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

M.Sc. Ling Hai

born in Sanya, Hainan, China

Oral examination: 23.02.2024

Characterization of tumor subpopulations in glioblastoma with single cell transcriptomics

Referees:

Prof. Dr. Benedikt Brors

Prof. Dr. Matthias Schlesner

To Chuantian Du

Acknowledgements

I would like to express my deepest thanks to my supervisor, Prof. Dr. Matthias Schlesner, who provided me with the opportunity to pursue a PhD in his group—the lovely BODA group. I am incredibly thankful for his scientific suggestions, thoughtful advice, and the freedom to conduct research, all of which have illuminated my PhD journey.

I want to thank Prof. Dr. Benedikt Brors for being my faculty supervisor and a member of my thesis advisory committee (TAC). I greatly appreciate his valuable feedback and for organizing the “MMK Tuesday Seminar”. I also want to thank the other members of my TAC, Prof. Dr. Oliver Stegle, and Dr. Carl Herrmann, for their feedback. Additionally, I am grateful to Prof. Dr. Peter Angel and Dr. Annarita Patrizi for their participation in my defense committee.

My gratitude belongs to my wonderful collaborators on the “GB Connectivity Signature” project, especially PD. Dr. Tobias Kessler and Dirk Hoffmann. I greatly enjoy working with them! Tobias is always available to help me. I also want to thank Prof. Dr. Wolfgang Wick for providing me with the opportunity to continue my work in his remarkable group — the Clinical Cooperation Unit Neurooncology — and as part of the “UNITE Glioblastoma” research program.

I would like to thank my amazing colleagues and friends for the enjoyable moments we shared together in Heidelberg. I feel so lucky to have met them!

I want to express my special thanks to my husband, Chuantian Du, for his unwavering love, support, and encouragement. I want to thank my son, Feng Du, for bringing immense fun and excitement to my life, taking me to a new stage of happiness. I also want to thank my father, Mingqiang Hai, my mother, Jianghong Sun, and my family in-law for their support and concern across the sea.

Abstract

Tumors are complex tissues with substantial intra-tumor heterogeneity, intricately linked to tumor progression and therapeutic resistance. Emerging single-cell RNA sequencing (scRNA-Seq) technologies empower researchers to elucidate these diverse tumor subpopulations. This thesis presents the characterization of a distinct glioblastoma (GB) cell population and introduces a novel bioinformatics tool designed to quantify similarity among cell populations.

Cell-to-cell connectivity through tumor microtubes (TMs) has been discovered among glioma tumor cells, conferring self-repair capabilities, augmenting therapy resistance, and driving tumor progression. Yet, a comprehensive molecular understanding and precise quantification of this connectivity have remained elusive. This study delves into the transcriptomic landscape of the highly connected glioma cell population using scRNA-Seq and RNA-Seq. I found that these highly connected cells exhibited a notable predominance of astrocyte-like (AC) and mesenchymal-like (MES) cell states, while lowly connected cells were characterized by a prevalence of neuronal progenitor-like (NPC) cell states. I established a 71-gene connectivity signature by comparing highly and lowly connected cells. A connectivity signature score (CSS) was developed based on the relative average expression levels of the connectivity signature. This CSS was then applied to several GB patient tumor scRNA-Seq and RNA-Seq datasets, consistently revealing higher CSS values for AC and MES cell states compared to NPC cell states. Furthermore, correlations were observed between CSS values and mesenchymal expression subtypes as well as between CSS values and the mutation status of NF1, PTEN, and TP53. One key finding is that higher CSS values were linked to poorer patient survival. Additionally, CHI3L1 — one of the connectivity signature genes — was identified as a robust marker for cell connectivity and a potential prognostic marker for GB patients. Investigating CHI3L1 overexpression RNA-Seq and proteomics datasets revealed that CHI3L1 upregulated multiple cell state markers and elevated CSS values. Notably, CHI3L1 overexpression also led to increased phosphorylation of the TM-connectivity marker GAP43.

In this thesis, I present a new bioinformatic tool named Interactive Explorer of Single-Cell Cluster Similarity (ieCS). This tool serves to link similar cell populations that share the same biological cell types/states across various donors or experimental conditions. ieCS utilizes an innovative metric to quantify similarity between cell populations. ieCS offers three distinct methods for identifying superclusters comprising similar cell

populations. Featuring a user-friendly graphical interface, ieCS enables interactive and intuitive visualization of these superclusters. In a demonstration dataset, ieCS accurately, robustly, and quickly identified superclusters across various experimental conditions.

In conclusion, this thesis characterizes the highly connected GB cell population and introduces a bioinformatics tool for mapping similar cell populations.

Zusammenfassung

Tumoren sind komplexe Gewebe mit erheblicher Heterogenität innerhalb des Tumors, die eng mit dem Fortschreiten des Tumors und der Therapieresistenz verbunden sind. Neue Technologien zur Einzelzell-RNA-Sequenzierung (scRNA-Seq) ermöglichen es Forschern, diese verschiedenen Tumorsubpopulationen aufzuklären. Diese Arbeit präsentiert die Charakterisierung einer bestimmten Glioblastom (GB)-Zellpopulation und stellt ein neuartiges Bioinformatik-Tool vor, das zur Quantifizierung der Ähnlichkeit zwischen Zellpopulationen entwickelt wurde.

Bei Gliomtumorzellen wurde eine Zell-zu-Zell-Konnektivität durch Tumormikroröhren (TMs) entdeckt, die Selbstreparaturfähigkeiten verleiht, die Therapieresistenz erhöht und das Fortschreiten des Tumors vorantreibt. Ein umfassendes molekulares Verständnis und eine genaue Quantifizierung dieser Konnektivität sind jedoch noch nicht möglich. Diese Studie befasst sich mit der transkriptomischen Landschaft der stark vernetzten Gliomzellpopulation mithilfe von scRNA-Seq und RNA-Seq. Ich fand heraus, dass diese stark verbundenen Zellen ein bemerkenswertes Vorherrschen von astrozytenähnlichen (AC) und mesenchymalen (MES) Zellzuständen aufwiesen, während schwach verbundene Zellen durch eine Prävalenz neuronaler Vorläufer-ähnlicher (NPC) Zellzustände gekennzeichnet waren. Ich habe eine 71-Gen-Konnektivitätssignatur erstellt, indem ich hoch und niedrig verbundene Zellen verglichen habe. Basierend auf den relativen durchschnittlichen Ausdrucksniveaus der Konnektivitätssignatur wurde ein Connectivity Signature Score (CSS) entwickelt. Dieses CSS wurde dann auf mehrere scRNA-Seq- und RNA-Seq-Datensätze von Tumortumoren in GB angewendet und ergab durchweg höhere CSS-Werte für AC- und MES-Zellzustände im Vergleich zu NPC-Zellzuständen. Darüber hinaus wurden Korrelationen zwischen CSS-Werten und mesenchymalen Expressionssubtypen sowie zwischen CSS-Werten und dem Mutationsstatus von NF1, PTEN und TP53 beobachtet. Eine wichtige Erkenntnis ist, dass höhere CSS-Werte mit einem schlechteren Patientenüberleben verbunden waren. Darüber hinaus wurde CHI3L1 – eines der Konnektivitätssignaturgene – als robuster Marker für die Zellkonnektivität und potenzieller prognostischer Marker für GB-Patienten identifiziert. Die Untersuchung der RNA-Seq- und Proteomics-Datensätze zur Überexpression von CHI3L1 ergab, dass CHI3L1 mehrere Zellzustandsmarker hochregulierte und die CSS-Werte erhöhte. Bemerkenswerterweise führte die Überexpression von CHI3L1 auch zu einer erhöhten Phosphorylierung des TM-Konnektivitätsmarkers GAP43.

In dieser Arbeit stelle ich ein neues bioinformatisches Tool namens Interactive Explorer of Single-Cell Cluster Similarity (ieCS) vor. Dieses Tool dient dazu, ähnliche Zellpopulationen zu verknüpfen, die über verschiedene Spender oder Versuchsbedingungen hinweg dieselben biologischen Zelltypen/-zustände aufweisen. ieCS nutzt eine innovative Metrik, um die Ähnlichkeit zwischen Zellpopulationen zu quantifizieren. ieCS bietet drei verschiedene Methoden zur Identifizierung von Superclustern mit ähnlichen Zellpopulationen. Mit einer benutzerfreundlichen grafischen Oberfläche ermöglicht ieCS eine interaktive und intuitive Visualisierung dieser Supercluster. In einem Demonstrationsdatensatz identifizierte ieCS Supercluster unter verschiedenen experimentellen Bedingungen präzise, zuverlässig und schnell.

Zusammenfassend charakterisiert diese Arbeit die stark vernetzte GB-Zellpopulation und stellt ein Bioinformatik-Tool zur Kartierung ähnlicher Zellpopulationen vor.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Full Name
2D	two-dimensional
AC	astrocytic-like tumor cell
CCU	Clinical Cooperation Unit
cDNA	complementary deoxyribonucleic acid
CL	classical expression subtype
CNV	copy number variation
Coxph	Cox proportional hazards regression survival analysis
CSS	connectivity signature score
DEG	differentially expressed gene
DEP	differentially expressed protein
DHC	directly hierarchical clustering
DKFZ	German Cancer Research Center
DPP	differential phosphorated protein
EANO	European association of neurooncology
eQTL	expression quantitative trait loci
FACS	fluorescence-activated cell sorting
FDR	false discovery rate
FPKM	fragments per kilobase million
GB	glioblastoma
GBmap	harmonized GB scRNA-Seq dataset
GBMSC	glioblastoma stem-like cells
GFP	green fluorescent protein
GHC	globally hierarchical clustering
GLASS	Glioma Longitudinal Analysis Cohort
GO	gene ontology
GTEx	Genotype-Tissue Expression
GUI	graphic user interface
ieCS	interactive explorer of single cell cluster similarity
KM	Kaplan-Meier survival analysis
MES	mesenchymal-like tumor cell
MPLSM	multiphoton laser scanning microscopy
MS	mesenchymal expression subtype
MSC	mesenchymal stem cells
NGS	next-generation sequencing
NPC	neuronal progenitor-like tumor cell
NPV	negative predictive value
OE	overexpression
OPC	oligodendrocyte progenitor-like cell
OS	overall survival
PC	principal component
PCA	principal component analysis
PCR	polymerase chain reaction
PDGCL	patient derived glioma cell line
PN	proneural expression subtype
PPV	positive predictive value
PVN	perivascular niche
Q1	the first quartile

Q2	the second quartile
Q3	the third quartile
Q4	the last quartile
RNA	ribonucleic acid
RNA-Seq	RNA sequencing
RTK	Receptor tyrosine kinase
scRNA-Seq	single cell RNA sequencing
SNN	shared nearest neighbor
SR101	sulforhodamine 101
TCGA	The Cancer Genome Atlas
TM	tumor microtube
TPM	transcripts per kilobase million
UMAP	uniform manifold approximation and projection
UMI	unique molecular identifier
WHO	world health organization
wt	wild type

1. Introduction

The text was written by Ling Hai. It has been proofread and edited by ChatGPT.

1.1 Transcriptomics

Transcriptomics is the study of eukaryotic transcriptomes, which represent the full set of ribonucleic acid (RNA) transcripts in a cell or tissue. The transcriptome provides a snapshot of cellular processes, enabling comparisons between different experimental conditions, diseases, tissues, species, times, or spaces. This sheds light on fundamental principles of gene expression in organisms. The applications of transcriptomics in biological research include:

- Biomarker discovery: By comparing transcriptomes between disease and non-disease samples, one can identify disease-associated gene signatures, gain insights into disease pathogenesis, and discover new biomarkers. This approach has been applied in various contexts, such as identifying transcriptomic biomarkers in cardiovascular disease (Pedrotty et al., 2012), prognostic biomarkers in chronic kidney disease (Ju et al., 2015), progressive and prognostic biomarker in melanoma (Raskin et al., 2013), carcinogenic biomarkers in lung cancer (Billatos et al., 2018), and specific biomarkers in cancers of unknown primary (Wei et al., 2014).
- Comparative transcriptomics: Comparative transcriptomics in mouse and human helps researchers understand when mouse models accurately represent human biology and what factors need consideration for optimizing the mouse model (Breschi et al., 2017). Comparative transcriptomics across distinct species, such as worms, flies and humans can reveal conserved co-expression modules (Gerstein et al., 2014).
- Studying cellular differentiation: Characterizing transcriptomes of stem cells and monitoring transcriptomic changes during cell differentiation improve the understanding of cell potency and regulation factors in different stages of differentiation. Examples include identifying functional features of mouse stem cells and early embryos (Sharov et al., 2003), stage-specific gene expression changes during human mammary cell commitment (Raouf et al., 2008), stage-specific gene transcription and alternative splicing during neural differentiation (Wu et al., 2010), regulators in the transition from embryonic stem cell to definitive endoderm (Chu et al., 2016), and endoderm differentiation-

associated expression quantitative trait loci (eQTL) and predictive markers of differentiation efficiency (Cuomo et al., 2020).

- Studying transcriptional regulation: Integrated analysis of gene expression and its regulatory elements, such as transcription factors, cofactors, enhancers, noncoding RNAs, and chromatin state, reveals transcriptional regulation mechanisms. Researchers have quantified both gene expression and transcription factor binding signals (Cheng et al., 2012), performed integrated analysis of transcriptome and chromatin accessibility in the same cell during neurogenesis (Chen S. et al., 2019), and integrated analysis of transcriptome, chromatin accessibility and surface marker abundance to identify specific regulatory features in mixed-phenotype acute leukemia (Granja et al., 2019).

1.1.1 RNA sequencing

RNA sequencing (RNA-Seq) is a high-throughput next-generation sequencing (NGS) technique used to quantify entire transcriptomes at the resolution of single bases (Wang Z. et al., 2009). In classical RNA-Seq experiments, the first step involves converting RNAs into complementary deoxyribonucleic acid (cDNA). Next, adaptors are added to the cDNA, and the resulting short cDNA sequences (usually 30-400 bp) are sequenced using NGS (Wang Z. et al., 2009).

The computational analysis of RNA-Seq data typically involves several steps, including quality control (Andrews et al., 2010), alignment (Dobin et al., 2013), gene expression quantification (Liao et al., 2014), differential expression analysis across experimental conditions (Love et al., 2014), identification of alternative splicing (Li et al., 2018), detection of gene fusions (Uhrig et al., 2021), eQTL mapping (Shabalin, 2012), and visualization (Gu et al., 2016). Best practices of RNA-Seq analysis have been reviewed by Conesa et al., 2016.

1.1.2 Single cell RNA sequencing

The swift advancement of sequencing techniques has propelled us into a new era of single-cell RNA sequencing (scRNA-Seq), allowing for the quantification of gene expression in individual cells rather than in bulk populations. This provides unprecedented opportunities for researchers to study various aspects, including the behavior of individual cells (Tang F. et al., 2009), cell differentiation processes (Cuomo

et al., 2020), the cell type composition of tissues (Darmanis et al., 2015), cellular response of specific subpopulations (Park et al., 2020), tumor heterogeneity (Neftel et al., 2019).

The typical steps involved in scRNA-Seq experiments are briefly described here, using protocols from full-length Smart-Seq2 (Picelli et al., 2014) and UMI-based 10x Genomics Chromium Single Cell 3' Solution (10x Genomics, 2020) as examples:

1) Capture single cells:
    - Smart-Seq2: Fluorescence-activated cell sorting (FACS) to sort cells into plates.
    - 10x Genomics Chromium: Cells are partitioned in a microfluidic platform.
2) Reverse transcription:
    - Smart-Seq2: Betaine (to improve cDNA yield) and oligo(dT) primer are used.
    - 10x Genomics Chromium: A primer containing read 1 sequencing primer, cell barcode, unique molecular identifier (UMI, which quantifies unique RNA molecules and reduces amplification bias), and oligo(dT) primer.
3) Amplify cDNA:
    - Both Smart-Seq2 and 10x Genomics Chromium use polymerase chain reaction (PCR).
4) Construct library:
    - Smart-Seq2: Tn5 transposase is used for tagmentation, followed by the addition of sample index, P5 Illumina primer and P7 Illumina primer.
    - 10x Genomics Chromium: Enzyme fragmentation is performed, and then the read 2 sequencing primer, sample index, P5 Illumina primer and P7 Illumina primer are added.
5) Sequencing:
    - Both Smart-Seq2 and 10x Genomics Chromium use paired-end NGS sequencing.
6) Transcript data:
    - Smart-Seq2: Yields hundreds of cells with full-length coverage and approximately 1000,000 reads per cell.
    - 10x Genomics Chromium: Captures 500-10,000 cells per sample with

3’-end coverage and a minimum of 20,000 reads per cell.

Some steps of the computational analysis for full-length scRNA-Seq data, like Smart-Seq2, can use the same tools originally designed for traditional RNA-Seq, such as read alignment, gene expression quantification and normalization. However, scRNA-Seq data presents additional concerns at the quality control step, where low-quality cells with degraded RNA need to be identified. For UMI-based 3’-end counting scRNA-Seq data, such as 10x Genomics Chromium, the normalization step requires adjustments (Stegle et al., 2015). Figure 1.1 illustrates the typical analysis workflow for scRNA-Seq data.

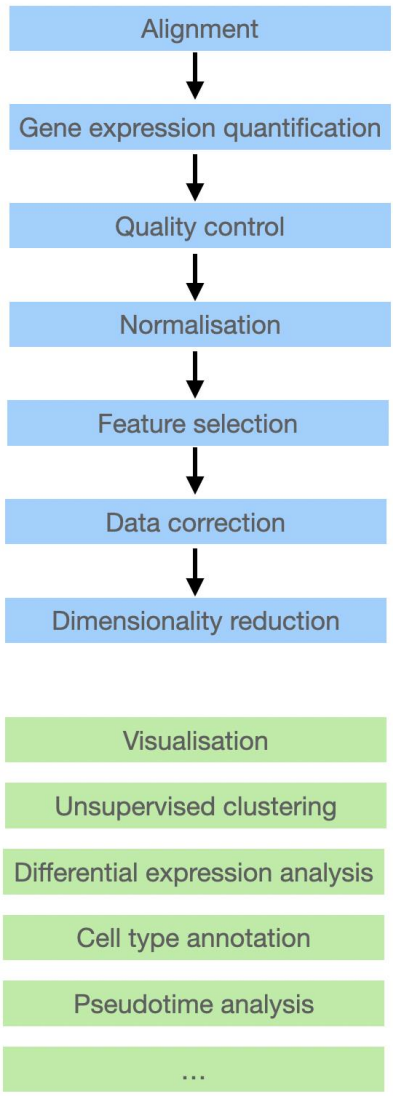


Figure 1.1 **The typical analysis workflow for scRNA-Seq data.** Blue boxes represent steps from the upstream analysis while green boxes represent steps in the downstream analysis.

The number of bioinformatics tools for scRNA-Seq data analysis has expanded remarkably, comprising approximately 1600 tools in 30 categories for various analysis tasks (https://www.scrna-tools.org/, November 2023). These tools cover alignment and expression quantification (Chen W. et al., 2020), quality control with doublet detection (Xi and Li, 2021), normalization (Chen W. et al., 2020), data correction (Chen W. et al., 2020; Tran et al., 2020), dimensionality reduction with principal component analysis (Tsuyuzaki et al., 2020), unsupervised clustering (Duò et al., 2020; Krzak et al., 2019), cell type annotation (Abdelaal et al., 2019), pseudotime analysis/trajectory analysis (Saelens et al., 2019), differential expression analysis (Wang T. et al., 2019), pathway analysis (Holland et al., 2020), visualization (Cakir et al., 2020), and more.

There are also convenient toolkits/pipelines for scRNA-Seq, such as Seurat (Stuart et al., 2019) in R, SCANPY (Wolf et al., 2018) in python and web-based analysis platforms (Gardeux et al., 2017; Zhu et al., 2017). Comprehensive benchmarking studies for pipelines and various tool combinations have been published (Chen W. et al., 2020; Tian et al., 2019; Vieth et al., 2019). A framework for benchmarking is also available (Germain et al., 2020). Best practices in scRNA-Seq analysis have been reviewed by Luecken and Theis, 2019.

The development of tools for scRNA-Seq data analysis is thriving, yet several challenges persist (Lähnemann et al., 2020; Poirion et al., 2016; Stegle et al., 2015). One grand challenge is to link cell subpopulations across different donors or experimental conditions. To tackle this, reseachers have employed various strategies:

- Integrating datasets: This approach involves identifying mutual nearest neighbors (MNN, Haghverdi et al., 2018), using “anchor” integration in Seurat (Stuart et al., 2019), employing iterative clustering in Harmony (Korsunsky et al., 2019) or adopting integrative non-negative matrix factorization in LIGER (Welch et al., 2019). The outputs are corrected/normalized matrices.
- Classifying nearest neighbors: This approach involves searching for nearest neighbors on unsupervised selected features using scmap (Kiselev et al., 2018) or mapping cells to a reference dataset based on correlation scores in SingleR (Aran et al., 2019). The outputs are cell labels with scores.
- Quantifying cell subpopulation similarity based on markers: This approach involves hierarchical clustering of binary-transformed markers in ClusterMap (Gao et al., 2019).

1.1.3 Tumor heterogeneity and scRNA-Seq

Tumors, as complex “tissues”, exhibit not only inter-tumor heterogeneity, which describes differences among patients with the same tumor type, but also intra-tumor heterogeneity, which describes variations within tumor cells in a single tumor. Intra-tumor heterogeneity encompasses morpho-histological differences, genomic instability (e.g., gene mutations and copy number alterations), epigenetic plasticity (e.g., DNA methylation and histone modification), and microenvironment interactions (Stanta and Bonin, 2018). It plays a crucial role in tumor progression, therapeutic resistance, and recurrences (Stanta and Bonin, 2018).

Various techniques have emerged to study tumor heterogeneity, such as multi-region sampling, autopsy sampling, longitudinal analysis with liquid biopsy, and single cell sequencing (Dagogo-Jack and Shaw, 2018). Among these, scRNA-Seq empowers researchers to dissect intra-tumor heterogeneity, enabling the study of diverse tumor microenvironments, multiple genomic subclones of tumor cells, and different physiological states of tumor cells (Levitin et al., 2018). Furthermore, scRNA-Seq can facilitate the identification and classification of tumor subpopulations with distinct transcriptional signatures, the detection of cancer stem cells, and the discovery of treatment-resistant tumor subpopulations (González-Silva et al., 2020).

1.2 Glioblastoma

Gliomas are the primary brain tumors, representing about 30% of all brain tumors and 80% of malignant brain tumors (Weller et al., 2015). Glioblastoma (GB) is the most common glioma, accounting for approximately 45% of all gliomas (Ostrom et al., 2014). GB is an aggressive tumor with a median survival of 15 months in treated patients (Koshy et al., 2012; Tamimi and Juweid, 2017).

1.2.1 Classification of glioma

Traditionally, gliomas were classified by histology into astrocytic, oligodendroglial or ependymal tumors (Weller et al., 2015). However, the rapid development of genetic profiling technologies has significantly improved the classification and treatment of

gliomas. Based on genetic features, gliomas can now be classified as follows (Weller et al., 2015):

- Isocitrate dehydrogenases (IDH) mutant and 1p/19q co-deleted tumors: These tumors with oligodendroglial morphology have the best outcome.
- IDH mutant and 1p/19q non-co-deleted tumors: These tumors with astrocytic morphology have intermediate outcome.
- IDH wild-type (wt) tumors: These tumors have the worst outcome.

In 2016, the World Health Organization (WHO) integrated both histological and genetic features to diagnose gliomas, resulting in the following classifications (Louis et al., 2016):

- Astrocytoma, oligoastrocytoma and oligodendroglioma based on histological appearance, further classified by molecular features, including IDH mutant and 1p/19q co-deleted oligodendroglioma and IDH mutant diffuse astrocytoma and IDH wt diffuse astrocytoma.
- Glioblastoma based on histological appearance, further classified as IDH mutant GB and IDH wt GB.

In 2021, the European Association of Neurooncology (EANO) suggested several molecular markers for glioma diagnosis (Weller et al., 2021):

- IDH1 R132 or IDH2 R172 mutation: Distinguishes IDH mutant gliomas from IDH wt GB and other IDH wt gliomas.
- 1p/19q codeletion: Distinguishes IDH mutant oligodendroglioma from IDH mutant astrocytoma.
- Histone H3K27M mutation: Defines diffuse midline glioma.
- Histone H3.3 G43R/V mutation: Defines diffuse hemispheric glioma.

1.2.2 Classification of glioblastoma

The 2016 WHO classification divides GB into two main subtypes: IDH wt GB and IDH mutant GB (Louis et al., 2016). Approximately 90% of GB cases are IDH wt GB, while around 10% are IDH mutant GB (Louis et al., 2016). IDH wt GB has a higher median age of diagnosis and worse overall survival (OS) compared to IDH mutant GB (62

years vs. 44 years; 15 months vs. 31 months; Louis et al., 2016).

In the current EANO guidelines, the classification of gliomas based on integrated histomolecular features refers to IDH mutant GB as IDH mutant astrocytoma (Weller et al., 2021). Within IDH wt GB, a DNA methylation-based classifier further groups GB into seven methylation subtypes: RTKI, RTKII, RTKIII, MES, MID, MYCN and G34 (Capper et al., 2018). Bulk RNA-Seq datasets of IDH wt GB has defined three expression subtypes (mesenchymal, classical and proneural), which have different immune microenvironments and survival times (Wang Q. et al., 2017).

Recent GB scRNA-Seq data have revealed several neurodevelopmental lineages and cellular states in GB tumors (Couturier et al., 2020; Neftel et al., 2019). The majority of GB tumors contain cells in four main cellular states (Neftel et al., 2019):

- Astrocyte-like (AC-like) cells express astrocytic markers (e.g., GFAP).
- Oligodendrocyte progenitor-like (OPC-like) cells express oligodendroglial lineage markers (e.g., OLIG1).
- Mesenchymal-like cells (MES-like) cells express mesenchymal genes (e.g., VIM). Within the MES-like state, there are two distinct expression meta-modules:
    a. MES1 is considered with hypoxia-independent and is characterized by high expression levels of specific genes, including CHI3L1, CD44 and APOE.
    b. MES2 is characterized by high expression levels of genes associated with hypoxia, such as HILPDA, as well as genes involved in glycolysis, such as ENO2 and LDHA.
- Neuronal progenitor-like (NPC-like) cells have high expression levels of neural stem cell markers (e.g., SOX4, SOX11, DCX and CD24). Within the NPC-like state, there are two distinct expression meta-modules:
    a. NPC1 has the potential to differentiate into OPC and characterized by high expression level of DLL1 and DLL3.
    b. NPC2 has the potential to differentiate into into neurons and characterized by high expressions of DLX5-AS1 and DLX6-AS1.

The proportions of these cell states vary between tumors and are influenced by genetic alterations (Neftel et al., 2019). Furthermore, single GB cells exhibit plasticity and can

transition between different cell states (Neftel et al., 2019). Figure 1.2 illustrates the four main cell states found in GB tumors.

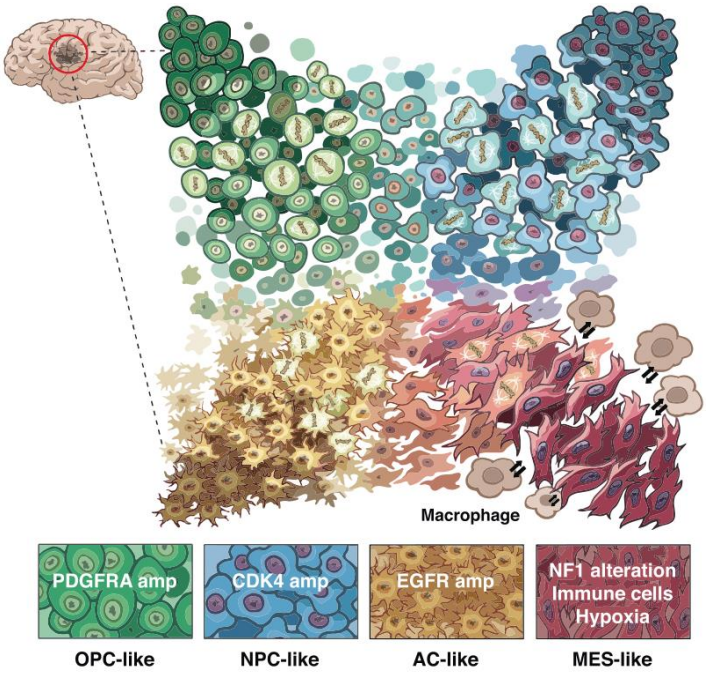


Figure 1.2 **The cell states in glioblastoma.** Green cells represent oligodendrocyte progenitor-like (OPC-like) tumor cells; blue cells represent neuronal progenitor-like (NPC-like) tumor cells; brown cells represent astrocytic-like (AC-like) tumor cells; red cells represent mesenchymal-like (MES-like) tumor cells. Cells with lighter or darker tones of specific color to indicate the strength of the specific cell state classification. Cell state transitions are indicated in cells between cell states. The cells with mitotic spindle indicate actively cycling cells. Four boxes at the bottom showing the genetic alterations and/or tumor microenvironment in cell sates: copy number amplification of PDGFRA, CDK4 and EGFR were associated with a high frequency of OPC-like, NPC-like, and AC-like cells, respectively. NF1 alteration is associated with MES-like cells. The microenvironment of MES-like cells is characterized by the presence of immune cells and hypoxia areas. Reprinted from Figure 7G in Neftel et al., 2019 with permission from Elsevier.

1.2.3 Cell-to-cell connectivity in glioblastoma

The standard treatment for GB includes surgery, radiotherapy, and chemotherapy with the alkylating agent temozolomide (Weller et al., 2015). However, despite these therapies, the survival of GB patients has shown very little improvement in the past decade, and drug resistance remains a major challenge (Haar et al., 2012). Various mechanisms contribute to therapeutic resistance in GB, including drug efflux, DNA damage repair, cancer stem cell, hypoxia, and miRNAs (Haar et al., 2012).

Recently, researchers have proposed another mechanism of therapeutic resistance

involving ultra-long membrane protrusions called tumor microtubes (TMs). GB tumor cells form a cell-to-cell network through these TMs, allowing them to invade, proliferate, and communicate with each other (Osswald et al., 2015). The interconnected tumor cells can exchange molecules through gap junctions present in the TMs, and the network can self-repair and protect TM-connected tumor cells from the effects of radiotherapy (Osswald et al., 2015). A further study has shown that TM-connected tumor cells are resistant not only to radiotherapy but also to chemotherapy and surgical lesions (Weil et al., 2017). Targeting the TM-connected tumor cell network has been proposed as a potential clinical application to overcome resistance in GB (Osswald et al., 2015; Weil et al., 2017).

The gene growth-associated protein 43 (GAP43) has been identified as a driver for TM formation (Osswald et al., 2015; Weil et al., 2017). Knockdown of GAP43 in GB tumor cells has been shown to reduced invasion speed, proliferation capacity and the number of TM connections (Osswald et al., 2015). Moreover, GAP43-deficient tumor cells fail to recolonize areas of surgical lesions (Weil et al., 2017).

Figure 1.3 illustrates the therapeutic resistance mechanisms, in which TM networks play a significant role. Additionally, studies have discovered synaptic ultrastructures between neurons and GB tumor cells located on TMs, termed neurogliomal synapses (Venkataramani et al., 2019). These synapses can activate the TM-connected tumor cell network and drive GB tumor cell invasion (Venkataramani et al., 2019).

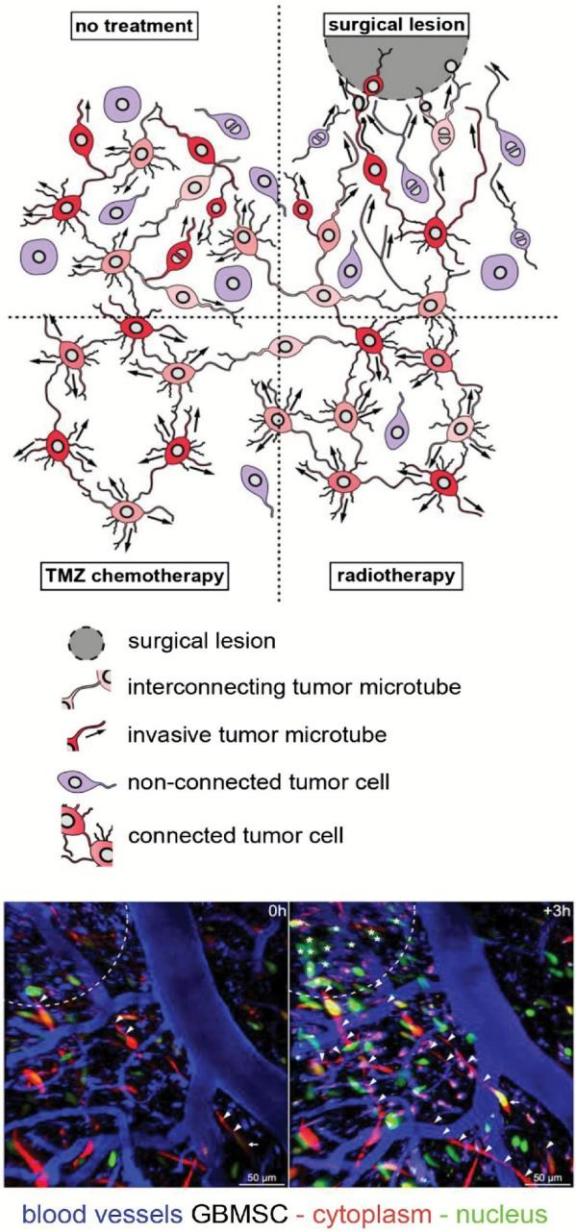


Figure 1.3 **TM network contributes to therapeutic resistance in GB.** Top panel, Tumor microtubes (TMs) interconnect tumor cells, promoting tumor invasion. After surgical resection, TMs extend and translocate nuclei to the surgical lesion area. After TMZ chemotherapy or radiotherapy, TM-connected tumor cells have a survival advantage, while TM-unconnected tumor cells are more susceptible to treatment-induced cell death. Bottom panel, TMs assist in the repair of surgical lesions. Xenografted glioblastoma stem-like cells (GBMSC) were observed using *in vivo* multiphoton laser scanning microscopy (MPLSM) at time point 0h and 3h, 7days after inducing a surgical lesion. The tumor cells extended TMs and were able to repopulate the lesion area. Arrowheads: extended TMs; Asterisks: transported tumor nuclei. Reprinted from Fig.1 in Lou, 2017 with permission from Oxford University Press.

To better understand the molecular features of TM-connected GB tumor cells, a method to distinguish TM-unconnected and TM-connected cells is needed. Since TM-connected GB cells exchange small molecules through gap junctions, a gap junction-permeable dye can be used to label TM-connected cells (Osswald et al., 2015).

Sulforhodamine 101 (SR101) is a red fluorescent dye that can spread through gap-junctional connections in astroglia and can monitor calcium level in cellular networks of astroglia and neurons (Nimmerjahn et al., 2004). Osswald et al. locally injected SR101 into tumors *in vivo* and then quantified the fluorescence intensity of SR101 in TM-connected and TM-unconnected cells under multiphoton laser scanning microscopy (MPLSM) (Osswald et al., 2015). The TM-connected tumor cells showed a higher SR101 fluorescence intensity than the TM-unconnected cells (Figure 1.4a, Osswald et al., 2015). A similar result was found in Venkataramani et al., 2019 (Figure 1.4b). As a systemic application, a similar intravital selection method (with SR101) was established to distinguish TM-connected and TM-unconnected GB tumor cells (Xie et al., 2021), and showed SR101 uptake was higher in TM-connected than TM-unconnected tumor cells (Figure 1.4c).

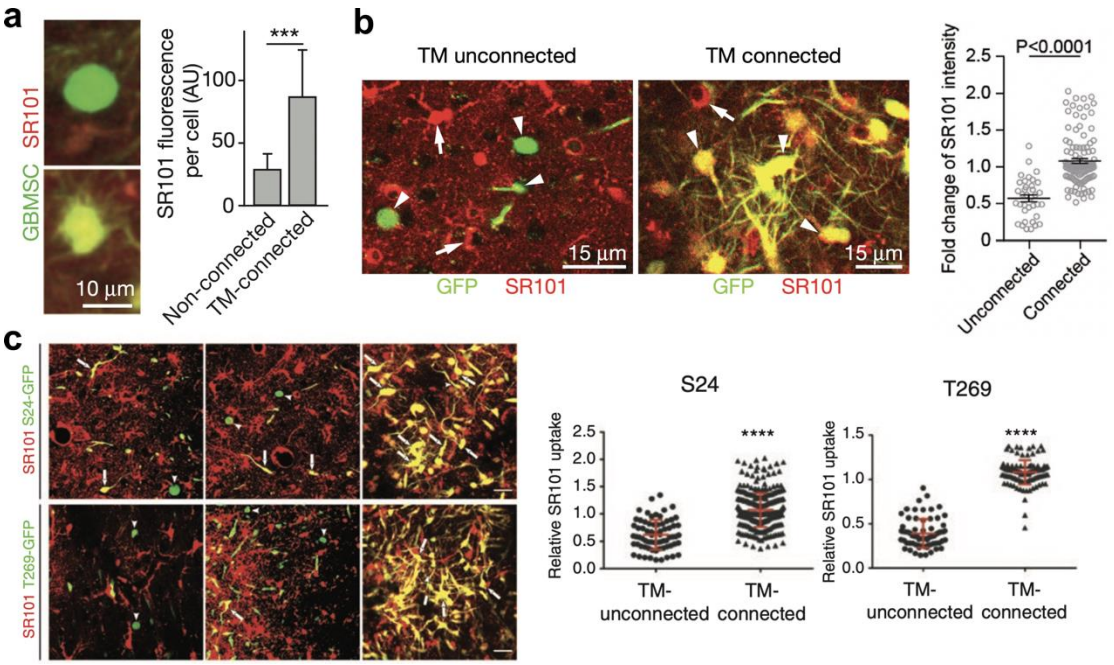


Figure 1.4 **SR101 uptake in TM-connected and TM-unconnected cells.** Glioma tumor cells are green fluorescent protein (GFP) tagged. Image data are obtained by in vivo MPLSM. a) SR101 uptake in a TM-unconnected glioma cell (Left top) and TM-connected glioma cell (Left bottom). Right, the quantification of SR101 uptake. 55 cells in three mice per condition; AU, arbitrary units; the p-value was obtained by Mann-Whitney U test; *** p-value < 0.001. Reprinted from Figure 2f in Osswald et al., 2015 with permission from Springer Nature. b) SR101 uptake in TM-unconnected cells (Left) and TM-connected cells (Middle). Arrowhead points to tumor cells; Arrow points to non-tumor cells. Right, the quantification of SR101 uptake. 116 TM-connected cells vs. 36 TM-unconnected cells from three mice. The p-value was obtained by Mann-Whitney U test. Reprinted from Fig. 2e and Extended Data Fig. 4l in Venkataramani et al., 2019 with permission from Springer Nature. c) Left, SR101 uptake in TM-unconnected tumor cells (arrowhead) and TM-connected tumor cells (arrow) from two cell lines. Right, quantification of SR101 uptake. Two-tailed unpaired Student's t test; *** p-value < 0.001. Reprinted from Figure 1c-e in Xie et al., 2021 with permission from Oxford

University Press.

1.3 Aim of the study

The primary goal of this study is to explore the transcriptome landscape of the TM-connected GB cell population using RNA-Seq and scRNA-Seq and establish a gene expression signature. This signature should not only identify the presence of TM-connected cells but also quantify the extent of their connectivity. The connectivity signature's robustness and efficacy will be rigorously assessed by applying it to multiple GB RNA-Seq and scRNA-Seq datasets. Furthermore, the study will explore potential correlations between the connectivity signature and various factors, including malignant cell states, prevalent gene mutations, GB expression subtypes, and patient survival outcomes. Moreover, efforts will be directed towards pinpointing a key marker that plays a pivotal role in the connectivity signature.

Additionally, the secondary aim of this study is to address the significant challenge of accurately identifying similar cell subpopulations that share the same biological types/states across diverse donors or experimental conditions. A more specific goal is to develop a user-friendly bioinformatics tool capable of quantifying the similarity between cell subpopulations effectively. This tool will enable precise, robust, and efficient mapping of these similar subpopulations.

Together, these aims endeavor to deepen our understanding of GB cell connectivity and the similarity within cell populations.

2. A connectivity signature in glioblastoma

Reseachers have demonstrated that approximately half of glioma cells were interconnected through tumor microtubes (TMs), forming a functional network that promotes tumor progression and exhibits resistant to therapy (Osswald et al., 2015; Weil et al., 2017; Venkataramani et al., 2019; Xie et al., 2021). However, there is a limited number of molecular markers for TM-connectivity that have been identified, and the quantification of connectivity in tumor samples has remained elusive.

In this section, I outline the development of a gene expression signature for the cell connectivity within glioblastoma (GB), referred to as the "connectivity signature". I introduce “connectivity signature score” (CSS) that measures the extent of connectivity using this signature. I validate the robustness and significance of the CSS various RNA-Seq and scRNA-Seq GB datasets. I conduct an examination of the relationships between the connectivity signature and cell states, as well as the association between the connectivity signature and patient survival. Additionally, I accomplish the identification and validation of a driver gene within the connectivity signature.

Contributions

The “wet lab” experiments were mainly conducted by Dirk C. Hoffmann from Clinical Cooperation Unit Neurooncology, German Cancer Research Center (DKFZ). Bioinformatic analyses and data visualization were performed by Ling Hai. This section mainly focuses on the presentation of bioinformatic analyses. The text was written by Ling Hai. It has been proofread and edited by ChatGPT.

2.1 Results

2.1.1 Development of connectivity signatures

To access the transcriptome landscape of TM-connected cells, the Sulforhodamine 101 (SR101) dye methodology was utilized to distinguish highly connected and lowly connected tumor cells in three xenografted patient-derived glioma cell lines (PDGCLs, Figure 2.1a). SR101 is a red fluorescent dye that can spread through gap-junctional connections in the cell network (Nimmerjahn et al., 2004). The fluorescence intensity of SR101 in cells can be quantified under multiphoton laser scanning microscopy (MPLSM) (Osswald et al., 2015). Tumor cells with high SR101 intensity ($SR101^{high}$) are considered highly connected, while tumor cells with low SR101 intensity ($SR101^{low}$)

are considered lowly connected (Osswald et al., 2015; Venkataramani et al., 2019; Xie et al., 2020; Hai & Hoffmann et al., 2021). The tumor cells were then FACS separated according to SR101 intensity and subjected to both RNA-Seq and scRNA-Seq (Figure 2.1a, Table 1).

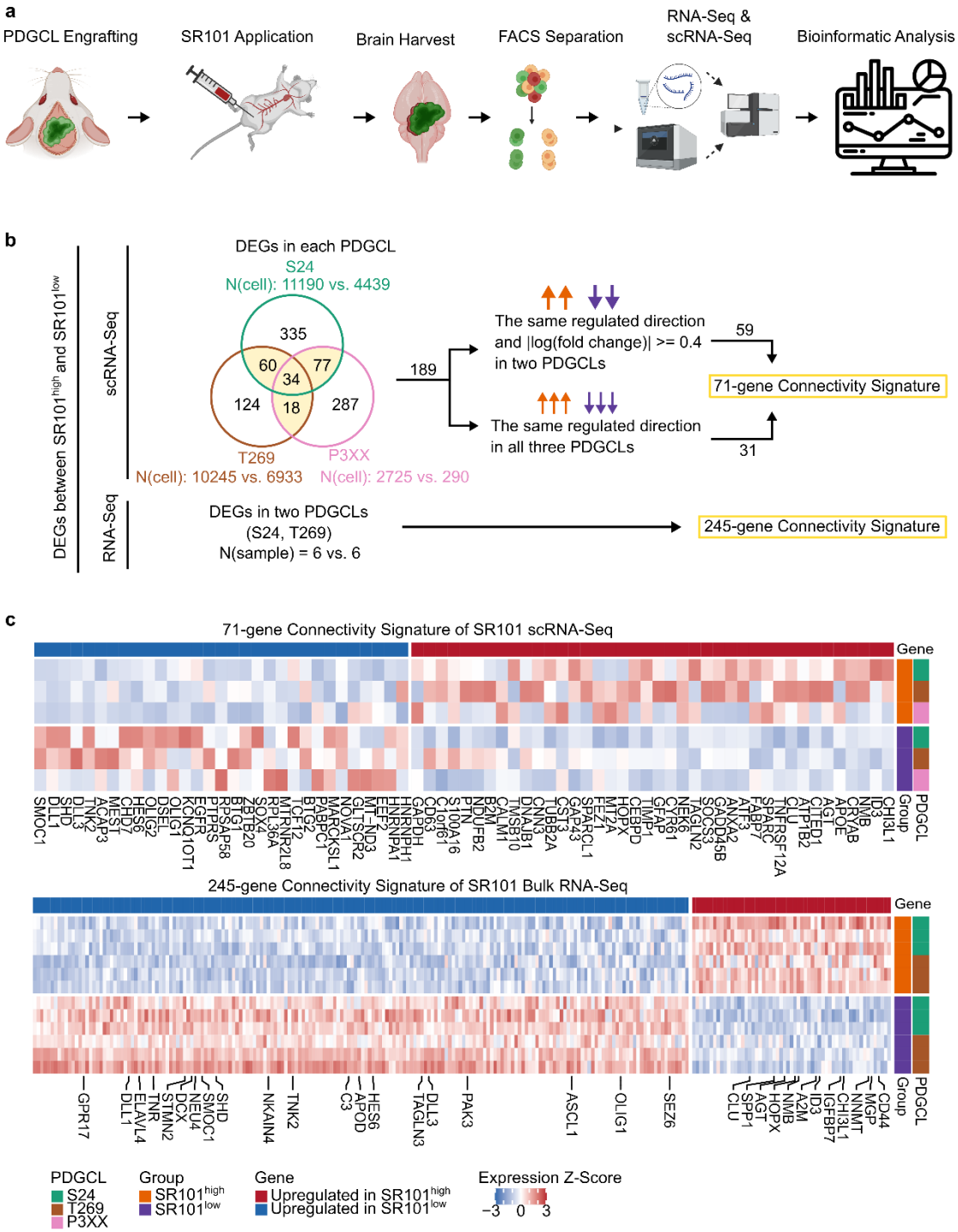


Figure 2.1 **Development of connectivity signatures.** a) Experimental design for the development of connectivity signatures. b) Computational strategy for the development of connectivity signatures from the SR101 scRNA-Seq and RNA-Seq datasets. Detailed in Methods. c) Heatmap showing the expression levels of connectivity

signature genes. Top: 71-gene connectivity signature derived from the SR101 scRNA-Seq dataset. Expression levels of 71 genes in cells of each sample were averaged, scaled to z-score, and winsorized at -3 and 3. Bottom: 245-gene connectivity signature derived from the SR101 RNA-Seq dataset. Expression levels of 245 genes in the samples were batch-corrected, scaled to z-score, and winsorized at -3 and 3. PDGCL, Patient derived glioma cell line. Figures were adapted from Hai & Hoffmann et al., 2021.

Table 1 **Overview of samples used for development of connectivity signatures.** PDGCL: Patient derived glioma cell line.

PDGCL	S24		T269		P3XX	
Group	High	Low	High	Low	High	Low
RNA-seq	3	3	3	3	0	0
scRNA-Seq	1	1	1	1	1	1

In the scRNA-Seq dataset, I obtained a total of 35,822 cells from three PDGCL models after quality controls (Table 2).

Table 2 **Properties of the SR101 scRNA-Seq dataset.** These numbers represent the data after undergoing quality controls to remove low-quality cells. Table was adapted from Hai & Hoffmann et al., 2021.

PDGCL	**Group**	**Genes (n)**	**UMIs (n)**	**Cells (n)**
S24	$SR101^{high}$	2876	7994	11190
S24	$SR101^{low}$	3153	8710	4439
T269	$SR101^{high}$	1380	2659	10245
T269	$SR101^{low}$	1590	3361	6933
P3XX	$SR101^{high}$	4718	23140	2725
P3XX	$SR101^{low}$	5230	26020	290

To gain molecular insights into cell-cell connectivity, I performed differential expression analyses between $SR101^{high}$ and $SR101^{low}$ samples in the SR101 RNA-Seq and scRNA-Seq datasets (Figure 2.1b):

In the scRNA-Seq dataset, to ensure an unbiased analysis in all three PDGCLs, I took several steps (Figure 2.1b): Firstly, I identified differentially expressed genes (DEGs) between $SR101^{high}$ and $SR101^{low}$ cells within each individual PDGCL separately using the FindMarker function in the Seurat package. Subsequently, I constructed a gene expression signature by including only those DEGs that satisfied one of the following criteria:

- DEGs that exhibited the same direction of regulation in all three PDGCLs.
- DEGs that demonstrated the same direction of regulation and absolute log2 fold change > 0.4 in at least two PDGCLs.

Finally, I obtained a 71-gene signature, referred to as “connectivity signature” (Figure 2.1c). Notably, this connectivity signature includes several known TM-connectivity genes such as GAP43 (Osswald et al., 2015; Weil et al., 2017; Venkataramani et al., 2022), DLL1 (Jung et al., 2021), DLL3 (Jung et al., 2021), and APOE (Venkataramani et al., 2019) (Figure 2.1c).

In the RNA-Seq dataset, I identified 245 DEGs that consistently showed dysregulation across two PDGCLs using the DESeq2 package (Figure 2.1b-c, Table 3).

Table 3 **The RNA-Seq-derived connectivity signature.** Table was adapted from Hai & Hoffmann et al., 2021.

Upregulated (*n* = 57): A2M, AC004485.3, AC009502.4, AGPAT9, AGT, AL031666.2, AP000345.1, APCDD1L, ARHGAP36, CARD16, CD44, CDK1, CHI3L1, CHRNA9, CLU, CNIH3, COL19A1, CPNE5, CYSTM1, DDO, F2RL2, GBP2, GJB2, GPER1, HBEGF, HEPH, HIST1H1B, HIST1H1D, HIST2H2AC, HOPX, ID3, IGFBP6, IGFBP7, LIF, LINC00551, LINC01057, LY96, MCHR1, MGP, NMB, NNMT, OTOS, PDCD1LG2, PLA2G5, PPEF1, RP11-283G6.4, RP11-3L21.2, RP11-483F11.7, SCN4B, SEMA3A, SPP1, SYNJ2, SYTL5, TMOD1, TRPC7, VAMP5, VSNL1
Downregulated (*n* = 188): AC007682.1, AC053503.11, AC114730.3, AC124944.5, ACTL6B, ADAMTS7, ADAMTSL2, ADAP2, ADCY7, AHSG, AMER2, AMZ1, APOD, ARHGAP24, ARHGDIB, ARPP21, ASCL1, ASXL3, ATP1A3, ATP1A4, ATP2B3, ATP8A1, B3GALT2, BCL11B, BCO2, BMF, BMP2, C14orf166B, C3, CA8, CAPN3, CARD10, CDH20, CDKN1C, CELF5, CHRM4, CKM, CLSTN2, CLVS2, COL20A1, CRB1, CTRC, CUX2, CX3CR1, CYP4F12, DCX, DDX26B, DGKE, DIRAS2, DLL1, DLL3, DOCK9, DSCAM, ELAVL4, ELFN2, EPB41, EPHB1, FAM105A, FBN3, FERMT1, FERMT3, FGD3, FGL2, FLRT1, FRAT1, FRMPD1, FTMT, GADD45G, GCH1, GPD1, GPR123-AS1, GPR17, GRID2, HCK, HDC, HES6, HID1, HMHA1, HSPA1L, IFITM10, IGSF9B, JAG2, KCNH8, KCNIP3, KCNJ2, KIF19, LAG3, LIMS2, LINC00925, LPIN3, MAP1LC3C, MARCH1 , MCF2, MFNG, MTRNR2L10, MTRNR2L6, MTRNR2L7, MTSS1, MUC4, MYCL, MYCN, MYH14, MYO7B, MYT1L, NAA11, NAALAD2, NEU4, NKAIN1, NKAIN4, NOD2, NSG2, OLIG1, PACSIN1, PAK3, PARP8, PCP4, PCSK2, PDE2A, PLCL1, PLCXD2, PLXDC1, PODN, PON3, PRKCZ, PSD2, PTAFR, PTPRJ, PTPRM, QPCT, RAB11FIP4, RAB33A, RASGEF1C, RNF144B, RP11-1055B8.3, RP11-134P9.3, RP11-161M6.2, RP11-231C18.1, RP11-309M23.1, RP11-328J6.1, RP11-430C7.5, RP11-85M11.2, RP11-90E5.1, RP13-735L24.1, SATB1, SEMA6B, SERPINB1, SERPINF1, SERPINF2, SEZ6, SHD, SIGLEC1, SLC16A10, SLC17A7, SLC29A3, SLC7A7, SLCO4A1, SLIT1, SMOC1, SNX20, SOX8, SPNS2, SRC, STEAP4, STMN2, STXBP2, STXBP6, SUSD3, TAGAP, TAGLN3, TBC1D9, TBX21, TEX38, TF, TM6SF1, TNFRSF11B, TNK2, TNR, TPBGL, TRPV4, TSPAN15, TTC24, TUSC3, UCP3, UNC13D, UNC5A, VIPR2, WIPF3, ZDHHC22

With these analyses, I obtained two sets of connectivity signatures. The further comparison between the scRNA-Seq-derived and RNA-Seq-derived connectivity signatures is presented in the next subsection.

2.1.2 Comparisons of RNA-Seq and scRNA-Seq-derived connectivity signatures

To compare the connectivity signatures obtained from two different transcriptome profiling methods, I examined the fold changes of genes between the $SR101^{high}$ and $SR101^{low}$ groups in both the RNA-Seq and scRNA-Seq datasets. Fold changes indicate the magnitude of expression change and the direction of regulation for each gene.

In the comparison of fold changes for all commonly detected genes, a moderate correlation (R = 0.44) was observed (Figure 2.2a). I hypothesized that this limited correlation is a result of the entirely different methodologies employed in bulk and single-cell sequencing. In scRNA-Seq, gene expression is quantified in individual cells, whereas bulk RNA-Seq comprises a mixed signal from various cell types. Additionally, gene drop-out events in the scRNA-Seq dataset may be influencing the correlation. By restricting the analysis to genes expressed in at least 10% of cells in the scRNA-Seq dataset, the correlation increased to R = 0.55 (Figure 2.2b). To address the potential influence of genes with insignificant changes between $SR101^{high}$ and $SR101^{low}$ groups, I only considered DEGs with adjusted p-value < 0.05 in both datasets, leading to a further increase in correlation to R = 0.71 (Figure 2.2c). Notably, the 13 overlapping genes in the two connectivity signatures exhibited a high correlation (R = 0.77, Figure 2.2d).

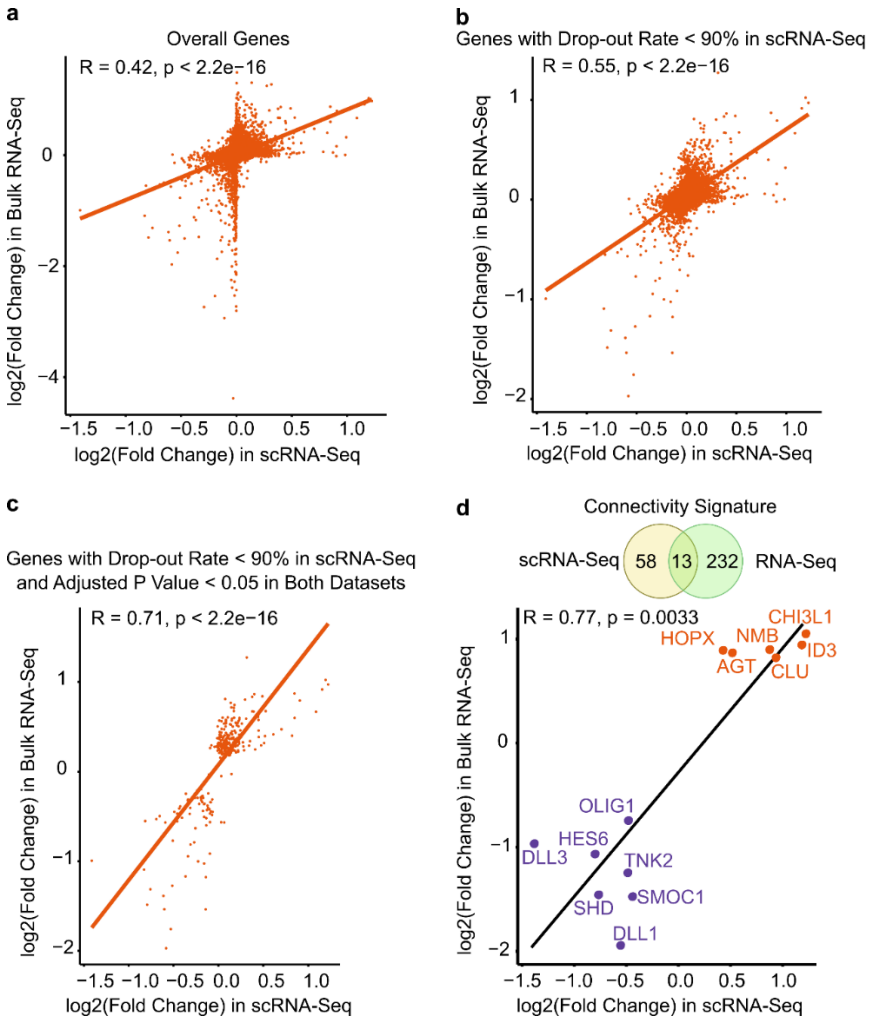


Figure 2.2 **Fold changes between highly and lowly connected samples in the SR101 scRNA-Seq and RNA-Seq datasets.** The log2 fold change of gene expression was calculated using the average expression of samples/cells between the $SR101^{high}$ and $SR101^{low}$ groups. Correlation coefficients were determined using the Spearman method. a) Overall, 16,759 genes were analyzed. b) 6,984 genes were expressed in more than 10% of cells in the scRNA-Seq dataset. c) 297 genes showed an adjusted p-value of less than 0.05 in both the RNA-Seq and scRNA-Seq datasets. d) There were 13 overlapping genes between the bulk and single-cell RNA-Seq-derived connectivity signatures. Orange color indicates upregulated genes in both datasets, while purple color indicates downregulated genes. Figures were adapted from Hai & Hoffmann et al., 2021.

Next, I compared the enriched gene ontologies (GOs) in both the SR101 RNA-Seq-derived and scRNA-Seq-derived connectivity signatures using the Metascape tool (Zhou et al., 2019). Although only 13 DEGs overlap between the two connectivity signatures (Figure 2.2d), it is noteworthy that the enriched GO terms associated with these two signatures demonstrate substantial consistency (Figure 2.3a). Additionally, a significant number of genes, although not directly overlapping within these two connectivity signatures, share the same GO terms (Figure 2.3b).

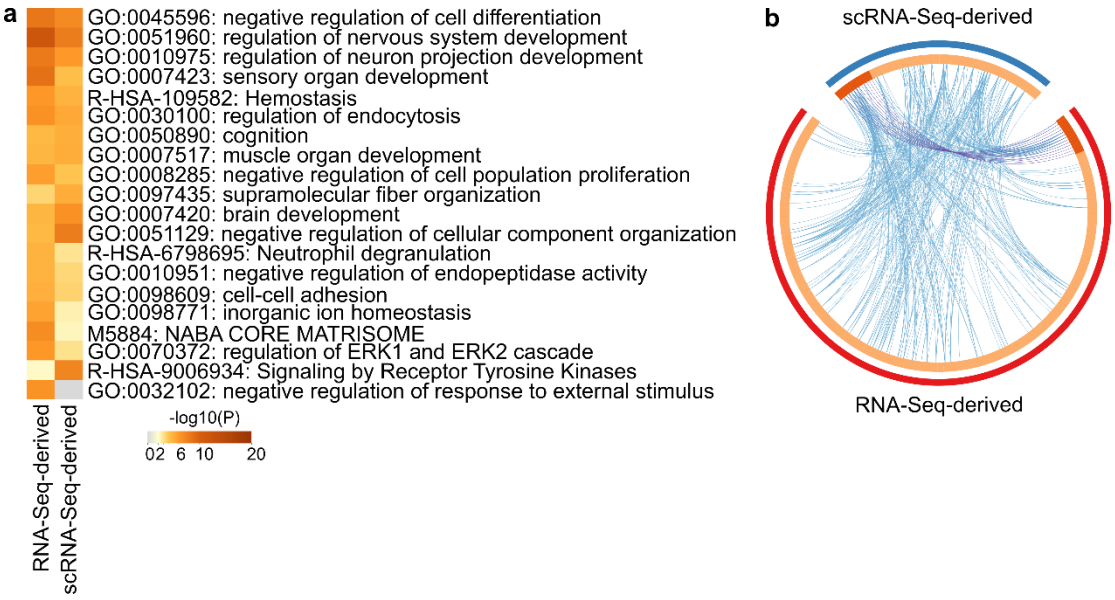


Figure 2.3 **Enriched gene ontologies in the connectivity signatures.** a-b) Analyses using Metascape (Zhou et al., 2019). a) Common enriched ontologies. b) Overlaping genes (purple lines) and genes shared the same GO terms (blue lines). The length of circular arc indicates the number of genes in two connectivity signatures. Outer circle represents the number of genes in each gene set. Inner circle highlights the number of overlapping genes (dark orange).

To computationally quantify the degree of connectivity in cells, I calculated a score based on the average relative expression of genes that constituted a connectivity signature, termed connectivity signature score (CSS), using the AddModuleScore function in the Seurat package. For each cell of the scRNA-Seq data or each sample of the RNA-Seq data, I calculated the CSS based on the RNA-Seq-derived or scRNA-Seq-derived connecvitiy signature. Remarkably, both RNA-Seq-derived and scRNA-Seq-derived CSSs can well-distinguish the $SR101^{high}$ and $SR101^{low}$ groups in each PDGCL in both RNA-Seq and scRNA-Seq datasets (Figure 2.4a-b). Furthermore, the RNA-Seq-derived and scRNA-Seq-derived CSSs exhibit a high correlation in the SR101 scRNA-Seq dataset (R = 0.87) as well as in The Cancer Genome Atlas (TCGA) GB RNA-Seq dataset (R = 0.89).

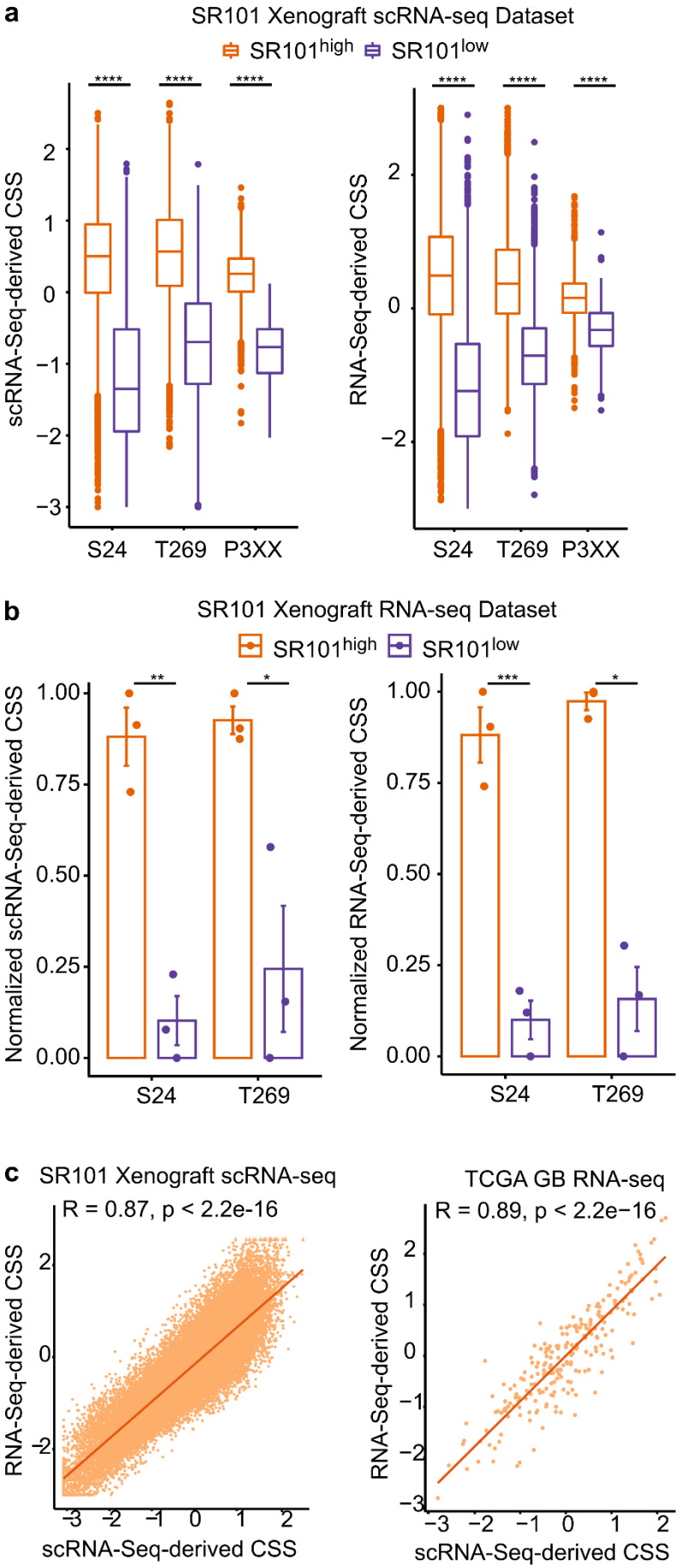

a
SR101 Xenograft scRNA-seq Dataset
SR101high SR101low
scRNA-Seq-derived CSS
RNA-Seq-derived CSS
****
S24 T269 P3XX
b
SR101 Xenograft RNA-seq Dataset
SR101high SR101low
Normalized scRNA-Seq-derived CSS
Normalized RNA-Seq-derived CSS
**
*
***
S24 T269
c
SR101 Xenograft scRNA-seq
R = 0.87, p < 2.2e-16
RNA-Seq-derived CSS
scRNA-Seq-derived CSS
TCGA GB RNA-seq
R = 0.89, p < 2.2e−16

Figure 2.4 **The scRNA-Seq-derived and RNA-Seq-derived CSSs**. a) The scRNA-Seq-derived CSS (Left) and RNA-Seq-derived CSS (Right) of cells in the $SR101^{high}$ and $SR101^{low}$ groups in each PDGCL in the SR101 scRNA-Seq. N = 35,822 cells. Tukey's boxplots display quartiles of data with whiskers drawn within 1.5 times the interquartile range (IQR). Mann-Whitney U test. b) The scRNA-Seq-derived CSS (Left) and RNA-Seq-derived CSS (Right) of cells in $SR101^{high}$ and $SR101^{low}$ groups in each PDGCL in the SR101 RNA-Seq. N = 3 per group per PDGCL. Barplots display means and standard errors (SE) of data. Two-sided paired t-test. c) Pearson correlation between scRNA-Seq-derived and RNA-Seq-derived CSS in the SR101 scRNA-Seq (Left) and TCGA GB RNA-Seq (Right) datasets. *, p-value < 0.05; **, p-value < 0.01; ***, p-value < 0.001. Figures were adapted from Hai & Hoffmann et al., 2021.

Next, I examined the CSS in cells of the SR101 scRNA-Seq dataset. I visualized the cells using Uniform Manifold Approximation and Projection (UMAP). UMAP can cluster cells with similar transcriptomic profiles together while preserving the global structure of the data to co-locate similar groups of cells. Cells from the same SR101 group were clustered together. while cells from different PDGCL were located far apart in the UMAPs (Figure 2.5a). I employed the "anchor" integration approach to remove the differences between PDGCLs, enabling cells from different PDGCLs to cluster together (Figure 2.5b). Both RNA-Seq-derived and scRNA-Seq-derived CSSs proved to be good indicators for the SR101 groups in cells in both the original and "anchor" integrated UMAPs (Figure 2.5a-b). The $SR101^{high}$ cells exhibited higher CSS values compared to those $SR101^{low}$ cells (Figure 2.5). Strikingly, cells located near the boundary between the $SR101^{high}$ and $SR101^{low}$ groups exhibited intermediate CSS values, showing a gradual and continuous increase in CSSs from $SR101^{high}$ to $SR101^{low}$ groups (Figure 2.5).

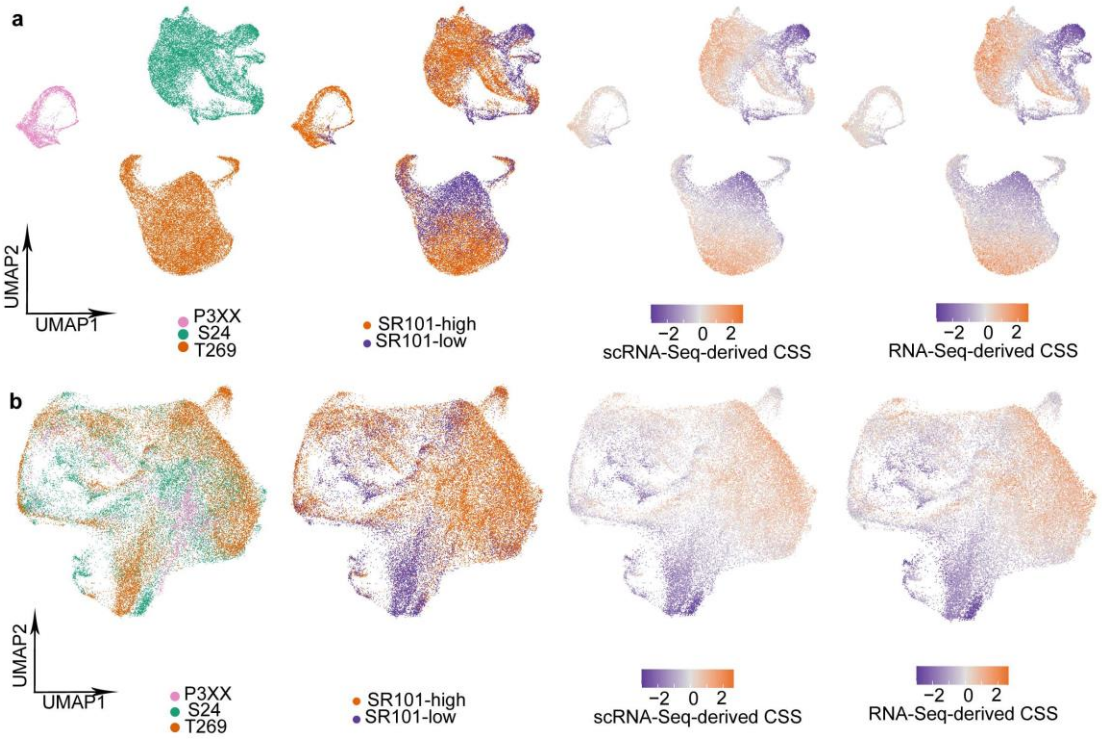


Figure 2.5 **The CSSs in the UMAPs of SR101 scRNA-Seq dataset.** a) Original UMAPs on single cells without integration. b) UMAPs on single cells with "anchor" integration. Cells are colored based on PDGCLs (Leftmost), SR101 groups (Second from Left), scRNA-Seq-derived CSSs (Third from Left) or RNA-Seq-derived CSSs (Rightmost). Figures were adapted from Hai & Hoffmann et al., 2021

To investigate the performance of CSS in predicting the group to which a cell belongs ($SR101^{high}$ or $SR101^{low}$), I calculated both RNA-Seq-derived and scRNA-Seq-derived CSSs for individual single cells in the SR101 scRNA-Seq dataset. Subsequently, I assigned cells to the $SR101^{high}$ group if they had positive CSS values or to the $SR101^{low}$ group if they had negative CSS values. I then compared the assigned groups with the FACS-sorted groups based on the SR101 fluorescence intensity. The scRNA-Seq-derived CSS yielded an 83% accuracy, while the RNA-Seq-derived CSS had 79% accuracy (Table 4). As a control, scores based on randomly selected genes resulted in a poor prediction performance (accuracy = 0.49, Table 4).

Table 4 **Prediction performances of RNA-Seq-derived and scRNA-Seq-derived CSSs.** PPV, positive predictive value; NPV, negative predictive value; Random score 1, randomly selected gene set with the same size as RNA-Seq-derived connectivity signature; Random score 2, randomly selected gene set with the same size as scRNA-Seq-derived connectivity signature. Figures were adapted from Hai & Hoffmann et al., 2021.

Metrics	RNA-Seq-derived CSS	scRNA-Seq-derived CSS	Random score 1	Random score 2
Accuracy	0.79	0.83	0.49	0.49
Sensitivity	0.77	0.95	0.47	0.48
Specificity	0.83	0.58	0.53	0.51
PPV	0.90	0.82	0.67	0.67
NPV	0.64	0.84	0.32	0.32

In this subsection, I compared the RNA-Seq-derived and scRNA-Seq-derived connectivity signatures at multiple aspects, including the gene expression fold change between the high and low SR101 groups in both SR101 RNA-Seq and scRNA-Seq datasets, the enriched GO terms of the connectivity signatures, and the comparisons between the RNA-Seq-derived and scRNA-Seq-derived CSSs. A high consistency was found in the two connectivity signatures. Since the scRNA-Seq-derived CSS yielded higher prediction accuracy, I will focus on characterizing the scRNA-Seq-derived connectivity signature and CSS, and refer to them as “connectivity signature” and “CSS” in the following subsections without explicitly mentioning “scRNA-Seq-derived”.

2.1.3 Connectivity and cell states

Recently, scRNA-Seq technology has enabled reseachers to further assign tumor cells to various cell states. Four major cell states (AC, MES, OPC, and NPC) were identified in GB (Neftel et al., 2019). I assigned these cell states to cells in the SR101 scRNA-Seq dataset. In the “anchor” integrated UMAP, cells with the same cell state were clustered together (Figure 2.6a). Interestingly, I found the cell state composition was different between the $SR101^{high}$ and $SR101^{low}$ groups (Figure 2.6b). The $SR101^{high}$ group contained high ratios of AC and MES cells, while the $SR101^{low}$ group had a high ratio of NPC cells (Figure 2.6b). Notably, the CSS were higher in AC and MES than NPC (Figure 2.6c).

Furthermore, when comparing to cell state markers identified in Neftel et al., 2019, I found approximately 50% of the genes in the connectivity signature were also cell states markers. Among the 40 upregulated genes of the connectivity signature, 10 were AC markers (AGT, ATP1B2, CLU, SPARC, FABP7, GFAP, HOPX, SPARCL1, CST3, and S100A16), seven were MES1 markers (CHI3L1, APOE, ANXA2, TAGLN2, TIMP1, MT2A, and S100A16), two were MES2 markers (ATF3 and ANXA2), and one was an NPC2 marker (TUBB2A). Additionally, there were 12 NPC1 markers

(MARCKSL1, BTG2, TCF12, SOX4, PTPRS, OLIG1, HES6, CHD7, MEST, DLL3, SHD, and DLL1), 1 OPC marker (OLIG1), and 1 NPC2 marker (SOX4) among the 31 downregulated genes in the connectivity signature.

When examining the expression level of the genes in the connectivity signature in each cell state, I found that the upregulated genes in connectivity signature were highly expressed in AC and MES cells, while the downregulated genes in the connectivity signature were highly expressed in NPC and OPC cells (Figure 2.6d).

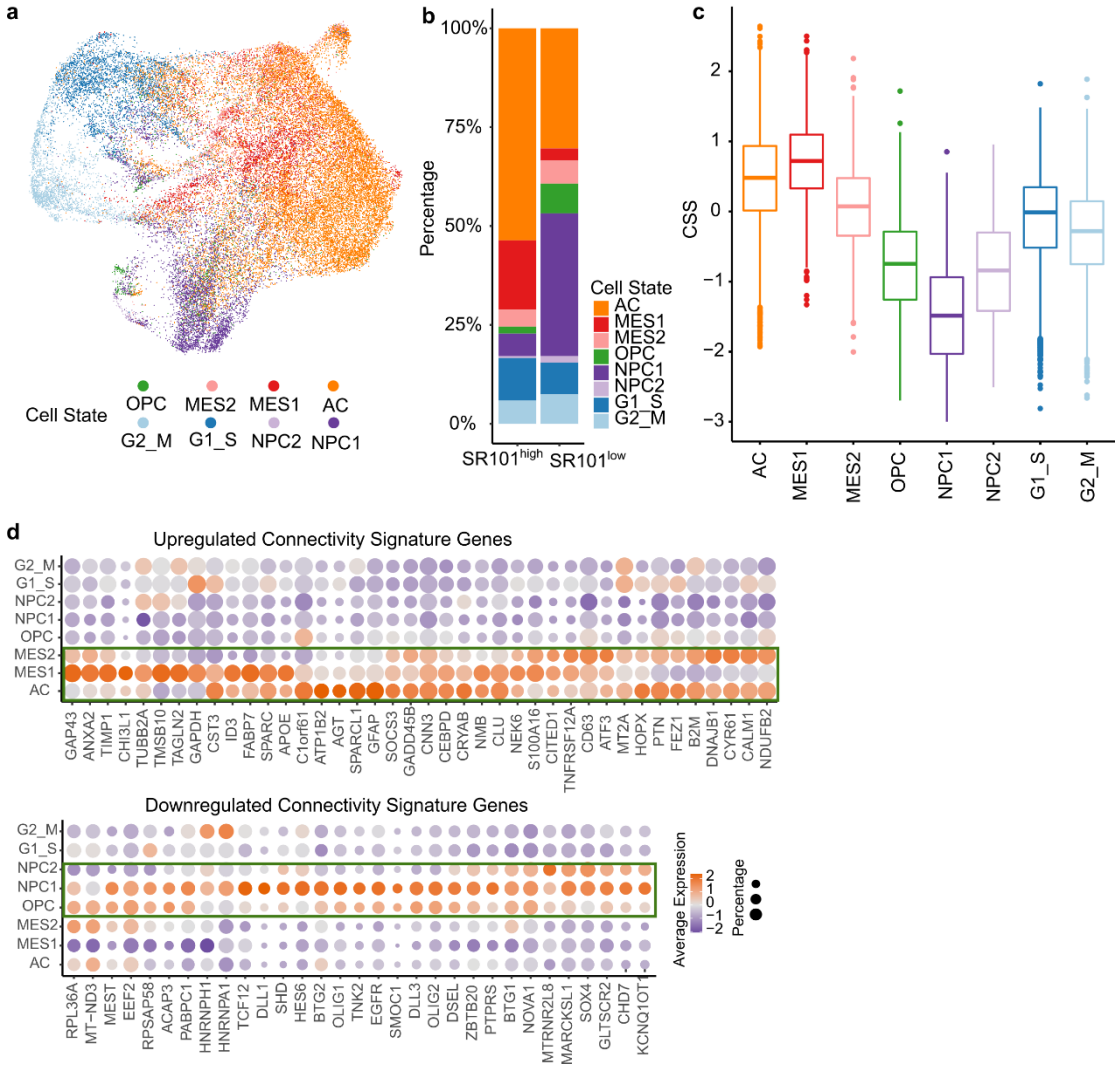


Figure 2.6 **Connectivity signature and cell state in the SR101 scRNA-Seq dataset.** a) UMAP colored by cell states. b) Percentage of cell states in the SR101 groups. c) CSSs of cells in each cell state. d) Average expression levels of upregulated (Top) and downregulated (Bottom) connectivity signature genes in each cell state. Expression levels were scaled to z-score across cell states. Dot size indicates percentage of expressed cells in each cell state. AC, astrocytic-like cell; MES, mesenchymal-like cell; NPC, neuronal progenitor-like cell; OPC, oligodendrocyte progenitor-like cell. Figures were adapted from Hai & Hoffmann et al., 2021.

To assess the cross-talk between cells, I identified ligand-receptor interactions in the SR101 scRNA-Seq dataset using CellChat (Jin et al. 2021). There are cell-cell communications among cell states in both $SR101^{high}$ and $SR101^{low}$ samples; however, the $SR101^{high}$ samples had a higher number of interations between cell states than the $SR101^{low}$ samples (Figure 2.7a-b). Several pathways exhibited distinct signaling patterns between $SR101^{high}$ and $SR101^{low}$ samples, including the NOTCH pathway (Figure 2.7c). Specifically, the NOTCH pathway was completely depleted in AC cells of $SR101^{high}$, while turned on in AC cells of $SR101^{low}$ cells (Figure 2.7c-e). Furthermore, the NOTCH pathway was highly activated in NPC1 cells (Figure 2.7c-e), which were enriched in $SR101^{low}$ samples (Figure 2.6b). Downregulation of the NOTCH pathway has found to promote TM-connections in a previous study (Jung et al., 2021).

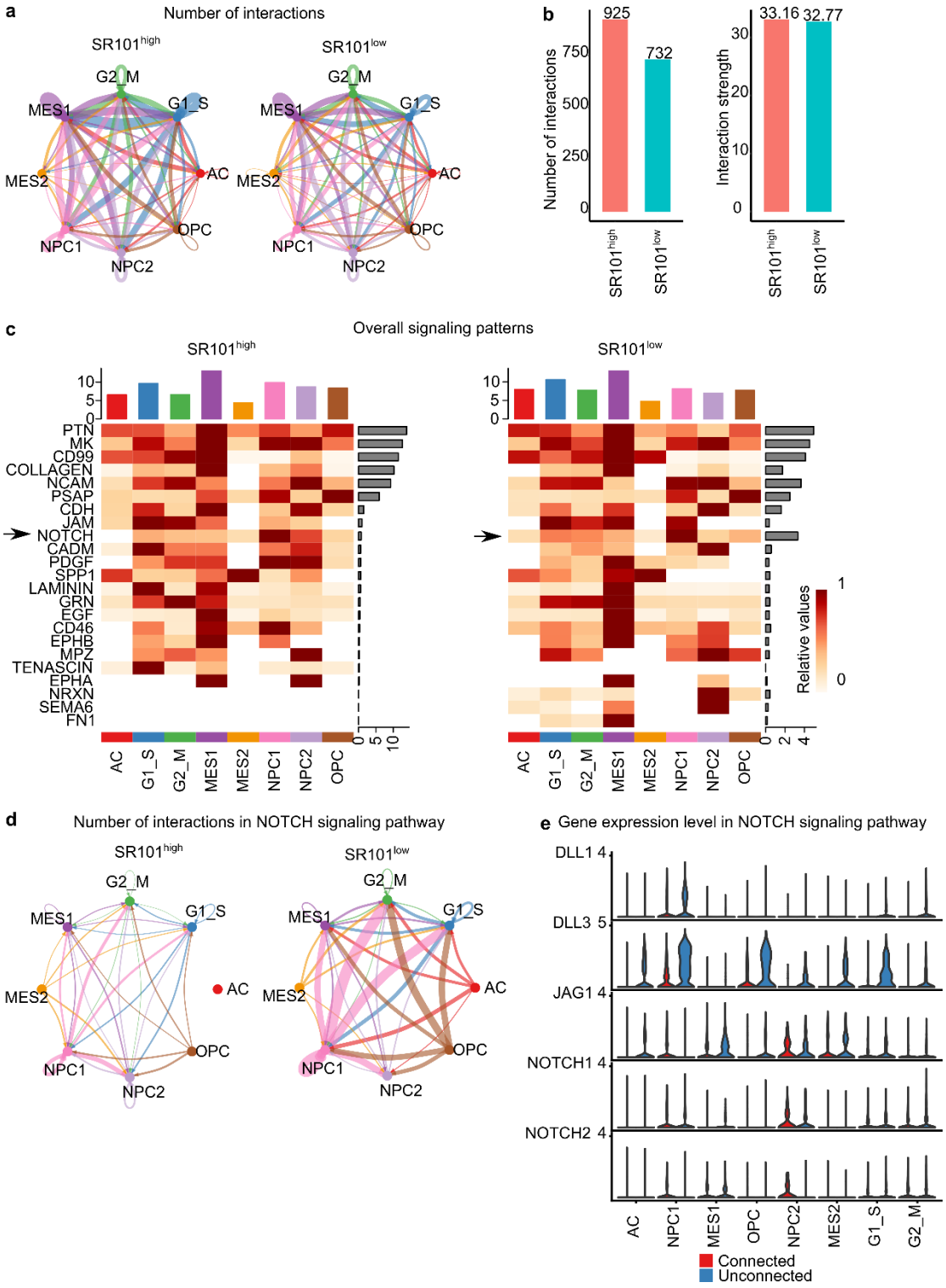


Figure 2.7 **Ligand-receptor interation among cell states in the two SR101 groups.** scRNA-Seq data from SR101^{high} and SR101^{low} sapmles was analyzed. a) The number of estimated cell interactions between cell states. The lines linking two cell states indicate an interation between these two cell states. The thickness of the lines indicates the relative number of interations between cell states. Each color represents the interactions from a specific cell state. b) The total number of estimated interactions (Left). The total interaction weights (Right). c) The overall signaling patterns of each cell state. Colors in the heatmap represent the relative number of interactions. The bar plot over columns represents the sum of the relative number of interactions in each cell state. The bar plot over rows represents the sum of the relative number of

interactions in each differentially interacted pathway. Arrows highlight the NOTCH signaling pathway. d) The number of interactions in NOTCH signaling pathway between cell states in $SR101^{high}$ samples (Left) and $SR101^{low}$ samples (Right). e) The expression levels of ligand/receptor genes involving in NOTCH signaling pathway.

To investigate the cell state dynamics in the SR101 groups, I applied RNA velocity estimation to the SR101 scRNA-Seq data using velocyto (La Manno et al. 2018) and scVelo (Bergen et al. 2020). In the $SR101^{high}$ samples, the transition flow of cells started from the cycling cells (G2_M), then went through various cell states, and finally reached AC as an endpoint (Figure 2.8). The AC cell state had a higher CSS (Figure 2.6c). This indicates that the $SR101^{high}$ cells tend to form a more connected and harmonized network. On the other hand, in the $SR101^{low}$ samples, the starting point was also G2_M, but there were multiple endpoints of the transition, including AC and NPC1 (Figure 2.8). The NPC1 cell state had a lower CSS (Figure 2.6c). This indicates that $SR101^{low}$ cells tend to form a loosely connected network with the potential to develop multiple cell states.

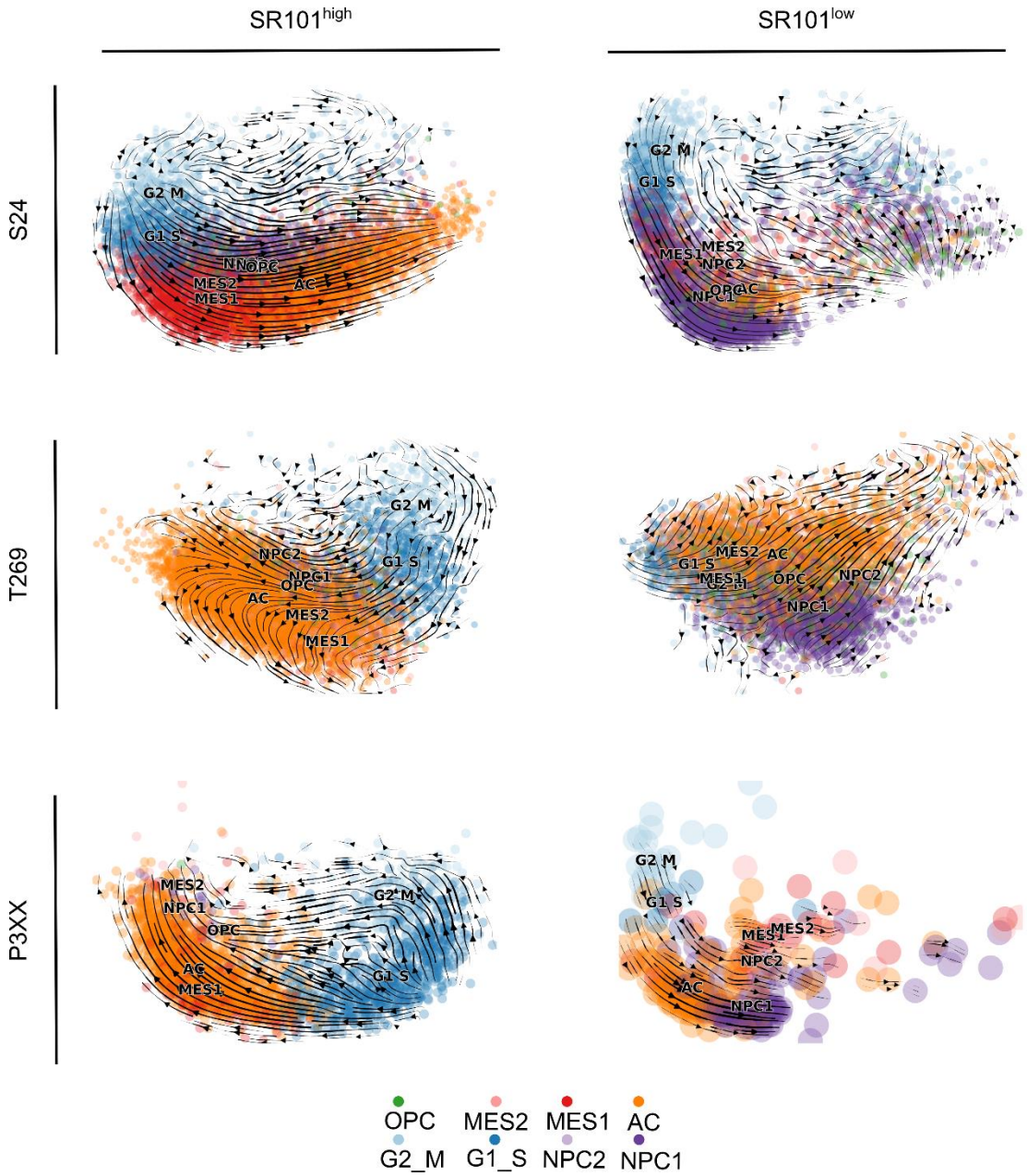


Figure 2.8 **RNA velocity in the two SR101 groups.** RNA velocities were projected onto the principle component analysis (PCA) plots of scRNA-Seq data of the SR101^{high} and SR101^{low} cells in each PDGCL. Cells were colored according to their cell states. The streamline indicates the velocity vector field, showing the direction of flow.

In this subsection, I investigated the relationship between connectivity and cell state in the SR101 scRNA-Seq dataset. I observed that SR101^{high} cells exhibited a high correlation with AC and MES cell states, while SR101^{low} cells were correlated with the NPC cell state. I will further evaluate the features of the connectivity signature in various datasets in the next subsections.

2.1.4 The connecivity signature in GB patient tumor scRNA-Seq datasets

I established the connectivity signature from the scRNA-Seq data of xenografted PDGCL models. To access the effectiveness of the connecivity signature in patient tumors, 21 IDH wt GB patient tumor samples were subjected to scRNA-Seq (Table 5). After quality controls, I obtained a total of 213,444 single cells (Table 5).

Table 5 **Properties of the patient tumor scRNA-Seq dataset.** N, the number; GB meth. subtype, glioblastoma methylation classifier in Capper et al., 2018. RTK, Receptor tyrosine kinase.

ID	Age	Sex	GB meth. subtype	Median gene (n)	Median count (n)	Cell (n)
T1	69	Male	RTK I	561	685	932
T2	61	Male	Mesenchymal	786	1116	16721
T3	68	Male	RTK I	634	794	6634
T4	61	Male	RTK II	943	1273	9029
T5	77	Male	RTK II	767	1006	6175
T6	73	Male	Mesenchymal	1020	1440	2744
T7	56	Male	RTK II	643	846	5009
T8	80	Female	Mesenchymal	584	679	1626
T9	67	Male	Mesenchymal	1393.5	2071	15092
T10	64	Male	Mesenchymal	1097	1653.5	11192
T11	44	Male	N/A	1198	1762	14588
T12	66	Male	RTK I	996	1381	15057
T13	54	Male	RTK I	1231	1946	5165
T14	69	Female	RTK II	1289	2053	11927
T15	53	Male	RTK II	998	1310	11533
T16	43	Female	RTK II	608	728	5830
T17	64	Male	Mesenchymal	1347.5	1917	19668
T18	56	Male	Mesenchymal	810	1102	8221
T19	55	Male	RTK I	1300.5	1890	13450
T20	32	Female	Mesenchymal	856	1055	15707
T21	55	Male	RTK II	1912	3024.5	17144

I observed a significant inter-tumor heterogeneity in the UMAPs of patient tumor

scRNA-Seq dataset (Figure 2.9a). To mitigate the differences across tumors, I employed “anchor” integration. This integration resulted in the merging of cells from different tumors in a UMAP plot (Figure 2.9b).

Furthermore, I identified 24 cell clusters through shared nearest neighbor (SNN) analysis in the integrated dataset (Figure 2.9c). To annotate these clusters, I collected maker genes of seven cell types in GBs from previous studies (Neftel et al., 2019; Zhang Y. et al., 2016; He et al., 2016). Applying these cell type markers to the patient tumor scRNA-Seq data, I calculated gene signature scores of each cell type in individual single cells, using the AddModuleScore function in the Seurat package (Figure 2.9d). Notably, specific clusters exhibited higher scores in particular cell types (Figure 2.9d). I annotated these clusters based on their highest cell type signature scores.

To validate the malignant and non-malignant clusters, I identified copy number variations in cell clusters (Figure 2.9e). As expected, the malignant clusters displayed amplification in chr7 and depletion in chr10, while non-malignant clusters showed no CNVs on these chromosomes (Figure 2.9e).

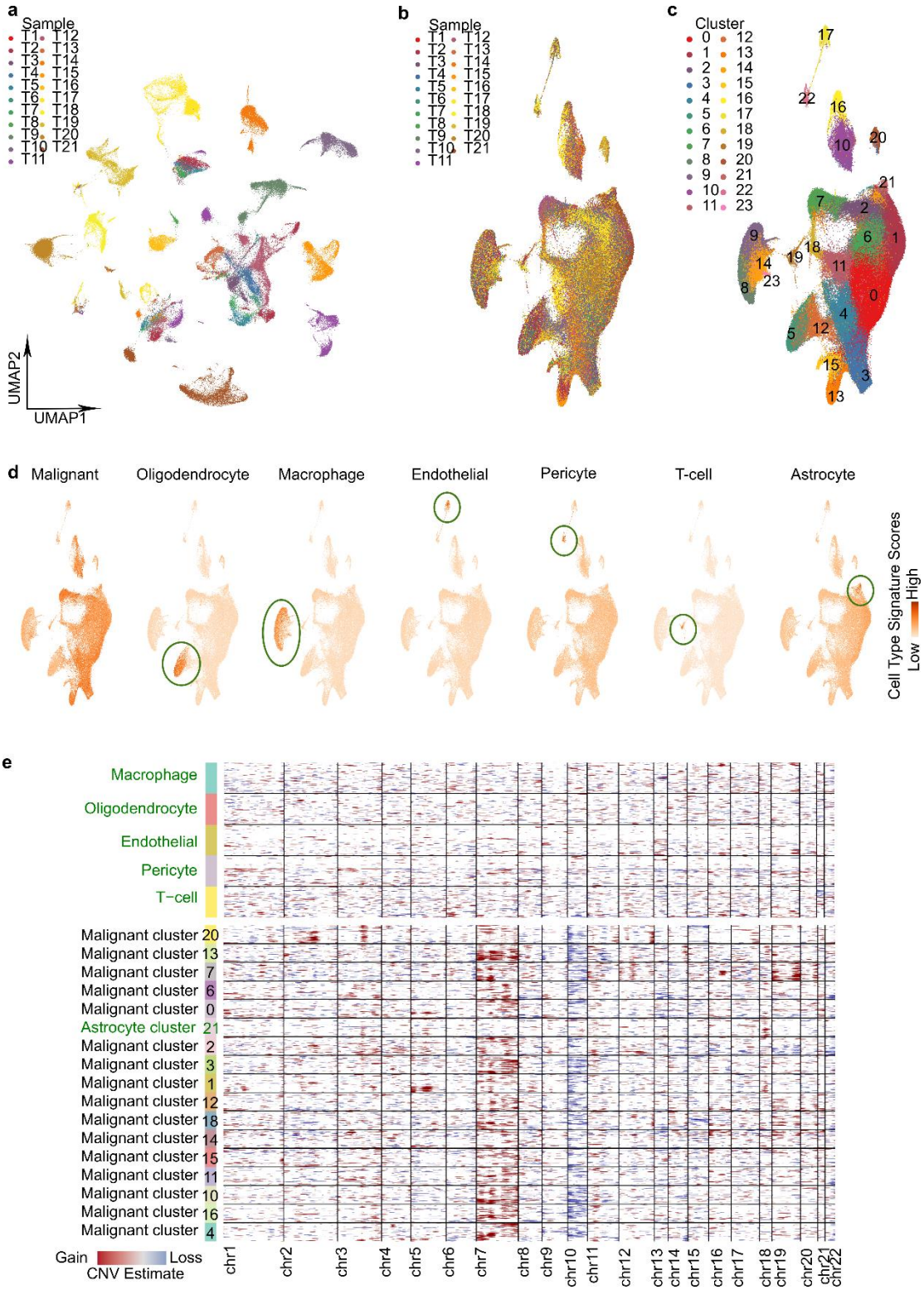


Figure 2.9 **The patient tumor sample scRNA-Seq dataset.** a-d) UMAPs of single cells in 21 GB patient tumor samples. a) Cells were colored by patient tumor samples. b) Cells were “anchor” integrated across samples and colored by patient tumor samples. c) Cells were “anchor” integrated across samples and colored by unsupervised clusters. d) Cells were “anchor” integrated across samples and colored by each cell type signature scores. Green circles highlight the cell subpopulation with the higest corresponding cell type signature score. e) Copy number variations (CNVs)

estimation in cell clusters in patient tumor scRNA-Seq dataset. Top, The non-malignant cell subpopulations as reference. Bottom, CNVs in cell clusters. Text in green indicates the non-malignant cell subpopulations. Figures were adapted from Hai & Hoffmann et al., 2021.

The assignment of cells to the seven cell types were annotated in UMAPs (Figure 2.10a-b). The cell type composition varied among the different patient tumor samples (Figure 2.10c). I identified marker genes corresponding to the assigned cell types in the patient tumor scRNA-Seq dataset (Figure 2.10d, Table 6). Notably, several typical markers were detected, including EGFR for malignant cells, MBP for oligodendrocytes, CD163 for macrophages, CLDN5 for endothelial cells, PDGFRB for pericytes, and CD2 for T-cells (Figure 2.10d, Table 6).

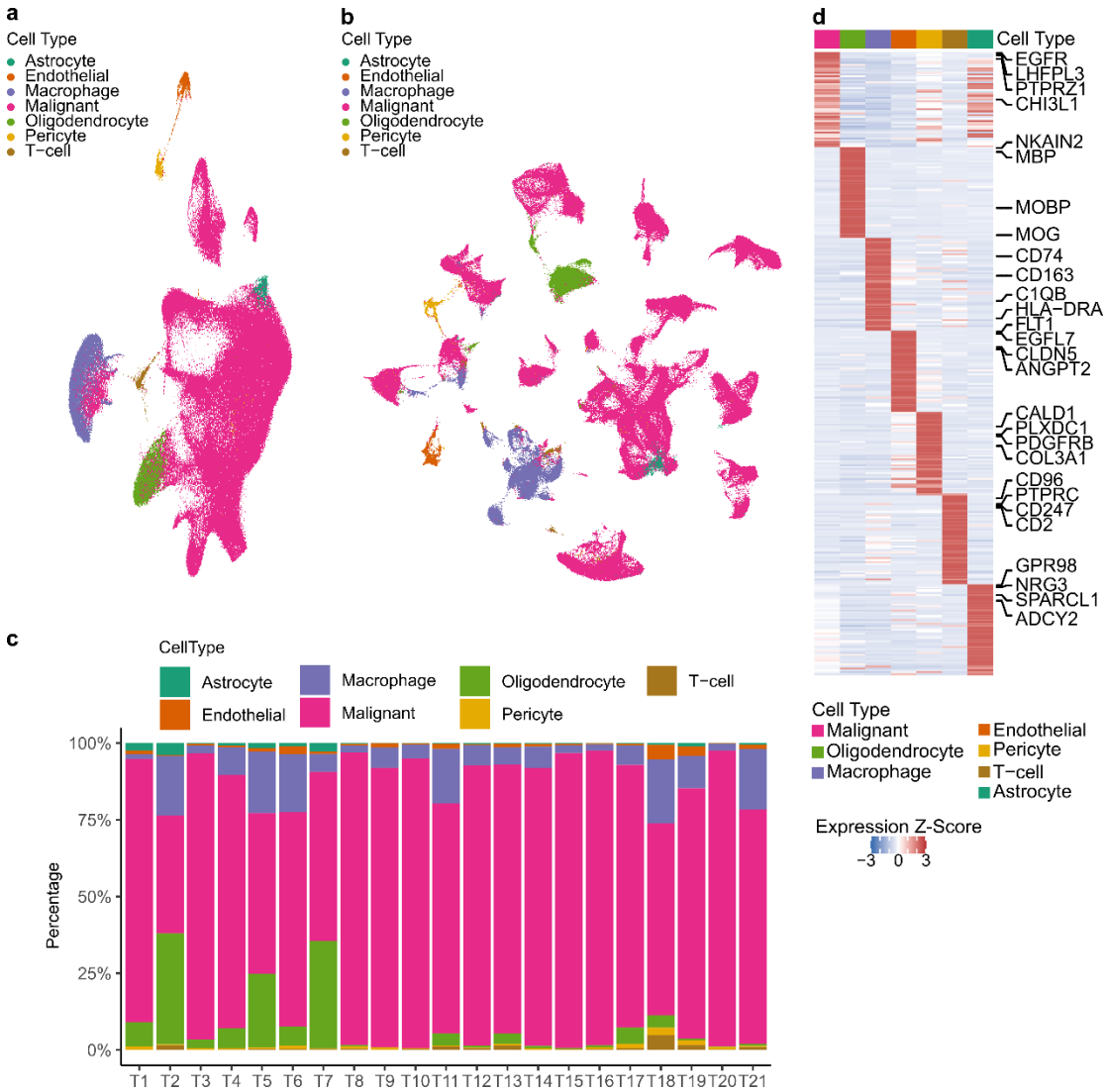


Figure 2.10 **Cell types in GB patient tumor samples.** The cells from 21 GB patient tumor sample scRNA-Seq were analyzed. a-b) UMAPs showing single cells colored by seven cell types with “anchor” integration (a) or without integration (b). c) The cell type composition in each patient tumor sample. d) Average expression levels of cell

type signatures. Expression levels were scaled to z-score and winsorized at -3 and 3. Figures were adapted from Hai & Hoffmann et al., 2021.

Table 6 **Cell type signatures identified in the patient tumor scRNA-Seq dataset.** Seven cell type signatures were obtained by differential expression analysis in the 21 GB patient tumor sample scRNA-Seq dataset. Table was adapted from Hai & Hoffmann et al., 2021.

Malignant (n = 49): EGFR, LHFPL3, SNTG1, PTPRZ1, RGS6, NRCAM, DPP6, ROBO2, CTNNA2, NRXN1, GRIK2, TNC, NOVA1, GLIS3, DGKG, SEC61G, GPM6A, RP11-40F8.2, LRP1B, MAP2, MEG3, KCND2, DCLK2, LPHN3, CHI3L1, VEGFA, NAV2, SLC4A4, SOX6, RORA, LSAMP, NLGN1, LINC00511, CSMD1, CDK14, NKAIN3, DLGAP1, TMEM178B, DGKB, SLC35F1, FMN2, PTN, RFX4, CADPS, TRIO, DENND2A, NLGN4X, ITGB8, TRIM9
Endothelial (n = 42): FLT1, EGFL7, ABCB1, INSR, VWF, ANO2, GALNT18, HSPG2, CLDN5, ANGPT2, PTPRB, ATP10A, DOCK9, MECOM, NOX4, GPR116, ERG, PTPRM, ELTD1, RASGRP3, PREX2, SORBS2, MYRIP, EPAS1, LPHN2, CTGF, SLC39A10, PLXNA2, HECW2, PLEKHG1, PTPRG, GRAPL, AC010084.1, MYO10, FLI1, LDB2, NOSTRIN, NOTCH4, ARL15, NR5A2, CALCRL, SLC7A5
Macrophage (n = 48): PLXDC2, TBXAS1, APBB1IP, FRMD4A, SLC11A1, SFMBT2, DOCK8, ARHGAP24, FYB, CD74, CSF2RA, SLCO2B1, ST6GAL1, MSR1, C10orf11, KCNQ3, SRGN, DOCK4, CPM, CD163, SAT1, MEF2C, MYO1F, MEF2A, ADAM28, FMN1, PIK3R5, RP11-556E13.1, C3, FCGBP, ATP8B4, CCL3, C1QB, SRGAP2, PALD1, MS4A6A, HLA-DRB1, SAMSN1, STAB1, SYK, RP11-624C23.1, HLA-DRA, RCSD1, RGS1, DENND3, INPP5D, MERTK, OLR1
Pericyte (n = 43): MIR4435-1HG, CCDC102B, EBF1, FN1, UACA, SLC38A11, CTD-3179P9.1, CALD1, GRM8, CACNA1C, PRR16, PLXDC1, PDGFRB, TRPC6, COL1A1, COL18A1, CDH6, COL3A1, RNF152, EDNRA, LAMC3, MIR143HG, RBPMS, LINC00152, LAMA4, GUCY1A2, CCDC3, ZEB1, ENPEP, LAMB1, SLIT3, EPS8, TPM1, DCN, IGFBP7, PTEN, GJC1, SVIL, COBLL1, ACTA2, RP11-649A16.1, NR2F2-AS1, INPP4B
Oligodendrocyte (n = 47): NKAIN2, ST18, MBP, PLP1, CTNNA3, MIR219-2, IL1RAPL1, TMEM144, RNF220, SPOCK3, EDIL3, SLC24A2, UNC5C, CLDN11, PEX5L, CERCAM, CNTNAP4, PIP4K2A, CNDP1, SLC44A1, MAP7, DOCK5, PLCL1, TF, KIRREL3, AK5, PCSK6, MAN2A1, C10orf90, SLC5A11, ANK3, MOBP, ENPP2, CARNS1, PLEKHH1, ABCA2, KCNMB4, TTLL7, KLHL32, ZNF536, KIAA1598, CDK18, MYRF, TMEFF2, DNM3, MOG, GRM3
T-cell (n = 46): SKAP1, CD96, THEMIS, SLFN12L, PTPRC, CD247, CD2, STAT4, AC105402.4, TC2N, CCL4, PARP8, SAMD3, CARD11, BCL11B, BCL2, AC104820.2, IKZF1, CCL5, PYHIN1, GRAP2, CCND3, ITGAL, HFM1, SYTL3, RHOH, KIAA1551, STK17B, FAM65B, MBNL1, CD97, IL7R, PDE3B, EMB, RNF213, CDC42SE2, GZMA, ITK, ACAP1, PRKCB, TNFAIP8, PRKCQ, CAMK4, LCP2, LCK, RUNX3
Astrocyte (n = 47): GPR98, NRG3, RNF219-AS1, GPC5, TPD52L1, SPARCL1, HPSE2, MGST1, ADCY2, MGAT4C, NEBL, PLEKHA5, RP11-627D16.1, FAM155A, SLC14A1, KCNN3, PAMR1, MAPK4, ABLIM1, MAOB, COL5A3, PITPNC1, CP, SORBS1, LINC01088, CTNND2, GABRB1, RANBP3L, DCLK1, AQP1, NTRK2, CNTN1, CD38, PRODH, SLC1A2, COLEC12, FUT9, AQP4-AS1, ARHGEF4, DTNA, EFEMP1, DNER, APOE, CDH20, GINS3, PARD3, CCDC85A

I further assigned GB cell states to the malignant cells in the patient tumor scRNA-Seq dataset. I observed that the cell state composition varied among tumors (Figure 2.11a). I calculated CSS in each malignant cell and grouped the cells into four groups based on the quartiles of their CSS values. I found that the percentage of these four CSS

groups in tumors varied (Figure 2.11b). Notably, the tumors exhibiting a higher fraction of AC and MES cells showed a higher proportion of the highest CSS group (Figure 2.11a-b).

To better illustrate the relationship between the CSS and cell state, I employed various visualization methods to highlight different aspects:

- Heatmap: I ordered individual single cells by their CSS values, visualizing their cell state signature scores alongside, and identified a pattern: cells with higher CSS values corresponded to higher AC and MES1 signature scores, whereas cells with lower CSS values displayed higher OPC and NPC1 signature scores (Figure 2.11c).
- UMAP: I visualized the cells in UMAP plots. Cells were clustered according to cell states in the UMAPs. AC and MES1 cells were in regions with higher CSS values, while OPC and NPC1 cells were positioned in areas with lower CSS values (Figure 2.11d).
- 2D plot: I projected the cells onto a two-dimensional (2D) scatterplot based on their cell state signature scores. Cells from the AC and MES states located in the upper part of the plot exhibited remarkably higher CSS values compared to cells in the lower part, which represented OPC and NPC cells (Figure 2.11e).

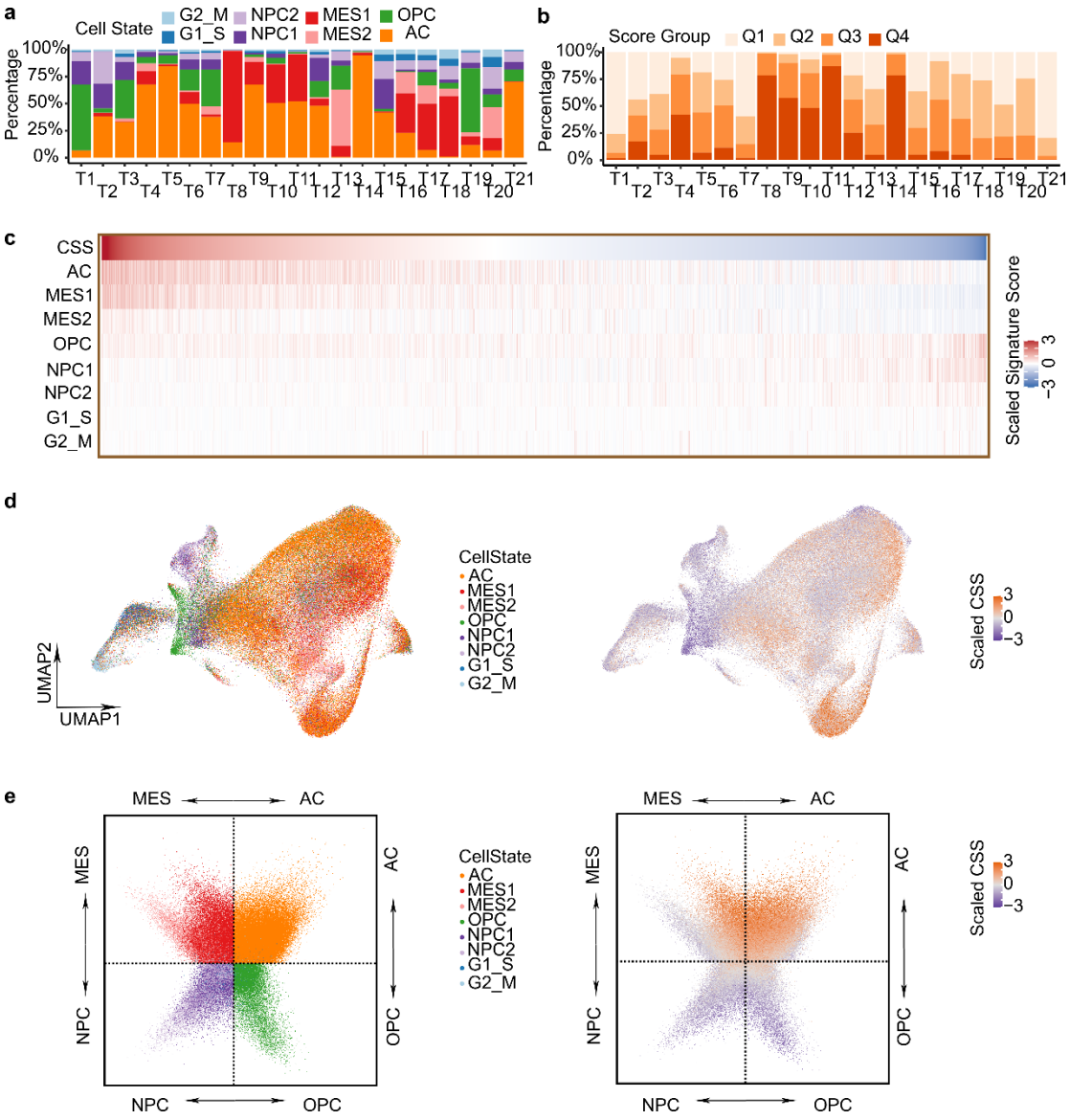


Figure 2.11 **CSS and cell state in GB patient malignant cells.** a) Percentage of cell states in patient tumor samples. b) Percentage of four CSS groups in patient tumor samples. Cells were categorized into four groups based on their CSS values' quartiles. Q1: Cells with the lowest 25% CSS values; Q2: Cells with 25% to 50% CSS values; Q3: Cells with 50% to 75% CSS values; Q4: Cells with the highest 75% CSS values. c) Heatmap depicting CSS and cell state signature scores in each cell. Cells were ordered based on their CSS values. Scores were scaled to z-score and winsorized at -3 and 3. d) UMAPs of cells with anchor integration. Cells were colored by their respective cell states (Left) and CSS values (Right). e) Two-dimensional (2D) embedding of cells based on cell state signature scores. The top-left corner represents cells enriched with higher MES scores, the top-right corner with higher AC scores, the bottom-left corner with higher NPC scores, and the bottom-right corner with enriched OPC scores. Cells were colored by cell states (Left) and CSS values (Right). Figures were adapted from Hai & Hoffmann et al., 2021.

I further investigated the relationship between CSS and cell states in a harmonized GB scRNA-Seq dataset known as "GBmap", which comprised 338,564 annotated cells obtained from 110 donors (Ruiz-Moreno et al., 2022). I applied CSS to all cells within

the GBmap dataset, and visualized the results using UMAPs (Figure 2.12a-c). I found that cells annotated as AC-like and MES-like in Ruiz-Moreno et al., 2022 were predominantly situated in the region exhibiting the highest CSS values, whereas NPC-like and OPC-like cells were clustered in the area with the lowest CSS values; other nonmalignant cells displayed intermediate CSS values (Figure 2.12a-c). Upon projecting the cells onto the 2D plot according to cell state signature scores, the upper part containing AC and MES cells displayed notably elevated CSS values (Figure 2.12d). These findings agree with the observations from the 21-sample GB patient tumor scRNA-Seq dataset (Figure 2.11).

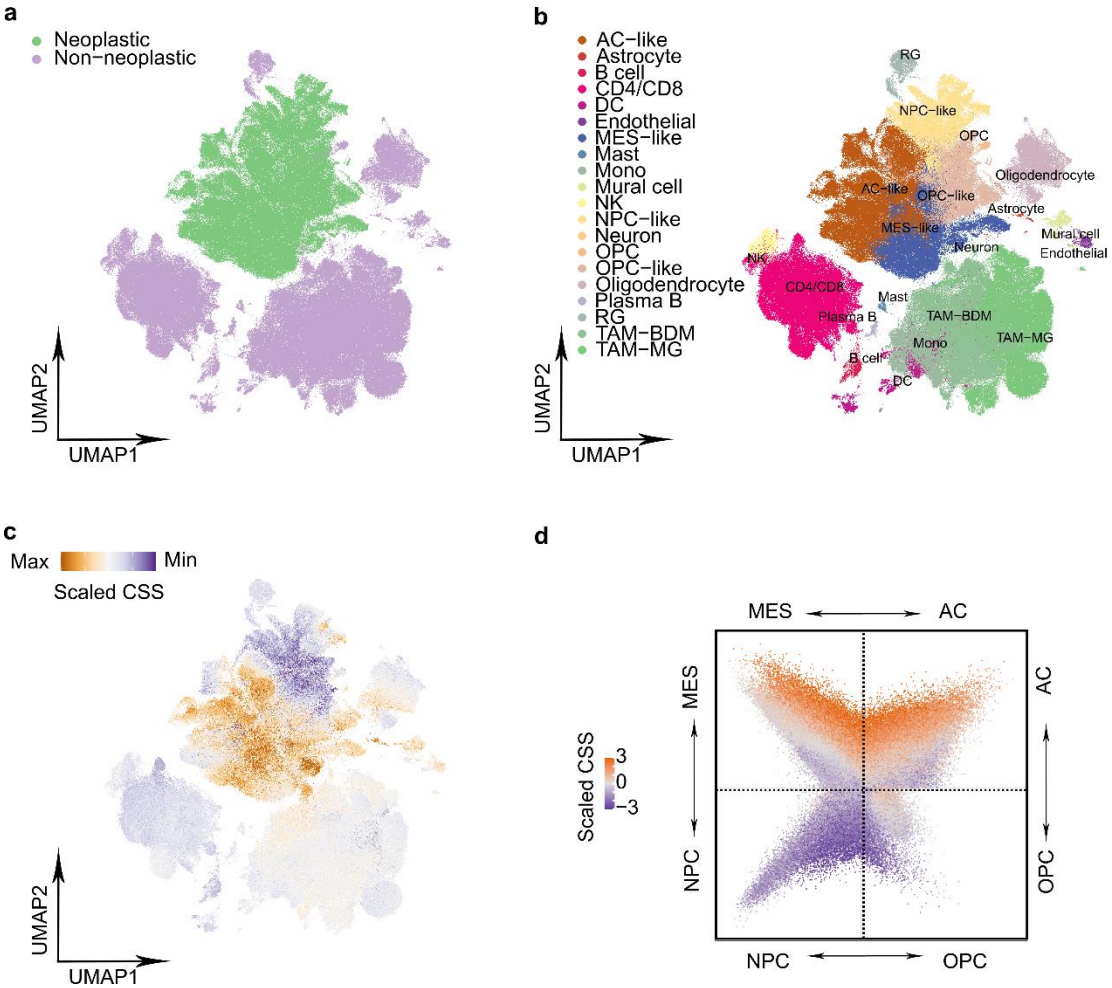


Figure 2.12 **CSS and cell type/state in the GBmap scRNA-Seq dataset.** A harmonized GB scRNA-Seq dataset named 'GBmap' comprised 338564 annotated cells from 110 donors (Ruiz-Moreno et al., 2022). a) A UMAP of all cells is displayed, with colors indicating malignant and nonmalignant annotations as per Ruiz-Moreno et al., 2022. b) A UMAP of all cells is displayed, with colors indicating cell type/state annotations as per Ruiz-Moreno et al., 2022. c) A UMAP of all cells colored based on CSS values. d) 2D embedding of all cells based on cell state signature scores and colored by CSS values.

Subsequently, I conducted tests to determine whether the cell type/state composition

within a sample influenced the CSS values. I found the compositions of cell type, cell state, and CSS groups exhibited significant heterogeneity across the samples within the GBmap dataset (Figure 2.13a-c). Despite this variability, I identified a discernible pattern: samples containing a higher proportion of CSS Q4 group were characterized by an increased proportion of malignant AC and MES cell states, while those within the CSS Q1 group displayed a higher prevalence of malignant NPC cell state (Figure 2.13a-b). Moreover, I found a strong correlation between the CSS groups and the proportions of malignant AC, MES and NPC cell states (Figure 2.13d). Additionally, the proportions of diverse immune cells exhibited a positive correlation with the CSS Q2 group (Figure 2.13d).

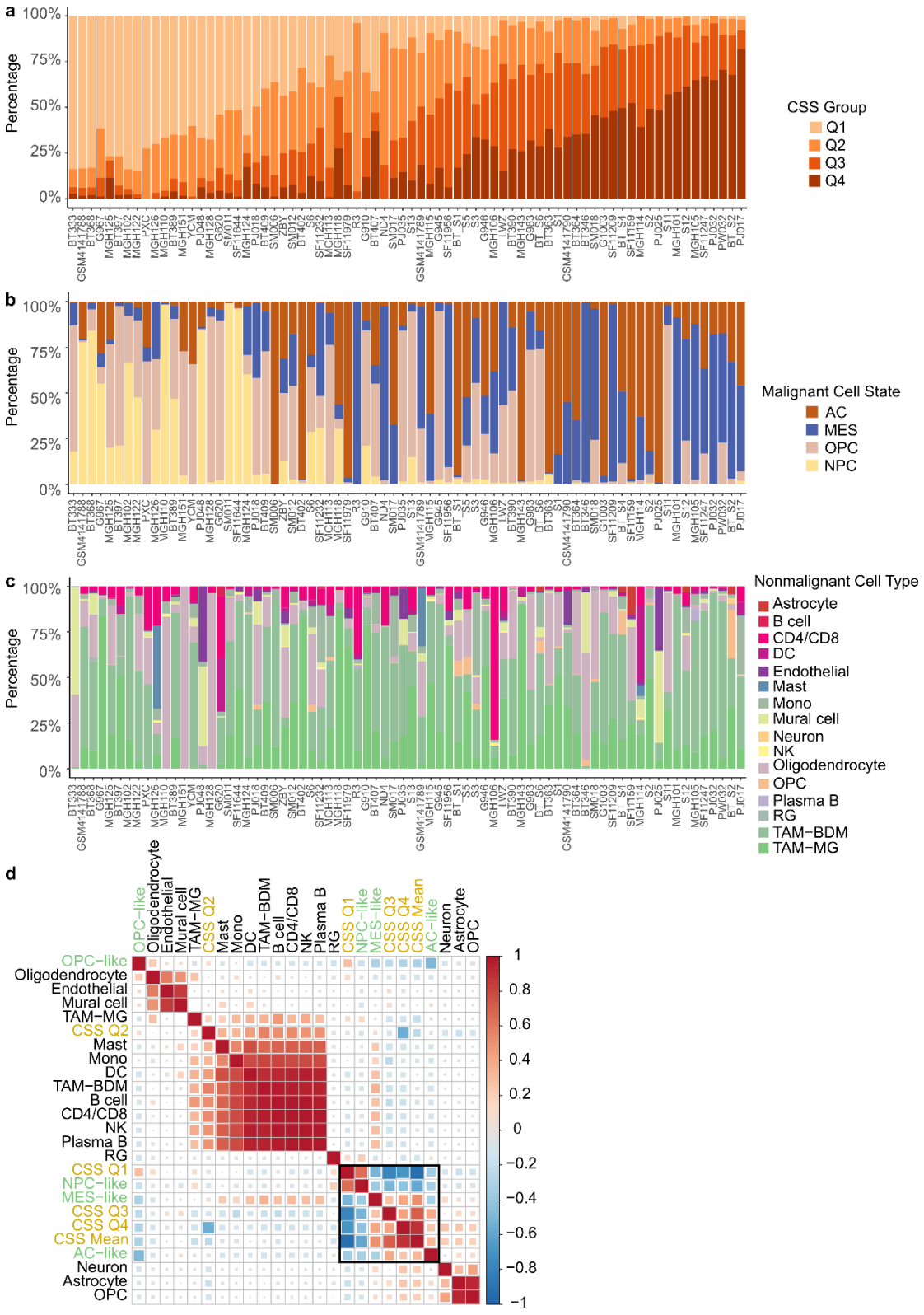


Figure 2.13 **CSS and cell type/state composition in the GBmap scRNA-Seq dataset.** 74 donors with at least 20 malignant cells and 20 nonmalignant cells from the GBmap scRNA-Seq dataset. a) Percentage of four CSS groups in 125486 malignant cells from 74 donors. Samples were sorted by mean CSS values. Cells categorized by quartiles of their CSS values. Q1: Cells with the lowest 25% CSS values; Q2: Cells with 25% to 50% CSS values; Q3: Cells with 50% to 75% CSS values; Q4: Cells with the highest 75% CSS values. b) Percentage of malignant cell states in 125486

malignant cells from 74 donors. c) Percentage of nonmalignant cell type in 108293 nonmalignant cells from 74 donors. d) Heatmap shows pearson correlation coefficient among the percentage of malignant cell states (green label), the percentage of nonmalignant cell types (black label), the percentage of CSS groups (orange label) and the mean of CSS values (oranger label) in each sample. The color and size of square indicates the Pearson correlation coefficient.

In this subsection, I delved deeper into the relationship between the connectivity signature and cell states using both 21-sample GB patient tumor scRNA-Seq dataset and the "GBmap" harmonized 110-patient tumor scRNA-Seq dataset. Despite the considerable heterogeneity in cell state composition among patients, I found that the AC and MES cell states exhibited a strong positive correlation with the CSS values. Conversely, the NPC cell state displayed a negative correlation with the CSS values. In the following subsection, I will further validate these relationships using bulk RNA-Seq datasets.

2.1.5 The connecivity signature in TCGA GB RNA-Seq dataset

Given that bulk RNA-Seq data is often more readily accessible compared to scRNA-Seq data, I assessed the effectiveness of the connectivity signature using RNA-Seq data from a cohort of 230 samples with GB IDH wt in TCGA.

I assigned three expression subtypes (mesenchymal [MS], classical [CL], and proneural [PN], as defined by Wang Q. et al., 2017) and the prevailing cell state (as defined by Neftel et al., 2019) to 230 samples in the TCGA GB RNA-Seq dataset. Among the samples belonging to the MS expression subtype, I found the MES1 cell state predominated. In the case of the CL subtype samples, the majority exhibited the AC cell state. Conversely, PN subtype samples displayed a variety of cell states, including AC, OPC, and NPC cell states (Figure 2.14a).

I applied the CSS to the samples. I found that the MS subtype samples exhibited the highest CSS values, the CL subtype samples displayed intermediate CSS values, and the PN subtype samples had the lowest CSS values (Figure 2.14b).

Furthermore, my investigation into the gene mutation status of the TCGA GB samples revealed a correlation between the CSS values and three genes—namely NF1, TP53, and PTEN (with FDR < 0.25 among genes mutated in at least 5% of samples, Figure 2.14c). Notably, I observed that samples harboring TP53 mutations (found in 24% of

GB samples) exhibited lower CSS values compared to TP53 wt samples (Figure 2.14c). This observation aligns with the notion that basal TP53 expression is essential for mesenchymal stem cells and for nanotube development in astrocytes (Boregowda et al. in 2018, Wang Y. et al. in 2011). Conversely, I found that samples carrying PTEN mutations (detected in 30% of GB samples) displayed higher CSS values compared to PTEN wt samples. This observation aligns with a previous study indicating that astrocytes overexpressing PTEN had shorter microtube protrusions (Hohensee et al. in 2017). Moreover, I noticed that samples with NF1 mutations (occurring in 15% of GB cases) exhibited higher CSS values. NF1 mutants have been associated with tumor invasiveness and the MS subtype (Fadhlullah et al. in 2019, Verhaak et al. in 2010). Interestingly, even among the samples belonging to the MS subtype, NF1 mutants still displayed higher CSS values, suggesting that NF1 correlates with CSS independently of the MS subtype.

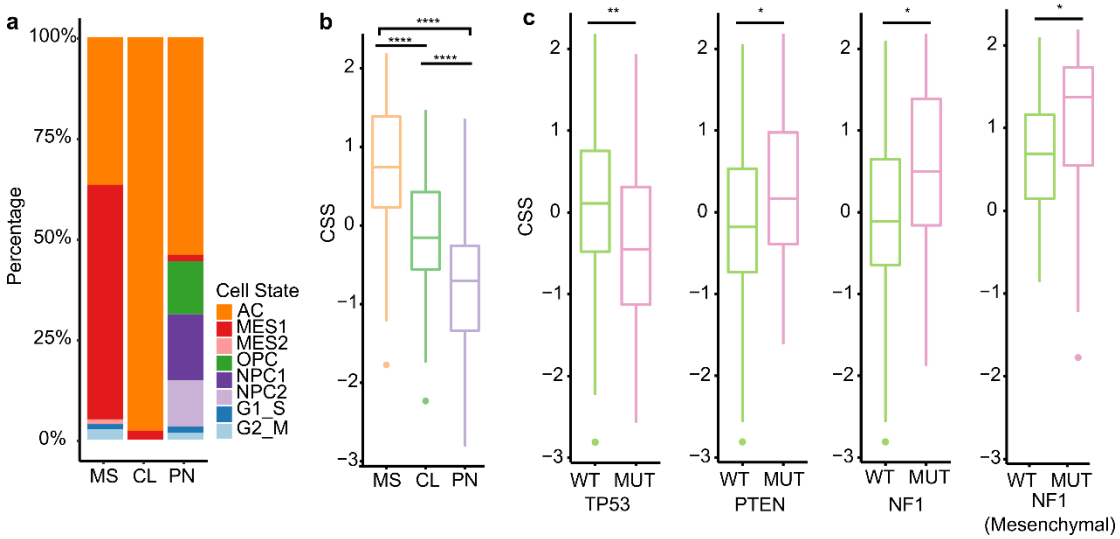


Figure 2.14 **CSS association with expression subtype and gene mutation.** 230-sample TCGA IDH wt GB cohort. a) Percentage of sample`s dominant cell state in three expression subtypes. 81 Mesenchymal (MS), 87 Classical (CL) and 62 Proneural (PN). b) CSS in three expression subtypes. c) CSS in mutation states of TP53 (57 mutants vs. 173 wts), PTEN (76 mutants vs. 154 wts), and NF1 (35 mutants vs. 195 wts; 16 mutants vs. 63 wts in MS subtype). *, p-value < 0.05; **, p-value < 0.01; ***, p-value < 0.001. Mann-Whitney U test. Figures were adapted from Hai & Hoffmann et al., 2021.

To explore the relationship between CSS and cell states in bulk RNA-Seq data, I employed bulk deconvolution methods to quantify the abundance of different cell types/states within the 230-sample TCGA GB RNA-Seq dataset.

I determined the proportions of different cell types/states in the bulk RNA-Seq data

using a signature matrix derived from 21-sample GB patient tumor scRNA-Seq data through CIBERSORTx (Newman et al., 2019). A large number of samples exhibited dominance by the AC cell state (Figure 2.15a). Notaly, correlation analysis revealed a positive association between CSS values and the proportion of MES1, while a negative correlation was observed with the NPC1 cell states (Figure 2.15b-c).

To validate the findings obtained from CIBERSORTx, I used an alternative bulk deconvolution method, GBMDeconvoluteR (Ajaib et al., 2023). GBMDeconvoluteR, designed specifically for GB, made use of multiple GB scRNA-Seq datasets as references (Ajaib et al., 2023). Consistent with the results from CIBERSORTx, I found that high MES abundance in the samples was associated with increased CSS values, while high NPC abundance showed a correlation with lower CSS values (Figure 2.15d-e). Additionally, leveraging the enhanced capabilities of GBMDeconvoluteR, which includes more detailed immune cell annotation and refines GB cell state markers specifically for bulk RNA-Seq, I observed a positive correlation between CSS values and various immune cells (Figure 2.15d).

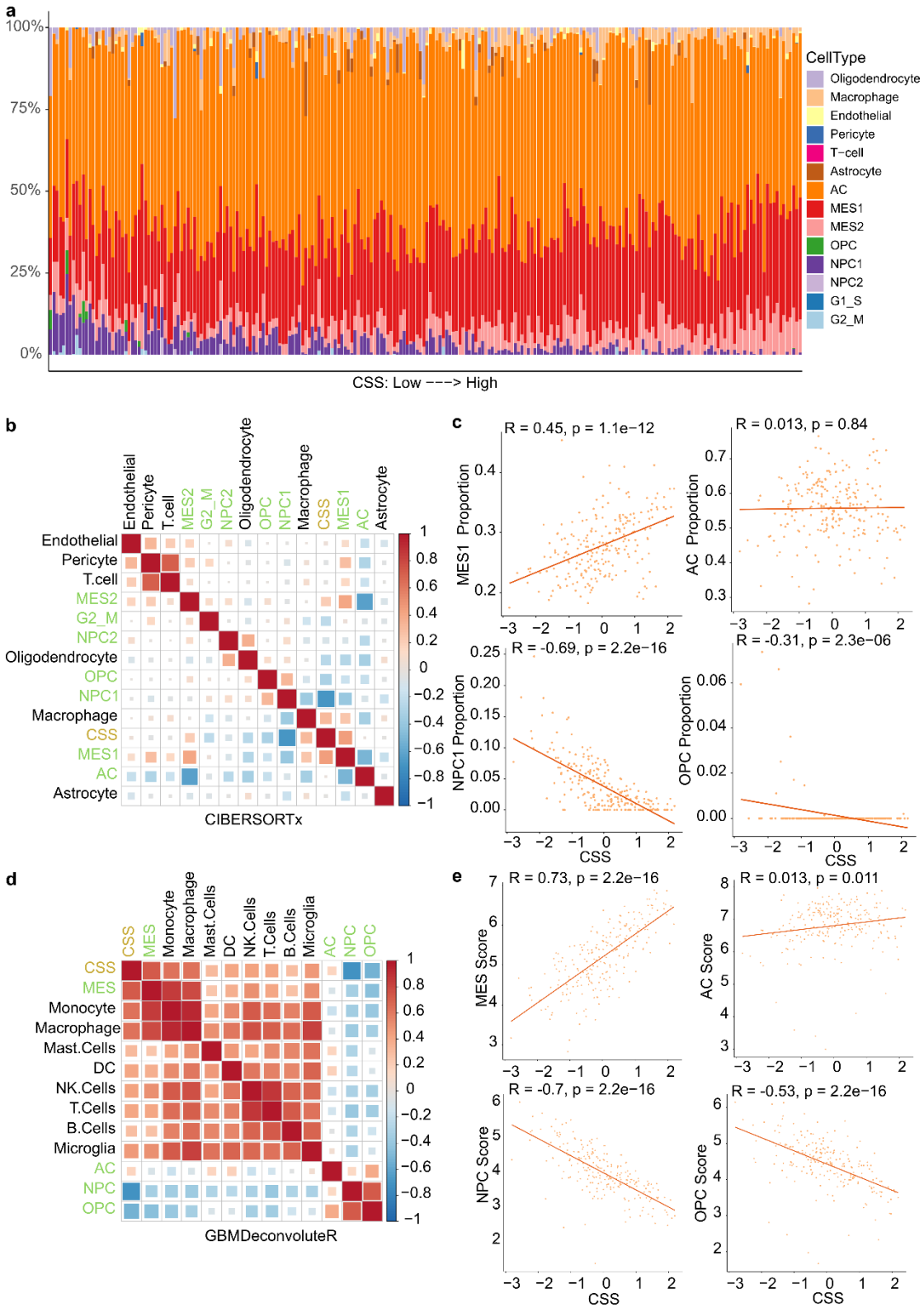


Figure 2.15 **CSS and deconvoluted cell type/state in the TCGA GB RNA-Seq dataset.** Bulk RNA-Seq data from 230-sample TCGA IDH wt GB cohort. a-c) The RNA-Seq data of samples were deconvoluted into various cell types/states using CIBERSORTx with the signature matrix generated from the scRNA-Seq data of 21-sample GB patient cohort. Samples were sorted by their CSS values. a) Cell type/state composition in samples. b) Pearson correlation coefficients among the percentage of malignant cell states (green label), the percentage of nonmalignant cell types (black label), the CSS values (oranger label) in each sample. The color and size of square

indicates the Pearson correlation coefficient. c) Scatterplots showing correlation between the proportions of cell states and CSS values. p value: Pearson correlation d-e) Deconvolution using GBMDeconvoluteR. d) Pearson correlation coefficient between cell types/states scores and CSS values. The color and size of square indicates the Pearson correlation coefficient. e) Scatterplots showing correlation between cell state scores and CSS values. p value: Pearson correlation.

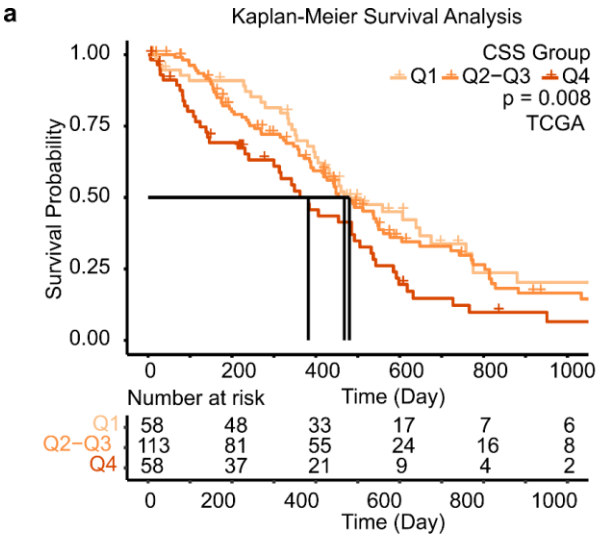
In this subsection, I employed CSS on 230-sample TCGA bulk RNA-Seq dataset and uncovered associations between CSS and three mutated genes. The correlation observed between CSS and deconvoluted cell types/states reaffirmed the conclusions drawn in the SR101 and patient sample scRNA-Seq dataset. This reaffirmation highlights the applicability of CSS to both bulk and single-cell RNA-Seq datasets. The forthcoming section will delve into an investigation of the relationship between the CSS and patient survival.

2.1.6 The connectivity signature and patient survival

Previous studies have demonstrated that TM-connected glioma cells are associated with therapy resistance (Osswald et al., 2015; Weil et al., 2017). Nevertheless, the effect of these cells on patient survival has yet to be fully elucidated.

To assess the impact of cell connectivity on patient survival, I applied CSS to the 230-sample TCGA IDH wt GB cohort. I classified the samples into three groups based on CSS quartiles (Q1, Q2-Q3, Q4). Kaplan-Meier (KM) survival analysis of the CSS groups demonstrated that the CSS Q1 group exhibited favorable survival, while Q4 group exhibited notably poorer survival outcomes (Figure 2.16a).

Furthermore, I conducted a Cox proportional hazards regression (Coxph) survival analysis using the continuous CSS values in samples. The results revealed a significant association between CSS values and overall patient survival (Figure 2.16b). Importantly, this association remained significant even after accounting for covariates such as age, gender, and expression subtypes (Figure 2.16b).



b Cox Proportional Hazards Regression Survival Analysis

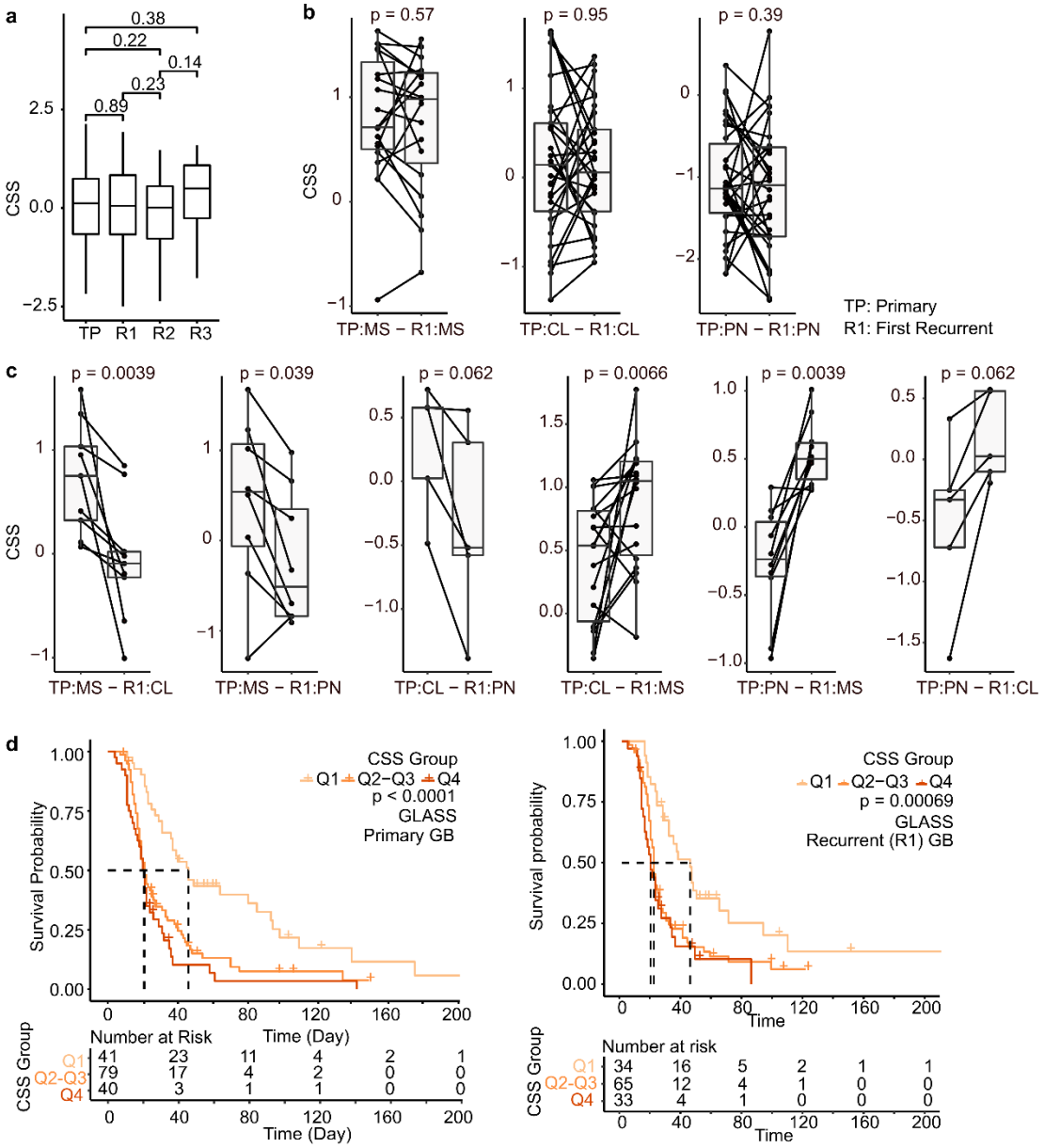
TCGA (n = 230)	Exp. coef. (95% int.)	p-value
CSS	1.279 (1.094-1.52)	0.000888
Adjusted for Age and Gender	1.221 (1.045-1.443)	0.00662
Adjusted for Age, Gender and Expression Subtype	1.293 (1.057-1.589)	0.0107

Figure 2.16 **CSS and patient survival in the TCGA GB cohort.** a) Kaplan-Meier (KM) survival analysis of 230 GB patients. Samples were categorized into three groups by quartiles of their CSS values. Q1: patients with the lowest 25% CSS values; Q2-Q3: patients with 25% to 75% CSS values; Q4: patients with the highest 75% CSS values. b) Cox proportional hazards regression (Coxph) survival analysis of 230 GB patients with CSS values. Exp. coef. (95% int.), exponentiated coefficients with 95% confidence intervals. Figures were adapted from Hai & Hoffmann et al., 2021.

To investigate CSS values during glioma evolution, I applied CSS to RNA-Seq data from the Glioma Longitudinal Analysis Cohort (GLASS, Varn et al., 2022). I found no significant changes of CSS values between primary and recurrent tumors (Figure 2.17a). In cases where patients had the same expression subtype in both primary and recurrent samples, CSS values did not have significant changes (Figure 2.17b). However, it is noteworthy that 49% of patients exhibited a switch in expression subtypes in recurrent tumors. When such switches occurred, corresponding changes were observed in CSS values (Figure 2.17c): CSS values decreased when a primary MS tumor transitioned to recurrent CL or PN tumor, whereas CSS values increased when primary CL or PN tumors switched to recurrent MS tumors (Figure 2.17c).

I conducted KM survival analysis on both primary and recurrent tumors, revealing a more favorable overall survival in the CSS Q1 group, whereas the Q4 group exhibited poorer survival outcomes (Figure 2.17d). Moreover, I applied Coxph survival analysis to assess the relationship between continuous CSS values and OS or surgery interval

in primary and recurrent tumors. The analysis revealed a correlation between high CSS values and shorter overall survival, as well as an association between high CSS values and a reduced interval to the next relapse after surgery (Figure 2.17e).



e

Cox Proportional Hazards Regression Survival Analysis

OS: Overal Survival; SI: Surgery Interval

Primary		Exp. coef. (95% int.)	p-value
CSS	OS	1.498 (1.270-1.766)	1.56×10
	SI	1.621 (1.371-1.915)	1.50×10^{-8}
Adjusted for Age and Gender	OS	1.388 (1.166-1.652)	0.00023
	SI	1.505 (1.267-1.787)	3.18×10^{-6}
Adjusted for Age, Gender and Expression Subtype	OS	1.295 (1.026-1.635)	0.0297
	SI	1.476 (1.184-1.84)	3.18×10^{-6}

Recurrent R1		Exp. coef. (95% int.)	p-value
CSS	OS	1.587 (1.295-1.944)	8.22×10^{-6}
	SI	1.685 (1.394-2.035)	6.34×10^{-8}
Adjusted for Age and Gender	OS	1.357 (1.094-1.683)	0.0055
	SI	1.588 (1.296-1.946)	8.48×10^{-6}
Adjusted for Age, Gender and Expression Subtype	OS	1.246 (0.992-1.565)	0.058
	SI	1.482 (1.199-1.832)	2.71×10^{-4}

Figure 2.17 **CSS in the GLASS primary and recurrent samples.** Primary and recurrent GB samples in Glioma Longitudinal Analysis Cohort (GLASS). a) CSS values

in primary (TP: n = 161) and recurrent (R1: n = 166, R2: n = 34 and R3: n = 10) tumors. Mann-Whitney U test. b-c) CSS values in TP and R1 tumors with the same expression subtypes (b) and the switched expression subtypes (c). Line indicates TP-R1 pair in the same patient. Paired Wilcoxon signed-rank test. d) KM survival analysis in primary (Left, n = 160) and recurrent tumors (Right, n = 132). Samples were categorized into three groups by quartiles of their CSS values. Q1: patients with the lowest 25% CSS values; Q2-Q3: patients with 25% to 75% CSS values; Q4: patients with the highest 75% CSS values. e) Coxph survival analysis of both overall survival (OS) and surgery interval (SI) in primary (Left, n = 160) and recurrent tumors (Right, n = 132) with CSS values. Exp. coef. (95% int.), exponentiated coefficients with 95% confidence intervals.

In this section, I examined the correlation between CSS values and patient survival in both the TCGA cohort and the GLASS primary and recurrent cohort. The findings indicated that higher CSS values were linked to poorer survival outcomes. These results suggest that the CSS could potentially serve as a valuable prognostic indicator. In the subsequent section, I will delve into the identification of the key gene within the connectivity signature.

2.1.7 CHI3L1 as a robust marker in connectivity

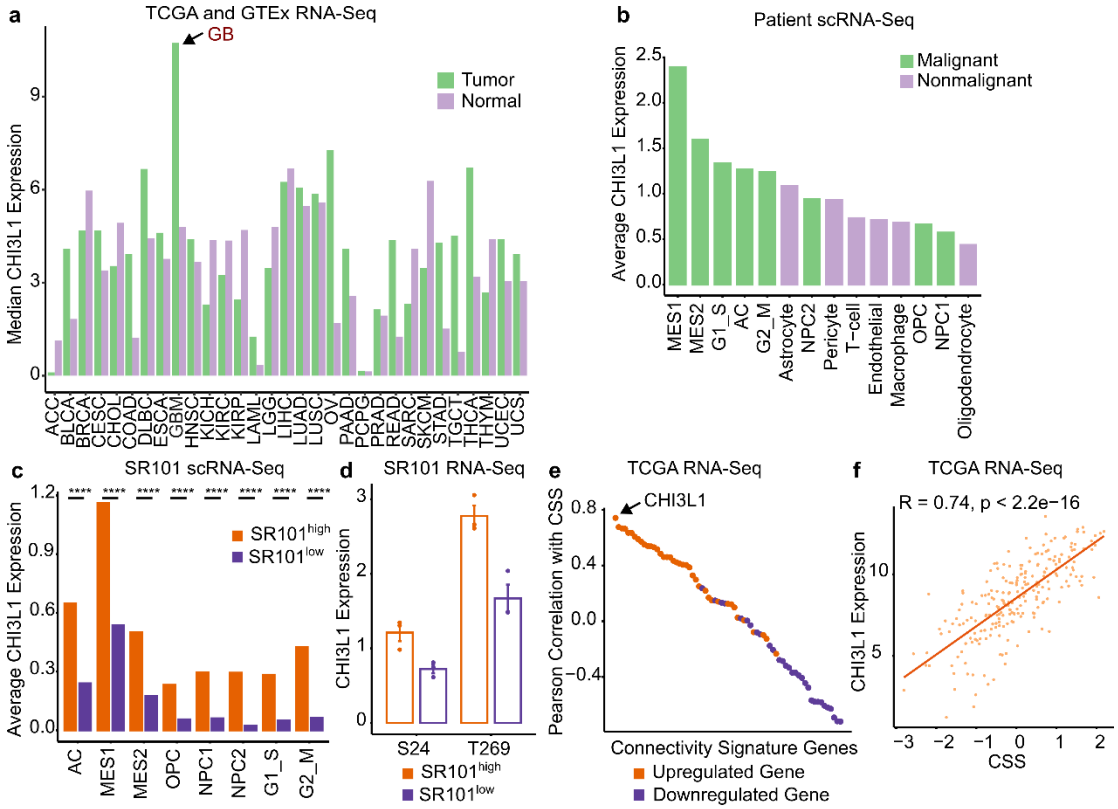
In the earlier subsections, CHI3L1 was identified as the most upregulated gene in $SR101^{high}$ cells as compared to $SR101^{low}$ cells in the scRNA-Seq dataset and was also notably upregulated in the $SR101^{high}$ samples in the RNA-seq dataset (Figure 2.1c, Table 3, Figure 2.2d). This suggests that CHI3L1 might play a role in connectivity. In this subsection, I delve into examining the role of CHI3L1 in GB.

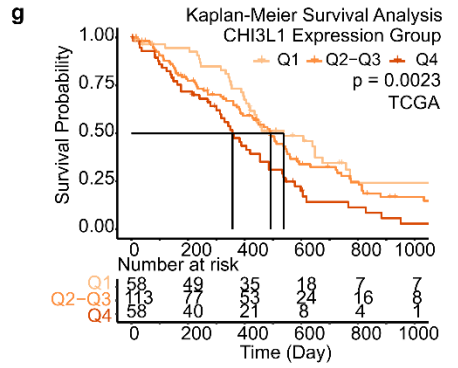
Firstly, I investigated the expression level of CHI3L1 in 31 tumor types and healthy tissues in the TCGA and Genotype-Tissue Expression (GTEx) RNA-Seq datasets (Figure 2.18a). I found CHI3L1 to be highly expressed in GB compared to other types of tumors and healthy tissues, consistent with the observation in the GB patient tumor scRNA-Seq dataset, where CHI3L1 exhibited higher expression in GB malignant cells (Figure 2.10d, Table 6). To be more specific, CHI3L1 had higher expression in MES, AC, and cycling cell states compared to NPC, OPC, and all nonmalignant cell types (Figure 2.18b). These findings suggest that CHI3L1 could serve as a GB marker.

To delve further, I found CHI3L1 exhibited significantly higher expression levels in $SR101^{high}$ cells compared to $SR101^{low}$ cells across each cell state in the SR101 scRNA-Seq dataset (Figure 2.18c), as well as in each PDGCL in the RNA-Seq dataset (Figure 2.18d). Additionally, CHI3L1 exhibited the highest correlation coefficient with the CSS

among the 71 genes in the connectivity signature (Figure 2.18e-f). These results suggest that CHI3L1 can also serve as a marker for the highly connected GB cells.

Furthermore, I observed higher expression levels of CHI3L1 linked to poorer survival outcomes in the KM survival analysis of TCGA GB samples (Figure 2.18g). This association remained significant even after adjusting for covariates such as age, gender, and expression subtypes, using the Coxph survival model (Figure 2.18h). Similar results were observed in the GLASS GB dataset, where patients with high CHI3L1 expression exhibited worse survival outcomes and a shorter interval to the next relapse after surgery. These results strongly indicate that CHI3L1 can serve as a prognostic marker for GB patients.

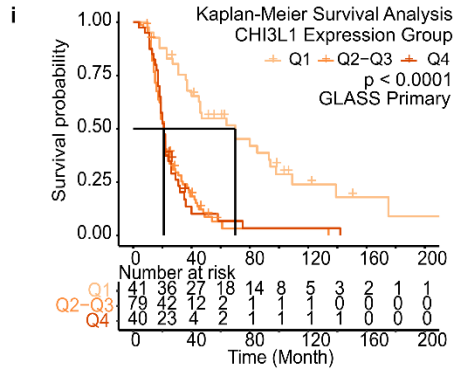




h

Cox Proportional Hazards Regression Survival Analysis

TCGA	Exp. coef. (95% int.)	p-value
CHI3L1	1.122 (1.047-1.202)	0.0011
Adjusted for Age and Gender	1.089 (1.015-1.169)	0.018
Adjusted for Age, Gender and Expression Subtype	1.098 (1.005-1.201)	0.0391



j

Cox Proportional Hazards Regression Survival Analysis

GLASS Primary		Exp. coef. (95% int.)	p-value
CHI3L1	OS	1.143 (1.091-1.198)	2.00E-08
	SI	1.165 (1.109-1.224)	1.37E-09
Adjusted for Age and Gender	OS	1.1103 (1.0547-1.169)	6.54E-05
	SI	1.1338 (1.076-1.195)	5.31E-04
Adjusted for Age, Gender and Expression Subtype	OS	1.0967 (1.0223-1.176)	9.97E-03
	SI	1.1331 (1.0579-1.214)	3.62E-04

Figure 2.18 **CHI3L1 association with GB malignant cell, connectivity and patient survival.** a) CHI3L1 expression levels (median of log2[TPM]) in TCGA patient tumor samples from 31 cancer types and GTEx normal samples from 31 tissues. Data from GEPIA webpage (http://gepia.cancer-pku.cn). b) CHI3L1 expression levels (mean of normalized counts) in the 21-sample patient tumor scRNA-Seq dataset. c) CHI3L1 expression levels (mean of normalized counts) in SR101 groups in each cell state in SR101 scRNA-Seq dataset. d) CHI3L1 expression levels (log2[FPKM]) in SR101 groups in each PDGCL in SR101 RNA-Seq dataset. e) Pearson correlation coefficient

between expression levels of 71 connectivity genes (log2[FPKM]) and CSS in the TCGA GB RNA-Seq dataset. f) Pearson correlation coefficient between CHI3L1 expression levels (log2[FPKM]) and CSS in the TCGA GB RNA-Seq dataset. g) KM survival analysis in the TCGA GB samples. Samples were categorized into three groups by quartiles of their CHI3L1 expression levels (log2[FPKM]). Q1: patients with the lowest 25% expression levels; Q2-Q3: patients with 25% to 75% expression levels; Q4: patients with the highest 75% expression levels. h) Coxph survival analysis in the TCGA GB samples with CHI3L1 expression levels (log2[FPKM]). Exp. coef. (95% int.), exponentiated coefficients with 95% confidence intervals. i) KM survival analysis in the GLASS primary samples. j) Coxph survival analysis in the GLASS primary samples. Figures were adapted from Hai & Hoffmann et al., 2021.

To assess the functional properties of CHI3L1, PDGCLs with overexpressed (OE) CHI3L1 were generated. These cells were then subjected to RNA-Seq, mass spectrometry-based proteomics and phosphoproteomics experiments for further evaluation.

I performed differential expression analyses in the RNA-Seq, proteomics and phosphoproteomics datasets of CHI3L1 OE samples (Figure 2.19a-c, Supplementary Table 1-3). In addition to the artificially overexpressed CHI3L1, several genes were overlapping between the DEGs from the RNA-Seq data and the differentially expressed proteins (DEPs) from the proteomics data. These genes included AC cell state markers such as SPARCL, SPARCL1, and CST3, OPC cell state marker FABP5, and NPC cell state marker UCHL1 (Neftel et al., 2019, Figure 2.19a-b). These findings suggest that the increased expression of CHI3L1 impacts various cell states. Notably, at the RNA level, CHI3L1 OE led to alterations in several connectivity signature genes, particularly the overlapping markers between scRNA-Seq-derived and RNA-Seq-derived connectivity signatures, such as AGT, NMB, HOPX, CLU, ID3, APOE, HES6, and DLL1 (Figure 2.19a, Figure 2.2d). At the protein level, six connectivity signature genes demonstrated altered expression (Figure 2.19b). Interestingly, GAP43, a previously identified TM-connectivity marker (Osswald et al., 2015; Weil et al., 2017), exhibited higher phosphorylation levels in CHI3L1 OE samples (Figure 2.19c).

The RNA and protein expression levels of CHI3L1 in GB patients exhibited a strong correlation (Figure 2.19d). The overlapping genes between DEGs and DEPs also displayed a strong correlation in their fold changes (Figure 2.19e). I computed CSS in both RNA-Seq and proteomics datasets from CHI3L1 OE samples, excluding the expression levels of artificially overexpressed CHI3L1. Remarkably, the CSS still exhibited a significant increase in the CHI3L1 OE samples (Figure 2.19f). This result indicates that CHI3L1 OE alone can indeed alter the CSS values. Additionally, the

CHI3L1 OE samples exhibited increased AC signature scores and decreased NPC1 signatue score at both RNA and protein levels (Figure 2.19g-h).

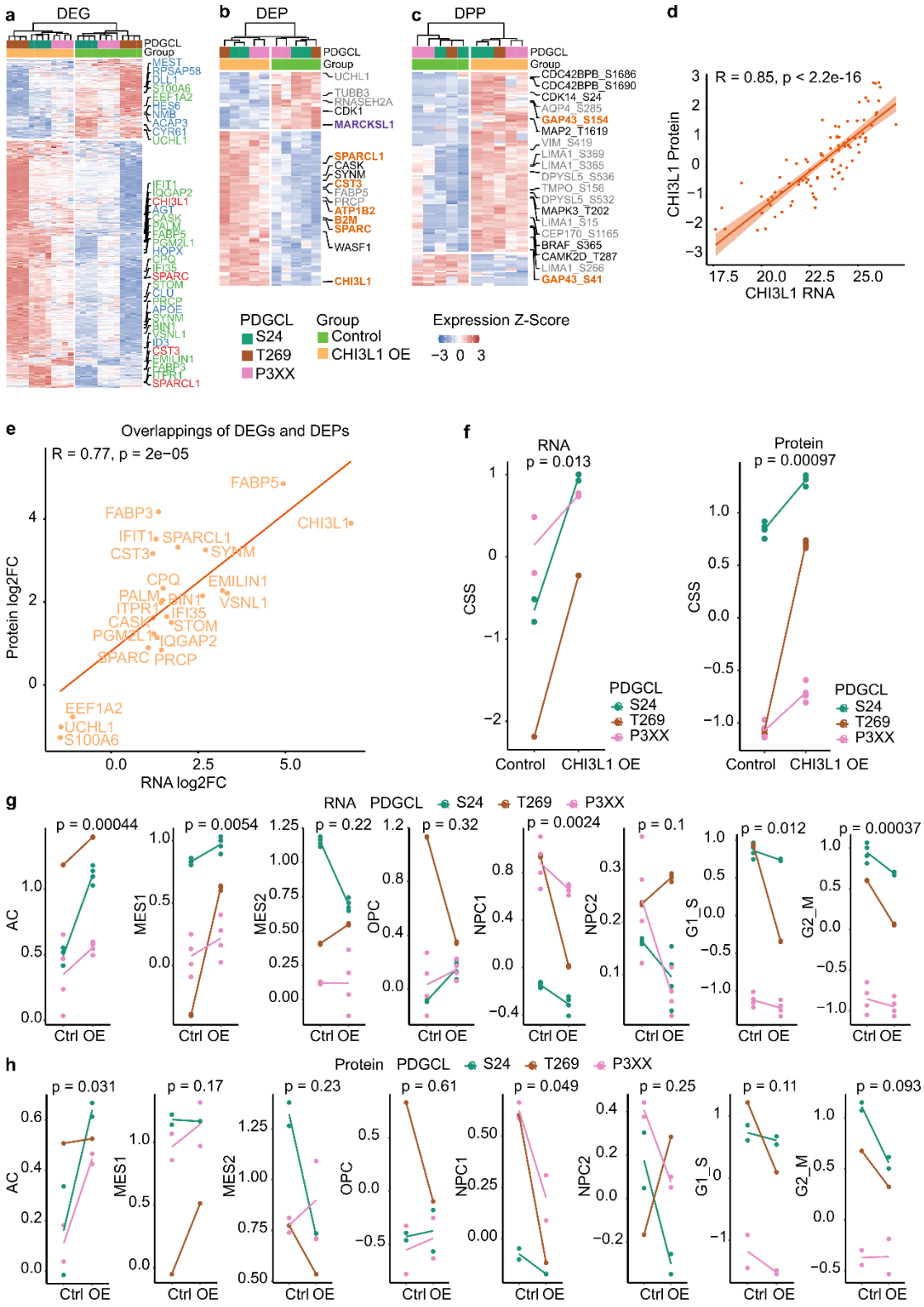


Figure 2.19 **RNA-Seq, proteomics and phosphoproteomics of CHI3L1 overexpressed PDGCLs.** a-c) Differential expression analyses between CHI3L1 overexpressed (OE) and control PDGCLs. a) Differentially expressed genes (DEGs)

in RNA-Seq data. Color coded labels: Blue: Overlapping genes with connectivity signature; Green: Overlapping genes with differential expressed proteins (DEPs); Red: Overlapping gene with both connectivity signature and DEPs. b) DEPs in mass spectrometry-based proteomics dataset. Color coded labels: Grey: Overlapping genes with cell state signatures; Black: Overlapping genes with kinases; Purple: Overlapping gene with downregulated connectivity signature genes; Orange: Overlapping gene with both upregulated connectivity signature genes. c) Differential phosphorated proteins (DPPs) in the phosphoproteomics dataset. Color codes the same as (c). d) Pearson correlation between paired RNA and protein expression levels of CHI3L1 in the 93-patient GB proteogenomic cohort (Wang L.B. et al., 2021). e) Pearson correlation of fold changes in the overlapping features between DEGs and DEPs. f-h) Signatue scores in CHI3L1 OE and control samples were calculated excluding CHI3L1 expression level due to the artificial overexpression. Line and color indicate PDGCL. Paired two-sided t-test. f) CSS in the RNA-Seq (Left) and proteomics dataset (Right). g) Cell state signature scores in the RNA-Seq dataset. h) Cell state signature scores in proteomics dataset.

Moreover, I conducted a comprehensive investigation into the enriched ontologies of CHI3L1 OE DEGs, DEPs, and DPPs. Notably, all three gene sets exhibited enrichment in "neuron projection development" and "cell junction organization" GO terms, both of which are relevant to the formation of TM-connectivity (Figure 2.20a). Intriguingly, the DEGs displayed enrichment in "tube morphogenesis," "regulation of trans-synaptic signaling" and "regulation of MAPK cascade" (Figure 2.20a). It's worth mentioning that a synaptic structure exists between neurons and glioma cells, situated on TMs (Venkataramani et al., 2019). Furthermore, MAPK and NF-kB pathways were found to be activated in a small population of highly active and TM-connected GB cells (Hausmann et al., 2023).

Among the DEGs, DEPs, and DPPs, there are several overlapping genes (Figure 2.20b). Notably, DEGs and DEPs were enriched in the transcription factors NFKB1 and RELA (Figure 2.20c). Prior research has established the association of NF-kB with microtubules (Rai et al., 2015, Hausmann et al., 2023), and it has been shown that CHI3L1 is a target of NF-kB (Hubner et al., 2020; Zhao et al., 2022). Furthermore, a kinase enrichment analysis was conducted. Remarkably, all three gene sets were enriched in the kinases SRC and FYN (Figure 2.20d). SRC has been demonstrated to promote GB tumor proliferation and activate the MAPK pathway (Ahluwalia et al., 2010). FYN, which belongs to the SRC family of kinases, has been implicated in synapse formation (Lim et al., 2009) and is associated with PTEN activity (Dey et al., 2008). Notably, PTEN-mutated GB samples exhibited higher CSS values compared to PTEN wt samples (Figure 2.14c). Taken together, these findings strongly suggest that CHI3L1 plays a pivotal role in TM-connectivity.

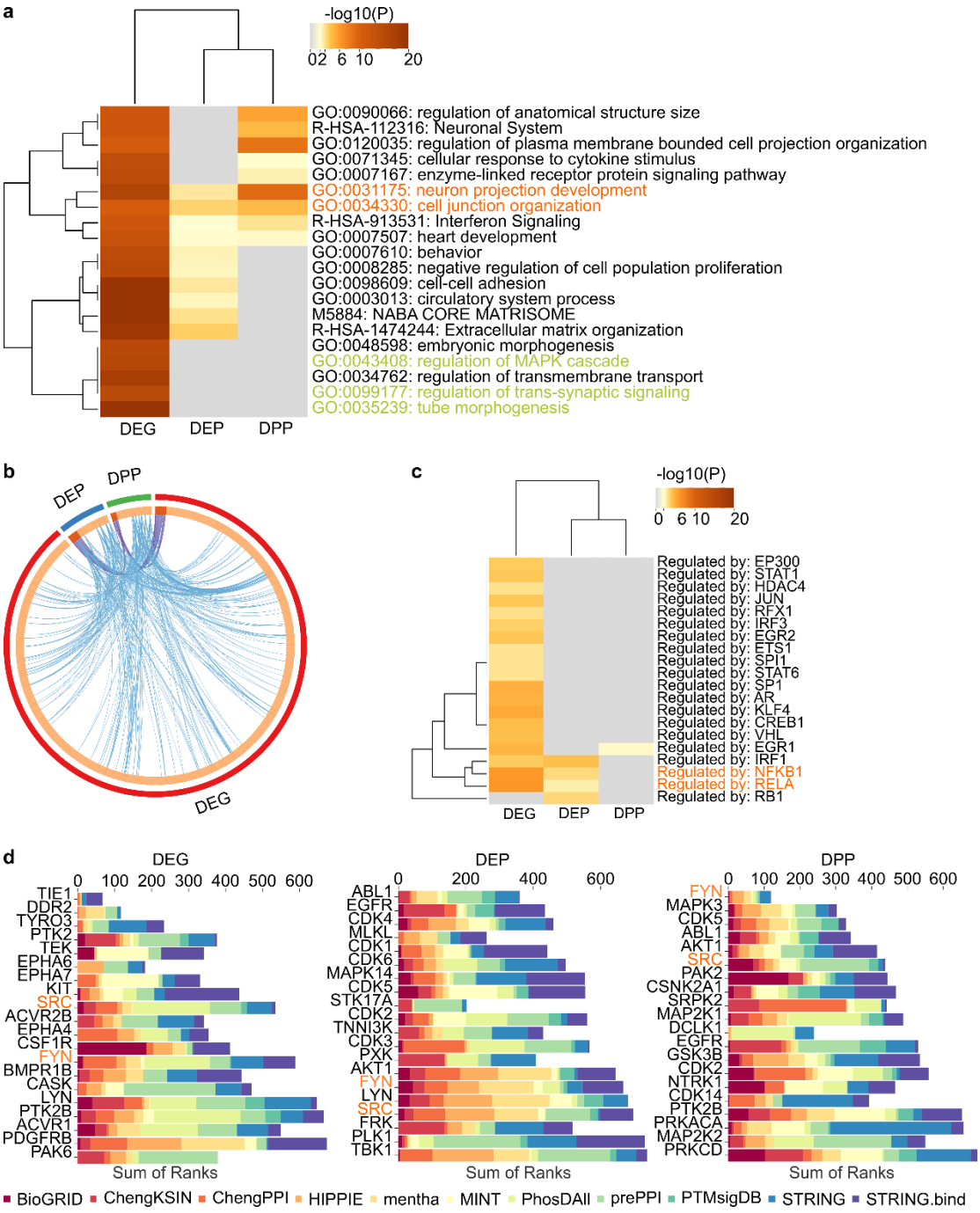


Figure 2.20 **Enriched ontologies of DEGs, DEPs, and DPPs in CHI3L1 overexpressed PDGCLs.** a-c) Analyses using Metascape (Zhou et al., 2019). a) Common enriched ontologies. b) Overlaping genes (purple lines) and genes in the overlapping ontology terms (blue lines). The length of circular arc indicates the number of genes. Outer circle represents the number of genes in each gene set. Inner circle highlights the number of overlapping genes (dark orange). c) Enriched transcriptional factors. d) Enriched kinases. Colors indicate protein-protein-interaction databases. Analyses in KEA3 (Kuleshov et al., 2021).

In this subsection, I introduced CHI3L1 as a robust marker for connectivity. I thoroughly investigated its functional properties across RNA, protein, and phosphorylation levels.

2.1.8 A web tool for data visualization

I designed a user-friendly web tool (https://connectivity-glioma.dkfz.de/) for visualizing gene expression data and metadata from the SR101 scRNA-Seq and 21-sample patient tumor scRNA-Seq datasets (Figure 2.21). This tool empowers users to explore the expression pattern of user-selected gene in the UMAPs of $SR101^{high}$ and $SR101^{low}$ cells, patient tumor malignant cells, or patient tumor cells. Additionally, users can compare specific gene's expression in the $SR101^{high}$ and $SR101^{low}$ groups or across different cell states/types using boxplots. Furthermore, users can investigate the correlation between specific gene's expression and CSS values through scatterplots.

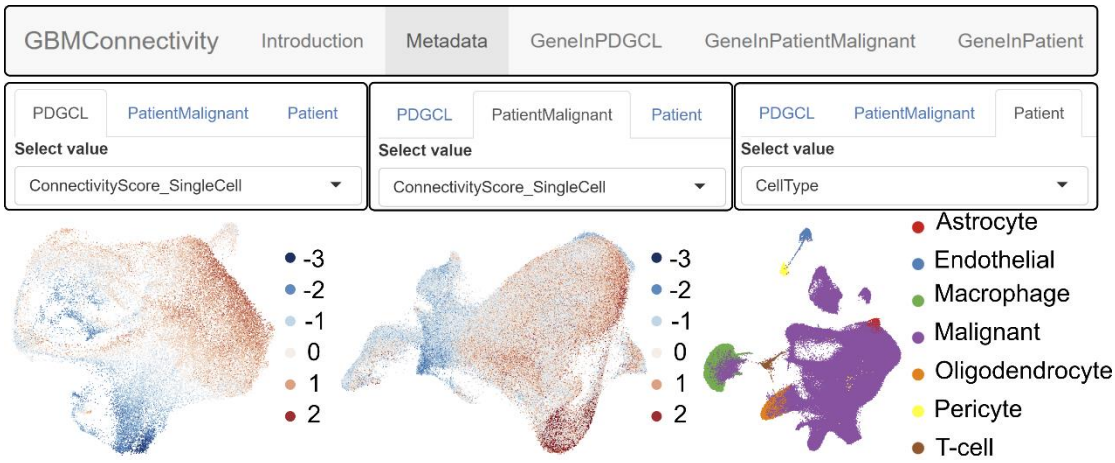


Figure 2.21 **Web tool interface for metadata visualized in UMAP.** Top: Navigation panel. Middle: Under "Metadata" tab, the metadata for the scRNA-Seq datasets are shown in separate sub-tabs: "PDGCL," "PatientMalignant," and "Patient." Bottom: Visualization examples. Bottom left: UMAP displays CSSs in the SR101 PDGCL cells under "PDGCL" subtab. Bottom middle: UMAP visualizes CSSs in the GB malignant cells under "PatientMalignant" subtab. Bottom right: UMAP presents cell type annotation in GB patient tumor cells under "Patient" subtab. The visualization also incorporates other metadata, such as RNA-Seq-derived CSSs, PDGCLs information, SR101 groups, and cell state annotation.

Users can interactively investigate the expression levels of genes within the $SR101^{high}$ and $SR101^{low}$ cells, as well as GB patient tumor cells in the UMAPs (Figure 2.22).

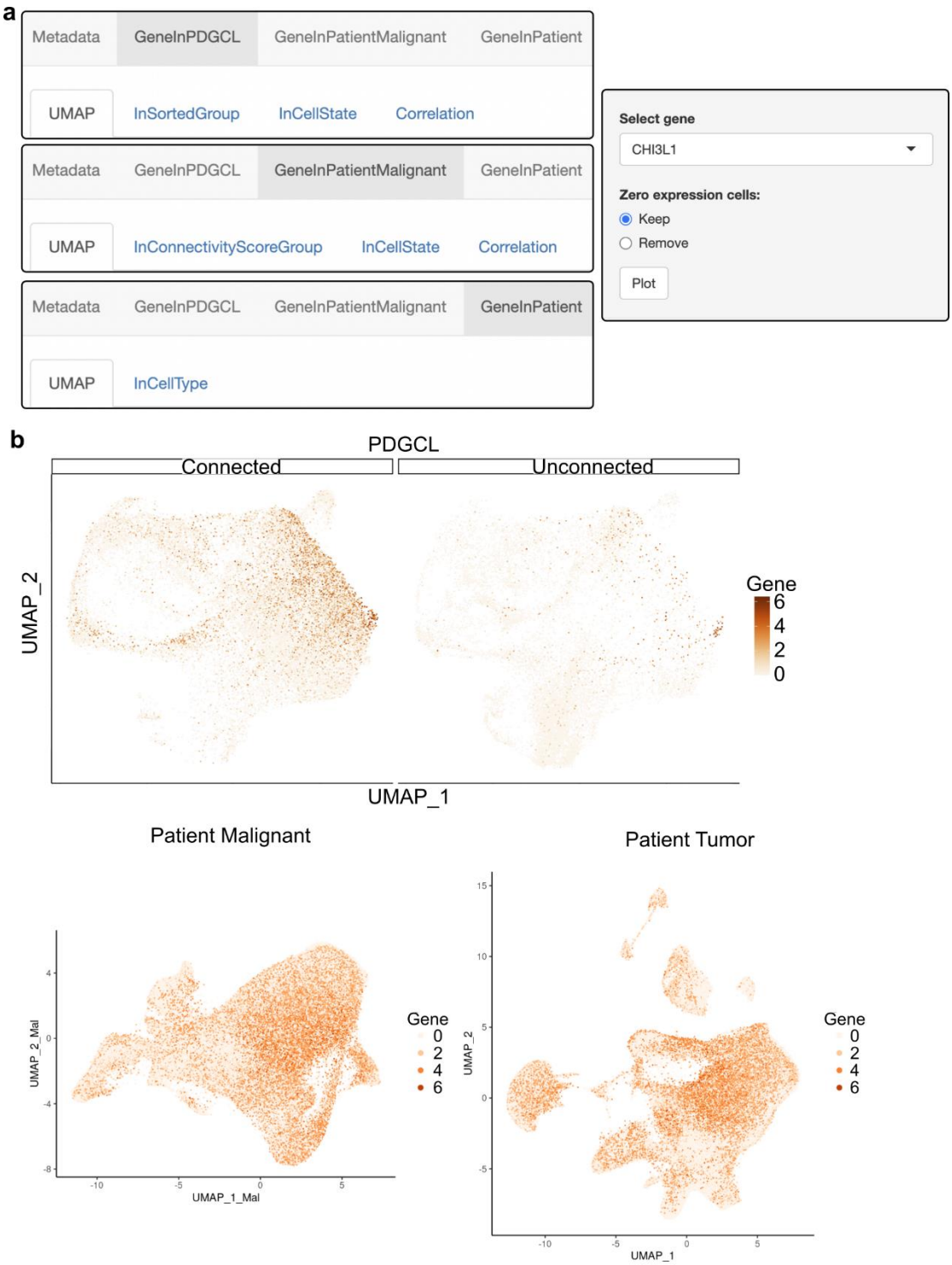


Figure 2.22 **Web tool interface for gene expression visualized in UMAP.** a) Navigation panels. Top, “GeneInPDGCL” tab allows gene expression level of SR101 scRNA-Seq data visualization in sub-tabs: “UMAP”, “InSortedGroup”, “InCellState” and “Correlation”. Middle: "GeneInPatientMalignant" tab visualizes gene expression levels in patient malignant cells in sub-tabs: "UMAP", "InConnectivityGroup," "InCellState," and "Correlation." Bottom: "GeneInPatient" tab presents gene expression levels in patient tumor scRNA-Seq in sub-tabs: "UMAP" and "InCellType". Right: User control panels for gene selection, and filtering cells with/without zero expression. b) Visualization examples. Top: UMAPs showcase expression levels of user-selected genes in $SR101^{high}$ and $SR101^{low}$ cells. Bottom left: UMAP displays gene expression levels in malignant cells. Bottom right: UMAP presents gene expression levels in the

patient tumor cells.

Furthermore, this tool facilitates the comparison of gene expression levels between SR101^{high} and SR101^{low} groups, as well as across various cell states within the SR101 scRNA-Seq dataset (Figure 2.23). Additionally, it enables the comparison of gene expression levels within the four CSS quartile groups, across different malignant cell states, and among various nonmalignant cell types within the GB patient tumor scRNA-Seq dataset (Figure 2.24). Moreover, users can access the correlations between the expression levels of genes and scRNA-Seq-derived or RNA-Seq-derived CSS in the SR101 and patient tumor scRNA-Seq dataset (Figure 2.23, Figure 2.24).

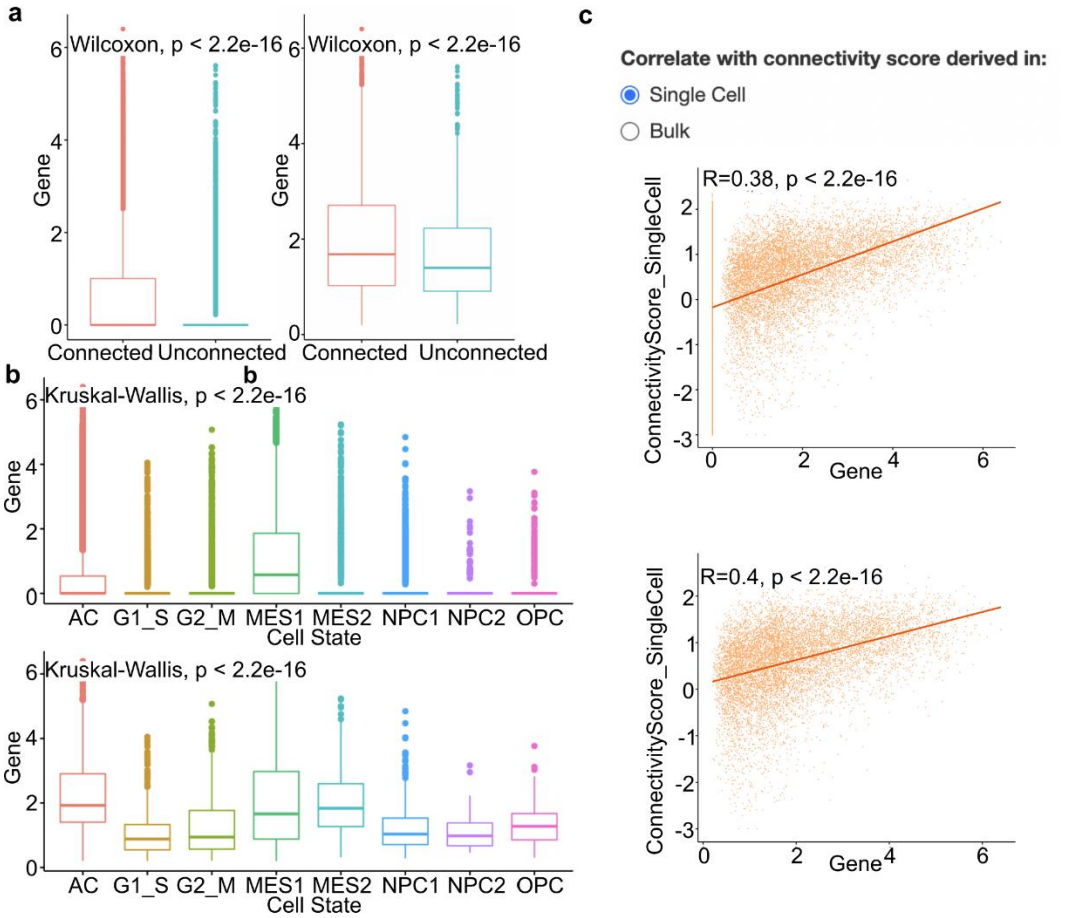


Figure 2.23 **Web tool interface for comparisons of gene expression in the SR101 scRNA-Seq dataset.** a) Box plots depict gene expression in the SR101^{high} and SR101^{low} groups, including cells with zero expression (left) and excluding cells with zero expression (right). P-values determined by Wilcoxon test. b) Box plots illustrate gene expression across cell states, including cells with zero expression (top) and excluding cells with zero expression (bottom). P-values derived from Kruskal-Wallis test. c) Scatterplots demonstrate correlations between gene expression and CSS: Top, scRNA-Seq-derived CSSs, including cells with zero expression. Bottom, scRNA-Seq-derived CSSs, excluding cells with zero expression. Pearson correlation coefficients were computed.

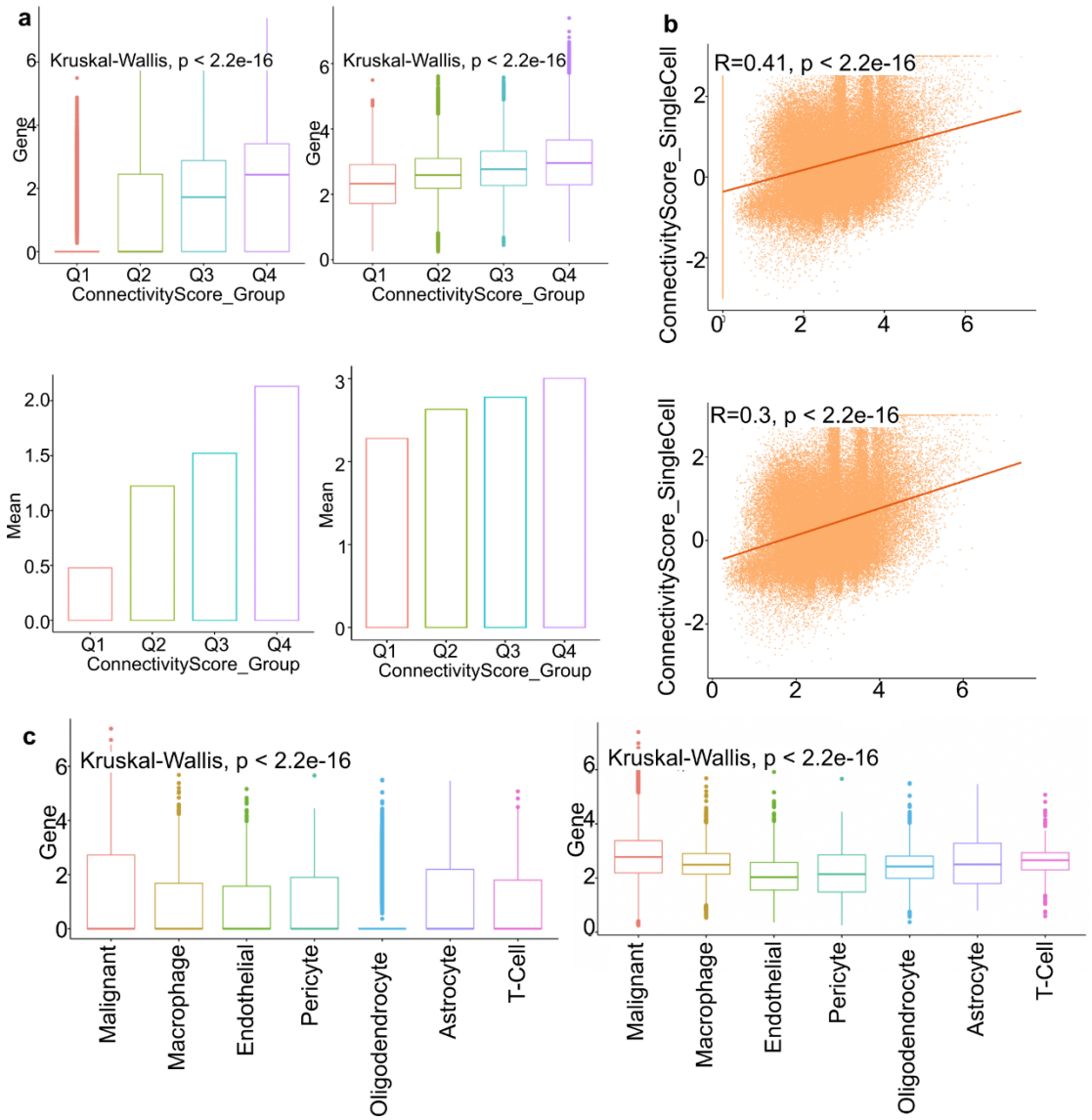


Figure 2.24 **Web tool interface for comparisons of gene expression in the GB patient tumor scRNA-Seq dataset.** a-b) Expression levels of user-selected genes in patient malignant cells. a) Box plots display gene expression in four CSS quartile groups, including cells with zero expression (top left) and excluding cells with zero expression (top right). P-values computed using Kruskal-Wallis test. Bar plots display mean expression in four CSS quartile groups, including cells with zero expression (bottom left) and excluding cells with zero expression (bottom right). b) Scatterplots demonstrate correlations between gene expression and CSSs: Top, scRNA-Seq-derived CSSs, including cells with zero expression. Bottom, scRNA-Seq-derived CSSs, excluding cells with zero expression. Pearson correlation coefficients were computed. c) Expression levels of user-selected genes in patient tumor scRNA-Seq dataset. Box plot represents gene expression across cell types, including zero expression cells (left) and without zero expression (right). P-value computed by Kruskal-Wallis test.

In this subsection, I introduced a web tool for the exploration of the SR101 and GB patient tumor scRNA-Seq datasets. Users can interactively investigate genes within two SR101 groups or across different GB cell states/types.

2.2 Methods

2.2.1 Data collection

2.2.1.1 In-house GB datasets

- **SR101 RNA-Seq dataset**: SR101 experiments were conducted following the protocols described in Osswald et. al., 2015 and Xie et. al., 2019. In brief, tGFP-tagged PDGCLs were injected into mouse brains. Subsequently, SR101 was administered via intraperitoneal injection in mice. The tumors were dissociated into single-cell suspensions, which were then subjected to FACS. The initial sorting step involved segregating doublets and non-viable cells from viable cells based on Calcein Violet and TO-PRO^{TM}-3 staining. Subsequently, human tumor cells were distinguished from non-malignant mouse cells based on tGFP intensity. Finally, tumor cells with high and low connectivity were segregated according to SR101 intensity. The obtained PDGCL cells underwent mRNA extraction using the RNeasy Micro Kit. Subsequently, RNA libraries for RNA-Seq were generated using the SMARTer Ultra Low Input RNA kit, following the manufacturer's guidelines. The resulting libraries were sequenced using an Illumina HiSeq 2000 sequencer, employing the 50 bp single-end mode. Sequencing was performed by the Genomics and Proteomics Core Facility (GPCF) at DKFZ. The sequencing reads were aligned to the human reference genome GRCh38 using STAR (v.2.5.3a, Dobin et al., 2013). Subsequently, a gene count matrix was generated using HTSeq-Counts (Anders, Pyl, and Huber 2015) against the GENCODE v.26 annotation. Genes that had a total count of less than 10 across all samples were excluded from further analysis. Dr. Ruifan Xie (R. X.) from Clinical Cooperation Unit (CCU) Neurooncology at DKFZ conducted this experiment.
- **SR101 scRNA-Seq dataset**: First, the SR101 experiment was conducted as the SR101 RNA-Seq dataset. Then, the FACS-separated highly and lowly connected tumor cells underwent the 10X Genomics scRNA-Seq protocol using the Chromium Next GEM Single Cell 3' GEM, Library & Gel Bead Kit v2, following the manufacturer's instructions. The generated libraries were subsequently sequenced on an Illumina HiSeq 4000 sequencer, utilizing the 150 bp paired-end mode. We achieved an approximate yield of 2 x 350 million reads per sample. The sequencing process was carried out by the GPCF. Dirk C Hoffmann (D. C. H.) from CCU Neurooncology conducted this experiment.
- **21-sample patient tumor scRNA-seq dataset**: A total of 21 frozen tumor

specimens were obtained from GB patients who underwent treatment at Heidelberg University Hospital. Single nuclei were isolated from these specimens and subsequently subjected to the scRNA-Seq. The processing methods was the same as that of the SR101 scRNA-Seq dataset. D. C. H conducted this experiment.

- **CHI3L1 overexpression RNA-Seq dataset**: PDGCL cells overexpressing CHI3L1 were generated by utilizing the control of the PGK1 promoter and transducing them with lentiviral particles. RNA was extracted using the RNeasy Mini Kit. Subsequently, RNA libraries for RNA-seq were prepared using the Illumina TruSeq Stranded RNA Library Prep Kit, following the manufacturer's protocols. The resulting libraries were then subjected to sequencing on an Illumina NovaSeq 6000 sequencer, using the 150 bp paired-end mode. The sequencing process was carried out by the GPCF. The sequencing data underwent processing at the Omics Data Core Facility (ODCF) at DKFZ, following the pipeline outlined in https://github.com/DKFZ-ODCF/RNAseqWorkflow. In brief, alignment of the reads against the human genome (1KGRef_PhiX) was conducted using STAR (v.2.5.3a, Dobin et al., 2013). Duplicate reads were identified and marked using Sambamba (v.0.6.5, Tarasov et al. 2015). For the generation of a gene-count matrix, FeatureCounts (Subread v.1.6.5, Liao et al., 2014) was employed. D. C. H. conducted this experiment.
- **CHI3L1 overexpression proteomics and phosphoproteomics dataset**: PDGCL cells overexpressing CHI3L1 underwent label-free mass spectrometry quantification. Proteins were digested using Trypsin/Lys-C enzymes. Subsequently, a sequential phosphopeptide SMOAC enrichment protocol was employed, which involved metal affinity chromatography using High-Select™ TiO2 in combination with Fe-NTA phosphopeptide enrichment kits. The resulting peptide samples were subjected to analysis using nanoflow LC-MS/MS, utilizing a Dionex 3000 nanoUHPLC connected to an Orbitrap Exploris mass spectrometer. The mass spectrometer operated in data-dependent acquisition mode. The raw mass spectrometry data was processed using MaxQuant (v.2.0.1.0, Cox and Mann 2008) software. Protein and phosphopeptide identifications were performed using the UniProt database UP000000589. FDR < 0.01 were applied at both the protein and peptide levels. For further analysis, phosphopeptides with phosphosite localization probabilities > 0.75 were selected. In the proteomics dataset, a total of 5,022 proteins were retained, with a median of 4,286 proteins obtained per sample. In the phosphoproteomics dataset, 12,799 phosphosites were kept, with a median of 8,520 phosphosites per sample. Gina Cebulla and Dr. Uwe Warnken

from CCU Neurooncology conducted this experiment.

2.2.1.2 Public GB datasets

- **TCGA 230-sample GB datasets** (https://www.cancer.gov/tcga): I downloaded the normalized gene expression matrix (log2FPKM) derived from RNA-Seq, somatic mutation, and CNV information obtained from whole exon sequencing, along with metadata including age, gender, and patient survival information for GDC TCGA Glioblastoma (146 samples) and Lower Grade Glioma (502 samples) cohorts from the UCSC Xena Hub (http://xena.ucsc.edu). I collected somatic mutations results from four variant calling pipelines: MuSE, MuTect2, VarScan2, and SomaticSniper. I excluded all synonymous variants. I re-classified patients as GB IDH wt if they lacked variants in both IDH1 and IDH2 according to these four variant calling pipelines and exhibited intact chromosome 1p/19q based on CNV information. A total of 230 samples fulfilled these criteria and were subsequently selected for further analysis.
- **GBmap harmonized 110-sample GB scRNA-Seq datasets** (Ruiz-Moreno et. al., 2022): I downloaded the gene expression matrix (Counts) derived from scRNA-Seq of 338,564 cells across 110 GB patients, along with metadata including cell annotations through the CELLxGENE data portal (https://cellxgene.cziscience.com/collections/999f2a15-3d7e-440b-96ae-2c806799c08c).
- **GLASS longitudinal 425-sample GB RNA-Seq datasets** (Varn et al., 2022, https://www.synapse.org/glass): I downloaded the normalized gene expression matrix (TPM) derived from RNA-Seq, along with metadata including age, gender, and patient survival information for 425 primary and recurrent GB samples from the Synapse data portal (https://www.synapse.org/glass).
- **GEPIA 31 tumor types and normal tissues RNA-Seq datasets** (Tang Z. et. al., 2017, http://gepia.cancer-pku.cn): I downloaded the median of normalized gene expression level (log2TPM) of CHI3L1 in each tumor type and normal tissue type from the GEPIA data portal (http://gepia.cancer-pku.cn), which encompasses 31 tumor types from TCGA and normal tissue samples from the genotype-tissue expression (GTEx) database.
- **GB 93-pair RNA-Seq and proteomics datasets** (Wang L.B. et. al., 2021): I obtained paired RNA-Seq data (normalized gene expression matrix [log2FPKM]) and proteomics data (normalized protein intensity values with log2 transformation)

from 93 GB patients from the supplementary data of Wang L.B. et al., 2021.

2.2.2 scRNA-Seq data processing

I generated the gene expression count matrices from the SR101 scRNA-Seq data using Cell Ranger (v.2.1.1, 10X Genomics, Zheng et al., 2017) against the pre-built human reference genome (Cell Ranger reference, hg19, v.1.2.0) with the default parameters.

For the patient tumor scRNA-Seq data, I generated the gene expression count matrices using Cell Ranger (v.3.0.1) against a custom pre-mRNA human reference genome created using the mkref function, following the official guidelines (Cell Ranger reference, hg19, v.1.2.0, available at https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/3.1/advanced/references) with the default parameters.

I filtered low-quality cells using the following uniform exclusion criteria:

1) Cells with a number of detected genes fewer than 200 or more than 8,000 were excluded.
2) Cells with a number of counts below 500 or exceeding 80,000 were excluded.
3) Cells contained more than 10% mitochondrial counts were excluded.

After the uniform exclusion process, I identified sample-wise outlier cells and subsequently removed if either the number of genes or counts exceeded three median absolute deviations (MADs) above the median, using the isOutliers function within the scater package (v.1.10.1, McCarthy et al., 2017).

For each sample, I estimated per-cell doublet scores and per-sample doublet score thresholds utilizing the Scrublet tool (v.0.2.1, Wolock et al., 2019) with its default parameters. If a doublet score threshold was situated between two peaks of the doublet score histogram, I adopted that specific threshold. Following this, I removed cells with a doublet score exceeding this threshold.

I further processed the resulting cells per sample using the Seurat package (v.3.1.5, Stuart et al., 2019) in the following steps with default parameters:

1) The gene expression counts were normalized using the NormalizeData function.
2) 2000 highly variable genes were identified using the FindVariableFeatures function.
3) The variation in the number of counts among cells was regressed out, and the resulting residuals were scaled and centered by the ScaleData function.
4) Dimensionality reduction of the data was performed by principal component analysis (PCA) using the RunPCA function.
5) The number of principal components (PCs) used for further analyses was determined using the ElbowPlot function.
6) Unsupervised clusters were identified using shared nearest neighbor (SNN) analysis in the PCA space through the FindNeighbors and FindClusters functions.
7) The data was visualized in UMAP using the RunUMAP function.

The visualization of the SNN clusters in UMAP allowed me to observe the presence of clusters that were notably distant from the majority of clusters. These isolated clusters were subsequently excluded from the analysis.

To tackle inter-tumor heterogeneity, I performed integration across different PDGCLs/patients using the "anchor" integration method through the Seurat package. The process involved the following steps:

1) Normalization and feature selection were performed for each PDGCL/patient as described above in steps 1-2.
2) Pairwise "anchors" were identified between PDGCL/patients using the FindIntegrationAnchors function.
3) Data integration was performed based on the identified "anchors" utilizing the IntegrateData function.
4) The integrated data underwent analogous processes as described in above steps 3-7, encompassing variance regression, scaling, centering, PCA dimensionality reduction, determination of the number of PCs, SNN clustering and UMAP visualization.

After the “anchor” integration, I identified an SNN cluster that expressed markers of two distinct cell types in the patient tumor scRNA-Seq dataset. Consequently, I excluded this cluster.

2.2.3 Computational development of connectivity signatures

Figure 2.1b depictes the schematic representation of the connectivity signature development.

For the SR101 scRNA-Seq dataset, I identified DEGs between $SR101^{high}$ and $SR101^{low}$ groups using the Seurat package with default parameters, following these steps:

1) DEGs were identified within each PDGCL individually through the FindMarkers function.
2) DEGs with adjusted p-value < 0.05 (after multiple testing correction) from all three PDGCLs were combined.
3) DEGs exhibiting consistent direction of regulation and an absolute log fold-change ≥ 0.4 in at least two PDGCLs were retained.
4) DEGs showing the same direction of regulation across all three PDGCLs were retained.

For the SR101 RNA-Seq dataset, I identified DEGs between $SR101^{high}$ and $SR101^{low}$ groups using the DESeq2 package (v.1.22.2, Love et al., 2014) with default parameters, in accordance with the following steps:

1) To ensure consistent DEGs across both PDGCLs, the design formula of the DESeqDataSet function included ~ PDGCL + Group.
2) Genes with a total count lower than 10 across all samples were filtered out.
3) Differential expression analysis was carried out using the DESeq function.
4) The log fold changes were shrunken using the apeglm method within the lfcShrink function.
5) DEGs with an adjusted p-value of < 0.05 and an absolute log2 fold-change ≥ 1 were retained.

2.2.4 Heatmap visualization of connectivity signature

For the SR101 scRNA-Seq dataset, I aggregated the normalized expression levels of the scRNA-Seq-derived connectivity signature in cells from each sample into a “pseudo bulk” using the AverageExpression function in Seurat. Subsequently, I scaled

and visualized this aggregated data as a heatmap using the ComplexHeatmap package (v.2.5.4, Gu et al., 2016).

For the SR101 RNA-Seq dataset, I transformed the expression levels of the RNA-Seq-derived connectivity signature using the vst function. After this transformation, I used the removeBatchEffect function from the limma package (v.3.36.5, Ritchie et al., 2015) to mitigate the variability associated with different PDGCLs. Subsequently, I scaled and visualized the adjusted expression levels in a heatmap using the ComplexHeatmap package.

2.2.5 Enrichment analysis

I determined the enriched GO terms in the RNA-Seq-derived and scRNA-Seq-derived connectivity signatures, and the enriched GO terms across DEGs, DEPs, and DPPs resulting from CHI3L1 overexpression, using Metascape (v3.5.20230501, Zhou et al., 2019). This analysis utilized the built-in database, including GO terms, KEGG pathway terms, canonical pathways, hallmark gene sets, TRRUST transcription factors, and more.

I identified the enriched kinases within the DEGs, DEPs, and DPPs of CHI3L1 overexpression datasets, using KEA3 (Kuleshov et al., 2021). This process involved utilizing a built-in database that includes STRING, ChengPPI, PTMsigDB, BioGRID, prePPI, and so on.

2.2.6 Connectivity signature score (CSS)

I calculated CSS for single cell/sample using the AddModuleScore function within Seurat. This function facilitates the computation of a score for each single cell or sample based on a specified gene set. This score is generated by considering the average expression levels of the given gene set and subsequently adjusting it by subtracting the aggregated expression of control gene sets. The control genes are chosen at random from predefined bins, organized according to the average expression levels of the genes.

CSS_{up} represents a score computed from the upregulated genes in the SR101^{high}

samples, while CSS_{down} represents a score calculated from the downregulated genes in the SR101^{high} samples using the AddModuleScore function. The CSS was calculated using the following formular:

$$\mathrm{CSS} = CSS_{up} - CSS_{down}$$

2.2.7 Performance of the CSS-based prediction

In the SR101 scRNA-Seq dataset, I predicted the SR101 group for each individual cell based on the CSS. If the CSS value was less than 0, I predicted the cell as SR101^{high}; otherwise, I predicted it as SR101^{low}. To assess the accuracy of these predictions, I generated a confusion matrix and various prediction metrics by comparing the number of cells predicted by CSS values with the number of cells sorted by FACS according to the SR101 staining intensity.

I employed the R package caret (v.6.0-80, Kuhn et al., 2018) to calculate prediction metrics such as accuracy, sensitivity, specificity, positive predictive value, and negative predictive value. I conducted this assessment for both the scRNA-Seq-derived and RNA-Seq-derived CSSs.

Additionally, I generated negative controls using scores generated from randomly selected gene sets. A total of 100 random gene sets were generated, with each gene set containing 71 randomly chosen genes (40 upregulated genes and 31 downregulated genes, the same size as the scRNA-Seq-derived connectivity signature). Similarly, another 100 random gene sets were created, each consisting of 245 randomly selected genes (57 upregulated genes and 188 downregulated genes, the same size as the RNA-Seq-derived connectivity signature). I utilized these gene sets to calculate scores and determine the average prediction metrics as part of the assessment process.

2.2.8 GB malignant cell state assignment

I extracted cell state-defining markers from a previously published GB scRNA-Seq dataset (Neftel et al., 2019). These markers were utilized to compute cell state signature scores for individual cells or samples, using the AddModuleScore function within Seurat. Each cell or sample was then assigned to a specific cell state based on

the cell state signature score that exhibited the highest value among all calculated cell state signature scores.

2.2.9 Ligand-receptor interaction in the SR101 scRNA-Seq dataset

I performed the inference and visualization of ligand-receptor interaction in the SR101 scRNA-Seq dataset using the CellChat package (v.1.6.1, Jin et al. 2021) with the built-in ligand-receptor interaction database.

The steps involved in this process are as follows:

1) Input Data: The normalized expression matrix and cell state annotations for both SR101^{high} and SR101^{low} groups.
2) Ligand-Receptor Identification: Over-expressed ligands or receptors were identified using the `identifyOverExpressedGenes` and `identifyOverExpressedInteractions` functions.
3) Expression Smoothing: Gene expression values were smoothed using a protein-protein interaction network through the `projectData` function.
4) Interaction Inference: The `computeCommunProb` function was used to infer ligand-receptor interaction probabilities. The `filterCommunication` function was then used to retain interactions involving more than 10 cells.
5) Signaling Pathway Inference: Interaction at the signaling pathway level was inferred using the `computeCommunProbPathway` function.
6) Aggregated Network: The aggregated ligand-receptor interaction network was created by summarizing interaction probabilities using the `aggregateNet` function.
7) Merging Objects: The SR101^{high} and SR101^{low} objects were merged using the `mergeCellChat` function. The maximum number of interactions was utilized to control node size and edge weights in figures.
8) Comparison Analysis: Various aspects of the comparison analysis between SR101^{high} and SR101^{low} groups were visualized, including comparisons of:
    - Total number of interactions and interaction strength.
    - Number of interactions among cell states.
    - Differential outgoing signaling in cell states visualized in a heatmap.
    - Number of interactions in the NOTCH pathway among cell states.
    - Expression levels of ligands and receptors in the NOTCH pathway

within each cell state.

2.2.10 RNA velocity in the SR101 scRNA-Seq dataset

I extracted the pre-mature (unspliced) and mature (spliced) mRNA count matrices within the SR101 scRNA-Seq dataset using velocyto (v.0.17.15, La Manno et al. 2018) with default settings. Then I processed these count matrices using scVelo with the default settings (v.0.2.4, Bergen et al. 2020) which included steps such as data filtering, normalization, dimensionality reduction (PCA), nearest neighbor identification, RNA velocity estimation, and embedding of RNA velocities as streamlines in the PCA space.

2.2.11 GB cell type annotation

I annotated cell clusters in the 21-sample GB patient tumor scRNA-Seq dataset with seven cell types (oligodendrocyte, macrophage, endothelial cell, pericyte, T cell, astrocyte, and malignant cell). This annotation involved the following steps:

1) SNN Graph and Clustering: I constructed a SNN graph in the "anchor" integrated 21-sample GB patient tumor dataset using the FindNeighbors function in Seurat. This graph was then used to identify 24 unsupervised clusters by employing the FindClusters function in Seurat.
2) Cell Type Marker Collection: I collected cell type markers from the following sources: I identified the top 100 markers for malignant cells, macrophages, T-cells, and oligodendrocytes from a published GB Smart-Seq2 dataset (Neftel et al., 2019) using the FindAllMarkers function in Seurat. Additionally, I collected 183 astrocyte markers from a healthy brain RNA-seq dataset (from the Supplementary Table of Zhang Y. et al., 2016), and I collected the top 100 markers for endothelial cells and pericytes from a brain mural cell RNA-seq dataset (from the Supplementary Table of He et al., 2016).
3) Cell Type Score Calculation: I calculated cell type scores for each cell based on the respective cell type markers using the AddModuleScore function in Seurat.
4) Cell Type Assignment to Clusters: I assigned cell types to cell clusters based on the MADs of cell type scores. Seven clusters were assigned to the five non-malignant cell types, namely oligodendrocytes, macrophages, endothelial cells, pericytes, and T cells (cluster 5 as oligodendrocytes, cluster 8, 9, and 23 as

macrophages, cluster 17 as endothelial cells, cluster 19 as T-cells, and cluster 22 as pericytes). These clusters exhibited high scores for their corresponding non-malignant cell types and low scores for malignant cells.

5) CNV Analysis and Validation: I estimated CNV in cells as a validation step for cell type assignment. Five non-malignant cell types and 17 unassigned clusters were downsampled to 500 cells per cluster and subjected to infercnv (v.1.2.1, Tickle et al., 2019) with the settings (cutoff = 0.1, cluster_by_groups = TRUE, denoise = TRUE, HMM = TRUE). Among them, 16 clusters exhibited significant CNVs in chr7 and chr10 along with higher malignant scores, leading to their annotation as malignant cells. Additionally, cluster 21, which showed no CNVs and the highest median of astrocyte scores, was annotated as astrocytes.
6) Cell Type Markers Identification: I identified the top 50 markers for each cell type in the GB patient tumor scRNA-Seq dataset using the FindAllMarkers function in Seurat.

2.2.12 Two-dimensional visualization of cells according to their cell state

Similar to Neftel et al., 2019, I visualized the single cells in a two-dimensional (2D) plot based on six cell state signature scores ($S_{AC}, S_{MES1}, S_{MES2}, S_{OPC}, S_{NPC1}, S_{NPC2}$). I employed the subsequent formula to calculate the X and Y coordinates for a given cell in the 2D plot:

$$Y = max(S_{AC}, S_{MES1}, S_{MES2}) - max(S_{OPC}, S_{NPC1}, S_{NPC2})$$
$$if\ Y > 0, then:\ X = S_{AC} - max(S_{MES1}, S_{MES2})$$
$$if\ Y \leq 0, then:\ X = S_{OPC} - max(S_{NPC1}, S_{NPC2})$$

As a result, cells characterized by MES1 and MES2 cell states occupy the top-left corner of the 2D plot; the AC cell state occupies the top-right corner; NPC1 and NPC2 states are located in the bottom-left corner, while the OPC state resides in the bottom-right corner.

2.2.13 GB expression subtype assignment

I assigned three expression subtypes (MS, CL, and PN) to the TCGA GB RNA-Seq datasets using single-sample Gene Set Enrichment Analysis (ssGSEA) as described

by Wang Q. et al. in 2017. This classification involved the following steps:

1) Input Data: The log2-transformed FPKM matrices derived from RNA-seq data.
2) Permutation: To generate null distributions for each gene set enrichment score, 100,000 permutations were performed.
3) Subtype Assignment: The subtype assignment was determined by selecting the subtype associated with the lowest empirical p-value.

2.2.14 Gene mutation and CSS in the TCGA GB cohort

I retrieved the somatic mutation data in the TCGA GB cohort from UCSC Xena database (as detailed in subsection 2.2.1.2), which included information from four variant calling pipelines (MuSE, MuTect2, VarScan2, and SomaticSniper). Excluding all synonymous variants, a total of 27 genes that exhibited variants in at least 5% of patients were chosen for the analysis.

For each of these 27 genes, I conducted Mann-Whitney U test to compare the CSS between wt patients and patients with mutation, using wilcox.test function within the stats package. To address the issue of multiple testing, I computed False Discovery Rate (FDR) using the p.adjust function.

2.2.15 Cell type/state deconvolution in the TCGA GB RNA-Seq dataset

I performed cell type/state deconvolution utilizing the CIBERSORTx tool (Newman et al., 2019, https://cibersortx.stanford.edu/) against a custom signature matrix. Specifically, I used the gene count matrix from randomly selected 50 cells of each cell type/state (including endothelial cells, pericytes, T cells, oligodendrocytes, astrocytes, macrophages, and malignant AC, MES1, MES2, OPC, NPC1, NPC2, G1_S, and G2_M cell states) within the 21-sample GB patient tumor scRNA-Seq dataset for CIBERSORTx to generate the signature matrix. Subsequently, I imputed cell fractions in the bulk RNA-Seq data based on this signature matrix using CIBERSORTx.

I employed an alternative approach for cell type/state deconvolution in the bulk RNA-Seq using a GB-specific deconvolution tool called GBMDeconvoluteR (v.1.5.0, Ajaib et al., 2023, https://gbmdeconvoluter.leeds.ac.uk/). This tool integrated GB cell type/state markers from multiple GB scRNA-Seq datasets to estimate abundance

scores for various cell types and states within each sample. The gene expression matrix (FPKM) served as input for GBMDeconvoluteR, resulting in the computation of abundance scores for cell populations, such as monocytes, macrophages, mast cells, dendritic cells, natural killer cells, T cells, B cells, microglia, and malignant AC, MES, NPC, and OPC cell states.

2.2.16 GB patient survival analysis

I classified the GB patients into three distinct groups based on quartiles of their CSS values or CHI3L1 expression levels. The categorization was as follows:

- Q1: samples with the lowest 25% values
- Q2-Q3: samples with values ranging from 25% to 75%
- Q4: samples with the highest 75% values

I conducted Kaplan-Meier survival analysis to assess the overall survival or surgical interval of patients within these three CSS or CHI3L1 groups. This analysis involved utilizing the survfit function from the survival package (v.3.1-12, Therneau, 2020) and visualizing the results using the ggsurvplot function from the survminer package (v.0.4.2, Kassambara and Kosinski, 2018).

To explore the potential association between the continuous CSS values or CHI3L1 expression levels and the overall survival or surgical interval of GB patients, I performed Cox proportional hazards regression (Coxph) survival analysis. This analysis employed the coxph function within the survival package. In the Coxph model, multiple covariates were adjusted, including age, gender, and expression subtype.

2.2.17 Differential gene expression analysis in the CHI3L1 OE RNA-Seq dataset

I determined the DEGs between CHI3L1 OE and control samples using the DESeq2 package. The procedure for identifying these DEGs was consistent with the methods outlined in the development of the connectivity signature from SR101 RNA-Seq dataset, as detailed in subsection 2.2.3.

2.2.18 Differential protein expression and phosphorylation analysis in the CHI3L1 OE proteomics dataset

I normalized the label-free quantification proteomics and phosphorproteomics data using the 'vsn' method, using the DEP package (v.1.14.0, Zhang X. et al. 2018).

The distribution of intensities and cumulative fraction of proteins in both proteomics and phosphoproteomics datasets suggested that proteins with missing values had lower intensities and might be under the detection threshold. To address these missing values, I employed the deterministic minimum (MinDet) method in the DEP package for missing value imputation. This approach involved substituting each missing value with the smallest detectable intensity (i.e., the 0.01 quantile) observed within each sample.

I identified DPPs between CHI3L1 OE and control samples using the test_diff function in the DEP package, applying a threshold of adjusted p value < 0.05 and an absolute log2 fold change > 1.5.

Given the substantial variations among PDGCLs within the proteomics dataset, I further mitigated these variations using the removeBatchEffect function from the limma package to the normalized and imputed matrix. Subsequently, I fitted a linear model to the corrected data through the lmFit function, and computed empirical Bayes statistics using the eBayes function. DEPs with adjusted p value < 0.05 and an absolute log2 fold change > 0.5 were kept.

2.2.19 Implementation of a web tool for data visualization

I developed a web tool using the Shiny framework (Chang et al., 2020) accessible at https://connectivity-glioma.dkfz.de/. The web tool incorporates metadata and normalized gene expression matrices from SR101 and patient tumor scRNA-Seq datasets. UMAPs were created using ggplot2 package. Boxplots with statistical tests and scatterplots with correlation coefficients were generated using the ggpubr package.

2.3 Discussion and Conclusions

Recently, researchers have discovered that half of glioma cells are interconnected

through ultra-long membrane protrusions known as tumor microtubes (TMs) (Osswald et al., 2015; Weil et al., 2017; Venkataramani et al., 2019; Xie et al., 2021). These cells form a cell-to-cell network that exhibits self-repair capabilities, resistance to surgical lesions, chemotherapy, and radiotherapy. Additionally, this network facilitates tumor cell invasion, proliferation, and communication (Osswald et al., 2015; Weil et al., 2017). This phenomenon holds significant clinical implications, as targeting the TM-connected tumor cell network could have therapeutic potential (Osswald et al., 2015; Winkler and Wick, 2018). However, despite its importance, there is a lack of comprehensive molecular understanding of this network and a quantification method for assessing cell connectivity within it.

2.3.1 The bulk and single-cell RNA-Seq-derived connectivity signatures

I conducted a comprehensive characterization of TM-connected glioblastoma cells, employing both bulk and single-cell transcriptome data analysis. From the SR101 scRNA-Seq data, I established a 71-gene connectivity signature. Additionally, utilizing SR101 bulk RNA-Seq data, I derived another 245-gene signature. These signatures capture the transcriptomic differences between TM-connected and TM-unconnected tumor cells.

Within these two signatures, 13 genes were found to overlap, constituting 5% of the RNA-Seq-derived signature and 18% of the scRNA-Seq-derived signature. This limited overlap likely arises from differences between bulk and single-cell sequencing methodologies. In scRNA-Seq, gene expression is quantified in individual cells, while bulk RNA-Seq contains a mixed signal from different cell types and averages expression across them.

Comparing fold changes in gene expression between highly connected and lowly connected samples from both scRNA-Seq and RNA-Seq data revealed low correlation across all genes. Since scRNA-Seq often experiences a high dropout rate and introduces greater variability, I excluded genes with a 95% dropout rate in the scRNA-Seq dataset and found that the correlation increased to a moderate level. Additionally, removing insignificantly regulated genes further enhanced the correlation. Notably, the 13 overlapping connectivity genes exhibited a strong correlation of fold changes. Moreover, the enriched pathways within the two connectivity signatures displayed remarkable similarity.

I introduced a score termed "CSS" (Connectivity Signature Score), which representing the relative expression of genes within the connectivity signature. The CSS based on the scRNA-Seq-derived signature, displayed a high correlation with RNA-Seq-derived CSS not only within the scRNA-Seq dataset but also in RNA-Seq dataset. These outcomes collectively indicate a robust consistency in connectivity signatures, even when derived from two distinct transcriptome sequencing techniques.

2.3.2 Connectivity and cell states

One notable advantage of scRNA-Seq is its ability to explore intra-tumor heterogeneity and characterize distinct cell states within tumors. A recent study revealed the presence of several diverse cell states within glioblastoma (Neftel et al., 2019). Therefore, I investigated the relationship between connectivity and these cell states.

I observed a noteworthy difference in cell state composition between highly connected ($SR101^{high}$) and lowly connected ($SR101^{low}$) samples. Notably, among the highly connected samples, there was a prevalence of AC and MES cell states. In contrast, the NPC cell state took precedence in the lowly connected samples. Interestingly, nearly 50% of the connectivity signature genes overlapped with the markers of specific cell states identified in Neftel et al., 2019. This overlap further emphasizes the interrelationship between connectivity and cell states.

In adidition, I applied RNA velocity to the SR101 scRNA-Seq dataset and found the highly connected and lowly connected cells underwent different possibilities of cell state transitions. Notably, the AC cell state represented the endpoint of cell state transitions among highly connected cells, while no specific cell state became the endpoint for lowly connected cells.

I also found evidence from the literature showing a relationship between cell connectivity and astrocytes and mesenchymal cells:

- Astrocytes, star-like glial cells in the brain, form intricate interconnected networks (Deemyad et al., 2018; Fields and Stevens-Graham, 2002; Mederos et al., 2018; Sul et al., 2004). Notably, astrocytes are abundant in gap junctions and express connexin Cx43 (Mederos et al., 2018). Within glioblastoma, TM

networks are linked through Cx43 gap junctions, contributing to their connectivity (Osswald et al., 2015).

- Mesenchymal cells, star-like cells, originate from the mesoderm within the head (Fish and Schneider, 2014). Mesenchymal stem cells (MSCs) can undergo transdifferentiation, acquiring neuron-like characteristics (Chu et al., 2006; Dilger et al., 2020). MSCs exhibit robust expression of Cx43, while this expression diminishes as the cells differentiate (though Cx43 still maintains a stronger presence compared to other connexins) (Dilger et al., 2020). Remarkably, bone marrow MSC transplantation enhances GAP43 expression and mitigates neurological deficits in cases of intracerebral hemorrhage (Cui et al., 2017). The differentiation process of MSCs influences microtubules and intermediate filaments (Saidova and Vorobjev, 2019), and both Cx43 and GAP43 play pivotal roles in the context of TM networks (Osswald et al., 2015).

Together, these results suggest an association between connectivity and AC and MES cell states.

2.3.3 CSS and cell state

While calculating CSS in the SR101 PDGCL-xenograft scRNA-Seq dataset, I observed higher CSS values in AC and MES cell states compared to NPC and OPC cell states. These consistent findings were also found in both the 21-sample GB patient tumor scRNA-Seq dataset and the GBmap 110-sample GB patient tumor scRNA-Seq dataset. This also demonstrates the effectiveness of the CSS in both xenograft models and GB patient tumors datasets.

Considering the abundance of bulk RNA-Seq datasets, which are often more readily accessible, I investigated the relationships between CSS and cell states in the TCGA 230-sample RNA-Seq dataset. This analysis involved the use of two bulk deconvolution methods—CIBERSORTx and GBDeconvoluteR. Encouragingly, I observed a positive correlation between CSS and the proportion of MES in samples, as well as a negative correlation between CSS and the proportion of NPC.

The consistent application of the CSS across multiple GB patient tumor scRNA-Seq and RNA-Seq datasets underscores the robustness and reliability of the connectivity signature.

2.3.4 CSS and expression subtype

Researchers have dedicated considerable efforts to classifying GB into several expression subtypes (Phillips et al., 2006; Verhaak et al., 2010; Wang Q. et al., 2017). Within the TCGA GB RNA-Seq dataset, three distinct expression subtypes (mesenchymal, classical, and proneural) were identified (Wang Q. et al., 2017). Therefore, I also investigated the relationship between CSS and these expression subtypes.

I found that mesenchymal samples exhibited the highest CSS values, classical samples showed medium CSS values, and proneural samples displayed the lowest CSS values.

The mesenchymal subtype, which is characterized by NF1 mutation and increased CHI3L1 expression (Behnan et al., 2019). In accordance with this, I have detected NF1-mutated GB tumors correlated with lower CSS values in the TCGA RNA-Seq dataset. In previous sections, CHI3L1 was identified as a key player in the connectivity signature.

The proneural subtype is characterized by TP53 mutation, increased OLIG2 expression, and NOTCH activation (Behnan et al., 2019). Notably, I observed that TP53-mutated GB tumors correlated with lower CSS values in the TCGA RNA-Seq dataset. OLIG2 was identified as a downregulated gene in the connectivity signature. Interestingly, when applying ligand-receptor cross-talk assessment in the SR101 scRNA-Seq dataset, I found a notable activation of the NOTCH signaling pathway in the lowly connected cells. A previous study also found that the NOTCH1 pathway is inhibited in tumor cells with TM connections (Jung et al., 2021).

In summary, these findings indicate a positive correlation between CSS and the mesenchymal subtype, along with a negative correlation between CSS and the proneural subtype.

2.3.5 CSS and patient survival

The impact of TM-connectivity on GB patient survival was not yet completely

addressed. In the analysis of TCGA GB cohort, I showed the correlation between high CSS levels and unfavorable patient survival. This relationship held true even after accounting for several confounding factors, including age, gender, and expression subtypes. Encouragingly, I obtained consistent results from the GLASS longitudinal cohort, which revealed that elevated CSS levels were linked to both reduced overall survival and shorter intervals until subsequent relapses in both primary and recurrent tumor samples.

These findings indicate that CSS could serve as a valuable prognostic factor for GB.

2.3.6 CHI3L1 as a novel marker for connectivity

The scRNA-Seq-derived connectivity signature contains several previously recognized TM-connectivity markers, including GAP43 (Osswald et al., 2015; Weil et al., 2017; Venkataramani et al., 2022), DLL1 (Jung et al., 2021), DLL3 (Jung et al., 2021), and APOE (Venkataramani et al., 2019). Additionally, the connectivity signature encompasses other genes with unexplored associations with TM-connectivity, thereby unveiling new opportunities for the identification of novel markers and potential therapeutic targets for connectivity.

In this study, CHI3L1 was the most significantly upregulated gene in both scRNA-Seq-derived and RNA-Seq-derived connectivity signatures. Notably, CHI3L1 displayed upregulation consistently in highly connected cells across various cell states. Furthermore, the expression level of CHI3L1 demonstrated the strongest correlation with the CSS compared to other genes within the connectivity signature.

To gain deeper insights into the functional role of CHI3L1 in connectivity, I analyzed the RNA-Seq, proteomics, and phosphoproteomics datasets of CHI3L1. I found the expression of CHI3L1 at both the RNA and protein levels exhibited a strong positive correlation. The overexpression of CHI3L1 led to the upregulation of multiple genes within the connectivity signature, as well as the upregulation of AC markers and the downregulation of NPC1 markers. These findings were observed at both the RNA and protein levels. Moreover, the overexpression of CHI3L1 prompted an increase in the CSS and the AC signature score, while concurrently causing a decrease in the NPC1 score. Remarkably, the overexpression of CHI3L1 was associated with higher phosphorylation level of GAP43, a marker of TM-connectivity.

Furthermore, the pathway enrichment analyses conducted on RNA-Seq, proteomics, and phosphoproteomics datasets of CHI3L1 overexpression revealed associations with neuron projection and cell junction, as well as MAPK and NFKB pathways. Interestingly, these findings are in concordance with the characteristics exhibited by highly connected GB pacemaker cells as elucidated by the work of Hausmann et al. in 2023.

In previous studies, although not specifically described in the context of GB, CHI3L1 has been demonstrated to play a role in connectivity in other cancer types. For instance, the knockdown of CHI3L1 suppresses cell migration and tube formation in endothelial cells (Kawada et al., 2012). Overexpression of CHI3L1 contributes to the invasion, migration, and growth of liver tumor cells, impacting cell-cell adhesion and adherent junction pathways (Qiu et al., 2018). In a breast tumorigenic epithelial cell line, CHI3L1 regulates migration, invasion, angiogenesis, and capillary-like network formation (Morera et al., 2019).

In summary, these findings indicate that CHI3L1 may serve as a novel marker for connectivity.

2.3.7 CHI3L1 as a prognostic marker for GB

In our patient tumor scRNA-Seq dataset, CHI3L1 was demonstrated high expression in GB malignant cells compared to nonmalignant cells. Previous study has shown higher CHI3L1 expression levels in gliomas compared to normal brain tissue (Ku et al. 2011).

Moreover, I found that high levels of CHI3L1 expression are linked to unfavorable survival outcomes in both the TCGA and GLASS GB datasets, even after adjusting for several covariates, including age, gender, and expression subtypes. This observation is consistent with previous findings that CHI3L1 expression is associated with patient survival in both low-grade glioma and GB (Steponaitis et al., 2016). Furthermore, CHI3L1 is known to regulate glioma tumor cell invasion, growth, and responses to anti-cancer drugs (Ku et al., 2011). CHI3L1 also plays a role in promoting angiogenesis and tumor cell proliferation in GB (Zhao et al., 2020).

These findings indicate that CHI3L1 could serve as a prognostic marker for GB.

In conclusion, this study characterizes highly connected GB cells using both scRNA-Seq and RNA-Seq. The established connectivity signature and CSS robustly illuminate the associations between connectivity and cell states, expression subtypes, gene mutations, and patient survival across various scRNA-Seq and RNA-Seq GB datasets. Furthermore, this study highlights the potential of the connectivity signature gene CHI3L1 as a novel marker for connectivity and a valuable prognostic marker for GB.

2.4 Outlook

In this study, I explored the transcriptome landscape of highly connected glioblastoma cell populations through scRNA-Seq and RNA-Seq. There are several facets that worth further exploration in future studies:

- The uncovered association between highly connected glioblastoma cells and tumor cell states prompts further investigation into the interactions of these highly connected tumor cells with the nonmalignant cells in the tumor microenvironment.
- The developed connectivity signature was evaluated across various GB scRNA-Seq and RNA-Seq datasets. The emerging *in situ* spatial single-cell genomics techniques that grant direct access to the transcriptome of TM-connected cells in their native environment could offer valuable insights.
- The demonstrated correlation between the connectivity signature and patient survival points to the potential use of the connectivity signature score as a cell connectivity indicator in clinical studies focused on disconnection treatments in GB patients.
- The identification of the connectivity signature gene CHI3L1 as a robust marker for connectivity and a valuable prognostic marker for glioblastoma suggests the need for further investigation into the functional mechanism by which CHI3L1 drives tumor microtube formation and shapes the cell connectivity network. Evaluating the potential of CHI3L1 in GB clinical studies also warrants further attention.

3. Interactive explorer of single cell cluster similarity

In heterogeneous scRNA-Seq datasets, comparing cell clusters across donors, experimental conditions, or sequencing technologies has remained a significant challenge. I introduce the interactive explorer of single cell cluster similarity (ieCS), an R package featuring a user-friendly graphical interface. ieCS is designed to address this challenge by applying a novel similarity metric and three methods to identify superclusters. These superclusters encompass cell clusters with similar behaviors, such as cell types, across heterogeneous datasets.

This section elaborates on the motivation behind the development of ieCS, describes the design and implementation of the tool, presents the outcomes of applying the tool to a demonstration dataset, and discusses the evaluations of the obtained results.

Contributions
Ling Hai developed, demonstrated and evaluated ieCS. The text was written by Ling Hai. It has been proofread and edited by ChatGPT.

3.1 Motivation

To illustrate the motivation behind the development of ieCS, I conducted scRNA-Seq standard analysis using a dataset encompassing 13,999 peripheral blood mononuclear cells (PBMCs) subjected to two distinct experimental conditions: interferon-beta (IFNB) stimulated and control conditions (Kang et al., 2018).

The standard analysis of scRNA-Seq data, performed using Seurat (Stuart et al., 2019, https://satijalab.org/seurat/articles/pbmc3k_tutorial), involved several steps: raw count data normalization, selecting highly variable genes as features, scaling the data, dimension reduction, unsupervised cluster identification, and UMAP-based data visualization. However, the resultant UMAPs did not exhibit coherent clustering of PBMCs from different experimental conditions (Figure 3.1a). Despite some cell clusters sharing the same cell type annotation (cell type annotation from Kang et al., 2018), they failed to cluster together (Figure 3.1b). These results are similar with the observations in the original paper (Figure 3a of Kang et al., 2018).

To address these condition-induced discrepancies, I separated and independently

analyzed the data for each condition (Figure 3.1c-f). This led to UMAPs with cells organized by cell types (cell type annotations got from Kang et al., 2018, Figure 3.1c, e). Subsequent unsupervised SNN clustering identified distinct cell clusters (Figure 3.1d, f). In total, there are 26 cell clusters across both conditions. I renamed the cell clusters based on the majority of cell type annotations for cells within each cluster (Table 7).

To link cell clusters of the same cell type across different conditions, I developed ieCS.

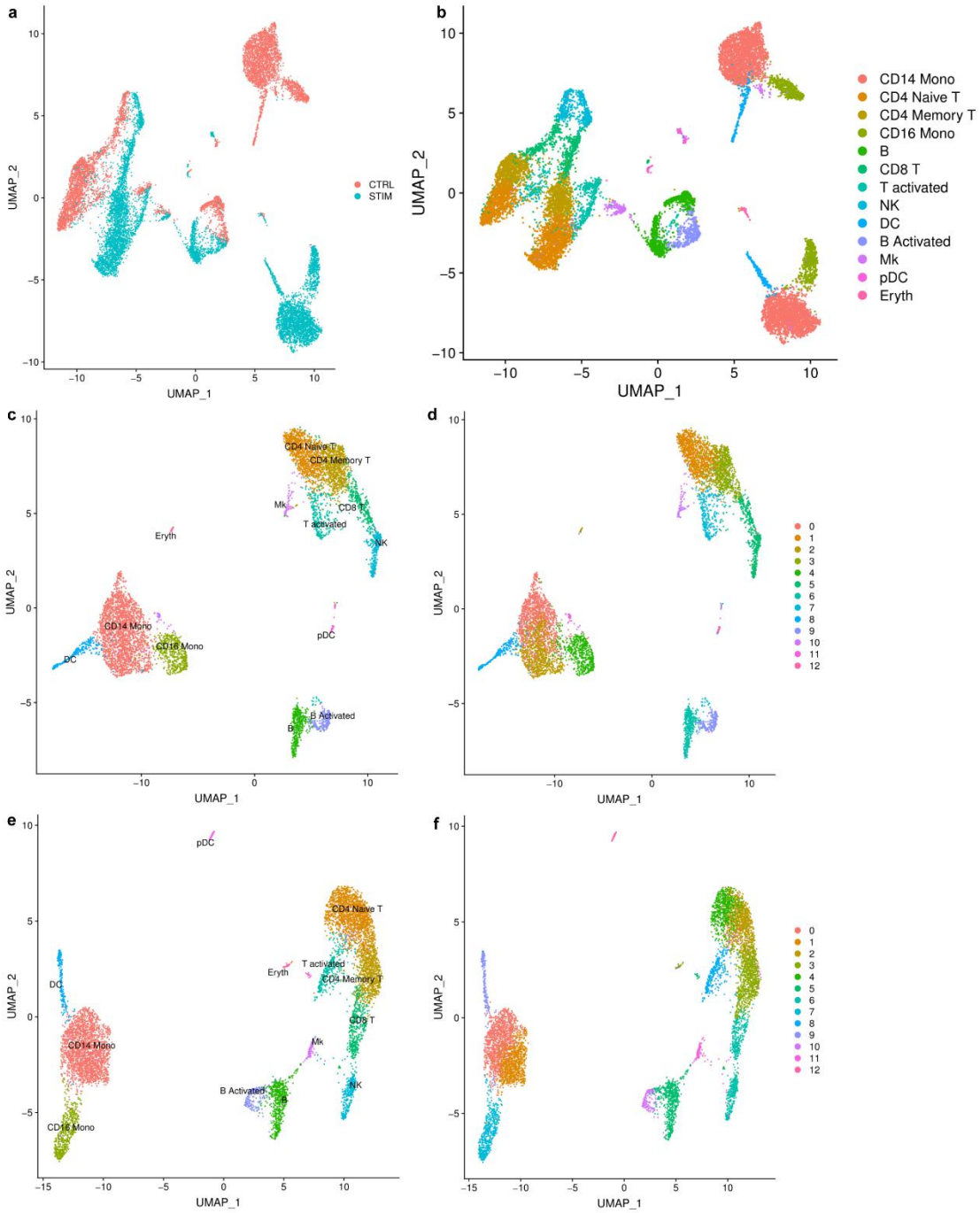


Figure 3.1 **UMAP visualization of the demonstration scRNA-Seq dataset**. a-b) Peripheral blood mononuclear cells (PBMCs) dataset (Kang et al., 2018). a) Colored by experimental conditions - STIM: interferon-beta (IFNB) stimulated; CTRL: control. b) Colored by cell type annotation. c-d) PBMCs under control condition. c) Colored by cell types. d) Colored by unsupervised clusters. e-f) PBMCs under IFNB stimulated condition. e) Colored by cell types. f) Colored by unsupervised clusters. Figures were adapted from Hai et al., in preparation.

Table 7 **The cell types and cell clusters in demo dataset.** Cell type labels were included in the demo dataset. The name of cell clusters was determined based on the

majority of cell type annotations for cells within each cluster.

Cell Type	Cell Cluster
B	CTRL_6_B; STIM_5_B
B_Activated	CTRL_9_B_Activated; STIM_10_B_Activated
CD14_Mono	CTRL_0_CD14_Mono; CTRL_2_CD14_Mono; STIM_0_CD14_Mono; STIM_1_CD14_Mono
CD16_Mono	CTRL_4_CD16_Mono; STIM_7_CD16_Mono
CD4_Memory_T	CTRL_3_CD4_Memory_T; STIM_3_CD4_Memory_T
CD4_Naive_T	CTRL_1_CD4_Naive_T; STIM_2_CD4_Naive_T; STIM_4_CD4_Naive_T
CD8_T	CTRL_5_NK-CD8_T; STIM_6_NK-CD8_T
DC	CTRL_8_DC; STIM_9_DC
Eryth	
Mk	CTRL_10_Mk; CTRL_11_Mk; STIM_11_Mk
NK	CTRL_5_NK-CD8_T; STIM_6_NK-CD8_T
pDC	CTRL_12_pDC; STIM_12_pDC
T_activated	CTRL_7_T_activated; STIM_8_T_activated

3.2 Design and Implementation

3.2.1 ieCS package

I developed ieCS using R language (R Core Team, 2018). The graphical user interface (GUI) of ieCS was built upon the Shiny framework (Chang et al., 2020).

Various functions in ieCS are based on a range of R packages, including: ape (Paradis and Schliep, 2018), collapsibleTree (Khan, 2018), DT (Xie et al., 2019), factoextra (Kassambara and Mundt, 2019), ggplot2 (Wickham, 2016), ggraph (Pedersen, 2019a), grid (R Core Team, 2018), gridExtra (Auguie, 2017), igraph (Csardi and Nepusz, 2006), pheatmap (Kolde, 2019), plotly (Sievert, 2020), plyr (Wickham, 2011), RColorBrewer (Neuwirth, 2014), stats (R Core Team, 2018), tidygraph (Pedersen, 2019b), methods (R Core Team, 2018), cluster (Maechler et al., 2021), colorspace (Zeileis et al., 2009),

dendextend (Galili, 2015), and ggpubr (Kassambara, 2020).

The source codes of ieCS are freely available on GitHub at https://github.com/L-Hai/ieCS/. The tool is provided as an R package, which simplifies installation and launch:

devtools::install_github("L-Hai/ieCS")

ieCS::run()

The workflow of ieCS is depicted in Figure 3.2. ieCS enables users to upload input files via the GUI. Subsequently, ieCS computes pairwise similarity scores among cell clusters and discerns superclusters within and between conditions. ieCS offers three methods (hierarchical clustering, network partitioning, and tree aggregation) to identify superclusters. Users could employ these methods in separate GUI tabs. Each tab allows users to select different parameters, facilitating interactive exploration of superclusters at varying degrees of similarity. In instances where markers for reference cell types are available, ieCS determins similarity scores between cell clusters and reference cell types. These scores could then aid in annotating the identified superclusters.

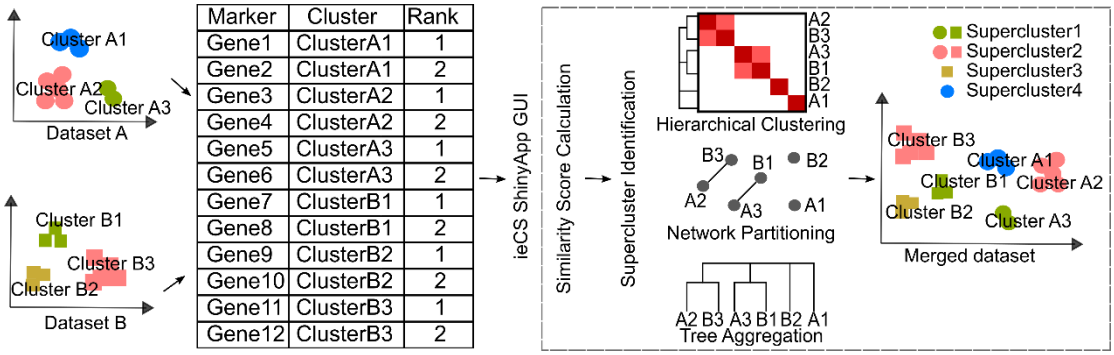


Figure 3.2 **Workflow of ieCS.** Progressing from left to right, the diagram portrays the sequence: Analyzing single datasets, identifying markers of cell clusters, providing to ieCS via ShinyApp's GUI as input. ieCS subsequently quantifies cell cluster similarity and identifies superclusters using three methods (hierarchical clustering, network partitioning, tree aggregation). Finally, ieCS visualizes superclusters using user-provided cell coordinates. Figures were adapted from Hai et al., in preparation.

The detailed description of the ieCS GUI, similarity score, three supercluster identification methods, and visualization will be provided in the following subsections.

3.2.2 ieCS GUI

The GUI of ieCS comprises six tabs:

- UploadData (Figure 3.3): This tab allows users to upload required files (Table 8) to perform similarity quantification between cell clusters.
- CSHierClust (Figure 3.4): In this tab, users can explore superclusters identified through global hierarchical clustering (GHC).
- CSHierClustDirect (Figure 3.5): This tab facilitates the exploration of superclusters identified through direct hierarchical clustering (DHC).
- CSNetwork (Figure 3.6): This tab is dedicated to exploring superclusters identified through network partitioning.
- CSTree (Figure 3.7): Users can explore superclusters identified through tree aggregation in this tab.
- CellEmbeddingPlot (Figure 3.8): Here, users can upload metadata (Table 9) and cell embedding files (Table 10) and visualize cells of superclusters in the cell embedding plots.

The details of similarity quantification and supercluster identification methods are presented in the following subsections.

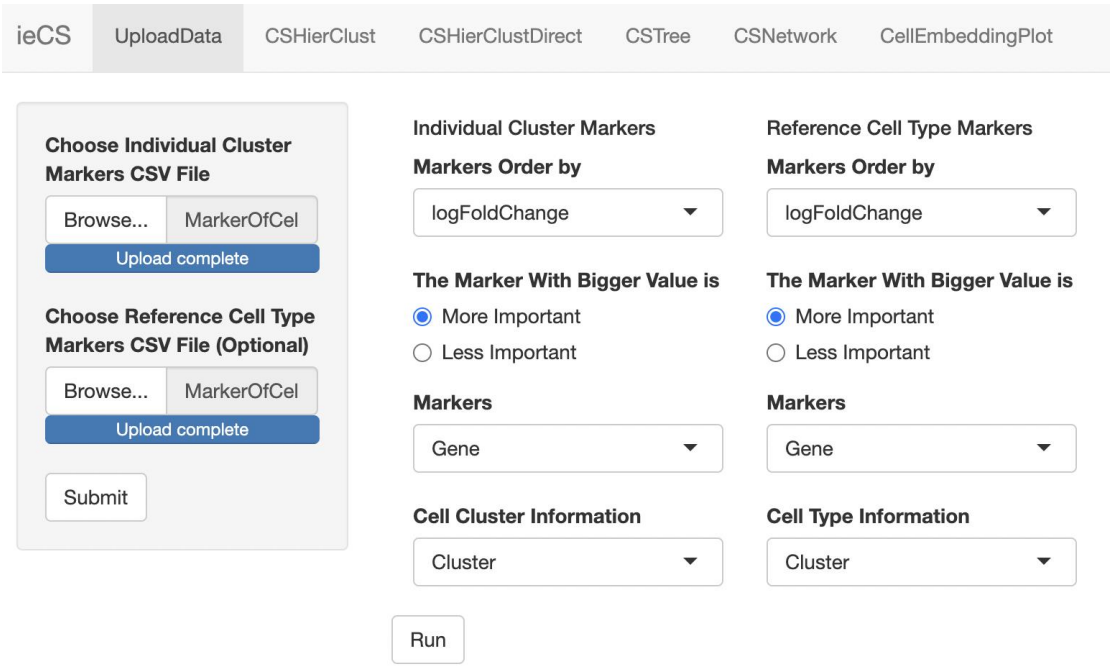


Figure 3.3 **UploadData tab for input data.** Within the UploadData tab, users upload input files in CSV format to ieCS. Upon clicking “Submit”, users configure column names in the files (Table 8), informing ieCS about marker order (these markers can be

ranked according to either fold change or p-value), markers, and cell clusters. Following configurations, clicking “Run” triggers ieCS to display the inputted details and initiate analysis. An option to upload cell type markers files as references for cell type annotation is available.

Table 8 **Input example of marker file for ieCS.** The marker file required the distinct markers of cell clusters, order of markers, cell cluster information. If a user wishes to supply a cell type reference for automated supercluster annotation, the input files required will adhere to the same format as this table.

Order (FoldChange)	Cell cluster	Marker
1.763792269	CTRL_0	S100A8
1.501863697	CTRL_0	S100A9
1.336291325	CTRL_0	CD14
1.477777015	CTRL_1	SELL
1.37907061	CTRL_1	GIMAP7
1.323339417	CTRL_1	LTB
1.878839211	STIM_1	CCL4
1.847322454	STIM_1	CCL3
1.702559738	STIM_1	SOD2

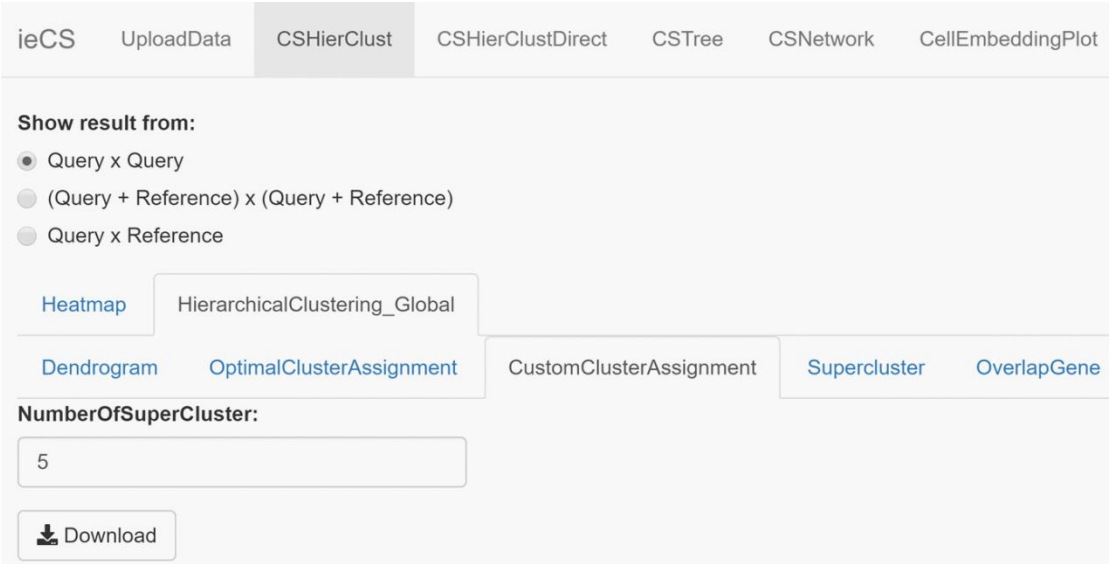


Figure 3.4 **CSHierClust tab for global hierarchical clustering (GHC).** Users can select various modes to calculate similarity scores between cell clusters, as detailed in subsection 3.2.3. They can then interactively identify superclusters using the GHC method in the subtabs: Heatmap, which displays a heatmap of similarity scores, and HierarchicalClustering_Global, where users can choose the optimal or a custom number of superclusters.

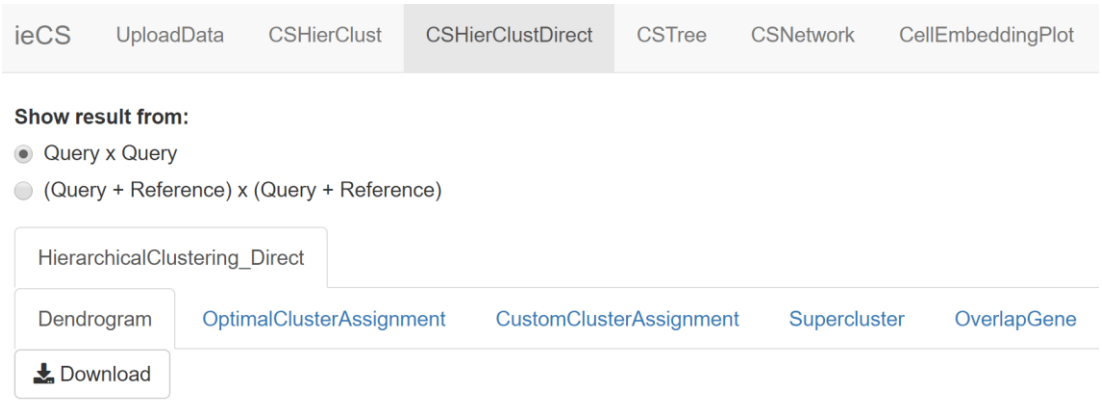


Figure 3.5 **CSHierClustDirect tab for direct hierarchical clustering (DHC).** Users can select various modes to calculate similarity scores between cell clusters, as detailed in subsection 3.2.3. They can then interactively identify superclusters using the DHC method in the subtab HierarchicalClustering_Direct, where users can choose the optimal or a custom number of superclusters.

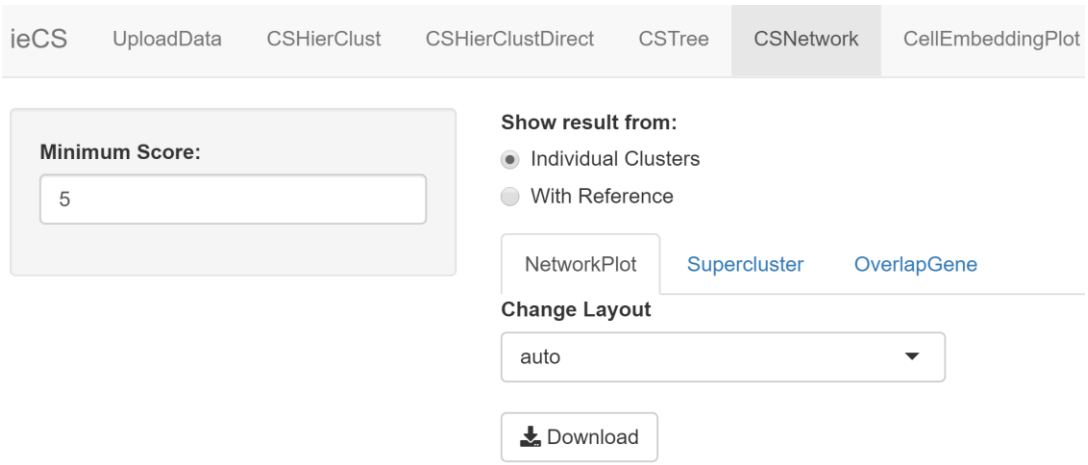


Figure 3.6 **CSNetwork tab for network partitioning.** In the left panel, users can select different cutoff values (minimum similarity scores) to generate a network of cell clusters. In the right panel, users can choose from various layouts to visualize the network of cell clusters. If a user has provided a cell type reference, they can opt to display the network showing both cell clusters and the reference.

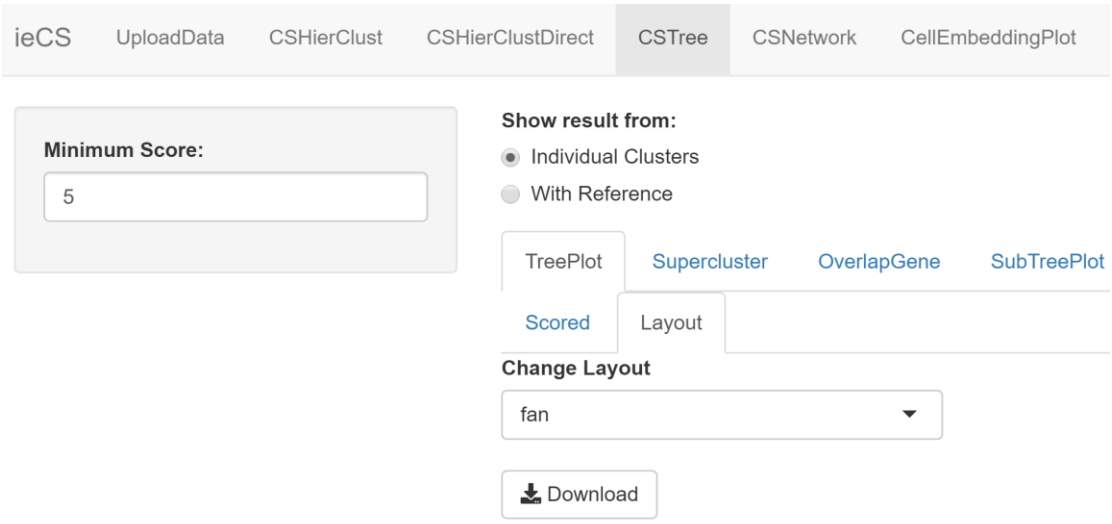


Figure 3.7 **CSTree tab for tree aggregation.** In the left panel, users can select different cutoff values (minimum similarity scores) to generate a tree of cell clusters. In the right panel, users can choose from various layouts to visualize the tree of cell clusters. If a user has provided a cell type reference, they can opt to display the tree showing both cell clusters and the reference.

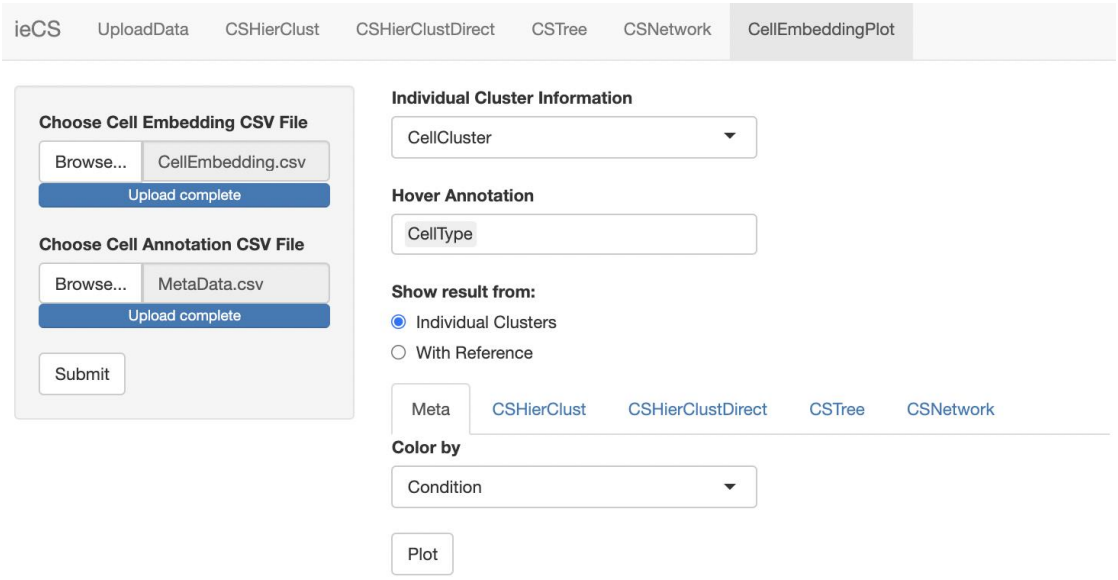


Figure 3.8 **CellEmbeddingPlot tab for cell visualization.** In the left panel, users can upload a cell coordination file (Table 9) and cell information (Table 10) in CSV format by clicking the "Submit" button. In the right panel, cluster information and hover annotations can be selected from columns within the cell information file. It is possible to color cells based on superclusters derived from the three supercluster identification methods, either with or without cell type markers. Once the configuration is set, users can generate the cell embedding plot by clicking the "Plot" button. The hover annotations for selected cells will be displayed upon mouse interaction.

Table 9 **Input example of cell coordination for ieCS.**

	UMAP_1	UMAP_2
AAACATACATTTCC.1	6.440331	7.502754
AAACATACCAGAAA.1	4.111667	8.648509
AAACATACCTCGCT.1	6.241261	8.087474
AAACATACCTGGTA.1	1.744034	3.217051
AAACATACGATGAA.1	-9.71896	1.293408
AAACATACGGCATT.1	5.879172	10.17724
AAACATACTGCGTA.1	-7.08594	1.232497
AAACATACTGCTGA.1	-10.1493	-0.05283

Table 10 **Input example of cell information for ieCS.**

	Condition	Annotation	Cell cluster
AAACATACATTTCC.1	IMMUNE_CTRL	CD14 Mono	CTRL_0
AAACATACCAGAAA.1	IMMUNE_CTRL	CD14 Mono	CTRL_2
AAACATACCTCGCT.1	IMMUNE_CTRL	CD14 Mono	CTRL_0
AAACATACCTGGTA.1	IMMUNE_CTRL	pDC	CTRL_12
AAACATACGATGAA.1	IMMUNE_CTRL	CD4 Memory T	CTRL_3
AAACATACGGCATT.1	IMMUNE_CTRL	CD14 Mono	CTRL_0
AAACATACTGCGTA.1	IMMUNE_CTRL	T activated	CTRL_7
AAACATACTGCTGA.1	IMMUNE_CTRL	CD4 Naive T	CTRL_1

3.2.3 Similarity score

I introduce the “similarity score”, an innovative metric for quantifying the similarity between two cell clusters. This score is determined by considering both the count of shared markers and the ranks of these shared markers within the two cell clusters. If two cell clusters exhibit numerous markers in common, especially with higher ranks, the resultant similarity score will be elevated, indicating greater similarity between these clusters.

The similarity score is calculated using the following equation:

$$S_{ij} = \sum_{k=1}^{n} \frac{50}{R_{ig_k} + R_{jg_k}}$$

In this equation, S_{ij} denotes the similarity score between cell clusters i and j. R_{ig_k} and R_{ig_k} correspond to the ranks of overlapping marker g_k within clusters i and j, respectively.

The ranks of a marker undergo a reciprocal transformation, assigning a higher value

to the top-ranked marker (indicating greater importance) and a smaller value to the lower-ranked marker (indicating lesser importance) (Figure 3.9a). In addition, the non-linear nature of the reciprocal transformation renders the value insensitive to the influence of markers with low ranks (Figure 3.9a).

The reciprocal value is multiplied by a scale factor of 50. This scaling allows the maximum similarity scores, indicating two identical cell clusters, to reach a score around 100 even with various numbers of markers (Figure 3.9b).

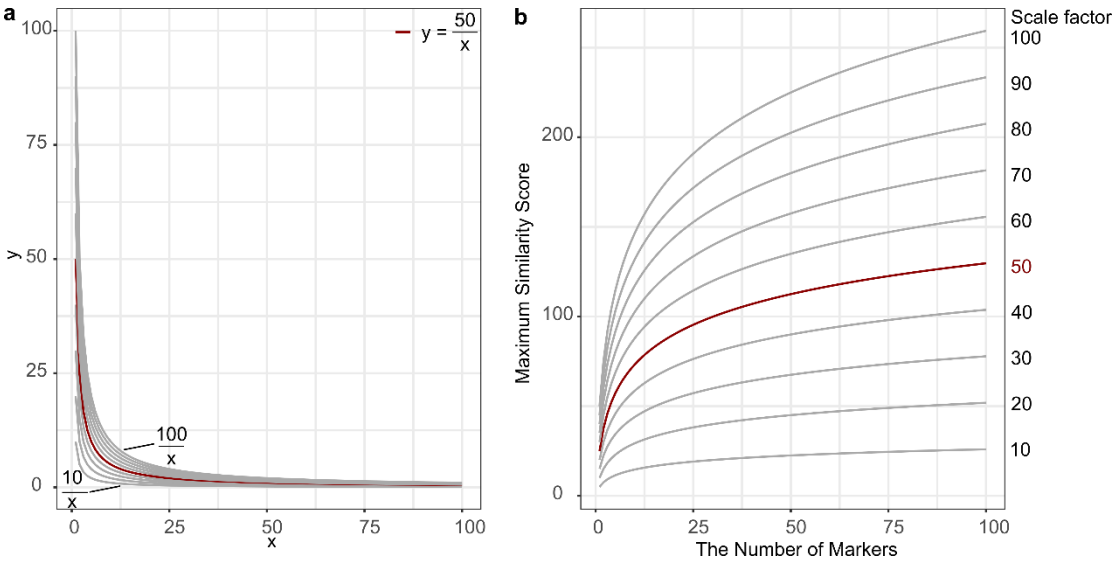


Figure 3.9 **Reciprocal transformation and scale factor in similarity score**. a) The curves depict reciprocal functions with scale factors of 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100. b) Each curve represents the maximum similarity score between two cell clusters with varying numbers of markers as input, and each curve corresponds to a specific scale factor indicated by the number on the right. Curves with a scale factor of 50 are highlighted in red.

In scenarios where cell type markers are provided as a reference (as Table 8) and uploaded to ieCS, the tool proceeds to calculate similarity scores between the cell types and the cell clusters. Two modes are available to calculate the similarity score matrix with reference:

A. (Query + Reference) x (Query + Reference)
B. Query x Reference

In mode A, the resulting similarity score matrix contains both cell types and cell clusters in both rows and columns. In mode B, the resulting similarity score matrix comprises cell clusters in its rows and cell types in its columns. In mode A, each element in the

matrix represents the similarity score between a cell type and a cell type, a cell type and a cell cluster, or a cell cluster and a cell cluster. In mode B, each element in the matrix represents the similarity score between a cell cluster and a cell type.

With the similarity score matrix, superclusters can be identified using hierarchical clustering, network partitioning, and tree aggregation methods in the ieCS. The details of these methods are shown in the following subsections.

3.2.4 Hierarchical clustering

I facilitated agglomerative hierarchical clustering with the complete-linkage algorithm to clustering cell clusters. There are two types of distance measures for hierarchical clustering in ieCS:

A. Global Hierarchical Clustering (GHC): Computing the Euclidean distance between two cell clusters based on the similarity score matrix using the following formula:

$$G(i,j) = \sqrt{\left(S_{i1} - S_{j1}\right)^2 + \left(S_{i2} - S_{j2}\right)^2 + \cdots + \left(S_{in} - S_{jn}\right)^2}$$

Here, $G(i,j)$ is the Euclidean distance between cell cluster i and j. S_{i1} represents the similarity score between cell cluster i and cell cluster 1. n is the total number of cell clusters.

I generated the distance matrix and performed hierarchical clustering using the pheatmap package.

B. Direct Hierarchical Clustering (DHC): Converting the similarity score matrix into a distance matrix using the following formula on each row of the similarity score matrix:

$$D(r,k) = 1 - \frac{S_{rk}}{Max(S_{r1}, S_{r2}, \ldots, S_{rn})}$$

Here, $D(r,k)$ is the distance mesure between cell cluster r (in the row r of the similarity score matrix) and cell cluster k. S_{rk} represents the similarity score between cell cluster r and cell cluster k. n is the total number of cell clusters. $Max(S_{r1}, S_{r2}, \ldots, S_{rn})$ represents the maximum score among the similarity score between cell cluster r and the other cell clusters.

I computed the distance matrix in R, then transformed it into a distance object using the as.dist function. Subsequently, I performed hierarchical clustering using the hclust function in the stats package.

In the CSHierClust tab of ieCS, the similarity score matrix is visualized in a heatmap with a dendrogram showing the GHC on the rows and columns using the pheatmap package. In the CSHierClust and CSHierClustDirect tab of ieCS, the hierarchical clustering of cell clusters is also displayed in the dendrogram using ggplot2, grid, stats, ggpubr, and factoextra packages.

To identify the optimal number of superclusters based on hierarchical clustering, I applied the Silhouette method. The silhouette width for a range of potential supercluster numbers was calculated using the cluster, stats, and factoextra package. Optimal number of superclusters was determined by selecting the number that resulted in the highest average Silhouette width.

In addition to employing the Silhouette method, ieCS provides users with the flexibility to interactively specify a total number of superclusters. Subsequently, ieCS assigns cell clusters into superclusters. For the assignment basing on GHC, the fviz_dend function from the factoextra package was used. For the assignment basing on DHC, the color_branches function from the dendextend package was used.

In the presence of uploaded cell type markers, cell clusters were assigned to specific cell types based on the highest similarity score. The resulting assignment was visually represented using the collapsibleTree package.

Additionally, ieCS provides access to overlapped genes among similar cell clusters within the same supercluster, allowing users the opportunity to manually annotate these superclusters.

3.2.5 Network partitioning

The similarity between cell clusters can be represented in a network, where each individual cell cluster is assigned as a node within the network. The edges linking these clusters are determined by the corresponding similarity scores.

The network of cell clusters is contructed utilizing the igraph package. Users possess the ability to reconstruct the network by interactively setting a minimum cutoff for similarity scores, allowing for the removal of certain edges between cell clusters.

The process of partitioning the network into distinct communities of cell clusters, which correspond to superclusters, was executed using the Louvain community detection algorithm via the cluster_louvain function available in the igraph package.

To provide visual insight into these superclusters, networks are visualized using the ggraph and tidygraph packages. Superclusters were visually represented using a color-coding scheme. Users have the flexibility to select various network layouts for the representation of cell clusters within the network. The available layout options include "auto," "fr," "kk," "gem," "dh," "graphopt," "mds," "drl," and "lgl" (using igraph layout algorithms via ggraph package, https://rdrr.io/cran/ggraph/man/layout_tbl_graph_igraph.html).

When a cell type reference is provided, the similarity score matrix for (Query + Reference) x (Query + Reference) is computed. Each individual cell type is also treated as a node within the network representation.

3.2.6 Tree aggregation

ieCS can organize cell clusters into a tree structure, where comprising cell clusters as the leaves, and the edges are indicative of the similarity scores between these cell clusters. The tree is contructed through the following steps:

1) User defines a cutoff, D_u, for the minimum similarity score to be considered.
2) Obtaining a set of similaritiy scores $Set(r)$ in the row r of the similarity score matrix. This set contains similarity scores between cell cluster r and the other cell clusters. $Set(r) = (S_{r1}, S_{r2}, \dots, S_{rn})$ Here, S_{r1} represents the similarity score between cell cluster r and cell cluster 1. n is the total number of cell clusters.
3) The similarity scores less than D_u in $Set(r)$ are removed, leading to the generation of a new set $Set_{D_u}(r)$.
4) Gradually increasing the cutoff and then removing the smaller similarity scores in $Set(r)$ results in a list of sets: $SetList(r) = Set_{D_u}(r), Set_{D_u+1}(r), Set_{D_u+2}(r),$ … , $Set_{Max(S_{r1},S_{r2},\dots,S_{rn})}(r)$.

5) Performing Step 2-4 for each row of the similarity score matrix results in n lists of sets: $SetList(1), SetList(2), \dots, SetList(n)$.

6) Ordering all the sets based on the cutoffs from D_u to the maximum similarity score.

e.g.:

$$\begin{gathered} Set_{D_u}(1), Set_{D_u}(2), \dots, Set_{D_u}(n) \\ Set_{D_u+1}(1), Set_{D_u+1}(2), \dots, Set_{D_u+1}(n) \\ \vdots \\ Set_{Max(S_{r1}, S_{r2}, \dots, S_{rn})}(r) \\ \vdots \\ Set_{Max(all)}(m) \end{gathered}$$

Here, assuming the maximum similarity score $Max(all)$ is in the row m.

7) Removing redundant sets.

e.g.: If $Set_d(r) = (S_{r1})$ and $Set_d(1) = (S_{1r})$,

since $S_{r1} = S_{1r}$ (both represent the similarity score between cell cluster r and 1), then $Set_d(r) = Set_d(1)$. $Set_d(1)$ and $Set_d(r)$ are redundant, one of them will be removed.

8) Agglomeratively merging sets: if a set with a lower cutoff contains all the cell clusters in a set with a higher cutoff. This merging process of cell clusters are stored in the Newick format for the construction of a tree.

e.g.: If there is a set with a cutoff of 80 containing only one element, cell cluster r and a ($Set_{80}(r) = S_{ra}$), and another set with a cutoff of 68 containing two elements: cell cluster r and a, and cell cluster r and b ($Set_{68}(r) = S_{ra,}S_{rb}$), the merging results in Newick format will be "(((r,a):80,b):68);".

9) The structure of tree in Newick format is contructed into a tree diagram utilizing the read.tree function in the ape package. Within the tree structure, cell clusters are portrayed as leaves and the similarity scores are marked on the edges connecting cell clusters.

e.g.,

$$D_u = 30$$

$$S_{rr} = S_{aa} = S_{bb} = S_{cc} = S_{dd} = 100$$

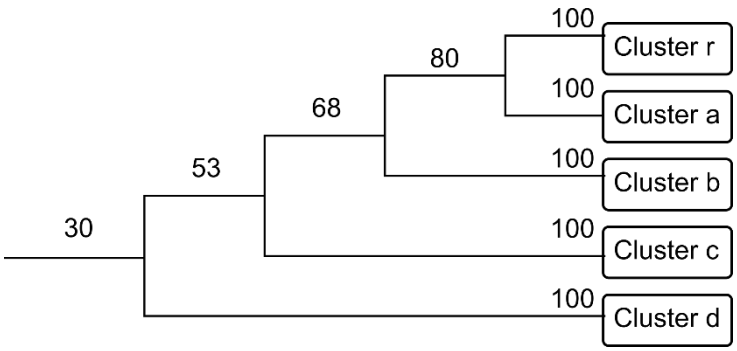
$$Set_{30}(r) = S_{ra,}S_{rb}, S_{rc}, S_{rd}$$

$$Set_{53}(r) = S_{ra,}S_{rb}, S_{rc}$$

$$Set_{68}(r) = S_{ra,}S_{rb}$$

$$Set_{80}(r) = S_{ra}$$

Here, S_{rr} represents the maximum similarity score of cell cluster r (the similarity score between cell cluster $r$and itself). The structure of tree in Newick format is "(((((r:100,a:100):80,b:100):68,c:100):53,d:100):30);". A tree diagram of the cell clusters can be presented as:



It indicates that these five cell clusters (r, a, b, c, d) have similarity scores higher than the cutoff, suggesting that they can form a supercluster.

Users possess the ability to reconstruct the tree by interactively setting a minimum cutoff for similarity scores, allowing for the removal of certain edges between cell clusters.

Apart from the score-marked trees, ieCS offers the functionality to color edges and leaves of trees based on superclusters. These trees can be visualized in five alternative layouts: "fan," "radial," "cladogram," "phylogram," and "unrooted" (using layout algorithms in ape package, https://cran.r-project.org/web/packages/ape/vignettes/DrawingPhylogenies.pdf).

When a cell type reference is provided, the similarity score matrix for (Query + Reference) x (Query + Reference) is computed. Each individual cell type is also treated as a leaf within the tree representation.

3.2.7 Cell visualization

Users can upload their cell embedding (Table 9) and metadata (Table 10) into ieCS. This allows cells to be visualized in the dedicated CellEmbeddingPlot tab. Superclusters that have been identified using the three identification methods can be visualized on these cell embedding plots using the plotly package. In the visual representations, cells belonging to the same supercluster are assigned a consistent color code. This color coding facilitates clear identification of cell clusters associated with a particular supercluster. The metadata and supercluster annotations of the cells will be displayed interactively when the user hovers their mouse over them.

3.3 Application Results

I applied the Kang et al., 2018 PBMCs dataset (Figure 3.1 in the subsection 3.1 Motivation) to demonstrate the usage of ieCS. The goal was to identify similar cell clusters across IFNB-stimulated and control conditions. Using the FindAllMarker function in the Seurat package, I identified the top 50 markers of each cell cluster in Figure 3.1d and f. These markers were then sorted by fold change and served as the required input for ieCS (in Table 8 format).

Additionally, for cell types assigned by Kang et al., 2018 (Figure 3.1b), the markers of each cell type were identified using the FindAllMarker function in the Seurat package. These markers served as the cell type reference for ieCS (in Table 8 format). The UMAP embedding information (Figure 3.1a) was also uploaded to ieCS (in Table 9-10 format).

In this section, I will present the results obtained from three supercluster identification methods on the demo dataset. The comparison and evaluation of these results among different methods will be discussed in the next section, 3.4 Evaluation and Discussion.

3.3.1 Hierarchical clustering

3.3.1.1 Global hierarchical clustering

In the CSHierClust tab of the ieCS webpage (Figure 3.4), ieCS visualized the similarity score matrix and the dendrograms of GHC through a heatmap (Figure 3.10). The heatmap displays cell clusters within the same cell types exhibit higher similarity scores and cluster together (Figure 3.10).

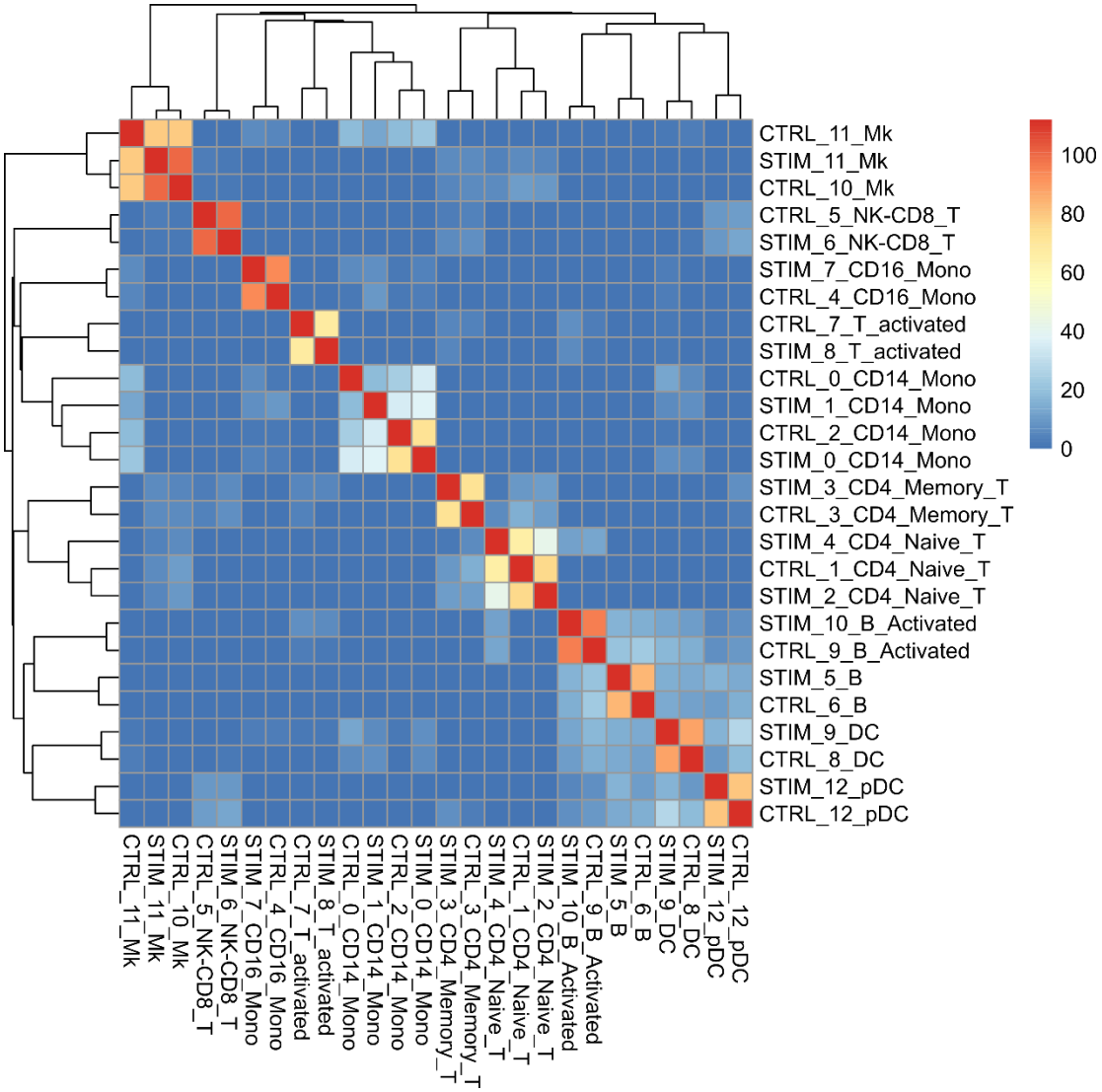


Figure 3.10 **Heatmap of the similarity score matrix and GHC dendrograms.** The color gradient ranges from blue to red, indicating low to high similarity scores between cell clusters. The dendrograms illustrates the global hierarchical clustering (GHC) results.

To determine the optimal number of superclusters, Silhouette method was applied. 11 superclusters exhibited the highest average Silhouette width were identified (Figure 3.11). Notably, the superclusters consist of cell clusters that share the same cell types from two different conditions (Figure 3.11).

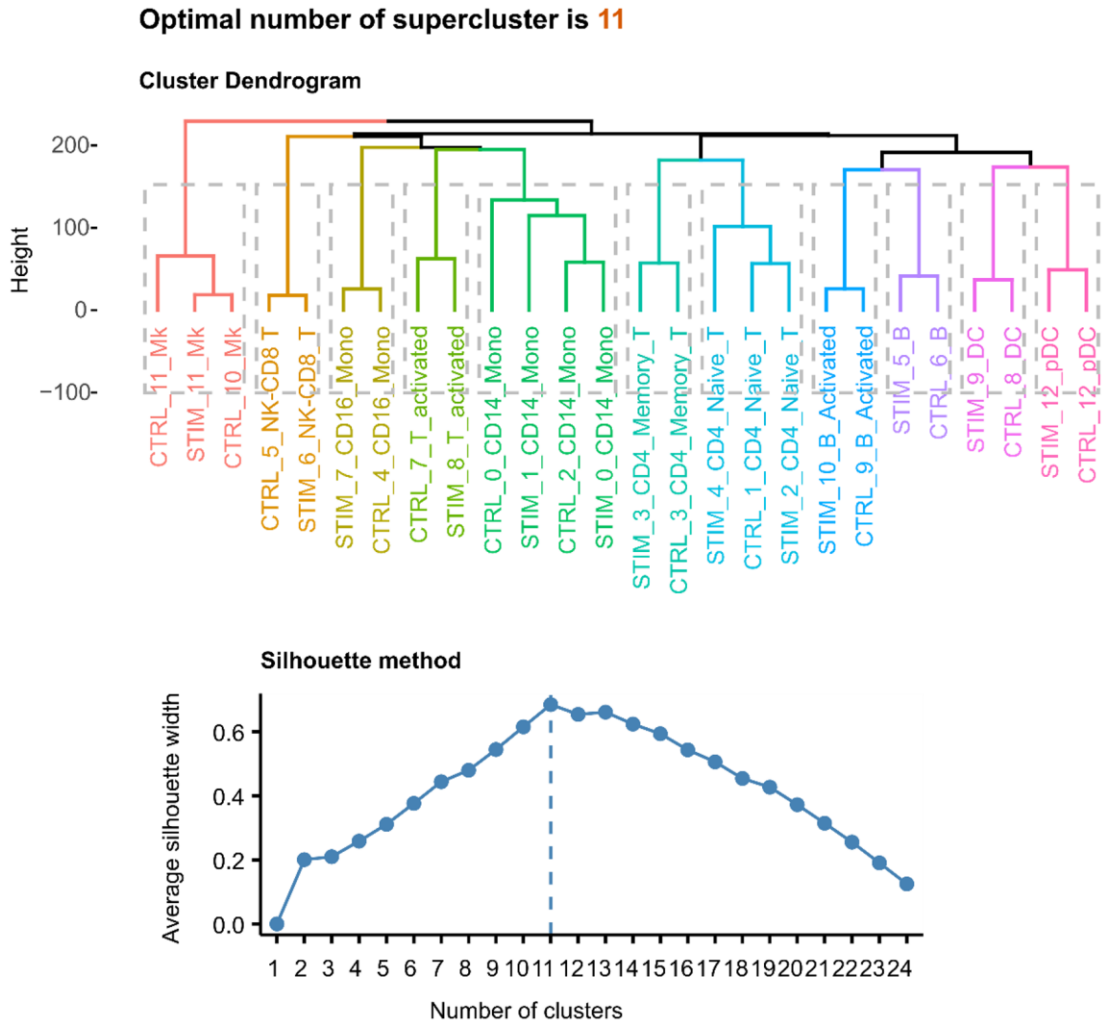


Figure 3.11 **Optimal number of superclusters in GHC.** Top panel displays the dendrogram resulting from GHC. Clusters are color-coded based on the optimal number of superclusters. The bottom panel showcases the average Silhouette width for varying numbers of superclusters.

Moreover, users possess the flexibility to define the number of superclusters, enabling ieCS to allocate cell clusters accordingly basing on the GHC dendrogram (Figure 3.12). ieCS can dynamically respond to user input, allowing for the interactive display of resulting superclusters.

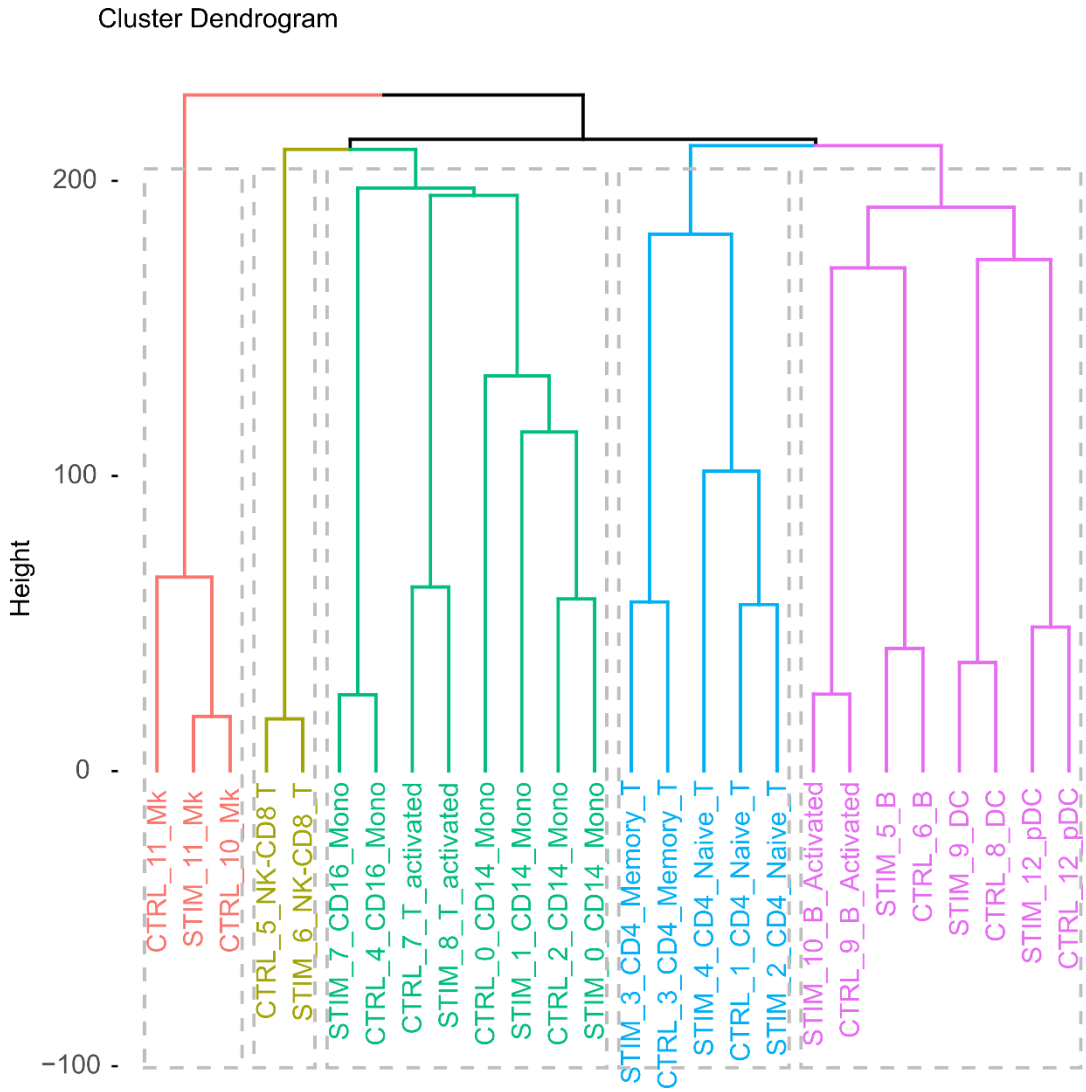


Figure 3.12 **Custom number of superclusters in GHC**. ieCS-assigned cell clusters organized into superclusters according to user-defined supercluster count of 5 superclusters. The dendrogram visualizes GHC, and clusters are differentiated by distinct colors representing the superclusters.

If cell type markers are uploaded onto ieCS as references (in Table 8 format), two modes are available to calculate the similarity score matrix:

A. (Query + Reference) x (Query + Reference)
B. Query x Reference

In mode A, the heatmap displays cell clusters and cell types in both rows and columns (Figure 3.13). Conversely, in mode B, the heatmap exhibits cell clusters in the rows and cell types in the columns (Figure 3.14). In mode B, ieCS can automatically assign a cell cluster to the specific cell type with the highest similarity score (Figure 3.15). Both modes yielded in the same optimal number of 11 superclusters (Figure 3.16-17).

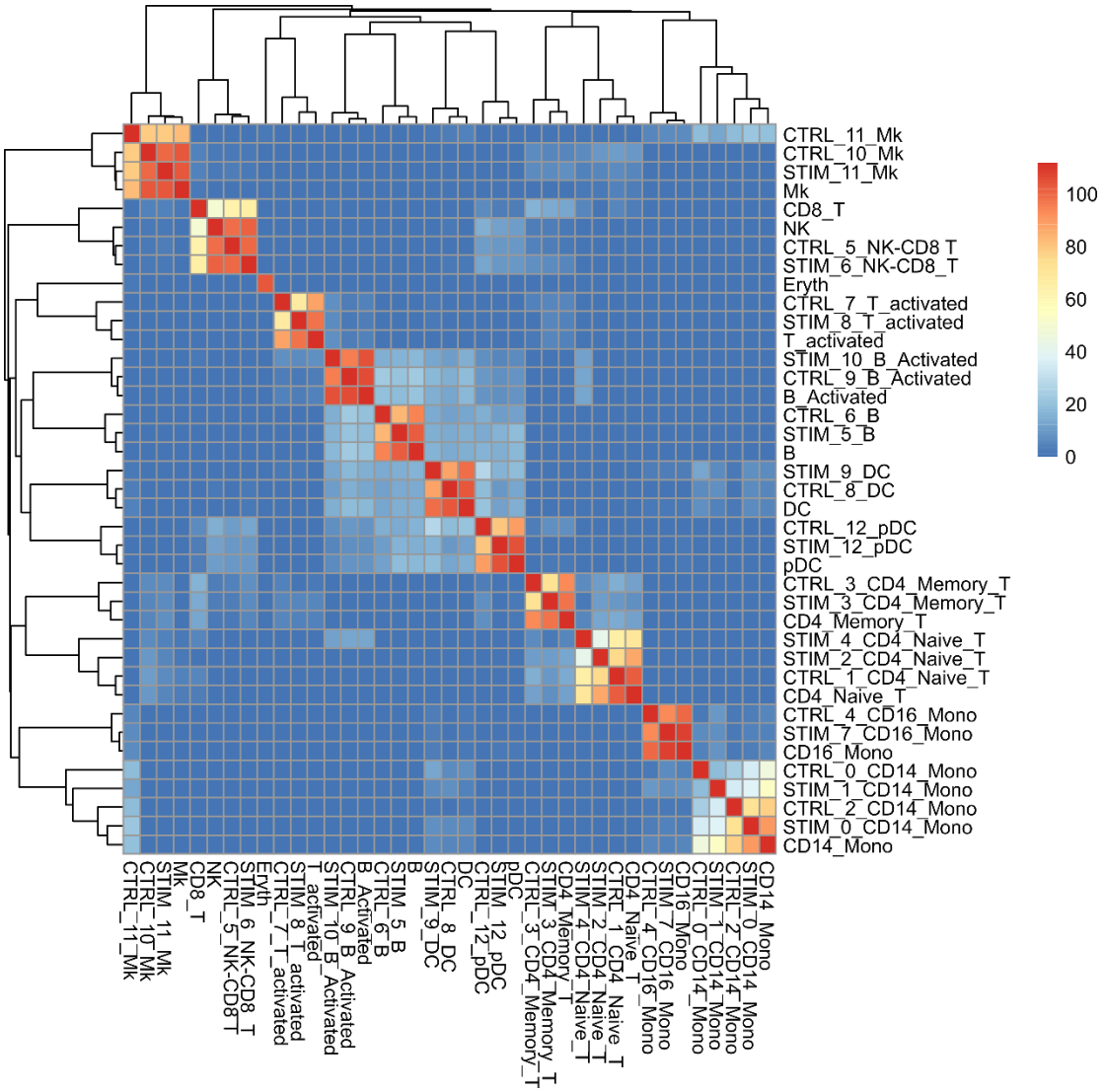


Figure 3.13 **Heatmap of the similarity score matrix in mode A and GHC dendrograms.** Mode A: (Query + Reference) x (Query + Reference). The color gradient ranges from blue to red, indicating low to high similarity scores. The dendrograms demonstrate the GHC arrangement of cell clusters and cell types.

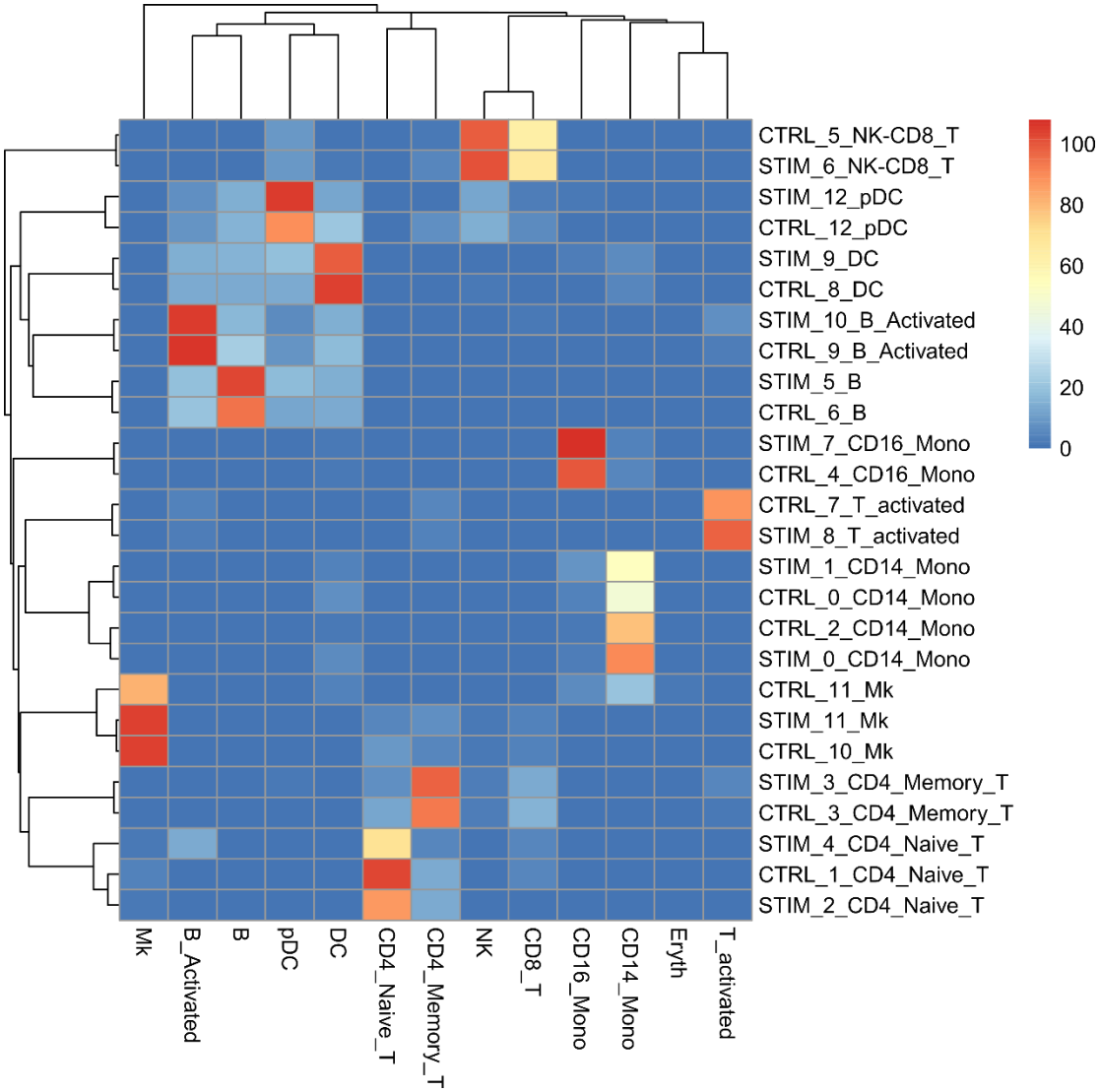


Figure 3.14 **Heatmap of the similarity score matrix in mode B and GHC dendrograms.** Mode B: Query x Reference. The color gradient ranges from blue to red, indicating low to high similarity scores. The dendrogram on the left demonstrates the GHC arrangement of cell clusters, while the dendrogram on the top demonstrates the GHC arrangement of cell types.

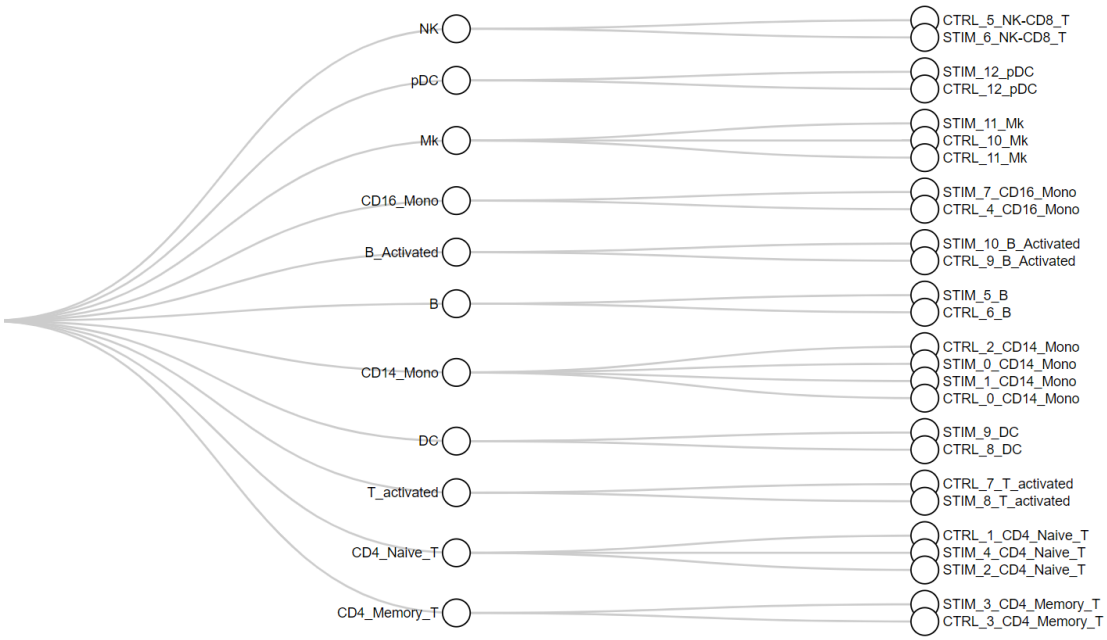


Figure 3.15 **Assignment of cell clusters to cell types in GHC with mode B.** Cell clusters are assigned to cell types based on the highest similarity scores in Mode B: Query x Reference.

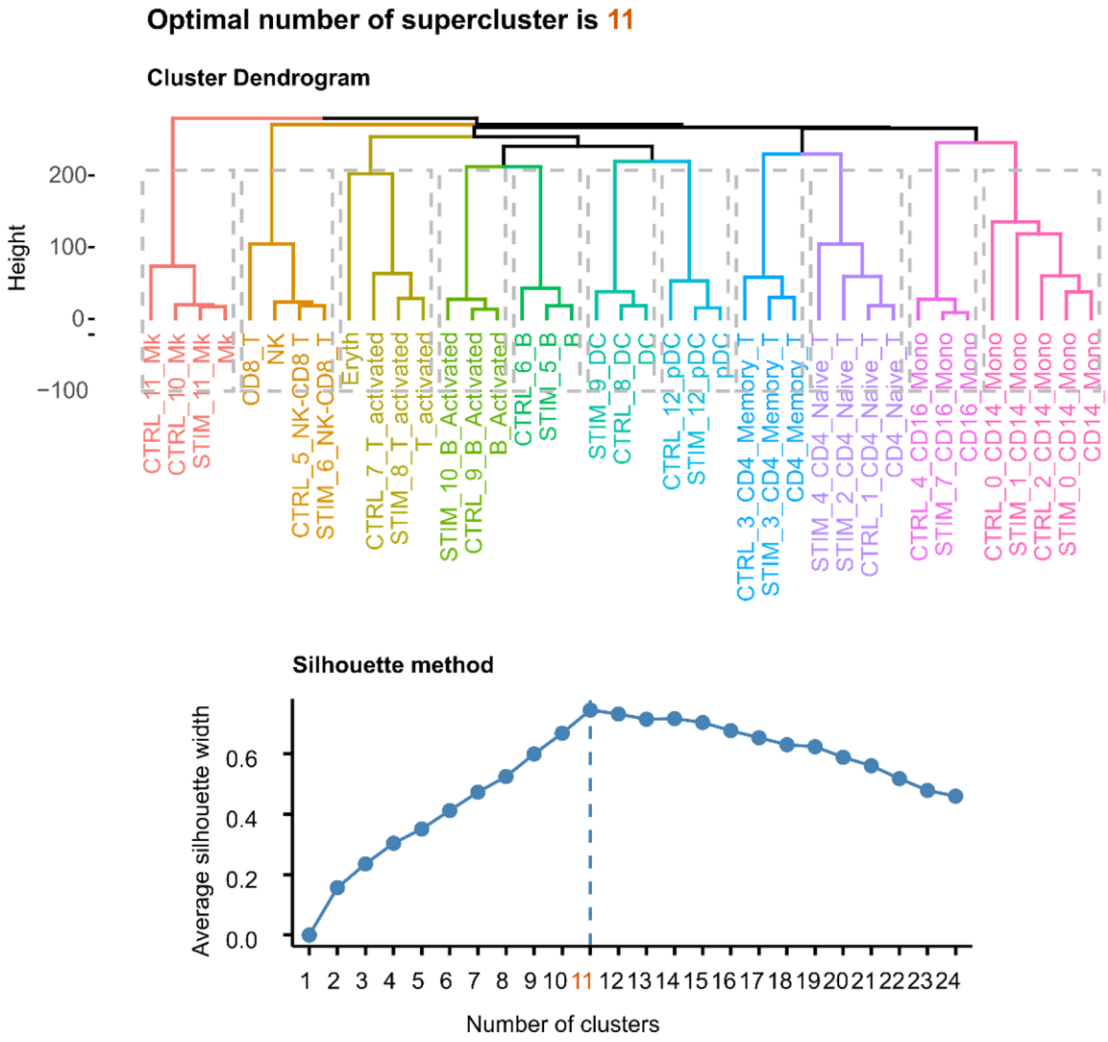


Figure 3.16 **Optimal number of superclusters in GHC with mode A.** Mode A: (Query + Reference) x (Query + Reference). Top: The dendrograms demonstrate the GHC arrangement of cell clusters and cell types. Clusters are color-coded according to the optimal number of superclusters. Bottom: The average Silhouette width for different supercluster counts.

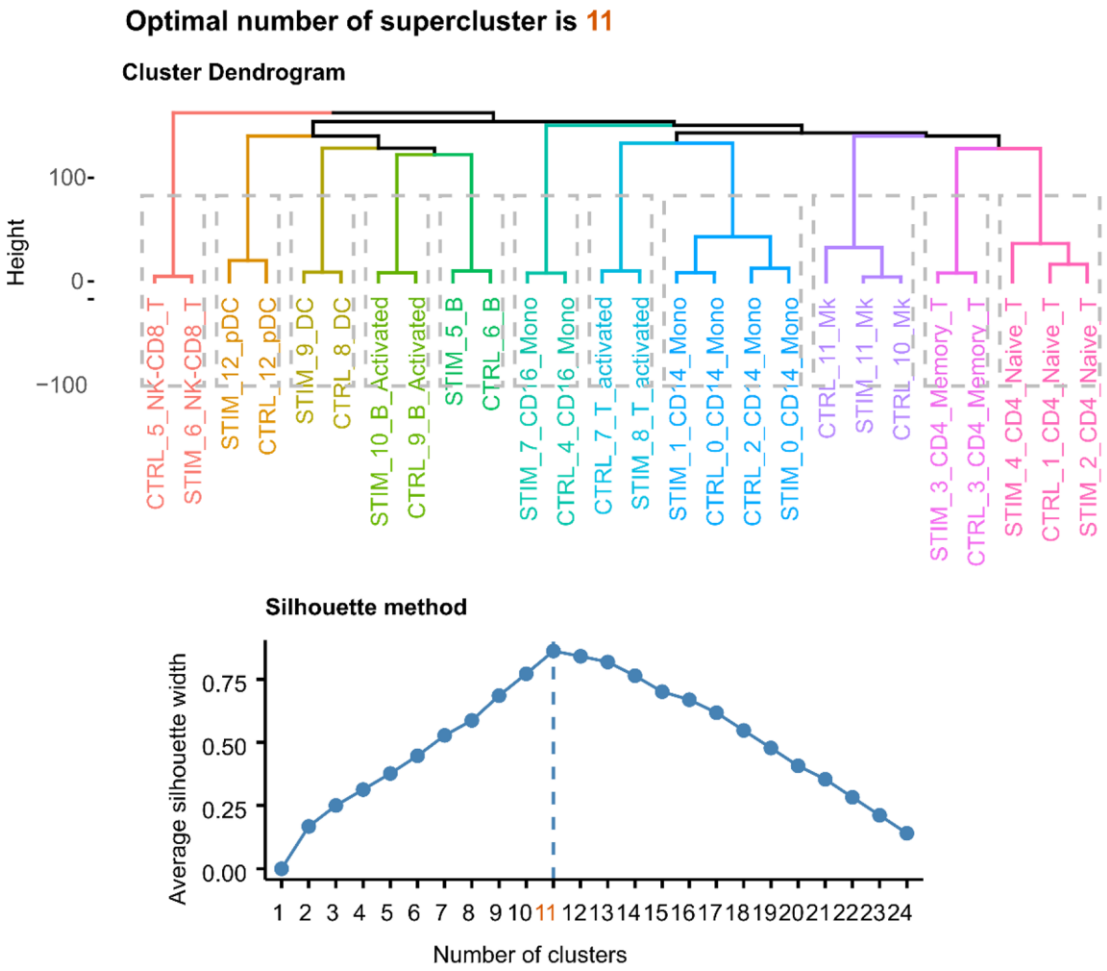


Figure 3.17 **Optimal number of superclusters in GHC with mode B.** Mode B: Query x Reference. Top: The dendrograms demonstrate the GHC arrangement of cell clusters. Clusters are color-coded according to the optimal number of superclusters. Bottom: The average Silhouette width for different supercluster counts.

If cell type reference is not available, ieCS provides the overlapping genes among cell clusters within the same superclusters for users to manually annotate the supercluster (Table 11).

Table 11 **The overlapping genes among cell clusters within the same superclusters.** Each row in the table represents a specific supercluster.

Cell Cluster	Overlapping Gene
CTRL_6_B; STIM_5_B	IGLL5; CD79A; MS4A1; IRF8; CD74; IGJ; C7orf50; BLNK; ID3; BANK1; RGS2; CD83; CD37; CXCR4; HVCN1; TNFRSF13B; HERPUD1; KIAA0226L; HLA-DQB1; HLA-DPB1; SYNGR2; HLA-DQA1; CHPT1; HLA-DPA1; TCL1A; CD79B; CDC37; EZR; RCSD1; ZFP36L1; PKIG; GNG7
CTRL_9_B_Activated; STIM_10_B_Activated	MIR155HG; MYC; HLA-DQA1; ID3; NME1; CD83; IRF8; NPM1; RAN; RANBP1; DUSP4; YBX1; PRMT1; FABP5; SRM; HSPE1; NHP2; HSPD1; REL; TVP23A; HLA-DQB1; NCL; RPL22L1; HERPUD1; HSP90AB1; DDX21; PMAIP1; C1QBP; SNRPD1; CD70; CKS2; SYNGR2; EBNA1BP2; EIF5A; PHB; DCTPP1; LY9
CTRL_0_CD14_Mono; CTRL_2_CD14_Mono; STIM_0_CD14_Mono; STIM_1_CD14_Mono	IL8; FTL; ANXA5; S100A9; CD63; C15orf48; S100A8; FCN1; FCER1G; LGALS1; TYMP
CTRL_4_CD16_Mono; STIM_7_CD16_Mono	VMO1; FCGR3A; MS4A7; MS4A4A; CXCL16; LST1; AIF1; C3AR1; FAM26F; CTSC; CD86; CFD; SLC31A2; SERPINA1; WARS; PILRA; TIMP1; GBP5; PLAC8; VASP; ATP1B3; ADA; RP11-290F20.3; FGL2; HN1; TNFSF10; GLUL; SNX10; GBP1; STXBP2; FCER1G; CST3; COTL1; CXCL10; IFITM3; IFITM2
CTRL_3_CD4_Memory_T; STIM_3_CD4_Memory_T	ALOX5AP; TRAT1; ZFP36L2; GPR171; PABPC1; TNFRSF4; CD2; SPOCK2; GPR183; IL32; CXCR4; FYN; CREM; CD7; IL7R; ITM2A; CD3D; RPL3; RARRES3; LEPROTL1; RPL14; PIK3R1; NR3C1; SESN3; FXYD5; RPS4X; CLEC2D; LCK; PTPRC; RPS18; RPSA; PBXIP1
CTRL_1_CD4_Naive_T; STIM_2_CD4_Naive_T; STIM_4_CD4_Naive_T	SELL; GIMAP7; LTB; CD3D; LEF1; LDHB; AES; GIMAP4; GIMAP5; CD3G; CCR7; GTF3A; RPS6; RPS15A; RPL32; RPL5; GIMAP1; RPL10A
CTRL_5_NK-CD8_T; STIM_6_NK-CD8_T	GNLY; NKG7; CCL5; GZMB; FGFBP2; APOBEC3G; CST7; CLIC3; GZMH; KLRD1; CTSW; GZMA; PRF1; CXCR3; HOPX; CHST12; TNFRSF18; KLRC1; RARRES3; C1orf21; SH2D2A; CD247; CD7; APMAP; DUSP2; LDHA; GCHFR; AOAH; XCL2; CD8A; GZMK; C12orf75; SLA2; CD2; C5orf56; FASLG; EVL; CD96; TIGIT
CTRL_8_DC; STIM_9_DC	HLA-DPB1; TXN; MARCKSL1; HLA-DQA1; HLA-DPA1; HLA-DRA; CD74; HLA-DRB1; HLA-DQB1; CST3; LYZ; SERPINB1; CD83; HLA-DMA; FABP5; GPR137B; RAB9A; CCL22; IDO1; HLA-DRB5; LSP1; ALDH2; ANXA2; ID2; SYNGR2; CD86; GPR183; CCR7; FAM49A; RAMP1; ACTB; CFP; BID; IL4I1
CTRL_10_Mk; CTRL_11_Mk; STIM_11_Mk	PPBP; PF4; GNG11; SDPR; NRGN; TUBB1; NCOA4; ACRBP; CLU; SPARC; MAP3K7CL; TREML1; TSC22D1; HIST1H2AC; RGS18; MYL9
CTRL_12_pDC; STIM_12_pDC	TSPAN13; TXN; PTGDS; GZMB; ITM2C; SEC61B; IGJ; HERPUD1; P2RY6; IRF8; TCL1A; CLIC3; SERPINF1; DNASE1L3; MAP1A; TCF4; PPP1R14B; PLD4; CD74; HLA-DMA; PARK7; HLA-DQB1; VAMP8; BIK; TRAF4; MZB1; CYB561A3; CTSC; PLA2G16
CTRL_7_T_activated; STIM_8_T_activated	HSPH1; HSPE1; CACYBP; GADD45B; CD69; HSPD1; SRSF7; SRSF2; HSPA8; ZFAND2A; RSRC2; YPEL5; DNAJB1; HSP90AB1; CLK1; CHORDC1; BIRC3; TCP1; SOD1; NOP58; HSPA1B; UBC; MRPL18; TSC22D3; DDIT4; SNHG15; UBB; DNAJB6; JUN; PIK3IP1; BTG1; SNHG8; STK17A; EIF5

3.3.1.2 Direct hierarchical clustering

In the CSHierClustDirect tab of the ieCS webpage (Figure 3.5), the dendrograms of the DHC method are displayed. The DHC of cell clusters led to the identification of 11 optimal superclusters (Figure 3.18). When a cell type reference was available, the DHC was run in mode A: (Query + Reference) x (Query + Reference), resulting in the

identification of 12 optimal superclusters (Figure 3.19). The DHC method also accepts a user-defined number for superclusters.

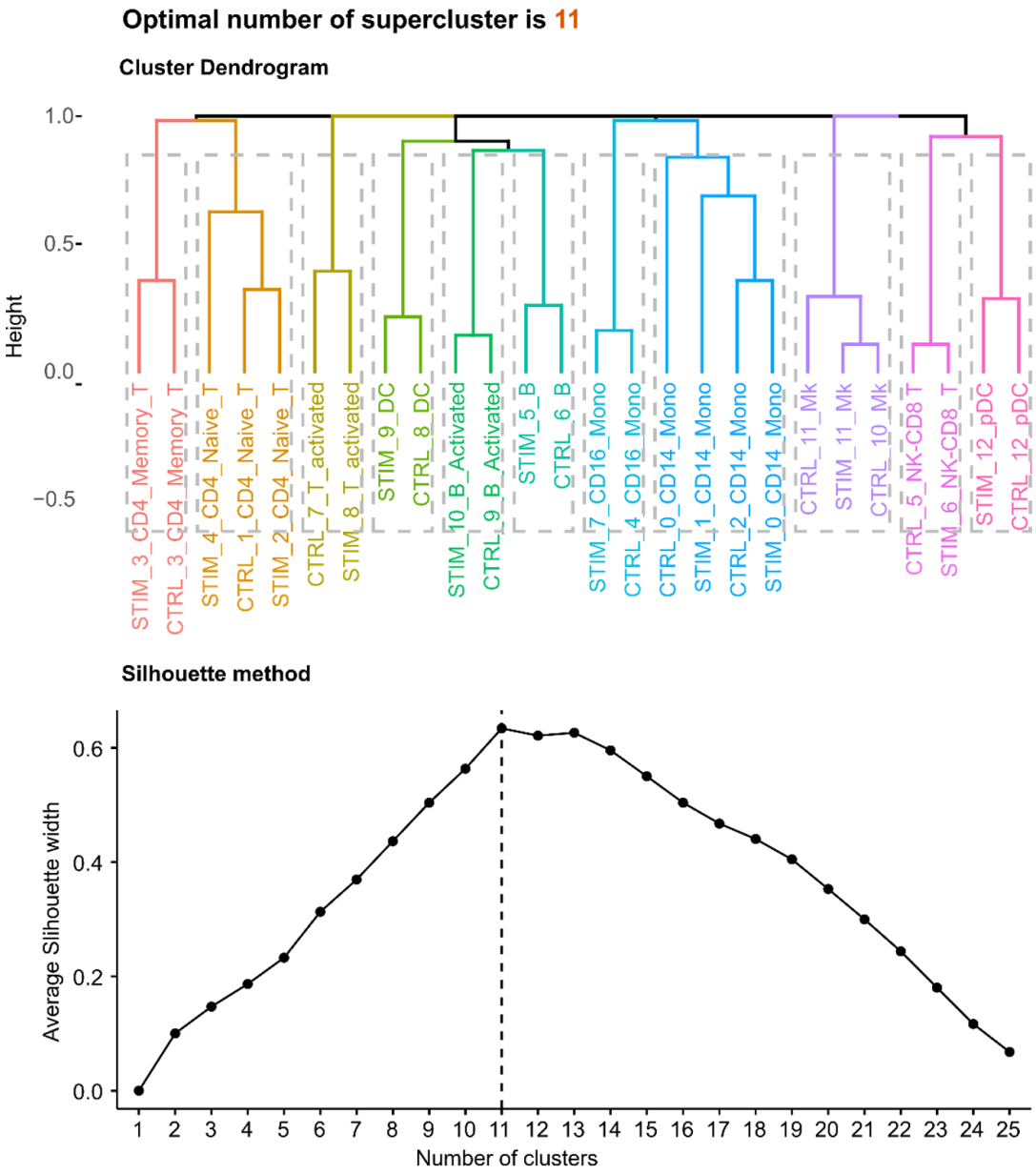


Figure 3.18 **Optimal number of superclusters in DHC.** Top: The dendrograms demonstrate the DHC arrangement of cell clusters. Clusters are color-coded according to the optimal number of superclusters. Bottom: The average Silhouette width for different supercluster counts.

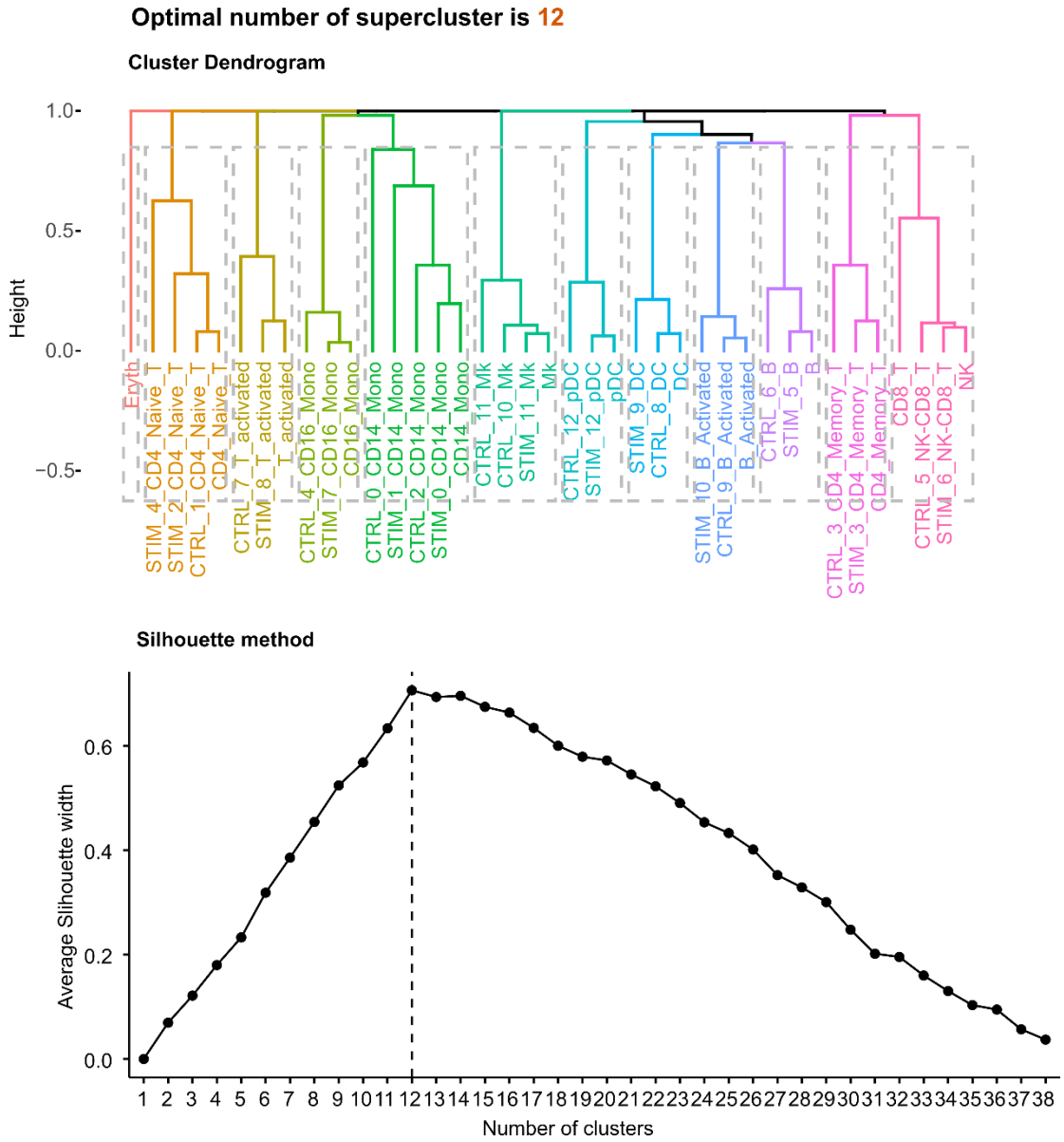


Figure 3.19 **Optimal number of superclusters in DHC with mode A.** Mode A: (Query + Reference) x (Query + Reference). Top: The dendrograms demonstrate the DHC arrangement of cell clusters and cell types. Clusters are color-coded according to the optimal number of superclusters. Bottom: The average Silhouette width for different supercluster counts.

In this subsection, I demonstrated how ieCS applied GHC and DHC to identify superclusters and determined the optimal number of superclusters. Users have the flexibility to specify a custom number of superclusters. For scenarios involving cell type markers as reference, ieCS can effectively assign cell clusters to corresponding cell types based on their similarity scores.

3.3.2 Network partitioning

In the *CSNetwork* tab of the ieCS webpage (Figure 3.6), ieCS constructes a network wherein cell clusters served as nodes, and the similarity scores between two cell clusters represented as an edge connecting these two cell clusters. User can define a cutoff to reconstruct the network of cell clusters. The edges with similarity scores lower than the cutoff will be removed.

Applying a cutoff of 5 resulted in the creation of a network where edges connected all cell clusters (Figure 3.20). Utilizing a network partitioning method - the Louvain algorithm, the network of cell clusters was partitioned into 7 superclusters, each representing a distinct grouping of similar cell clusters (Figure 3.20).

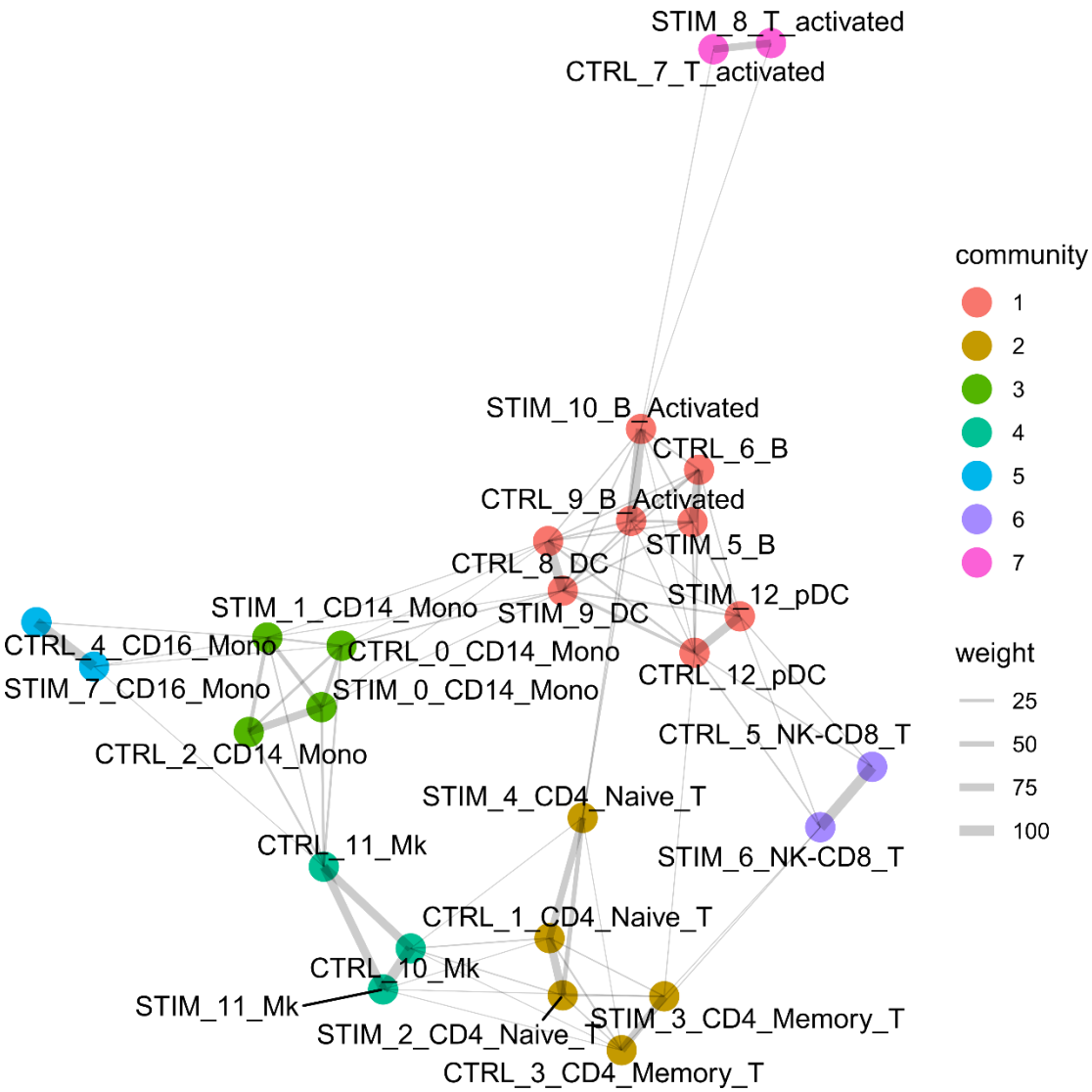


Figure 3.20 **A network of cell clusters at a cutoff of 5.** Each cell cluster is represented as a node, and the similarity scores serve as the weights of the edges

connecting these clusters. The network is enriched with colors that correspond to the superclusters identified through the Louvain algorithm.

By applying a cutoff of 30, I observed a sparser network (Figure 3.21). Employing the Louvain algorithm, ieCS subsequently partitioned the network of cell clusters into 11 distinct superclusters. Notably, these superclusters were composed of cell clusters that share the same cell types across IFNB stimulated and control conditions (Figure 3.21).

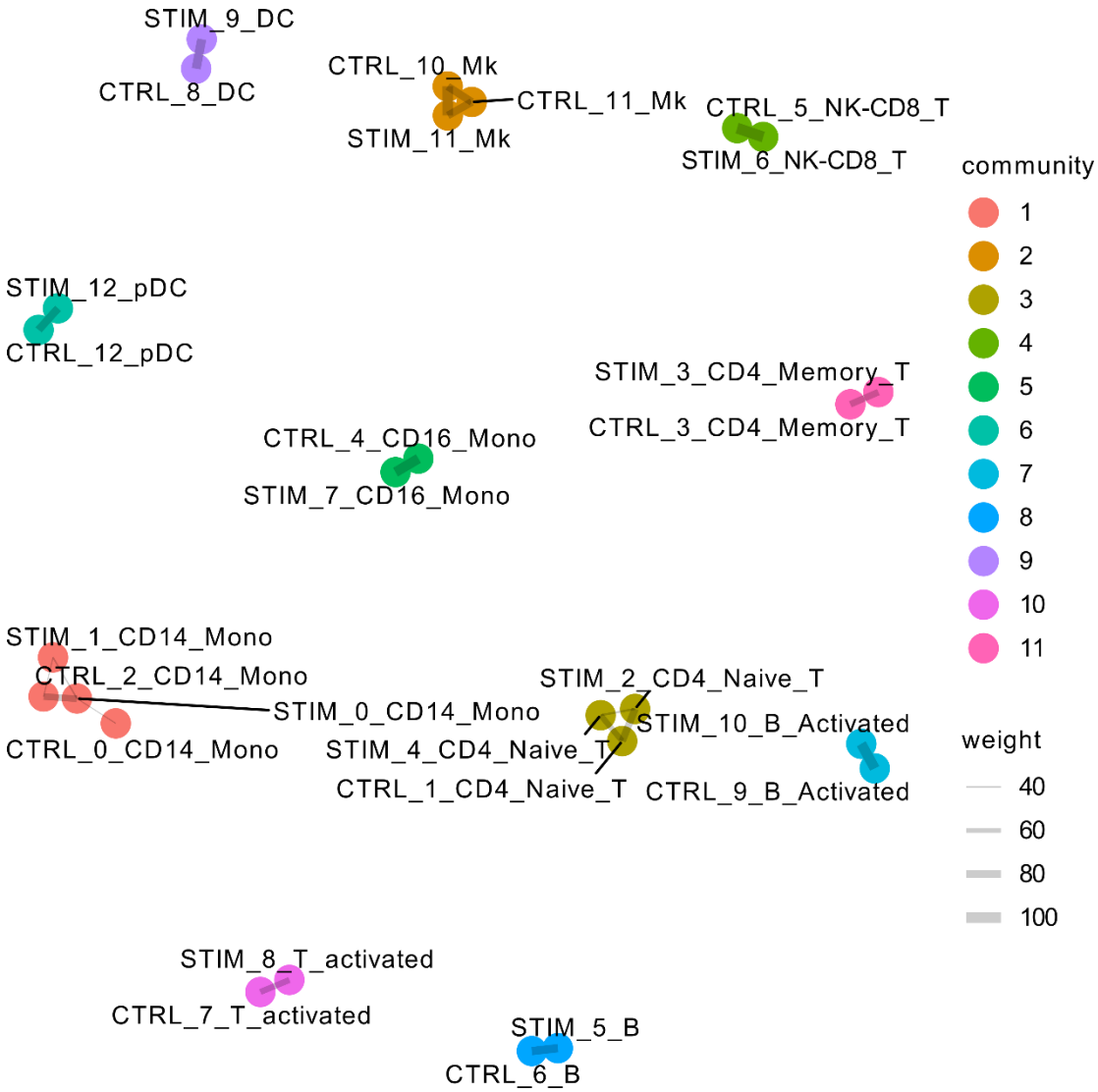


Figure 3.21 **A network of cell clusters at a cutoff of 30.** Each cell cluster is represented as a node, and the similarity scores serve as the weights of the edges connecting these clusters. The network is enriched with colors that correspond to the superclusters identified through the Louvain algorithm.

In situations where cell type markers are available, ieCS constructes a network that encompassed both cell clusters and cell types. At a cutoff of 5, the network was partitioning into 8 superclusters (Figure 3.22).

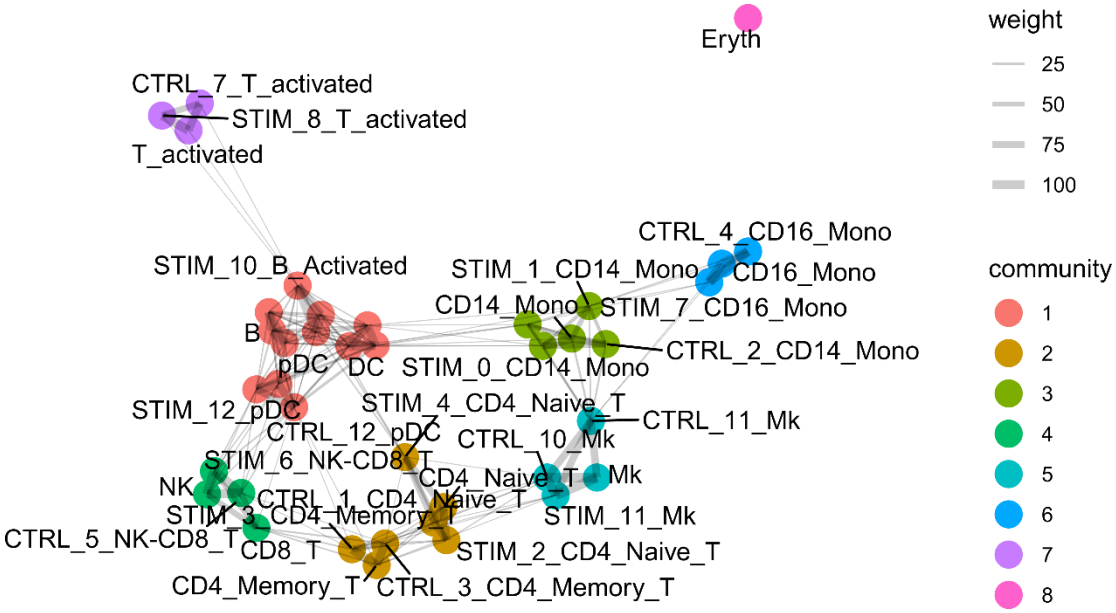


Figure 3.22 **A network of cell clusters and cell types at a cutoff of 5.** Each cell cluster and cell type are represented as a node, and the similarity scores serve as the weights of the edges connecting these clusters. The network is enriched with colors that correspond to the superclusters identified through the Louvain algorithm.

Applying a cutoff of 30, I found 12 distinct superclusters (Figure 3.23). Remarkably, these superclusters consist of cell clusters that share the same cell types as well as corresponding reference cell types (Figure 3.23).

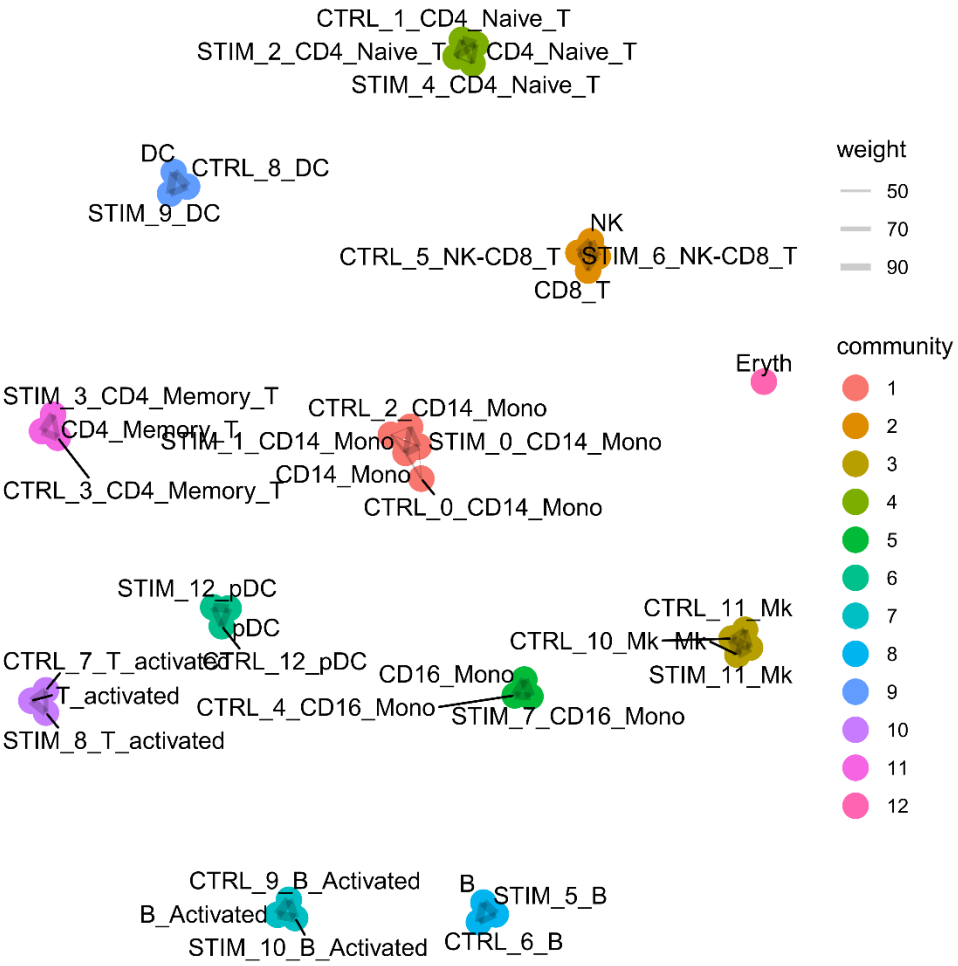


Figure 3.23 **A network of cell clusters and cell types at a cutoff of 30.** Each cell cluster and cell type are represented as a node, and the similarity scores serve as the weights of the edges connecting these clusters. The network is enriched with colors that correspond to the superclusters identified through the Louvain algorithm.

In this subsection, I demonstrated the visualization of cell clusters within a network. Subsequently, I employed a network partitioning technique implemented in ieCS to identify superclusters. Within the demo dataset, these superclusters were formed by cell clusters sharing identical cell types. Furthermore, users retain the flexibility to define a minimum similarity score cutoff for the reconstruction of the network.

3.3.3 Tree aggregation

In the *CSTree* tab of the ieCS webpage (Figure 3.7), ieCS utilizes a tree aggregation

method to visualize and identify superclusters. Cell clusters are served as the leaves, and the edges are indicative of the similarity scores between these cell clusters. Users have the capability to dynamically reconstruct trees using varying cutoffs.

Utilizing a cutoff of 5, a total of seven subtrees (representing superclusters) were constructed (Figure 3.24).

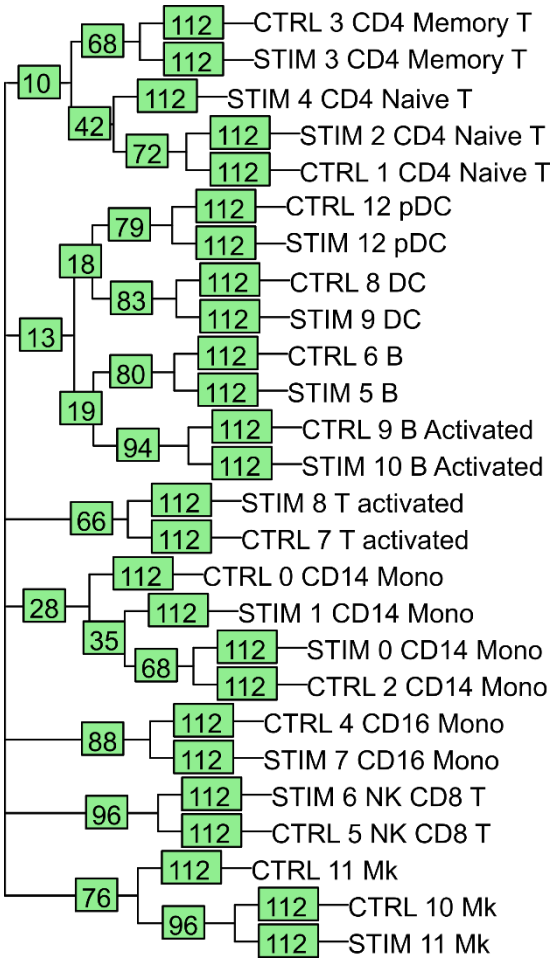


Figure 3.24 **A tree of cell clusters at a cutoff of 5.** Seven superclusters (subtrees) were identified. Cell clusters are depicted as leaves, and the edges between these clusters are determined by their similarity scores. These similarity scores are then labeled on the corresponding edges. The similarity score of 112 represents the similarity score between the cell cluster and itself (the maximum similarity score). This score corresponds to the scenario where the cell clusters had 50 markers, as indicated by the red curve in Figure 3.9b (when $x = 50$, $y \approx 112$).

ieCS offers various tree visualization options. Figure 3.25 illustrates the identification of 11 superclusters at a cutoff of 30, visualized in a "fan" tree style.

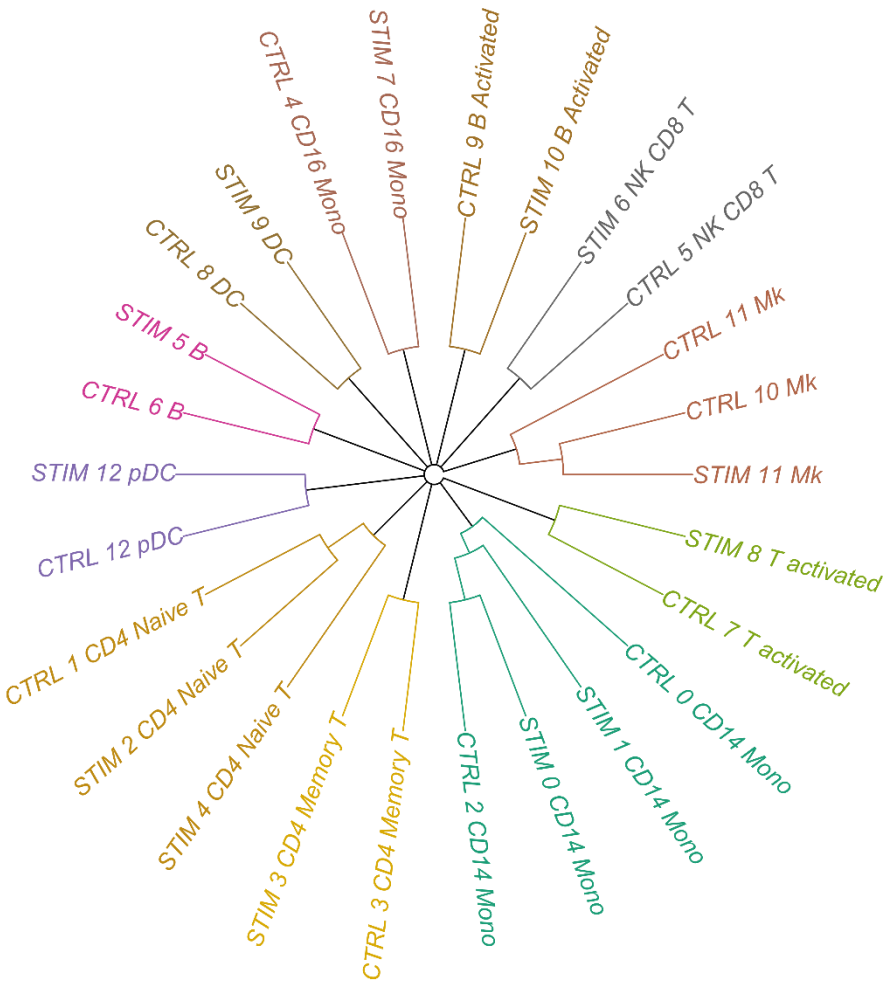


Figure 3.25 **Tree aggregated of cell clusters at a cutoff of 30.** Eleven superclusters (subtrees) have been identified and visualized in a “fan” style. Each subtree is represented by a distinct color.

When reference cell types were accessible, a total of 8 subtrees (representing superclusters) were constructed using a cutoff of 5 (Figure 3.26).

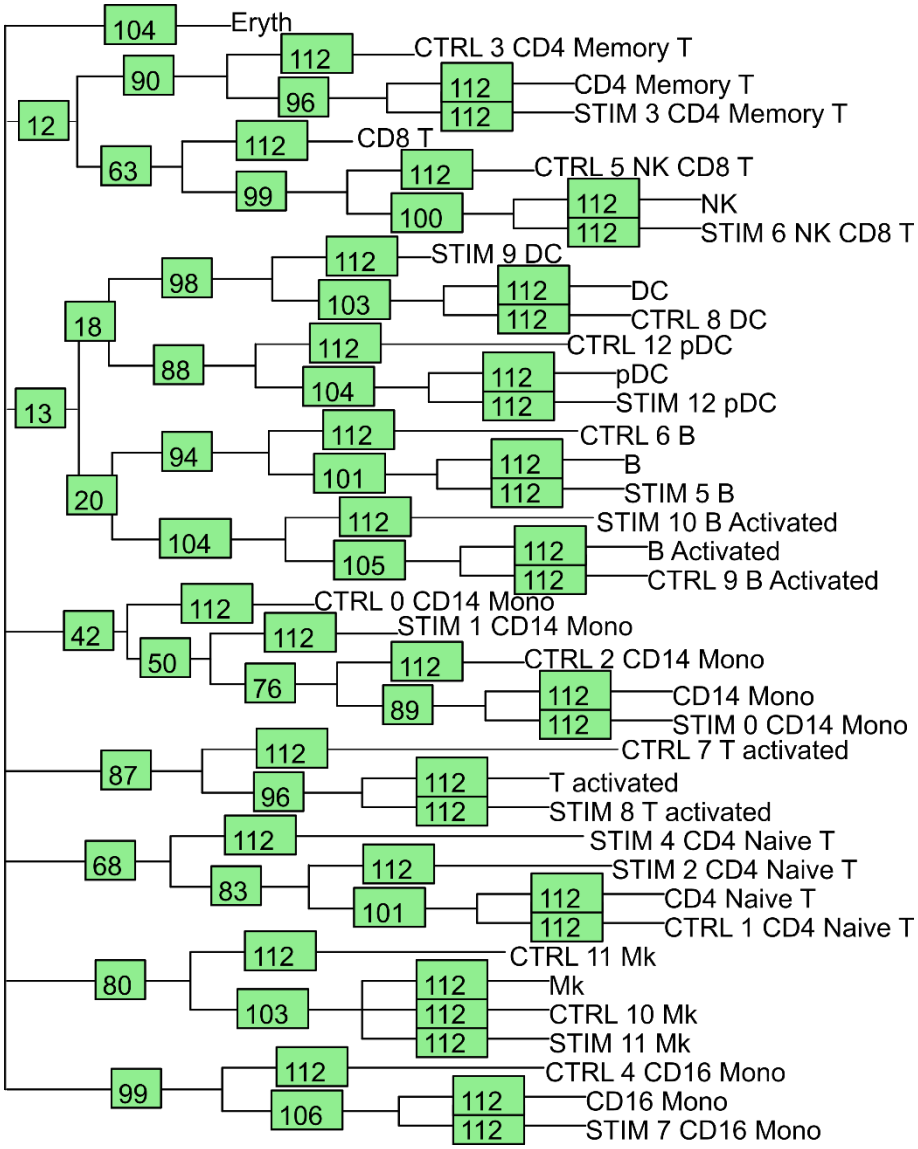


Figure 3.26 **A tree of cell clusters and cell types at a cutoff of 5.** Cell clusters were depicted as leaves, and the edges between these clusters were determined by their similarity scores. These similarity scores were then labeled on the corresponding edges. Eight superclusters (subtrees) were identified.

Figure 3.27 showcases the identification of 12 superclusters at a cutoff of 30, accompanied by cell type references. These are visualized in a "fan" tree style.

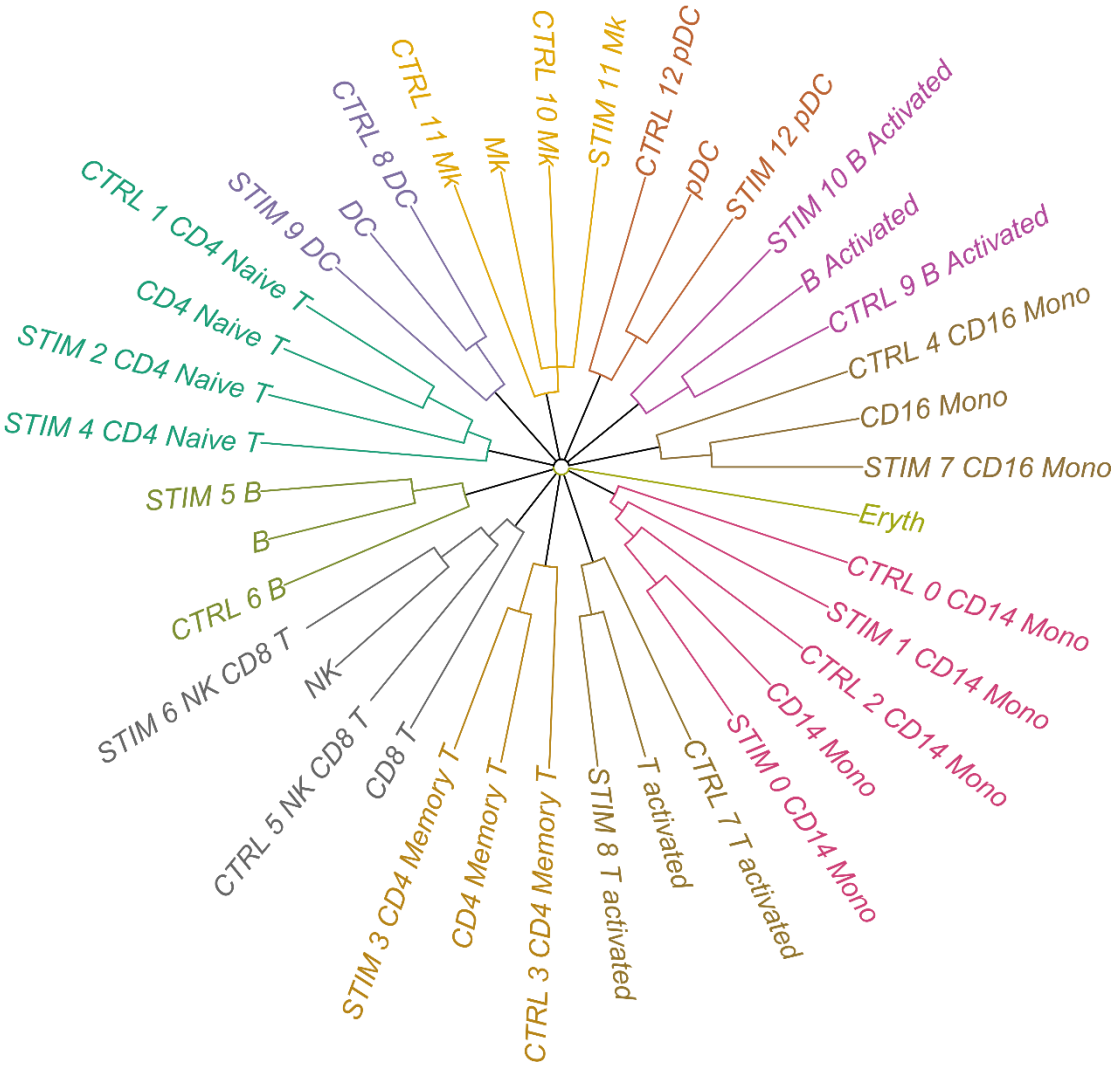


Figure 3.27 **A tree of cell clusters and cell types at a cutoff of 30.** Twelve superclusters (subtrees) have been identified and visualized in a “fan” style. Each subtree is represented by a distinct color.

In this subsection, I introduced a novel tree aggregation method. Superclusters were visualized as subtrees, and users can interactively reconstruct the tree of cell clusters using a defined cutoff for the minimum similarity score.

3.3.4 Cell visualization

In the "CellEmbeddingPlot" section of the ieCS webpage (Figure 3.8), ieCS facilitates the utilization of user-provided cell coordinate files, such as UMAP embeddings (Table 9), for the purpose of visualizing cells. This feature also enables users to provide cell-specific information (Table 10) and hover the annotation over individual cells. Users can apply color coding to cells based on either user-selected cell attributes or the superclusters generated from supercluster identification methods, which can be observed in the cell embedding plots interatively (Figure 3.28, Figure 3.29).

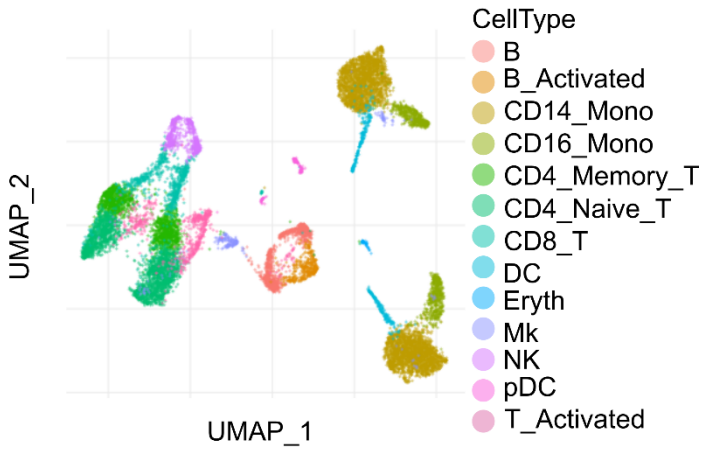


Figure 3.28 **Cell embedding plot colored by cell types.**

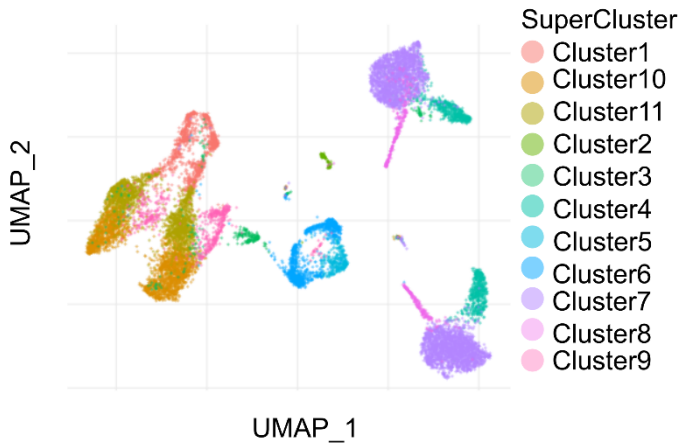


Figure 3.29 **Cell embedding plot colored by superclusters of GHC.**

Within this subsection, I introduced an additional functionality of ieCS focused on the visualization of both individual cells and the outcomes of supercluster analysis.

3.4 Evaluation and Discussion

ieCS is developed in the R language (R Core Team, 2018) using the Shiny package (Chang et al., 2020). It is open-source and available on GitHub (https://github.com/L-Hai/ieCS/). ieCS is packaged as an R package, making it easily distributable and installable. Launching the ieCS GUI requires just one command (ieCS::run()). Users

can upload inputs and explore superclusters interactively. ieCS achieves rapid similarity quantification; for instance, in the demo analysis of scRNA-Seq data consisting of 26 cell clusters (Kang et al., 2018), the process takes only about 10 seconds.

With its user-friendly GUI, ieCS facilitates easy and interactive supercluster identification. Three methods are provided for supercluster identification: Hierarchical Clustering, Network Partitioning, and Tree Aggregation. Users can select these methods in separate GUI tabs. Within each method, users can adjust parameters to explore superclusters with varying degrees of similarity. Increasing the similarity cutoff results in smaller superclusters containing fewer cell clusters, but with higher similarity within the supercluster. ieCS instantly responds to user inputs, displaying results directly on web pages. All three methods allow for user-provided cell type markers to automatically annotate the cell clusters. In cases where no cell type markers are available, the overlapping markers of cell clusters within the same supercluster will be provided to assist in manual supercluster annotation. The similarity matrix is visualized as a heatmap. Superclusters are presented as dendrograms in the hierarchical clustering method, network graphs in the network partitioning method, and tree graphs in the tree aggregation method. Users can also visualize superclusters on a provided cell embedding plot. This variety of methods for supercluster identification and visualization provides comprehensive insights into cell cluster similarity.

To comprehensively evaluate the results of the different supercluster identification methods, I undertook three assessments from varying perspectives.

The first assessment aimed to determine whether the supercluster identification methods could link similar cell clusters (annotated as the same cell type) across IFNB-stimulated and control conditions.

The results in subsection 3.3 Application Results, showed that all three supercluster identification methods can successfully link similar cell clusters across different conditions and identified 11 superclusters. Each supercluster contains cell clusters from the same cell types (Table 12). However, achieving this goal involves different processes for each method. In the hierarchical clustering method, the optimal number of superclusters was determined using the Silhouette method. For the network or tree presentation of cell clusters, a minimum similarity score cutoff of 30 was applied to construct the network or tree. Subsequently, for the network partitioning method, the

optimal number of superclusters was determined using the Louvain community detection algorithm. Regarding the tree aggregation method, the optimal number of superclusters was determined by agglomeratively merging sets of cell clusters.

Table 12 **The cell clusters within the same superclusters at the optimal setting.** Each row in the table represents a specific supercluster, while each column corresponds to a supercluster identification method. Results from the Silhouette method-based optimal number of superclusters in GHC (Figure 3.10) and DHC (Figure 3.17), and a cutoff of 30 in network partioning (Figure 3.21) and tree aggregation (Figure 3.25).

GHC	DHC	Network Partitioning	Tree Aggregation
CTRL_6_B; STIM_5_B	CTRL_6_B; STIM_5_B	CTRL_6_B; STIM_5_B	CTRL_6_B; STIM_5_B
CTRL_9_B_Activated; STIM_10_B_Activated	CTRL_9_B_Activated; STIM_10_B_Activated	CTRL_9_B_Activated; STIM_10_B_Activated	CTRL_9_B_Activated; STIM_10_B_Activated
CTRL_0_CD14_Mono; CTRL_2_CD14_Mono; STIM_0_CD14_Mono; STIM_1_CD14_Mono	CTRL_0_CD14_Mono; CTRL_2_CD14_Mono; STIM_0_CD14_Mono; STIM_1_CD14_Mono	CTRL_0_CD14_Mono; CTRL_2_CD14_Mono; STIM_0_CD14_Mono; STIM_1_CD14_Mono	CTRL_0_CD14_Mono; CTRL_2_CD14_Mono; STIM_0_CD14_Mono; STIM_1_CD14_Mono
CTRL_4_CD16_Mono; STIM_7_CD16_Mono	CTRL_4_CD16_Mono; STIM_7_CD16_Mono	CTRL_4_CD16_Mono; STIM_7_CD16_Mono	CTRL_4_CD16_Mono; STIM_7_CD16_Mono
CTRL_3_CD4_Memory_T; STIM_3_CD4_Memory_T	CTRL_3_CD4_Memory_T; STIM_3_CD4_Memory_T	CTRL_3_CD4_Memory_T; STIM_3_CD4_Memory_T	CTRL_3_CD4_Memory_T; STIM_3_CD4_Memory_T
CTRL_1_CD4_Naive_T; STIM_2_CD4_Naive_T; STIM_4_CD4_Naive_T	CTRL_1_CD4_Naive_T; STIM_2_CD4_Naive_T; STIM_4_CD4_Naive_T	CTRL_1_CD4_Naive_T; STIM_2_CD4_Naive_T; STIM_4_CD4_Naive_T	CTRL_1_CD4_Naive_T; STIM_2_CD4_Naive_T; STIM_4_CD4_Naive_T
CTRL_5_NK-CD8_T; STIM_6_NK-CD8_T	CTRL_5_NK-CD8_T; STIM_6_NK-CD8_T	CTRL_5_NK-CD8_T; STIM_6_NK-CD8_T	CTRL_5_NK-CD8_T; STIM_6_NK-CD8_T
CTRL_8_DC; STIM_9_DC	CTRL_8_DC; STIM_9_DC	CTRL_8_DC; STIM_9_DC	CTRL_8_DC; STIM_9_DC
CTRL_10_Mk; CTRL_11_Mk; STIM_11_Mk	CTRL_10_Mk; CTRL_11_Mk; STIM_11_Mk	CTRL_10_Mk; CTRL_11_Mk; STIM_11_Mk	CTRL_10_Mk; CTRL_11_Mk; STIM_11_Mk
CTRL_12_pDC; STIM_12_pDC	CTRL_12_pDC; STIM_12_pDC	CTRL_12_pDC; STIM_12_pDC	CTRL_12_pDC; STIM_12_pDC
CTRL_7_T_activated; STIM_8_T_activated	CTRL_7_T_activated; STIM_8_T_activated	CTRL_7_T_activated; STIM_8_T_activated	CTRL_7_T_activated; STIM_8_T_activated

Although the optimal outcomes suggest a total of 11 superclusters, users have the option to instruct ieCS to generate a different number of superclusters. For instance, in the GHC method, ieCS produced 5 superclusters (Figure 3.11), each containing a larger number of cell clusters (Figure 3.11). While these cell clusters within the same supercluster might originate from various cell types, they still exhibit some degree of similarity; for example, CD4 Naïve T and CD Memory T were found within the same supercluster (Figure 3.11). Similarly, users retain the ability to adjust the minimum similarity score cutoff to reconstruct the superclusters in both the network partitioning method (Figure 3.20) and the tree aggregation method (Figure 3.24).

By gradually increasing the custom number or cutoff of superclusters, users can observe different groupings of cell clusters. This helps users gain a deeper understanding of the relationships among cell clusters.

I carried out the second assessment to evaluate the performance of ieCS at lower similarity degree (Table 13). At a cutoff of 5, both network partitioning and tree aggregation methods yielded identical outcomes. Across all supercluster identification methods, cell clusters originating from CD4 Naïve T and CD Memory T were grouped within the same supercluster (Table 13). Superclusters encompassing cell clusters from T activated were identified by three methods, except for GHC, where cell clusters from T activated were merged with those from CD14 Mono and CD16 Mono (Table 13). Three superclusters (one containing NK and CD8 T, one containing Mk, and one containing DC, pDC, B, and B Activated) were identified by three methods, excluding the DHC method (Table 13).

Under optimal settings in the first assessment, all methods yielded uniform results in which each supercluster comprised cell clusters sharing the same cell types. However, when tested at lower similarity settings, different supercluster identification methods exhibited some unique characteristics.

Table 13 **The cell clusters within the same superclusters at a lower similarity setting.** Each row in the table represents a specific supercluster, while each column corresponds to a supercluster identification method. A custom of 5 superclusters in GHC and DHC. A minimum similarity score cutoff of 5 in network partitioning and tree aggregation.

GHC	DHC	Network Partitioning	Tree Aggregation
CTRL_1_CD4_Naive_T; STIM_3_CD4_Memory_T; STIM_4_CD4_Naive_T; STIM_2_CD4_Naive_T; CTRL_3_CD4_Memory_T	CTRL_1_CD4_Naive_T; STIM_3_CD4_Memory_T; STIM_4_CD4_Naive_T; STIM_2_CD4_Naive_T; CTRL_3_CD4_Memory_T	CTRL_1_CD4_Naive_T; STIM_3_CD4_Memory_T; STIM_4_CD4_Naive_T; STIM_2_CD4_Naive_T; CTRL_3_CD4_Memory_T	CTRL_1_CD4_Naive_T; STIM_3_CD4_Memory_T; STIM_4_CD4_Naive_T; STIM_2_CD4_Naive_T; CTRL_3_CD4_Memory_T
CTRL_5_NK-CD8_T; STIM_6_NK-CD8_T	CTRL_5_NK-CD8_T; STIM_6_NK-CD8_T; STIM_12_pDC; STIM_11_Mk; CTRL_10_Mk; CTRL_12_pDC; CTRL_11_Mk	CTRL_5_NK-CD8_T; STIM_6_NK-CD8_T	CTRL_5_NK-CD8_T; STIM_6_NK-CD8_T
STIM_11_Mk; CTRL_10_Mk; CTRL_11_Mk		STIM_11_Mk; CTRL_10_Mk; CTRL_11_Mk	STIM_11_Mk; CTRL_10_Mk; CTRL_11_Mk
STIM_12_pDC; CTRL_12_pDC; STIM_10_B_Activated; CTRL_9_B_Activated; STIM_5_B; CTRL_6_B; STIM_9_DC; CTRL_8_DC	STIM_10_B_Activated; CTRL_9_B_Activated; STIM_5_B; CTRL_6_B; STIM_9_DC; CTRL_8_DC	STIM_12_pDC; CTRL_12_pDC; STIM_10_B_Activated; CTRL_9_B_Activated; STIM_5_B; CTRL_6_B; STIM_9_DC; CTRL_8_DC	STIM_12_pDC; CTRL_12_pDC; STIM_10_B_Activated; CTRL_9_B_Activated; STIM_5_B; CTRL_6_B; STIM_9_DC; CTRL_8_DC
	CTRL_7_T_activated; STIM_8_T_activated	CTRL_7_T_activated; STIM_8_T_activated	CTRL_7_T_activated; STIM_8_T_activated
STIM_7_CD16_Mono; CTRL_4_CD16_Mono; CTRL_2_CD14_Mono; CTRL_7_T_activated; STIM_0_CD14_Mono; STIM_1_CD14_Mono; CTRL_0_CD14_Mono; STIM_8_T_activated	STIM_7_CD16_Mono; CTRL_4_CD16_Mono; CTRL_2_CD14_Mono; STIM_0_CD14_Mono; STIM_1_CD14_Mono; CTRL_0_CD14_Mono	CTRL_2_CD14_Mono; STIM_0_CD14_Mono; STIM_1_CD14_Mono; CTRL_0_CD14_Mono	CTRL_2_CD14_Mono; STIM_0_CD14_Mono; STIM_1_CD14_Mono; CTRL_0_CD14_Mono
		STIM_7_CD16_Mono; CTRL_4_CD16_Mono	STIM_7_CD16_Mono; CTRL_4_CD16_Mono

In the scenario that cell type markers were accessible, I applied the third assessment. Under optimal settings, the similar cell clusters within superclusters were consistent across several supercluster identification methods, with the exception of GHC, which grouped Eryth cells into the T_activated superclusters (Table 14). When examining the heatmap of the similarity score matrix of cell clusters and cell types, Eryth displayed little similarity to the other cell clusters and types (Figure 3.13). The dendrogram of GHC (Figures 3.12 and 3.15) suggests that when a number of 12 superclusters is set, Eryth will become a standalone supercluster. However, the Silhouette method failed to recommend this number of superclusters in GHC.

Table 14 **The cell clusters and cell types within the same superclusters at the optimal settings.** Each row in the table represents a specific supercluster, while each column corresponds to a supercluster identification method. Results from the Silhouette method-based optimal number of superclusters in GHC (Figure 3.16) and DHC (Figure 3.19) methods, and a cutoff of 30 in network partioning (Figure 3.23) and

tree aggregation (Figure 3.27).

GHC	DHC	Network Partitioning	Tree Aggregation
B; CTRL_6_B; STIM_5_B	B; CTRL_6_B; STIM_5_B	B; CTRL_6_B; STIM_5_B	B; CTRL_6_B; STIM_5_B
B_Activated; CTRL_9_B_Activated; STIM_10_B_Activated	B_Activated; CTRL_9_B_Activated; STIM_10_B_Activated	B_Activated; CTRL_9_B_Activated; STIM_10_B_Activated	B_Activated; CTRL_9_B_Activated; STIM_10_B_Activated
CD14_Mono; CTRL_0_CD14_Mono; CTRL_2_CD14_Mono; STIM_0_CD14_Mono; STIM_1_CD14_Mono	CD14_Mono; CTRL_0_CD14_Mono; CTRL_2_CD14_Mono; STIM_0_CD14_Mono; STIM_1_CD14_Mono	CD14_Mono; CTRL_0_CD14_Mono; CTRL_2_CD14_Mono; STIM_0_CD14_Mono; STIM_1_CD14_Mono	CD14_Mono; CTRL_0_CD14_Mono; CTRL_2_CD14_Mono; STIM_0_CD14_Mono; STIM_1_CD14_Mono
CD16_Mono; CTRL_4_CD16_Mono; STIM_7_CD16_Mono	CD16_Mono; CTRL_4_CD16_Mono; STIM_7_CD16_Mono	CD16_Mono; CTRL_4_CD16_Mono; STIM_7_CD16_Mono	CD16_Mono; CTRL_4_CD16_Mono; STIM_7_CD16_Mono
CD4_Memory_T; CTRL_3_CD4_Memory_T; STIM_3_CD4_Memory_T	CD4_Memory_T; CTRL_3_CD4_Memory_T; STIM_3_CD4_Memory_T	CD4_Memory_T; CTRL_3_CD4_Memory_T; STIM_3_CD4_Memory_T	CD4_Memory_T; CTRL_3_CD4_Memory_T; STIM_3_CD4_Memory_T
CD4_Naive_T; CTRL_1_CD4_Naive_T; STIM_2_CD4_Naive_T; STIM_4_CD4_Naive_T	CD4_Naive_T; CTRL_1_CD4_Naive_T; STIM_2_CD4_Naive_T; STIM_4_CD4_Naive_T	CD4_Naive_T; CTRL_1_CD4_Naive_T; STIM_2_CD4_Naive_T; STIM_4_CD4_Naive_T	CD4_Naive_T; CTRL_1_CD4_Naive_T; STIM_2_CD4_Naive_T; STIM_4_CD4_Naive_T
CD8_T; NK; STIM_6_NK-CD8_T; CTRL_5_NK-CD8_T	CD8_T; NK; STIM_6_NK-CD8_T; CTRL_5_NK-CD8_T	CD8_T; NK; STIM_6_NK-CD8_T; CTRL_5_NK-CD8_T	CD8_T; NK; STIM_6_NK-CD8_T; CTRL_5_NK-CD8_T
DC; CTRL_8_DC; STIM_9_DC	DC; CTRL_8_DC; STIM_9_DC	DC; CTRL_8_DC; STIM_9_DC	DC; CTRL_8_DC; STIM_9_DC
Mk; CTRL_10_Mk; CTRL_11_Mk; STIM_11_Mk	Mk; CTRL_10_Mk; CTRL_11_Mk; STIM_11_Mk	Mk; CTRL_10_Mk; CTRL_11_Mk; STIM_11_Mk	Mk; CTRL_10_Mk; CTRL_11_Mk; STIM_11_Mk
pDC; CTRL_12_pDC; STIM_12_pDC	pDC; CTRL_12_pDC; STIM_12_pDC	pDC; CTRL_12_pDC; STIM_12_pDC	pDC; CTRL_12_pDC; STIM_12_pDC
Eryth; T_activated; CTRL_7_T_activated; STIM_8_T_activated	T_activated; CTRL_7_T_activated; STIM_8_T_activated	T_activated; CTRL_7_T_activated; STIM_8_T_activated	T_activated; CTRL_7_T_activated; STIM_8_T_activated
	Eryth	Eryth	Eryth

In summary, the three assessments indicate that, in analyses of demo scRNA-Seq data (Kang et al., 2018), all three supercluster identification methods successfully identified cell clusters that shared the same or similar biological cell types.

To achieve this, ieCS employs a novel metric to quantify the similarity between two cell clusters. It requires markers of cell clusters and their ranks as inputs, with the rank of a marker indicating its importance. This similarity metric considers not only the number of shared markers between cell clusters but also the ranks of those markers. When two cell clusters share numerous high-ranked markers, they are deemed highly similar.

While Gao et al. previously quantified similarity through marker expression binarization, yielding significant results (Gao et al., 2019), this approach doesn't consider marker importance. In contrast, ieCS allows users to define marker importance by utilizing metrics like p-values, fold-changes, or other criteria during marker identification. Moreover, ieCS provides the capability for users to upload cell type markers as a reference for annotating cell clusters, streamlining the calculation of similarity between cell clusters and cell types.

3.5 Limits and Outlook

In the application of ieCS for quantifying cell population similarity in heterogeneous scRNA-Seq datasets, there are several aspects that could be enhanced to make ieCS more powerful in the future:

- ieCS introduces a novel metric for quantifying the similarity between cell subpopulations, relying on user-provided cell cluster markers and marker importance ranks. The accuracy of the clustering of cell and marker identification significantly influences the similarity quantification. Integrating reliable marker identification methods into ieCS could enhance the precision of similarity calculations.
- ieCS offers three supercluster identification methods and empowers users to select methods and set parameters, enabling the interactive construction of superclusters at varying levels of similarity. To further improve user experience, integrating automated parameter tuning methods that determine the optimal supercluster compositions could enhance ieCS's efficiency.
- ieCS accommodates user-provided cell type markers as a reference for annotating superclusters. Integrating a cell type database into ieCS to facilitate automated supercluster annotation could greatly enhance its functionality.

4. Own Publications

Publications related to this thesis

Hai, L.*, Hoffmann, D.C.*, Mandelbaum, H., Xie, R., Ito, J., Jung, E., Weil, S., Sievers, P., Venkataramani, V., Azorin, D.D., Ernst, K., Reibold, D., Will, R., Suvà, M.L., Herold-Mende, C., Sahm, F., Winkler, F., Schlesner, M., Wick, W.,Kessler, T. 2021. A connectivity signature for glioblastoma. **bioRxiv**: 2021.11.07.465791. https://doi.org/10.1101/2021.11.07.465791. *: co-first authors.

Hai, L., Kessler, T., Wick, W., Schlesner, M. ieCS: interactive explorer of single cell cluster similarity. **In Preparation**.

Hausmann, D., Hoffmann, D.C., Venkataramani, V., Jung, E., Horschitz, S., Tetzlaff, S.K., Jabali, A., **Hai, L.**, Kessler, T., Azorin, D.D., Weil, S., Kourtesakis, A., Sievers, P., Habel, A., Breckwoldt, M.O., Karreman, M.A., Ratliff, M., Messmer, J.M., Yang, Y., Reyhan, E., Wendler, S., Lob, C., Mayer, C., Figarella, K., Osswald, M., Solecki, G., Sahm, F., Garaschuk, O., Kuner, T., Koch, P., Schlesner, M., Wick, W.,Winkler, F. 2023. Autonomous rhythmic activity in glioma networks drives brain tumour growth. **Nature** 613 (7942): 179-186. https://doi.org/10.1038/s41586-022-05520-4.

Xie, R., Kessler, T., Grosch, J., **Hai, L.**, Venkataramani, V., Huang, L., Hoffmann, D.C., Solecki, G., Ratliff, M., Schlesner, M., Wick, W., Winkler, F. 2021. Tumor cell network integration in glioma represents a stemness feature. **Neuro Oncol** 23 (5): 757-769. https://doi.org/10.1093/neuonc/noaa275.

Other publications during my PhD study

Alhalabi, O.T., Fletcher, M.N.C., Hielscher, T., Kessler, T., Lokumcu, T., Baumgartner, U., Wittmann, E., Schlue, S., Gottmann, M., Rahman, S., **Hai, L.**, Hansen-Palmus, L., Puccio, L., Nakano, I., Herold-Mende, C., Day, B.W., Wick, W., Sahm, F., Phillips, E.,Goidts, V. 2022. A novel patient stratification strategy to enhance the therapeutic efficacy of dasatinib in glioblastoma. **Neuro Oncol** 24 (1): 39-51. https://doi.org/10.1093/neuonc/noab158.

Gengenbacher, N., Singhal, M., Mogler, C., **Hai, L.**, Milde, L., Pari, A.A.A., Besemfelder, E., Fricke, C., Baumann, D., Gehrs, S., Utikal, J., Felcht, M., Hu, J., Schlesner, M., Offringa, R., Chintharlapalli, S.R.,Augustin, H.G. 2021. Timed Ang2-Targeted Therapy Identifies the Angiopoietin-Tie Pathway as Key Regulator of Fatal Lymphogenous Metastasis. **Cancer Discov** 11 (2): 424-445. https://doi.org/10.1158/2159-8290.CD-20-0122.

Kaulen, L.D., Denisova, E., Hinz, F., **Hai, L.**, Friedel, D., Henegariu, O., Hoffmann, D.C., Ito, J., Kourtesakis, A., Lehnert, P., Doubrovinskaia, S., Karschnia, P., von Baumgarten, L., Kessler, T., Baehring, J.M., Brors, B., Sahm, F.,Wick, W. 2023. Integrated genetic analyses of immunodeficiency-associated Epstein-Barr virus- (EBV) positive primary CNS lymphomas. **Acta Neuropathol** 146 (3): 499-514. https://doi.org/10.1007/s00401-023-02613-w.

Kessler, T., Berberich, A., Casalini, B., Druschler, K., Ostermann, H., Dormann, A., Walter, S., **Hai, L.**, Schlesner, M., Herold-Mende, C., Jungk, C., Unterberg, A., Bendszus, M., Sahm, K., von Deimling, A., Winkler, F., Platten, M., Wick, W., Sahm, F.,Wick, A. 2020. Molecular profiling-based decision for targeted therapies in IDH wild-type glioblastoma. **Neurooncol Adv** 2 (1): vdz060. https://doi.org/10.1093/noajnl/vdz060.

Kessler, T., Schrimpf, D., Doerner, L., **Hai, L.**, Kaulen, L.D., Ito, J., van den Bent, M., Taphoorn, M., Brandes, A.A., Idbaih, A., Domont, J., Clement, P.M., Campone, M., Bendszus, M., von Deimling, A., Sahm, F., Platten, M., Wick, W.,Wick, A. 2023. Prognostic markers of DNA methylation and NGS sequencing in

progressive glioblastoma from the EORTC-26101 trial. **Clin Cancer Res.** https://doi.org/10.1158/1078-0432.CCR-23-0926.

Ratliff, M., Karimian-Jazi, K., Hoffmann, D.C., Rauschenbach, L., Simon, M., **Hai, L.**, Mandelbaum, H., Schubert, M.C., Kessler, T., Uhlig, S., Azorin, D.D., Jung, E., Osswald, M., Solecki, G., Maros, M.E., Venkataramani, V., Glas, M., Etminan, N., Scheffler, B., Wick, W., Winkler, F. 2023. Individual glioblastoma cells harbor both proliferative and invasive capabilities during tumor progression. **Neuro Oncol**. https://doi.org/10.1093/neuonc/noad109.

Singhal, M., Gengenbacher, N., Abdul Pari, A.A., Kamiyama, M., **Hai, L.**, Kuhn, B.J., Kallenberg, D.M., Kulkarni, S.R., Camilli, C., Preuss, S.F., Leuchs, B., Mogler, C., Espinet, E., Besemfelder, E., Heide, D., Heikenwalder, M., Sprick, M.R., Trumpp, A., Krijgsveld, J., Schlesner, M., Hu, J., Moss, S.E., Greenwood, J.,Augustin, H.G. 2021. Temporal multi-omics identifies LRG1 as a vascular niche instructor of metastasis. **Sci Transl Med** 13 (609): eabe6805. https://doi.org/10.1126/scitranslmed.abe6805.

Venkataramani, V., Yang, Y., Schubert, M.C., Reyhan, E., Tetzlaff, S.K., Wissmann, N., Botz, M., Soyka, S.J., Beretta, C.A., Pramatarov, R.L., Fankhauser, L., Garofano, L., Freudenberg, A., Wagner, J., Tanev, D.I., Ratliff, M., Xie, R., Kessler, T., Hoffmann, D.C., **Hai, L.**, Dorflinger, Y., Hoppe, S., Yabo, Y.A., Golebiewska, A., Niclou, S.P., Sahm, F., Lasorella, A., Slowik, M., Doring, L., Iavarone, A., Wick, W., Kuner, T.,Winkler, F. 2022. Glioblastoma hijacks neuronal mechanisms for brain invasion. **Cell** 185 (16): 2899-2917 e31. https://doi.org/10.1016/j.cell.2022.06.054.

5. References

10x Genomics, 2020. Chromium Single Cell 3' Reagent Kits User Guide (v3 Chemistry) -User Guide -Library Prep -Single Cell Gene Expression -Official 10x Genomics Support [WWW Document]. URL https://support.10xgenomics.com/single-cell-gene-expression/library-prep/doc/user-guide-chromium-single-cell-3-reagent-kits-user-guide-v3-chemistry (accessed 3.31.21).

Abdelaal, T., Michielsen, L., Cats, D., Hoogduin, D., Mei, H., Reinders, M.J.T., Mahfouz, A., 2019. A comparison of automatic cell identification methods for single-cell RNA sequencing data. Genome Biol 20, 194. https://doi.org/10.1186/s13059-019-1795-z

Ahluwalia, M.S., de Groot, J., Liu, W.M.,Gladson, C.L. 2010. Targeting SRC in glioblastoma tumors and brain metastases: rationale and preclinical studies. Cancer Lett 298 (2): 139-49. https://doi.org/10.1016/j.canlet.2010.08.014.

Ajaib, S., Lodha, D., Pollock, S., Hemmings, G., Finetti, M.A., Gusnanto, A., Chakrabarty, A., Ismail, A., Wilson, E., Varn, F.S., Hunter, B., Filby, A., Brockman, A.A., McDonald, D., Verhaak, R.G.W., Ihrie, R.A., Stead, L.F. 2023. GBMdeconvoluteR accurately infers proportions of neoplastic and immune cell populations from bulk glioblastoma transcriptomics data. Neuro Oncol 25 (7): 1236-1248. https://doi.org/10.1093/neuonc/noad021.

Anders, S., Pyl, P.T.,Huber, W. 2015. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics 31 (2): 166-9. https://doi.org/10.1093/bioinformatics/btu638.

Andrews, S., others, 2010. FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom.

Aran, D., Looney, A.P., Liu, L., Wu, E., Fong, V., Hsu, A., Chak, S., Naikawadi, R.P., Wolters, P.J., Abate, A.R., Butte, A.J., Bhattacharya, M., 2019. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nature Immunology 20, 163–172. https://doi.org/10.1038/s41590-018-0276-y

Auguie, B., 2017. gridExtra: Miscellaneous Functions for "Grid" Graphics.

Behnan, J., Finocchiaro, G., Hanna, G., 2019. The landscape of the mesenchymal signature in brain tumours. Brain 142, 847–866. https://doi.org/10.1093/brain/awz044

Bergen, V., Lange, M., Peidli, S., Wolf, F.A.,Theis, F.J. 2020. Generalizing RNA velocity to transient cell states through dynamical modeling. Nat Biotechnol 38 (12): 1408-1414. https://doi.org/10.1038/s41587-020-0591-3.

Billatos, E., Vick, J.L., Lenburg, M.E., Spira, A.E., 2018. The Airway Transcriptome as a Biomarker for Early Lung Cancer Detection. Clin Cancer Res 24, 2984–2992. https://doi.org/10.1158/1078-0432.CCR-16-3187

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

Boregowda, S.V., Krishnappa, V., Strivelli, J., Haga, C.L., Booker, C.N., Phinney, D.G., 2018. Basal p53 expression is indispensable for mesenchymal stem cell integrity. Cell Death & Differentiation 25, 679–692. https://doi.org/10.1038/s41418-017-0004-4

Breschi, A., Gingeras, T.R., Guigó, R., 2017. Comparative transcriptomics in human and mouse. Nature Reviews Genetics 18, 425–440. https://doi.org/10.1038/nrg.2017.19

Cakir, B., Prete, M., Huang, N., van Dongen, S., Pir, P., Kiselev, V.Y., 2020. Comparison of visualization tools for single-cell RNAseq data. NAR Genomics and Bioinformatics 2. https://doi.org/10.1093/nargab/lqaa052

Capper, D., Jones, D.T.W., Sill, M., Hovestadt, V., Schrimpf, D., Sturm, D., Koelsche, C., Sahm, F., Chavez, L., Reuss, D.E., Kratz, A., Wefers, A.K., Huang, K., Pajtler, K.W., Schweizer, L., Stichel, D., Olar, A., Engel, N.W., Lindenberg, K., Harter, P.N., Braczynski, A.K., Plate, K.H., Dohmen, H., Garvalov, B.K., Coras, R., Hölsken, A., Hewer, E., Bewerunge-Hudler, M., Schick, M., Fischer, R., Beschorner, R., Schittenhelm, J., Staszewski, O., Wani, K., Varlet, P., Pages, M., Temming, P.,

Lohmann, D., Selt, F., Witt, H., Milde, T., Witt, O., Aronica, E., Giangaspero, F., Rushing, E., Scheurlen, W., Geisenberger, C., Rodriguez, F.J., Becker, A., Preusser, M., Haberler, C., Bjerkvig, R., Cryan, J., Farrell, M., Deckert, M., Hench, J., Frank, S., Serrano, J., Kannan, K., Tsirigos, A., Brück, W., Hofer, S., Brehmer, S., Seiz-Rosenhagen, M., Hänggi, D., Hans, V., Rozsnoki, S., Hansford, J.R., Kohlhof, P., Kristensen, B.W., Lechner, M., Lopes, B., Mawrin, C., Ketter, R., Kulozik, A., Khatib, Z., Heppner, F., Koch, A., Jouvet, A., Keohane, C., Mühleisen, H., Mueller, W., Pohl, U., Prinz, M., Benner, A., Zapatka, M., Gottardo, N.G., Driever, P.H., Kramm, C.M., Müller, H.L., Rutkowski, S., von Hoff, K., Frühwald, M.C., Gnekow, A., Fleischhack, G., Tippelt, S., Calaminus, G., Monoranu, C.-M., Perry, A., Jones, C., Jacques, T.S., Radlwimmer, B., Gessi, M., Pietsch, T., Schramm, J., Schackert, G., Westphal, M., Reifenberger, G., Wesseling, P., Weller, M., Collins, V.P., Blümcke, I., Bendszus, M., Debus, J., Huang, A., Jabado, N., Northcott, P.A., Paulus, W., Gajjar, A., Robinson, G.W., Taylor, M.D., Jaunmuktane, Z., Ryzhova, M., Platten, M., Unterberg, A., Wick, W., Karajannis, M.A., Mittelbronn, M., Acker, T., Hartmann, C., Aldape, K., Schüller, U., Buslei, R., Lichter, P., Kool, M., Herold-Mende, C., Ellison, D.W., Hasselblatt, M., Snuderl, M., Brandner, S., Korshunov, A., von Deimling, A., Pfister, S.M., 2018. DNA methylation-based classification of central nervous system tumours. Nature 555, 469–474. https://doi.org/10.1038/nature26000

Chang, W., Cheng, J., Allaire, J.J., Xie, Y., McPherson, J., 2020. shiny: Web Application Framework for R.

Chen, S., Lake, B.B., Zhang, K., 2019. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. Nature Biotechnology 37, 1452–1457. https://doi.org/10.1038/s41587-019-0290-0

Chen, W., Zhao, Y., Chen, X., Yang, Z., Xu, X., Bi, Y., Chen, V., Li, J., Choi, H., Ernest, B., Tran, B., Mehta, M., Kumar, P., Farmer, A., Mir, A., Mehra, U.A., Li, J.-L., Moos, M., Xiao, W., Wang, C., 2020. A multicenter study benchmarking single-cell RNA sequencing technologies using reference samples. Nature Biotechnology 1–12. https://doi.org/10.1038/s41587-020-00748-9

Cheng, C., Alexander, R., Min, R., Leng, J., Yip, K.Y., Rozowsky, J., Yan, K.-K., Dong, X., Djebali, S., Ruan, Y., Davis, C.A., Carninci, P., Lassman, T., Gingeras, T.R., Guigó, R., Birney, E., Weng, Z., Snyder, M., Gerstein, M., 2012. Understanding transcriptional regulation by integrative analysis of transcription factor binding data. Genome Res. 22, 1658–1667. https://doi.org/10.1101/gr.136838.111

Chu, L.-F., Leng, N., Zhang, J., Hou, Z., Mamott, D., Vereide, D.T., Choi, J., Kendziorski, C., Stewart, R., Thomson, J.A., 2016. Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. Genome Biology 17, 173. https://doi.org/10.1186/s13059-016-1033-x

Chu, M.-S., Chang, C.-F., Yang, C.-C., Bau, Y.-C., Ho, L.L.-T., Hung, S.-C., 2006. Signalling pathway in the induction of neurite outgrowth in human mesenchymal stem cells. Cell Signal 18, 519–530. https://doi.org/10.1016/j.cellsig.2005.05.018

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M.W., Gaffney, D.J., Elo, L.L., Zhang, X., Mortazavi, A., 2016. A survey of best practices for RNA-seq data analysis. Genome Biology 17, 13. https://doi.org/10.1186/s13059-016-0881-8

Cox, J.,Mann, M. 2008. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol 26 (12): 1367-72. https://doi.org/10.1038/nbt.1511.

Couturier, C.P., Ayyadhury, S., Le, P.U., Nadaf, J., Monlong, J., Riva, G., Allache, R., Baig, S., Yan, X., Bourgey, M., Lee, C., Wang, Y.C.D., Wee Yong, V., Guiot, M.-C., Najafabadi, H., Misic, B., Antel, J., Bourque, G., Ragoussis, J., Petrecca, K., 2020. Single-cell RNA-seq reveals that glioblastoma recapitulates a normal neurodevelopmental hierarchy. Nature Communications 11, 3406. https://doi.org/10.1038/s41467-020-17186-5

Csardi, G., Nepusz, T., 2006. The igraph software package for complex network research.

InterJournal Complex Systems, 1695.
Cui, J., Cui, C., Cui, Y., Li, R., Sheng, H., Jiang, X., Tian, Y., Wang, K., Gao, J., 2017. Bone Marrow Mesenchymal Stem Cell Transplantation Increases GAP-43 Expression via ERK1/2 and PI3K/Akt Pathways in Intracerebral Hemorrhage. CPB 42, 137–144. https://doi.org/10.1159/000477122
Cuomo, A.S.E., Seaton, D.D., McCarthy, D.J., Martinez, I., Bonder, M.J., Garcia-Bernardo, J., Amatya, S., Madrigal, P., Isaacson, A., Buettner, F., Knights, A., Natarajan, K.N., Vallier, L., Marioni, J.C., Chhatriwala, M., Stegle, O., 2020. Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression. Nature Communications 11, 810. https://doi.org/10.1038/s41467-020-14457-z
Dagogo-Jack, I., Shaw, A.T., 2018. Tumour heterogeneity and resistance to cancer therapies. Nature Reviews Clinical Oncology 15, 81–94. https://doi.org/10.1038/nrclinonc.2017.166
Darmanis, S., Sloan, S.A., Zhang, Y., Enge, M., Caneda, C., Shuer, L.M., Gephart, M.G.H., Barres, B.A., Quake, S.R., 2015. A survey of human brain transcriptome diversity at the single cell level. PNAS 112, 7285–7290. https://doi.org/10.1073/pnas.1507125112
Deemyad, T., Lüthi, J., Spruston, N., 2018. Astrocytes integrate and drive action potential firing in inhibitory subnetworks. Nature Communications 9, 4336. https://doi.org/10.1038/s41467-018-06338-3
Dey, N., Crosswell, H.E., De, P., Parsons, R., Peng, Q., Su, J.D.,Durden, D.L. 2008. The protein phosphatase activity of PTEN regulates SRC family kinases and controls glioma migration. Cancer Res 68 (6): 1862-71. https://doi.org/10.1158/0008-5472.CAN-07-1182.
Dilger, N., Neehus, A.-L., Grieger, K., Hoffmann, A., Menssen, M., Ngezahayo, A., 2020. Gap Junction Dependent Cell Communication Is Modulated During Transdifferentiation of Mesenchymal Stem/Stromal Cells Towards Neuron-Like Cells. Front. Cell Dev. Biol. 8. https://doi.org/10.3389/fcell.2020.00869
Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R., 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21. https://doi.org/10.1093/bioinformatics/bts635
Duò, A., Robinson, M.D., Soneson, C., 2020. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. F1000Res 7, 1141. https://doi.org/10.12688/f1000research.15666.3
Fadhlullah, S.F.B., Halim, N.B.A., Yeo, J.Y.T., Ho, R.L.Y., Um, P., Ang, B.T., Tang, C., Ng, W.H., Virshup, D.M., Ho, I.A.W., 2019. Pathogenic mutations in neurofibromin identifies a leucine-rich domain regulating glioma cell invasiveness. Oncogene 38, 5367–5380. https://doi.org/10.1038/s41388-019-0809-3
Fields, R.D., Stevens-Graham, B., 2002. New Insights into Neuron-Glia Communication. Science 298, 556–562. https://doi.org/10.1126/science.298.5593.556
Fish, J.L., Schneider, R.A., 2014. Chapter 6 - Neural Crest-Mediated Tissue Interactions During Craniofacial Development: The Origins of Species-Specific Pattern, in: Trainor, P.A. (Ed.), Neural Crest Cells. Academic Press, Boston, pp. 101–124. https://doi.org/10.1016/B978-0-12-401730-6.00007-7
Galili, T., 2015. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. Bioinformatics 31, 3718–3720. https://doi.org/10.1093/bioinformatics/btv428
Gao, X., Hu, D., Gogol, M., Li, H., 2019. ClusterMap: compare multiple single cell RNA-Seq datasets across different experimental conditions. Bioinformatics 35, 3038–3045. https://doi.org/10.1093/bioinformatics/btz024
Gardeux, V., David, F.P.A., Shajkofci, A., Schwalie, P.C., Deplancke, B., 2017. ASAP: a web-based platform for the analysis and interactive visualization of single-cell RNA-seq data. Bioinformatics 33, 3123–3125. https://doi.org/10.1093/bioinformatics/btx337
Geng, B., Pan, J., Zhao, T., Ji, J., Zhang, C., Che, Y., Yang, J., Shi, H., Li, J., Zhou, H., Mu, X., Xu, C., Wang, C., Xu, Y., Liu, Z., Wen, H., You, Q., 2018. Chitinase 3-like 1-CD44

interaction promotes metastasis and epithelial-to-mesenchymal transition through β-catenin/Erk/Akt signaling in gastric cancer. Journal of Experimental & Clinical Cancer Research 37, 208. https://doi.org/10.1186/s13046-018-0876-2

Germain, P.-L., Sonrel, A., Robinson, M.D., 2020. pipeComp, a general framework for the evaluation of computational pipelines, reveals performant single cell RNA-seq preprocessing tools. Genome Biology 21, 227. https://doi.org/10.1186/s13059-020-02136-7

Gerstein, M.B., Rozowsky, J., Yan, K.-K., Wang, D., Cheng, C., Brown, J.B., Davis, C.A., Hillier, L., Sisu, C., Li, J.J., Pei, B., Harmanci, A.O., Duff, M.O., Djebali, S., Alexander, R.P., Alver, B.H., Auerbach, R., Bell, K., Bickel, P.J., Boeck, M.E., Boley, N.P., Booth, B.W., Cherbas, L., Cherbas, P., Di, C., Dobin, A., Drenkow, J., Ewing, B., Fang, G., Fastuca, M., Feingold, E.A., Frankish, A., Gao, G., Good, P.J., Guigó, R., Hammonds, A., Harrow, J., Hoskins, R.A., Howald, C., Hu, L., Huang, H., Hubbard, T.J.P., Huynh, C., Jha, S., Kasper, D., Kato, M., Kaufman, T.C., Kitchen, R.R., Ladewig, E., Lagarde, J., Lai, E., Leng, J., Lu, Z., MacCoss, M., May, G., McWhirter, R., Merrihew, G., Miller, D.M., Mortazavi, A., Murad, R., Oliver, B., Olson, S., Park, P.J., Pazin, M.J., Perrimon, N., Pervouchine, D., Reinke, V., Reymond, A., Robinson, G., Samsonova, A., Saunders, G.I., Schlesinger, F., Sethi, A., Slack, F.J., Spencer, W.C., Stoiber, M.H., Strasbourger, P., Tanzer, A., Thompson, O.A., Wan, K.H., Wang, G., Wang, H., Watkins, K.L., Wen, J., Wen, K., Xue, C., Yang, L., Yip, K., Zaleski, C., Zhang, Y., Zheng, H., Brenner, S.E., Graveley, B.R., Celniker, S.E., Gingeras, T.R., Waterston, R., 2014. Comparative analysis of the transcriptome across distant species. Nature 512, 445–448. https://doi.org/10.1038/nature13424

González-Silva, L., Quevedo, L., Varela, I., 2020. Tumor Functional Heterogeneity Unraveled by scRNA-seq Technologies. Trends in Cancer 6, 13–19. https://doi.org/10.1016/j.trecan.2019.11.010

Granja, J.M., Klemm, S., McGinnis, L.M., Kathiria, A.S., Mezger, A., Corces, M.R., Parks, B., Gars, E., Liedtke, M., Zheng, G.X.Y., Chang, H.Y., Majeti, R., Greenleaf, W.J., 2019. Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. Nature Biotechnology 37, 1458–1465. https://doi.org/10.1038/s41587-019-0332-7

Gu, Z., Eils, R., Schlesner, M., 2016. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics.

Haar, C.P., Hebbar, P., Wallace, G.C., Das, A., Vandergrift, W.A., Smith, J.A., Giglio, P., Patel, S.J., Ray, S.K., Banik, N.L., 2012. Drug Resistance in Glioblastoma: A Mini Review. Neurochem Res 37, 1192–1200. https://doi.org/10.1007/s11064-011-0701-1

Haghverdi, L., Lun, A.T.L., Morgan, M.D., Marioni, J.C., 2018. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nature Biotechnology 36, 421–427. https://doi.org/10.1038/nbt.4091

Hai, L., Hoffmann, D.C., Mandelbaum, H., Xie, R., Ito, J., Jung, E., Weil, S., Sievers, P., Venkataramani, V., Azorin, D.D., Ernst, K., Reibold, D., Will, R., Suvà, M.L., Herold-Mende, C., Sahm, F., Winkler, F., Schlesner, M., Wick, W.,Kessler, T. 2021. A connectivity signature for glioblastoma. bioRxiv: 2021.11.07.465791. https://doi.org/10.1101/2021.11.07.465791.

Hai, L., Kessler, T., Wick, W., Schlesner, M., ieCS: interactive explorer of single cell cluster similarity. In Preparation.

Hausmann, D., Hoffmann, D.C., Venkataramani, V., Jung, E., Horschitz, S., Tetzlaff, S.K., Jabali, A., Hai, L., Kessler, T., Azoŕin, D.D., Weil, S., Kourtesakis, A., Sievers, P., Habel, A., Breckwoldt, M.O., Karreman, M.A., Ratliff, M., Messmer, J.M., Yang, Y., Reyhan, E., Wendler, S., Löb, C., Mayer, C., Figarella, K., Osswald, M., Solecki, G., Sahm, F., Garaschuk, O., Kuner, T., Koch, P., Schlesner, M., Wick, W.,Winkler, F. 2023. Autonomous rhythmic activity in glioma networks drives brain tumour growth. Nature 613 (7942): 179-186. https://doi.org/10.1038/s41586-022-05520-4.

Hay, S.B., Ferchen, K., Chetal, K., Grimes, H.L., Salomonis, N., 2018a. The Human Cell Atlas bone marrow single-cell interactive web portal. Experimental Hematology 68, 51–61.

https://doi.org/10.1016/j.exphem.2018.09.004
He, L., Vanlandewijck, M., Raschperger, E., Andaloussi Mäe, M., Jung, B., Lebouvier, T., Ando, K., Hofmann, J., Keller, A., Betsholtz, C., 2016. Analysis of the brain mural cell transcriptome. Scientific Reports 6, 35108. https://doi.org/10.1038/srep35108
Hohensee, I., Chuang, H.-N., Grottke, A., Werner, S., Schulte, A., Horn, S., Lamszus, K., Bartkowiak, K., Witzel, I., Westphal, M., Matschke, J., Glatzel, M., Jücker, M., Pukrop, T., Pantel, K.,Wikman, H. 2016. PTEN mediates the cross talk between breast and glial cells in brain metastases leading to rapid disease progression. Oncotarget 8 (4) https://doi.org/10.18632/oncotarget.14047.
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., Szalai, B., Saez-Rodriguez, J., 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
Hubner, K., Karwelat, D., Pietsch, E., Beinborn, I., Winterberg, S., Bedenbender, K., Benedikter, B.J., Schmeck, B.,Vollmeister, E. 2020. NF-kappaB-mediated inhibition of microRNA-149-5p regulates Chitinase-3-like 1 expression in human airway epithelial cells. Cell Signal 67: 109498. https://doi.org/10.1016/j.cellsig.2019.109498.
Jin, S., Guerrero-Juarez, C.F., Zhang, L., Chang, I., Ramos, R., Kuan, C.H., Myung, P., Plikus, M.V.,Nie, Q. 2021. Inference and analysis of cell-cell communication using CellChat. Nat Commun 12 (1): 1088. https://doi.org/10.1038/s41467-021-21246-9.
Ju, W., Nair, V., Smith, S., Zhu, L., Shedden, K., Song, P.X.K., Mariani, L.H., Eichinger, F.H., Berthier, C.C., Randolph, A., Lai, J.Y.-C., Zhou, Y., Hawkins, J.J., Bitzer, M., Sampson, M.G., Thier, M., Solier, C., Duran-Pacheco, G.C., Duchateau-Nguyen, G., Essioux, L., Schott, B., Formentini, I., Magnone, M.C., Bobadilla, M., Cohen, C.D., Bagnasco, S.M., Barisoni, L., Lv, J., Zhang, H., Wang, H.-Y., Brosius, F.C., Gadegbeku, C.A., Kretzler, M., for the ERCB, C.-P., 2015. Tissue transcriptome-driven identification of epidermal growth factor as a chronic kidney disease biomarker. Science Translational Medicine 7, 316ra193-316ra193. https://doi.org/10.1126/scitranslmed.aac7071
Jung, E., Osswald, M., Ratliff, M., Dogan, H., Xie, R., Weil, S., Hoffmann, D.C., Kurz, F.T., Kessler, T., Heiland, S., von Deimling, A., Sahm, F., Wick, W., Winkler, F., 2021. Tumor cell plasticity, heterogeneity, and resistance in crucial microenvironmental niches in glioma. Nature Communications 12, 1014. https://doi.org/10.1038/s41467-021-21117-3
Kang, H.M., Subramaniam, M., Targ, S., Nguyen, M., Maliskova, L., McCarthy, E., Wan, E., Wong, S., Byrnes, L., Lanata, C.M., Gate, R.E., Mostafavi, S., Marson, A., Zaitlen, N., Criswell, L.A., Ye, C.J., 2018. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. nature biotechnology 36, 10.
Kassambara, A., 2020. ggpubr: "ggplot2" Based Publication Ready Plots.
Kassambara, A., Kosinski, M., 2018. survminer: Drawing Survival Curves using "ggplot2."
Kassambara, A., Mundt, F., 2019. factoextra: Extract and Visualize the Results of Multivariate Data Analyses.
Kawada, M., Seno, H., Kanda, K., Nakanishi, Y., Akitake, R., Komekado, H., Kawada, K., Sakai, Y., Mizoguchi, E., Chiba, T., 2012. Chitinase 3-like 1 promotes macrophage recruitment and angiogenesis in colorectal cancer. Oncogene 31, 3111–3123. https://doi.org/10.1038/onc.2011.498
Khan, A., 2018, collapsibleTree: Interactive Collapsible Tree Diagrams using "D3.js." R package. https://cran.r-project.org/web/packages/collapsibleTree/
Kiselev, V.Y., Yiu, A., Hemberg, M., 2018. scmap: projection of single-cell RNA-seq data across data sets. Nature Methods 15, 359–362. https://doi.org/10.1038/nmeth.4644
Kolde, R., 2019. pheatmap: Pretty Heatmaps. R package. https://cran.r-project.org/web/packages/pheatmap/
Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P., Raychaudhuri, S., 2019. Fast, sensitive and accurate integration

of single-cell data with Harmony. Nature Methods 16, 1289–1296. https://doi.org/10.1038/s41592-019-0619-0

Koshy, M., Villano, J.L., Dolecek, T.A., Howard, A., Mahmood, U., Chmura, S.J., Weichselbaum, R.R., McCarthy, B.J., 2012. Improved survival time trends for glioblastoma using the SEER 17 population-based registries. J Neurooncol 107, 207–212. https://doi.org/10.1007/s11060-011-0738-7

Krzak, M., Raykov, Y., Boukouvalas, A., Cutillo, L., Angelini, C., 2019. Benchmark and Parameter Sensitivity Analysis of Single-Cell RNA Sequencing Clustering Methods. Front. Genet. 10. https://doi.org/10.3389/fgene.2019.01253

Ku, B.M., Lee, Y.K., Ryu, J., Jeong, J.Y., Choi, J., Eun, K.M., Shin, H.Y., Kim, D.G., Hwang, E.M., Yoo, J. cheal, Park, J.-Y., Roh, G.S., Kim, H.J., Cho, G.J., Choi, W.S., Paek, S.H., Kang, S.S., 2011. CHI3L1 (YKL-40) is expressed in human gliomas and regulates the invasion, growth and survival of glioma cells. International Journal of Cancer 128, 1316–1326. https://doi.org/10.1002/ijc.25466

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., Team, the R.C., Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C., Hunt, T., 2018. caret: Classification and Regression Training.

Kuleshov, M.V., Xie, Z., London, A.B.K., Yang, J., Evangelista, J.E., Lachmann, A., Shu, I., Torre, D.,Ma'ayan, A. 2021. KEA3: improved kinase enrichment analysis via data integration. Nucleic Acids Res 49 (W1): W304-W316. https://doi.org/10.1093/nar/gkab359.

La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M.E., Lonnerberg, P., Furlan, A., Fan, J., Borm, L.E., Liu, Z., van Bruggen, D., Guo, J., He, X., Barker, R., Sundstrom, E., Castelo-Branco, G., Cramer, P., Adameyko, I., Linnarsson, S.,Kharchenko, P.V. 2018. RNA velocity of single cells. Nature 560 (7719): 494-498. https://doi.org/10.1038/s41586-018-0414-6.

Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D.J., Hicks, S.C., Robinson, M.D., Vallejos, C.A., Campbell, K.R., Beerenwinkel, N., Mahfouz, A., Pinello, L., Skums, P., Stamatakis, A., Attolini, C.S.-O., Aparicio, S., Baaijens, J., Balvert, M., Barbanson, B. de, Cappuccio, A., Corleone, G., Dutilh, B.E., Florescu, M., Guryev, V., Holmer, R., Jahn, K., Lobo, T.J., Keizer, E.M., Khatri, I., Kielbasa, S.M., Korbel, J.O., Kozlov, A.M., Kuo, T.-H., Lelieveldt, B.P.F., Mandoiu, I.I., Marioni, J.C., Marschall, T., Mölder, F., Niknejad, A., Raczkowski, L., Reinders, M., Ridder, J. de, Saliba, A.-E., Somarakis, A., Stegle, O., Theis, F.J., Yang, H., Zelikovsky, A., McHardy, A.C., Raphael, B.J., Shah, S.P., Schönhuth, A., 2020. Eleven grand challenges in single-cell data science. Genome Biology 21, 31. https://doi.org/10.1186/s13059-020-1926-6

Levitin, H.M., Yuan, J., Sims, P.A., 2018. Single-Cell Transcriptomic Analysis of Tumor Heterogeneity. Trends in Cancer, Special Issue: Physical Sciences in Oncology 4, 264–268. https://doi.org/10.1016/j.trecan.2018.02.003

Li, Y.I., Knowles, D.A., Humphrey, J., Barbeira, A.N., Dickinson, S.P., Im, H.K., Pritchard, J.K., 2018. Annotation-free quantification of RNA splicing using LeafCutter. Nature Genetics 50, 151–158. https://doi.org/10.1038/s41588-017-0004-9

Liao, Y., Smyth, G.K., Shi, W., 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30, 923–930. https://doi.org/10.1093/bioinformatics/btt656

Lim, S.H., Kwon, S.K., Lee, M.K., Moon, J., Jeong, D.G., Park, E., Kim, S.J., Park, B.C., Lee, S.C., Ryu, S.E., Yu, D.Y., Chung, B.H., Kim, E., Myung, P.K.,Lee, J.R. 2009. Synapse formation regulated by protein tyrosine phosphatase receptor T through interaction with cell adhesion molecules and Fyn. EMBO J 28 (22): 3564-78. https://doi.org/10.1038/emboj.2009.289.

Lou, E., 2017. Can you hear them now? Tumor microtubes form cellular communication networks that protect gliomas from surgical lesions and chemotherapy treatments. Neuro Oncol 19, 1289–1291. https://doi.org/10.1093/neuonc/nox103

Louis, D.N., Perry, A., Reifenberger, G., von Deimling, A., Figarella-Branger, D., Cavenee, W.K., Ohgaki, H., Wiestler, O.D., Kleihues, P., Ellison, D.W., 2016. The 2016 World

Health Organization Classification of Tumors of the Central Nervous System: a summary. Acta Neuropathol 131, 803–820. https://doi.org/10.1007/s00401-016-1545-1

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

Luecken, M.D., Theis, F.J., 2019. Current best practices in single-cell RNA-seq analysis: a tutorial. Molecular Systems Biology 15, e8746. https://doi.org/10.15252/msb.20188746

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K., 2021. cluster: Cluster Analysis Basics and Extensions.

McCarthy, D.J., Campbell, K.R., Lun, A.T.L., Wills, Q.F., 2017. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. Bioinformatics 33, 1179–1186. https://doi.org/10.1093/bioinformatics/btw777

Mederos, S., González-Arias, C., Perea, G., 2018. Astrocyte–Neuron Networks: A Multilane Highway of Signaling for Homeostatic Brain Function. Front. Synaptic Neurosci. 10. https://doi.org/10.3389/fnsyn.2018.00045

Morera, E., Steinhäuser, S.S., Budkova, Z., Ingthorsson, S., Kricker, J., Krueger, A., Traustadottir, G.A., Gudjonsson, T., 2019. YKL-40/CHI3L1 facilitates migration and invasion in HER2 overexpressing breast epithelial progenitor cells and generates a niche for capillary-like network formation. In Vitro Cell.Dev.Biol.-Animal 55, 838–853. https://doi.org/10.1007/s11626-019-00403-x

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., Dewitt, J., Gritsch, S., Perez, E.M., Gonzalez Castro, L.N., Lan, X., Druck, N., Rodman, C., Dionne, D., Kaplan, A., Bertalan, M.S., Small, J., Pelton, K., Becker, S., Bonal, D., Nguyen, Q.-D., Servis, R.L., Fung, J.M., Mylvaganam, R., Mayr, L., Gojo, J., Haberler, C., Geyeregger, R., Czech, T., Slavc, I., Nahed, B.V., Curry, W.T., Carter, B.S., Wakimoto, H., Brastianos, P.K., Batchelor, T.T., Stemmer-Rachamimov, A., Martinez-Lage, M., Frosch, M.P., Stamenkovic, I., Riggi, N., Rheinbay, E., Monje, M., Rozenblatt-Rosen, O., Cahill, D.P., Patel, A.P., Hunter, T., Verma, I.M., Ligon, K.L., Louis, D.N., Regev, A., Bernstein, B.E., Tirosh, I., Suvà, M.L., 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

Neuwirth, E., 2014. RColorBrewer: ColorBrewer Palettes.

Newman, A.M., Steen, C.B., Liu, C.L., Gentles, A.J., Chaudhuri, A.A., Scherer, F., Khodadoust, M.S., Esfahani, M.S., Luca, B.A., Steiner, D., Diehn, M.,Alizadeh, A.A. 2019. Determining cell type abundance and expression from bulk tissues with digital cytometry. Nature Biotechnology 37 (7): 773-782. https://doi.org/10.1038/s41587-019-0114-2.

Nimmerjahn, A., Kirchhoff, F., Kerr, J.N.D., Helmchen, F., 2004. Sulforhodamine 101 as a specific marker of astroglia in the neocortex in vivo. Nature Methods 1, 31–37. https://doi.org/10.1038/nmeth706

Osswald, M., Jung, E., Sahm, F., Solecki, G., Venkataramani, V., Blaes, J., Weil, S., Horstmann, H., Wiestler, B., Syed, M., Huang, L., Ratliff, M., Karimian Jazi, K., Kurz, F.T., Schmenger, T., Lemke, D., Gömmel, M., Pauli, M., Liao, Y., Häring, P., Pusch, S., Herl, V., Steinhäuser, C., Krunic, D., Jarahian, M., Miletic, H., Berghoff, A.S., Griesbeck, O., Kalamakis, G., Garaschuk, O., Preusser, M., Weiss, S., Liu, H., Heiland, S., Platten, M., Huber, P.E., Kuner, T., von Deimling, A., Wick, W., Winkler, F., 2015. Brain tumour cells interconnect to a functional and resistant network. Nature 528, 93–98. https://doi.org/10.1038/nature16071

Ostrom, Q.T., Bauchet, L., Davis, F.G., Deltour, I., Fisher, J.L., Langer, C.E., Pekmezci, M., Schwartzbaum, J.A., Turner, M.C., Walsh, K.M., Wrensch, M.R., Barnholtz-Sloan, J.S., 2014. The epidemiology of glioma in adults: a “state of the science” review. Neuro-Oncology 16, 896–913. https://doi.org/10.1093/neuonc/nou087

Paradis, E., Schliep, K., 2018. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 35, 526–528.

Park, S.R., Namkoong, S., Friesen, L., Cho, C.-S., Zhang, Z.Z., Chen, Y.-C., Yoon, E., Kim, C.H., Kwak, H., Kang, H.M., Lee, J.H., 2020. Single-Cell Transcriptome Analysis of Colon Cancer Cell Response to 5-Fluorouracil-Induced DNA Damage. Cell Reports 32, 108077. https://doi.org/10.1016/j.celrep.2020.108077

Pedersen, T.L., 2019a. ggraph: An Implementation of Grammar of Graphics for Graphs and Networks.

Pedersen, T.L., 2019b. tidygraph: A Tidy API for Graph Manipulation.

Pedrotty, D.M., Morley, M.P., Cappola, T.P., 2012. Transcriptomic Biomarkers of Cardiovascular Disease. Prog Cardiovasc Dis 55, 64–69. https://doi.org/10.1016/j.pcad.2012.06.003

Phillips, H.S., Kharbanda, S., Chen, R., Forrest, W.F., Soriano, R.H., Wu, T.D., Misra, A., Nigro, J.M., Colman, H., Soroceanu, L., Williams, P.M., Modrusan, Z., Feuerstein, B.G., Aldape, K., 2006. Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. Cancer Cell 9, 157–173. https://doi.org/10.1016/j.ccr.2006.02.019

Picelli, S., Faridani, O.R., Björklund, Å.K., Winberg, G., Sagasser, S., Sandberg, R., 2014. Full-length RNA-seq from single cells using Smart-Seq2. Nature Protocols 9, 171–181. https://doi.org/10.1038/nprot.2014.006

Poirion, O.B., Zhu, X., Ching, T., Garmire, L., 2016. Single-Cell Transcriptomics Bioinformatics and Computational Challenges. Front. Genet. 7. https://doi.org/10.3389/fgene.2016.00163

Qiu, Q.-C., Wang, L., Jin, S.-S., Liu, G.-F., Liu, J., Ma, L., Mao, R.-F., Ma, Y.-Y., Zhao, N., Chen, M., Lin, B.-Y., 2018. CHI3L1 promotes tumor progression by activating TGF-β signaling pathway in hepatocellular carcinoma. Sci Rep 8. https://doi.org/10.1038/s41598-018-33239-8

R Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rai, A., Kapoor, S., Singh, S., Chatterji, B.P.,Panda, D. 2015. Transcription factor NF-kappaB associates with microtubules and stimulates apoptosis in response to suppression of microtubule dynamics in MCF-7 cells. Biochem Pharmacol 93 (3): 277-89. https://doi.org/10.1016/j.bcp.2014.12.007.

Raouf, A., Zhao, Y., To, K., Stingl, J., Delaney, A., Barbara, M., Iscove, N., Jones, S., McKinney, S., Emerman, J., Aparicio, S., Marra, M., Eaves, C., 2008. Transcriptome Analysis of the Normal Human Mammary Cell Commitment and Differentiation Process. Cell Stem Cell 3, 109–118. https://doi.org/10.1016/j.stem.2008.05.018

Raskin, L., Fullen, D.R., Giordano, T.J., Thomas, D.G., Frohm, M.L., Cha, K.B., Ahn, J., Mukherjee, B., Johnson, T.M., Gruber, S.B., 2013. Transcriptome Profiling Identifies HMGA2 as a Biomarker of Melanoma Progression and Prognosis. Journal of Investigative Dermatology 133, 2585–2592. https://doi.org/10.1038/jid.2013.197

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., Smyth, G.K., 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research 43, e47.

Rousseeuw, P.J., 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Ruiz-Moreno, C., Salas, S.M., Samuelsson, E., Brandner, S., Kranendonk, M.E., Nilsson, M. and Stunnenberg, H.G., 2022. Harmonized single-cell landscape, intercellular crosstalk and tumor architecture of glioblastoma. BioRxiv, pp.2022-08. https://doi.org/10.1101/2022.08.27.505439

Saelens, W., Cannoodt, R., Todorov, H., Saeys, Y., 2019. A comparison of single-cell trajectory inference methods. Nature Biotechnology 37, 547–554. https://doi.org/10.1038/s41587-019-0071-9

Saidova, A.A., Vorobjev, I.A., 2019. Lineage Commitment, Signaling Pathways, and the

Cytoskeleton Systems in Mesenchymal Stem Cells. Tissue Engineering Part B: Reviews 26, 13–25. https://doi.org/10.1089/ten.teb.2019.0250
Shabalin, A.A., 2012. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics 28, 1353–1358. https://doi.org/10.1093/bioinformatics/bts163
Sharov, A.A., Piao, Y., Matoba, R., Dudekula, D.B., Qian, Y., VanBuren, V., Falco, G., Martin, P.R., Stagg, C.A., Bassey, U.C., Wang, Y., Carter, M.G., Hamatani, T., Aiba, K., Akutsu, H., Sharova, L., Tanaka, T.S., Kimber, W.L., Yoshikawa, T., Jaradat, S.A., Pantano, S., Nagaraja, R., Boheler, K.R., Taub, D., Hodes, R.J., Longo, D.L., Schlessinger, D., Keller, J., Klotz, E., Kelsoe, G., Umezawa, A., Vescovi, A.L., Rossant, J., Kunath, T., Hogan, B.L.M., Curci, A., D'Urso, M., Kelso, J., Hide, W., Ko, M.S.H., 2003. Transcriptome Analysis of Mouse Stem Cells and Early Embryos. PLOS Biology 1, e74. https://doi.org/10.1371/journal.pbio.0000074
Sievert, C., 2020. Interactive Web-Based Data Visualization with R, plotly, and shiny. Chapman and Hall/CRC.
Stanta, G., Bonin, S., 2018. Overview on Clinical Relevance of Intra-Tumor Heterogeneity. Front. Med. 5. https://doi.org/10.3389/fmed.2018.00085
Stegle, O., Teichmann, S.A., Marioni, J.C., 2015. Computational and analytical challenges in single-cell transcriptomics. Nature Reviews Genetics 16, 133–145. https://doi.org/10.1038/nrg3833
Steponaitis, G., Skiriutė, D., Kazlauskas, A., Golubickaitė, I., Stakaitis, R., Tamašauskas, A., Vaitkienė, P., 2016. High CHI3L1 expression is associated with glioma patient survival. Diagn Pathol 11. https://doi.org/10.1186/s13000-016-0492-4
Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., Hao, Y., Stoeckius, M., Smibert, P., Satija, R., 2019. Comprehensive Integration of Single-Cell Data. Cell 177, 1888-1902.e21. https://doi.org/10.1016/j.cell.2019.05.031
Sul, J.-Y., Orosz, G., Givens, R.S., Haydon, P.G., 2004. Astrocytic Connectivity in the Hippocampus. Neuron Glia Biol 1, 3–11. https://doi.org/10.1017/s1740925x04000031
Tamimi, A.F., Juweid, M., 2017. Epidemiology and Outcome of Glioblastoma, in: De Vleeschouwer, S. (Ed.), Glioblastoma. Codon Publications, Brisbane (AU).
Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B.B., Siddiqui, A., Lao, K., Surani, M.A., 2009. mRNA-Seq whole-transcriptome analysis of a single cell. Nature Methods 6, 377–382. https://doi.org/10.1038/nmeth.1315
Tang, Z., Li, C., Kang, B., Gao, G., Li, C.,Zhang, Z. 2017. GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. Nucleic Acids Res 45 (W1): W98-W102. https://doi.org/10.1093/nar/gkx247.
Tarasov, A., Vilella, A.J., Cuppen, E., Nijman, I.J.,Prins, P. 2015. Sambamba: fast processing of NGS alignment formats. Bioinformatics 31 (12): 2032-4. https://doi.org/10.1093/bioinformatics/btv098.
Therneau, T.M., 2020. A Package for Survival Analysis in R.
Tian, L., Dong, X., Freytag, S., Lê Cao, K.-A., Su, S., JalalAbadi, A., Amann-Zalcenstein, D., Weber, T.S., Seidi, A., Jabbari, J.S., Naik, S.H., Ritchie, M.E., 2019. Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. Nature Methods 16, 479–487. https://doi.org/10.1038/s41592-019-0425-8
Tickle, T., Tirosh, I., Georgescu, C., Brown, M., Haas, B., 2019. inferCNV of the Trinity CTAT Project. Klarman Cell Observatory, Broad Institute of MIT and Harvard, Cambridge, MA, USA.
Tran, H.T.N., Ang, K.S., Chevrier, M., Zhang, X., Lee, N.Y.S., Goh, M., Chen, J., 2020. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. Genome Biology 21, 12. https://doi.org/10.1186/s13059-019-1850-9
Tsuyuzaki, K., Sato, H., Sato, K., Nikaido, I., 2020. Benchmarking principal component analysis for large-scale single-cell RNA-sequencing. Genome Biol 21, 9. https://doi.org/10.1186/s13059-019-1900-3
Uhrig, S., Ellermann, J., Walther, T., Burkhardt, P., Fröhlich, M., Hutter, B., Toprak, U.H., Neumann, O., Stenzinger, A., Scholl, C., Fröhling, S., Brors, B., 2021. Accurate and

efficient detection of gene fusions from RNA sequencing data. Genome Res. 31, 448–460. https://doi.org/10.1101/gr.257246.119

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., Brat, D., Gould, P.V., Woehrer, A., Aldape, K., Ismail, A., Sivajothi, S.K., Barthel, F.P., Kim, H., Kocakavuk, E., Ahmed, N., White, K., Datta, I., Moon, H.E., Pollock, S., Goldfarb, C., Lee, G.H., Garofano, L., Anderson, K.J., Nehar-Belaid, D., Barnholtz-Sloan, J.S., Bakas, S., Byrne, A.T., D'Angelo, F., Gan, H.K., Khasraw, M., Migliozzi, S., Ormond, D.R., Paek, S.H., Van Meir, E.G., Walenkamp, A.M.E., Watts, C., Weiss, T., Weller, M., Palucka, K., Stead, L.F., Poisson, L.M., Noushmehr, H., Iavarone, A., Verhaak, R.G.W.,Consortium, G. 2022. Glioma progression is shaped by genetic evolution and microenvironment interactions. Cell 185 (12): 2184-2199 e16. https://doi.org/10.1016/j.cell.2022.04.038.

Venkataramani, V., Tanev, D.I., Strahle, C., Studier-Fischer, A., Fankhauser, L., Kessler, T., Körber, C., Kardorff, M., Ratliff, M., Xie, R., Horstmann, H., Messer, M., Paik, S.P., Knabbe, J., Sahm, F., Kurz, F.T., Acikgöz, A.A., Herrmannsdörfer, F., Agarwal, A., Bergles, D.E., Chalmers, A., Miletic, H., Turcan, S., Mawrin, C., Hänggi, D., Liu, H.-K., Wick, W., Winkler, F., Kuner, T., 2019. Glutamatergic synaptic input to glioma cells drives brain tumour progression. Nature 573, 532–538. https://doi.org/10.1038/s41586-019-1564-x

Venkataramani, V., Schneider, M., Giordano, F.A., Kuner, T., Wick, W., Herrlinger, U. and Winkler, F., 2022. Disconnecting multicellular networks in brain tumours. Nature Reviews Cancer, 22(8), pp.481-491. https://doi.org/10.1038/s41568-022-00475-0

Verhaak, R.G.W., Hoadley, K.A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M.D., Miller, C.R., Ding, L., Golub, T., Mesirov, J.P., Alexe, G., Lawrence, M., O'Kelly, M., Tamayo, P., Weir, B.A., Gabriel, S., Winckler, W., Gupta, S., Jakkula, L., Feiler, H.S., Hodgson, J.G., James, C.D., Sarkaria, J.N., Brennan, C., Kahn, A., Spellman, P.T., Wilson, R.K., Speed, T.P., Gray, J.W., Meyerson, M., Getz, G., Perou, C.M., Hayes, D.N., 2010. Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell 17, 98–110. https://doi.org/10.1016/j.ccr.2009.12.020

Vieth, B., Parekh, S., Ziegenhain, C., Enard, W., Hellmann, I., 2019. A systematic evaluation of single cell RNA-seq analysis pipelines. Nature Communications 10, 4667. https://doi.org/10.1038/s41467-019-12266-7

Wang, L.B., Karpova, A., Gritsenko, M.A., Kyle, J.E., Cao, S., Li, Y., Rykunov, D., Colaprico, A., Rothstein, J.H., Hong, R., Stathias, V., Cornwell, M., Petralia, F., Wu, Y., Reva, B., Krug, K., Pugliese, P., Kawaler, E., Olsen, L.K., Liang, W.W., Song, X., Dou, Y., Wendl, M.C., Caravan, W., Liu, W., Cui Zhou, D., Ji, J., Tsai, C.F., Petyuk, V.A., Moon, J., Ma, W., Chu, R.K., Weitz, K.K., Moore, R.J., Monroe, M.E., Zhao, R., Yang, X., Yoo, S., Krek, A., Demopoulos, A., Zhu, H., Wyczalkowski, M.A., McMichael, J.F., Henderson, B.L., Lindgren, C.M., Boekweg, H., Lu, S., Baral, J., Yao, L., Stratton, K.G., Bramer, L.M., Zink, E., Couvillion, S.P., Bloodsworth, K.J., Satpathy, S., Sieh, W., Boca, S.M., Schurer, S., Chen, F., Wiznerowicz, M., Ketchum, K.A., Boja, E.S., Kinsinger, C.R., Robles, A.I., Hiltke, T., Thiagarajan, M., Nesvizhskii, A.I., Zhang, B., Mani, D.R., Ceccarelli, M., Chen, X.S., Cottingham, S.L., Li, Q.K., Kim, A.H., Fenyo, D., Ruggles, K.V., Rodriguez, H., Mesri, M., Payne, S.H., Resnick, A.C., Wang, P., Smith, R.D., Iavarone, A., Chheda, M.G., Barnholtz-Sloan, J.S., Rodland, K.D., Liu, T., Ding, L.,Clinical Proteomic Tumor Analysis, C. 2021. Proteogenomic and metabolomic characterization of human glioblastoma. Cancer Cell 39 (4): 509-528 e20. https://doi.org/10.1016/j.ccell.2021.01.006.

Wang, Q., Hu, B., Hu, X., Kim, H., Squatrito, M., Scarpace, L., deCarvalho, A.C., Lyu, S., Li, P., Li, Y., Barthel, F., Cho, H.J., Lin, Y.-H., Satani, N., Martinez-Ledesma, E., Zheng, S., Chang, E., Sauvé, C.-E.G., Olar, A., Lan, Z.D., Finocchiaro, G., Phillips, J.J., Berger, M.S., Gabrusiewicz, K.R., Wang, G., Eskilsson, E., Hu, J., Mikkelsen, T., DePinho, R.A., Muller, F., Heimberger, A.B., Sulman, E.P., Nam, D.-H., Verhaak, R.G.W., 2017. Tumor Evolution of Glioma-Intrinsic Gene Expression Subtypes

Associates with Immunological Changes in the Microenvironment. Cancer Cell 32, 42-56.e6. https://doi.org/10.1016/j.ccell.2017.06.003
Wang, T., Li, B., Nelson, C.E., 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
Wang, Y., Cui, J., Sun, X., Zhang, Y., 2011. Tunneling-nanotube development in astrocytes depends on p53 activation. Cell Death & Differentiation 18, 732–742. https://doi.org/10.1038/cdd.2010.147
Wang, Z., Gerstein, M., Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics 10, 57–63. https://doi.org/10.1038/nrg2484
Wei, I.H., Shi, Y., Jiang, H., Kumar-Sinha, C., Chinnaiyan, A.M., 2014. RNA-Seq Accurately Identifies Cancer Biomarker Signatures to Distinguish Tissue of Origin. Neoplasia 16, 918–927. https://doi.org/10.1016/j.neo.2014.09.007
Weil, S., Osswald, M., Solecki, G., Grosch, J., Jung, E., Lemke, D., Ratliff, M., Hänggi, D., Wick, W., Winkler, F., 2017. Tumor microtubes convey resistance to surgical lesions and chemotherapy in gliomas. Neuro-Oncology 19, 1316–1326. https://doi.org/10.1093/neuonc/nox070
Welch, J.D., Kozareva, V., Ferreira, A., Vanderburg, C., Martin, C., Macosko, E.Z., 2019. Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. Cell 177, 1873-1887.e17. https://doi.org/10.1016/j.cell.2019.05.006
Weller, M., van den Bent, M., Preusser, M., Le Rhun, E., Tonn, J.C., Minniti, G., Bendszus, M., Balana, C., Chinot, O., Dirven, L., French, P., Hegi, M.E., Jakola, A.S., Platten, M., Roth, P., Rudà, R., Short, S., Smits, M., Taphoorn, M.J.B., von Deimling, A., Westphal, M., Soffietti, R., Reifenberger, G., Wick, W., 2021. EANO guidelines on the diagnosis and treatment of diffuse gliomas of adulthood. Nature Reviews Clinical Oncology 18, 170–186. https://doi.org/10.1038/s41571-020-00447-z
Weller, M., Wick, W., Aldape, K., Brada, M., Berger, M., Pfister, S.M., Nishikawa, R., Rosenthal, M., Wen, P.Y., Stupp, R., Reifenberger, G., 2015. Glioma. Nature Reviews Disease Primers 1, 1–18. https://doi.org/10.1038/nrdp.2015.17
Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.
Wickham, H., 2011. The Split-Apply-Combine Strategy for Data Analysis. Journal of Statistical Software 40, 1–29.
Winkler, F., Wick, W., 2018. Harmful networks in the brain and beyond. Science 359, 1100–1101. https://doi.org/10.1126/science.aar5555
Wolf, F.A., Angerer, P., Theis, F.J., 2018. SCANPY: large-scale single-cell gene expression data analysis. Genome Biology 19, 15. https://doi.org/10.1186/s13059-017-1382-0
Wolock, S.L., Lopez, R., Klein, A.M., 2019. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. Cell Systems 8, 281-291.e9. https://doi.org/10.1016/j.cels.2018.11.005
Wu, J.Q., Habegger, L., Noisa, P., Szekely, A., Qiu, C., Hutchison, S., Raha, D., Egholm, M., Lin, H., Weissman, S., Cui, W., Gerstein, M., Snyder, M., 2010. Dynamic transcriptomes during neural differentiation of human embryonic stem cells revealed by short, long, and paired-end sequencing. PNAS 107, 5254–5259. https://doi.org/10.1073/pnas.0914114107
Xi, N.M., Li, J.J., 2021. Benchmarking Computational Doublet-Detection Methods for Single-Cell RNA Sequencing Data. Cell Systems 12, 176-194.e6. https://doi.org/10.1016/j.cels.2020.11.008
Xie, R., Kessler, T., Grosch, J., Hai, L., Venkataramani, V., Huang, L., Hoffmann, D.C., Solecki, G., Ratliff, M., Schlesner, M., Wick, W., Winkler, F., 2021. Tumor cell network integration in glioma represents a stemness feature. Neuro Oncol 23(5), 757–769. https://doi.org/10.1093/neuonc/noaa275
Xie, Y., Cheng, J., Tan, X., 2019. DT: A Wrapper of the JavaScript Library "DataTables."
Zappia, L., Phipson, B., Oshlack, A., 2018. Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. PLOS Computational Biology 14, e1006245. https://doi.org/10.1371/journal.pcbi.1006245

Zeileis, A., Hornik, K., Murrell, P., 2009. Escaping RGBland: Selecting Colors for Statistical Graphics. Computational Statistics & Data Analysis 53, 3259–3270. https://doi.org/10.1016/j.csda.2008.11.033

Zhang, X., Smits, A.H., van Tilburg, G.B., Ovaa, H., Huber, W.,Vermeulen, M. 2018. Proteome-wide identification of ubiquitin interactions using UbIA-MS. Nat Protoc 13 (3): 530-550. https://doi.org/10.1038/nprot.2017.147.

Zhang, Y., Sloan, S.A., Clarke, L.E., Caneda, C., Plaza, C.A., Blumenthal, P.D., Vogel, H., Steinberg, G.K., Edwards, M.S.B., Li, G., Duncan, J.A., Cheshier, S.H., Shuer, L.M., Chang, E.F., Grant, G.A., Gephart, M.G.H., Barres, B.A., 2016. Purification and Characterization of Progenitor and Mature Human Astrocytes Reveals Transcriptional and Functional Differences with Mouse. Neuron 89, 37–53. https://doi.org/10.1016/j.neuron.2015.11.013

Zhao, T., Su, Z., Li, Y., Zhang, X., You, Q., 2020. Chitinase-3 like-protein-1 function and its role in diseases. Signal Transduction and Targeted Therapy 5, 1–20. https://doi.org/10.1038/s41392-020-00303-7

Zhao, T., Zeng, J., Xu, Y., Su, Z., Chong, Y., Ling, T., Xu, H., Shi, H., Zhu, M., Mo, Q., Huang, X., Li, Y., Zhang, X., Ni, H.,You, Q. 2022. Chitinase-3 like-protein-1 promotes glioma progression via the NF-kappaB signaling pathway and tumor microenvironment reprogramming. Theranostics 12 (16): 6989-7008. https://doi.org/10.7150/thno.75069.

Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., Gregory, M.T., Shuga, J., Montesclaros, L., Underwood, J.G., Masquelier, D.A., Nishimura, S.Y., Schnall-Levin, M., Wyatt, P.W., Hindson, C.M., Bharadwaj, R., Wong, A., Ness, K.D., Beppu, L.W., Deeg, H.J., McFarland, C., Loeb, K.R., Valente, W.J., Ericson, N.G., Stevens, E.A., Radich, J.P., Mikkelsen, T.S., Hindson, B.J., Bielas, J.H., 2017. Massively parallel digital transcriptional profiling of single cells. Nature Communications 8, 14049. https://doi.org/10.1038/ncomms14049

Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A.H., Tanaseichuk, O., Benner, C.,Chanda, S.K. 2019. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat Commun 10 (1): 1523. https://doi.org/10.1038/s41467-019-09234-6.

Zhu, X., Wolfgruber, T.K., Tasato, A., Arisdakessian, C., Garmire, D.G., Garmire, L.X., 2017. Granatum: a graphical single-cell RNA-Seq analysis pipeline for genomics scientists. Genome Medicine 9, 108. https://doi.org/10.1186/s13073-017-0492-3

6. Supplementary Tables

Supplementary Table 1 **2978 DEGs in CHI3L1 OE RNA-Seq data.**

2252 upregulated DEGs

C5orf38, EDN3, GAL, MED15P1, PXDN, FOLH1, ADARB2, COL3A1, CPXM2, ZNF626, GYPC, ADAMTS18, ST6GAL2, FABP5P7, RP11-171N4.1, COLEC12, AQP4, DPT, GPRC5C, DNAJC5B, RP11-597A11.1, LINC01016, NHLRC1, TUBB8P8, TLX1NB, CACNA1C, TEK, OTOR, AP000344.3, ADCY2, RP3-410C9.2, KCNQ5, NF1P4, SERTAD4, MXRA5, ST6GAL2-IT1, UNC5C, RIMS4, RP11-347H15.1, GBP4, RP11-91P17.1, CHST9, GRM4, KAL1, RUFY4, CHST15, CHI3L1, CTC-264K15.6, RP11-8L8.2, MIR143HG, DNAJC15, RP11-138J23.1, EFEMP1, TDO2, LRRN3, VIT, AC092614.2, KCNG4, RP11-367G6.3, RP11-6N13.1, SMTNL2, SMOC2, NF1P6, CTD-2022H16.1, PHACTR3, RXRG, AP004372.1, DNTT, ANO1, EGFLAM, RP11-558A11.1, HAND1, COL5A3, CTD-2022H16.2, RIN3, RP11-1259L22.1, IL1RN, CTD-2553C6.1, CTB-78F1.2, GPRIN3, TIMP3, FLRT2, PCDH8, CTD-2339F6.1, RP11-344F13.1, ME3, PLEK, KCNA1, IGFN1, TUBBP5, ST8SIA2, GPR101, LPAR1, MXRA5P1, GPR26, RP1-214M20.2, AFF3, RP11-449J10.1, DUSP27, SCN4A, CXorf28, RP11-799P8.1, CTB-26E19.1, LMX1B, GREM2, STAT4, ATP12A, AC009093.1, RP4-796I8.1, THRB, RP11-1112C15.2, ADRA1B, RP4-610C12.3, RP11-143J24.1, RP5-1139B12.3, RP11-720L2.4, FRAS1, MAMDC2, CTD-2022H16.3, CASP1, KCNA5, TBX3, ITGA8, SLN, AKR1C1, GBP3, GBP1, FAM50B, ACAN, SCN10A, CYTL1, RP11-245J24.1, RTP4, NCOR1P1, TENM2, EMX2OS, ROR2, IGSF21, SAMD5, AKR1C3, DMRT2, CCDC42, SGIP1, AMER2, GABRG3, RP11-124N19.3, DHX58, PCDH20, CPPED1, ALPK2, SORBS2, FABP5, IKZF3, MYO1B, CEACAMP10, APOD, ZNF679, RP11-391M7.3, AC109583.1, INHBA, RP11-321L2.1, ADH1C, MIR4635, ZSWIM2, SEMA3E, AP002856.5, CHRDL1, RP4-610C12.4, TNS1, LRRC32, FAM149A, DACT2, XKR8, ZSWIM5P2, SH3GL3, OLFM4, BGN, RP11-749H20.4, STMN2, WSCD2, ACSM5, TNFAIP8L3, NF1P8, COL8A1, FOXI2, C11orf87, DSG2, TLR4, ATP1A2, TFAP2C, NBEAP1, RET, SLC22A3, NANOGP1, IFI44, PSORS1C3, ADAMTS2, FAM90A20P, ANKRD33, C8orf48, FMO2, RP11-94B19.6, AC012123.1, AC096669.3, RP11-396O20.1, SLC17A8, ACSS3, CNTN2, RP11-348F1.2, LGR6, UGT3A2, AQP4-AS1, HSD11B1, IRX2, AC013271.3, AC092667.2, CMKLR1, CDH13, SLC1A2, AC017076.5, MGAT3, AC109826.1, PTGER2, SYNDIG1, IL15, RERG, AMIGO2, NTN4, MYL3, CACNG3, CNTN1, FAM189A1, CARD16, TAAR3, APOBEC3G, UST, CTC-353G13.1, SCML4, CCL24, GPR27, ELAVL2, PDGFRA, SERTAD4-AS1, RIT2, PCSK2, EPSTI1, CCDC8, OAS1, TLX1, LBX1-AS1, WNT5A,

GABRA3, GOLGA8J, OLFM2, MIR145, RP11-43F13.4, AP000345.1, SYT4, MT1A, CTC-321K16.1, RP11-1002K11.1, RP4-813D12.3, GLDN, RP11-578F21.12, ARSE, FGL2, RP11-100M12.3, F13A1, HLA-DPA1, APOBEC3D, GALNT9, COL1A1, RBBP8NL, TRAJ24, VCAM1, TMEM100, AKR1C2, FAM26E, PDLIM3, CACNA1D, RP11-155L15.1, DSC1, RP11-683L23.1, SLFN5, MDFI, C17orf72, PMAIP1, SHISA3, CLEC2L, RP4-745K6.1, ALX3, WFDC1, GABRG2, TRIM6, RP11-173M11.2, SLC6A7, HTR7, RP11-561N12.5, PRMT8, RP11-693N9.2, CPXM1, EVA1A, UBE2QL1, FMN1, CXCL14, RP11-648F7.1, RASL12, ICAM2, MRVI1, AF186192.5, TLR3, CRYGD, AL512428.1, IL10RA, RP11-531H8.1, BAI1, IL20RA, LBX1, RBMS3, ZNF423, GBP1P1, AIM1, RP11-297H3.4, CD79B, TNFSF10, NHS, LPAR4, PTGS2, RP11-94B19.5, CTD-2325B11.1, RP5-1139B12.2, ASS1P1, CDX1, SLC12A7, C10orf105, ISLR, ARHGDIB, TEPP, FMOD, DIO2, SLITRK2, SPTLC3, FABP4, SALL3, AC007464.1, AP001065.15, RP11-710C12.1, RP3-522D1.1, RP11-449D8.2, KLHL14, XPNPEP2, RP11-314P15.2, NRG1, USP26, RP11-3N2.13, WDR86-AS1, FUNDC2P2, AC104654.2, KRT79, DAB2, NPS, COL23A1, AC007750.5, RP11-148B18.4, MYOM3, RP11-805L22.3, SLC18A1, RP11-1101H11.1, LINC00621, RP11-449D8.5, AC073135.3, SLC14A1, MAT1A, RP4-798P15.3, FAIM3, HTR6, RGS1, AC093627.8, MAGEE2, BLNK, CES5A, RP11-426L16.8, ACY3, RP11-449D8.1, TRIM29, NTM, DOK5, SHROOM4, CTD-2521M24.5, TLR5, RP11-459E5.1, MRGPRF, LGR5, AC100802.3, AF186192.1, VSNL1, ZNF831, PDCD1LG2, PRUNE2, GH1, RP11-123K19.1, COL26A1, AMER2-AS1, SNAP25-AS1, ROR1, EID3, RP11-728E14.3, ENPP6, KIAA1462, ASPA, RP11-335O13.8, AP000344.4, UNC13A, SAMD9L, FOXF1, ETNPPL, RP1-71H24.1, RP5-944M2.3, TFAP2B, RAD51AP2, KCNJ12, SUSD3, AP003025.2, BX119917.1, SLC24A2, RP11-94L15.2, GRM3, RP11-356N1.2, RP5-1104E15.6, KANK4, CTB-78F1.1, EMILIN1, LINC00277, IFITM1, RP11-214C8.2, TCEAL5, LONRF2, VAT1L, WNT5A-AS1, HLA-DPB1, GPR111, RP11-453M23.1, NUS1P2, RP11-430H10.1, ADRA1A, LINC00681, FAM65B, PCDP1, TMEM47, CTD-3247F14.2, RP11-429A20.4, TNFAIP3, GPR12, CTD-2377D24.6, AQP7P2, GBP5, TUSC5, CTD-3162L10.4, CTD-2151A2.1, DDR2, ATP8A1, CDK15, PLVAP, CD70, LTBP1, SVIL, NR1H4, RP11-392O17.1, CYP1B1, THBD, VSTM2A, FGF13, CTSS, RP5-1119D9.4, CTD-2023N9.1, DOCK10, GADL1, RP1-45C12.1, MAGEC2, SLC9A9, BST2, KCNG3, RRN3P2, RP1-290I10.5, KIF19, SH2D4A, NUDT7, CLIC5, SOD3, IL1RAPL2, RP5-1043L13.1, PTPRVP, XAF1, AC138623.1, AC004562.1, OASL, SPINT1, MYOG, AC007106.1, RP11-875O11.3, MAP7D2, PLXDC2, ERG, ARHGAP24, WWC1, CTD-2521M24.9,

GAREM, ART5, LINC00284, LAMP5, POU2F3, COL6A3, ZNF521, NEUROD4, RP11-845M18.7, CHMP4C, CHRM3, FAM189A2, GGTA1P, SNAI1, HLX, NKX2-4, COL12A1, AC099344.2, RPL29P19, PTPRO, LINC00536, HERC6, RP11-342A23.1, BATF2, ANGPTL7, SLCO2B1, MTUS2, NCKAP1L, GBP6, ADAMTS19, KCNH5, C1orf222, SCUBE1, TTC40, HBE1, HOXB13, SERPINI1, GPR115, TMOD1, IL22RA1, CD55, LAYN, SCN2B, KIF21B, RP1-244F24.1, ATP9A, RP11-429A20.3, LY6E, RP11-886D15.1, EBF1, HTRA3, MCHR1, VSIG1, RNF135, SYNM, NTN1, CTD-3157E16.1, PINLYP, AKR1C7P, RP11-363E6.3, KCNN3, RP3-467K16.2, OR5B19P, COX6B1P4, BPIFB1, HOXB2, GABRG1, DPYD, C3orf36, LRRC3C, RBP1, FAM153A, XX-CR54.3, TNFRSF14, RP11-328N19.1, AMPH, GRAMD1B, PLCH1, RP11-431J24.2, ADRB1, GALNT5, PPARG, NRXN3, TMEM244, KLHL41, RP11-538I12.3, AC011294.3, PDGFB, ALOX12B, TMC1, IL24, IL1R1, PARP10, BIN1, BTBD17, MGLL, CLDN10-AS1, PDGFRB, GVINP1, ALCAM, NT5C1A, CASQ2, GFRA2, ZNF334, RBP4, RP11-104L21.2, SPOCK2, CCDC144NL, RP11-565P22.2, MYLIP, LINC00173, LYN, FOLR1, LOXL1-AS1, RP4-701O16.5, SNPH, PRRG4, MIR199A2, MAATS1, FAM26F, CMAHP, COL6A2, PCP4L1, TRIM34, EGR3, OVCH1, DSCAML1, HMGB1P7, FAM110B, CCK, ARHGAP25, COL14A1, AC013275.2, CARD6, SCUBE2, DHRS2, CSF1R, DCHS2, ZMYND15, SECTM1, PCDH18, RP11-478K15.6, RP11-213G2.2, PCED1B, RP11-283I3.2, RP11-332H18.5, RP11-1049A21.2, RP11-492A10.1, TSPAN19, CCDC110, IL15RA, UNC5B-AS1, SORCS2, NME5, RP11-305L7.6, RP11-89K11.1, RP11-1277A3.2, SARDH, RP11-179A16.2, RP11-45A16.4, IGFBP5, WIPF3, GYG2, CLMP, ALDH1L1, CTD-2207P18.1, ANKEF1, PLAC9, SLC26A4-AS1, ESRRG, SEMA7A, OAF, MAGEB17, GRPR, CTB-134H23.3, KIF12, RP11-262D11.2, SLC29A3, AC006445.7, RP11-145M4.3, RP11-930P14.2, RP11-849I19.1, STOX2, JPH2, LBH, RP11-157J24.1, GUCY2D, PTPN3, TWIST2, RP11-363E6.4, KIAA1755, AC010984.3, SAMD9, RP11-132E11.2, RN7SL417P, DNAJC12, CYP1B1-AS1, PHLDA2, GOLGA6L7P, FCN2, RP11-429P3.3, MMEL1, AC006946.16, MATN3, FGFR2, GS1-304P7.1, POU2AF1, TRIM22, NMUR1, RP11-435I10.3, RP11-301L8.2, KLKP1, PAX9, SV2C, CECR2, IFNB1, AC009110.1, HLA-AS1, RP11-514D23.3, NFE2, SEC16B, HOXD1, FOXD2, C3orf67, ZSWIM5, MICAL2, EVC2, KIAA0040, ABCA8, ETV7, GPNMB, WDR86, AC018865.8, PCA3, HOXC-AS5, PTPRT, RP1-71H24.4, HTATIP2, BZRAP1, AP001471.1, RP11-292E2.2, CMTM8, LRCOL1, RP11-94C24.8, RP11-553P9.3, RTN4RL2, EGFL7, RP11-268G13.1, LAMC3, POMC, RP11-75C9.1, AC093627.7, ALG1L15P, RP11-54D18.2, RP11-279O9.4, CACNG5, PPEF1, RP11-81H14.2,

C3orf80, HEYL, CTD-2562J17.7, RP13-977J11.8, NTNG2, RP11-119J18.1, GALNT6, IRF6, RP11-104L21.3, HLA-U, IFNLR1, RP11-295M18.2, MISP, NYNRIN, RP11-445P17.8, DDX25, CDR1, LNX1, AC106874.1, RP11-536C10.10, SCN7A, NR1I2, MAP3K7CL, LINC00639, RP11-221N13.3, TNNI2, SPON1, LIPC, TDGF1, RUNX2, CD84, CLEC3A, IFIT3, CRISPLD1, TESC, C2CD4A, PTPRJ, AC079613.1, ANGPT4, CD248, FAP, AC096669.1, MAP3K8, PROSER2, CREB3L3, DISP2, RP11-766F14.2, BTBD11, RP11-809C9.2, ICOSLG, INSC, SDC2, RFPL2, CTD-2526M8.2, INA, CCDC3, C14orf180, HIST1H2APS3, LINC00648, ADIRF-AS1, DOC2GP, FZD9, PMEPA1, AC092162.1, DMD, HEY1, TAL1, CYB5R1, RIPPLY3, RP11-1042B17.5, TRAJ6, C16orf74, DNM3OS, MAFB, JPH4, RP11-676F20.1, PHF11, CECR1, RP11-521O16.1, LGI4, CD24P4, TINAGL1, RP11-428G5.7, NMRK1, STRA6, LOXL1, AC073636.1, HNF4A, HOXB6, ST8SIA4, RP11-1365D11.1, RP11-10N16.2, RYR3, GMPR, SNAP25, GSN, RP11-295M18.6, QPCT, TMEM229B, TSPEAR-AS1, HSPB8, CTD-2521M24.4, RP11-162J8.3, HERC5, IFIH1, ACE2, CTD-2054N24.2, WNT1, NPR1, RN7SL405P, HOXB-AS1, HIST1H1T, RP11-70C1.1, RP11-770E5.3, C8orf22, RP11-671M22.6, OAS2, DSE, CTSO, GRIK4, BLACE, MYH6, C9orf135-AS1, LINC00620, UNC93B7, AC096574.5, SLC7A2, EMP1, GRIA1, FAM225A, SLC26A7, PTGES, MUSK, IFIT1B, KLHDC7A, PPP4R1L, RP5-1139I1.1, CTC-501O10.1, AL845321.1, RP11-344E13.3, RP11-740P5.3, 45175, SCN4B, APCDD1L, TAF4B, RP11-679B19.2, KB-1184D12.1, DKK3, ALDH3A1, RP5-1107A17.4, PKDCC, AC011239.1, FXYD7, VDR, PRRX1, CTD-3065J16.6, TRIM31, CXCL9, KCNK15, CCND1, CTD-2313J23.1, AC096669.2, ZFP92, SCARA5, LPPR3, AC087393.1, TMTC1, RN7SKP88, SLC10A6, PRKAG2-AS1, GPR64, RP1-93H18.7, SERPINE2, RP11-325F22.2, CLDN2, CTA-268H5.14, CTD-3074O7.7, MX1, SLC51B, CTD-3076O17.1, TMCO4, RP11-416N2.4, RP11-16P20.3, DNM1P51, SPARCL1, RP3-395M20.8, PRRX2, RP5-1063M23.1, RP11-141E13.1, LINC00957, RP11-76E17.4, OR2B4P, SLC4A10, HS3ST3B1, RP5-1185H19.2, RP11-540O11.7, CCL5, FAM182A, TPD52L1, SVEP1, CXCR4, DNAH2, RARRES3, FBLL1, RFPL4B, FILIP1, SALL4, LEKR1, RP11-655M14.13, RP11-296L22.7, CLMN, AC079776.3, EPHA7, GSG1L, MEOX1, PANX2, RP11-507M3.1, GSN-AS1, RP11-141A19.1, PAX5, GRHL3, NTSR2, TRAJ26, RP11-720N19.1, MAGI1-IT1, CTC-548K16.1, IER3, RP11-128M1.1, CD274, SLC27A3, BTF3P2, LRRC3B, FAM83H, MYBPH, TENM1, CATSPER1, CCR4, HDAC1P1, FAM114A1, RNF182, RP11-923I11.5, AKR1B15, ATP8B1, SFRP2, SLC7A3, OR1L3, RP11-441F2.2, OBP2A, RP11-467K18.2, FUT5, RP4-737E23.2, RP11-175K6.1, TRPC7, SEMA3D, RASEF,

HOXC13, RP11-739G5.1, ARRB1, SLC37A1, HEPH, THRB-AS1, COL9A1, PLSCR4, RHAG, PARD3B, RP11-348F1.3, ALDH5A1, RP11-508N12.3, LINC00943, LRRC14B, SLC39A8, LYPD6B, RP11-340E6.1, RP11-357H14.17, AFF2, ARHGAP22, KCNF1, C5orf49, SRC, CACNA1C-AS1, RBPMS2, RP11-263F14.3, RP11-489D6.2, ANGPT2, RGPD1, LINC00355, RP11-230G5.2, PI3, KLF5, STAB1, FKBP6, CTC-260E6.9, CD7, RP11-115D19.1, SYN1, HLA-K, INPP5D, RNASEL, CRIP3, ZNF280A, SNX25P1, RP11-184M15.1, DGKE, TMSB4XP3, DRD5P2, MGC4294, LINC00643, AC019068.2, ZNF214, TRPV3, HOXB5, LYPD5, TNFSF9, SLC25A48, TLR6, ACTL6B, HCP5, RP11-6O2.4, PAK7, CTB-60B18.12, RP11-609L3.1, MT1G, EFTUD1P1, CTD-2547L24.3, FAM19A3, CTB-178M22.2, DDX60, AC139768.1, CNPY1, CHD3, CECR7, MTL5, Z83001.1, CALML3-AS1, APOE, NECAB2, AKR1B10, ABCA6, BDKRB2, PTGIR, RP11-344B5.4, CTD-2247C11.1, KCNA7, LINC00589, SMCO4, SEMA3F, ARHGDIG, BMP2, STOM, ANKS1B, RP11-401P9.5, FZD7, FAM90A27P, XXbac-BPG248L24.12, KCNC4, RP11-187O7.3, CD8A, APOL1, GCNT1, FUCA1P1, HLA-G, ITGA3, AC010149.4, ASTN2, AP001625.6, RP11-680H20.2, RP11-445O16.2, TUBA4A, ENG, RP11-718B12.2, SULT1A1, GAS2L2, NEURL, PARP15, RP11-295G24.5, MRVI1-AS1, AC009518.4, ADAM28, CXCL10, TSHR, NRXN1, ACSL5, RP11-112L6.4, FOXD2-AS1, PABPC3, CCL26, ZNF29P, DYSF, HOXB3, UBA7, TNXB, ABCD2, ALPL, FAIM2, NHSL2, RP11-277P12.10, TRANK1, ANK1, RINL, OR2H1, SP140, CTSF, GPR124, GPA33, TRAJ25, SEPP1, PAPPA-AS2, RP11-93L9.1, ERC2, PCDHB5, THEG, STRC, RP11-273G15.2, ALOXE3, FAM163A, SMAD6, NPR2, CMPK2, C10orf54, KB-1639H6.4, RP11-366L20.2, RP11-718B12.5, CCDC120, TMEM173, LAMC2, COL1A2, NKX3-1, RP11-785G17.1, RP11-347E10.1, SMTNL1, TMEM156, SYNPO2, RASSF3, RP11-844P9.2, RP11-14C22.6, ELMOD1, UCN2, FBXO39, SCGB1D2, CYP1A1, ITGA9, KIAA1377, RP11-923I11.7, RP11-718B12.1, CDH4, RP11-561C5.7, RP11-196G18.3, RP11-532F6.3, RP11-867G2.8, TPBG, RP11-88I21.1, TSPAN16, LGALS2, RP11-408O19.3, RP3-473L9.4, RN7SL834P, AQP7P4, MAGEB6, AQP7P1, SNHG18, FUT6, BTBD3, PPIEL, CRYGN, IFI35, CTD-2609K8.3, CCDC141, RP11-350G8.5, SLCO2A1, CSAG4, HES5, FTLP12, BMX, ISM2, LRIT2, AC079776.2, TEX15, CLU, RP5-968J1.1, COPZ2, RP11-362F19.1, MIR3151, HCG4P7, PTH1R, SHC1P2, TSPAN1, IFITM3, KLHL29, AC109309.4, RP1-232P20.1, RP11-64P14.7, PRRT1, MYRF, PRR15, KCNK3, HMGN2P40, ADCY5, PADI4, CTD-3154N5.1, C11orf96, AC010745.1, RHBDL2, RNA5SP259, COL25A1, MDS2, XIRP1, MIR663A, RP11-19E11.1, RNF212, EXD1, FENDRR, CPLX2, HEPACAM, TMEM27, IFIT2, ID3,

PADI2, ADAMTSL3, RP11-381O7.3, LAMB1, SOX18, SLC5A4, CYP4X1, KY, RP13-631K18.2, AC087380.14, ZNF90, SEL1L3, C2orf50, RP11-116O18.3, ZNF718, FAM226B, AC007000.10, SGSM1, OR2B6, RDH12, OR1H1P, GABRR1, CORO6, RP11-21L19.1, PIP5K1B, BTN3A3, CD1D, FOXH1, ADAM33, RAB3IL1, SLC25A21, RP11-973H7.1, RPH3A, AFAP1L2, VWA1, RP11-731K22.1, SIMC1, ESYT3, SERP2, MAGEA12, CREG1, PHOSPHO1, RP11-276E17.2, HIST2H2AA4, SIX1, GNAS-AS1, PARP8, HLA-DOB, AC072052.7, LRRC37A5P, FMO5, HS3ST5, IGF2, RGPD2, RP11-240M16.1, LTBP2, CTC-340D7.1, C2orf62, CSPG4P5, ADRA2A, CTD-2311M21.2, RP3-510L9.1, CPQ, PRR15L, RP4-794H19.4, KIAA1199, ZNF204P, CT64, SLURP1, PALM, AC079776.1, LURAP1, RP11-145M4.2, FLNC, RP11-95P13.1, MYO6, RAB3D, ATRNL1, RP11-277B15.2, RP11-363N22.2, C19orf35, SP140L, AC011747.7, RP13-977J11.2, CTB-35F21.4, PARP14, FGF12, CPNE5, NAV3, IFI6, RP11-64K12.10, SULT2B1, GS1-304P7.2, ABCG4, COL15A1, FGF13-AS1, GBP2, C15orf37, TRIM55, TLR10, ERP27, PLCE1-AS1, HELZ2, RP11-680F20.11, SRSF12, RP11-757G1.6, BMS1P17, CTD-2611K5.5, RP11-356I2.1, RAET1E, DUOXA1, SNTG1, MSTN, TTBK1, TMEM216, INTS4L1, FMO4, LLNLR-299G3.1, PRCP, UNC93B3, PARK2, RP11-374A4.1, SATB1, RP11-268P4.5, RP11-355I22.2, RP1-79C4.4, RP11-266O8.1, MICALCL, ADAMTS16, RP11-4B14.3, RP1-150O5.3, AF196972.9, RP11-3L10.3, CCL2, SYT12, FAM183A, ITPR1, ADAMTS10, NRK, EMID1, COLCA2, SLC2A10, RP11-579D7.8, PRSS56, GSTT1, GUCY2C, EDARADD, CILP, SLC13A3, FAM66A, MAOA, RP11-196E1.3, FABP9, SLC43A3, ADCY8, RNF152, UBE2E2, ROBO4, LINC00877, MTNR1B, CD83, ZNF572, MYH7, CTD-2267D19.3, FMO3, THSD4, MTTP, SPESP1, RUSC2, TEX11, MGAT4A, C15orf65, THSD7B, NBL1, CYB5R2, KCND1, SYNGR4, KIRREL3-AS1, GIMAP1, RP1-90G24.11, FBLN2, RASGRP2, PNMA3, PARP12, MYO3B, HOXB-AS2, AC016735.2, AARD, ULK4P3, ABCC11, PABPC1P7, RP11-2G1.1, HSH2D, SIGLEC15, RP11-407A16.1, GOLGA8M, CXCL11, APOBEC3F, PTK2B, CTB-60B18.18, VASN, RP11-507K2.3, SCML2P2, AL121578.2, RPL6P7, RP1-90G24.10, OGFRL1, POU5F1, ARG1, STX1B, RP3-416H24.1, FAM46B, GRIN2A, RN7SL614P, GPLD1, PCOLCE2, HOXD-AS1, PDE1B, HCLS1, SLCO3A1, GPR68, NBPF3, RP11-347C12.3, RP1-290F12.3, RN7SL316P, AC079305.11, RP11-328P23.2, RS1, RP11-255E6.5, RP11-561C5.5, MLKL, KCTD16, RP11-1000B6.3, AC083949.1, LIPH, AP001092.4, OR13E1P, TNFSF12, CNRIP1, RP11-806K15.1, OR5BK1P, RP11-21A7A.3, DNM1P47, TBC1D10C, ABLIM2, HOXB-AS3, FABP3, RANBP3L, CSMD1, EFHD1, HIF3A, RP11-266L9.6, GOLGA8A, CNNM1, LAMA4, RP11-678G14.3, NKG7, TIE1,

SCPEP1, RIC3, CENPV, CASP4, SPINT2, RHOXF1, HOPX, RPL7AP33, MFAP2, RP1-69M21.2, NRP1, CCDC170, ADCY7, DHH, AL021917.1, RP11-402L6.1, AC007000.12, COX6A2, CARM1P1, RP11-326C3.2, PXDNL, HOTAIR, COL17A1, SLC22A15, CTC-260E6.11, THEMIS2, RP11-403I13.5, AC109333.10, MAP6, RP11-332L8.1, LINC00487, CDSN, LINC00202-1, FAM13A, RGS4, IQGAP2, SRL, RP11-396C23.2, FRG1B, OSTN, UMODL1, FOXP2, RP11-554A11.8, AF186192.6, HSPD1P7, GOLGA8S, IFITM10, CTD-3145H4.1, COCH, LRRC43, KCNC1, NFE4, PGM5, AP001626.2, RP1-86C11.7, MMP28, PRKCQ, PROCR, HLA-B, HIST1H2BD, RP11-167H9.4, LOXL2, MAFA, RP1-161N10.1, CTC-329D1.3, RP11-23J9.4, ZNF492, RP11-70P17.1, AHNAK, AC140481.1, LDLRAD2, GPR1, ACSL6, FOXC1, C8orf31, TRIM21, OR2W6P, RP11-394I13.1, HERC2P3, GPC3, SLC14A2, RP11-46I8.3, RP11-369K16.1, SERPINF1, IZUMO4, ZNF890P, SLC26A9, CNTN4, PRX, SLC22A5, RP11-15B24.5, IFIT1, AC109642.1, RP11-227D2.3, LGALS12, RP11-1191J2.5, PCOLCE, RP11-495P10.3, MAGED1, GBX1, RP11-182J1.15, CTC-360J11.4, CTD-2534I21.9, TRPV2, SP100, RP11-31L23.3, RP11-297M9.2, EHD3, LDHAL6EP, RP1-313L4.3, DLX5, SNUPN, HOXB9, IL11RA, IMPA1P, CTD-2227E11.1, RASGRP1, SYT6, TMEM63C, WI2-2334D6.1, AP000330.8, RP11-177H13.2, RP11-472N13.2, C16orf96, CXCL16, RP11-175B12.2, RP11-247C2.2, GFI1, DDX58, RP11-110I1.12, RASD2, MYO15A, RP11-157J24.2, AC007391.2, RP11-492E3.2, IFITM5, COL5A1-AS1, AC145343.2, PPAPDC1A, DENND1C, CES4A, AGT, KCNG2, ADORA2B, AC104699.1, CD300A, CD22, RP11-89M16.1, RP11-483H20.6, C1orf145, RNU6-446P, SEMA3C, RP11-326C3.11, CTD-2384A14.1, AC024704.2, DGKI, PRICKLE1, RP11-365O16.3, GOLGA6A, TMEM246, CCDC33, RP1-125I3.2, FCGBP, HOXB7, RP11-107N15.1, CTD-2555A7.2, RP13-314C10.5, CRYGA, STOML3, CASP12, RP11-485G7.5, ZIC2, CLCA2, PSCA, RP11-351N6.1, CDHR1, PRKCD, RP11-27N21.3, CTD-3092A11.2, ZNF663P, AC115115.3, TRIM50, RP11-462G12.1, ACBD4, CCDC175, PGM2L1, RP11-461O7.1, KLF9, COL6A1, RNU6-353P, RD3, SLFN13, AC010733.4, NKX2-5, NLRP6, TAGLN, VEPH1, SLC46A1, KLHDC9, FRMPD2, CTD-2589M5.4, CTD-2277K2.1, PEBP4, RHOD, KCNAB3, HRH2, NPM2, ECHDC2, GREB1, SLC6A17, FBXO25, RP11-356I2.4, RP11-112L6.3, RTN1, AHNAK2, LY75, RP11-319G9.5, DENND2D, CTD-2542C24.2, RP11-275O4.3, CNTFR, NFAM1, NAT1, RP5-1039K5.16, IMPG1, B3GAT2, KITLG, RP11-214O1.2, CDK18, DNASE2B, AC009950.2, MAPK8IP2, ZNF883, TBC1D3P5, AC096559.1, SMPD4P1, GOLGA8Q, TMEM221, DNAH3, CTC-367J11.1, BTN3A2, ARHGAP26-IT1, CASK, FOXP3, PIGR, TMEM88, RP11-345J13.1, RP11-145M4.1, WI2-81516E3.1,

RNF222, AC005559.3, RGCC, VAMP5, FXYD1, RP11-149I23.3, RP5-905G11.3, FGD5, AC024592.9, PLA2G16, SAMHD1, IRX3, BSND, FCGR1A, RP11-534L20.5, CST3, CYP19A1, PDIA2, TRNP1, AC062028.1, TEKT5, MIR298, SAR1P1, FREM1, LITAF, DACT1, RP11-387H17.4, TSPYL2, RP11-89B16.1, LST1, RP11-395D3.1, SLC40A1, USP18, RP11-150O12.6, FAM196A, ISG15, RP11-863K10.2, AP000783.1, RP1-249H1.4, RP11-1M18.1, RP11-445H22.3, RP11-578F21.9, ALDH2, TSC22D1, RNU6-548P, TOM1L1, OTOP2, SLFN11, CTB-60E11.9, TPTE2P1, CILP2, RP11-848P1.9, COL28A1, ARHGEF7-AS2, TMEM150C, SCN11A, ABCA4, EPHX2, PPP1R2P1, A3GALT2, GIPC2, MKRN7P, ETS2, CNTFR-AS1, NIPA2P4, ZAP70, CCDC87, AC007000.11, RP11-736K20.5, GDPD5, RP11-284F21.8, LRRD1, CREB3L1, RP5-915N17.3, CASP5, RP11-281O15.4, AC098973.2, ARHGAP26-AS1, MEIS3P1, RP11-315I14.2, KB-1460A1.2, DYNC2H1, CDC42EP3, AKAP7, RP11-770E5.2, AC118278.1, CCDC148, RP11-495P10.2, CAPG, PTPRD, COL4A6, XYLT1, RP11-25I15.3, PLCH1-AS1, RP11-413E6.1, FMO1, RP5-916L7.2, SP7, RP11-293M10.6, RP11-231G15.2, PSTPIP2, ARHGAP26, APOL6, FAM19A5, C9orf135, AC139100.3, MICF, APOH, SLC35D3, FZD4, BSPRY, GBGT1, OR52N4, RP11-480N24.4, NXF3, ATHL1, RP11-399D6.2, GPR18, RNU1-122P, SDR42E2, HSPA2, CXorf67, CTD-3074O7.2, DDO, KCNQ1, RP11-579D7.4, ZNF541, IL1R2, RP11-430B1.2, CTD-2517M22.14, RP11-368I23.3, CSGALNACT1, COL19A1, DAAM2, UPK2, CDC42EP5, CTD-2315E11.1, RASGRP3, NBPF2P, SMIM2-IT1, ALDH8A1, WHAMMP2, PHKA2-AS1, LINC00341, ID2, RP11-495P10.5, ADC, RP11-243A14.1, RP11-1180F24.1, MT-TT, HIST2H2BF, GIPR, CTD-3092A11.1, RP11-1094H24.4, SRGAP1, EIF4A1P12, RP11-657O9.1, CLDN10, SCN5A, AC015849.13, AL133493.2, CRB3, ZNF835, AMOTL1, CCT8P1, PCDHA1, C6orf223, NDRG4, CAMTA1-IT1, RP11-25K19.1, BTN3A1, NKX3-2, MLPH, RP11-68I3.10, SCTR, KCNK10, PCDHGA8, FLRT3, MAGED4, ADIG, SH3GL2, CTSH, GUSBP5, FAM228A, HIST1H4H, C9orf173, AC010987.6, ZNF454, PC, FAM71E1, KNDC1, AC009518.2, AC068831.3, RP11-21A7A.4, RP11-236F9.2, RP11-8H2.1, TMPRSS3, RP11-91J19.4, MTOR-AS1, EXOC3L4, PSORS1C2, HIST1H1D, USP43, PLA2G4C, RP11-451M19.3, RP11-535M15.2, GBX2, RP11-712L6.5, FOXD3, PATL2, STAC2, MR1, IL20RB-AS1, SEC14L4, RP11-158I9.5, RP11-54D18.3, HUNK-AS1, FCGR2B, RP11-326C3.14, CA4, RP11-81K2.1, LRRC8B, RP11-466G12.3, CCNB3, RRBP1, MAP1LC3A, C19orf45, RP13-152O15.5, CHST1, PDCD4-AS1, RP11-401P9.4, AC011816.1, WNT10B, RP1-56K13.5, CTD-3064H18.2, FAM180B, AK8, RP11-923I11.6, ZNF311, AC115522.3, RP11-753H16.5, ZNF618, SLC25A43, RP11-113K21.4,

GMCL1P1, IMMP2L, CDC37P1, RNF144B, RP11-279N8.1, AC005618.6, GPR152, RP5-1050E16.2, RYR1, HLA-DMB, DEFB109P1, ISM1, SLC43A1, ATP2B2, RP11-355I22.7, AC008073.9, ADORA1, RP11-247A12.2, FAM47E-STBD1, KLF8, TAPSAR1, DRD5P1, AC068282.3, FIBIN, C1orf226, PRSS42, KLHDC1, MKNK1-AS1, ULK4P2, KIF16B, RP11-508P1.2, CYP46A1, COL11A1, 45173, INSR, PITPNM1, RP11-227D13.1, RP11-569A11.1, NRG2, GRB7, RNU6-729P, AC073343.1, SELV, ONECUT2, INTS4L2, GRM6, LINC00882, OTOL1, PLCE1, PDYN, AC005532.5, ANKRD55, DIRC3, DYNLRB2, L29074.3, CES3, AC010092.1, RASSF5, PSORS1C1, AC010641.1, PKNOX2, KIAA1467, RP11-318M2.2, OVOL1, MIR3936, CTB-111H14.1, CATSPER2, SPARC, DENND3, RP11-179A10.1, MKI67IPP4, AP000351.8, GPR4, SOX30, RP11-426L16.9, RP11-678G14.2, DEPTOR, HCAR1, CMP21-97G8.2, CTC-241F20.3, NTF4, RNU6-1085P, CTD-2258A20.5, UBQLNL, MYBPC1, FAM181B, RP11-473M20.5, TNFRSF10D, TEX14, FAM66B, C2, ROCK1P1, SORCS1, SESN3, CCR1, DLX2, DTX1, TRAJ23, RP11-441F2.5, SULT1C2, ULK2, RP1-1J6.2, FUCA1, STXBP6, UBBP4, TBX15, C14orf132, TRPS1, PRPH2, RP11-819M15.1, CSF1, CXorf36, C6orf141, SPINK2, RDM1, FAM49A, AGBL2, ABCC8, RBM44, XXbac-BPG55C20.7, ATP5F1P5, DUSP13, YPEL4, AC034220.3, C17orf105, NGFR, CDON, CDH1, RPSAP52, CARNS1, ASAP3, RP5-882C2.2, BEAN1, LINC00594, ZNF354C, UTS2B, DLL4, CTC-510F12.4, TRIM5, CHRM3-AS2, RP4-764D2.1, DCDC2, KLHL7-AS1, C20orf166-AS1, CNN2P3, CTD-2623N2.5, RAB34, TJP2, TMEM59L, RP11-150D20.5, RP11-782C8.3, ZNF483, SCGB2B2, TNFRSF21, PRICKLE3, ID1, TRH, RP11-495O10.1, PYGM, RCN3, AC114788.2, SPTBN2, PLS1, SLC12A3, PPFIBP1, NFATC4, FAM185BP, CTNNA3, RP11-396C23.4, CD4, SLC37A2, AC003102.3, AC007386.4, RP11-864I4.4, RP11-439A17.9, MT1M, PLCXD2, TSSC2, RP1-179N16.6, PCDHB17, BACH1-IT2, EMP3, SDAD1P1, RP11-736K20.6, IQSEC3, CTD-2189E23.1, RP11-675F6.4, RP11-517P14.2, SYCE3, RP11-482M8.1, ROBO2, CACNA1E, GSDMD, PPP1R1B, RP11-43N5.1, RP11-753H16.3, ZNF391, SIRT4, FRMPD1, PRKAG2, NSD1, BTBD19, SLC2A4, RP11-182J1.16, RP11-1079K10.2, RGS18, ZIC5, PKD1L2, RP11-142A5.1, RIMS3, RP11-94C24.6, RP11-44M6.1, RNA5SP385, SIK3-IT1, FAT3, ADCY4, LPP-AS1, TMEM171, PTCHD4, RP11-757A13.1, FSIP1

726 downregulated DEGs

CEP55, RPL13P12, NETO2, RP4-575N6.2, EXO1, RP11-437L7.1, AC138655.6, CA14, EZR, CDHR2, LTBR, SNORD125, E2F7, RPL18A, ZWINT, CTD-2316B1.1, KCNRG, RPL10P3, SYCP2L, RP11-63N8.3, RNU6-1091P, ZBED3, ZNF552,

NPPA, NEO1, GSTT2B, RP11-116N8.1, TRAV30, RRM2, RPLP0, ARC, ENKUR, AC008278.2, CH17-12M21.1, RP11-436D10.3, ZNF331, RP11-282K24.1, RP3-395M20.2, LRRC34, RAB32, RP5-826L7.1, CEP152, ELOVL6, DRP2, CTC-529L17.1, KIF18B, PSD, MIR4673, EEF1A1P12, RP11-35O15.1, AOC1, GAPDHP63, RP11-215A21.2, XXbac-BPG170G13.32, SCARNA2, ZC3H12B, AC132008.1, TNNT1, RP1-127H14.3, RIIAD1, BAG1, TRAV41, RP11-175B9.3, RP5-1024G6.7, RN7SL535P, AL162151.3, SNORA47, CENPW, CTD-2044J15.2, NCKAP5, CDH5, AC007163.6, CDC45, AP000289.6, MT-TS1, NANOS3, PLEKHS1, RNU7-128P, RNF219-AS1, MCM5, PCDHB12, E2F2, IFITM4P, SPDYE2B, RPL18AP3, RP11-16E18.1, SDK1, RP11-375I20.6, RP11-224O19.2, RP11-556K13.1, PCDHA5, RP11-650L12.2, CTB-63M22.1, AC096677.1, NT5C3AP1, TBATA, RPS2P46, RP11-259K15.2, LRRIQ4, RP11-782C8.5, CTD-3088G3.4, SMARCA5-AS1, SLC46A3, ABHD17AP3, OPN3, AC016708.2, C22orf31, AC004471.9, RP11-737O24.5, PAPLN, RP11-453F18__B.1, HCCAT5, CSRP2, ST6GALNAC5, CTC-543D15.1, BPIFB4, PRH2, RP1-137D17.1, AC138123.2, ZNF219, RP11-157F20.3, RP11-288K12.1, CTD-2561B21.3, RP11-728K20.2, EIF4E1B, YBX1P1, FAM21B, AC018804.6, EPGN, CTD-2192J16.15, AC007773.1, RP11-524C21.1, BEX1, CAMK2A, AC004381.7, ASB9P1, RP11-15A1.3, GPRIN1, AVP, PCDHA2, DDC, RN7SL771P, ZMAT1, ARHGEF28, RPSAP15, EEF1A2, AC006042.8, KIAA0101, RP11-811P12.3, CTHRC1, PTPRB, RP11-227D13.4, C2orf88, GAPDHP40, CAMK2N1, RPLP1, SMIM3, BNIP3P1, API5P1, RP11-774D14.1, RP11-4C20.4, RPL37P1, RP11-159N11.3, ZNF101P2, RP5-1172A22.1, RP11-1024P17.1, IL36G, PDLIM1, RP11-613M5.2, NMB, RPL10, ANKRD10-IT1, RP11-936I5.1, MYCL, PTP4A3, HUS1B, SFRP4, CTA-85E5.10, ENC1, RP11-73C9.1, HIST1H1B, RP11-676F20.2, RP11-561N12.7, TUBB2B, FOXM1, TRAV27, CTC-467M3.1, RP13-672B3.5, RP11-84A1.1, SCARNA10, RP5-827C21.1, LINC00618, ACAP3, AC114803.3, RP5-1098D14.1, DNER, TMEM255B, RP11-417J8.6, PUSL1, RP11-150L8.4, FAM115C, RP11-820K3.3, RP11-36B15.1, AC011247.3, AC006946.15, RP11-328M4.2, HOTTIP, RP11-290L1.3, MDFIC, AC074289.1, RP11-594N15.3, RP11-751K21.1, RP11-114H23.3, RPS2, RPS3AP3, SNHG8, PLLP, WDR72, RP11-641D5.1, AC107021.1, CTD-2372A4.1, CR381653.1, RP1-155D22.1, HIST1H2BB, AC005795.1, AC083863.5, RP11-43A14.1, HERC2P10, RPS2P5, CTC-575D19.1, MSH4, MIR4786, RP11-480I12.5, RP11-815J4.7, ACN9, RPS2P55, AP000936.1, RP11-274H2.5, RP4-694A7.4, TRAV28, NAV2-AS5, FGD4, RNA5SP323, AXL, RP11-195E2.4, RP11-567O16.1, RP11-286H14.4, SCARB1, MIR146A, MYOC, ZNF850, CHRNA5, HSD11B2,

AL009178.1, AK3P3, RPSAP58, RP11-376O6.2, RP11-525G13.2, HIST3H2BA, RP4-555D20.4, RP11-204L24.2, RP11-432F4.2, TMEM190, RP11-187C18.3, RP1-128O3.6, PCDHB7, RP11-86H7.6, C2orf70, DRD2, SNORA84, IGHV1OR15-1, AC144449.1, CTC-563A5.4, RP4-676L2.1, TMEM74, VTRNA1-2, IGSF5, AC007375.1, RP11-145A3.1, RP11-699L21.2, FOXB1, PNPLA3, PLAU, CASC6, DNAH14, RP11-1007I13.4, TAS2R6, AC068538.4, GAS5, HIST1H2AJ, RP11-360L9.7, RP11-93O7.5, AC073072.5, SMARCE1P2, MIMT1, COX6B2, MFAP3L, TRDV2, ACTBP7, RNA5SP493, RP11-588H23.3, CTC-265N9.1, TNS3, PTENP1, MIR3648, IKZF2, AC009005.2, RP11-561B11.6, RP11-33A14.1, MAP1LC3C, RP3-395M20.3, TGIF2P1, PPIAP29, MYO18B, CLDN7, RP11-448P19.1, MND1, RNU1-2, RP11-285C1.2, FAM127B, PROS1, SCEL, MEGF11, RP11-779O18.2, PRRG1, RP11-708B6.2, AC003003.5, RP11-9E13.4, AP001046.5, RP11-758M4.1, IL6, TAGLN3, ARL5AP3, REG1A, SCNN1G, RP11-647P12.1, CTD-3148I10.1, LRRN4, RP11-125B21.2, GOLGA8UP, RPL21P132, PABPC1P4, GPR52, RP1-93I3.1, RP11-488C13.4, AC097713.2, BCHE, RP11-696N14.1, RPL12P37, TAS2R5, KMO, DISC1-IT1, BX470102.3, CASC8, SNORD3B-2, LINC00511, RP11-746P2.3, GP6, PLOD2, MACC1, IL1RAP, JAM2, MYB, IGHG2, AC004410.3, CTNND2, RP11-644C3.1, RP11-131L23.2, PNLIPRP3, RP1-125I3.4, SUN5, KIAA1658, AC009061.1, ADH6, TRAV33, RP11-678B3.2, MB21D2, RAD21L1, PCDHA4, KCNMB4, NCAM2, AC019221.4, FAM19A2, THEM4, SLC7A5, IGHG3, RP11-512H23.2, RP4-555D20.2, CCR9, RENBP, CDH19, NLGN4Y-AS1, AC012066.1, RP11-829H16.3, RP1-182D15.2, MAGEA8, SNORA80B, S100A2, KISS1R, CPLX1, SH3TC2, RP11-536O18.1, RP1-40E16.9, RP11-1275H24.1, MGST1, TRAM1L1, RP11-1275H24.3, UCHL1, ADAMTS14, DBNDD2, RPL29P14, EFNA5, RN7SKP230, ITGA1, RP11-849F2.4, CTD-2303H24.2, CA9, RFX4, S100A6, OPCML-IT1, RP11-698N11.2, SLC44A5, NANOGP4, RP11-17E2.2, GPC6, RP11-647O20.1, RP11-434O22.1, C10orf55, ARL4C, C19orf81, RNA5SP442, IGF1R, NAP1L3, RP11-693J15.5, CHADL, RP11-351J23.2, CYR61, CDH11, TMEM169, PI15, RP11-381N20.2, RP11-82L7.4, KRT18P63, RP11-525A16.4, RN7SL354P, AC010731.2, RP11-338O1.2, RP11-219J21.1, MEST, AMOTL2, AC021218.2, ZNF536, RP11-439E19.7, CLGN, RP11-867G23.10, RP1-97D16.1, RP11-390F4.2, AKT3, NHLH1, CTNNA2, KCNH7, PLA2G3, LPCAT2, GOLGA6L2, MROH2A, FRMD4A, AC061992.2, CTD-2049O4.1, RP11-713P17.3, RPS4XP6, CNN2, RP11-269F20.1, RP11-351I21.11, RN7SKP296, GLIPR1L1, RNU1-47P, RPL10P1, RP11-284F21.7, AC073109.2, RP13-870H17.3, S100A5, RP11-838N2.4, CTA-392C11.1, CLVS1, RP5-1178H5.2, RP11-185E8.1, RP11-65J3.14, RP11-390P2.2,

TCEB1P18, RNVU1-15, GNAI1, NPTX2, ECM1, ASB9, RP5-998N21.7, LOXL4, RP11-352D3.2, RP11-317P15.5, DGCR11, MAGEA10, CCDC144B, H2AFY2, CNR1, RP11-524C21.2, RP4-668G5.1, TMEM192, RPL32P31, CYP4F12, RP11-40F8.2, FGF11, RP11-106A1.3, RN7SK, GCOM2, PDZD2, PCAT1, RNU6-623P, RP11-212I21.4, GPR158, PCDHB11, S1PR4, HOXA11-AS, RP4-718D20.3, PHEX-AS1, RP11-227D13.2, AC093677.1, TBXAS1, DTHD1, RP11-907D1.1, RP11-351J23.1, LRRC4, MIR138-1, RP11-139K4.2, DPY19L2, RP11-15A1.2, POU5F1B, AC068535.3, RP11-429B14.4, RP11-218L14.4, INSM2, AC104777.4, VDAC3P1, B3GALT2, EPB41L4A, AC012360.6, RP11-567J24.4, MAGEA8-AS1, C10orf107, RP11-564C4.6, RP11-739N20.2, TERC, RP11-434I12.2, AC005754.8, TSC22D3, ACOT1, AC093668.1, GPR63, NAV2-IT1, HES6, LINC00896, RP11-681N23.1, RGS8, BARHL1, NOTCH4, RP5-857K21.4, CD200, NPY, AR, AC060834.2, RP11-539I5.1, NEU4, KCNJ8, RP11-881M11.1, AIF1L, RP1-170O19.14, FDPSP8, TMLHE-AS1, RP13-439H18.4, CREG2, KRT8P3, TMTC2, LRRTM3, AC005754.7, NDNF, MARVELD1, MID2, OTOGL, AL353791.1, MN1, ARPP21, PID1, RP11-348J24.2, MYO5B, RP11-85M11.2, PDLIM5, AGAP1-IT1, MIR378A, EPB41L4A-AS2, WBSCR17, RPS3AP44, PCDH10, SNCB, RYR2, COBL, DLL1, RP11-170M17.1, RP1-102K2.8, PCDHB3, LINC00928, TSPAN12, VCAN, HMX3, LINC00404, GRM8, GLDC, HIST1H3F, GPR126, DLG1-AS1, AC019118.2, USP53, RP11-1336O20.2, VN1R67P, TMPRSS11D, RP11-435B5.3, PKIA, NUTM2E, CTA-992D9.7, UCA1, RP11-309M7.1, CTC-529P8.1, ARHGAP15, RP11-523L20.1, RP11-214L19.1, SLC30A3, SERPINB8, CASC15, PNPT1P1, AC010145.4, AC097713.4, FHOD3, RPE65, RP5-1033K19.2, RP11-241F15.9, RP11-563N12.2, RP11-838N2.5, RP11-90K6.1, KCNH8, RDH10, NNMT, CXorf57, CADM2-AS1, RP11-405A12.1, ERBB4, RP11-807H7.2, ZDHHC2, OPCML, ARF1P2, MFNG, LIF, RP11-408A13.2, DACH1, RP11-706J10.3, SLITRK3, LINC00158, TLL2, MIAT, AC097713.3, TEKT4P2, CTA-796E4.3, LINC00925, RP11-380P13.2, LINC00659, CTD-2245F17.3, SVIP, RP3-404K8.2, LPHN3, ACKR3, CTD-2023N9.3, SLC35F3, HPSE, RP11-379L18.1, RP11-390F4.6, LINC00403, PTCHD1, KRT8P48, CFI, UGT8, RP11-402J6.1, RP11-282I1.1, SEZ6L, CTA-796E4.4, EPHA5, TRBV26OR9-2, RP11-27G24.1, YBX3, RP11-986E7.7, DOK6, GLDCP1, RP11-807H7.1, RP11-52L5.6, CDH10, NOL4, AC005537.2, POU4F1, LYPD1, RP5-1029K10.4, MECOM, KCNIP1, ASXL3, GJB1, RP4-724E13.2, VENTXP4, FUT9, RP11-97N19.2, RP11-592B15.3, RP11-410K21.2, TRAV38-1, RP11-404J23.1, S100A9, RP11-384P7.7, RP11-419C19.2, S100A8, RP11-280G9.1, RP11-509A17.3, LPHN2, CADM2, RP11-603B24.1

Supplementary Table 2 **123 DEPs in CHI3L1 OE proteomics data.**

90 upregulated DEPs
CHI3L1, OPTN, SPARCL1, IQGAP2, CHPF, B3GALTL, RIC8A, TAP1, UBXN1, PRCP, WASF1, EEA1, MYH10, EPB41L5, SWAP70, LGALSL, PGM2L1, TMTC3, STXBP1, NOS1AP, IFIT1, BIN1, TANC1, FABP3, CPNE2, ITPR1, ARMT1, FKBP2, CPQ, STIM1, SDR39U1, CRELD1, LIMS1, IGFBP2, ATP1B2, UBE2L6, HIST1H2AC;HIST3H2A;HIST1H2AB, COMMD9, SERPINH1, CUTA, ACTR1B, KIFAP3, CST3, OXR1, LLGL1, RALB, SPARC, SPATS2L, GGACT, EMILIN1, FABP5, SYNM, ELAC2, RFX5, KLC4, ACSF2, C11orf73, STOM, VSNL1, PSMB9, HSF1, IFI35, FKBP7, CPT2, DNMT3A, GYS1, TAP2, PALM, PTPRG, SNRPD1, SLC27A1, FKBP10, OSBPL6, ADRBK1, PDIA3, GTF2E2, WBSCR16, CASK, SSR1, ARMCX2, MINK1, C8orf82, TRMT1L, LAMC1, MPRIP, HIBCH, SELH, B2M, DAK, INPPL1
33 downregulated DEPs
WDFY1, NRD1, CXADR, DPY19L1, TMX2, TFRC, MPP6, ADO, ITGB8, ENDOG, MARCKSL1, MZT2A;MZT2B, PLXNB2, EEF1A2, CDK1, COX5A, LACTB, SMARCAD1, RB1, PBXIP1, S100A6, CCDC97, HN1L, STAM2, CKMT1A, RNASEH2A, TIMM13, NCSTN, SOGA3, BCCIP, TUBB3, LIMD1, UCHL1

Supplementary Table 3 **152 DPPs in CHI3L1 OE phosphoproteomics data.**

129 upregulated DPPs
AQP4_S285_1, GAP43_S41_1, MAPT_T720_3, AHNAK_S4995_1, SRRM1_S715_3, DCLK2_S308_1, EPB41L1_S510_1, DPYSL5_S532_1, TMPO_S156_2, BCLAF1_S422_1, NDRG3_S331_1, MAP2_T1619_1, CTTN_S11_1, NHSL2_S801_1, DIP2C_S89_1, STIM1_S519_1, SLC4A4_S257_2, SLC4A4_T254_2, ITPR1_S1598_1, PTPN11_S591_1, GAP43_S154_2, NHSL2_S576_1, CPT1B_T745_1, CTNND1_T869_1, PSIP1_T115_1, MAP7D1_S116_2, LIN37_S138_1, BRD9_S56_2, DBNL_S232_1, NHSL2_S1214_1, CTNND1_S230_1, NHSL2_S1072_1, SUN2_S12_1, SPATS2L_S526_1, CAMK2D_T287_2, LMNB1_S408_2, BRAF_S365_1, SPATS2L_S531_1, STIM1_S401_1, EML4_S171_1, PICALM_S16_1, URI1_S418_1, STIM1_S521_1, CXCR4_S339_1, PPP1R12C_T560_1, PLEC_S4386_2, GAB2_S480_1, STIM1_S668_1, MAPK3_T202_1, GAB2_S264_1, AHNAK_S5400_1, EPS8_S685_1, DIP2B_S100_1, DIP2A_S94_1, CTTN_S150_1, EIF4G1_S1077_1, EIF3A_S1262_1, LIMA1_S369_3, LIMA1_S365_3, SRGAP2_S1027_1, LEMD3_S261_1, PRKAR2A_T54_1,

PALM_S124_1, DPYSL5_S536_2, CABLES2_S130_1, LIMA1_S15_1, SRGAP2_S1013_1, OXR1_S16_1, NFATC2IP_S338_1, 44448_S332_1, EPHA3_S968_1, LIMA1_S266_1, CHN1_T192_1, RANBP2_S1456_1, CDC37L1_S88_1, TOP1_S10_1, SPATS2L_S467_1, ENPP7_S245_3, ENPP7_T244_3, ENPP7_T250_3, ENPP7_T251_3, SIPA1L1_S161_1, MYH10_S1935_1, EPB41L5_S348_1, RAPH1_S980_1, APC_S2449_2, APC_T2442_2, DEK_S71_1, ARPIN_S2_1, NOP16_S16_1, CASP7_S47_1, BCLAF1_S531_2, BCLAF1_S525_2, ACIN1_S240_1, CEP170_S1165_2, MTSS1L_S634_1, DBNL_T291_1, RGL2_S619_1, VIM_S419_1, SRSF12_S219_3, SRSF12_S223_3, SRSF12_S227_3, MTSS1L_T391_1, ETV6_S439_1, ARHGAP21_S495_1, ZNF521_S605_1, RAB1A;RAB1B_T75_1, ARFGAP2_S240_1, LSM14A_S374_1, VAPB_S156_2, APBB1_S517_1, CDK14_S24_1, CCNY_S21_1, NOL4L_S295_1, CSDE1_T761_1, CEP170;PLIN5_S1160_2, COIL_T122_1, KIRREL_S574_1, TOM1L1_S323_1, CDC42BPB_S1686_2, CDC42BPB_S1690_2, ADCY9_S1257_1, MMP14_S577_1, KIF13B_S1410_1, MPRIP_T542_1, AGAP1_T836_1, ZNF521_S273_1, PSMG1_T18_1, KLC2_S610_1
23 downregulated DPPs
SPTAN1_S33_1, SRRM1_S560_1, RBM15_S257_2, RBM15_S259_2, PPAN_S238_1, SIPA1L1_S1255_1, ZEB2_S731_1, GBF1_S352_1, SPEN_S1222_1, CCNL1_S335_2, MCC_S294_1, TCOF1_S1111_1, FRMD4A_S604_1, LSM14A_S183_2, LSM14A_S192_2, BAZ1B_S161_1, MSL3_S400_1, RTN4_S184_2, TDP1_T496_2, SCG2_S532_2, SRP72_S621_2, REPS1_S482_1, CENPA_T21_2