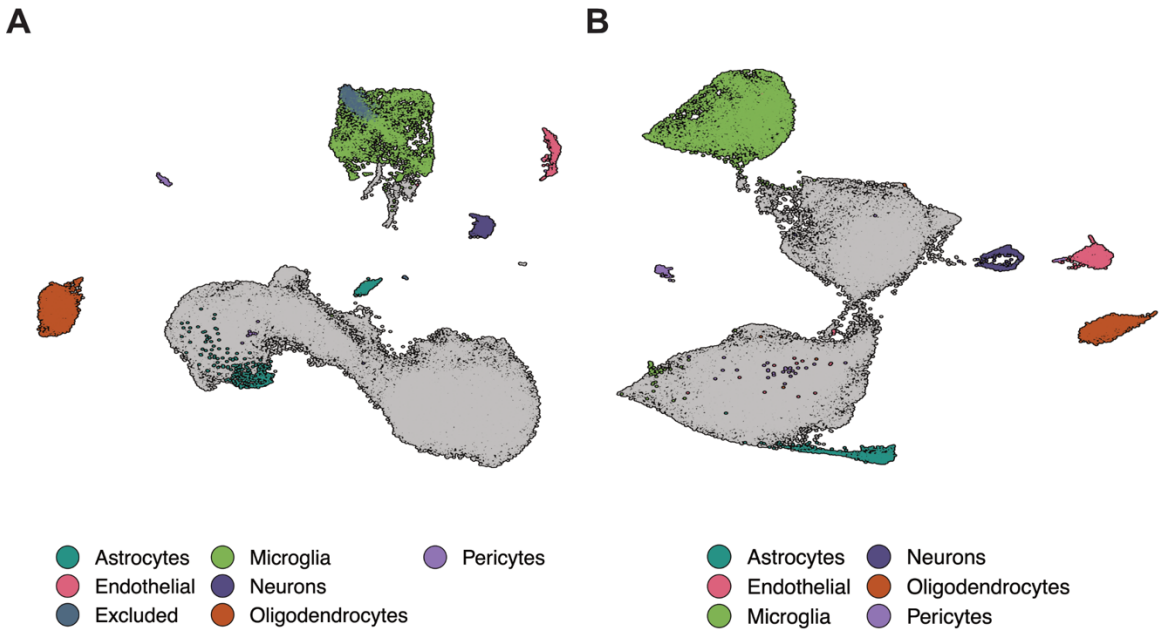


Figure 28: Integration with Harmony on primary datasets. Integrated UMAP representation of primary oligodendroglioma **(A)** and astrocytoma **(B)**. The effect of individual patients is removed during integration. Grey cells are part of the tumor compartment. UMAP1, x-axis; UMAP2, y-axis.

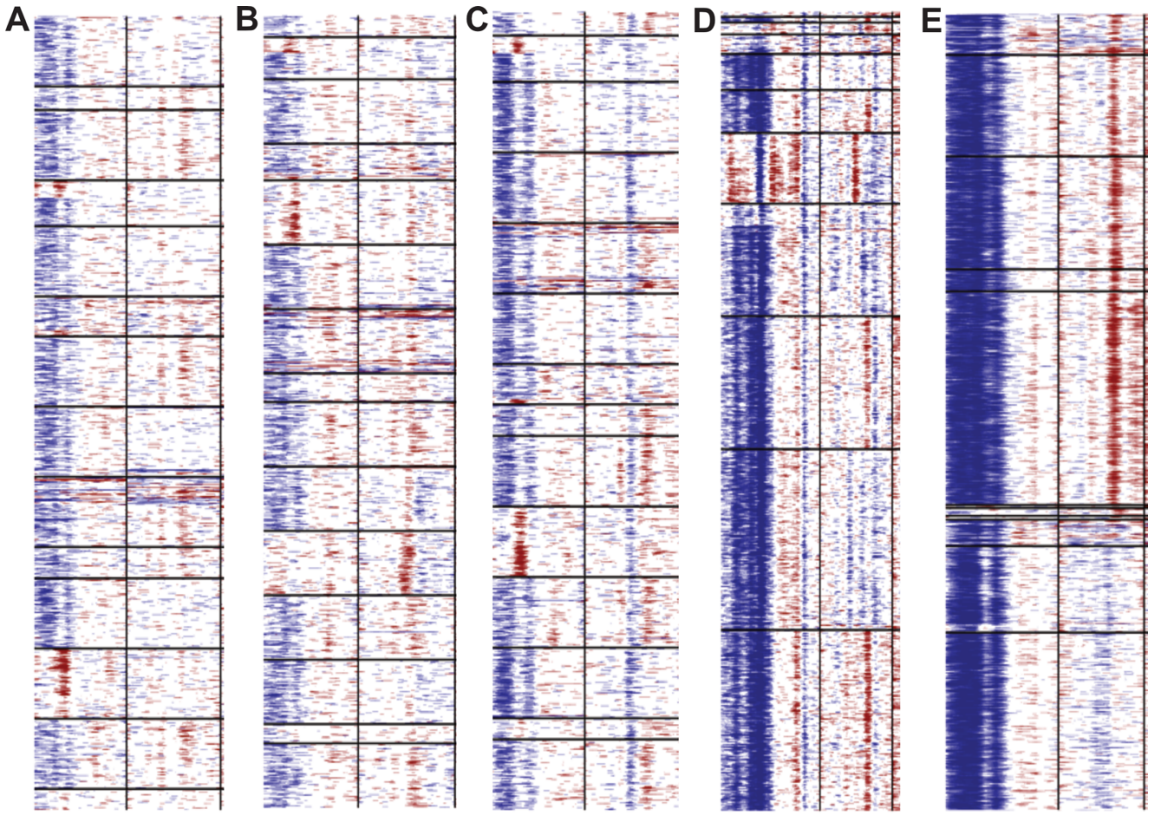


Figure 29: Enhancing the sensitivity of inferCNV for 10X datasets. Different setups were tested in an inferCNV run. Each panel represents a zoomed-in cut into chromosome 1p scores as a comparison metric. Reference panels and cluster annotations are not included. From left to right: **(A)** using microglia and endothelial cells as reference, **(B)** only endothelial cells as reference, **(C)** including microglia and endothelial cells as reference but treating them as a single cluster, **(D)** using oligodendrocyte cells and computing metacells, **(E)** and using oligodendrocyte cells, computing metacells, and setting the window size to 201 genes. Each row is a cell in the dataset and each column depicts the selected window size of genes. Chromosome gains are depicted as red, while chromosome loses are shown as blue.

Both approaches yielded marginal improvements compared to the first scenario. Subsequently, to increase the sensitivity of the analysis, I aggregated the count data of five different cells from the same cluster into metacells. The use of metacells greatly increased the sensitivity of the CNV calling, allowing for a clear detection of chromosome 1p deletion (Figure 29D). Finally, despite experimenting with increasing the number of genes accounted for in a single window, there were no notable improvements observed (Figure 29E). Consequently, based on these iterative analyses, I determined that generating metacells without modifying the window size (Figure 29D) yielded optimal results.

This configuration was thus adopted for the final CNV inference used for downstream analyses. The results indicated that, in the context of oligodendrogliomas, chromosome 1p codeletion could be reliably inferred and used to identify tumor cells. However, the same level of clarity as not observed for chromosome 19q deletion, which exhibited an overall diminished signal and, in certain instances, lacked deletion scores even when corresponding 1p deletion was present (Figure 30A). Conversely, in the case of astrocytomas, no recurrent CNV event was identified across patients (Figure 30B). Therefore, to infer the tumorigenic state of the cells within my datasets, the patient composition of the clusters in a merged UMAP, along with enrichment in marker gene sets were utilized for astrocytomas, while also relying on the chromosome 1p deletion scores for oligodendrogliomas.

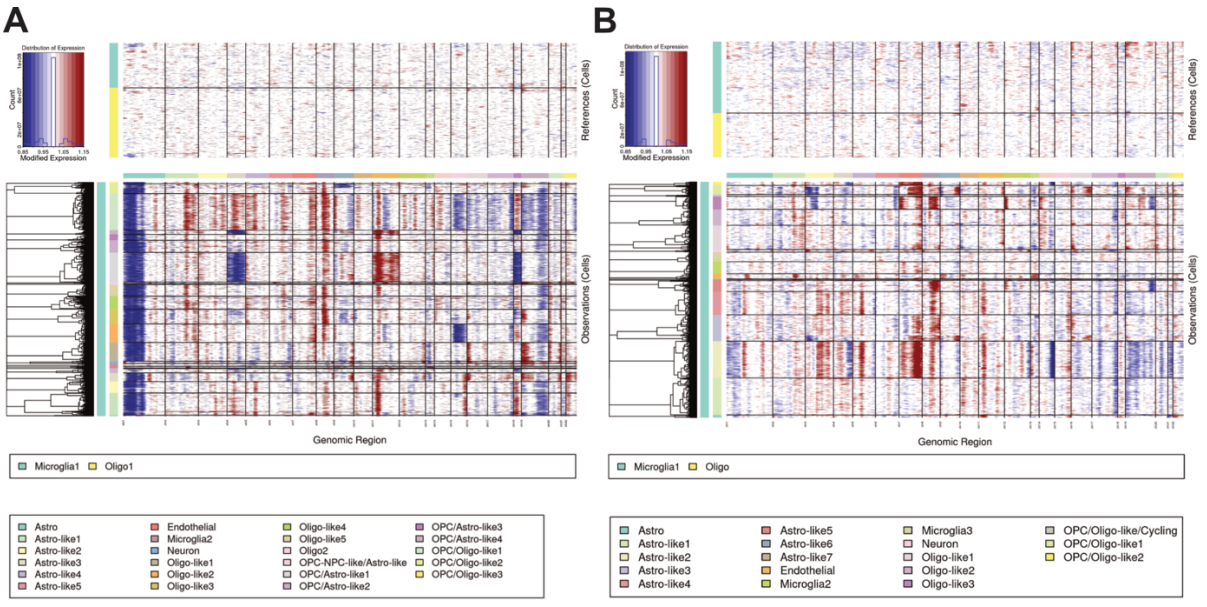


Figure 30: "Output of inferCNV for OD and AS. (A and B) Heatmap of CNV profiles inferred from snRNA-seq from oligodendrogliomas **(A)** and astrocytomas **(B)**. Each row corresponds to a nucleus, ordered by initial cluster labeling from merged data. Red indicates gain and blue indicates loss." Adapted from Blanco-Carmona, *et al.*[189]

Inferring recurrent transcriptional programs across tumor cells from different patients

The next step in the analysis is to identify the different tumor cell populations within the dataset. Typically, there are several approaches, broadly classified as either supervised or unsupervised methods, to achieve this goal. Supervised annotation methods aim to identify tumor cell populations based on existing knowledge, often relying on gene marker sets known to be associated with specific phenotypes. On the other hand, unsupervised methods aim to uncover patterns within the expression data in the dataset, which can then be analyzed to assign phenotypes. Previously, it has been established that IDH-mutant gliomas harbor three primary tumor cell types, typically exhibiting astro-like, oligo-like and stemness-like phenotypes[135,252]. With this knowledge in mind, I sought to ascertain whether these cell populations were also present in my datasets. To do so, I computed enrichment scores on the marker genes reported in Venteicher, *et al.*[252] Out of the three expected tumor cell populations, I identified clusters enriched in oligo-like and astro-like gene marker sets, while enrichment for stemness markers was not observable in my datasets (Figure 31A-B).

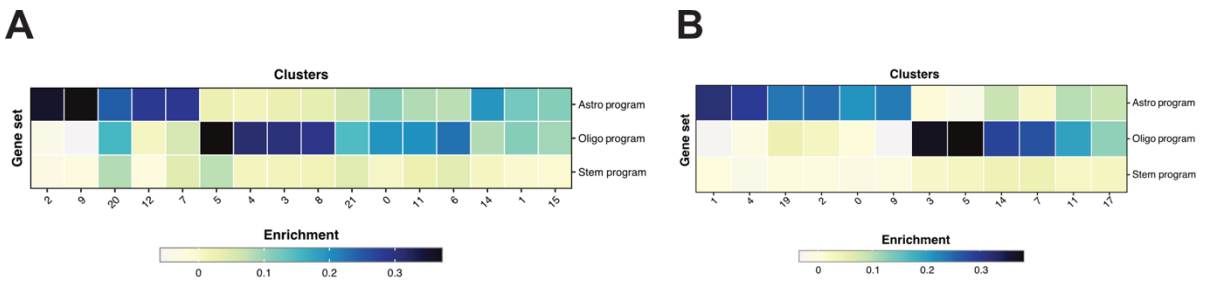


Figure 31: Gradient enrichment of astro and oligo programs. Enrichment in the gene programs reported in Venteicher, *et al.*[252] across tumor clusters for both primary oligodendrogliomas **(A)** and astrocytomas **(B)**. Clusters, x-axis; Gene sets, y-axis.

Furthermore, in both tumor populations, the enrichment in astro-like and oligo-like markers manifested in a gradient fashion, distinctly forming two ends of a spectrum. This suggests that the tumor cells may potentially differentiate along these two trajectories, with cells at various differentiation stages captured at the time of sequencing.

NMF analysis: choice of rank and retrieval of NMF metaprograms

To further explore tumor heterogeneity in IDH-mutant gliomas beyond the described astro-like and oligo-like tumor populations, I employed NMF. The method yielded four sets of correlated NMF programs for oligodendrogliomas and three for astrocytomas (Figure 32A-B). Interestingly, all NMF metaprograms identified in astrocytomas exhibited gene-wise overlap

with those in oligodendrogliomas, encompassing OPC-like (*OPCML*, *DSCAM*), astro-like (*NGR3*, *SPARCL1*, *ADGRV1*) and cycling (*MKI67*, *CENPK*) phenotypes.

The NMF metaprograms recapitulating astro-like and OPC-like phenotypes closely resembled those described in Venteicher, *et al.*[252] The fourth NMF metaprogram in oligodendrogliomas featured stemness genes (*OLIG1*), oncogenes (*ETV1*), elongation factors (*EEF2*, *EEF1A1*) and ribosomal genes, which I termed ribosomal enriched (RE). This represented a novel tumor population in the field of IDH-mutant gliomas. Consequently, downstream analyses will explore the biological nature of the RE population.

Permutation testing to overcome gradient-based enrichment on cell annotation

To annotate the tumor cells in both oligodendrogliomas and astrocytomas datasets based on the retrieved NMF metaprograms, I calculated enrichment scores. The enrichment in each NMF metaprogram but the cycling exhibited a gradient pattern, suggesting that tumor cells would progressively transition towards each of the phenotypes associated with the NMF metaprograms (Figure 33A-B).

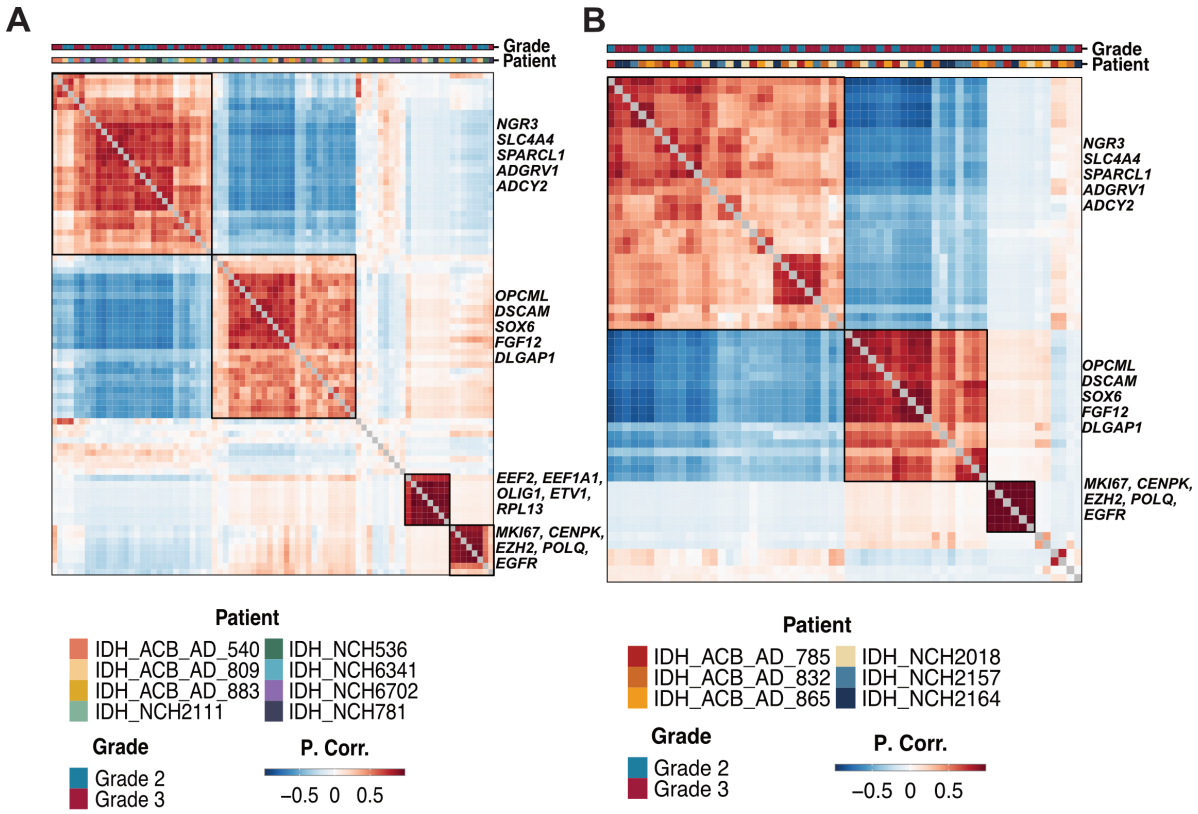


Figure 32: "NMF metaprograms for OD and AS. (A and B) Pearson's correlation scores for individual NMF programs (rows and columns) in oligodendrogliomas (n = 8) **(A)** and astrocytomas (n = 6) **(B).** Metaprograms were identified by hierarchical clustering of individual NMF programs." Reprinted from Blanco-Carmona, *et al.*[189]

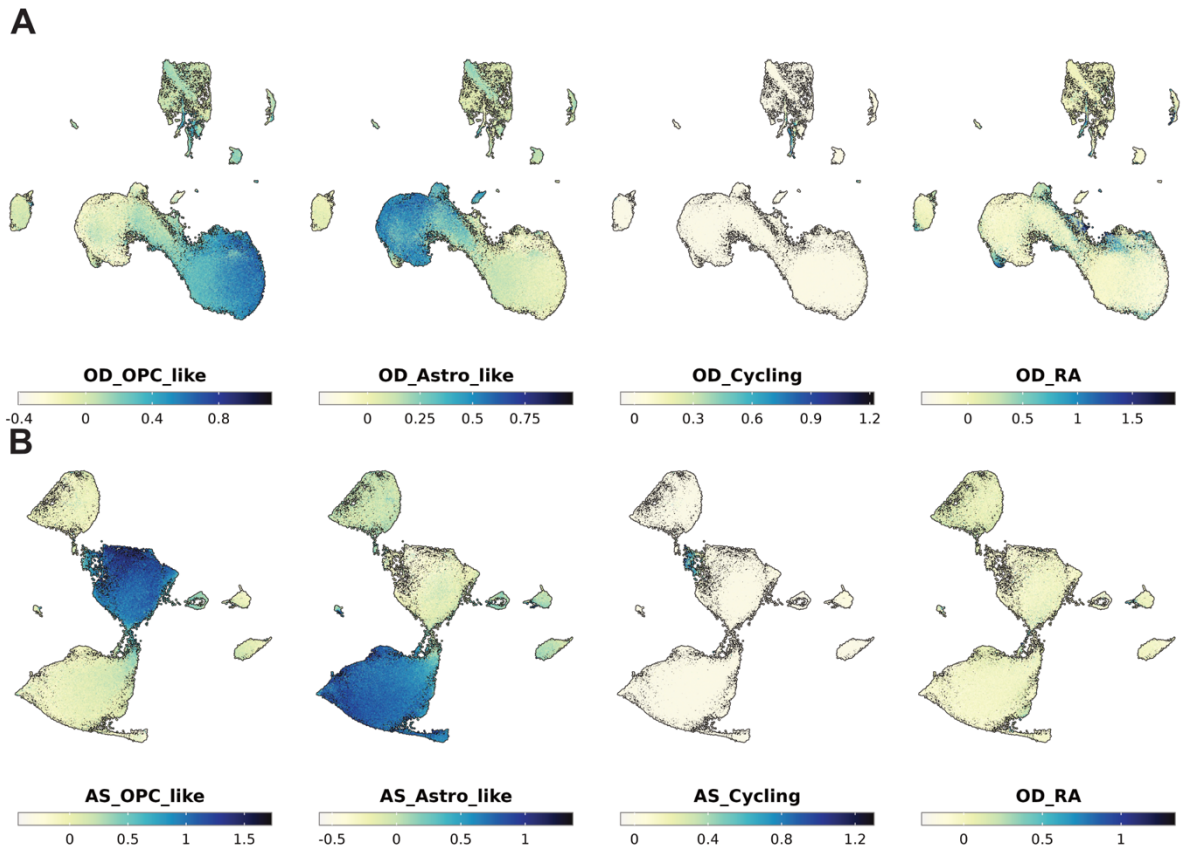


Figure 33: Enrichment in NMF metaprograms. Enrichment scores for each NMF metaprogram retrieved for both oligodendrogliomas **(A)** and astrocytomas **(B)**. AS: Astrocytoma, OD: Oligodendroglioma, RA: Ribosomal active (later renamed as RE). UMAP1, x-axis; UMAP2, y-axis.

Given the absence of clear enrichment within specific cell clusters, it became a necessity to devise a method for statistically selecting cells based on enrichment for the NMF metaprograms. Statistical significance was crucial, as I intended to compare tumor cell type proportions across IDH-mutant glioma subtypes later in the analysis. The method I devised centered around permutation testing, which compares the distribution of enrichment scores to that of a control gene set. This approach provides a corrected p-value to each cell for each NMF metaprogram.

Tumor cell type annotation was conducted in two stages. Initially, cells were classified into either astro-like or OPC-like as a baseline phenotype. Subsequently, they were assigned to either cycling or RE if deemed significant. Cells failing to meet the significance threshold for any of the NMF metaprograms were designated as gradient cells, indicating their position in the middle of the differentiation trajectory and thus insufficiently differentiated into any of the retrieved NMF metaprograms. This resulted in a tumor annotation in which, for both oligodendroglioma and astrocytoma datasets, astro-like and OPC-like tumor populations

occupied opposite ends of the UMAP, with gradient cells in between and a cluster of cycling cells detached from the rest of tumor cells. RE cells were dispersed across the UMAP, failing to form a single cluster (Figure 34A-B). Despite being predominantly present in a single patient, RE cells were also observed in lesser proportions in other patients in both oligodendroglioma and astrocytoma datasets (Figure 34C). Across IDH-mutant glioma subtypes, tumor populations exhibited a similar expression profile for key marker genes of each NMF metaprogram (Figure 34D).

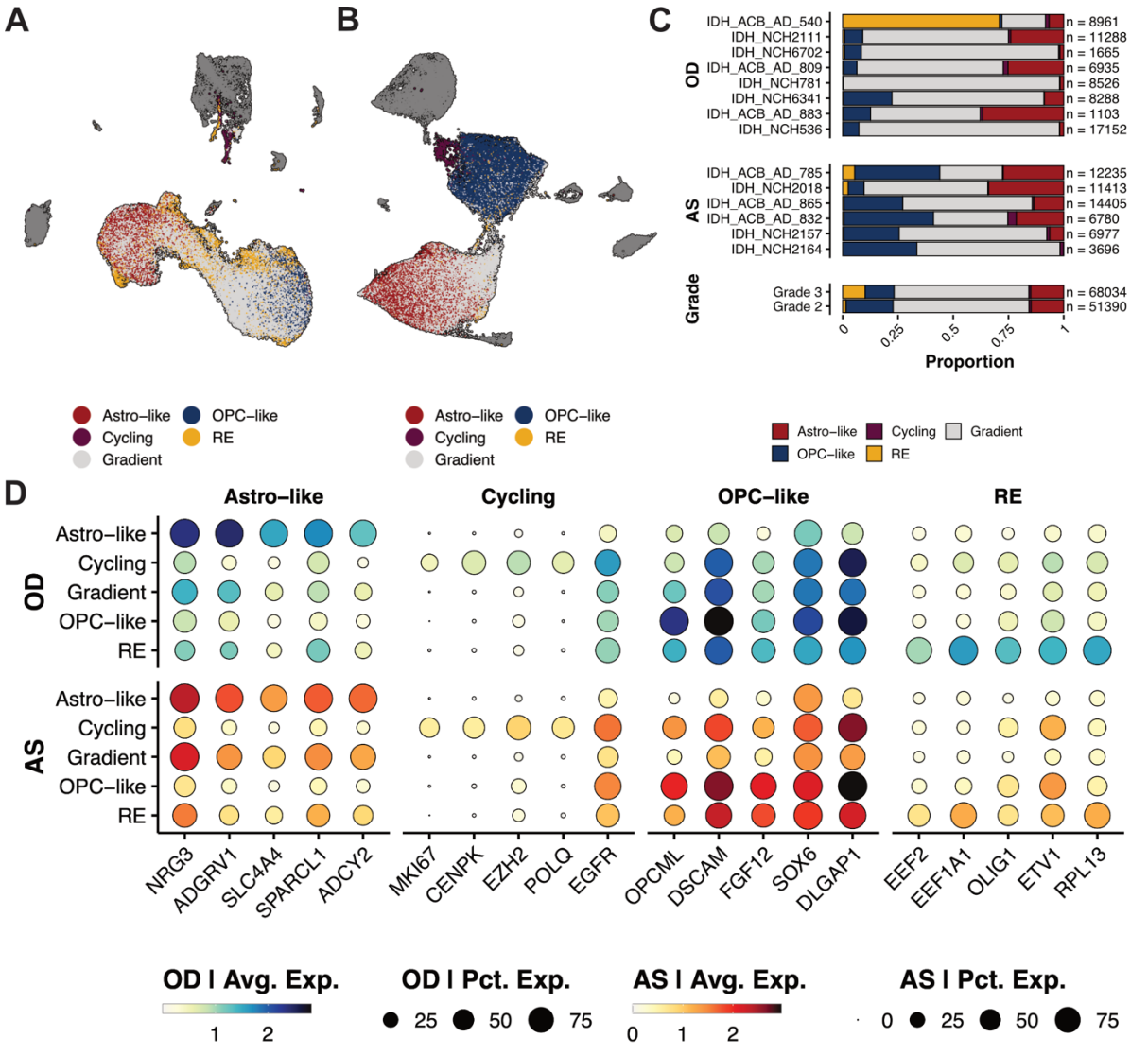


Figure 34: "Overview of tumor annotation for OD and AS. (A and B) UMAP embedding of oligodendrogliomas **(A)** and astrocytomas **(B)**, colored by NMF metaprograms, gradient cells, and TME (dark gray). UMAP1, x-axis; UMAP2, y-axis. **(C)** Bar plot showing tumor population proportions across individual samples, grouped by subtype and grade. Bars are arranged by descending RE proportion. UMAP1, x-axis; UMAP2, y-axis. **(D)** Dot plot displaying selected five marker genes for each NMF-derived tumor population for both oligodendrogliomas and astrocytomas." Reprinted from Blanco-Carmona, *et al.*[189]

The biological nature of the gradient population

Despite being the result of applying an annotation method based on statistical testing, the gradient tumor cell population still comprised the majority of the tumor cells. Therefore, while downstream analyses will primarily focus on the other tumor cell populations, it was important to validate the biological nature of the gradient cells.

To accomplish this, gradient cells for both primary (oligodendroglioma and astrocytoma) and primary-recurrent datasets (see below) were isolated, reclustered, and a new UMAP embedding was generated. Subsequently, for each dataset, enrichment scores for the NMF metaprograms and the marker genes sets from Venteicher, *et al.*[252] were computed.

The results revealed that enrichment scores for OPC-like and astro-like metaprograms also exhibited a gradient pattern (Figure 35A-C), further confirming the lineage nature of these cells. This validation reinforces the decision to focus on the remaining, significantly enriched, tumor cell populations in downstream analyses.

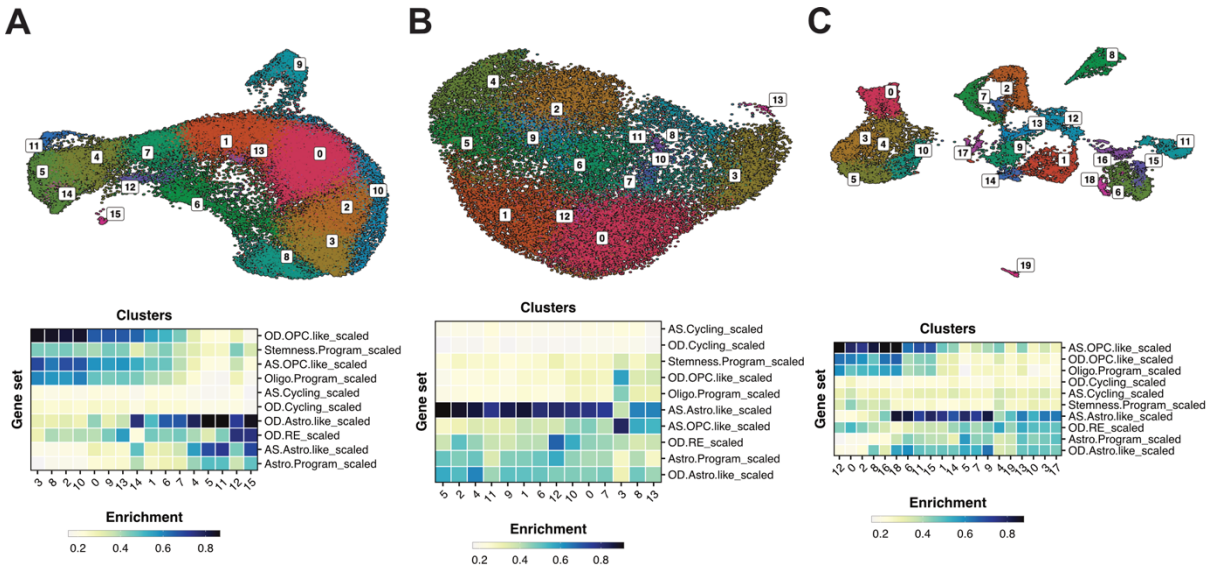


Figure 35: "Understanding the biological nature of the Gradient tumor cell population. (A, B, and C) Gradient subset of the tumor subpopulation for primary oligodendrogliomas **(A)**, primary astrocytomas **(B)** and paired astrocytomas **(C)**. Cells are re-normalized, and dimensional reduction is computed, generating a new UMAP embedding and clustering (top). For each cluster, enrichment scores for the NMF metaprograms and the programs described in Venteicher, *et al.*[252] are computed, and displayed as an enrichment heatmap (bottom). UMAP1, x-axis; UMAP2, y-axis." Reprinted from Blanco-Carmona, *et al.*[189]

The RE tumor cell population: a true biological tumor cell population or a bias in the analysis

To ascertain whether the presence of the RE population stemmed from a bias in the analysis or constituted a biological finding, I initially examined the distribution of UMI, genes and mitochondrial RNA across each cell population. This analysis confirmed that the RE population exhibited a distribution similar to that of the other high-quality cell populations within the datasets (Figure 36A-C).

Subsequently, I recomputed NMF in the oligodendroglioma dataset under various conditions: excluding mitochondrial genes (OD/AS_), excluding mitochondrial and ribosomal genes (N_), and excluding mitochondrial and ribosomal genes together with removing the patient accounting for the majority of RE cells (N2_). In each iteration, four NMF metaprograms were obtained. Jaccard similarity analysis revealed a high concordance between each iteration of OPC-like, astro-like and RE NMF metaprograms (Figure 36D). With the consistency of the RE NMF metaprogram confirmed across iterations, I proceeded to apply the permutation testing method to identify cells enriched for each iteration of the RE metaprogram. Jaccard similarity comparing the different sets of statistically enriched cells in the primary oligodendroglioma, primary astrocytoma and primary-recurrent astrocytoma datasets (see below) demonstrated that mainly the same subset of cells were selected in each iteration and dataset (Figure 36E-G). These results support the validity of the RE population as a biological entity rather than as an artifact of the analysis.

Validation of the presence of RE population in external datasets

Given the novelty and potential relevance of the RE population in future research in IDH-mutant gliomas, I aimed to ascertain the presence of the RE population beyond my datasets. For this, I employed several approaches.

Deconvolution analysis

The following analysis was performed by I.H. and M.E.B. in Blanco-Carmona*, et al.*[189] The cancer genome atlas (TCGA) and the Chinese glioma genome atlas (CGGA) IDH-mutant glioma bulk transcriptomics datasets were deconvoluted using Spotlight[101]. Both cohorts were divided into oligodendroglioma or astrocytoma patient samples.

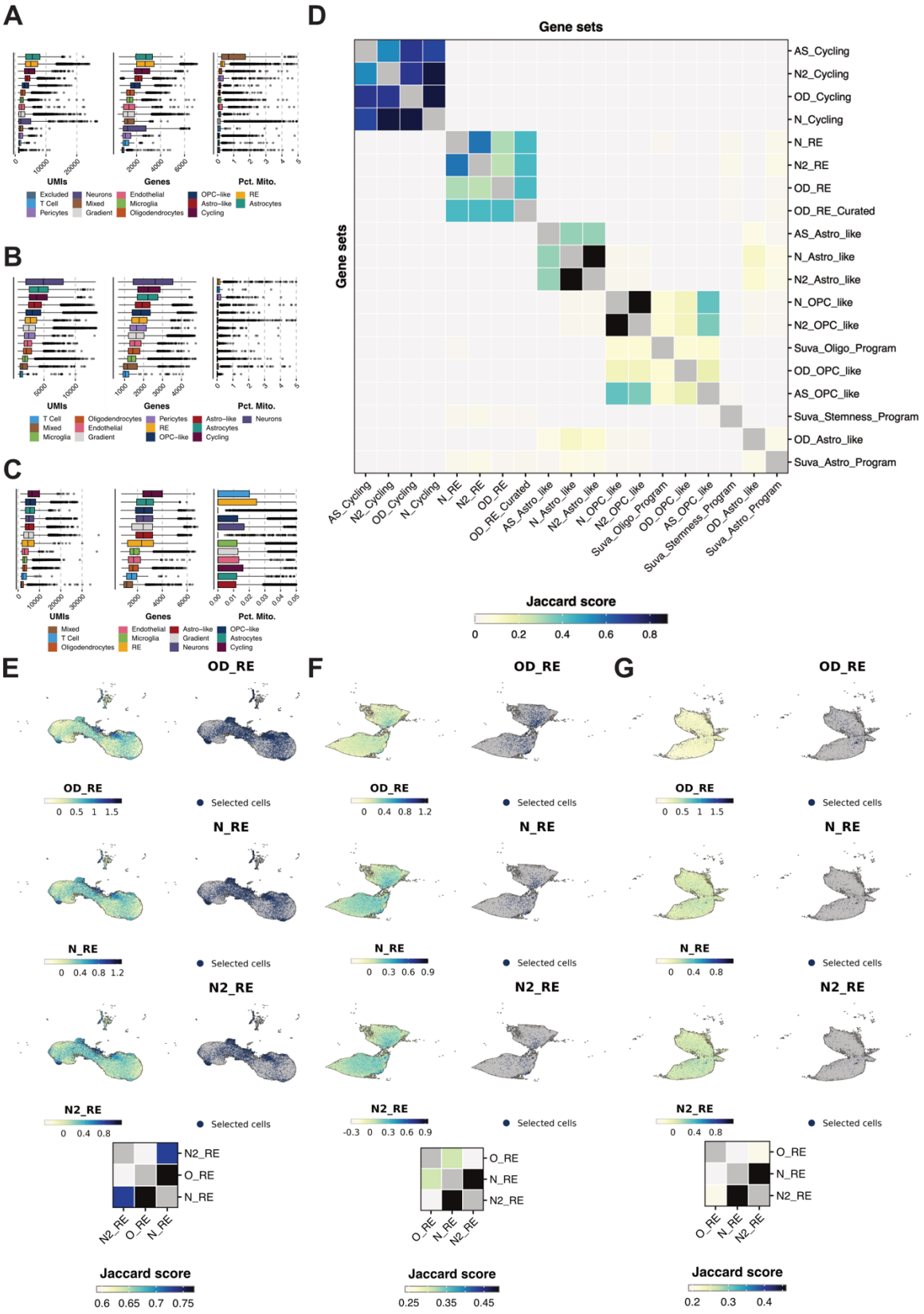


Figure 36: "Determining RE as a true biological tumor population, not arising from bias effects. (A-C) Boxplots showing from left to right, distribution of UMIs per cell, genes per cell and percentage of mitochondrial RNA per cell are shown for primary oligodendroglioma **(A)**, primary astrocytoma **(B)** and paired astrocytoma **(C)** tumors. **(D)** Correlation matrix depicting the Jaccard similarities between the NMF metaprograms and the Astrocyte-like, Oligo-like and Stemness programs in Venteicher, *et al.*[252]. Three iterations of NMF have been computed, represented by the prefixes in the metaprograms names: OD_ and AS_ metaprograms belong to the first iteration,

containing ribosomal genes. N_ metaprograms correspond to the metaprograms retrieved prior removal of ribosomal genes. N2_ metaprograms refer to the metaprograms retrieved prior removal of ribosomal genes and exclusion of sample IDH_ACB_AD_540, which has a significantly higher proportion of RE. OD_RE_Curated metaprogram contains the genes present across all three RE metaprogram iterations. Jaccard similarities depict a high consensus between all homologue iterations of NMF metaprograms. **(E-G)** Permutation testing selection method based on the enrichment scores for the three different RE metaprograms retrieved including ribosomal genes (OD_RE), excluding them (N_RE) and excluding the sample with the highest proportion of RE cells, IDH_ACB_AD_540 (N2_RE) is used in the tumor cells of oligodendrogliomas **(E)** and astrocytomas **(F)** and paired astrocytomas **(G)**. Jaccard similarities between the selected cells for each of the NMF metaprograms is shown (bottom). Enrichment scores are shown on the left and the selected cells on the right. UMAP1, x-axis; UMAP2, y-axis." Reprinted from Blanco-Carmona, *et al.*[189]

The inferred proportions revealed the presence of the RE population in both TCGA and CGGA oligodendroglioma cohorts, with the exception of a single patient in CGGA. However, not all astrocytomas datasets exhibited detection of the RE population (Figure 37A-D). Nonetheless, these findings suggest that the RE population is present in oligodendrogliomas and partially in astrocytomas, to varying degrees.

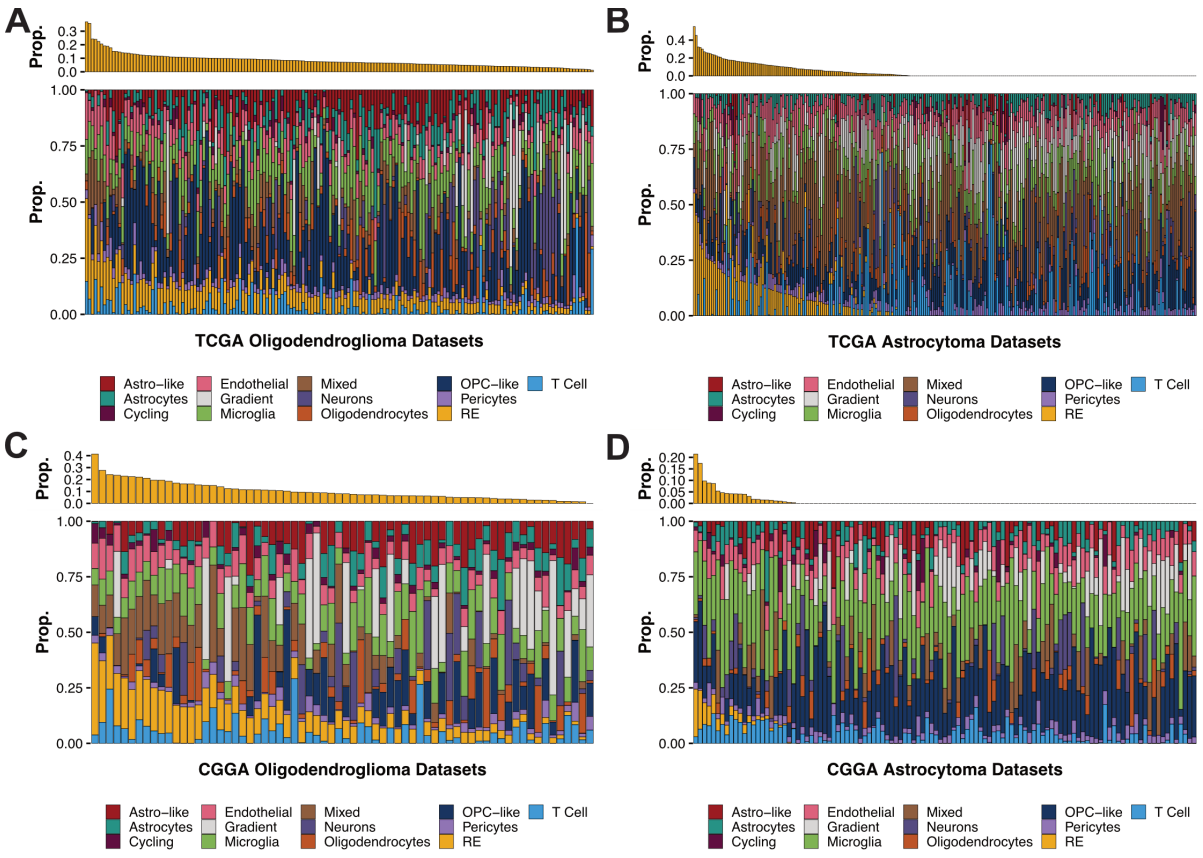


Figure 37: "Deconvolution of publicly available bulk glioma RNA-seq datasets (TCGA and CGGA). (A) SPOTlight results of TCGA (the Cancer Genome Atlas) OD IDH mutant glioma cohort (on top, proportion of RE, ordered. on the bottom, all proportions ordered also by descending RE). **(B)** SPOTlight results of TCGA AS IDH mutant glioma cohort (on top, proportion of RE, ordered. on the bottom, all proportions ordered also by descending RE) **(C)** SPOTlight results of CGGA OD (Chinese Glioma Genome Atlas) IDH mutant glioma cohort (on top, proportion of RE, ordered. on the bottom, all proportions ordered also by descending RE). **(D)** SPOTlight results of CGGA AS IDH mutant glioma cohort (on top, proportion of RE, ordered. on the bottom, all proportions ordered also by descending RE)." Reprinted from Blanco-Carmona, *et al.*[189]

Cell activities based on custom prior-knowledge networks

Furthermore, I sourced published fresh tissue single-cell datasets (scRNAseq) for both oligodendrogliomas[135] and astrocytomas[252]. In both cases, I generated a UMAP embedding and computed enrichment scores for the RE NMF metaprogram. In contrast with my frozen tissue datasets (snRNAseq), enrichment scores in these datasets exhibited a normal distribution (Shapiro test, p <= 0.05) (`Figure 38A-B, top`). To define the RE population, I selected a threshold of enrichment scores such that the probability of finding a cell with a score higher than the threshold was 5% (`Figure 38A-B, top right`). Once the RE population was defined, I aimed to assess its specificity to the gene marker sets retrieved from the primary datasets.

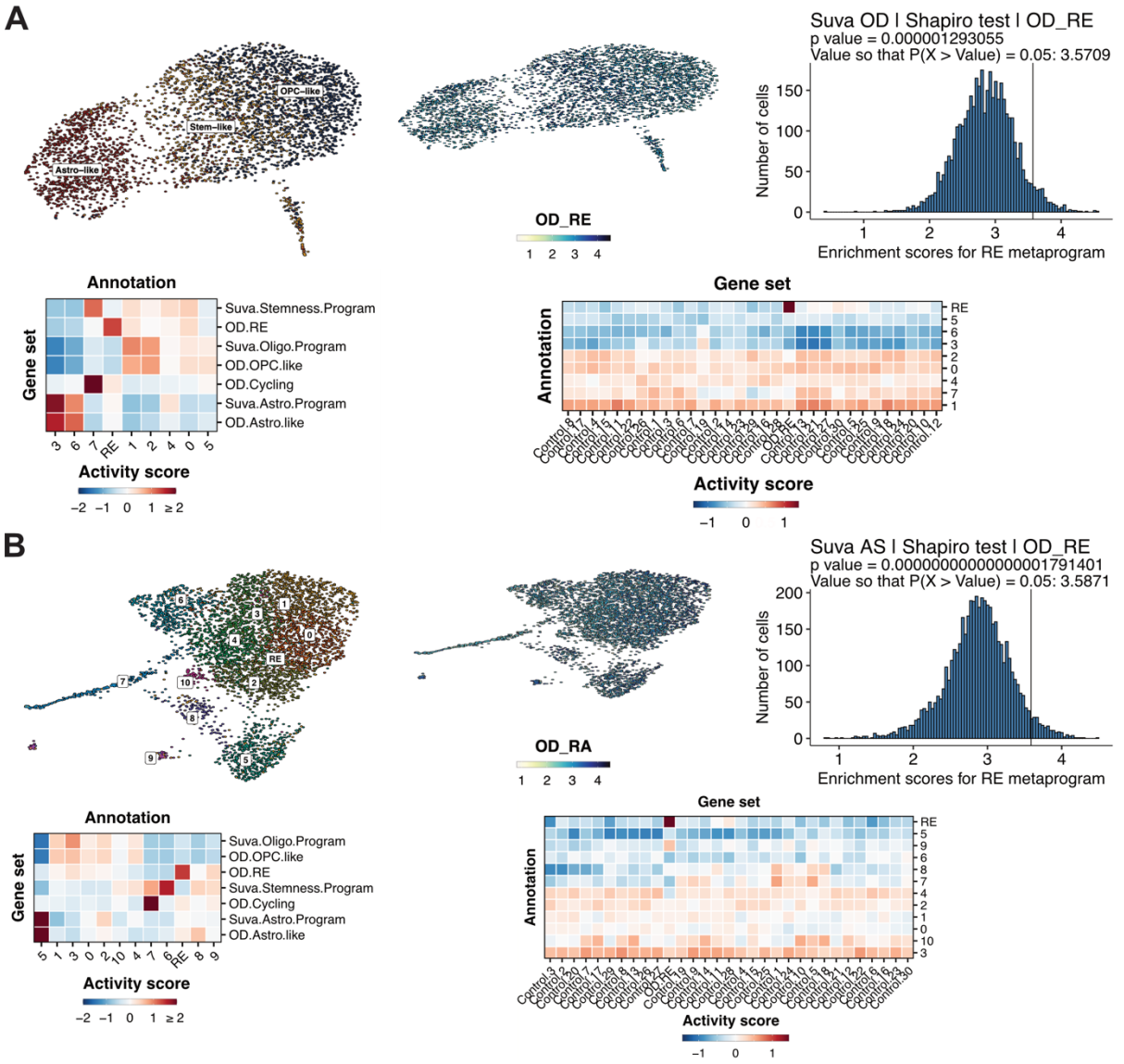


Figure 38: "The RE metaprogram in publicly available scRNA-seq datasets from IDH mutant gliomas, and the stemness score per patient in snRNA-sq dataset from oligodendrogliomas and astrocytomas. Datasets from publicly available oligodendrogliomas[135] **(A)** and astrocytomas[252] **(B)** are used to determine the presence of the

RE metaprogram. UMAP representation of the tumor cells is shown together with the enrichment scores for the RE metaprogram (top, left and middle). The distribution of enrichment scores is tested for normality using Shapiro test (top-right) and the value so that the probability of finding an enrichment score higher is 5% is used to select the RE population. The activity of the different cell clusters and the RE population towards the NMF metaprograms and the publicly available programs is computed by using decoupleR. Activity scores are scaled and centered and displayed grouped by cell population (bottom-left). To query the robustness of RE metaprogram towards the RE population, 50 different gene sets of equal size are generated by randomly selecting genes in the same bin of expression as the genes in the RE metaprogram. Activity scores are computed, scaled, and centered, and displayed grouped by cell population. UMAP1, x-axis; UMAP2, y- axis." Reprinted from Blanco-Carmona*, et al.*[189]

To accomplish this, activity scores were computed based on a custom prior knowledge network generated using the NMF metaprograms and the tumor programs from Venteicher, *et al.*[252] The results unveiled specific clusters in both datasets that were activated in diverse gene sets, including RE metaprogram (`Figure 38A-B, bottom-left`). Moreover, to ensure the validity of these results and rule out the possibility of observing them due to chance, the robustness of the analysis was assessed by generating 30 control gene sets containing genes whose expression levels were similar to those in the NMF RE metaprogram.

Subsequently, activity scores were computed based on a custom prior knowledge network that contained the RE metaprogram and the control gene sets. The findings revealed that, for both datasets, the defined RE population exhibited the highest activation in the NMF RE metaprogram, contrasting with the other control gene sets (`Figure 38A-B, bottom right`). Collectively, these results demonstrate that the RE population can indeed be identified in external datasets, further underlining its potential biological relevance.

Reference mapping

Finally, I aimed to contextualize the RE population within the context of development and glioblastoma lineages. The overarching hypothesis was that since the RE and cycling cells were annotated as secondary phenotypes, the primary phenotype of either OPC-like and astro-like should still be present. Therefore, performing reference mapping onto datasets containing either astrocytes/oligodendrocytes or astro-like/OPC-like cells would result in the RE cells mapping onto these populations.

To achieve this, `Azimuth` was used to perform reference mapping of the oligodendroglioma and astrocytoma datasets against a single-cell atlas of glioblastoma (GBmap)[255] and a human fetal development atlas[256]. In GBmap, RE cells primarily mapped onto OPC-like cells, while in

the fetal brain atlas, RE cells mapped equally to oligodendrocytes and astrocytes (Figure 39A-D). These findings suggest a similarity between RE cells and an oligodendrocyte program.

Immunohistochemistry

To validate the previous findings, IHC staining was conducted by J.T., M.C. and P.L.P. in Blanco-Carmona, *et al.*[189] in a validation cohort comprising 37 patient samples (oligodendroglioma = 22, astrocytoma = 15). The results revealed a higher degree of staining in oligodendrogliomas compared to astrocytomas (Figure 40A-B).

Additionally, a statistically significant increase in EEF2 levels with tumor grade was observed in astrocytomas, while a statistically significant increase of EEF1A1 levels in oligodendrogliomas compared to astrocytomas was also noted (Figure 40C). Notably, IHC staining revealed spatial heterogeneity, particularly pronounced in astrocytomas compared to oligodendrogliomas (Figure 40D-E).

Investigating the stemness profile of tumor cell populations through diffusion maps

Once the RE population was established as a biological finding in my datasets, I set out to determine its role in the tumor biology of IDH-mutant gliomas. To achieve this, I first investigated whether this population harbored a stem-like phenotype. Diffusion maps based on a universe of genes containing the programs reported in Venteicher, *et al.*[252] were computed. The results showcased that diffusion component one (DC_1) distinctly separated OPC-like and astro-like cells, while diffusion component two (DC_2) separated RE cells from cycling cells (Figure 41).

Hence, I hypothesized that DC_2 would comprise a potential stemness phenotype. To assess this, I computed enrichment scores for all three programs from Venteicher, *et al.*[252] and scaled and centered the scores for clarity. The scores were visualized as a heatmap, with the X-axis representing cells ordered by their position along DC_2, and the Y-axis displaying the enrichment scores for each program. This analysis revealed the highest enrichment of the Stemness program at the end of the DC_2, where RE cells were located, suggesting that RE cells harbor a stemness phenotype (Figure 41B).

Pathway and TF activity profiles of tumor cell populations

Finally, to provide a more comprehensive understanding of the tumor populations present in the primary datasets, I conducted functional enrichment analysis at both the pathway and transcription factor level using `decoupleR`[85], together with published prior knowledge networks[80,81]. The resulting activities revealed a depletion of p53 signature in the cycling cells (`Figure 42A`). More interestingly, there was an enrichment of FOXM1 regulon in cycling cells but not in RE cells (`Figure 42B`). This transcription factor is known to be associated with proliferation[257]. In light of these results, the RE tumor cell population seems to be a non-cycling stem-like tumor population.

Querying accessibility profiles of tumor cell populations

The following analysis was conducted by I.H. in Blanco-Carmona*, et al.*[189] Next, I aimed to compare the chromatin accessibility profiles of the different tumor populations in oligodendrogliomas and astrocytomas. To accomplish this, I.H. was granted access to the snATACseq cohort of the IDH-mutant glioma primary datasets (oligodendroglioma = 4, astrocytoma = 5) (`Figure 26A`). She performed label transfer using the tumor annotation I generated from the transcriptomic data of the primary datasets, excluding the gradient population. I argued that including the gradient population, that encompasses a spectrum of cells amidst differentiation into astro-like or OPC-like, would be detrimental to the efficiency of the label transfer.

The results revealed that all populations (both TME and tumor-based) could be retrieved in the snATACseq data of both oligodendroglioma and astrocytoma datasets (`Figure 43A-C`). Notably, the RE population was consistently identified across all patients, defining it as a separate entity from the other tumor populations based on chromatin accessibility. Correlation analysis of the highly variable peaks showed a positive correlation between OPC-like and RE tumor cell populations (`Figure 44A`). This was also reflected at the level of transcription factor enrichment analysis, demonstrating similarities between RE and OPC-like tumor cell populations across IDH-mutant glioma subtypes (`Figure 44B-C`).

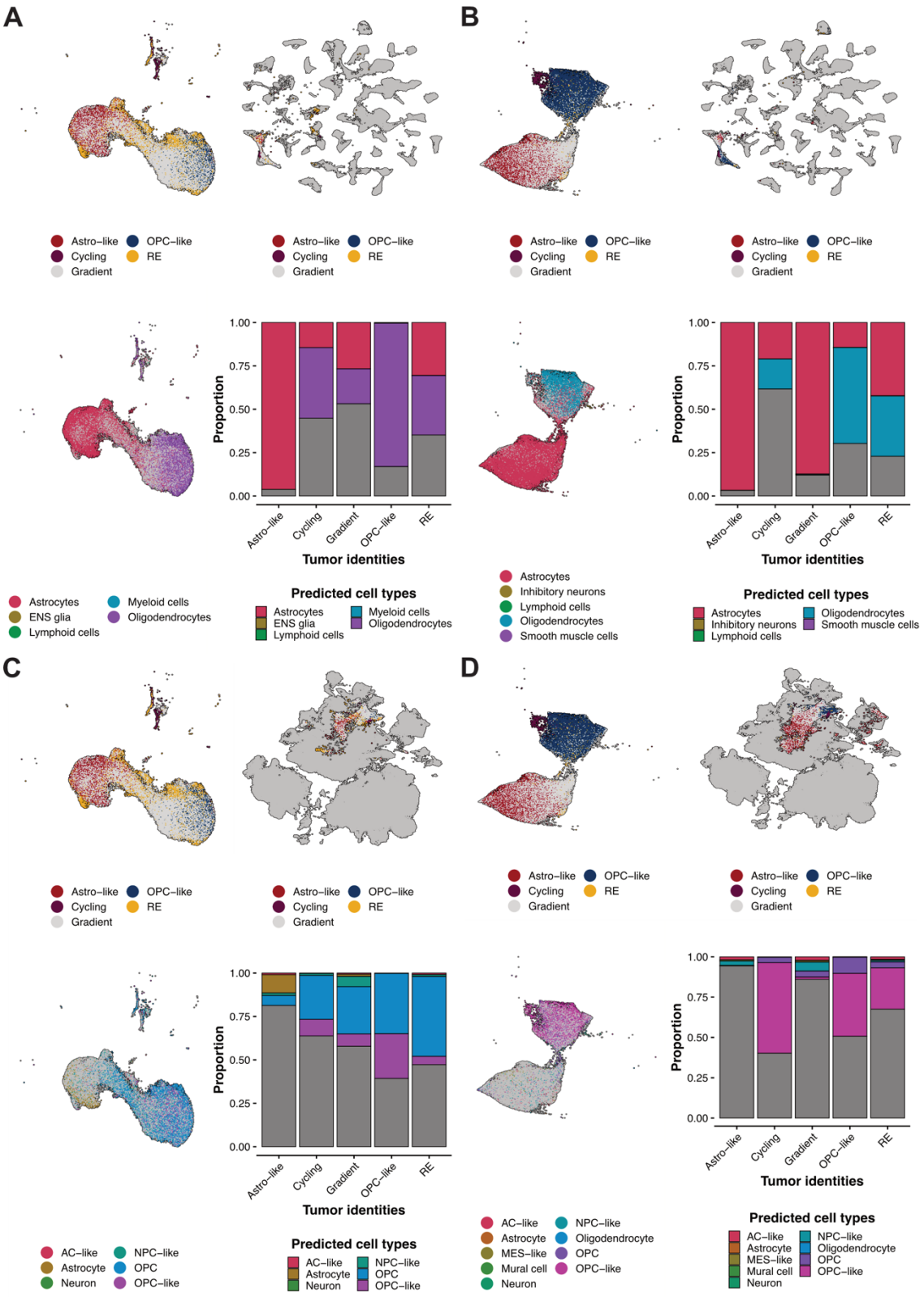


Figure 39: "Reference mapping using Azimuth. (A, B) Tumor subsets for primary oligodendrogliomas **(A)** and primary astrocytomas **(B)** are mapped onto the fetus reference from Azimuth. **(C, D)** Tumor subsets for primary oligodendrogliomas **(C)** and primary astrocytomas (D) are mapped onto the GBmap reference. For each variation, original UMAP embedding is shown on the top-left, cells mapped onto the reference UMAP on the top-right, cells re-labelled based on the reference mapping on the bottom left and a bar plot of the proportions of the new annotation per each original tumor entity on the bottom-right. UMAP1, x-axis; UMAP2, y-axis." Reprinted from Blanco-Carmona, *et al.*[189]

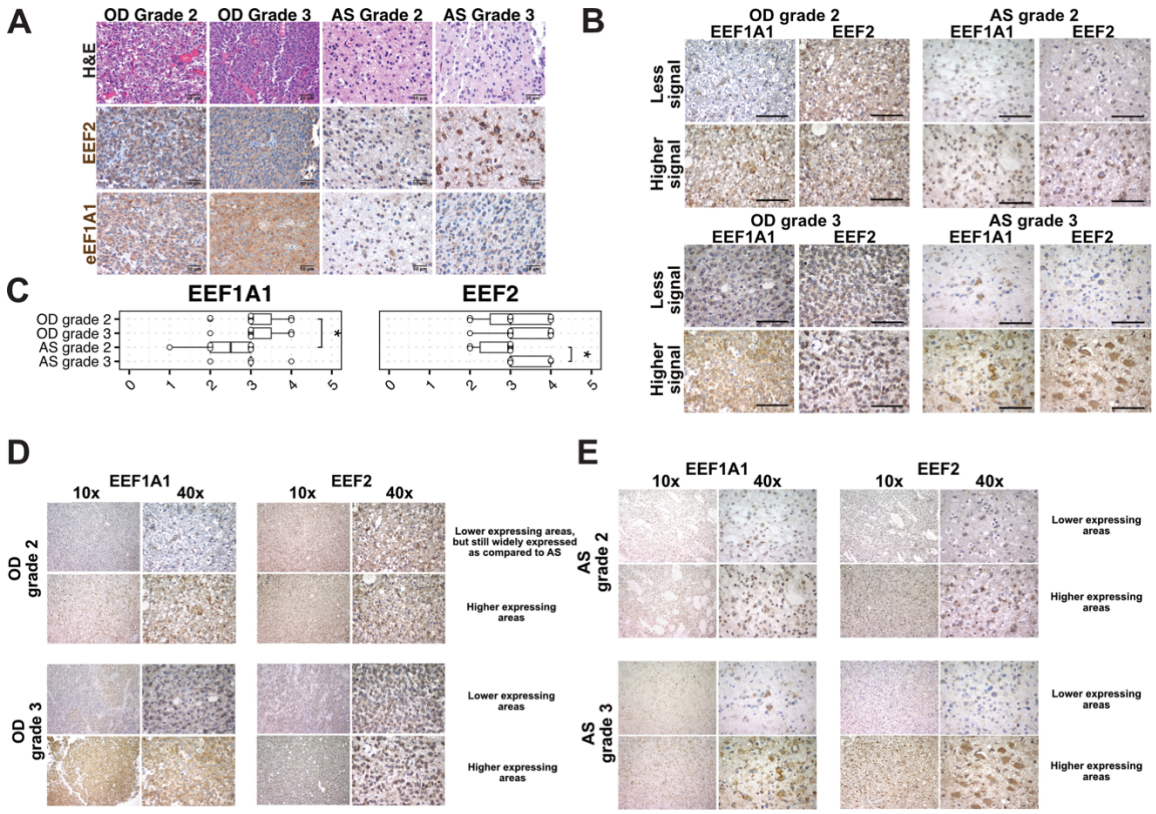


Figure 40: "Representative IHC staining for EEF1A1 and eEF2 in IDH mutant gliomas. (A) Representative IHC staining for H&E and two RE population markers, EEF2 and EEF1A1, for oligodendroglioma and astrocytoma samples, separated by grade. Oligodendroglioma (OD) n = 22 (11 grade 2, 11 grade 3); astrocytoma (AS) n = 15 (10 grade 2, 5 grade 3). Scale bars represent 50 µm. **(B)** Spatial distribution of EEF2 and EEF1A1 (40x). Scale bars represent 50 µm. **(C)** Plots showing semi-quantitative histological scores (0 = 0%, 1 = 0%–5%, 2 = 6%–29%, 3 = 30%–69%, and 4 >70% positive staining) for EEF2 (top) and EEF1A1 (bottom). Top: a qualitative increase in EEF2 staining in grade 3 compared with grade 2 AS tumors. Bottom: a qualitative increase in EEF1A1 staining in grade 2 OD tumors compared with grade 2 AS tumors. Wilcoxon rank-sum test was performed to test for significance, *$p < 0.05$. **(D, E)** Representative IHC staining for EEF1A1 and eEF2 population marker genes in grade 2 (top) and grade 3 (bottom) oligodendrogliomas **(D)** and astrocytomas **(E)**. Images are shown as 10x and 40x to show the spatial distribution of EEF1A1 and EEF2 in both tumor types. OD, oligodendroglioma; AS, astrocytoma." Panels A, B, D and E were generated by J.T., M.C. and P.L.P. in Blanco-Carmona, *et al*.[189] Reprinted from Blanco-Carmona, *et al.*[189]

Characterizing the immune microenvironment of IDH-mutant gliomas

After examining the cell type composition of the primary datasets, I noticed a consistent presence of microglia cells across patients. Typically, clusters designated as microglia in single-cell datasets can be further subdivided into different cell fractions belonging to tumor associated macrophages (TAMs). Previous studies have indicated that IDH-mutant gliomas harbor TAMs, including border-associated macrophages (BAMs), bone-marrow-derived macrophages (BMD) and tissue-resident microglia, as their most abundant TME cell type[258]. Based on this knowledge, I decided to further investigate into the inherent heterogeneity of the microglia populations in the primary samples.

For this, I received assistance from I.H., that conducted the bioinformatics analyses, and J.N.S., that provided the biological expertise to accurately characterize the different TAM subpopulations. My role involved the discussion of the results and the generation of the figures.

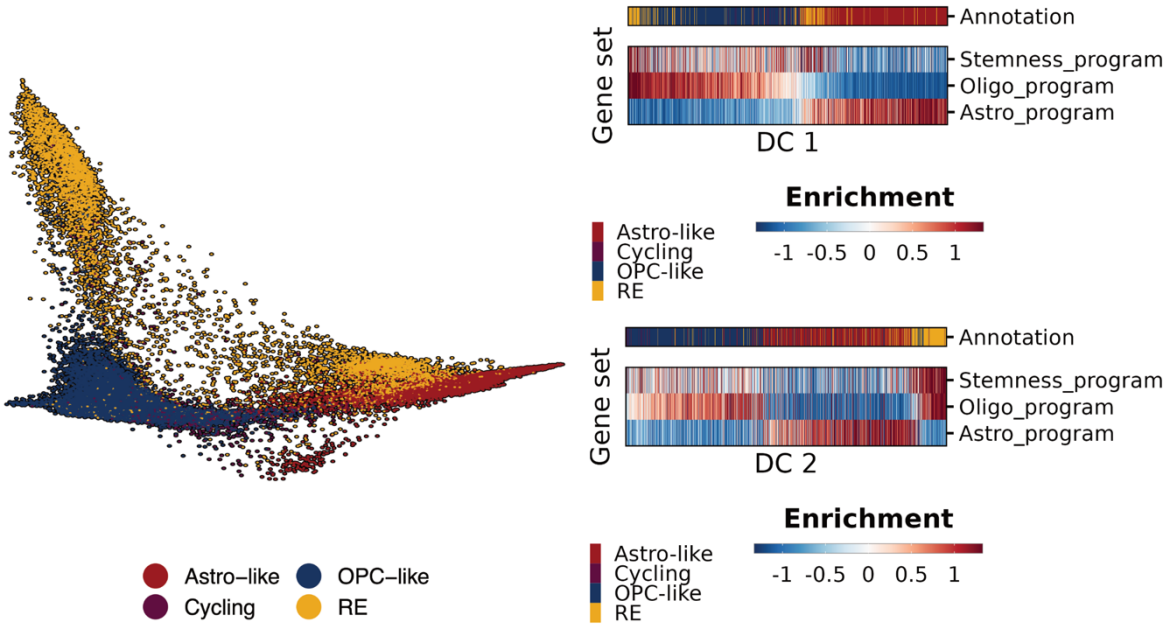


Figure 41: "Stemness profile in RE population. Diffusion map visualization of tumor cells in OD and AS tumors, based on publicly available astro-like, oligo-like, and stemness program markers.[252] Right: scaled and centered enrichment scores for NMF metaprograms. Cells are ordered along diffusion components 1 (DC_1) and 2 (DC_2); gradient cells were excluded for clarity." Reprinted from Blanco-Carmona, *et al.*[189]

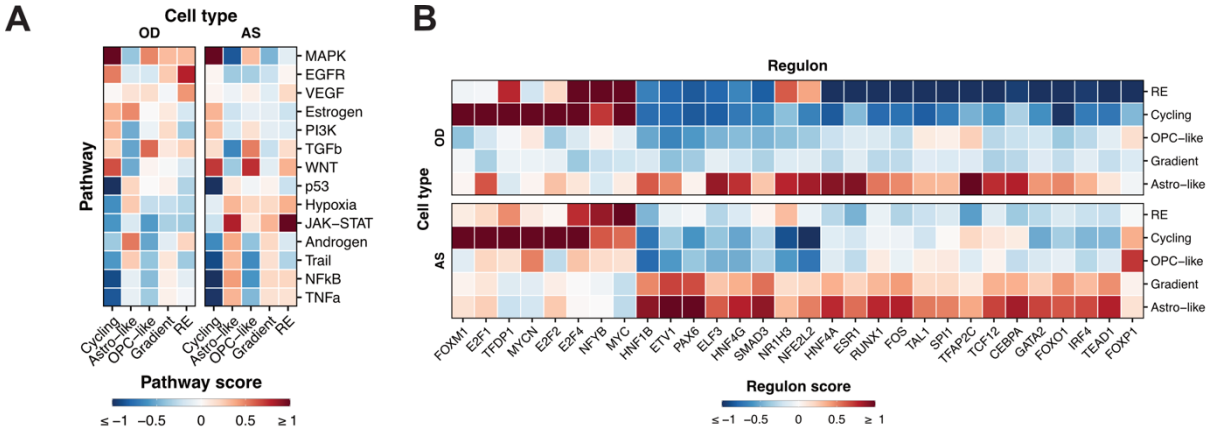


Figure 42: "Pathway and TF enrichment analysis. Scaled and centered pathway **(A)** and regulon **(B)** activity scores (normalized weighted) for OD and AS tumor cell populations." Reprinted from Blanco-Carmona, *et al.*[189]

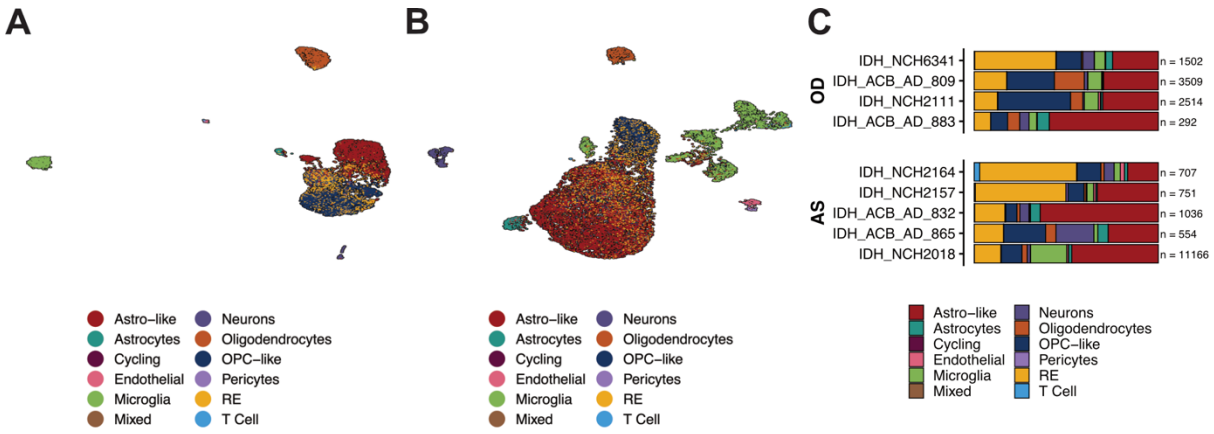


Figure 43: "Cluster annotation in snATACseq datasets. (A and B) UMAP embedding of OD **(A)** and AS **(B)** snATAC-seq data (OD n = 4, AS n = 5) colored by labels transferred from snRNA-seq datasets. UMAP1, x axis; UMAP2, y axis. **(C)** Bar plot showing proportion of cell types in snATAC-seq data." Reprinted from Blanco-Carmona, *et al.*[189]

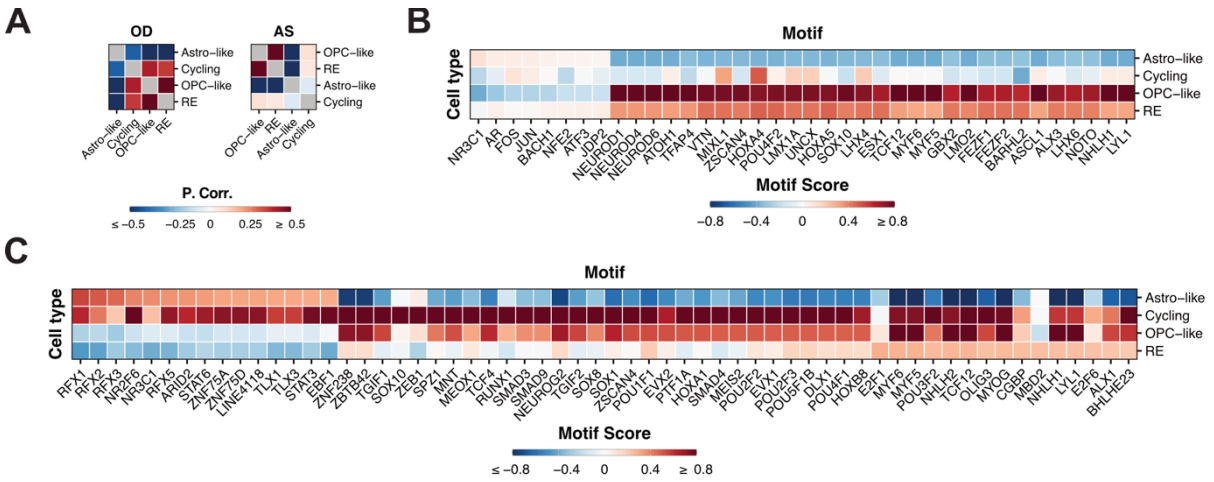


Figure 44: "Motif and TF enrichment analysis. (A) Pearson correlation of highly variable peaks from snATAC-seq data in OD and AS tumors. **(B and C)** Heatmap showing the top significantly enriched transcription factor motifs in AS **(B)** and OD tumors **(C)**. Scores are scaled and centered for clarity." Reprinted from Blanco-Carmona, *et al.*[189]

Supervised annotation of TAM subpopulations

First, microglia cells from both oligodendroglioma and astrocytoma datasets were aggregated, normalized and integrated. J.N.S. curated gene marker sets from the literature[259,260] that were utilized to label different TAM subpopulations (Figure 45A-B). A total of 10 subpopulations were identified, including BAMs, BMD Anti-Inflammatory TAMs, Microglia (Mg) Activated, Mg Homeostatic, Mg Interferon gamma (INFγ) TAM, Mg Inflammatory ICAM1+, Mg Inflammatory TAMs, Mg, Phagocytic, Mg Resident-like TAMs and Mg Stressed TAMs (Figure 45C). Comparisons across IDH-mutant glioma subtypes revealed significant shifts in TAM composition. Specifically, Mg Homeostatic cells were more prevalent in astrocytomas (p = 0.01,

Wilcoxon test), while BAM (p = 0.048, Wilcoxon test) and BMD anti-inflammatory TAMs (p = 0.048, Wilcoxon test), and Mg Phagocytic TAM (p = 0.012, Wilcoxon test) were more prevalent in oligodendrogliomas (Figure 45D). These findings provided initial confirmation of differential TAM compositions across IDH-mutant glioma subtypes.

Pro- and anti-inflammatory TAM composition across IDH-mutant gliomas subtypes

Given the differential prevalence of TAM subpopulations associated with pro- and anti-inflammatory phenotypes between IDH-mutant glioma subtypes, J.N.S. and I sought to determine whether these subpopulations were also differentially enriched in literature-based gene marker sets for such phenotypes[261,262]. To address this, enrichment scores for gene marker sets for pro- and inti-inflammatory phenotypes were calculated across TAM subpopulations. These scores were visualized as a function of the density of the neighboring cells that were also enriched for these markers. The results revealed no significant differences for either pro- or anti-inflammatory across IDH-mutant glioma subtypes (Figure 46A-B).

The crosstalk between TAM subpopulations and tumor cell types across IDH-mutant gliomas subtypes

In collaboration with J.N.S., M.E.B. and I.H., the interplay between the different TAM subpopulations and the tumor compartment was characterized. Ligand-receptor interactions were inferred using `liana`[90]. Subsequently, J.N.S. filtered the significant interactions to those with a biological relevance to the project.

In oligodendrogliomas, biologically relevant interactions comprised BMP pathway proteins originating from OPC-like tumor cell population and PGF-NRP1/2 from Astro-like tumor cell population towards TAMs. Other widespread interactions such as SIRPA-CD47 pointed towards a reduction of the phagocytic functions of innate immune cells[263]. Similarly, WNT5A-FZD3 and WNT5A-PTPRK were significant across all TAMs. Interestingly, CSF1-CSF1R was particularly prominent in astro-like tumor population of astrocytomas (Figure 47A). Further investigation into the ligand-receptor interactions from TAMs to tumor subpopulations revealed TNF in oligodendrogliomas and DLL1 in astrocytomas (Figure 47B).

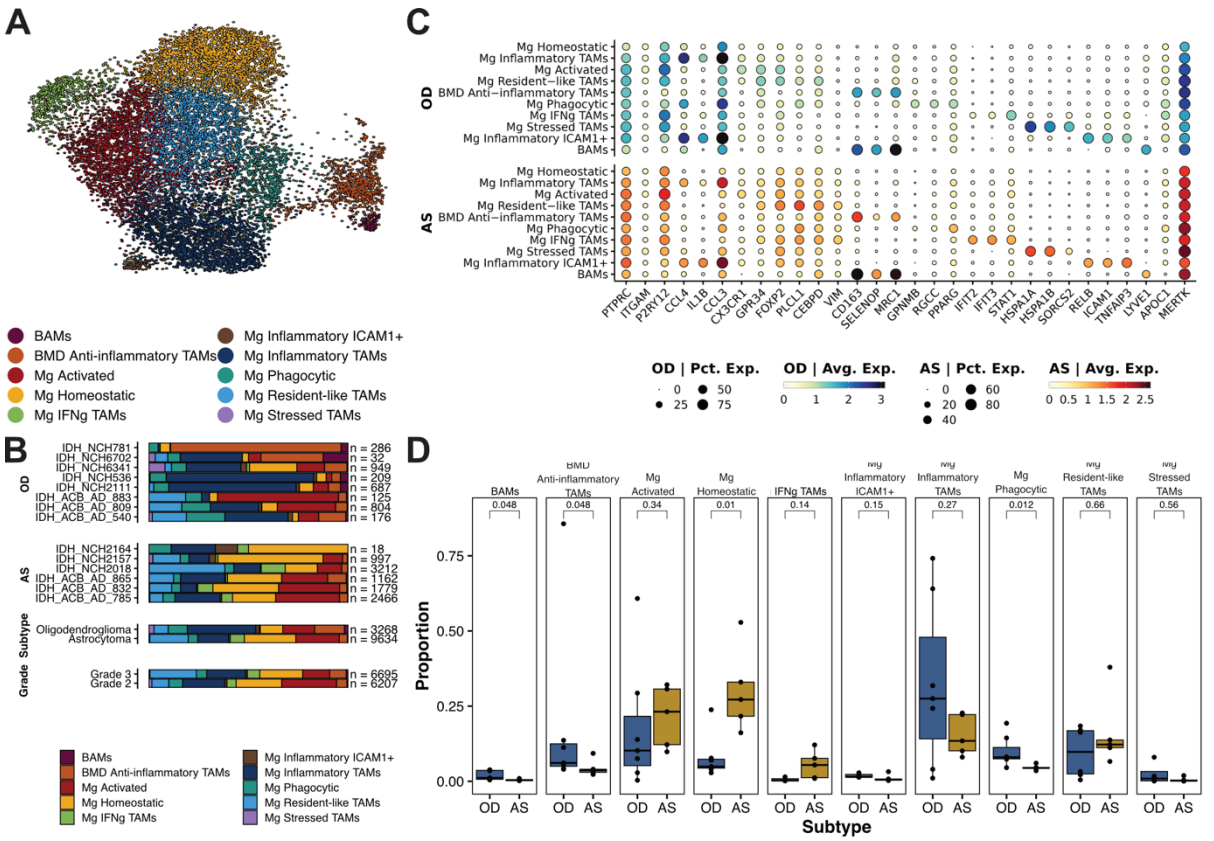


Figure 45: "Microglia TAM subpopulation characterization. (A) UMAP embedding of integrated microglia from snRNA-seq data of primary OD and AS tumors, colored by assigned TAM subpopulations. UMAP1, x axis; UMAP2, y axis. **(B)** Bar plots indicating TAM subpopulation proportions, separated by tumor subtype and grade. **(C)** Dot plot showing three key marker genes for the TAM subpopulations in oligodendrogliomas (top) and astrocytomas (bottom). OD, oligodendroglioma; AS, astrocytoma. **(D)** Boxplots showing TAM subpopulation proportions by tumor type. OD, oligodendroglioma; AS, astrocytoma. Wilcoxon rank-sum test was performed to test for significance." Reprinted from Blanco-Carmona*, et al.*[189]

Investigating the effect of tumor grade and recurrence in IDH-mutant gliomas

I next sought to understand the biology of IDH-mutant gliomas in the context of tumor grade and tumor recurrence. For this purpose, a dataset comprising primary and recurrent astrocytomas (n = 12) was generated (Figure 48A). The analysis of this dataset was done collaboratively: I conducted tumor cell type annotation and I.H. and J.N.S. analyzed and characterized TAM subpopulations.

Similar to the primary samples, TME cell populations were identified based on enrichment in gene marker sets from PanglaoDB[191]. Tumor cell populations were identified through permutation testing utilizing the NMF metaprograms retrieved from the primary samples (Figure 48B). While the RE population was identified across tumor pairs, neither the RE, astro-like nor OPC-like tumor cell populations exhibited significant changes at recurrence.

However, there was a noticeable trend of decreasing astro-like and OPC-like populations and increasing cycling and gradient proportions with tumor grade (Figure 48C). To further validate these findings, IHC staining for EEF2 and EEF1A1 was conducted by J.T., M.C. and P.L.P. in Blanco-Carmona*, et al.*[189] on a cohort of 12 patients encompassing both primary and recurrent tumors (oligodendroglioma = 6, astrocytoma = 6) (Figure 48D). The results revealed elevated levels of both EEF2 and EEF1A1 in recurrent astrocytomas compared to primary tissue. Interestingly, the levels of these proteins remained consistently high in both primary and recurrent oligodendrogliomas (Figure 48E).

Following this, microglia cells were isolated from the dataset and reclustered. Using the same marker sets as in the primary samples, J.N.S. identified the same TAM subpopulations, with the exception of BAMs (Figure 49A-B). This observation was consistent with BAMs being exclusively identified in oligodendroglioma cells in the primary datasets. Upon analyzing cell type proportions across pairs, it was noted that pairs 2 and 5 had a very low number of cells in one of the pairs, resulting in their exclusion from further comparisons (Figure 49C). Examining the shifts in proportions at recurrence revealed no significant increase or decrease of TAM subpopulations (paired Wilcoxon test), indicating that TAMs remain consistent after tumor recurrence. However, several trends were observed, which may become statistically significant with a larger sample size. These trends included decreased Mg Homeostatic and increased Mg Hypoxic, BMD anti-inflammatory TAMs and Mg IFNγ at recurrence.

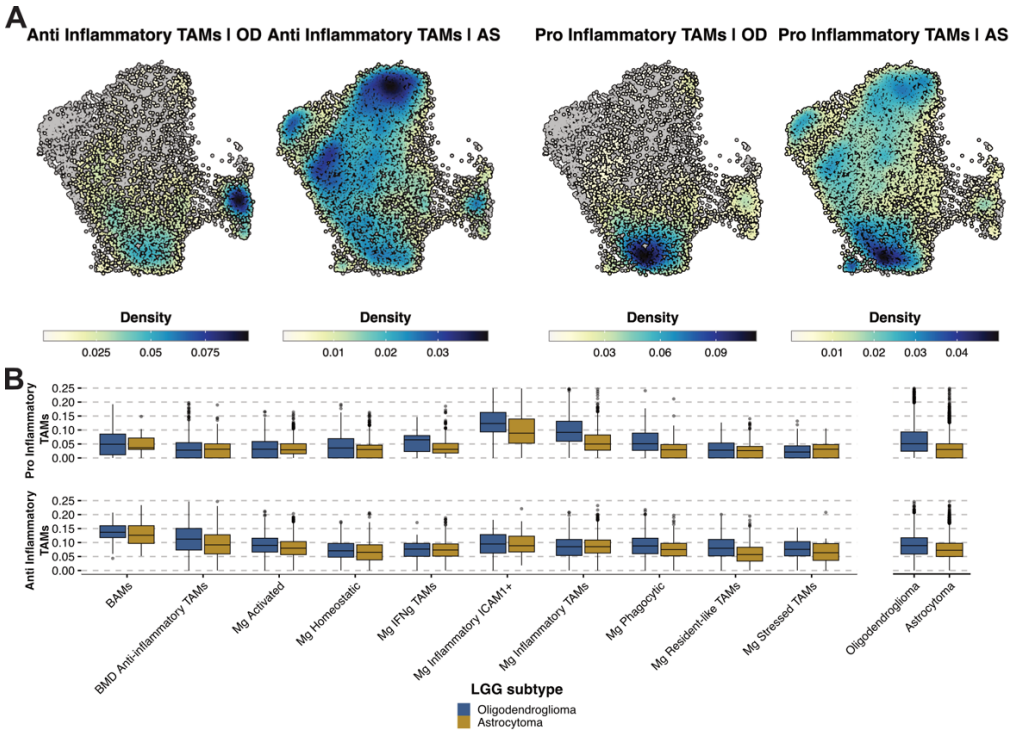


Figure 46: "Defining Microglia pro- and anti-inflammatory. (A) Density gradient of pro-inflammatory and anti-inflammatory signatures in TAMs of astrocytomas and oligodendrogliomas. OD, oligodendroglioma; AS, astrocytoma. **(B)** Boxplots showing pro-inflammatory (top) and anti-inflammatory (bottom) scores per TAM subpopulation in oligodendrogliomas and astrocytomas." Reprinted from Blanco-Carmona, *et al.*[189]

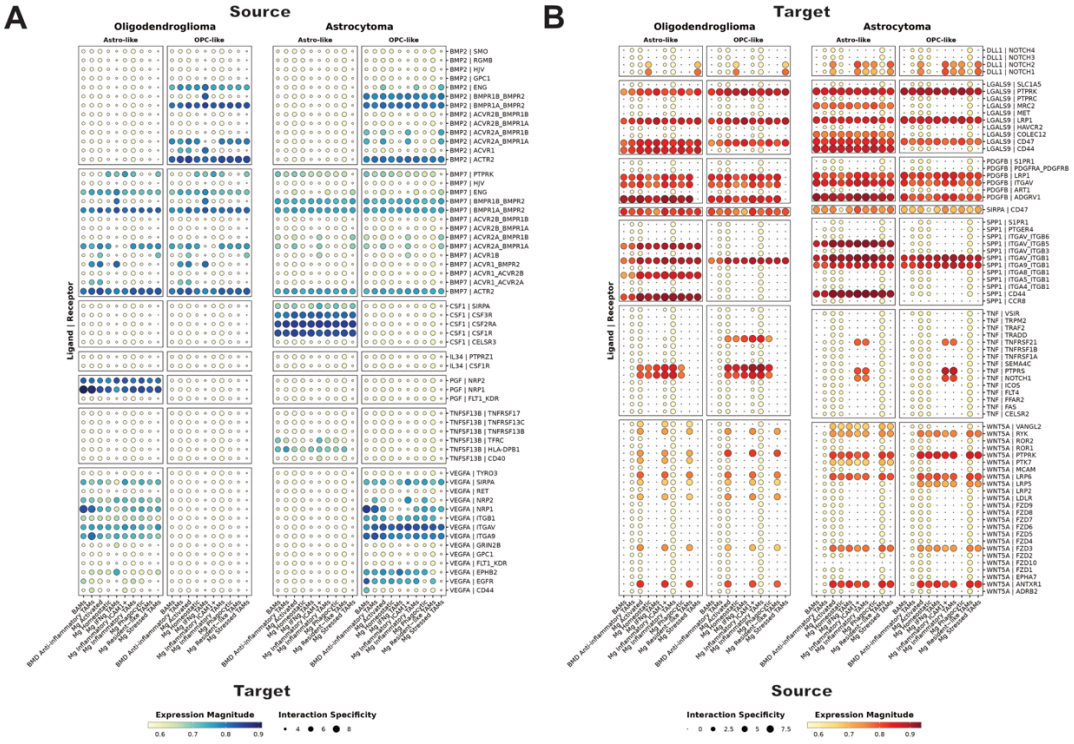


Figure 47: "LR interactions. (A and B) Selected statistically significant receptor-ligand interactions between OPC-like and astro-like tumor populations (source) and the TAM subpopulations (target) **(A)** and TAM subpopulations (source) and OPC-like and astro-like tumor cells (target) **(B)**. Dot size represents significance (adjusted p values); dot color reflects expression magnitude (means of average expression level of the interacting pair of genes)." Adapted from Blanco-Carmona, *et al.*[189]

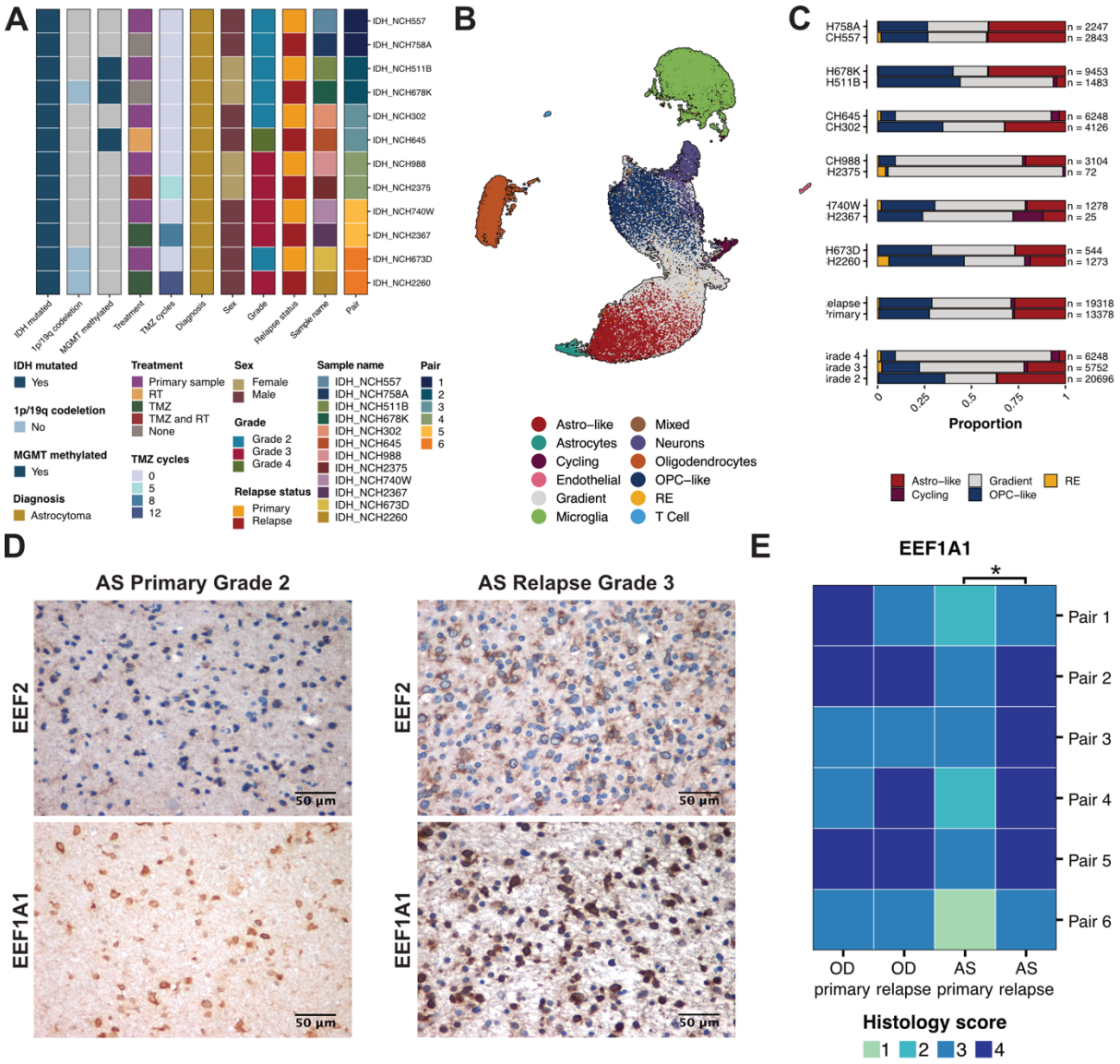


Figure 48: "snRNA-seq of paired primary and recurrent AS cohorts indicates shifts in tumor populations associated with grade. (A) Clinical and molecular characteristics of 6 paired primary and recurrent IDH-mutant AS cohorts for snRNA-seq. **(B)** UMAP embedding of integrated snRNA-seq data from paired AS tumors, colored by assigned cell types. UMAP1, x axis; UMAP2, y axis. **(C)** Bar plots showing tumor population proportions, separated by pairs, relapse status, and grade. **(D)** Representative IHC staining for RE markers EEF2 and EEF1A1 in paired primary (grade 2) and relapse (grade 3) AS tumors. Images are captured at 40x original magnification (scale bar, 50 µm). OD: oligodendroglioma. AS: Astrocytoma. **(E)** Heatmap showing semi-quantitative EEF1A1 histological scores (0 = 0%, 1 = 0%–5%, 2 = 6%–29%, 3 = 30%–69%, 4 >70% positive staining) for six pairs of primary and recurrent OD and AS tumors. Wilcoxon signed-rank test for paired samples was performed to test for significance, *p < 0.05. OD: oligodendroglioma. AS: Astrocytoma." Panel D was generated by J.T., M.C. and P.L.P. in *Blanco-Carmona, et al.*[189] Reprinted from Blanco-Carmona*, et al.*[189]

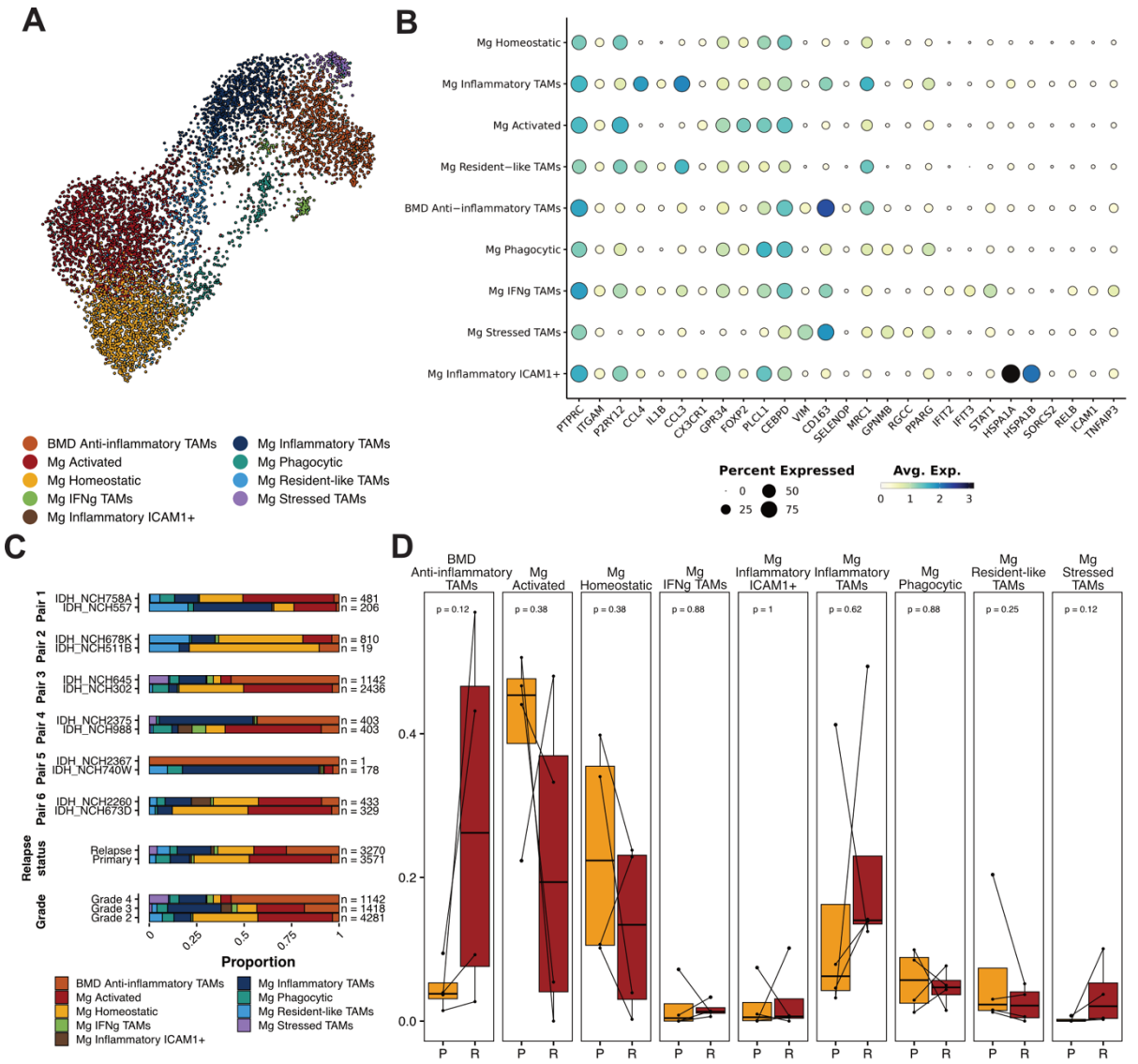


Figure 49: "TAM subpopulations in paired datasets. (A) UMAP embedding of integrated microglia population of paired primary and recurrent AS cohorts, colored by assigned TAM subpopulations. UMAP1, x axis; UMAP2, y axis. **(B)** Dot plot showing three key marker genes for the TAM subpopulations in the paired datasets. **(C)** Bar plots showing TAM subpopulation proportions separated by pairs, relapse status, and grade. **(D)** Boxplots showing TAM subpopulation proportions separated by relapse status. P, primary; R, relapse." Reprinted from Blanco-Carmona, *et al.*[189]

Discussion

Over recent years, several research projects have explored the heterogeneity of IDH-mutant gliomas[135,252]. However, these studies were conducted on independent cohorts focusing exclusively on either oligodendroglioma or astrocytoma tumor cases. Consequently, a comprehensive and comparative study between both IDH-mutant gliomas subtypes remained unfulfilled. In this project, I aimed to address this gap by systematically analyzing snRNAseq and snATACseq datasets from two independent cohorts comprising of primary IDH-mutant gliomas (oligodendroglioma and astrocytoma) and primary-recurrent astrocytoma pairs. Expanding on the obtained results and to elucidate the composition of tumor and TAM cell populations, single-cell data was complemented with IHC staining on a separate cohort of primary and recurrent oligodendrogliomas and astrocytomas. This approach enabled comparisons across IDH-mutant gliomas subtypes and WHO tumor grades.

Following a throughout analysis of the tumor cell populations across both oligodendrogliomas and astrocytomas, in line with previous research[135,214], I observed that tumor cells primarily segregated into cycling cells, forming a distinct cluster after cell integration, and tumor cells exhibiting expression profiles resembling that or healthy astrocytes and oligodendrocyte precursor cells (referred to as astro-like and OPC-like). These populations were located in opposite ends of the UMAP embedding, effectively forming a gradient of differentiation of the tumor cells from more undifferentiated towards either of astro-like or OPC-like tumor cell populations. This hypothesis was further supported by a permutation testing approach designed to statistically select cells enriched for either tumor cell population. The cells located at the two ends of the gradient were identified as the most significantly enriched, categorizing the remaining cells as undifferentiated (termed “gradient”). While the gradient population represented cells that did not surpass the significance thresholds applied to the permutation testing analysis, further analysis on this tumor population alone revealed a gradient-like enrichment for astro-like and OPC-like marker genes. This finding supports previous knowledge that tumor cells in IDH-mutant gliomas differentiate into astrocytic and oligodendrocytic lineages.[135,214]

Through non-negative matrix factorization (NMF), in addition to the aforementioned tumor cell populations, I identified a novel tumor cell population in both oligodendroglioma and

astrocytomas exhibiting stem-like properties coupled with high expression of ribosomal genes and elongation factors, but lacking expression of genes associated with cell proliferation, termed ribosomal enriched (RE). The presence of this population was confirmed by immunostaining of protein markers EEF1A1 and EEF2 in independent IDH-mutant gliomas cohorts, revealing a broader distribution of both markers despite the relatively low frequency of RE population in the snRNAseq cohort. The discrepancy between the snRNAseq and immunostaining results may stem from the limited number of cells expressing the entirety of the NMF metaprogram, as opposed to the two protein markers assessed through immunostaining, along with potential spatial features within the tumors. Nonetheless, the association between ribosomal composition and spatial heterogeneity has been recently noted in glioblastoma research[264], and the function of ribosome biogenesis in IDH-mutant gliomas remains to be fully explored. Gene set enrichment analysis across tumor cell populations unveiled high activity of the MYC/MYCN regulon in the RE population, indicative of enhanced biosynthetic properties due to the association of MYC and EEF1A1[265]. Jointly, snRNAseq and immunostaining experiments suggested a higher prevalence of the RE population in oligodendrogliomas compared to astrocytomas, a finding validated through deconvolution of IDH-mutant gliomas datasets from TCGA and CGGA, where almost the entire oligodendroglioma cohort exhibited expression of the RE population. Hence, therapy approaches promoting differentiation targeting the RE population may be a potential avenue of research[266,267], as during tumor cell differentiation biosynthetic capabilities are reduced[268].

Identifying tumor progression drivers in IDH-mutant gliomas through comparative analysis between primary and recurrent tumors remains an ongoing challenge, with consortiums like the Glioma Longitudinal Analysis (GLASS) currently investigating glioma dynamics[269,270]. In this project, I aimed to gain insights into glioma progression through the comparative analysis of primary and recurrent astrocytomas. The results revealed that tumor cell populations remained consistent between primary and recurrent tumor pairs, regardless of treatment (radiotherapy, temozolomide or none), aligning with recent studies indicating stable tumor cell populations at recurrence[271]. This underscores these tumor cell populations as potential therapeutic targets. Furthermore, functional studies on these populations paired with master regulator protein analysis[272] could unveil essential mechanisms maintaining these tumor cell populations at recurrence.

Despite the fact that IDH-mutant gliomas exhibit tumor cell populations the expression profile of which resemble that of OPCs, their exact role in the biology of IDH-mutant gliomas is yet to be fully characterized. Thus, understanding the bidirectional interplay between tumor cell populations and tumor microenvironment is crucial for developing effective therapeutical strategies. In this project, I aimed to elucidate the crosstalk between tumor cells and the microenvironment by inferring ligand-receptor interactions between astro-like and OPC-like tumor cell populations and the distinct TAMs compartments characterized in the primary IDH-mutant glioma cohort. The results suggested that the astro-like tumor cell population in astrocytomas promotes a pro-inflammatory microenvironment primarily through IFNγ signaling. This includes the induction of phosphorylation of STAT1, highlighting the pivotal role of IFNγ in astrocytoma biology. Consequently, therapy approaches targeting these tumor cell populations may lead to alterations in the tumor microenvironment.

Although several studies have focused on understanding the nature and composition of TAMs in IDH-mutant gliomas[244,273], no comparative analysis focusing on whether TAM composition varies between oligodendrogliomas and astrocytomas is available. To address this gap, in this project, I sought to characterize potential changes in TAM composition across IDH-mutant gliomas subtypes. The results revealed an overall higher proportion of Mg-derived TAMs compared to BMD-derived TAMs, consistent with previous studies[273]. Furthermore, a differential TAM composition between oligodendrogliomas and astrocytomas was observed. Paired with immunostaining, the results highlighted an increased proportion of inflammatory TAMs expressing p-STAT1 in astrocytomas, a regulator of TAM inflammatory response also associated with neuronal damage[274,275].

Given the differential tumor-specific molecular alterations between IDH-mutant gliomas subtypes, further research is needed to determine their contributions to myeloid diversity. Among the inferred ligand-receptor interactions, the interaction between PGF and NRP1 was particularly interesting. PGF, a member of the VEGF family, is involved in cancer pathogenesis and immune modulation[276], playing a role in medulloblastoma growth via its signaling with Nrp1[277]. Despite the presence of both proteins being validated in primary tissue by immunostaining, follow-up studies targeting these interaction pairs are needed.

Altogether, in this study I have addressed the tumor heterogeneity inherent in IDH-mutant gliomas through a comparative analysis between oligodendrogliomas and astrocytomas, characterizing and expanding on the different tumor cell populations, determining TAM composition across IDH-mutant gliomas subtypes, inferring the crosstalk between tumor and TAM cell populations, and assessing the consistency of tumor and TAM cell populations at tumor recurrence. These findings represent a valuable contribution to our understanding of the biology of IDH-mutant gliomas and lay a solid foundation for further research in this field.

Epilogue

Recent method developments have allowed for easier and more affordable molecular characterization of different tumor types, resulting in a shift towards including molecular features into CNS tumor classification. This is evidenced by the 5th edition of the WHO CNS tumor classification, where many novel tumor types and subtypes have emerged as a result. Therefore, understanding the biological differences underlining these novel tumor types and subtypes has become a necessity in the recent times.

Consequently, single-cell technologies have powered multitude of comparative studies aimed at filling this gap in knowledge. These studies, including the IDH-mutant glioma and ATRT projects I researched on, typically encompass several Omics layers, such as transcriptomics and chromatin accessibility. While being fundamental to characterize the commonalities and differential characteristics across tumor subtypes, single-cell technologies can often be very descriptive, resulting in the discovery and characterization of novel tumor cell types, such as the RE cells in IDH-mutant gliomas or the IPC-like population in ATRTs. These findings hold promise for the development of subtype-specific therapies, albeit still require additional validation beyond what bioinformatic analyses of single-cell data alone can achieve, typically involving immunostaining of key protein markers to validate the presence of a cell population and the use of preclinical models such as cell lines, organoid models or animal models for subsequent hypothesis-driven experiments.

Nonetheless, a paradigm shift is currently taking place as a result of the development of spatial transcriptomics methodologies, which provide insights into the spatial arrangement of the cells of different cell populations within the tumor tissue. Altogether, spatial transcriptomics offers a new layer of possibilities, not only allowing for the validation of these novel tumor cell populations but also for the examination of the crosstalk between different tumor cells and between the tumor and its microenvironment, through the co-expression of ligand-receptor pairs together with the spatial disposition of the different cell types. Moreover, experimental setup is a crucial aspect in spatial transcriptomics, as it can enable the comparative analysis of tumors with varying degree of immune infiltration, or the evaluation of longitudinal changes occurring at tumor recurrence, among others.

All in all, single-cell technologies are reshaping the way studies are carried out, particularly in the field of cancer research. As technologies develop, so does their associated software tools, resulting in a stream of framework and method-specific tools that smooth the adoption of single-cell analyses by the scientific community. With the popularity of single-cell technologies at its peak, not only data analysis is facilitated by the software tools but also new solutions that streamline the generation of high-quality data visualization figures are becoming available, for which I contributed with the development of my R package, `SCpubr`.

While my work as a PhD student has arrived to its end, as a scientist and bioinformatician I am thrilled to experience the coming of a new era in both tumor research and bioinformatics. May these new avenues of research pave the way towards more specific and effective therapy treatments.

Publications

Blanco-Carmona, E.*, Paasen, I.*, He, J.*, Büllesbach, A., Buh, L. J., Federico, A., Liu, I., Young, M., Kildisiute, G., Behjati, S., Vibhakar, R., Donson, A., Foreman, N., Hovestadt, V., Shaw, M., Chi, N. S., Frühwald, M., Korshunov, A., Hasselblatt, M., Hoving, W. E., Jäger, N., Johann, D. P., Pfister, M. S., Filbin, M., Drost, J, Kool, M. **Multi-omics sequencing of atypical teratoid/rhabdoid tumors unveils a shared rhabdoid ground-state population promoting subtype-specific differentiation trajectories.** *Manuscript in preparation. Provisional title.*
* Joint first authors.

Blanco-Carmona, E.*, Narayanan, A.*, Hernandez, I.*, Nieto, J.C.*, Elosua-Bayes, M., Sun, X., Schmidt, C., Pamir, N., Özduman, K., Herold-Mende, C., et al. (2023). **Tumor heterogeneity and tumor-microglia interactions in primary and recurrent IDH1-mutant gliomas.** Cell Rep Med, 101249.
* Joint first authors.

Blanco-Carmona, E. (2022). **Generating publication ready visualizations for Single Cell transcriptomics using SCpubr.** bioRxiv. *Manuscript in preparation.*

Narayanan, A., **Blanco-Carmona, E.**, Demirdizen, E., Sun, X., Herold-Mende, C., Schlesner, M., and Turcan, S. (2020). **Nuclei isolation from fresh frozen brain tumors for single-nucleus rna-seq and atac-seq.** Journal of Visualized Experiments 2020, 1–14.

Acknowledgements

This PhD experience has been an amazing journey, both scientifically and at a personal level, where I was able to cross paths with many different people. In this section, I will try to put into words the gratitude I feel towards each of you.

SUPERVISORS

Prof. Dr. Marcel Kool: Thank you for being an excellent supervisor. Not only you always promoted a work environment in which I could make independent decisions regarding my project, but also pushed me to improve weak areas such as my skills in scientific writing and translating results into biological insights. You have been a supervisor one can always rely on and has always been open to discuss any problems. I am very grateful that you took me under your supervision when I was on the search for another research group during the pandemic. Furthermore, I am immensely thankful to you for proofreading my thesis. It has been an honor being your PhD student.

Prof. Dr. Matthias Schlesner: Thank you for seeing the potential in me and recruiting me as a PhD student. You have been an excellent supervisor, and I have always been in awe at the group environment and dynamics that you promote. Thank you for the continued supervision in the IDH-mutant glioma project, even after your transition to Augsburg. It has been an honor spending part of my PhD working under your supervision.

Dr. Natalie Jäger: Thank you for interviewing me and considering me for a PhD position involving the ATRT project. Thanks to your positive opinion, I was able to continue my PhD in Heidelberg during the pandemic. I really enjoyed being part of your group dynamics and your supervision style. Thank you for mentoring me and supervising my work on the ATRT project.

Dr. med. Pascal Johann: Thank you also for your positive opinion during my interviews for the ATRT project. During the time you supervised me, it was a pleasure working with you and analyzing the results together. Your passion for research was truly contagious. Thank you for easing my transition into the group and helping me out during the initial stages of the ATRT project.

DEFENSE COMMITTEE AND TAC MEMBERS:

Prof. Dr. Benedikt Brors: Thank you for being my first examiner and a TAC committee member. Your feedback during TAC meetings as well as during the thesis writing has been very helpful to progress in my PhD. I also want to thank you for always being open to discuss thesis writing-related matters, especially without prior schedule.

Dr. Şevin Turcan: Thank you for being my second examiner and participating in my TAC meetings. Furthermore, it has been an honor working with you in the IDH-mutant glioma project. Your passion for research and understanding about the technical problems commonly arising in bioinformatics have made analyzing the IDH-mutant gliomas datasets such an enjoyable experience. Thank you for always being open to discuss any issues and for proofreading the IDH-mutant glioma part of the thesis.

Dr. Christiane Opitz: Thank you for being part of my PhD defense committee.

Prof. Dr. Marcel Kool: Thank you for being part of my TAC and PhD defense committees.

Prof. Dr. Matthias Schlesner: Thank you for being part of my TAC committee.

Dr. Natalie Jäger: Thank you for being part of my TAC committee.

RESEARCH GROUPS

B062: First and foremost, I want to thank **Prof. Dr. med. Stefan Pfister** for granting me the opportunity to be part of the Division of pediatric neurooncology at DKFZ. It has been an honor to work under several of the divisions within your group. Thank you for the always insightful feedback during seminars and in the ISPNO 2022 presentation. Next, I want to thank both of the research groups I had the pleasure to work with during my time here: **BO62 | Kool** and **B062 | Jäger**, for the good times we have shared together, both scientifically and during social gatherings. Thanks for accepting me in the group and helping me out when needed. I would like to especially thank **Anne**, **Aylin**, **Apurva** and **Shanzheng** for always reaching out and helping me with any problems I had and to **Beni** for his endless anime recommendations. Also, I would like to thank the rest of the team for always being so nice to me.

B240 | BODA: Although brief, my stay in B240 was truly memorable. It was a pleasure to get to know all the people in this group, and share moments and laugh together. Specially, during corona times, the virtual gatherings playing different games always brought me a smile. I would like to especially thank Christian and Daniel for helping me get started in the group and with single-cell transcriptomic analyses and to the rest for being always so nice to me.

SOCIAL GROUPS

I would like thank the different groups of friends I have the luck to be part of:

LUGAs: What an office environment! Areeba, Christina, Dina, and later on Zaira. I cannot sufficiently express how special and unique sharing the office with you has been. All the moments and laughs we have shared together. I can very confidently say that I will not experience this sense of togetherness and belonging again in a work environment. You truly made my everyday PhD life unforgettable. For this, I will always keep these memories close to my heart. Thank you for everything!

Boulder Truppe: Thanks for being the coolest group of people to hang out with when bouldering. Anastasiia, Anastasia, Carine, Julia, Kai, Killian, Laia, Micha, Paul and Sven. With you, I found out the beauty of a new hobby, one that has become a big part of my personality. I am very grateful to share moments and overhanging climbs with all of you.

Otto's flotte Truppe: Bea, Christina, Daniela, Giulia, Jonas, Nicola, Paul, Sandro, Sonia, Yassin and Vivien. The pandemic brought us all together, and together we canoed down the river in a 3-canoe docked formation! The moments and parties we shared over the course of the last years, are memories that I will never forget. Thank you for being such an amazing group of people!

Ese de ahí se pincha: Amparo, Laura, María Pilar, Patricia and Salva. No matter how much time has passed and how much it has weighted on me the fact that I have become an expat, you always remind me that I have a place in Valencia with you. For this, I cannot thank you enough. Thanks for always being there every time I came back to Valencia and had a life crisis. Thanks for reminding me of why I chose this path. Thanks for supporting me even when my decisions took me away from you. You are the best childhood friends one can ever ask for.

El consejo de las 3 pa*as: Emilio and Marc. While our lives took different directions ever since we finished the Bachelor, we have always remained true to the friendship we built during these years. Many years later, I can proudly say that no amount of time will ever change that. Cheers to another completed chapter, and cheers to the fact that the only person who did not pursue a PhD is the richest among us by far.

The Mala*as: You know who you are. Anastasia, Carine and Laia. The crazy people I cannot see myself climbing without. You make an essential and irreplaceable part of my life in Heidelberg. Meeting you has been one of the best experiences of my life. That's how much I appreciate all of you. There are no words to explain how much good you do to me. Special mention here to Ahmed, Hannah, Parnian and Kai, which are a great complementary mix to the prior crazy trio, and that always bring me smiles and are always a source of good conversations.

Uncanny things: The name of this group used to be nicer and more seasonal, but I gave it my personal touch and remained like this ever since. Bea and Christina. The most responsible people I know. The happiest PhD students I have ever met. You are both a role model to me. You have taught me so much about how to life a happy life, and I still have much more to learn from you! I hope you both keep inspiring me for many more years.

Spanish gang: Celia, Pau and Sonia. We started this journey together, went through a pandemic together, and will finish the PhD also very close one from the other. With you, I have not only found great friends, but also a family here in Germany. I lack the words to describe how profoundly, deeply grateful I am for having shared these five years of PhD together with you. No matter where our paths will take us next, I have no doubts that we will remain true to the bonds we forged here.

PEOPLE

Five years is a long period of time. Enough to cross paths with many people. Enough for them to leave an imprint on you. Here, I would like to thank each and every one of you individually. This ended up shorter than I anticipated (believe it or not), as I really tried my best to keep it only to PhD-related matters. Find your names arranged alphabetically, for clarity:

Albert: It was great that we both lived in Germany for a while. Getting to meet you outside Valencia was a very nice experience and it helped me reconnect with my roots. Thank you for accompanying me during most of my PhD time and being such an amazing friend! While I am sure our friendship will continue no matter where we are, I am curious whether we will end up living close by in another foreign country in the future!

Alexandra: You have been for sure one of the biggest surprises of the recent months. It is great to have connected with you and I am very proud of the friendship we share. I cannot thank you enough for being there for me, listening to me and sheltering me during these months of thesis writing. Let's celebrate this thesis with an overhanging climb! Thank you for being such a caring and amazing friend!

Amparo: Chinchorro! Thank you for always being there every time I go back to Valencia. It is nice to see that there are things that never change, and that is how I feel back at home every time I meet you again. I am very grateful to have such an amazing friend!

Anastasiia: Thank you for showing me kindness and reaching out to me when I was not going through nice moments during thesis writing. I am really thankful to you that you decided to meet me to make me smile! Thank you for being such an amazing friend! I cannot wait to hear the next crazy story that happened in your life. For many more bouldering sessions together!

Anastasia: The Greek Bulder! What to say to you? You are an amazing person. Do not ever let anything change that personality of yours. You are pure impulse, pure randomness. And I love it! It is very refreshing being around people like you. I cannot thank you enough for bringing me joy every time we see each other, for being there in the good and not so good times and for being so upfront and honest about anything you think. I am sure we will continue to share more experiences together in the future! Thank you for being such an amazing friend!

Areeba: You know, after all my efforts, I really did not manage to break that *britishness* of yours. I really think you can take that as a win! Sharing the office with you has been one of the greatest things out of my PhD time. You've taught me a lot and I really do not comprehend how you can manage everything you do so flawlessly. You are an amazing friend, Areeba. I cannot thank you enough for all the times you've been there for me and all the things I have learnt from you. *Tank* you for proofreading the acknowledgments of this thesis!

Ashwin: Thank you for being such an amazing collaborator. Your passion for research is truly contagious. It was an honor working alongside you in this project. Thank you for proofreading the IDH-mutant glioma introduction of this thesis.

Bea (trice): Who would have known that, what a potato chip blessing started, would end up in such a beautiful friendship? You are amazing, Bea. You are such a source of motivation and positive attitude that I am genuinely envious about how you cope with life. A long time ago, I set out myself to learn as much as I could from your personality and way of thinking. Thank you for being such a source of inspiration and a pillar supporting me throughout the whole PhD. I am very grateful for the bond we share, and I have zero doubts we have many more great experiences together to share in the future! Thank you for being such an amazing friend!

Bea (Moltasa): Thank you for always finding the time to meet me when I was back in Valencia. You reminded me every time that I had a place and a home to get back to if things would not work out here in Germany. No matter how much time passes, when we meet again it always feels like always. You are an amazing friend, Moltasilla. Thank you for being part of my life!

Carine: Thank you for being so nice, so thoughtful and so peaceful! You are such a great source of positivity and energy, do not ever let anything change that! You became an essential part of my life in Heidelberg, and one of the big driving factors preventing me from going insane during this thesis writing phase. I am very thankful to you for all the good you do to me, and for the beautiful friendship we've built together! You know, annoying you might be perhaps one of the things that brings me joy the most. For many more times together! Don't you think I will ever let you escape from my annoyingness! Thank you for being such an amazing friend!

Carmen: Thank you for all the walks we had during the pandemic! You were a great support during those times. It was great to connect with you again, and I am really looking forward to our next catch-up meeting!

Celia: Thank you for being there since the very beginning of the PhD. Thank you for offering shelter every single time I needed it, and for being part of this little family we created here in Heidelberg. We went through a lot together, so much that at this point it feels that we can deal with anything the PhD has for us, together. What the endless Catan nights during the pandemic started, I am sure will carry over for many, many years. Thank you for being an integral part of

my life here in Heidelberg an such an amazing friend! Thank you for proofreading the Prologue and Epilogue section of this thesis!

Christina: You are the sole reason I did not go insane in the office during all these years. I owe you so many things, you have taught me so much. I am very happy to have experienced the PhD journey alongside you! I am very proud of the distinct groups of friends we are part of and all the adventures we have had together. You have such an overwhelmingly positive personality, that it is really contagious. I have tried to learn as much from you as possible on how to cope with life events the positive way. Thank you for always being there for me and being such an amazing friend! Thank you for helping me correctly translate the abstract to German and for proofreading the SCpubr chapter of the thesis!

Clara: You did not expect such an oversharing from my side when you asked me how I was doing, and here we are, with such a beautiful friendship as a result! I am very happy to have crossed paths with you! You have supported me and have been a safe wall preventing me from going insane during thesis writing. I am immensely grateful for all the talks we have shared together, for bringing a smile to my face every single time we meet together and for doing me so good. Thanks for being such an amazing friend!

Diego: Thank you for adapting to my busy schedule every time I went back home! Meeting you from time to time was truly nice, especially after discovering that we both adopted the same new hobby! Every time we go climbing together it feels as when we were 17 years old and you were my running coach!

Dina: Thank you for being part of the amazing office environment that we generated together! Thank you for all the laughs and great moments and plans we organized together as LUGAs.

Emilio: Thank you for demonstrating me that, no matter how much time pass, you will always have time for me to catch up whenever I go back to Valencia! It always feels like we last met yesterday every time we gather together. Thanks for being such an amazing friend!

Eva: Thank you for being such a great support during these last months! Meeting with you has helped me a lot to get through the thesis writing phase, and I am always very happy to share moments with you. Thank you for always being so comprehensive with me, and lending an ear to listen to my problems! Thank you for being such an amazing friend!

Guille: Thank you for always arranging a bit of your time to meet me when I come back to Valencia! Every single time we meet, I get inspired by you. You are so ahead in the field of personal development that listening to you talking about how you set up goals and a system to achieve them is truly fascinating. Every time I am about to see you, I am always eager to hear about your recent successes. Thank you for demonstrating me that our friendship remains untouched no matter the time! Thank you for being such an amazing friend!

Javi: Thank you for supporting me during the thesis writing step. You have always lent an ear whenever I needed it, even if the times were also rough for you, for which I am truly grateful. Thank you for introducing me to the concept of AFK games, which has kept my gaming soul at ease all these months I could actually not invest any time into videogames. Finally, thank you for helping me out with the logo design of my R package! Thank you for being such an amazing friend!

Johannes: Thank you for always being such a positive influence and a great company to hang out with! You are an amazing friend, Johannes. Thank you for always cheering me up and being there for me in the bad moments. I am very happy I crossed paths with you!

Kai: Thank you for offering your help whenever I needed it and listening to my own problems. I really appreciate you reaching out for me and checking on me during these months of thesis writing! I am very happy I crossed paths with you!

Laia: Thank you for being such an important part of my life in Heidelberg. I really enjoy spending time with you, connecting with you and doing plans together. I really appreciated your support and assertiveness during the months of thesis writing. Despite not being easy, you gave it your best to be there for me, and for that I am immensely grateful. Thank you for being such an amazing friend! For many more adventures together!

Laura: Laureta! Thank you for being always so supportive and caring for me. While us living in different cities and countries rather than Valencia has made it almost impossible for us to coincide at the same time back home, that has not stopped our friendship to be as good as always. Thank you for being such an amazing friend!

Lavinia: Ragazza! Thank you for being my wonderful salsa partner! I really enjoyed the time we spent together in Heidelberg, and I am very proud we both together hosted the greatest birthday party ever! While our paths separated again, I am sure the friendship we built in Heidelberg will remain. Thank you for all the moments, memories and support I received from you all this time! Thank you for being such an amazing friend!

Lucía: There are not enough words to describe how much you have helped me. Thanks to you, I was able to understand and start to take ownership of the insecurities that I have, that resulted in my recurrent anxiety periods. You are an amazing therapist, and I am tremendously grateful Salva connected us together. With your guidance, I was able to go through the thesis writing phase without constantly falling into negative loops. Not only that, but you helped me realize the things I could actually learn from this period of my life. Thank you for everything, Lucía!

Marc: Thank you for accepting my random phone calls without any prior warning all these years! You have been a pillar supporting me throughout my PhD. You were happy on my successes and supportive on the hard times. You have always made time to meet me when I was in Valencia, no matter how tight the schedule was! Thank you for always making me laugh, especially on the hard times, and for always having an endless collection of anime series to recommend me. I am immensely grateful for having such a caring and amazing friend as you by my side. For many more experiences together!

María Pilar: Pilarica! Thank you for always being there and supporting me every time I reached out to you. Thank you for making me feel back at home every time I went back to Valencia. Thank you for your encouraging words and for being such an amazing friend!

Marta: Martita! Getting to know you has been one of the biggest surprises in my PhD. You are one of the most cheerful and happiest people I have ever met! You inspire me to be better, to work on myself and learn to enjoy the small things in life. You always manage to bring out the best part of me and to connect with the things that really matter. You have been an essential support all this time, and I cannot state how much I owe you. Thank you for being such a wonderful and amazing friend and an irreplaceable part of my life here in Heidelberg. For more paellitas and moments together in the future!

Micha: Thank you for always being supportive with me and listening to my problems! You are such a great person, Micha. Do not let anybody ever change the immense good that is within you. You are an amazing friend and an inspiration for me to become a better person. I am very happy to have crossed paths with you during the PhD. For more moments together!

Mireia: Over the years, you have demonstrated to me that I can count on you for whatever I need. Moving to another country did not change this fact. Thank you for always being there for me, despite the difficult and sometimes crazy schedules that you lately have. Every single time we have a call I get reminded how amazing of a friend you are, and how grateful I am to have crossed paths with you! For many more moments together!

Nicola: We met at the very beginning of your PhDs and here we are, five years later! Thank you for the great experiences we have shared together. Thank you for listening to me and giving me advice in all the dramas I have had over the years. I am very happy we crossed paths with you, Nico. I have learnt a lot from you and I am very happy to have somebody as amazing as you as a friend!

Oscar: Thank you for all the coffee breaks we have shared over the last months, for listening to my problems and giving advice as a senior bioinformatician! It is really refreshing to have you around in the Mathematikon. I am very happy I crossed paths with you! Thank you for proofreading the Prologue and Epilogue sections of this thesis!

Patricia: Thank you for supporting me and listening to my problems and deep talks every time I come back to Valencia. It is great to see how, as each of us grows with the years, we are able to see each other's progress. It is truly fantastic! Thank you for being such an amazing friend!

Pau: Senyor! Thank you for being an integral part of my family here in Heidelberg. We have been through the whole PhD together and nothing makes me prouder! Seeing your passion for research is admirable. But even crazier is the fact that every Catan game is different because the economy is ever changing! Thank you for always being there for me, I am really grateful to have somebody as amazing as you as a friend. When you become a PI, if you ever need a senior scientist for figure designing, hire me! Thank you for proofreading the SCpubr thesis chapter.

Salva: You know, I think this thesis ended up being very chu-chu-chuuuuuuli! When I decided to move to another country, we came up with the resolve to strengthen our friendship and, as a result, it has become very solid nowadays. I am very thankful for all the times you listened to my problems, and tried to give me some hints on what to work on next. Thank you for helping me find the best therapist, in order to start taking ownership of my insecurities. You are an amazing friend, Salva, and I am very proud and thankful for being yours!

Sonia: Thank you for being part of my family here in Germany! You are the major reason I have not gone insane during these years. Especially, during the pandemic, you were a driving force of positivity. You've been together with me since the beginning of the PhD and have witnessed and supported me in every up and down this journey had to offer me. You are amazing and irreplaceable, and the friendship we have is one of the best things Germany has provided me. Thank you for everything, Sonia. I could not ask for a better friend as you!

Sonja: My thesis writing buddy! As promised, here we are, at the other side of this process, happier than ever! I am very proud of what you have accomplished! Thanks for supporting me during my writing months!

Zaira: Thank you for supporting me during several of my major anxiety breakdowns over the last years, I really appreciate all the effort you put into making me break out of the loop. I am really happy to have crossed paths with you, shared the office together, and be friends with you! If you read this, I hope one day you resume reading "Omniscient reader". It's really good!

FAMILY

In this section, I would like to thank my family. For this, I will switch to Spanish.

Mamá: Muchísimas gracias por todo el apoyo que me has proporcionado estos años. Siempre le has quitado el hierro a todos los problemas que te he ido comentado y has hecho que me centre en lo importante. Te lo dije cuando empecé esta etapa y te lo digo también ahora: esto no hubiera sido posible si no hubieses sido tan persistente en que aprendiese inglés cuando era pequeño. Te agradezco de corazón todo el esfuerzo y cariño que me has dedicado estos estos años. ¡Gracias por todo!

Papá: Muchísimas gracias por todo el apoyo que me has dado estos años. Especialmente, muchas gracias por ser la contraparte positiva en todas aquellas llamadas tras el trabajo en el que yo estaba harto y muy negativo, muchas veces dudando de mis propias capacidades. Siempre me ayudaste a no perder la perspectiva de lo importante y a que las cosas no suelen ser tan malas como uno se las imagina. ¡Gracias por todo!

Francisco: Muchísimas gracias por abrirme la puerta de tu casa siempre que quisiera y por preocuparte por mí a lo largo del doctorado. Eres un gran hermano, y no sólo en el sentido literal de la frase. Muchas gracias por ofrecerme tus consejos a nivel laboral y por recordarme todos estos años que, pasara lo que pasara, siempre tendría a mi familia cerca para lo que necesitase. ¡Gracias por todo!

Jules: Muchísimas gracias por ser una persona tan alegre y risueña. Quedar contigo siempre es muy enriquecedor, porque siempre promueves un ambiente muy positivo. Muchas gracias por cuidar de mí y por ayudarme con toda la burocracia alemana. Eres una persona maravillosa. ¡Qué ganas de que te conviertas en parte de esta familia! ¡Gracias por todo!

Tengo una familia maravillosa, que cree en mí y que me apoya en todo lo que hago. No puedo expresar lo suficiente lo agradecido que estoy de teneros. De la misma manera, me gustaría agradecer el amor y apoyo incondicional que he recibido del resto de mi familia: de la abuela Francisca, la tía Júlia, mis primo/as y tíos/as. Asimismo, recordar en este momento tan especial y significativo, a aquellos que ya no están: el abuelo Paco, la abuela Carmen, el abuelo Cristóbal, el tío Alfonso, y mi perrita, Inca. Siempre os llevo en mi corazón.

FINAL THOUGHTS

Finally, I would like to bring closure to the acknowledgements with a personal thought. I began this chapter in my life thriving for new experiences and, in a sense, wanting to find myself along the way. Despite things not being always easy, going through a pandemic, and having to face my insecurities, I managed to persevere and finish this chapter.

The experiences I gathered, both at personal and professional level, together with the people I met and the life lessons I learnt during this chapter, are truly invaluable. I would definitely repeat this journey. Thank you to everybody for being part of it, and to me for adventuring to start it on the first place.

Finalmente, me gustaría acabar los agradecimientos con una reflexión personal. Empecé esta etapa con muchas ganas de vivir experiencias nuevas y, de alguna manera, encontrarme a mí mismo. Pese a que las cosas no fueron siempre fáciles, con una pandemia de por medio, y teniendo que enfrentarme de frente con mis inseguridades, conseguí perseverar y acabar el doctorado.

Las experiencias a nivel personal y profesional, las personas encontradas por el camino y las lecciones vitales aprendidas durante esta etapa, son invaluables. Volvería a repetir esta experiencia. ¡Muchas gracias a todos por formar parte de ella y a mí por aventurarme a empezarla!

Bibliography

1. Organisation mondiale de la santé and Centre international de recherche sur le cancer eds. (2021). Central nervous system tumours 5th ed. (International agency for research on cancer).

2. Ariffin, H., Hainaut, P., Puzio-Kuter, A., Choong, S.S., Chan, A.S.L., Tolkunov, D., Rajagopal, G., Kang, W., Lim, L.L.W., Krishnan, S., et al. (2014). Whole-genome sequencing analysis of phenotypic heterogeneity and anticipation in Li–Fraumeni cancer predisposition syndrome. Proc. Natl. Acad. Sci. U. S. A. *111*, 15497–15501. https://doi.org/10.1073/pnas.1417322111.

3. Kolodziejczak, A.S., Guerrini-Rousseau, L., Planchon, J.M., Ecker, J., Selt, F., Mynarek, M., Obrecht, D., Sill, M., Autry, R.J., Stutheit-Zhao, E., et al. (2023). Clinical outcome of pediatric medulloblastoma patients with Li-Fraumeni syndrome. Neuro-Oncol. *25*, 2273–2286. https://doi.org/10.1093/neuonc/noad114.

4. Yan, H., Parsons, D.W., Jin, G., McLendon, R., Rasheed, B.A., Yuan, W., Kos, I., Batinic-Haberle, I., Jones, S., Riggins, G.J., et al. (2009). IDH1 and IDH2 mutations in gliomas. N. Engl. J. Med. *360*, 765–773. https://doi.org/10.1056/NEJMoa0808710.

5. Versteege, I., Sévenet, N., Lange, J., Rousseau-Merck, M.-F., Ambros, P., Handgretinger, R., Aurias, A., and Delattre, O. (1998). Truncating mutations of hSNF5/INI1 in aggressive paediatric cancer. Nature *394*, 203–206. https://doi.org/10.1038/28212.

6. Hasselblatt, M., Nagel, I., Oyen, F., Bartelheim, K., Russell, R.B., Schüller, U., Junckerstorff, R., Rosenblum, M., Alassiri, A.H., Rossi, S., et al. (2014). SMARCA4-mutated atypical teratoid/rhabdoid tumors are associated with inherited germline alterations and poor prognosis. Acta Neuropathol. (Berl.) *128*, 453–456. https://doi.org/10.1007/s00401-014-1323-x.

7. Louis, D.N., Perry, A., Wesseling, P., Brat, D.J., Cree, I.A., Figarella-Branger, D., Hawkins, C., Ng, H.K., Pfister, S.M., Reifenberger, G., et al. (2021). The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. Neuro-Oncol. *23*, 1231–1251. https://doi.org/10.1093/neuonc/noab106.

8. Capper, D., Jones, D.T.W., Sill, M., Hovestadt, V., Schrimpf, D., Sturm, D., Koelsche, C., Sahm, F., Chavez, L., Reuss, D.E., et al. (2018). DNA methylation-based classification of central nervous system tumours. Nature *555*, 469–474. https://doi.org/10.1038/nature26000.

9. Fernandez, A.F., Assenov, Y., Martin-Subero, J.I., Balint, B., Siebert, R., Taniguchi, H., Yamamoto, H., Hidalgo, M., Tan, A.-C., Galm, O., et al. (2012). A DNA methylation fingerprint of 1628 human samples. Genome Res. *22*, 407–419. https://doi.org/10.1101/gr.119867.110.

10. Hovestadt, V., Jones, D.T.W., Picelli, S., Wang, W., Kool, M., Northcott, P.A., Sultan, M., Stachurski, K., Ryzhova, M., Warnatz, H.-J., et al. (2014). Decoding the regulatory landscape

of medulloblastoma using DNA methylation sequencing. Nature *510*, 537–541. https://doi.org/10.1038/nature13268.

11. Moran, S., Martínez-Cardús, A., Sayols, S., Musulén, E., Balañá, C., Estival-Gonzalez, A., Moutinho, C., Heyn, H., Diaz-Lagares, A., de Moura, M.C., et al. (2016). Epigenetic profiling to classify cancer of unknown primary: a multicentre, retrospective analysis. Lancet Oncol. *17*, 1386–1395. https://doi.org/10.1016/S1470-2045(16)30297-2.

12. Ostrom, Q.T., Price, M., Neff, C., Cioffi, G., Waite, K.A., Kruchko, C., and Barnholtz-Sloan, J.S. (2023). CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2016—2020. Neuro-Oncol. *25*, iv1–iv99. https://doi.org/10.1093/neuonc/noad149.

13. Ostrom, Q.T., Price, M., Ryan, K., Edelson, J., Neff, C., Cioffi, G., Waite, K.A., Kruchko, C., and Barnholtz-Sloan, J.S. (2022). CBTRUS Statistical Report: Pediatric Brain Tumor Foundation Childhood and Adolescent Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2014–2018. Neuro-Oncol. *24*, iii1–iii38. https://doi.org/10.1093/neuonc/noac161.

14. Price, M., Ryan, K., Shoaf, M.L., Neff, C., Iorgulescu, J.B., Landi, D.B., Cioffi, G., Waite, K.A., Kruchko, C., Barnholtz-Sloan, J.S., et al. (2024). Childhood, adolescent, and adult primary brain and central nervous system tumor statistics for practicing healthcare providers in neuro-oncology, CBTRUS 2015-2019. Neuro-Oncol. Pract. *11*, 5–25. https://doi.org/10.1093/nop/npad061.

15. International Classification of Diseases for Oncology, 3rd Edition (ICD-O-3) https://www.who.int/standards/classifications/other-classifications/international-classification-of-diseases-for-oncology.

16. Heumos, L., Schaar, A.C., Lance, C., Litinetskaya, A., Drost, F., Zappia, L., Lücken, M.D., Strobl, D.C., Henao, J., Curion, F., et al. (2023). Best practices for single-cell analysis across modalities. Nat. Rev. Genet. *24*, 550–572. https://doi.org/10.1038/s41576-023-00586-w.

17. Baek, S., and Lee, I. (2020). Single-cell ATAC sequencing analysis: From data preprocessing to hypothesis generation. Comput. Struct. Biotechnol. J. *18*, 1429–1439. https://doi.org/10.1016/j.csbj.2020.06.012.

18. Han, A., Glanville, J., Hansmann, L., and Davis, M.M. (2014). Linking T-cell receptor sequence to functional phenotype at the single-cell level. Nat. Biotechnol. *32*, 684–692. https://doi.org/10.1038/nbt.2938.

19. Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P.K., Swerdlow, H., Satija, R., and Smibert, P. (2017). Simultaneous epitope and transcriptome measurement in single cells. Nat. Methods *14*, 865–868. https://doi.org/10.1038/nmeth.4380.

20. Larsson, L., Frisén, J., and Lundeberg, J. (2021). Spatially resolved transcriptomics adds a new dimension to genomics. Nat. Methods *18*, 15–18. https://doi.org/10.1038/s41592-020-01038-7.

21. Wen, L., Li, G., Huang, T., Geng, W., Pei, H., Yang, J., Zhu, M., Zhang, P., Hou, R., Tian, G., et al. (2022). Single-cell technologies: From research to application. Innov. Camb. Mass *3*, 100342. https://doi.org/10.1016/j.xinn.2022.100342.

22. Zappia, L., and Theis, F.J. (2021). Over 1000 tools reveal trends in the single-cell RNA-seq analysis landscape. Genome Biol. *22*, 301. https://doi.org/10.1186/s13059-021-02519-4.

23. Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., et al. (2021). Integrated analysis of multimodal single-cell data. Cell *184*, 3573-3587.e29. https://doi.org/10.1016/j.cell.2021.04.048.

24. Wolf, F.A., Angerer, P., and Theis, F.J. (2018). SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. *19*, 15. https://doi.org/10.1186/s13059-017-1382-0.

25. Slyper, M., Porter, C.B.M., Ashenberg, O., Waldman, J., Drokhlyansky, E., Wakiro, I., Smillie, C., Smith-Rosario, G., Wu, J., Dionne, D., et al. (2020). A single-cell and single-nucleus RNA-Seq toolbox for fresh and frozen human tumors. Nat. Med. *26*, 792–802. https://doi.org/10.1038/s41591-020-0844-1.

26. Andrews, T.S., and Hemberg, M. (2018). Identifying cell populations with scRNASeq. Mol. Aspects Med. *59*, 114–122. https://doi.org/10.1016/j.mam.2017.07.002.

27. Luecken, M.D., and Theis, F.J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. Mol. Syst. Biol. *15*, e8746. https://doi.org/10.15252/msb.20188746.

28. Amezquita, R.A., Lun, A.T.L., Becht, E., Carey, V.J., Carpp, L.N., Geistlinger, L., Marini, F., Rue-Albrecht, K., Risso, D., Soneson, C., et al. (2020). Orchestrating single-cell analysis with Bioconductor. Nat. Methods *17*, 137–145. https://doi.org/10.1038/s41592-019-0654-x.

29. Kharchenko, P.V. (2021). The triumphs and limitations of computational methods for scRNA-seq. Nat. Methods *18*, 723–732. https://doi.org/10.1038/s41592-021-01171-x.

30. Mereu, E., Lafzi, A., Moutinho, C., Ziegenhain, C., McCarthy, D.J., Álvarez-Varela, A., Batlle, E., Sagar, null, Grün, D., Lau, J.K., et al. (2020). Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. Nat. Biotechnol. *38*, 747–755. https://doi.org/10.1038/s41587-020-0469-4.

31. Stegle, O., Teichmann, S.A., and Marioni, J.C. (2015). Computational and analytical challenges in single-cell transcriptomics. Nat. Rev. Genet. *16*, 133–145. https://doi.org/10.1038/nrg3833.

32. Galluzzi, L., Kepp, O., and Kroemer, G. (2012). Mitochondria: master regulators of danger signalling. Nat. Rev. Mol. Cell Biol. *13*, 780–788. https://doi.org/10.1038/nrm3479.

33. Detmer, S.A., and Chan, D.C. (2007). Functions and dysfunctions of mitochondrial dynamics. Nat. Rev. Mol. Cell Biol. *8*, 870–879. https://doi.org/10.1038/nrm2275.

34. Seurat - Guided Clustering Tutorial.

35. Young, M.D., and Behjati, S. (2020). SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. GigaScience *9*, giaa151. https://doi.org/10.1093/gigascience/giaa151.

36. Fleming, S.J., Chaffin, M.D., Arduini, A., Akkad, A.-D., Banks, E., Marioni, J.C., Philippakis, A.A., Ellinor, P.T., and Babadi, M. (2023). Unsupervised removal of systematic background noise from droplet-based single-cell experiments using CellBender. Nat. Methods *20*, 1323–1335. https://doi.org/10.1038/s41592-023-01943-7.

37. Yang, S., Corbett, S.E., Koga, Y., Wang, Z., Johnson, W.E., Yajima, M., and Campbell, J.D. (2020). Decontamination of ambient RNA in single-cell RNA-seq with DecontX. Genome Biol. *21*, 57. https://doi.org/10.1186/s13059-020-1950-6.

38. Xi, N.M., and Li, J.J. (2021). Protocol for executing and benchmarking eight computational doublet-detection methods in single-cell RNA sequencing data analysis. STAR Protoc. *2*, 100699. https://doi.org/10.1016/j.xpro.2021.100699.

39. Xi, N.M., and Li, J.J. (2021). Benchmarking Computational Doublet-Detection Methods for Single-Cell RNA Sequencing Data. Cell Syst. *12*, 176-194.e6. https://doi.org/10.1016/j.cels.2020.11.008.

40. Germain, P.-L., Lun, A., Garcia Meixide, C., Macnair, W., and Robinson, M.D. (2021). Doublet identification in single-cell sequencing data using scDblFinder. F1000Research *10*, 979. https://doi.org/10.12688/f1000research.73600.2.

41. McGinnis, C.S., Murrow, L.M., and Gartner, Z.J. (2019). DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. Cell Syst. *8*, 329-337.e4. https://doi.org/10.1016/j.cels.2019.03.003.

42. Neavin, D., Senabouth, A., Arora, H., Lee, J.T.H., Ripoll-Cladellas, A., sc-eQTLGen Consortium, Franke, L., Prabhakar, S., Ye, C.J., McCarthy, D.J., et al. (2024). Demuxafy: improvement in droplet assignment by integrating multiple single-cell demultiplexing and doublet detection methods. Genome Biol. *25*, 94. https://doi.org/10.1186/s13059-024-03224-8.

43. Vallejos, C.A., Risso, D., Scialdone, A., Dudoit, S., and Marioni, J.C. (2017). Normalizing single-cell RNA sequencing data: challenges and opportunities. Nat. Methods *14*, 565–571. https://doi.org/10.1038/nmeth.4292.

44. Ahlmann-Eltze, C., and Huber, W. (2023). Comparison of transformations for single-cell RNA-seq data. Nat. Methods *20*, 665–672. https://doi.org/10.1038/s41592-023-01814-1.

45. Lun, A.T.L., McCarthy, D.J., and Marioni, J.C. (2016). A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. F1000Research *5*, 2122. https://doi.org/10.12688/f1000research.9501.2.

46. Lause, J., Berens, P., and Kobak, D. (2021). Analytic Pearson residuals for normalization of single-cell RNA-seq UMI data. Genome Biol. *22*, 258. https://doi.org/10.1186/s13059-021-02451-7.

47. Hafemeister, C., and Satija, R. (2019). Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol. *20*, 296. https://doi.org/10.1186/s13059-019-1874-1.

48. Choudhary, S., and Satija, R. (2022). Comparison and evaluation of statistical error models for scRNA-seq. Genome Biol. *23*, 27. https://doi.org/10.1186/s13059-021-02584-9.

49. Sheng, J., and Li, W.V. (2021). Selecting gene features for unsupervised analysis of single-cell gene expression data. Brief. Bioinform. *22*, bbab295. https://doi.org/10.1093/bib/bbab295.

50. McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018). UMAP: Uniform Manifold Approximation and Projection. J. Open Source Softw. *3*, 861. https://doi.org/10.21105/joss.00861.

51. Maaten, L. van der, and Hinton, G. (2008). Visualizing Data using t-SNE. J. Mach. Learn. Res. *9*, 2579–2605.

52. Moon, K.R., van Dijk, D., Wang, Z., Gigante, S., Burkhardt, D.B., Chen, W.S., Yim, K., Elzen, A. van den, Hirn, M.J., Coifman, R.R., et al. (2019). Visualizing structure and transitions in high-dimensional biological data. Nat. Biotechnol. *37*, 1482–1492. https://doi.org/10.1038/s41587-019-0336-3.

53. Chari, T., and Pachter, L. (2023). The specious art of single-cell genomics. PLoS Comput. Biol. *19*, e1011288. https://doi.org/10.1371/journal.pcbi.1011288.

54. Luecken, M.D., Büttner, M., Chaichoompu, K., Danese, A., Interlandi, M., Mueller, M.F., Strobl, D.C., Zappia, L., Dugas, M., Colomé-Tatché, M., et al. (2022). Benchmarking atlas-level data integration in single-cell genomics. Nat. Methods *19*, 41–50. https://doi.org/10.1038/s41592-021-01336-8.

55. Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat. Biotechnol. *36*, 411–420. https://doi.org/10.1038/nbt.4096.

56. Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P., and Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. Nat. Methods *16*, 1289–1296. https://doi.org/10.1038/s41592-019-0619-0.

57. Xu, C., Lopez, R., Mehlman, E., Regier, J., Jordan, M.I., and Yosef, N. (2021). Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. Mol. Syst. Biol. *17*, e9620. https://doi.org/10.15252/msb.20209620.

58. Lopez, R., Regier, J., Cole, M.B., Jordan, M.I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. Nat. Methods *15*, 1053–1058. https://doi.org/10.1038/s41592-018-0229-2.

59. Lotfollahi, M., Wolf, F.A., and Theis, F.J. (2019). scGen predicts single-cell perturbation responses. Nat. Methods *16*, 715–721. https://doi.org/10.1038/s41592-019-0494-8.

60. Hie, B., Bryson, B., and Berger, B. (2019). Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. Nat. Biotechnol. *37*, 685–691. https://doi.org/10.1038/s41587-019-0113-3.

61. Zheng, S.C., Stein-O'Brien, G., Augustin, J.J., Slosberg, J., Carosso, G.A., Winer, B., Shin, G., Bjornsson, H.T., Goff, L.A., and Hansen, K.D. (2022). Universal prediction of cell-cycle position using transfer learning. Genome Biol. *23*, 41. https://doi.org/10.1186/s13059-021-02581-y.

62. Chervov, A., and Zinovyev, A. (2022). Computational challenges of cell cycle analysis using single cell transcriptomics. Preprint at arXiv, https://doi.org/10.48550/arXiv.2208.05229 https://doi.org/10.48550/arXiv.2208.05229.

63. Duò, A., Robinson, M.D., and Soneson, C. (2018). A systematic performance evaluation of clustering methods for single-cell RNA-seq data. F1000Research *7*, 1141. https://doi.org/10.12688/f1000research.15666.3.

64. Freytag, S., Tian, L., Lönnstedt, I., Ng, M., and Bahlo, M. (2018). Comparison of clustering tools in R for medium-sized 10x Genomics single-cell RNA-sequencing data. F1000Research *7*, 1297. https://doi.org/10.12688/f1000research.15809.2.

65. Traag, V.A., Waltman, L., and van Eck, N.J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. Sci. Rep. *9*, 1–12. https://doi.org/10.1038/s41598-019-41695-z.

66. Clarke, Z.A., Andrews, T.S., Atif, J., Pouyabahar, D., Innes, B.T., MacParland, S.A., and Bader, G.D. (2021). Tutorial: guidelines for annotating single-cell transcriptomic maps using automated and manual methods. Nat. Protoc. *16*, 2749–2764. https://doi.org/10.1038/s41596-021-00534-0.

67. Lotfollahi, M., Naghipourfar, M., Luecken, M.D., Khajavi, M., Büttner, M., Wagenstetter, M., Avsec, Ž., Gayoso, A., Yosef, N., Interlandi, M., et al. (2022). Mapping single-cell data to reference atlases by transfer learning. Nat. Biotechnol. *40*, 121–130. https://doi.org/10.1038/s41587-021-01001-7.

68. Kang, J.B., Nathan, A., Weinand, K., Zhang, F., Millard, N., Rumker, L., Moody, D.B., Korsunsky, I., and Raychaudhuri, S. (2021). Efficient and precise single-cell reference atlas mapping with Symphony. Nat. Commun. *12*, 5890. https://doi.org/10.1038/s41467-021-25957-x.

69. Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinforma. Oxf. Engl. *26*, 139–140. https://doi.org/10.1093/bioinformatics/btp616.

70. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550. https://doi.org/10.1186/s13059-014-0550-8.

71. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47. https://doi.org/10.1093/nar/gkv007.

72. Wang, T., Li, B., Nelson, C.E., and Nabavi, S. (2019). Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. BMC Bioinformatics *20*, 40. https://doi.org/10.1186/s12859-019-2599-6.

73. Das, S., Rai, A., Merchant, M.L., Cave, M.C., and Rai, S.N. (2021). A Comprehensive Survey of Statistical Approaches for Differential Expression Analysis in Single-Cell RNA Sequencing Studies. Genes *12*, 1947. https://doi.org/10.3390/genes12121947.

74. Soneson, C., and Robinson, M.D. (2018). Bias, robustness and scalability in single-cell differential expression analysis. Nat. Methods *15*, 255–261. https://doi.org/10.1038/nmeth.4612.

75. Squair, J.W., Gautier, M., Kathe, C., Anderson, M.A., James, N.D., Hutson, T.H., Hudelle, R., Qaiser, T., Matson, K.J.E., Barraud, Q., et al. (2021). Confronting false discoveries in single-cell differential expression. Nat. Commun. *12*, 5692. https://doi.org/10.1038/s41467-021-25960-2.

76. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. *25*, 25–29. https://doi.org/10.1038/75556.

77. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. *45*, D353–D361. https://doi.org/10.1093/nar/gkw1092.

78. Gillespie, M., Jassal, B., Stephan, R., Milacic, M., Rothfels, K., Senff-Ribeiro, A., Griss, J., Sevilla, C., Matthews, L., Gong, C., et al. (2022). The reactome pathway knowledgebase 2022. Nucleic Acids Res. *50*, D687–D692. https://doi.org/10.1093/nar/gkab1028.

79. Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., and Mesirov, J.P. (2011). Molecular signatures database (MSigDB) 3.0. Bioinforma. Oxf. Engl. *27*, 1739–1740. https://doi.org/10.1093/bioinformatics/btr260.

80. Schubert, M., Klinger, B., Klünemann, M., Sieber, A., Uhlitz, F., Sauer, S., Garnett, M.J., Blüthgen, N., and Saez-Rodriguez, J. (2018). Perturbation-response genes reveal signaling footprints in cancer gene expression. Nat. Commun. *9*, 1–11. https://doi.org/10.1038/s41467-017-02391-6.

81. Garcia-Alonso, L., Holland, C.H., Ibrahim, M.M., Turei, D., and Saez-Rodriguez, J. (2019). Benchmark and integration of resources for the estimation of human transcription factor activities. Genome Res. *29*, 1363–1375. https://doi.org/10.1101/gr.240663.118.

82. Müller-Dott, S., Tsirvouli, E., Vazquez, M., Ramirez Flores, R.O., Badia-i-Mompel, P., Fallegger, R., Türei, D., Lægreid, A., and Saez-Rodriguez, J. (2023). Expanding the coverage of regulons from high-confidence prior knowledge for accurate estimation of transcription factor activities. Nucleic Acids Res. *51*, 10934–10949. https://doi.org/10.1093/nar/gkad841.

83. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U. S. A. *102*, 15545–15550. https://doi.org/10.1073/pnas.0506580102.

84. Hänzelmann, S., Castelo, R., and Guinney, J. (2013). GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics *14*, 7. https://doi.org/10.1186/1471-2105-14-7.

85. Badia-i-Mompel, P., Vélez Santiago, J., Braunger, J., Geiss, C., Dimitrov, D., Müller-Dott, S., Taus, P., Dugourd, A., Holland, C.H., Ramirez Flores, R.O., et al. (2022). decoupleR: ensemble of computational methods to infer biological activities from omics data. Bioinforma. Adv. *2*, vbac016. https://doi.org/10.1093/bioadv/vbac016.

86. Holland, C.H., Tanevski, J., Perales-Patón, J., Gleixner, J., Kumar, M.P., Mereu, E., Joughin, B.A., Stegle, O., Lauffenburger, D.A., Heyn, H., et al. (2020). Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data. Genome Biol. *21*, 36. https://doi.org/10.1186/s13059-020-1949-z.

87. Jin, S., Guerrero-Juarez, C.F., Zhang, L., Chang, I., Ramos, R., Kuan, C.-H., Myung, P., Plikus, M.V., and Nie, Q. (2021). Inference and analysis of cell-cell communication using CellChat. Nat. Commun. *12*, 1088. https://doi.org/10.1038/s41467-021-21246-9.

88. Efremova, M., Vento-Tormo, M., Teichmann, S.A., and Vento-Tormo, R. (2020). CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. Nat. Protoc. *15*, 1484–1506. https://doi.org/10.1038/s41596-020-0292-x.

89. Cabello-Aguilar, S., Alame, M., Kon-Sun-Tack, F., Fau, C., Lacroix, M., and Colinge, J. (2020). SingleCellSignalR: inference of intercellular networks from single-cell transcriptomics. Nucleic Acids Res. *48*, e55. https://doi.org/10.1093/nar/gkaa183.

90. Dimitrov, D., Türei, D., Garrido-Rodriguez, M., Burmedi, P.L., Nagai, J.S., Boys, C., Ramirez Flores, R.O., Kim, H., Szalai, B., Costa, I.G., et al. (2022). Comparison of methods and resources for cell-cell communication inference from single-cell RNA-Seq data. Nat. Commun. *13*, 3224. https://doi.org/10.1038/s41467-022-30755-0.

91. Dimitrov, D., Schäfer, P.S.L., Farr, E., Mier, P.R., Lobentanzer, S., Dugourd, A., Tanevski, J., Flores, R.O.R., and Saez-Rodriguez, J. (2023). LIANA+: an all-in-one cell-cell communication framework. Preprint at bioRxiv, https://doi.org/10.1101/2023.08.19.553863 https://doi.org/10.1101/2023.08.19.553863.

92. Huang, R., Huang, X., Stegle, O., and Huang, Y. (2023). Robust analysis of allele-specific copy number variations from scRNA-seq data with XClone. Preprint at bioRxiv, https://doi.org/10.1101/2023.04.03.535352 https://doi.org/10.1101/2023.04.03.535352.

93. Mahdipour-Shirayeh, A., Erdmann, N., Leung-Hagesteijn, C., and Tiedemann, R.E. (2022). sciCNV: high-throughput paired profiling of transcriptomes and DNA copy number variations at single-cell resolution. Brief. Bioinform. *23*, bbab413. https://doi.org/10.1093/bib/bbab413.

94. Gao, R., Bai, S., Henderson, Y.C., Lin, Y., Schalck, A., Yan, Y., Kumar, T., Hu, M., Sei, E., Davis, A., et al. (2021). Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. Nat. Biotechnol. *39*, 599–608. https://doi.org/10.1038/s41587-020-00795-2.

95. inferCNV of the Trinity CTAT Project.

96. Momeni, K., Ghorbian, S., Ahmadpour, E., and Sharifi, R. (2023). Unraveling the complexity: understanding the deconvolutions of RNA-seq data. Transl. Med. Commun. *8*, 21. https://doi.org/10.1186/s41231-023-00154-8.

97. Avila Cobos, F., Vandesompele, J., Mestdagh, P., and De Preter, K. (2018). Computational deconvolution of transcriptomics data from mixed cell populations. Bioinformatics *34*, 1969–1979. https://doi.org/10.1093/bioinformatics/bty019.

98. Chen, B., Khodadoust, M.S., Liu, C.L., Newman, A.M., and Alizadeh, A.A. (2018). Profiling tumor infiltrating immune cells with CIBERSORT. Methods Mol. Biol. Clifton NJ *1711*, 243–259. https://doi.org/10.1007/978-1-4939-7493-1_12.

99. Wang, X., Park, J., Susztak, K., Zhang, N.R., and Li, M. (2019). Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. Nat. Commun. *10*, 380. https://doi.org/10.1038/s41467-018-08023-x.

100. 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. Nat. Commun. *11*, 5650. https://doi.org/10.1038/s41467-020-19015-1.

101. Elosua-Bayes, M., Nieto, P., Mereu, E., Gut, I., and Heyn, H. (2021). SPOTlight: seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. Nucleic Acids Res. *49*, e50. https://doi.org/10.1093/nar/gkab043.

102. Stein-O'Brien, G.L., Arora, R., Culhane, A.C., Favorov, A.V., Garmire, L.X., Greene, C.S., Goff, L.A., Li, Y., Ngom, A., Ochs, M.F., et al. (2018). Enter the Matrix: Factorization Uncovers Knowledge from Omics. Trends Genet. *34*, 790–805. https://doi.org/10.1016/j.tig.2018.07.003.

103. Hovestadt, V., Smith, K.S., Bihannic, L., Filbin, M.G., Shaw, M.L., Baumgartner, A., DeWitt, J.C., Groves, A., Mayr, L., Weisman, H.R., et al. (2019). Resolving medulloblastoma cellular architecture by single-cell genomics. Nature *572*, 74–79. https://doi.org/10.1038/s41586-019-1434-6.

104. Barkley, D., Moncada, R., Pour, M., Liberman, D.A., Dryg, I., Werba, G., Wang, W., Baron, M., Rao, A., Xia, B., et al. (2022). Cancer cell states recur across tumor types and form specific interactions with the tumor microenvironment. Nat. Genet. *54*, 1192–1201. https://doi.org/10.1038/s41588-022-01141-9.

105. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods *10*, 1213–1218. https://doi.org/10.1038/nmeth.2688.

106. Goryshin, I.Y., and Reznikoff, W.S. (1998). Tn5 in vitro transposition. J. Biol. Chem. *273*, 7367–7374. https://doi.org/10.1074/jbc.273.13.7367.

107. Chen, H., Lareau, C., Andreani, T., Vinyard, M.E., Garcia, S.P., Clement, K., Andrade-Navarro, M.A., Buenrostro, J.D., and Pinello, L. (2019). Assessment of computational methods for the analysis of single-cell ATAC-seq data. Genome Biol. *20*, 241. https://doi.org/10.1186/s13059-019-1854-5.

108. Fang, R., Preissl, S., Li, Y., Hou, X., Lucero, J., Wang, X., Motamedi, A., Shiau, A.K., Zhou, X., Xie, F., et al. (2021). Comprehensive analysis of single cell ATAC-seq data with SnapATAC. Nat. Commun. *12*, 1337. https://doi.org/10.1038/s41467-021-21583-9.

109. Stuart, T., Srivastava, A., Madad, S., Lareau, C.A., and Satija, R. (2021). Single-cell chromatin state analysis with Signac. Nat. Methods *18*, 1333–1341. https://doi.org/10.1038/s41592-021-01282-5.

110. Ou, J., Liu, H., Yu, J., Kelliher, M.A., Castilla, L.H., Lawson, N.D., and Zhu, L.J. (2018). ATACseqQC: a Bioconductor package for post-alignment quality assessment of ATAC-seq data. BMC Genomics *19*, 169. https://doi.org/10.1186/s12864-018-4559-3.

111. Amemiya, H.M., Kundaje, A., and Boyle, A.P. (2019). The ENCODE Blacklist: Identification of Problematic Regions of the Genome. Sci. Rep. *9*, 1–5. https://doi.org/10.1038/s41598-019-45839-z.

112. Thibodeau, A., Eroglu, A., McGinnis, C.S., Lawlor, N., Nehar-Belaid, D., Kursawe, R., Marches, R., Conrad, D.N., Kuchel, G.A., Gartner, Z.J., et al. (2021). AMULET: a novel read count-based method for effective multiplet detection from single nucleus ATAC-seq data. Genome Biol. *22*, 252. https://doi.org/10.1186/s13059-021-02469-x.

113. Granja, J.M., Corces, M.R., Pierce, S.E., Bagdatli, S.T., Choudhry, H., Chang, H.Y., and Greenleaf, W.J. (2021). ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. Nat. Genet. *53*, 403–411. https://doi.org/10.1038/s41588-021-00790-6.

114. Bravo González-Blas, C., Minnoye, L., Papasokrati, D., Aibar, S., Hulselmans, G., Christiaens, V., Davie, K., Wouters, J., and Aerts, S. (2019). cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. Nat. Methods *16*, 397–400. https://doi.org/10.1038/s41592-019-0367-1.

115. Martens, L.D., Fischer, D.S., Yépez, V.A., Theis, F.J., and Gagneur, J. (2024). Modeling fragment counts improves single-cell ATAC-seq analysis. Nat. Methods *21*, 28–31. https://doi.org/10.1038/s41592-023-02112-6.

116. Gontarz, P., Fu, S., Xing, X., Liu, S., Miao, B., Bazylianska, V., Sharma, A., Madden, P., Cates, K., Yoo, A., et al. (2020). Comparison of differential accessibility analysis strategies for ATAC-seq data. Sci. Rep. *10*, 10150. https://doi.org/10.1038/s41598-020-66998-4.

117. Schep, A.N., Wu, B., Buenrostro, J.D., and Greenleaf, W.J. (2017). chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. Nat. Methods *14*, 975–978. https://doi.org/10.1038/nmeth.4401.

118. Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive Integration of Single-Cell Data. Cell *177*, 1888-1902.e21. https://doi.org/10.1016/j.cell.2019.05.031.

119. Marsh, S., Salmon, M., and Hoffman, P. (2024). samuel-marsh/scCustomize: Version 2.1.2. Version v2.1.2 (Zenodo). https://doi.org/10.5281/zenodo.10724532 https://doi.org/10.5281/zenodo.10724532.

120. Bunis, D.G., Andrews, J., Fragiadakis, G.K., Burt, T.D., and Sirota, M. (2021). dittoSeq: universal user-friendly single-cell and bulk RNA sequencing visualization toolkit. Bioinformatics *36*, 5535–5536. https://doi.org/10.1093/bioinformatics/btaa1011.

121. Rue-Albrecht, K., Marini, F., Soneson, C., and Lun, A.T.L. (2018). iSEE: Interactive SummarizedExperiment Explorer. F1000Research *7*, 741. https://doi.org/10.12688/f1000research.14966.1.

122. Gonzalez-Velasco, O. (2024). LotOfCells: data visualization and statistics of single cell metadata. Preprint at bioRxiv, https://doi.org/10.1101/2024.05.23.595582 https://doi.org/10.1101/2024.05.23.595582.

123. Wu, H., Villalobos, R.G., Yao, X., Reilly, D., Chen, T., Rankin, M., Myshkin, E., Breyer, M.D., and Humphreys, B.D. (2022). Mapping the single-cell transcriptomic response of murine diabetic kidney disease to therapies. Cell Metab. *34*, 1064-1078.e6. https://doi.org/10.1016/j.cmet.2022.05.010.

124. ggplot2: Elegant Graphics for Data Analysis (3e) https://ggplot2-book.org/.

125. Keryan ArtStation. https://www.artstation.com/keryan96.

126. 10k Human PBMCs, 3’ v3.1, Chromium X 10x Genomics. https://www.10xgenomics.com/resources/datasets/10k-human-pbmcs-3-ht-v3-1-chromium-x-3-1-high.

127. Blondel, V.D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. J. Stat. Mech. Theory Exp. *2008*, 10008. https://doi.org/10.1088/1742-5468/2008/10/P10008.

128. R Packages (2e) https://r-pkgs.org/.

129. In-Line Documentation for R https://roxygen2.r-lib.org/.

130. Wickham, H. (2011). testthat: Get Started with Testing. R J. *3*, 5. https://doi.org/10.32614/RJ-2011-002.

131. Tools to Make Developing R Packages Easier https://devtools.r-lib.org/.

132. Allaire, J.J., Teague, C., Xie, Y., and Dervieux, C. (2022). Quarto. (Zenodo). https://doi.org/10.5281/zenodo.5960048 https://doi.org/10.5281/zenodo.5960048.

133. Blanco-Carmona, E. (2022). Generating publication ready visualizations for Single Cell transcriptomics using SCpubr. bioRxiv. https://doi.org/10.1101/2022.02.28.482303.

134. Blanco-Carmona, Enrique SCpubr. https://enblacar.github.io/SCpubr-book/.

135. Tirosh, I., Venteicher, A.S., Hebert, C., Escalante, L.E., Patel, A.P., Yizhak, K., Fisher, J.M., Rodman, C., Mount, C., Filbin, M.G., et al. (2016). Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. Nature *539*, 309–313. https://doi.org/10.1038/nature20123.

136. Neftel, C., Laffy, J., Filbin, M.G., Hara, T., Shore, M.E., Rahme, G.J., Richman, A.R., Silverbush, D., Shaw, M.L., Hebert, C.M., et al. (2019). An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma. Cell *178*, 835-849.e21. https://doi.org/10.1016/j.cell.2019.06.024.

137. Finetti, M.A., Grabovska, Y., Bailey, S., and Williamson, D. (2020). Translational genomics of malignant rhabdoid tumours: Current impact and future possibilities. Semin. Cancer Biol. *61*, 30–41. https://doi.org/10.1016/j.semcancer.2019.12.017.

138. Versteege, I., Sévenet, N., Lange, J., Rousseau-Merck, M.F., Ambros, P., Handgretinger, R., Aurias, A., and Delattre, O. (1998). Truncating mutations of hSNF5/INI1 in aggressive paediatric cancer. Nature *394*, 203–206. https://doi.org/10.1038/28212.

139. Ho, B., Johann, P.D., Johann, P.D., Johann, P.D., Grabovska, Y., De Dieu Andrianteranagna, M.J., De Dieu Andrianteranagna, M.J., Yao, F., Frühwald, M., Hasselblatt, M., et al. (2020). Molecular subgrouping of atypical teratoid/rhabdoid tumors - A reinvestigation and current consensus. Neuro-Oncol. *22*, 613–624. https://doi.org/10.1093/neuonc/noz235.

140. Hasselblatt, M., Isken, S., Linge, A., Eikmeier, K., Jeibmann, A., Oyen, F., Nagel, I., Richter, J., Bartelheim, K., Kordes, U., et al. (2013). High-resolution genomic analysis suggests the absence of recurrent genomic alterations other than SMARCB1 aberrations in

atypical teratoid/rhabdoid tumors. Genes. Chromosomes Cancer *52*, 185–190. https://doi.org/10.1002/gcc.22018.

141. Lee, R.S., Stewart, C., Carter, S.L., Ambrogio, L., Cibulskis, K., Sougnez, C., Lawrence, M.S., Auclair, D., Mora, J., Golub, T.R., et al. (2012). A remarkably simple genome underlies highly malignant pediatric rhabdoid cancers. J. Clin. Invest. *122*, 2983–2988. https://doi.org/10.1172/JCI64400.

142. Chi, S.N., Zimmerman, M.A., Yao, X., Cohen, K.J., Burger, P., Biegel, J.A., Rorke-Adams, L.B., Fisher, M.J., Janss, A., Mazewski, C., et al. (2009). Intensive multimodality treatment for children with newly diagnosed CNS atypical teratoid rhabdoid tumor. J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol. *27*, 385–389. https://doi.org/10.1200/JCO.2008.18.7724.

143. Hasselblatt, M., Gesk, S., Oyen, F., Rossi, S., Viscardi, E., Giangaspero, F., Giannini, C., Judkins, A.R., Frühwald, M.C., Obser, T., et al. (2011). Nonsense mutation and inactivation of SMARCA4 (BRG1) in an atypical teratoid/rhabdoid tumor showing retained SMARCB1 (INI1) expression. Am. J. Surg. Pathol. *35*, 933–935. https://doi.org/10.1097/PAS.0b013e3182196a39.

144. Johann, P.D., Erkek, S., Zapatka, M., Kerl, K., Buchhalter, I., Hovestadt, V., Jones, D.T.W., Sturm, D., Hermann, C., Segura Wang, M., et al. (2016). Atypical Teratoid/Rhabdoid Tumors Are Comprised of Three Epigenetic Subgroups with Distinct Enhancer Landscapes. Cancer Cell *29*, 379–393. https://doi.org/10.1016/j.ccell.2016.02.001.

145. Roberts, C.W., Galusha, S.A., McMenamin, M.E., Fletcher, C.D., and Orkin, S.H. (2000). Haploinsufficiency of Snf5 (integrase interactor 1) predisposes to malignant rhabdoid tumors in mice. Proc. Natl. Acad. Sci. U. S. A. *97*, 13796–13800. https://doi.org/10.1073/pnas.250492697.

146. Wang, X., Lee, R.S., Alver, B.H., Haswell, J.R., Wang, S., Mieczkowski, J., Drier, Y., Gillespie, S.M., Archer, T.C., Wu, J.N., et al. (2017). SMARCB1-mediated SWI/SNF complex function is essential for enhancer regulation. Nat. Genet. *49*, 289–295. https://doi.org/10.1038/ng.3746.

147. Alver, B.H., Kim, K.H., Lu, P., Wang, X., Manchester, H.E., Wang, W., Haswell, J.R., Park, P.J., and Roberts, C.W.M. (2017). The SWI/SNF chromatin remodelling complex is required for maintenance of lineage specific enhancers. Nat. Commun. *8*, 14648. https://doi.org/10.1038/ncomms14648.

148. Mittal, P., and Roberts, C.W.M. (2020). The SWI/SNF complex in cancer - biology, biomarkers and therapy. Nat. Rev. Clin. Oncol. *17*, 435–448. https://doi.org/10.1038/s41571-020-0357-3.

149. Kwon, H., Imbalzano, A.N., Khavari, P.A., Kingston, R.E., and Green, M.R. (1994). Nucleosome disruption and enhancement of activator binding by a human SW1/SNF complex. Nature *370*, 477–481. https://doi.org/10.1038/370477a0.

150. Mashtalir, N., D'Avino, A.R., Michel, B.C., Luo, J., Pan, J., Otto, J.E., Zullow, H.J., McKenzie, Z.M., Kubiak, R.L., St Pierre, R., et al. (2018). Modular Organization and Assembly

of SWI/SNF Family Chromatin Remodeling Complexes. Cell *175*, 1272-1288.e20. https://doi.org/10.1016/j.cell.2018.09.032.

151. Nesvick, C.L., Lafay-Cousin, L., Raghunathan, A., Bouffet, E., Huang, A.A., and Daniels, D.J. (2020). Atypical teratoid rhabdoid tumor: molecular insights and translation to novel therapeutics. J. Neurooncol. *150*, 47–56. https://doi.org/10.1007/s11060-020-03639-w.

152. Shain, A.H., and Pollack, J.R. (2013). The spectrum of SWI/SNF mutations, ubiquitous in human cancers. PloS One *8*, e55119. https://doi.org/10.1371/journal.pone.0055119.

153. Brennan, B., Stiller, C., and Bourdeaut, F. (2013). Extracranial rhabdoid tumours: what we have learned so far and future directions. Lancet Oncol. *14*, e329-336. https://doi.org/10.1016/S1470-2045(13)70088-3.

154. Erkek, S., Johann, P.D., Finetti, M.A., Drosos, Y., Chou, H.-C., Zapatka, M., Sturm, D., Jones, D.T.W., Korshunov, A., Rhyzova, M., et al. (2019). Comprehensive Analysis of Chromatin States in Atypical Teratoid/Rhabdoid Tumor Identifies Diverging Roles for SWI/SNF and Polycomb in Gene Regulation. Cancer Cell *35*, 95-110.e8. https://doi.org/10.1016/j.ccell.2018.11.014.

155. Alimova, I., Birks, D.K., Harris, P.S., Knipstein, J.A., Venkataraman, S., Marquez, V.E., Foreman, Nicholas K., and Vibhakar, R. (2013). Inhibition of EZH2 suppresses self-renewal and induces radiation sensitivity in atypical rhabdoid teratoid tumor cells. Neuro-Oncol. *15*, 149–160. https://doi.org/10.1093/neuonc/nos285.

156. Kadoch, C., Hargreaves, D.C., Hodges, C., Elias, L., Ho, L., Ranish, J., and Crabtree, G.R. (2013). Proteomic and bioinformatic analysis of mammalian SWI/SNF complexes identifies extensive roles in human malignancy. Nat. Genet. *45*, 592–601. https://doi.org/10.1038/ng.2628.

157. Moreno, N., and Kerl, K. (2016). Preclinical Evaluation of Combined Targeted Approaches in Malignant Rhabdoid Tumors. Anticancer Res. *36*, 3883–3887.

158. Nakayama, R.T., Pulice, J.L., Valencia, A.M., McBride, M.J., McKenzie, Z.M., Gillespie, M.A., Ku, W.L., Teng, M., Cui, K., Williams, R.T., et al. (2017). SMARCB1 is required for widespread BAF complex-mediated activation of enhancers and bivalent promoters. Nat. Genet. *49*, 1613–1623. https://doi.org/10.1038/ng.3958.

159. Chun, H.J.E., Johann, P.D., Milne, K., Zapatka, M., Buellesbach, A., Ishaque, N., Iskar, M., Erkek, S., Wei, L., Tessier-Cloutier, B., et al. (2019). Identification and Analyses of Extra-Cranial and Cranial Rhabdoid Tumor Molecular Subgroups Reveal Tumors with Cytotoxic T Cell Infiltration. Cell Rep. *29*, 2338-2354.e7. https://doi.org/10.1016/j.celrep.2019.10.013.

160. Holdhof, D., Johann, P.D., Spohn, M., Bockmayr, M., Safaei, S., Joshi, P., Masliah-Planchon, J., Ho, B., Andrianteranagna, M., Bourdeaut, F., et al. (2021). Atypical teratoid/rhabdoid tumors (ATRTs) with SMARCA4 mutation are molecularly distinct from SMARCB1-deficient cases. Acta Neuropathol. (Berl.) *141*, 291–301. https://doi.org/10.1007/s00401-020-02250-7.

161. Tief, K., Schmidt, A., Aguzzi, A., and Beermann, F. (1996). Tyrosinase is a new marker for cell populations in the mouse neural tube. Dev. Dyn. *205*, 445–456. https://doi.org/10.1002/(SICI)1097-0177(199604)205:4<445::AID-AJA8>3.0.CO;2-I.

162. Johann, P.D., Hovestadt, V., Thomas, C., Jeibmann, A., Heß, K., Bens, S., Oyen, F., Hawkins, C., Pierson, C.R., Aldape, K., et al. (2017). Cribriform neuroepithelial tumor: molecular characterization of a SMARCB1-deficient non-rhabdoid tumor with favorable long-term outcome. Brain Pathol. *27*, 411–418. https://doi.org/10.1111/bpa.12413.

163. Torchia, J., Golbourn, B., Feng, S., Ho, K.C., Sin-Chan, P., Vasiljevic, A., Norman, J.D., Guilhamon, P., Garzia, L., Agamez, N.R., et al. (2016). Integrated (epi)-Genomic Analyses Identify Subgroup-Specific Therapeutic Targets in CNS Rhabdoid Tumors. Cancer Cell *30*, 891–908. https://doi.org/10.1016/j.ccell.2016.11.003.

164. Frühwald, M.C., Hasselblatt, M., Nemes, K., Bens, S., Steinbügl, M., Johann, P.D., Kerl, K., Hauser, P., Quiroga, E., Solano-Paez, P., et al. (2020). Age and DNA methylation subgroup as potential independent risk factors for treatment stratification in children with atypical teratoid/rhabdoid tumors. Neuro-Oncol. *22*, 1006–1017. https://doi.org/10.1093/neuonc/noz244.

165. Bookhout, C., Bouldin, T.W., and Ellison, D.W. (2018). Atypical teratoid/rhabdoid tumor with retained INI1 (SMARCB1) expression and loss of BRG1 (SMARCA4). Neuropathol. Off. J. Jpn. Soc. Neuropathol. *38*, 305–308. https://doi.org/10.1111/neup.12452.

166. Federico, A., Thomas, C., Miskiewicz, K., Woltering, N., Zin, F., Nemes, K., Bison, B., Johann, P.D., Hawes, D., Bens, S., et al. (2022). ATRT–SHH comprises three molecular subgroups with characteristic clinical and histopathological features and prognostic significance. Acta Neuropathol. (Berl.) *143*, 697–711. https://doi.org/10.1007/s00401-022-02424-5.

167. Jagani, Z., Mora-Blanco, E.L., Sansam, C.G., McKenna, E.S., Wilson, B., Chen, D., Klekota, J., Tamayo, P., Nguyen, P.T.L., Tolstorukov, M., et al. (2010). Loss of the tumor suppressor Snf5 leads to aberrant activation of the Hedgehog-Gli pathway. Nat. Med. *16*, 1429–1434. https://doi.org/10.1038/nm.2251.

168. Gao, F., Shi, L., Russin, J., Zeng, L., Chang, X., He, S., Chen, T.C., Giannotta, S.L., Weisenberger, D.J., Zada, G., et al. (2013). DNA methylation in the malignant transformation of meningiomas. PloS One *8*, e54114. https://doi.org/10.1371/journal.pone.0054114.

169. Notaro, S., Reimer, D., Fiegl, H., Schmid, G., Wiedemair, A., Rössler, J., Marth, C., and Zeimet, A.G. (2016). Evaluation of folate receptor 1 (FOLR1) mRNA expression, its specific promoter methylation and global DNA hypomethylation in type I and type II ovarian cancers. BMC Cancer *16*, 589. https://doi.org/10.1186/s12885-016-2637-y.

170. Sive, J.I., Feber, A., Smith, D., Quinn, J., Beck, S., and Yong, K. (2016). Global hypomethylation in myeloma is associated with poor prognosis. Br. J. Haematol. *172*, 473–475. https://doi.org/10.1111/bjh.13506.

171. Madakashira, B.P., and Sadler, K.C. (2017). DNA Methylation, Nuclear Organization, and Cancer. Front. Genet. *8*, 76. https://doi.org/10.3389/fgene.2017.00076.

172. Van Tongelen, A., Loriot, A., and De Smet, C. (2017). Oncogenic roles of DNA hypomethylation through the activation of cancer-germline genes. Cancer Lett. *396*, 130–137. https://doi.org/10.1016/j.canlet.2017.03.029.

173. Nakada, M., Hayashi, Y., and Hamada, J. (2011). Role of Eph/ephrin tyrosine kinase in malignant glioma. Neuro-Oncol. *13*, 1163–1170. https://doi.org/10.1093/neuonc/nor102.

174. Giorgio, C., Zanotti, I., Lodola, A., and Tognolini, M. (2020). Ephrin or not? Six tough questions on Eph targeting. Expert Opin. Ther. Targets *24*, 403–415. https://doi.org/10.1080/14728222.2020.1745187.

175. Jain, R., Jain, D., Liu, Q., Bartosinska, B., Wang, J., Schumann, D., Kauschke, S.G., Eickelmann, P., Piemonti, L., Gray, N.S., et al. (2013). Pharmacological inhibition of Eph receptors enhances glucose-stimulated insulin secretion from mouse and human pancreatic islets. Diabetologia *56*, 1350–1355. https://doi.org/10.1007/s00125-013-2877-1.

176. Wang, X., Zhang, M., Ping, F., Liu, H., Sun, J., Wang, Y., Shen, A., Ding, J., and Geng, M. (2019). Identification and Therapeutic Intervention of Coactivated Anaplastic Lymphoma Kinase, Fibroblast Growth Factor Receptor 2, and Ephrin Type-A Receptor 5 Kinases in Hepatocellular Carcinoma. Hepatol. Baltim. Md *69*, 573–586. https://doi.org/10.1002/hep.29792.

177. Allen, J.C., Judkins, A.R., Rosenblum, M.K., and Biegel, J.A. (2006). Atypical teratoid/rhabdoid tumor evolving from an optic pathway ganglioglioma: Case study. Neuro-Oncol. *8*, 79–82.

178. Bertrand, A., Rondenet, C., Masliah-Planchon, J., Leblond, P., de la Fourchardière, A., Pissaloux, D., Aït-Raïs, K., Lequin, D., Jouvet, A., Freneaux, P., et al. (2018). Rhabdoid component emerging as a subclonal evolution of paediatric glioneuronal tumours. Neuropathol. Appl. Neurobiol. *44*, 224–228. https://doi.org/10.1111/nan.12379.

179. Bozzai, B., Hasselblatt, M., Turányi, E., Frühwald, M.C., Siebert, R., Bens, S., Schneppenheim, R., Kool, M., Stelczer, G., Hortobágyi, T., et al. (2017). Atypical teratoid/rhabdoid tumor arising in a malignant glioma. Pediatr. Blood Cancer *64*, 96–99. https://doi.org/10.1002/pbc.26173.

180. Nobusawa, S., Hirato, J., Sugai, T., Okura, N., Yamazaki, T., Yamada, S., Ikota, H., Nakazato, Y., and Yokoo, H. (2016). Atypical Teratoid/Rhabdoid Tumor (AT/RT) Arising From Ependymoma: A Type of AT/RT Secondarily Developing From Other Primary Central Nervous System Tumors. J. Neuropathol. Exp. Neurol. *75*, 167–174. https://doi.org/10.1093/jnen/nlv017.

181. Ginn, K.F., and Gajjar, A. (2012). Atypical Teratoid Rhabdoid Tumor: Current Therapy and Future Directions. Front. Oncol. *2*. https://doi.org/10.3389/fonc.2012.00114.

182. Nemes, K., Johann, P.D., Steinbügl, M., Gruhle, M., Bens, S., Kachanov, D., Teleshova, M., Hauser, P., Simon, T., Tippelt, S., et al. (2022). Infants and Newborns with Atypical Teratoid Rhabdoid Tumors (ATRT) and Extracranial Malignant Rhabdoid Tumors (eMRT) in the EU-RHAB Registry: A Unique and Challenging Population. Cancers *14*, 2185. https://doi.org/10.3390/cancers14092185.

183. Gastberger, K., Fincke, V.E., Mucha, M., Siebert, R., Hasselblatt, M., and Frühwald, M.C. (2023). Current Molecular and Clinical Landscape of ATRT – The Link to Future Therapies. Cancer Manag. Res. *15*, 1369–1393. https://doi.org/10.2147/CMAR.S379451.

184. Lange, M., Bergen, V., Klein, M., Setty, M., Reuter, B., Bakhti, M., Lickert, H., Ansari, M., Schniering, J., Schiller, H.B., et al. (2022). CellRank for directed single-cell fate mapping. Nat. Methods *19*, 159–170. https://doi.org/10.1038/s41592-021-01346-6.

185. Drokhlyansky, E., Smillie, C.S., Wittenberghe, N.V., Ericsson, M., Griffin, G.K., Dionne, D., Cuoco, M.S., Goder-Reiser, M.N., Sharova, T., Aguirre, A.J., et al. (2019). The enteric nervous system of the human and mouse colon at a single-cell resolution. Preprint at bioRxiv, https://doi.org/10.1101/746743 https://doi.org/10.1101/746743.

186. What is Cell Ranger? -Software -Single Cell Gene Expression -Official 10x Genomics Support https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/what-is-cell-ranger.

187. Hao, Y., Stuart, T., Kowalski, M.H., Choudhary, S., Hoffman, P., Hartman, A., Srivastava, A., Molla, G., Madad, S., Fernandez-Granda, C., et al. (2023). Dictionary learning for integrative, multimodal and scalable single-cell analysis. Nat. Biotechnol., 1–12. https://doi.org/10.1038/s41587-023-01767-y.

188. Wolock, S.L., Lopez, R., and Klein, A.M. (2019). Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. Cell Syst. *8*, 281-291.e9. https://doi.org/10.1016/j.cels.2018.11.005.

189. Blanco-Carmona, E., Narayanan, A., Hernandez, I., Nieto, J.C., Elosua-Bayes, M., Sun, X., Schmidt, C., Pamir, N., Özduman, K., Herold-Mende, C., et al. (2023). Tumor heterogeneity and tumor-microglia interactions in primary and recurrent IDH1-mutant gliomas. Cell Rep. Med., 101249. https://doi.org/10.1016/j.xcrm.2023.101249.

190. Andreatta, M., and Carmona, S.J. (2021). UCell: Robust and scalable single-cell gene signature scoring. Comput. Struct. Biotechnol. J. *19*, 3796–3798. https://doi.org/10.1016/j.csbj.2021.06.043.

191. Franzén, O., Gan, L.-M., and Björkegren, J.L.M. (2019). PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. Database *2019*. https://doi.org/10.1093/database/baz046.

192. Braun, E., Danan-Gotthold, M., Borm, L.E., Vinsland, E., Lee, K.W., Lönnerberg, P., Hu, L., Li, X., He, X., Andrusivová, Ž., et al. (2022). Comprehensive cell atlas of the first-trimester developing human brain. Preprint at bioRxiv, https://doi.org/10.1101/2022.10.24.513487 https://doi.org/10.1101/2022.10.24.513487.

193. Dai, M., Pei, X., and Wang, X.-J. (2022). Accurate and fast cell marker gene identification with COSG. Brief. Bioinform. *23*, bbab579. https://doi.org/10.1093/bib/bbab579.

194. Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A.H., Tanaseichuk, O., Benner, C., and Chanda, S.K. (2019). Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat. Commun. *10*, 1523. https://doi.org/10.1038/s41467-019-09234-6.

195. Jawaid, W. (2021). An R interface to the Enrichr database.

196. Xenium Panels 10x Genomics. https://www.10xgenomics.com/products/xenium-panels.

197. Tang, M., and Verhaak, R.G. (2016). A Molecular Take on Malignant Rhabdoid Tumors. Trends Cancer *2*, 217–218. https://doi.org/10.1016/j.trecan.2016.04.003.

198. Graf, M., Interlandi, M., Moreno, N., Holdhof, D., Göbel, C., Melcher, V., Mertins, J., Albert, T.K., Kastrati, D., Alfert, A., et al. (2022). Single-cell transcriptomics identifies potential cells of origin of MYC rhabdoid tumors. Nat. Commun. *13*, 1544. https://doi.org/10.1038/s41467-022-29152-4.

199. Duronio, R.J., and Xiong, Y. (2013). Signaling Pathways that Control Cell Proliferation. Cold Spring Harb. Perspect. Biol. *5*, a008904. https://doi.org/10.1101/cshperspect.a008904.

200. Torchia, J., Picard, D., Lafay-Cousin, L., Hawkins, C.E., Kim, S.K., Letourneau, L., Ra, Y.S., Ho, K.C., Chan, T.S.Y., Sin-Chan, P., et al. (2015). Molecular subgroups of atypical teratoid rhabdoid tumours in children: An integrated genomic and clinicopathological analysis. Lancet Oncol. *16*, 569–582. https://doi.org/10.1016/S1470-2045(15)70114-2.

201. Johann, P.D., Erkek, S., Zapatka, M., Kerl, K., Buchhalter, I., Hovestadt, V., Jones, D.T.W., Sturm, D., Hermann, C., Segura Wang, M., et al. (2016). Atypical Teratoid/Rhabdoid Tumors Are Comprised of Three Epigenetic Subgroups with Distinct Enhancer Landscapes. Cancer Cell *29*, 379–393. https://doi.org/10.1016/j.ccell.2016.02.001.

202. Lobón-Iglesias, M.-J., Andrianteranagna, M., Han, Z.-Y., Chauvin, C., Masliah-Planchon, J., Manriquez, V., Tauziede-Espariat, A., Turczynski, S., Bouarich-Bourimi, R., Frah, M., et al. (2023). Imaging and multi-omics datasets converge to define different neural progenitor origins for ATRT-SHH subgroups. Nat. Commun. *14*, 6669. https://doi.org/10.1038/s41467-023-42371-7.

203. Ostrom, Q.T., Patil, N., Cioffi, G., Waite, K., Kruchko, C., and Barnholtz-Sloan, J.S. (2020). CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2013-2017. Neuro-Oncol. *22*, iv1–iv96. https://doi.org/10.1093/neuonc/noaa200.

204. Balss, J., Meyer, J., Mueller, W., Korshunov, A., Hartmann, C., and von Deimling, A. (2008). Analysis of the IDH1 codon 132 mutation in brain tumors. Acta Neuropathol. (Berl.) *116*, 597–602. https://doi.org/10.1007/s00401-008-0455-2.

205. Hartmann, C., Meyer, J., Balss, J., Capper, D., Mueller, W., Christians, A., Felsberg, J., Wolter, M., Mawrin, C., Wick, W., et al. (2009). Type and frequency of IDH1 and IDH2 mutations are related to astrocytic and oligodendroglial differentiation and age: a study of 1,010 diffuse gliomas. Acta Neuropathol. (Berl.) *118*, 469–474. https://doi.org/10.1007/s00401-009-0561-9.

206. Yan, H., Parsons, D.W., Jin, G., McLendon, R., Rasheed, B.A., Yuan, W., Kos, I., Batinic-Haberle, I., Jones, S., Riggins, G.J., et al. (2009). IDH1 and IDH2 mutations in gliomas. N. Engl. J. Med. *360*, 765–773. https://doi.org/10.1056/NEJMoa0808710.

207. Dang, L., White, D.W., Gross, S., Bennett, B.D., Bittinger, M.A., Driggers, E.M., Fantin, V.R., Jang, H.G., Jin, S., Keenan, M.C., et al. (2009). Cancer-associated IDH1 mutations produce 2-hydroxyglutarate. Nature *462*, 739–744. https://doi.org/10.1038/nature08617.

208. Koivunen, P., Lee, S., Duncan, C.G., Lopez, G., Lu, G., Ramkissoon, S., Losman, J.A., Joensuu, P., Bergmann, U., Gross, S., et al. (2012). Transformation by the (R)-enantiomer of 2-hydroxyglutarate linked to EGLN activation. Nature *483*, 484–488. https://doi.org/10.1038/nature10898.

209. Zagzag, D., Zhong, H., Scalzitti, J.M., Laughner, E., Simons, J.W., and Semenza, G.L. (2000). Expression of hypoxia-inducible factor 1alpha in brain tumors: association with angiogenesis, invasion, and progression. Cancer *88*, 2606–2618.

210. Chowdhury, R., Yeoh, K.K., Tian, Y.-M., Hillringhaus, L., Bagg, E.A., Rose, N.R., Leung, I.K.H., Li, X.S., Woon, E.C.Y., Yang, M., et al. (2011). The oncometabolite 2-hydroxyglutarate inhibits histone lysine demethylases. EMBO Rep. *12*, 463–469. https://doi.org/10.1038/embor.2011.43.

211. Xu, W., Yang, H., Liu, Y., Yang, Y., Wang, P., Kim, S.-H., Ito, S., Yang, C., Wang, P., Xiao, M.-T., et al. (2011). Oncometabolite 2-hydroxyglutarate is a competitive inhibitor of α-ketoglutarate-dependent dioxygenases. Cancer Cell *19*, 17–30. https://doi.org/10.1016/j.ccr.2010.12.014.

212. Turcan, S., Rohle, D., Goenka, A., Walsh, L.A., Fang, F., Yilmaz, E., Campos, C., Fabius, A.W.M., Lu, C., Ward, P.S., et al. (2012). IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype. Nature *483*, 479–483. https://doi.org/10.1038/nature10866.

213. Lu, C., Ward, P.S., Kapoor, G.S., Rohle, D., Turcan, S., Abdel-Wahab, O., Edwards, C.R., Khanin, R., Figueroa, M.E., Melnick, A., et al. (2012). IDH mutation impairs histone demethylation and results in a block to cell differentiation. Nature *483*, 474–478. https://doi.org/10.1038/nature10860.

214. Venteicher, A.S., Tirosh, I., Hebert, C., Yizhak, K., Neftel, C., Filbin, M.G., Hovestadt, V., Escalante, L.E., Shaw, M.L., Rodman, C., et al. (2017). Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq. Science *355*, eaai8478. https://doi.org/10.1126/science.aai8478.

215. de Souza, C.F., Sabedot, T.S., Malta, T.M., Stetson, L., Morozova, O., Sokolov, A., Laird, P.W., Wiznerowicz, M., Iavarone, A., Snyder, J., et al. (2018). A Distinct DNA Methylation

Shift in a Subset of Glioma CpG Island Methylator Phenotypes during Tumor Recurrence. Cell Rep. *23*, 637–651. https://doi.org/10.1016/j.celrep.2018.03.107.

216. Ceccarelli, M., Barthel, F.P., Malta, T.M., Sabedot, T.S., Salama, S.R., Murray, B.A., Morozova, O., Newton, Y., Radenbaugh, A., Pagnotta, S.M., et al. (2016). Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma. Cell *164*, 550–563. https://doi.org/10.1016/j.cell.2015.12.028.

217. Miller, J.J., Gonzalez Castro, L.N., McBrayer, S., Weller, M., Cloughesy, T., Portnow, J., Andronesi, O., Barnholtz-Sloan, J.S., Baumert, B.G., Berger, M.S., et al. (2023). Isocitrate dehydrogenase (IDH) mutant gliomas: A Society for Neuro-Oncology (SNO) consensus review on diagnosis, management, and future directions. Neuro-Oncol. *25*, 4–25. https://doi.org/10.1093/neuonc/noac207.

218. Flavahan, W.A., Drier, Y., Liau, B.B., Gillespie, S.M., Venteicher, A.S., Stemmer-Rachamimov, A.O., Suvà, M.L., and Bernstein, B.E. (2016). Insulator dysfunction and oncogene activation in IDH mutant gliomas. Nature *529*, 110–114. https://doi.org/10.1038/nature16490.

219. Nakamura, M., Watanabe, T., Yonekawa, Y., Kleihues, P., and Ohgaki, H. (2001). Promoter methylation of the DNA repair gene MGMT in astrocytomas is frequently associated with G:C --> A:T mutations of the TP53 tumor suppressor gene. Carcinogenesis *22*, 1715–1719. https://doi.org/10.1093/carcin/22.10.1715.

220. Esteller, M., Garcia-Foncillas, J., Andion, E., Goodman, S.N., Hidalgo, O.F., Vanaclocha, V., Baylin, S.B., and Herman, J.G. (2000). Inactivation of the DNA-repair gene MGMT and the clinical response of gliomas to alkylating agents. N. Engl. J. Med. *343*, 1350–1354. https://doi.org/10.1056/NEJM200011093431901.

221. Brat, D.J., Verhaak, R.G.W., Aldape, K.D., Yung, W.K.A., Salama, S.R., Cooper, L.A.D., Rheinbay, E., Miller, C.R., Vitucci, M., Morozova, O., et al. (2015). Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. N. Engl. J. Med. *372*, 2481–2498. https://doi.org/10.1056/NEJMoa1402121.

222. Reuss, D.E., Mamatjan, Y., Schrimpf, D., Capper, D., Hovestadt, V., Kratz, A., Sahm, F., Koelsche, C., Korshunov, A., Olar, A., et al. (2015). IDH mutant diffuse and anaplastic astrocytomas have similar age at presentation and little difference in survival: a grading problem for WHO. Acta Neuropathol. (Berl.) *129*, 867–873. https://doi.org/10.1007/s00401-015-1438-8.

223. Jiao, Y., Killela, P.J., Reitman, Z.J., Rasheed, A.B., Heaphy, C.M., de Wilde, R.F., Rodriguez, F.J., Rosemberg, S., Oba-Shinjo, S.M., Nagahashi Marie, S.K., et al. (2012). Frequent ATRX, CIC, FUBP1 and IDH1 mutations refine the classification of malignant gliomas. Oncotarget *3*, 709–722. https://doi.org/10.18632/oncotarget.588.

224. Heaphy, C.M., de Wilde, R.F., Jiao, Y., Klein, A.P., Edil, B.H., Shi, C., Bettegowda, C., Rodriguez, F.J., Eberhart, C.G., Hebbar, S., et al. (2011). Altered telomeres in tumors with ATRX and DAXX mutations. Science *333*, 425. https://doi.org/10.1126/science.1207313.

225. Conte, D., Huh, M., Goodall, E., Delorme, M., Parks, R.J., and Picketts, D.J. (2012). Loss of Atrx sensitizes cells to DNA damaging agents through p53-mediated death pathways. PloS One *7*, e52167. https://doi.org/10.1371/journal.pone.0052167.

226. Jenkins, R.B., Blair, H., Ballman, K.V., Giannini, C., Arusell, R.M., Law, M., Flynn, H., Passe, S., Felten, S., Brown, P.D., et al. (2006). A t(1;19)(q10;p10) mediates the combined deletions of 1p and 19q and predicts a better prognosis of patients with oligodendroglioma. Cancer Res. *66*, 9852–9861. https://doi.org/10.1158/0008-5472.CAN-06-1796.

227. Ostrom, Q.T., Cioffi, G., Gittleman, H., Patil, N., Waite, K., Kruchko, C., and Barnholtz-Sloan, J.S. (2019). CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2012-2016. Neuro-Oncol. *21*, v1–v100. https://doi.org/10.1093/neuonc/noz150.

228. Arita, H., Narita, Y., Fukushima, S., Tateishi, K., Matsushita, Y., Yoshida, A., Miyakita, Y., Ohno, M., Collins, V.P., Kawahara, N., et al. (2013). Upregulating mutations in the TERT promoter commonly occur in adult malignant gliomas and are strongly associated with total 1p19q loss. Acta Neuropathol. (Berl.) *126*, 267–276. https://doi.org/10.1007/s00401-013-1141-6.

229. Killela, P.J., Reitman, Z.J., Jiao, Y., Bettegowda, C., Agrawal, N., Diaz, L.A., Friedman, A.H., Friedman, H., Gallia, G.L., Giovanella, B.C., et al. (2013). TERT promoter mutations occur frequently in gliomas and a subset of tumors derived from cells with low rates of self-renewal. Proc. Natl. Acad. Sci. U. S. A. *110*, 6021–6026. https://doi.org/10.1073/pnas.1303607110.

230. Bettegowda, C., Agrawal, N., Jiao, Y., Sausen, M., Wood, L.D., Hruban, R.H., Rodriguez, F.J., Cahill, D.P., McLendon, R., Riggins, G., et al. (2011). Mutations in CIC and FUBP1 contribute to human oligodendroglioma. Science *333*, 1453–1455. https://doi.org/10.1126/science.1210557.

231. Hafezi, F., and Perez Bercoff, D. (2020). The Solo Play of TERT Promoter Mutations. Cells *9*, 749. https://doi.org/10.3390/cells9030749.

232. Suzuki, H., Aoki, K., Chiba, K., Sato, Y., Shiozawa, Y., Shiraishi, Y., Shimamura, T., Niida, A., Motomura, K., Ohka, F., et al. (2015). Mutational landscape and clonal architecture in grade II and III gliomas. Nat. Genet. *47*, 458–468. https://doi.org/10.1038/ng.3273.

233. Bell, R.J.A., Rube, H.T., Kreig, A., Mancini, A., Fouse, S.D., Nagarajan, R.P., Choi, S., Hong, C., He, D., Pekmezci, M., et al. (2015). Cancer. The transcription factor GABP selectively binds and activates the mutant TERT promoter in cancer. Science *348*, 1036–1039. https://doi.org/10.1126/science.aab0015.

234. Wong, D., and Yip, S. (2020). Making heads or tails - the emergence of capicua (CIC) as an important multifunctional tumour suppressor. J. Pathol. *250*, 532–540. https://doi.org/10.1002/path.5400.

235. Ahmad, S.T., Rogers, A.D., Chen, M.J., Dixit, R., Adnani, L., Frankiw, L.S., Lawn, S.O., Blough, M.D., Alshehri, M., Wu, W., et al. (2019). Capicua regulates neural stem cell proliferation and lineage specification through control of Ets factors. Nat. Commun. *10*, 2000. https://doi.org/10.1038/s41467-019-09949-6.

236. Yang, R., Chen, L.H., Hansen, L.J., Carpenter, A.B., Moure, C.J., Liu, H., Pirozzi, C.J., Diplas, B.H., Waitkus, M.S., Greer, P.K., et al. (2017). Cic Loss Promotes Gliomagenesis via Aberrant Neural Stem Cell Proliferation and Differentiation. Cancer Res. *77*, 6097–6108. https://doi.org/10.1158/0008-5472.CAN-17-1018.

237. Chittaranjan, S., Chan, S., Yang, C., Yang, K.C., Chen, V., Moradian, A., Firme, M., Song, J., Go, N.E., Blough, M.D., et al. (2014). Mutations in CIC and IDH1 cooperatively regulate 2-hydroxyglutarate levels and cell clonogenicity. Oncotarget *5*, 7960–7979.

238. Rabenhorst, U., Thalheimer, F.B., Gerlach, K., Kijonka, M., Böhm, S., Krause, D.S., Vauti, F., Arnold, H.-H., Schroeder, T., Schnütgen, F., et al. (2015). Single-Stranded DNA-Binding Transcriptional Regulator FUBP1 Is Essential for Fetal and Adult Hematopoietic Stem Cell Self-Renewal. Cell Rep. *11*, 1847–1855. https://doi.org/10.1016/j.celrep.2015.05.038.

239. Elman, J.S., Ni, T.K., Mengwasser, K.E., Jin, D., Wronski, A., Elledge, S.J., and Kuperwasser, C. (2019). Identification of FUBP1 as a Long Tail Cancer Driver and Widespread Regulator of Tumor Suppressor and Oncogene Alternative Splicing. Cell Rep. *28*, 3435-3449.e5. https://doi.org/10.1016/j.celrep.2019.08.060.

240. Chan, A.K.-Y., Pang, J.C.-S., Chung, N.Y.-F., Li, K.K.-W., Poon, W.S., Chan, D.T.-M., Shi, Z., Chen, L., Zhou, L., and Ng, H.-K. (2014). Loss of CIC and FUBP1 expressions are potential markers of shorter time to recurrence in oligodendroglial tumors. Mod. Pathol. Off. J. U. S. Can. Acad. Pathol. Inc *27*, 332–342. https://doi.org/10.1038/modpathol.2013.165.

241. Zlatescu, M.C., TehraniYazdi, A., Sasaki, H., Megyesi, J.F., Betensky, R.A., Louis, D.N., and Cairncross, J.G. (2001). Tumor location and growth pattern correlate with genetic signature in oligodendroglial neoplasms. Cancer Res. *61*, 6713–6715.

242. Stockhammer, F., Misch, M., Helms, H.-J., Lengler, U., Prall, F., von Deimling, A., and Hartmann, C. (2012). IDH1/2 mutations in WHO grade II astrocytomas associated with localization and seizure as the initial symptom. Seizure *21*, 194–197. https://doi.org/10.1016/j.seizure.2011.12.007.

243. Zong, H., Parada, L.F., and Baker, S.J. (2015). Cell of origin for malignant gliomas and its implication in therapeutic development. Cold Spring Harb. Perspect. Biol. *7*, a020610. https://doi.org/10.1101/cshperspect.a020610.

244. Klemm, F., Maas, R.R., Bowman, R.L., Kornete, M., Soukup, K., Nassiri, S., Brouland, J.-P., Iacobuzio-Donahue, C.A., Brennan, C., Tabar, V., et al. (2020). Interrogation of the Microenvironmental Landscape in Brain Tumors Reveals Disease-Specific Alterations of Immune Cells. Cell *181*, 1643-1660.e17. https://doi.org/10.1016/j.cell.2020.05.007.

245. Friedrich, M., Sankowski, R., Bunse, L., Kilian, M., Green, E., Ramallo Guevara, C., Pusch, S., Poschet, G., Sanghvi, K., Hahn, M., et al. (2021). Tryptophan metabolism drives dynamic

immunosuppressive myeloid states in IDH-mutant gliomas. Nat. Cancer *2*, 723–740. https://doi.org/10.1038/s43018-021-00201-z.

246. Creating a Reference Package with cellranger mkref -Software -Single Cell Gene Expression -Official 10x Genomics Support https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/3.1/advanced/references?src=website&lss=blog&cnm=none&cid=NULL.

247. Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., et al. (2021). Integrated analysis of multimodal single-cell data. Cell *184*, 3573-3587.e29. https://doi.org/10.1016/j.cell.2021.04.048.

248. scrublet/examples/scrublet_basics.ipynb at master · swolock/scrublet GitHub. https://github.com/swolock/scrublet/blob/master/examples/scrublet_basics.ipynb.

249. SCTransform on multiple batches · Issue #55 · satijalab/sctransform GitHub. https://github.com/satijalab/sctransform/issues/55.

250. R, G., and C, S. (2010). A flexible R package for nonnegative matrix factorization. BMC Bioinformatics *11*. https://doi.org/10.1186/1471-2105-11-367.

251. Benjamini, Y., and Hochberg, Y. (1995). Controlling The False Discovery Rate - A Practical And Powerful Approach To Multiple Testing. J R. Stat. Soc Ser. B *57*, 289–300. https://doi.org/10.2307/2346101.

252. Venteicher, A.S., Tirosh, I., Hebert, C., Yizhak, K., Neftel, C., Filbin, M.G., Hovestadt, V., Escalante, L.E., Shaw, M.L., Rodman, C., et al. (2017). Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq. Science *355*, eaai8478. https://doi.org/10.1126/science.aai8478.

253. Angerer, P., Haghverdi, L., Büttner, M., Theis, F.J., Marr, C., and Buettner, F. (2016). destiny: diffusion maps for large-scale single-cell data in R. Bioinformatics *32*, 1241–1243. https://doi.org/10.1093/bioinformatics/btv715.

254. EnsDb.Hsapiens.v86 Bioconductor. http://bioconductor.org/packages/EnsDb.Hsapiens.v86/.

255. Ruiz-Moreno, C., Salas, S.M., Samuelsson, E., Brandner, S., Kranendonk, M.E.G., Nilsson, M., and Stunnenberg, H.G. (2022). Harmonized single-cell landscape, intercellular crosstalk and tumor architecture of glioblastoma. Preprint at bioRxiv, https://doi.org/10.1101/2022.08.27.505439 https://doi.org/10.1101/2022.08.27.505439.

256. Cao, J., O'Day, D.R., Pliner, H.A., Kingsley, P.D., Deng, M., Daza, R.M., Zager, M.A., Aldinger, K.A., Blecher-Gonen, R., Zhang, F., et al. (2020). A human cell atlas of fetal gene expression. Science *370*, eaba7721. https://doi.org/10.1126/science.aba7721.

257. Wierstra, I., and Alves, J. (2007). FOXM1, a typical proliferation-associated transcription factor. Biol. Chem. *388*, 1257–1274. https://doi.org/10.1515/BC.2007.159.

258. Hambardzumyan, D., Gutmann, D.H., and Kettenmann, H. (2016). The role of microglia and macrophages in glioma maintenance and progression. Nat. Neurosci. *19*, 20–27. https://doi.org/10.1038/nn.4185.

259. Pombo Antunes, A.R., Scheyltjens, I., Lodi, F., Messiaen, J., Antoranz, A., Duerinck, J., Kancheva, D., Martens, L., De Vlaminck, K., Van Hove, H., et al. (2021). Single-cell profiling of myeloid cells in glioblastoma across species and disease stage reveals macrophage competition and specialization. Nat. Neurosci. *24*, 595–610. https://doi.org/10.1038/s41593-020-00789-y.

260. Paolicelli, R.C., Sierra, A., Stevens, B., Tremblay, M.-E., Aguzzi, A., Ajami, B., Amit, I., Audinat, E., Bechmann, I., Bennett, M., et al. (2022). Microglia states and nomenclature: A field at its crossroads. Neuron *110*, 3458–3483. https://doi.org/10.1016/j.neuron.2022.10.020.

261. Gerganova, G., Riddell, A., and Miller, A.A. (2022). CNS border-associated macrophages in the homeostatic and ischaemic brain. Pharmacol. Ther. *240*, 108220. https://doi.org/10.1016/j.pharmthera.2022.108220.

262. Sun, R., and Jiang, H. (2024). Border-associated macrophages in the central nervous system. J. Neuroinflammation *21*, 67. https://doi.org/10.1186/s12974-024-03059-x.

263. Morrissey, M.A., Kern, N., and Vale, R.D. (2020). CD47 Ligation Repositions the Inhibitory Receptor SIRPA to Suppress Integrin Activation and Phagocytosis. Immunity *53*, 290-302.e6. https://doi.org/10.1016/j.immuni.2020.07.008.

264. Larionova, T.D., Bastola, S., Aksinina, T.E., Anufrieva, K.S., Wang, J., Shender, V.O., Andreev, D.E., Kovalenko, T.F., Arapidi, G.P., Shnaider, P.V., et al. (2022). Alternative RNA splicing modulates ribosomal composition and determines the spatial phenotype of glioblastoma cells. Nat. Cell Biol. *24*, 1541–1557. https://doi.org/10.1038/s41556-022-00994-w.

265. Li, M., Yang, L., Chan, A.K.N., Pokharel, S.P., Liu, Q., Mattson, N., Xu, X., Chang, W.-H., Miyashita, K., Singh, P., et al. (2023). Epigenetic Control of Translation Checkpoint and Tumor Progression via RUVBL1-EEF1A1 Axis. Adv. Sci. Weinh. Baden-Wurtt. Ger. *10*, e2206584. https://doi.org/10.1002/advs.202206584.

266. Park, J.-W., Sahm, F., Steffl, B., Arrillaga-Romany, I., Cahill, D., Monje, M., Herold-Mende, C., Wick, W., and Turcan, Ş. (2021). TERT and DNMT1 expression predict sensitivity to decitabine in gliomas. Neuro-Oncol. *23*, 76–87. https://doi.org/10.1093/neuonc/noaa207.

267. Turcan, S., Fabius, A.W.M., Borodovsky, A., Pedraza, A., Brennan, C., Huse, J., Viale, A., Riggins, G.J., and Chan, T.A. (2013). Efficient induction of differentiation and growth inhibition in IDH1 mutant glioma cells by the DNMT Inhibitor Decitabine. Oncotarget *4*, 1729–1736. https://doi.org/10.18632/oncotarget.1412.

268. Morral, C., Stanisavljevic, J., Hernando-Momblona, X., Mereu, E., Álvarez-Varela, A., Cortina, C., Stork, D., Slebe, F., Turon, G., Whissell, G., et al. (2020). Zonation of Ribosomal

DNA Transcription Defines a Stem Cell Hierarchy in Colorectal Cancer. Cell Stem Cell *26*, 845-861.e12. https://doi.org/10.1016/j.stem.2020.04.012.

269. GLASS Consortium (2018). Glioma through the looking GLASS: molecular evolution of diffuse gliomas and the Glioma Longitudinal Analysis Consortium. Neuro-Oncol. *20*, 873–884. https://doi.org/10.1093/neuonc/noy020.

270. Barthel, F.P., Johnson, K.C., Varn, F.S., Moskalik, A.D., Tanner, G., Kocakavuk, E., Anderson, K.J., Abiola, O., Aldape, K., Alfaro, K.D., et al. (2019). Longitudinal molecular trajectories of diffuse glioma in adults. Nature *576*, 112–120. https://doi.org/10.1038/s41586-019-1775-1.

271. Varn, F.S., Johnson, K.C., Martinek, J., Huse, J.T., Nasrallah, M.P., Wesseling, P., Cooper, L.A.D., Malta, T.M., Wade, T.E., Sabedot, T.S., et al. (2022). Glioma progression is shaped by genetic evolution and microenvironment interactions. Cell *185*, 2184-2199.e16. https://doi.org/10.1016/j.cell.2022.04.038.

272. Paull, E.O., Aytes, A., Jones, S.J., Subramaniam, P.S., Giorgi, F.M., Douglass, E.F., Tagore, S., Chu, B., Vasciaveo, A., Zheng, S., et al. (2021). A modular master regulator landscape controls cancer transcriptional identity. Cell *184*, 334-351.e20. https://doi.org/10.1016/j.cell.2020.11.045.

273. Friebel, E., Kapolou, K., Unger, S., Núñez, N.G., Utz, S., Rushing, E.J., Regli, L., Weller, M., Greter, M., Tugues, S., et al. (2020). Single-Cell Mapping of Human Brain Cancer Reveals Tumor-Specific Instruction of Tissue-Invading Leukocytes. Cell *181*, 1626-1642.e20. https://doi.org/10.1016/j.cell.2020.04.055.

274. Butturini, E., Boriero, D., Carcereri de Prati, A., and Mariotto, S. (2019). STAT1 drives M1 microglia activation and neuroinflammation under hypoxia. Arch. Biochem. Biophys. *669*, 22–30. https://doi.org/10.1016/j.abb.2019.05.011.

275. Hu, X., Herrero, C., Li, W.-P., Antoniv, T.T., Falck-Pedersen, E., Koch, A.E., Woods, J.M., Haines, G.K., and Ivashkiv, L.B. (2002). Sensitization of IFN-gamma Jak-STAT signaling during macrophage activation. Nat. Immunol. *3*, 859–866. https://doi.org/10.1038/ni828.

276. Smith, G.T., Radin, D.P., and Tsirka, S.E. (2022). From protein-protein interactions to immune modulation: Therapeutic prospects of targeting Neuropilin-1 in high-grade glioma. Front. Immunol. *13*, 958620. https://doi.org/10.3389/fimmu.2022.958620.

277. Snuderl, M., Batista, A., Kirkpatrick, N.D., Ruiz de Almodovar, C., Riedemann, L., Walsh, E.C., Anolik, R., Huang, Y., Martin, J.D., Kamoun, W., et al. (2013). Targeting placental growth factor/neuropilin 1 pathway inhibits growth and spread of medulloblastoma. Cell *152*, 1065–1076. https://doi.org/10.1016/j.cell.2013.01.036.